Response functions, trading strategies, and random matrices: Analysis of large fluctuations and correlations in stock price diffusion

I n a u g u r a l - D i s s e r t a t i o n

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Philipp Weber
aus Bonn

Köln

2007

Berichterstatter: Priv.-Doz. Dr. Bernd Rosenow
Prof. Dr. Dietrich Stauffer

Tag der mündlichen Prüfung: 22. März 2007

Abstract.— In this thesis, we analyze and explain various properties of stock price changes.

The change of a stock price in a given time interval is composed of many price changes resulting from single trades. Thus, the up and down movements of a stock price can be seen as analogous to the classic diffusion of a particle: if the particle moves due to random collisions with other particles, the displacement after some time is determined by the sum of the displacements between the collisions.

Under certain conditions, the values of such sums are Gaussian distributed. In contrast, extreme price movements such as those of "Black Monday" in 1987, when the S&P500 index fell by about 20% within one day, are so much larger than ordinary price movements that they cannot be accounted for by a Gaussian distribution. One could classify such events as "outliers" that reflect an abnormal market behavior. However, our empirical analysis reveals self-similar features in the time series of price changes, meaning that price changes exhibit similar characteristics on many scales. In particular, a huge price change induces a series of large price changes whose rate decreases over the following months. In a similar way, some of these subsequent large price changes themselves induce further series of intermediate price changes in the following days. Hence, the mechanisms connected to huge price changes seem to be similar for smaller price changes, raising the possibility that these same mechanisms might also underlie ordinary price movements. This picture is supported by the widely accepted finding that the tail of the distribution of stock returns, i.e. changes of the logarithm of the stock price, follows a power law that describes intermediate returns as well as extreme events.

Though extreme returns seem to be "ordinary" in the sense that they are connected to the same mechanisms as smaller returns, it is still an open question how returns can occur that are much larger than can be accounted for by a Gaussian distribution. In fixed time intervals, where the large price movements described above take place, the return is determined by two factors: the number of trades in the respective interval and the magnitude of the returns due to single trades (tick returns). In order to better distinguish between these two effects, we focus on intervals with a fixed number of trades, rather than on intervals defined by their actual length in units of time. Interestingly, also here we find unusually large returns, resulting from the concurrence of two things: (i) in the respective interval, the average tick return is large and (ii) most trades change the price in the same direction. We show that a statistical model incorporating the average tick return and the direction of tick returns can reproduce the

distribution of stock returns in the studied intervals.

While this analysis explains in detail *how* large stock returns are composed, we examine in a further study *why* these strong returns occur. It is a reasonable assumption that, besides the influence of news, prices change in response to an imbalance between supply and demand. This imbalance can be quantified by volume imbalance, defined as the difference between the volume (number of shares) of buy and sell orders in a given time interval. On a given volume imbalance, the stock price reacts with a price change that is determined by the price impact function. We reconstruct this function in each time interval from data containing information about all orders present in the market. Here, we show that the time-varying slope of the price impact function is responsible for very large returns. Though in each time interval the price moves due to the volume imbalance, extremely large returns occur only when the price impact function is steeper than average.

If prices change in response to trades, there seems to be a paradox: the signs of orders, indicating whether it is a buy or a sell order, are long-term correlated, whereas the returns resulting from the execution of these orders exhibit only short-term correlations with a characteristic time of a few minutes. In order to understand this paradox, we model trading strategies and show that uncorrelated stock price changes appear naturally as soon as someone uses the correlations in the orders to make profit.

After studying time correlations in the returns, we also investigate a tool that can be used to analyze cross-correlations between finite time series. Since their length is limited, even uncorrelated time series exhibit spurious cross-correlations resulting from random co-movements that do not reflect the real interactions. We show that a hypothesis test based on random matrix theory can distinguish spurious correlations from real correlations, which we demonstrate using numerical simulations.

Zusammenfassung.— Diese Dissertation untersucht und erklärt verschiedene Eigenschaften von Aktienkursänderungen.

Die Gesamtänderung eines Aktienkurses in einer gewissen Zeit setzt sich aus vielen Preisänderungen zusammen, die durch einzelne Transaktionen hervorgerufen werden. Somit können die Auf- und Abbewegungen eines Aktienkurses mit der klassischen Diffusion eines Teilchens verglichen werden: Stöße mit anderen Teilchen führen hier zu einer zufälligen Bewegung, wobei die nach einer gewissen Zeit zurückgelegte Strecke durch die Summe der Strecken zwischen den einzelnen Stößen bestimmt ist.

Unter bestimmten Voraussetzungen sind die Werte einer solchen Summe gaußverteilt. Im Gegensatz dazu stehen extreme Kursänderungen, die so viel größer sind als alltägliche Kursschwankungen, dass sie nicht von einer Gaußverteilung beschrieben werden können. Man könnte solche Beispiele wie den „Schwarzen Montag" im Jahr 1987, als der S&P500-Index innerhalb eines Tages um etwa 20% fiel, für Ausreißer halten, die ein unnatürliches Verhalten des Marktes widerspiegeln. In einer empirischen Analyse finden wir allerdings selbstähnliche Merkmale in der Zeitreihe von Aktienkursänderungen, die also auf vielen Skalen ähnliche Eigenschaften aufweist: Eine riesige Kursänderung verursacht eine Reihe weiterer großer Kursänderungen, deren Rate in den folgenden Monaten langsam abfällt. Auf ähnliche Weise bewirken einige dieser nachfolgenden großen Kursänderungen wiederum mittlere Kursänderungen, deren abfallende Rate für einige Tage nachweisbar ist. Die Mechanismen in Verbindung mit extremen Kursausschlägen scheinen also ähnlich denen von mittleren Schwankungen zu sein, was vermuten lässt, dass sich auch alltägliche Kursänderungen auf ähnliche Weise verhalten. In dieses Bild passt die bekannte, auf breite Akzeptanz stoßende Entdeckung, dass der Rand der Verteilung von Renditen, d.h. Änderungen des logarithmierten Aktienkurses, wie ein Potenzgesetz abfällt, welches mittlere Renditen genauso beschreibt wie Extremereignisse.

Obwohl also Extremereignisse „normal" zu sein scheinen in dem Sinne, dass sie mit den gleichen Mechanismen zusammenhängen wie kleinere Kursbewegungen, ist es noch immer eine ungeklärte Frage, wie Renditen entstehen, die viel größer als in einer Gaußverteilung sind. In festen Zeitintervallen, in denen die oben beschriebenen Kursschwankungen stattfinden, hängt die Kursänderung von zwei Beiträgen ab: von der Anzahl der Transaktionen im Zeitintervall und von der Größe der Kursänderungen infolge einzelner Transaktionen. Um zwischen diesen Effekten besser trennen zu können, untersuchen wir zuerst Intervalle mit einer konstanten Anzahl von Transaktionen, im Gegensatz zu einer festen Länge in der Zeit. Interessanterweise findet man hier noch immer

ungewöhnlich große Kursänderungen, die aus dem gleichzeitigen Auftreten von zwei Dingen resultieren: (i) in dem jeweiligen Intervall ist die mittlere Preisänderung durch eine einzelne Transaktion (die Tickpreisänderung) besonders groß, und (ii) die meisten Transaktionen ändern den Preis in die gleiche Richtung. Die Verteilung der Gesamtrenditen in den untersuchten Intervallen kann von einem statistischen Modell reproduziert werden, das auf den Verteilungen der mittleren Tickpreisänderung und der Richtung der Tickpreisänderungen basiert.

Während diese Analyse im Detail beschreibt, *wie* sich große Aktienkursänderungen zusammensetzen, arbeiten wir in einer weiteren Untersuchung heraus, *warum* es zu diesen großen Kursänderungen kommt. Eine mögliche Annahme ist, dass Preise sich, neben dem Einfluss von Nachrichten, als Antwort auf ein Ungleichgewicht in Angebot und Nachfrage ändern. Dieses Ungleichgewicht lässt sich durch das Volumenungleichgewicht quantifizieren, welches die Differenz zwischen dem Volumen (Anzahl von Aktien) an Kauf- und Verkaufaufträgen (Orders) in einem Zeitintervall beschreibt. Der Aktienkurs reagiert auf ein gegebenes Ungleichgewicht durch eine Preisänderung, welche durch die Preiseinwirkungsfunktion bestimmt ist. Wir rekonstruieren diese Funktion in jedem Zeitintervall aus Daten über alle im Markt vorhandenen Orders. Dabei zeigen wir, dass die zeitlichen Änderungen der Preiseinwirkungsfunktion für das Auftreten außergewöhnlich großer Kursänderungen verantwortlich sind. Obwohl zu jedem Zeitpunkt die Kursänderung durch das Volumenungleichgewicht hervorgerufen wird, treten besonders große Kursänderungen nur dann auf, wenn die Preiseinwirkungsfunktion überdurchschnittlich steil ist.

Wenn Aktienkurse sich als Antwort auf das Ausführen von Orders ändern, scheint ein Paradoxon zu entstehen: Orders, bzw. deren Vorzeichen, die angeben, ob es sich um Kauf- oder Verkauf-Orders handelt, sind langreichweitig korreliert, wohingegen die durch die Orders hervorgerufenen Kursänderungen lediglich kurzreichweitige Korrelationen aufweisen. Um dieses Paradoxon zu lösen, modellieren wir Handelsstrategien und zeigen, dass unkorrelierte Aktienkursänderungen aus korrelierten Orders auf natürliche Weise entstehen, sobald jemand die Korrelationen in den Orders zur Steigerung seines Gewinns verwendet.

Nach der Untersuchung zeitlicher Korrelationen von Kursänderungen analysieren wir auch Kreuzkorrelationen zwischen endlichen Zeitreihen. Auf Grund ihrer begrenzten Länge führen zufällige Gleichbewegungen zwischen den Zeitreihen zu künstlichen Korrelationen, die nicht die tatsächlichen Wechselwirkungen widerspiegeln. Wir stellen einen auf Zufallsmatrixtheorie basierenden Test vor, der zwischen echten und unechten Korrelationen in endlichen Zeitreihen unterscheiden kann, was mit numerischen Simulationen demonstriert wird.

Contents

1 Introduction

In recent years, the stock market has received a great deal of attention from the general public. Since the "New Economy" boom in the late 1990s [1, 2, 3, 4], the values of the world's major stock market indices have become a common feature on daily news shows, and many people pay close attention to the ups and downs of the "Dow Jones", "Nikkei", or "DAX".

The usual motivation for investing in the stock market is a desire to obtain a return that is larger than the return yielded by a riskless investment such as government bonds. However, this additional return can be gained only by exposing one's investment to the risk that is inherent in large fluctuations of the stock price [5]. For example, market crashes such as those of October 1929 or 1987 show that stock prices can fall drastically within a matter of hours, which might be dangerous not only for individual investors but even for the economy as a whole [6, 7, 8, 9, 10, 11]. The mechanisms underlying such large price changes are thus an important object of research.

Economists as well as physicists have studied stock price movements in the past, revealing many properties of stock price changes [12, 13]. Of particular interest for physicists was the discovery of power law tails in the distribution of stock returns (i.e. changes of the logarithm of the price) [14, 15, 16, 17], that describe also extreme price movements such as stock market crashes.

In physics, power laws appear when a system is close to a phase transition [18, 19]. For example, when a magnet is cooled down so that its temperature approaches the critical (Curie) temperature, long-range correlations emerge so that the magnetizations of a large number of subsystems are coupled. The resulting collective behavior of the subsystems leads to large global fluctuations and a strong response to an external influence such as a magnetic field. This response is quantified by the susceptibility, which, together with other quantities, diverges according to a power law when the temperature of the system approaches the critical temperature. The existence of power laws in critical phenomena suggests that there might also be such fundamental mechanisms in the stock market causing the discovered power law distribution of returns.

In this thesis, we analyze large stock price changes, first from a descriptive

point of view and then with a more explanatory approach using the concept of response functions. Here, we study the price impact function [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], which quantifies how the price changes in response to trades. In addition, we focus on correlations in the return time series of single stocks and indices as well as correlations between returns of different stocks.

1.1 Properties of stock returns

In the following, we describe in more detail some of the manifold patterns that have been found in stock price movements. These patterns pertain to the distribution of stock returns as well as time correlations in the return time series.

1.1.1 Distribution of stock returns

Returns on a "macroscopic" scale such as days are the sum of many returns on smaller scales, i.e. minute returns or even returns due to single trades (tick returns). When a variable is calculated by the summation of many random numbers, the central limit theorem states that this variable will be asymptotically distributed like a Gaussian if the summands are independent and have a finite variance. Hence, the Gaussian (or normal) distribution is a reasonable starting point when studying such variables. Indeed, more than one hundred years ago Bachelier [32] modeled price changes as independent, normally distributed random numbers. Later, it was found that price changes are better described by a log-normal distribution, which means that changes of the logarithm of the price, i.e. returns, are normally distributed. The assumption of such a distribution led, for example, to the famous Black-Scholes formula [33] for option pricing [1].

However, market crashes like the ones in 1929 or 1987 show that there are fluctuations much larger than those found in a normal distribution. These extreme price movements used to be called "outliers" since they did not agree with existing theory. This is in striking contrast to physics, where a theory must be valid for all data points – an outlier (if it is not due to a measurement error) that contradicts the theory is not to be discarded, but it is rather the reason for attempts to create a better theory that can also explain this outlier. Indeed, many studies show that the distribution of stock returns exhibits fat tails, indicating that large returns are much more probable than in a Gaussian distribution [14, 15, 16, 17, 34, 35, 36, 37, 38, 39, 40, 41]. In addition, the functional

[1] The Black-Scholes formula dating from 1973 has since been adapted to many "stylized facts" of financial markets, including non-Gaussian returns. An overview of these more recent advances in option pricing is given in [5].

form of the distribution stays similar (stable) if the return is aggregated on very different time scales from seconds to months.

These findings led to the idea that stock returns might have a Lévy stable distribution [14, 37, 42, 43]. Such a distribution corresponds to the generalization of the central limit theorem [44], which also applies to random processes with infinite variance. In particular, if a random process has power law tails $P(x) \sim x^{-(\alpha+1)}$ with $\alpha < 2$, summation of these (independent) random variables leads to a distribution that converges to a Lévy-stable process characterized by α. For $\alpha > 2$, the variance exists and the limiting distribution is a Gaussian. In a Lévy-stable process, which is also called "Lévy flight" or "stable Paretian", the tail of the distribution of the sum is determined by large events in the underlying process, i.e. a large jump of the sum results from an extreme jump in one of the summands.

In contrast to the idea of a Lévy-stable process, later empirical studies find evidence that the distribution of stock returns has a finite variance and a tail that follows a power law $P(x) \sim x^{-(\alpha+1)}$ with $\alpha \approx 3$, suggesting that the distribution is not a stable Paretian [16, 17, 35, 36, 39, 45, 46, 47, 48, 49, 50]. In addition, for very large time scales, the return distribution seems to approximate a Gaussian [39]. Though the tail of the return distribution is currently the object of great interest [27, 51, 52, 53], there is still no general consensus about the "true" mechanism behind large returns.

1.1.2 Stock return correlations

One condition of the aforementioned limit theorems is that the underlying stochastic processes are independent. Indeed, it has been known for a long time that returns have only weak linear correlations, which was later quantified as an exponential decay with a characteristic time of around four minutes [65, 66]. In the light of these small correlations, assuming that returns are independent seems to be a good first order approximation.

Though the return itself has only weak autocorrelations, it has been found that returns are in fact not independent: absolute or squared returns, which in economics are measures of volatility, exhibit long-term memory [15, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70]. Returns seem to remember their past (absolute) value, resulting in time periods of high volatility and other periods when the volatility is low. In economics, this phenomenon is called "volatility clustering". The long-term memory has been known qualitatively for some time, as it was found that the autocorrelation function of absolute returns has a "slow" [57] or "very slow" [58] decay in time. More recently, attempts

were made to quantify this slow decay. For example, Ding *et al.* [62] fit the autocorrelation function of the absolute daily S&P500 index return with a combination of an exponential function and a power law, while Dacorogna *et al.* [63] find a power law exponent between 0.2 and 0.3 for absolute 20 minute returns of different exchange rates. Liu *et al.* [65, 66] study the absolute one-minute S&P500 index return in a 13-year period. Using detrended fluctuation analysis and power spectrum analysis, the authors find that the autocorrelation function can be described by two different power laws with a crossover time of about 1.5 days.

1.2 Cross-correlations and portfolio optimization

Understanding the distribution of price fluctuations and the correlations within the return time series is only part of the picture. If one wants to estimate the risk of an investment, it can be indeed important to know the return distribution in order to prepare for a certain risk so that large price changes do not lead to bankruptcy. However, when managing a portfolio of a variety of stocks, it is possible to actually lower the overall risk of the portfolio.

To this end, it is important to correctly estimate and then minimize the correlations between the stocks. For instance, if the portfolio consists exclusively of stocks from the energy sector, a drop in the price of oil could lower the profits of every company in the portfolio, resulting in a collective drop of their stock prices. In contrast, if one also holds stocks from other sectors that are less influenced by the price of oil, then the loss will be small in relation to the entire invested capital. Hence, when correlations are minimized, so too is the likelihood that a single market event will drastically decrease the value of the portfolio. In other words, minimizing the correlations in a portfolio by diversifying (i.e. investing in many different assets with small cross-correlations), substantially reduces the risk for the invested capital.

A theoretical basis for diversification has been established by Markowitz in his pioneering work on the calculation of efficient investments in the stock market [71]. According to Markowitz, an investment is efficient if for a given return the risk for the invested capital is minimal. Here, "risk" refers to the variance of the portfolio, as opposed to the risk of an extreme price change as discussed above. For efficient diversification and thus a minimal risk, one has to evaluate the cross-correlations between the price changes of all stocks in the portfolio, since these correlations determine the variance, and thus the risk, of the portfolio as a whole.

In order to estimate the correlations within the portfolio, one can calculate its correlation matrix from historical data, which contains the correlation coefficients between all the portfolio's stocks. However, when estimating these correlations, one has to deal with the "curse of dimensionality": if one has only short time series so that their length is comparable to the number of considered stocks, the number of input parameters is of the same order as the number of estimated correlation coefficients, resulting in large estimation errors. This leads to a dilemma: on the one hand, one needs sufficiently long time series to have enough data for a reliable calculation of the correlations. On the other hand, correlations change over time so that one cannot extend the calculation over a long period. For short time series, even totally uncorrelated returns lead to a correlation matrix that deviates significantly from the unit matrix. Hence, due to the limited length of the considered time series, random co-movements of the stock prices lead to spurious correlations that do not reflect the real interactions. Random matrix theory [72] can help distinguish these artificial correlations from "real" correlations by comparing the empirically found correlations with correlations of randomly generated time series of the same length.

1.3 Organization of the chapters

In the beginning of this thesis, we study large price fluctuations from macroscopic to more microscopic points of view in order to reveal information about the underlying mechanisms.

In chapter 2, we analyze time periods after stock market crashes. Similar to the Omori law for earthquakes, a shock (i.e. a large price change) is followed by aftershocks, and the rate of aftershocks larger than a given threshold decreases over time according to a power law with exponent one. Surprisingly, some of these aftershocks themselves initiate a similar power law decay on a smaller scale. This occurrence of crashes on all scales, where each crash is followed by its own aftershocks, can be related to the memory in volatility.

This similarity between price changes on different scales suggests that they are connected to the same mechanisms. Hence, the study of large price changes might reveal information about the general mechanisms underlying the movements of stock prices. In chapter 3, we examine the factors that lead to large price changes in intervals with a fixed number of trades. By using such intervals rather than fixed time intervals, we eliminate the direct influence of the trading frequency, thereby isolating other factors for a more detailed study. We show that large price movements can be modeled using the average tick return and

the difference between the number of positive and negative tick returns in the respective interval.

In chapter 4, we change the point of view. Instead of describing how large returns are composed from the microscopic quantities, we focus on the question why these large returns occur. To this end, we analyze response functions in time intervals of five minutes. Here, we calculate the expected price impact of an order, measuring how the price changes in reaction to a certain buy or sell volume (i.e. the number of traded shares). Unlike in chapter 3, we study not only the price change induced by a trade, but also take into account the whole order book, including complete information about all orders present in the market at a given time. We show that fluctuations of the price impact play an important role in the occurrence of large price changes.

Chapter 5 seeks to explain why returns are basically uncorrelated. This lack of correlations is surprising because prices change as a result of order execution, and order signs, indicating buy or sell orders, are strongly (long-term) correlated [31, 73]. In this chapter, we analyze two different trading strategies and show that uncorrelated stock returns emerge from correlated orders when the order correlations are used to increase the profit of the trader.

In chapter 6, we study a method of testing whether an empirical correlation matrix contains significant correlations as opposed to spurious correlations caused by the limited length of empirical time series. The test compares properties of the empirically found correlation matrix with average properties of random matrices. We analyze this test by the use of Monte Carlo simulations and show that its properties for finite samples can be improved by adjustments obtained from the numerical simulations.

Appendix A gives detailed information about the data sets studied and the programming methods used for this work.

2 Crashes and subcrashes – Omori law on all scales

In this and the following two chapters, we will try to understand the mechanisms underlying stock price movements by studying large price changes from different points of view. While in chapter 3 and 4 we will analyze large price changes on relatively short scales of single trades or five minute intervals, in this chapter we focus on market crashes and their effects on the price movements in the months after the crash.

Such time periods after major stock market crashes were relatively recently studied by Lillo and Mantegna [74] who find that here the stock market behaves similar to earthquakes: the rate of volatilities (i.e. absolute returns) larger than a given threshold q decreases like a power law with an exponent close to one, analogous to the classic Omori law describing the aftershocks following a large earthquake [75].

In this chapter, we show that the Omori law holds not only after significant market crashes, but also after "intermediate shocks". Moreover, we find self-similar features in the volatility. Specifically, within the aftercrash period (called "Omori process" as it is characterized by the Omori law) there are smaller shocks that themselves behave like the Omori law on smaller scales. We call these shocks subcrashes, which can be considered as "new crashes on a smaller scale", followed by their own aftershocks [76].

Our results suggest that Omori processes might be present on all scales and therefore constitute an important part of the mechanism underlying price fluctuations. Having this in mind, we study the relation between Omori processes and the long-term memory in volatility. Here, we do not only analyze the volatility itself, but focus on volatility return intervals, the time between two consecutive events with volatilities larger than a given threshold. Recent studies [77, 78, 79, 80] show that this analysis can reveal more information about the temporal structure of the volatility time series. They find that, similar to the volatility, return intervals display memory and volatility clustering, and also scaling properties for different thresholds, which seem to be universal for differ-

ent time scales and markets [77, 78, 79, 80]. This behavior is similar to what is found in earthquakes [81] and climate [82, 83]. Due to the scaling properties, it is possible to analyze the statistics of return intervals for different thresholds by studying only the behavior of small fluctuations occurring very frequently, which have good statistics. The results can then be applied to the more interesting but rare extreme events.

Our analysis shows that the memory in volatility return intervals after large market crashes is indeed related to the Omori law. Specifically, if we perform appropriate detrending, the return intervals show significantly less memory, but some memory still exists, independent of the large market crash. We also show that at least part of this "remaining memory" can be described by the self-similar subcrashes: if we remove also Omori processes due to subcrashes, the memory is further reduced. We also analyze the memory in the volatility time series and show that removing the influence of the major crash and some of its subcrashes reduces the memory in the data set. However, some memory still remains so that these crashes cannot account for the entire memory, raising the possibility that the "remaining memory" is due to other subcrashes whose influence was not removed.

This chapter is organized as follows. Section 2.1 presents information about the analyzed data. In section 2.2 we show and discuss the mechanism based on Omori processes on different scales. In section 2.3 we study the memory in return intervals induced by large and intermediate shocks. In section 2.4 we analyze the influence of crashes on the volatility memory, and section 2.5 gives a summary of the results.

2.1 Description of data sets

In order to capture a variety of market crashes, we analyze three different data sets. More specific information about the studied data sets is given in appendix A.

- (i) We study the one minute return time series of the S&P500 index from 1984 to 1989. Here, we analyze the aftercrash period in the 15,000 trading minutes (approximately two months) after "Black Monday", 19 October 1987, as well as after a smaller crash on 11 September 1986. We also analyze the time after several other smaller market crashes within the entire data set.
- (ii) The second data set consists of the TAQ data base of the year 1997 which is provided by the NYSE and contains all trades and quotes for

all stocks traded at NYSE, NASDAQ, and AMEX. We choose the 100 most frequently traded stocks and calculate an index by a summation of the normalized prices of each stock. From this index, we calculate a one minute return time series for our analysis, which we analyze in the approximately two months after the crash on 27 October 1997.

- (iii) As an example of a crash that is clearly due to an external event, we also study the one minute return series of General Electric (GE) stock in the three months after 11 September 2001.

For all three data sets, we calculate the volatility as the absolute value of the one minute return

$$G_{\Delta t} = \ln S(t + \Delta t) - \ln S(t) \tag{2.1}$$

with $\Delta t = 1\text{min}$, normalized by the standard deviation

$$\sigma_G = \sqrt{\langle G^2 \rangle - \langle G \rangle^2} \tag{2.2}$$

of the entire period. Hence, in this chapter the volatility and also the threshold q, as well as many quantities in the following chapters, are measured in units of the standard deviation σ_G.

2.2 Omori law on different scales

Lillo and Mantegna [74] showed that the Omori law [75] for earthquakes also holds after crashes of large magnitude in financial markets, so that the rate $n(t)$ of events with volatility larger than a given threshold q decays as a power law

$$n(t) = kt^{-\Omega} \quad , \tag{2.3}$$

where Ω is around one for large q and k is a parameter characterizing the amplitude of the rate $n(t)$. For estimating the parameter k and the exponent Ω, we use the cumulative number $N(t)$ of events larger than q, given by

$$N(t) = \int_0^t n(t')dt' = k\frac{1}{1-\Omega}t^{1-\Omega} \quad . \tag{2.4}$$

We study the Omori law on different time scales. Figure 2.1 shows the cumulative rate $N(t)$ above (a) $q = 3$ and (b,c) $q = 4$ compared to the volatility in time periods following three significant market crashes in (a) 1986, (b) 1987, and (c) 1997. The volatility is smoothed by a moving average over 60 minutes in order to remove insignificant fluctuations. The large shock in the beginning of the time interval is followed by aftershocks, which induces an Omori-like behavior of $N(t)$ (Omori process), shown by the dashed lines representing a power law

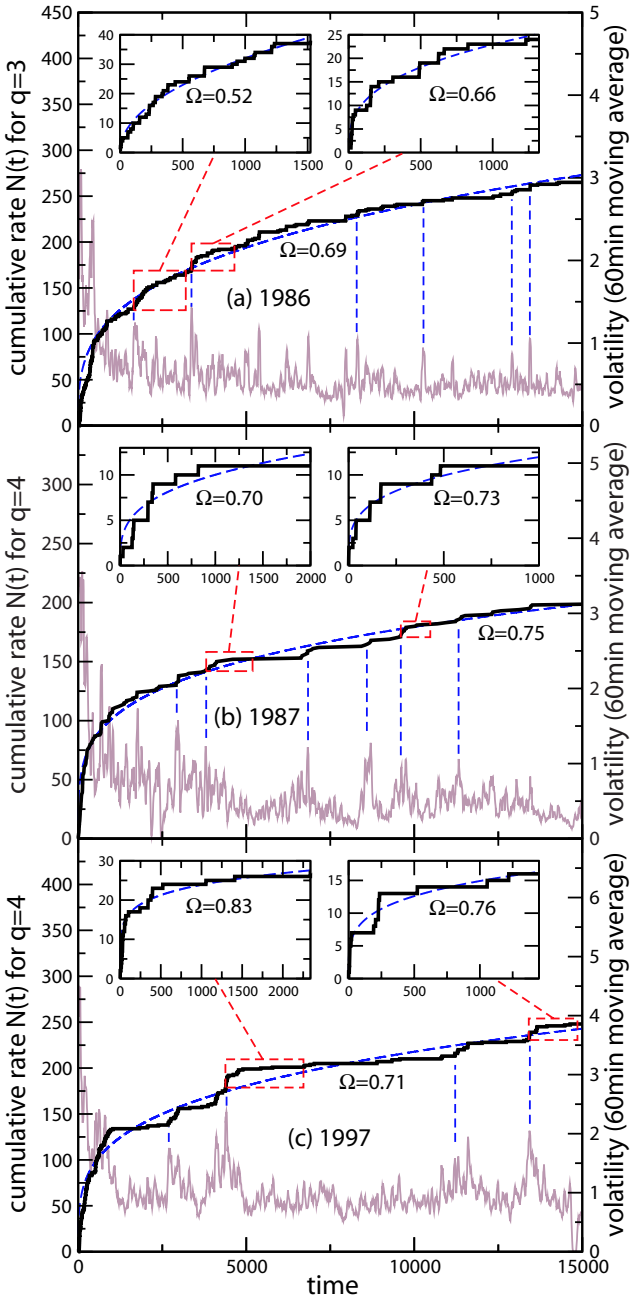

Figure 2.1: Comparison between volatility and the cumulative rate $N(t)$ of volatilities (absolute one minute returns) larger than a threshold q. The plots show the 15,000 minutes (approximately two months) after the market crashes on (a) 11 September 1986, with $q = 3$, (b) 19 October 1987, with $q = 4$, and (c) 27 October 27 1997, with $q = 4$. The volatility is displayed as a moving average over 60 minutes in order to suppress insignificant fluctuations. The insets show the self-similarity of the data set, meaning that while the big crash in the beginning induces a behavior following the Omori law, some of the aftershocks induce again a similar behavior on a smaller scale.

fit. However, as seen in the insets of Fig. 2.1 many of these aftershocks seem to behave like "real" crashes with their own aftershocks (subcrashes), but on a smaller scale (shown by vertical lines). The insets show that a closer look into many of these subcrashes reveals a similar pattern as the Omori law on large scales. The exponent Ω is often smaller after smaller crashes, which is consistent with the finding that the power law decay of the volatility after smaller shocks has a smaller exponent than after large crashes [84]. Below we explore the possibility that the self-similarity of the volatility (where the Omori law is present on different scales) is directly related to the memory.

2.3 Return interval memory after crashes and subcrashes

In order to explore the memory effects of the Omori law, we first analyze time periods after very large market crashes. Specifically, we study the memory in the volatility return intervals, which form a sequence of time intervals $\tau(t)$ between two consecutive events with volatilities larger than a given threshold q [77, 78, 79, 80]. We next show that the influence of the Omori law on $\tau(t)$ can be estimated by comparing the original $\tau(t)$ with a detrended time series $\tilde{\tau}(t)$ which is independent of the market crash. We fit the cumulative rate $N(t)$ in the period after a market crash with a power law according to Eq. (2.4), thus obtaining the parameter k and the exponent Ω for the rate $n(t)$ [74]. Using $n(t)$, we can detrend the return interval time series $\tau(t)$ by rescaling by $n(t)$ [85]

$$\tilde{\tau}(t) = \tau(t)n(t) \quad . \tag{2.5}$$

The rational for this detrending is the following: immediately after the crash we have a large rate $n(t)$ of high volatilities so that the return intervals $\tau(t)$ are very short. Later, the rate of high volatilities becomes small while the return intervals get large. According to Eq. (2.5), high (low) rates and small (large) return intervals cancel each other so that $\tilde{\tau}(t)$ is detrended and thus independent of the existence of the crash, since the trend caused by the crash is no longer present.

The relation between the Omori law and the short-term memory in the return interval time series can be studied by analyzing the conditional expectation value $\langle \tau(t)|\tau_0 \rangle$ of the return interval series $\tau(t)$ conditioned on the previous return interval τ_0 [77, 78], for both the original return intervals $\tau(t)$ and the detrended time series $\tilde{\tau}(t)$. In Fig. 2.2 (left column), $\langle \tau(t)|\tau_0 \rangle$ is plotted against τ_0. Both quantities are normalized by the average return interval $\langle \tau \rangle$, for return intervals

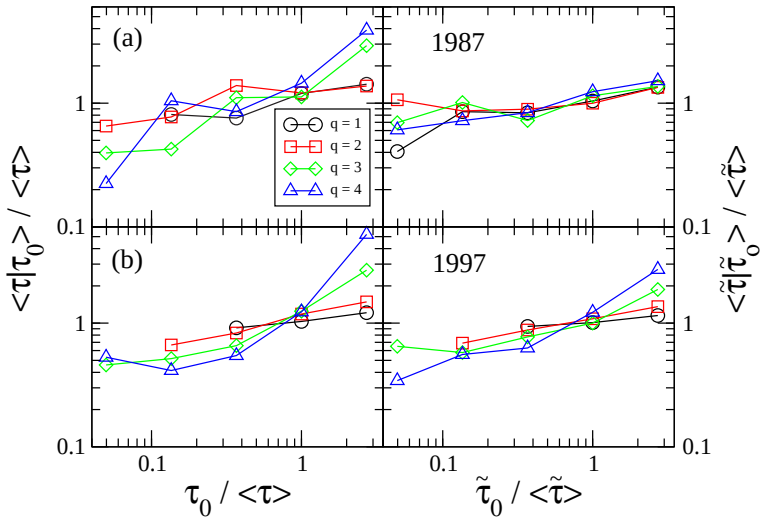

Figure 2.2: Memory in volatility return intervals for different thresholds before (left column) and after (right column) detrending the time series according to Eq. (2.5). The analysis is shown for (a) the S&P500 index in the two months after the crash on 19 October 1987 and (b) an index calculated from the 100 most frequently traded stocks from the TAQ data base after the crash of 27 October 1997. Removing the Omori law reduces the memory in the data sets, but some memory still exists.

after the crashes in (a) October 1987 and (b) October 1997. The deviations from a horizontal line at $\langle \tau(t)|\tau_0\rangle = 1$ for all thresholds show memory: large (small) values of τ_0 are more likely to be followed by large (small) values of $\tau(t)$. The slopes of the curves for the detrended time series $\tilde{\tau}$ are significantly less steep (right column), indicating that detrending the Omori law from the time series significantly reduces the memory, but some of the memory still remains, which might be due to the Omori process still present on smaller scales (see Fig. 2.1).

In addition to the effect of the major crash, we can also analyze the influence of Omori processes after subcrashes on smaller scales. To this end, we further detrend the time series by removing some subcrashes and test whether the memory is further reduced. After identifying the subcrashes [1], we detrend the

[1] To properly identify subcrashes that can be removed from the records, we filter the time series with an appropriate criteria for each data set. For the S&P500 index time series, including the crashes from 1986 and 1987, we define a subcrash as an event where the 60 minute moving average of the one minute volatility exceeds one standard deviation (corresponding to a much larger one minute volatility burst). We also require at least 500 minutes to the next subcrash (events within 100 minutes are considered as the same subcrash). For the data from 1997, we analyze the ten minute moving average, and a

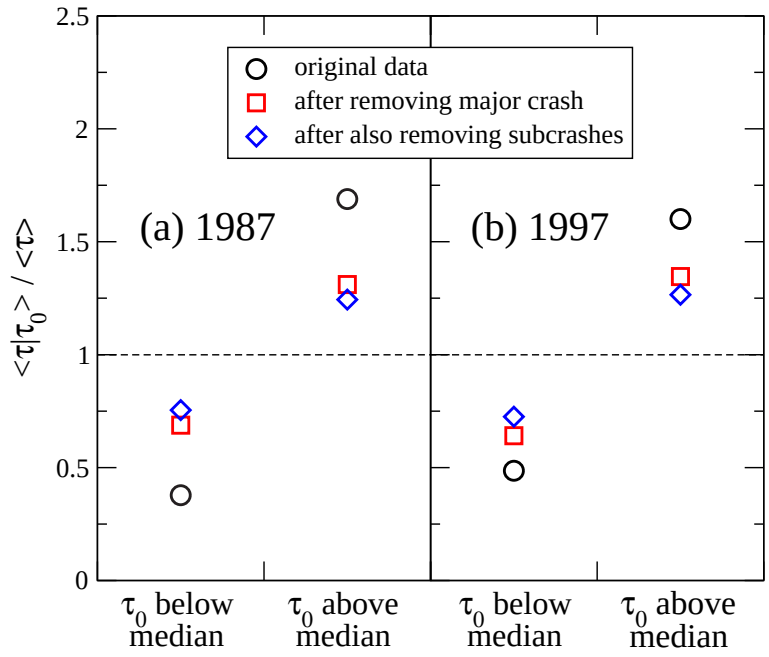

Figure 2.3: Memory in volatility return intervals for threshold $q = 3$ for (a) the S&P500 index in the two months after the crash on 19 October 1987 and (b) for an index calculated from the 100 most frequently traded stocks from the TAQ data base after the crash of 27 October 1997. The conditional expectation value $\langle \tau|\tau_0\rangle / \langle\tau\rangle$ conditioned on the previous return interval τ_0 is smaller than one if τ_0 is below the median while $\langle \tau|\tau_0\rangle / \langle\tau\rangle > 1$ if τ_0 is above the median, indicating the memory in the records (circles). The effect weakens upon detrending the time series by removing the influence of the major crash (squares) and even further when removing some subcrashes (diamonds).

return intervals $\tau(t)$ by removing the Omori process due to the major crash as well as the Omori processes induced by the subcrashes. To this end, we estimate the parameters k and Ω in Eq. (2.3) for the rate $n(t)$ after the major crash as well as for the rate $n_s(t)$ in the 1000 minutes following each subcrash (or the time to the next subcrash, if smaller). Note that $n_s(t)$ is calculated from the detrended return intervals $\tilde{\tau}(t)$. Then, the double detrended return interval time series is given by

$$\tilde{\tilde{\tau}}(t) = \begin{cases} n_s(t)\tilde{\tau}(t) & \text{in time following a subcrash} \\ \tilde{\tau}(t) & \text{otherwise.} \end{cases} \tag{2.6}$$

In order to improve the statistics for testing the effect of removing also subcrashes on the memory, we plot in Fig. 2.3 the conditional expectation value

subcrash has to exceed 2.5 standard deviations. The other parameters are the same as for the S&P500 data.

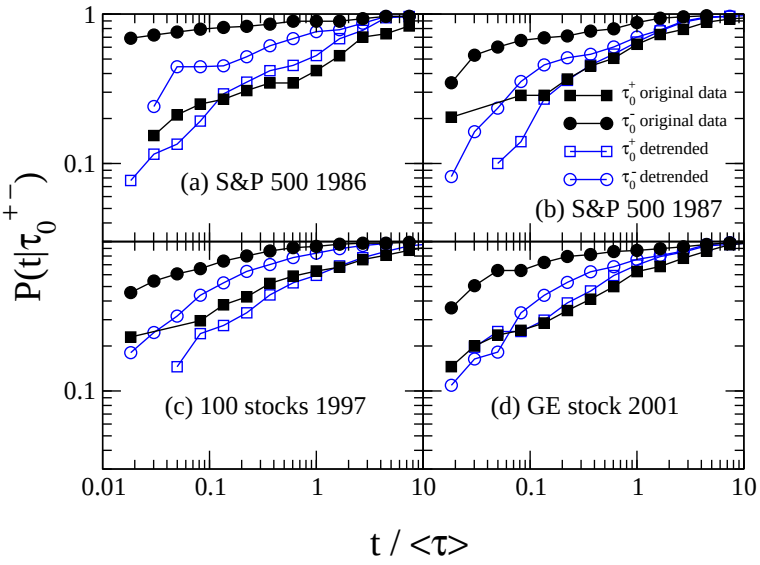

Figure 2.4: Probability $P(t|\tau_0)$ that after a return interval τ_0 the next volatility larger than a threshold $q = 4$ ($q = 3$ in (d)) occurs within time t. Here, τ_0 belongs to either the 25% smallest values (τ_0^-, circles) or the 25% largest values (τ_0^+, squares) of τ. The memory in the original time series (filled symbols) is reduced after removing the influence of the major crash by detrending according to Eq. (2.5) (open symbols), but some of the memory still remains. The results are shown for (a) the S&P500 index after a crash on 11 September 1986, (b) the S&P500 index after the crash on 19 October 1987, (c) an index created from the 100 most frequently traded stocks from the TAQ database after the crash on 27 October 1997 and (d) GE stock after 11 September 2001.

$\langle\tau|\tau_0\rangle / \langle\tau\rangle$ for only two τ_0 intervals: τ_0 below and τ_0 above the median of τ. We see in Fig. 2.3 that when τ_0 is below the median, $\langle\tau|\tau_0\rangle / \langle\tau\rangle < 1$, while $\langle\tau|\tau_0\rangle / \langle\tau\rangle > 1$ for τ_0 above the median. This indicates the memory in the records, and also shows that the memory in the original records (circles) weakens upon detrending the time series by removing the influence of the major crash (squares) and further weakens when also some subcrashes are removed (diamonds). Hence, not only a large market crash but also smaller subcrashes contribute to the memory in return intervals.

To further investigate the effect of removing the memory induced by aftershocks, we analyze the probability $P(t)$ that after an event larger than a certain volatility q the next volatility larger than q appears within a time t [79, 81, 83]. In order to study the memory, we plot the conditional probability $P(t|\tau_0)$ for different values of the preceding return interval τ_0. Figure 2.4 shows $P(t|\tau_0)$ for $q = 2$ under the condition that the preceding return interval τ_0^- belongs to the smallest 25%

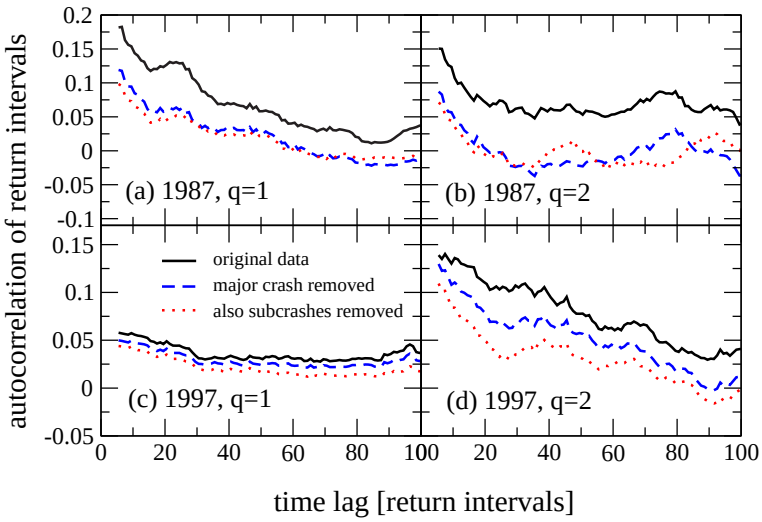

Figure 2.5: Autocorrelation function of the return interval time series for threshold (a,c) $q = 1$ and (b,d) $q = 2$. The first row (a,b) shows results from the S&P500 index in the three months after the market crash on October 19, 1987, while the second row (c,d) results from an index created from the 100 most frequently traded stocks from the TAQ database after the crash on 27 October 1997. The Omori law due to the market crash (original data, solid lines) induces correlations leading to an offset in the autocorrelation function which is removed in the detrended $\tilde{\tau}$ (dashed lines), but the data still shows some long-term correlations even after removing the influence of the Omori law. However, after further detrending with respect to some subcrashes (dotted line), the autocorrelation is further reduced. All lines are smoothed by a moving average over ten return intervals.

of the return intervals or that the preceding return interval τ_0^+ belongs to the largest 25%. The memory in the time series leads to a splitting of the curves because after larger return intervals (squares) the time to the next volatility above q is usually large, while it is short after small return intervals (circles). After removing the influence of the major crash by detrending, the curves get closer, indicating a reduced memory, but also here some memory still remains.

To test the long-term memory effects of the Omori process on the volatility return intervals we study their autocorrelation function. For two different time series $x(t)$ and $y(t)$, the correlation function quantifies the correlations at a time lag Δ as

$$c(x(t), y(t), \Delta) = \frac{\langle x(t)y(t+\Delta)\rangle - \langle x(t)\rangle\langle y(t)\rangle}{\sigma_x \sigma_y} \quad , \tag{2.7}$$

where σ_x and σ_y denote the standard deviations of $x(t)$ and $y(t)$. If the time

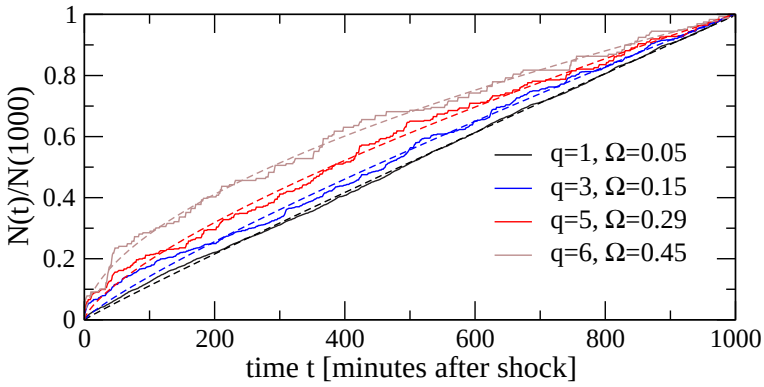

Figure 2.6: Cumulative rate $N(t)$ of events larger than a threshold q averaged over the 1000 minutes after 22 shocks between $11\sigma_G$ and $16\sigma_G$ in the S&P500 one minute time series of the years 1984 to 1989. The data for each shock is normalized by $N(1000)$ in order to make different shocks comparable irrespective of the current trading activity. The cumulative rate can be well fitted by a power law according to Eq. (2.4). The exponent grows from $\Omega = 0.05$ to $\Omega = 0.45$ for $q = 1 \ldots 6$.

series $x(t)$ and $y(t)$ are identical, one obtains the autocorrelation function

$$c(x(t), \Delta) = \frac{\langle x(t)x(t+\Delta)\rangle - \langle x(t)\rangle^2}{\sigma_x^2} \quad . \tag{2.8}$$

Figure 2.5 shows the autocorrelation function of return intervals after the market crashes in 1987 and 1997 for two different thresholds $q = 1$ and $q = 2$. For both thresholds, we see that there exists a significant correlation even between return intervals 100 steps apart, which corresponds to approximately $2-5$ days in 1987 ($0.5-2$ days in 1997) since the average return intervals are $\langle\tau(q=1)\rangle = 6.33$min and $\langle\tau(q=2)\rangle = 17.4$min in 1987 and $\langle\tau(q=1)\rangle = 2.47$min and $\langle\tau(q=2)\rangle = 7.66$min in 1997. If we now remove the effect of the Omori process due to the market crash by detrending according to Eq. (2.5), the memory in the detrended sequence $\tilde{\tau}$ is reduced significantly, as we see in the dashed curves of Fig. 2.5. The dotted lines show that removing also the influence of some subcrashes according to Eq. (2.6) further reduces the memory.

So far, we showed indications that within the time period after a big crash there might exist smaller crashes that behave in a similar way. The question arises whether such subcrashes are only typical after a big crash or whether they appear in all time periods independent of the existence of a big crash. To test this, we study 22 crashes of sizes between 11 and 16 standard deviations in the S&P500 time series from 1984 to 1989. These crashes are considerably smaller than the huge crashes of more than 30 standard deviations in a one

minute interval studied above. We analyze the cumulative rate $N(t)$ in the 1000 trading minutes following these smaller crashes. In order to make different crashes comparable irrespective of the current trading activity, we normalize the cumulative rate $N(t)$ by $N(1000)$. Figure 2.6 shows this normalized rate $N(t)/N(1000)$ averaged over all aftershock periods [2]. For different thresholds q, $N(t)/N(1000)$ can be fit with a power law, Eq. (2.4). The exponent Ω increases with the threshold, but is generally smaller than the exponents found after very large shocks. Our results for the *rate* decay are analogous to volatility studies [84, 86] where the exponent characterizing the *volatility* decay depends on the magnitude of the shock [84]. These results indicate that relatively small crashes have similar Omori processes which may lead to memory effects.

2.4 Memory in Volatility after crashes and subcrashes

In the previous sections, we showed that the memory in return intervals decreases when we remove effects due to Omori processes. Since the studied return intervals $\tau(t)$ are derived from the volatility time series $v(t)$, it would be interesting to test whether the memory in $v(t)$ is also affected by Omori processes. Thus, we next analyze the memory in the volatility time series directly. It is known that a market crash induces a power law decay of the approximate form

$$v_{\mathrm{PL}}(t) \equiv v_0 t^{-\beta} \tag{2.9}$$

with an exponent $\beta \approx 0.2 - 0.3$ [74, 84]. In order to study the memory induced by this decay, we compare the original time series $v(t)$ to a detrended one

$$\tilde{v}(t) \equiv \frac{v(t)}{v_{\mathrm{PL}}(t)} \tag{2.10}$$

so that $\tilde{v}(t)$ does not depend on the market crash.

We use second order detrended fluctuation analysis (DFA2) [87, 88, 89] to study the long-term memory in the volatility [15, 54, 55, 56, 58, 57, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70]. In DFA2, the deviations $F(s)$ (root mean square fluctuations) from a second degree polynomial fit of the profile

$$y(t) = \sum_{t'=0}^{t} v(t') \tag{2.11}$$

as a function of different scales s (time windows) reveal information about the memory. If $F(s) \sim s^{\alpha}$, the autocorrelation exponent γ of the time series is

[2] The average only includes crashes where the volatility exceeds the threshold q at least five times during the studied time period of 1000 minutes. For e.g. $q = 6$, there are 11 crashes that satisfy this criteria.

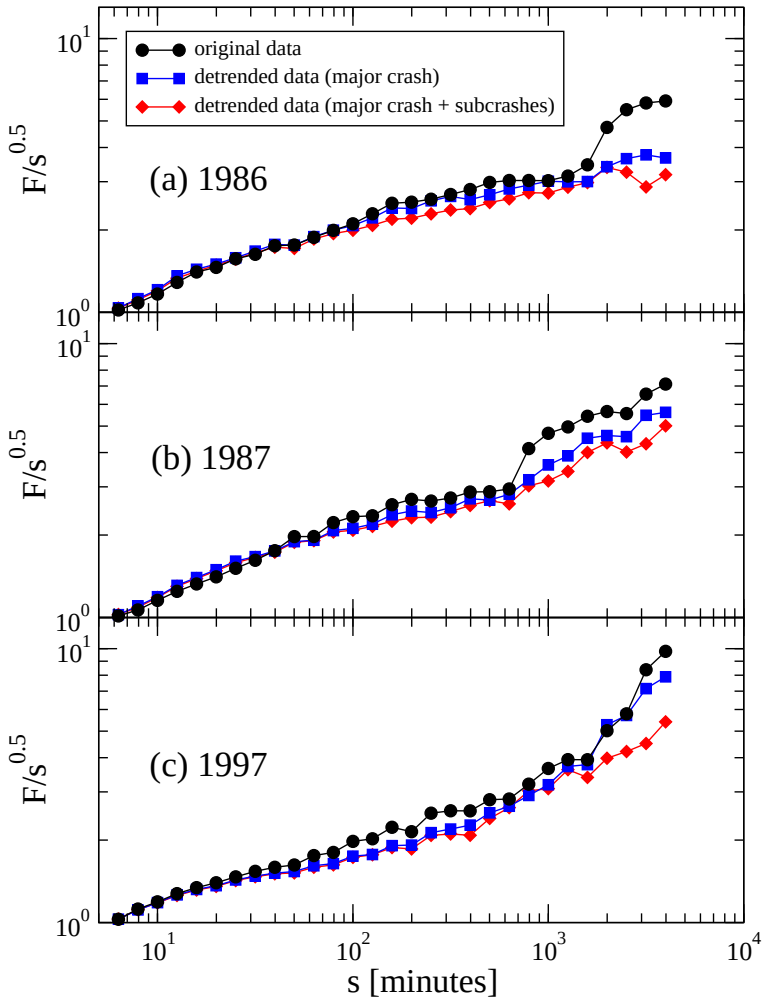

Figure 2.7: Root mean square fluctuation $F(s)$ obtained by the second order DFA method (DFA2) for the volatility in the 15,000 minutes following market crashes in (a) the S&P500 index on 11 September 1986 and (b) on 19 October 1987, as well as (c) the market crash on 27 October 1997 for an index created from TAQ-data for 100 stocks. $F(s)$ is divided by $s^{0.5}$ to clarify the deviation from uncorrelated data. Compared to the original volatility $v(t)$ (circles), the memory is reduced in the detrended records $\tilde{v}(t)$ (squares), and even further after also detrending some subcrashes in $\tilde{\tilde{v}}(t)$ (diamonds).

related to the exponent α by $\alpha = 1 - \gamma/2$. For $\alpha > 0.5$, the time series is long-range correlated, it is anti-correlated for $\alpha < 0.5$, and $\alpha = 0.5$ indicates no long-range correlations. Figure 2.7 shows $\log(F(s)/s^{0.5})$ plotted against $\log s$ for 15,000 trading minutes after three different market crashes of 1986, 1987, and 1997. With no long-term correlations, the function would be constant, while a positive slope indicates long-term correlations. For all crashes, the original time series (circles) shows an increased slope on large time scales. After detrending according to Eq. (2.10) and replacing $v(t')$ by $\tilde{v}(t')$ in Eq. (2.11), the curve (squares) gets less steep, indicating a reduction of the memory (the curves are shifted so that they start at the same point).

As described before, there are also subcrashes which may induce their own power law decay on a smaller scale – not only in the rate, but also in the volatility. In order to analyze the memory due to these subcrashes, we further detrend the time series and test whether the memory is reduced even further. To this end, we fit the detrended volatility $\tilde{v}(t)$ in the 1000 minutes following each subcrash (or the time to the next subcrash, if shorter) with a power law $\tilde{v}_{\mathrm{PL}}$ according to Eq. (2.9). Then, we further detrend $\tilde{v}(t)$ in these regions using Eq. (2.10) for $\tilde{v}(t)$ instead of $v(t)$. The DFA2 curve for the double detrended time series $\tilde{\tilde{v}}(t) \equiv \tilde{v}/\tilde{v}_{\mathrm{PL}}$ is also shown in Fig. 2.7. The decrease in the slope shows that the memory is further reduced after removing the influence of the subcrashes. However, we clearly see that removing the trends induced by a market crash as well as subcrashes only slightly reduces the memory in the volatility on quite small scales ($s < 60$min).

The effect of removing subcrashes on the long-term correlations of volatility is seen better in Fig. 2.8. Here, we compare the autocorrelation functions of the detrended volatility $\tilde{v}(t)$ and the double detrended volatility $\tilde{\tilde{v}}(t)$ after also removing subcrashes. It is seen that generally the autocorrelation of $\tilde{\tilde{v}}(t)$ is smaller than of $\tilde{v}(t)$, which indicates that the Omori processes after subcrashes also contain some memory. Our results are in agreement with the findings of Borland and Bouchaud [90], who recently presented a multi-timescale model that can account for both volatility clustering and Omori type laws.

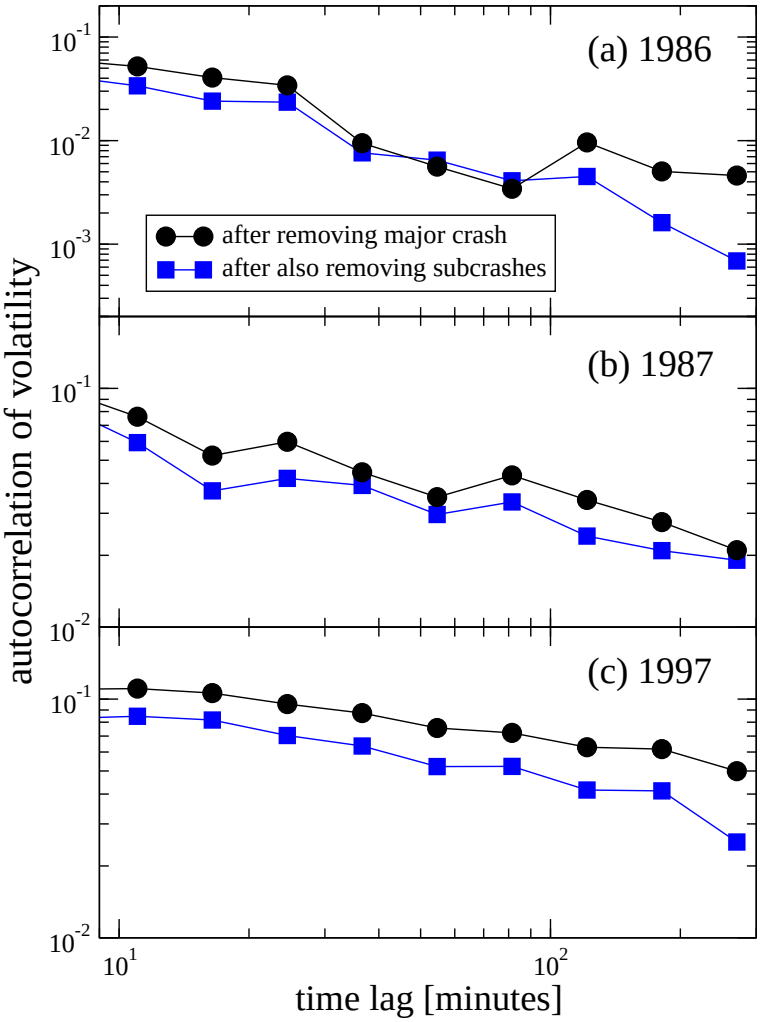

Figure 2.8: Autocorrelation function of the volatility time series after detrending. Compared to the volatility time series after only detrending the major crash (circles), detrending subcrashes (squares) further reduces the autocorrelations. The results are shown for (a) the S&P500 index after a crash on 11 September 1986, (b) the S&P500 index after the crash on 19 October 1987, (c) an index created from the 100 most frequently traded stocks from the TAQ database after the crash on 27 October 1997. The autocorrelation function of the original volatility time series is not shown because it is not meaningful as it is dominated by the influence of the market crash.

2.5 Summary

We find that Omori processes after market crashes exist not only on very large scales, but a similar behavior is also induced by less significant shocks. Moreover, we find that such Omori processes on different scales can occur within the same time period. This leads to self-similar features of the volatility time series, meaning that some of the aftershocks of a large crash can be considered as subcrashes that themselves initiate Omori processes on a smaller scale.

This result suggests a mechanism that might be present on all scales, not only after large market crashes. Indeed, we find that a significant amount of memory is induced by this self-similarity with crashes and subcrashes, which suggests that a large part of the memory in volatility might be due to Omori processes on different scales.

3 Analysis of aggregated tick returns

In the previous chapter, we studied Omori processes after huge market crashes, affecting the volatility over several months. We found that these crashes have an equivalent on smaller scales, suggesting that there might be a generic mechanism that can be related to the memory in volatility. In the present chapter, we want to focus on a much smaller scale and study not the time after large price changes, but try to understand the mechanism that leads to such large returns.

In the introduction of this thesis, we already described the practical as well as theoretical relevance of the fat tailed distribution of stock price changes. In practice, using the correct distribution can help find more accurate models and can lead to a better risk estimation when the probability for extreme events is known [12, 5]. From a theoretical point of view, the power law distribution is reminiscent of critical phenomena and universality, suggesting that there might be a universal mechanism leading to this distribution. Finding this mechanism could lead to a better understanding of financial markets and reveal important constraints for modeling financial time series.

In this chapter, we focus on the analysis of tick returns, i.e. returns due to a single trade. Farmer *et al.* find that the distribution of tick returns is similar to the distribution of returns aggregated on longer time scales, exhibiting power law tails $P(x) \sim x^{-(\alpha+1)}$ with the same tail exponent [52]. Although the exponent is outside the Lévy regime $0 < \alpha < 2$, the authors argue that similar to a Lévy flight, both distributions are caused by the same microscopic mechanism, so that large aggregate returns are due to single exceptionally large tick returns. Plerou *et al.* describe the price movements as a diffusion process with a fluctuating diffusion constant and relate the distribution of aggregate returns to the distribution of the variance of the tick returns [91].

We investigate the transition from tick returns to returns aggregated in intervals with a larger number of trades. It is well documented (e.g., in [92, 93]) that the number of trades in a time interval is an important determinant of the aggregate return. However, the trading frequency alone cannot account for the observed fat tailed distribution of aggregate returns [91, 94]. Thus, we remove the direct influence of the trading frequency by analyzing intervals with a constant number

of trades so that effects due to other quantities like the absolute tick return are more clearly visible.

Similar to the work of Plerou *et al.* [91], this study examines price movements as a diffusion process [95]. Our results for intervals with a constant number of trades confirm some of their findings for time intervals, specifically the result that the mean square of the tick return (here the mean absolute tick return) is an important determinant for large aggregate returns. However, our study goes considerably beyond this work. While Plerou *et al.* compare the exponents of the distributions and conclude that the power law tails of the aggregate return are due to the distribution of the variance of the tick returns, we actually study the intervals with the largest aggregate returns and check which quantities lead to these specific events. In this way, we can directly study the influence of each quantity on the aggregate return. Using this information, we also present a statistical model illustrating the mechanism leading to large price fluctuations.

Moreover, we find that the tick return size (absolute tick return) can well characterize an interval of many trades because it is long-term correlated in tick time (compare [62, 63, 65, 66, 67, 68, 96, 97, 98, 99]). According to the central limit theorem, independent tick returns would in aggregation lead to Gaussian-distributed returns, but due to the correlations, the fluctuations of the mean tick return size lead to the non-Gaussian behavior of the aggregate return. In this picture, large aggregate returns do not occur because of a few very large tick returns, but rather when the average tick return is large, so that even Gaussian fluctuations in the direction of the trades can lead to aggregate returns larger than in a Gaussian distribution.

The remainder of this chapter is organized as follows. Section 3.1 shows our model for the price diffusion process, in section 3.2 we describe the data set used for this study, section 3.3 shows the influence of the tick return size on the aggregate return while section 3.4 focuses on the influence of differences in the direction of tick returns (number difference). Section 3.5 compares the number difference and the flow of market orders and in section 3.6 we present a statistical model which approximates the distribution of aggregate returns. We conclude with a discussion of our results in section 3.7.

3.1 Model

We study intervals with a fixed number of $N = 100$ trades. If the price of a stock before the ith trade is s_i, we define the return due to a single trade, the

tick return, as

$$\delta g_i = \ln(s_{i+1}) - \ln(s_i) \quad . \tag{3.1}$$

The interval I_j contains all N trades with index i between jN and $(j+1)N$, so the aggregate return G_j is given by the sum over all δg_i with $i \epsilon I_j$:

$$G_j = \sum_{i \epsilon I_j} \delta g_i \quad . \tag{3.2}$$

We want to discuss two special cases in order to analyze the mechanism leading to large aggregate returns G_j. In the first case, G_j is dominated by one (or a few) extremely large $\delta g_{i_0}^{max}$, so that

$$G_j = \delta g_{i_0}^{max} + \sum_{i \epsilon I_j, i \neq i_0} \delta g_i \approx \delta g_{i_0}^{max} \quad . \tag{3.3}$$

Thus, G_j becomes large if $\delta g_{i_0}^{max}$ is exceptionally large.

In the second case, we assume that there is no extremely large tick return dominating the aggregate return, so that we focus on the average size Δg_j of the non-zero tick returns, which is defined by

$$\Delta g_j = \frac{1}{n_j} \sum_{\delta g_i \neq 0, i \epsilon I_j} |\delta g_i| \quad . \tag{3.4}$$

Here, n_j is the number of $\delta g_i \neq 0$ in the interval I_j. Neglecting asymmetries in the δg_i, we can replace all $\delta g_i \neq 0$ by $\mathrm{sgn}(\delta g_i)\Delta g_j$ and approximate the aggregate return by

$$G_j \approx \Delta g_j \sum_{\delta g_i \neq 0, i \epsilon I_j} \mathrm{sgn}(\delta g_i) = \Delta g_j \Delta N_j \quad , \tag{3.5}$$

where $\Delta N_j = \sum_{\delta g_i \neq 0, i \epsilon I_j} \mathrm{sgn}(\delta g_i)$ is the number difference. Similarly, G_j can be described as a diffusion process with

$$\left\langle G_j^2 \right\rangle \approx D_j N \quad , \tag{3.6}$$

where the diffusion constant $D_j = \frac{n_j}{N} \Delta g_j^2$ varies due to the varying step width Δg_j and the number n_j of nonzero tick returns.

In the approximation given by Eq. (3.5), we can study the influence of the mean size of the tick returns as well as asymmetries in their direction. A large aggregate return can occur if the price moves more often in one direction than in the other. Thus, with large temporary correlations between the signs, even small tick returns could compose a large G_j. On the other hand, if Δg_j is larger, even a small asymmetry in the signs can lead to a large return.

The two approximations given in Eqs. (3.3) and (3.5) are analyzed in sections 3.3 and 3.4 of this chapter, but in section 3.6 we also consider the error term neglected in Eq. (3.5). An exact formulation is written

$$G_j = \Delta g_j \Delta N_j + \frac{2n_j^+ n_j^-}{n_j}(\Delta g_j^+ - \Delta g_j^-) \tag{3.7}$$

where Δg_j^+ and Δg_j^- are the average tick returns in the positive and negative directions while n_j^+ and n_j^- are the numbers of non-negative tick returns in the positive and negative directions.

3.2 Data analysis

We analyzed the order book data of the year 2002 from Island ECN for the ten most frequently traded stocks [100]. Since the Island ECN is a secondary market where only part of the whole stock volume is traded, we also studied the index fund QQQ which was mainly traded via Island until September 2002. Since our results for the ten stocks and QQQ are similar, we find no evidence that secondary market characteristics of Island affect our analysis negatively. More detailed information about the studied data set is given in appendix A.

In an electronic market place like Island, people can place limit orders to buy or to sell at a given or better price (limit price), which is specified in the order. These orders are stored in the order book, and they are only executed when the actual stock price reaches the limit price. A trade is initiated by a market order indicating that someone wants to buy or sell immediately at the best available price. Such a market order executes the limit orders offering the best prices until the number of shares specified in the market order is traded.

Our data set contains information about every limit order so that we are able to reproduce the market situation at each instant of time. We combine those limit order executions with identical time stamps as they reflect the same market order. Therefore, we can analyze the impact of each single market order on the price. In this analysis, the price s_i is defined as the midquote price $s_i = \frac{1}{2}(s_i^{\text{bid}} + s_i^{\text{ask}})$, which is the mean of the quotes, i.e. the best available buy limit price s_i^{bid} (bid price) and sell limit price s_i^{ask} (ask price). We study intervals with a fixed number of $N = 100$ market orders and have approximately 100,000 intervals in our data set for ten stocks. Thus, on average a 100 trade interval corresponds to about ten minutes, but the trading frequency fluctuates strongly so that 100 trades can correspond to time intervals with very different lengths.

We determine the midquote price s_i just before the execution of the ith market order. Since most trades change the price just by the size of the gap between

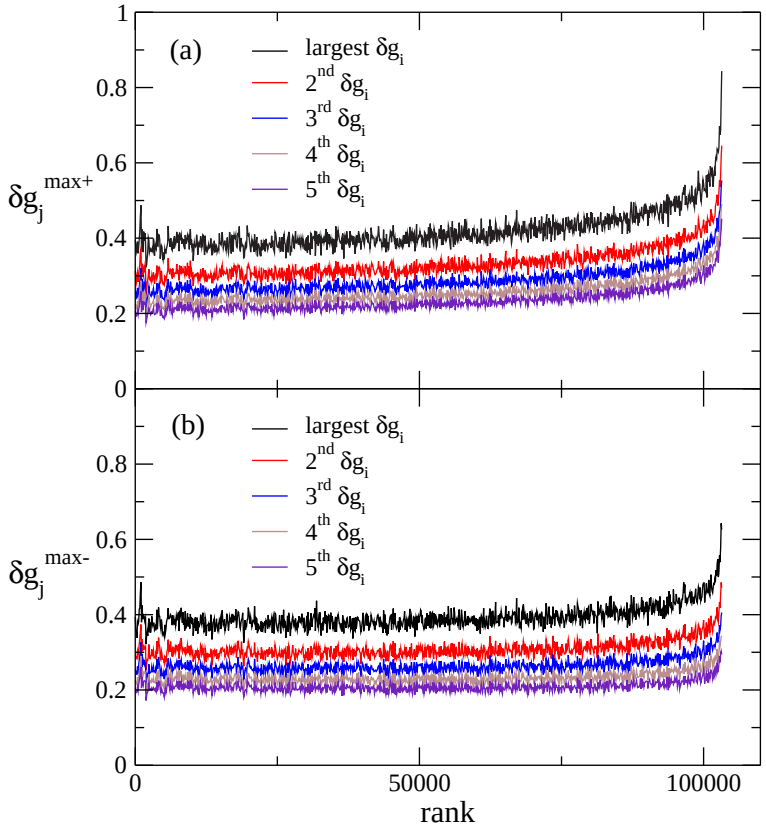

Figure 3.1: Five largest price changes (a) δg_j^{max+} and (b) δg_j^{max-} due to a single trade with (a) the same and (b) the opposite sign as the aggregate return in that 100 tick interval, plotted against the rank of the corresponding aggregate return $|G_j|$ for the combined data of ten Nasdaq stocks in 2002 (smoothed by averaging over 100 intervals). For large $|G_j|$, the size of the δg_j^{max+} increases by a factor of two while the increase in the δg_j^{max-} is slightly smaller. The sum over all five δg_j^{max+} reaches more than three standard deviations for intervals with extremely large $|G_j|$, but the fluctuations in the opposite direction are almost equally large.

the best and second best limit prices [52], the tick return δg_i corresponds to the gap size. We note that the price can (and often does) change between two consecutive market orders due to placement or cancelation of limit orders so that δg_i does not provide a direct estimate of the gap size. We normalize the tick returns δg_i by the standard deviation of the aggregate return G_j for each stock individually so that we can combine the results for different stocks.

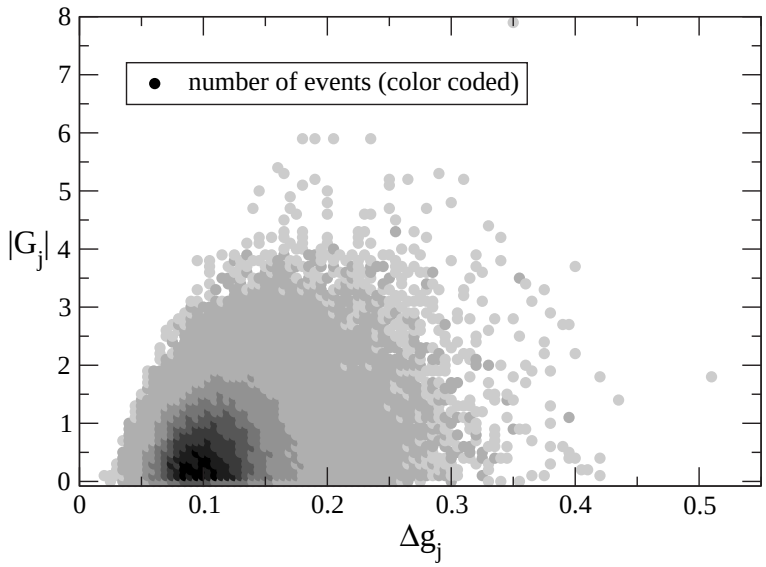

Figure 3.2: Density plot of the 100-trade return $|G_j|$ of ten Nasdaq stocks against the average return of a single trade Δg_j for each interval. The Points are coded from light gray to black indicating the number of events from 1 to more than 500. A linear regression has only a small correlation coefficient $R^2 = 0.07$.

3.3 Influence of the size of tick returns

First, we investigate the question whether large tick returns caused by large gaps in the order book can be responsible for large aggregate returns. To this end, we start with the approximation shown in Eq. (3.3) where a few extremely large tick returns (corresponding to some very large gaps in the order book) lead to a very large aggregate return G_j. In order to test this hypothesis, we analyze the five largest tick returns δg_j^{max+} with the same sign as the aggregate return G_j (i.e. the five largest positive δg_i if $G_j > 0$ and the five largest negative δg_i for $G_j < 0$) in each time interval. To this end, we sort the intervals by $|G_j|$ and plot the δg_j^{max+} against the rank of the interval according to its return $|G_j|$.

Figure 3.1(a) shows the values of these δg_j^{max+} in intervals with small $G_j \approx 0$ on the left while the values for large returns exceeding five standard deviations can be found on the right. Since there are large fluctuations in the data, we smoothed the curves by averaging over 100 intervals. The δg_j^{max+} grow by a factor of two between small and very large returns $|G_j|$. When aggregated, these five largest δg_j^{max+} can reach about three standard deviations, which is almost half of the largest aggregate returns.

In Fig. 3.1(b), we plot the five largest tick returns δg_j^{max-} with the opposite

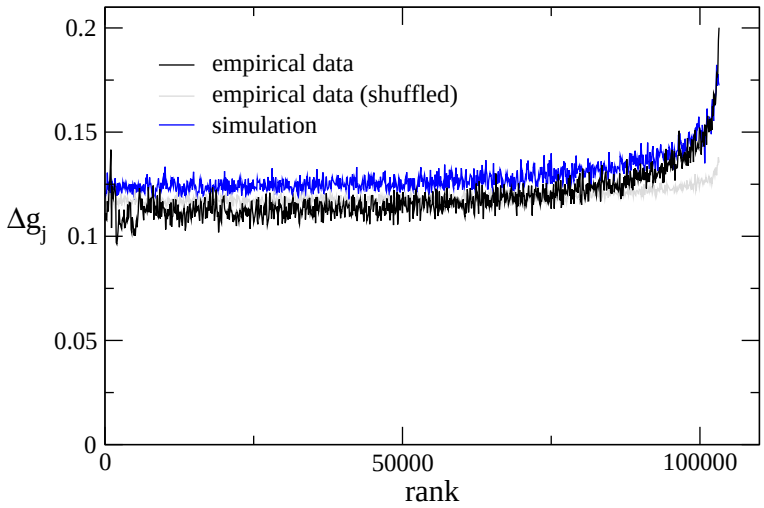

Figure 3.3: Black curve: average tick return Δg_j of ten Nasdaq stocks plotted against the rank of the corresponding aggregate return $|G_j|$, smoothed by averaging over 100 intervals. Going from the smallest returns $|G_j| \approx 0$ to returns larger than five standard deviations, the mean tick return Δg_j increases by a factor of two. Light gray curve: after shuffling the tick returns for each stock, the same curve is only slightly increased for the largest aggregate returns, the effect is much smaller than for the original data. Blue curve (or dark gray): the simulation according to the statistical model discussed in section 3.6 shows a similar behavior as the empirical data, but in the simulation Δg_j is a little larger than the empirical one except for the largest $|G_j|$ where the simulated Δg_j is slightly smaller than the empirical mean tick return.

direction as the aggregate return against their rank. The δg_j^{max-} behave similarly to the δg_j^{max+}, though the increase for large aggregate returns is slightly weaker. However, even for the largest aggregate returns the difference between the δg_j^{max+} and δg_j^{max-} is rather small, so that in addition to the large tick returns in the direction of the aggregate return there are also large tick returns with the opposite sign, reducing the aggregate return.

Our findings suggest that in the data set studied single exceptionally large tick returns might not be the generic mechanism leading to large aggregate returns. This result seems to contradict the experience that there often are price "jumps" due to new public information, e.g. earnings announcements or monetary policy announcements. However, these jumps usually appear together with a largely increased trading activity (volume), so that there are many trades occurring within a short time. Hence, a 100-tick interval could correspond to a rather short time period, so that its aggregate return might look like a price "jump"

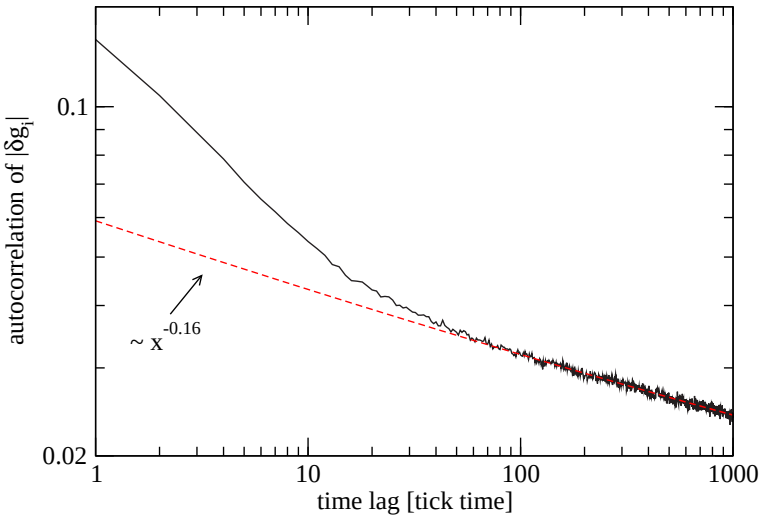

Figure 3.4: Autocorrelation function of the absolute value of the tick return $|\delta g_i|$ averaged over the data of ten Nasdaq stocks in 2002. The function shows a power law decay in tick time proportional to $\Delta i^{-0.16}$ for large Δi.

in real time but actually consists of many not too large tick returns, which is visible in tick time. This interesting subject could be investigated in a further analysis.

In the following, we want to focus not on the extreme tick returns, but on the influence of their mean value. More precisely, we analyze Eq. (3.5) and the mean tick return Δg_j of all non-zero $|\delta g_i|$ in the interval I_j as defined in Eq. (3.4). A density plot of $|G_j|$ against Δg_j is shown in Fig. 3.2. It seems that extremely large returns G_j correspond to larger average tick returns Δg_j, but the broad distribution suggests that the explanatory power of Δg_j alone for the aggregate return G_j is small, which is confirmed by the low correlation coefficient $R^2 = 0.07$ of a linear regression.

In order to clarify the relation between the extreme values of $|G_j|$ and Δg_j, we sort the intervals by $|G_j|$ and plot Δg_j against the rank of the interval according to its return $|G_j|$. In Fig. 3.3 (black curve), we see that large returns $|G_j|$ coincide with larger tick returns as Δg_j changes by a factor of two from very low aggregate returns to large returns of several standard deviations. In comparison with the largest tick returns δg_j^{max+} shown in Fig. 3.1, the change of a factor of two is similar, but the mean Δg_j is two to four times smaller than the largest tick returns.

This finding can be explained by the presence of autocorrelations in the time series of δg_i, which can be illustrated when we shuffle the data for each stock by exchanging each tick return with another tick return randomly chosen from

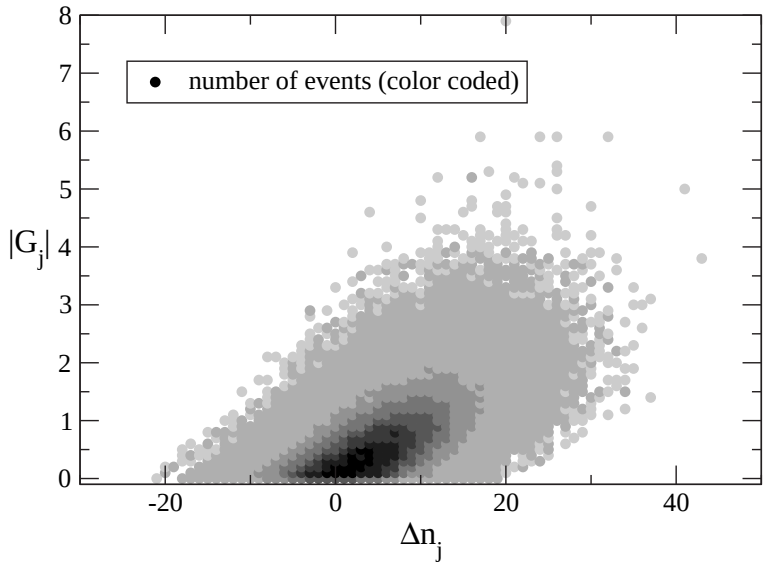

Figure 3.5: Density plot of the aggregate return $|G_j|$ against the difference Δn_j between the number of tick returns with the same and with the opposite direction as the aggregate return, for ten Nasdaq stocks. The points are coded from light gray to black indicating the number of events from 1 to more than 600. A linear regression has a large correlation coefficient $R^2 = 0.32$.

the entire time series. The light gray curve in Fig. 3.3 shows that for shuffled data Δg_j increases only marginally for large aggregate returns, suggesting that autocorrelations of the tick returns have a strong influence on the mean tick return size Δg_j. Indeed, we find that the absolute values $|\delta g_i|$ of the tick return are long-range correlated in tick time with a correlation function decaying like $\Delta i^{-0.16}$ for large time lags $\Delta i = |i_1 - i_2|$, as shown in Fig. 3.4. If these correlations are destroyed by shuffling, in each interval of 100 trades only a few large tick returns remain so that the average over these 100 tick returns approximates the global mean of all tick returns in the data set.

In contrast, in the empirical, unshuffled data correlations lead to intervals where many tick returns are large, so that the average tick return size is also large. The average tick return size Δg_j can well characterize the interval only because these autocorrelations exist. It turns out that the increase of Δg_j by a factor of two is the main effect where the original empirical data deviate significantly from shuffled data. Hence, we suggest that fluctuations of the tick return size are responsible for the non-Gaussian fluctuations of the aggregate return.

Using Eq. (3.5), we can estimate whether the change by a factor of two of the average tick return alone is enough to explain large aggregate returns G_j of more than five standard deviations. To this end, we focus on the intervals with the

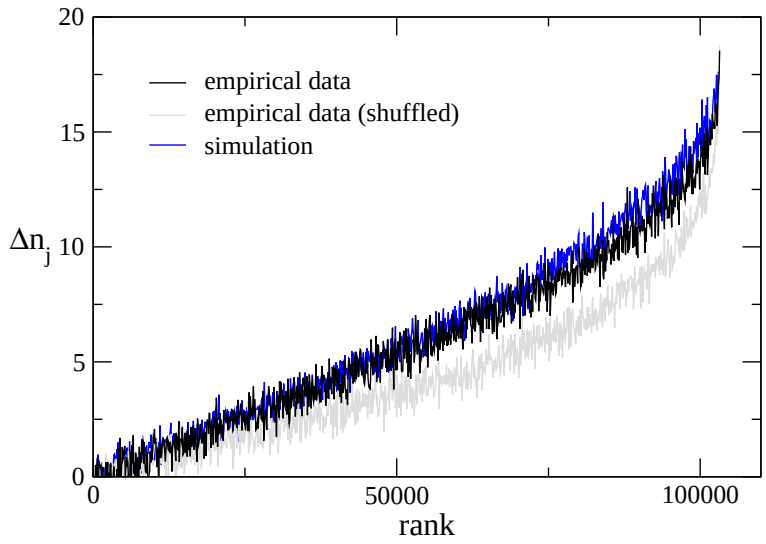

Figure 3.6: Black curve: the sign-adapted number difference Δn_j is plotted against the rank according to the aggregate return $|G_j|$ for ten Nasdaq stocks, smoothed by averaging over 100 intervals. Δn_j grows from zero to 18. The relation between Δn_j and the rank seems to be linear except for the largest 15% of the aggregate returns. A simulation [blue curve (or dark gray)] using a normal distribution for ΔN_j leads to nearly the same dependance on the rank. For shuffled data (light gray curve), the curve is slightly flatter, but the difference is not large.

50 largest aggregate returns ranging from approximately four to almost eight standard deviations. Here, we find that Δg_j fluctuates between 0.14 and 0.35. Assuming uncorrelated returns, ΔN_j should be of the order $\sqrt{N} \approx 10$ if each trade would lead to a price change, but normal fluctuations could well lead to ΔN_j twice as large as $\sqrt{N}$, so that large tick returns together with fluctuations in the number difference could explain the large aggregate returns we find in our data set.

Thus, we find that in intervals with 100 trades large $|G_j|$ do not mainly depend on single extremely large tick returns. It rather turns out that correlations between the tick returns lead to large average tick returns Δg_j in an interval, and the fluctuations of Δg_j can account for the non-Gaussian distribution of the aggregate returns.

3.4 Number difference

The diffusion process of aggregate returns is not only influenced by the step width (i.e. the tick return size), but also by the direction of the steps. Therefore, we now analyze the influence of the number difference ΔN_j in Eq. (3.5). In order to treat positive and negative aggregate returns in the same analysis, it is useful to replace ΔN_j by the sign-adapted number difference

$$\Delta n_j = \mathrm{sgn}(G_j)\Delta N_j \quad . \tag{3.8}$$

A positive value of Δn_j indicates that the price tends to move in one specific direction leading to an aggregate return with the same sign. Δn_j can be negative if there are a few large tick returns determining the direction of the aggregate return, but also many small tick returns with the opposite direction which do not affect the aggregate return very much. Figure 3.5 shows a density plot of the aggregate return $|G_j|$ against the sign-adapted number difference Δn_j. A linear regression yields an R^2 of 0.32, a large correlation coefficient confirming the visual impression that Δn_j and $|G_j|$ are strongly connected. We can also see that Δn_j is mostly positive for large returns G_j, so that each large price change is accompanied by a certain sign-adapted number difference Δn_j.

We now plot, in Fig. 3.6, Δn_j against the rank according to $|G_j|$. We find that except for the largest (approximately 15%) of the aggregate returns, Δn_j grows linearly with the rank while in Fig. 3.3 Δg_j remained almost constant in that region. For the largest ranks, Δn increases more rapidly, so that all in all the smoothed curve (averaged over 100 intervals) grows from zero to 18 between very small and extremely large aggregate returns. Thus, in intervals with very large returns there are approximately 18 trades pushing the price in one direction (assuming that all other trades cancel each other), so that even with rather small tick returns this can lead to large returns in aggregation. Focusing on the 50 largest G_j, we find that Δn_j ranges from 4 to 41, most of them clearly above the expected standard deviation of 10 when assuming uncorrelated returns and $n_j = N$.

Thus, the fluctuations of Δn_j around the mean value are crucial for getting large aggregate returns. The number difference seems to be the main mechanism affecting the aggregate return since it changes much more drastically than the tick return size when the aggregate return increases. On the other hand, when we compare the results to the analysis with shuffled data (light gray curve in Fig. 3.6), it turns out that this effect is very similar to what happens with random price changes. Hence, the basic movement of the aggregate return seems to depend mostly on the number difference, but the non-Gaussian large

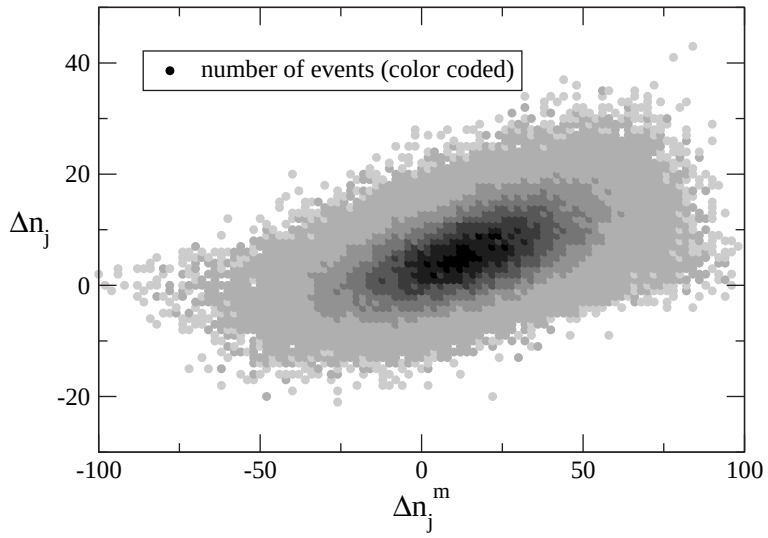

Figure 3.7: Comparison between sign-adapted number difference Δn_j and market order difference Δn_j^m for ten Nasdaq stocks. The Points are coded from light gray to black indicating the number of events from 1 to more than 200. The correlation coefficient of a linear regression yields $R^2 = 0.29$, thus there is a strong connection between the two quantities. On the other hand, the events scatter widely so that small Δn are often linked with large Δn_j^m and vice versa.

aggregate price changes only occur if the tick returns are large.

3.5 Market order signs and direction of tick returns

It is known that the signs of market orders are strongly correlated [31, 73] which means that there is a large probability that a buy market order will be followed by another buy market order. Thus, it is probable that large number differences in the direction of tick returns are caused by large numbers of equally signed market orders. In order to analyze the relation between the number difference and the market order flow, we define the difference Δn_j^m between the number n_j^{m+} of market orders with the same direction as G_j and the market orders with opposite direction n_j^{m-}:

$$\Delta n_j^m = n_j^{m+} - n_j^{m-} \quad . \tag{3.9}$$

In Fig. 3.7 we plot the sign-adapted number difference Δn_j against the market order difference Δn_j^m. We find a strong correlation between Δn_j and Δn_j^m; a linear regression yields a correlation coefficient R^2 of 0.29. However, there

are also large fluctuations suggesting that the number difference is also due to order book dynamics, namely limit order placement and cancelation as well as asymmetries in the order book. A model for price formation due to these quantities was recently proposed by Mike and Farmer [101].

3.6 Distribution of aggregate returns and a statistical model

In the first part of this chapter, we analyzed the mechanism leading to large aggregate returns and showed that the varying step width Δg_j accounts for the non-Gaussian behavior of the diffusion process of price movements. Now we want to use our results in a statistical model and reproduce the cumulative distribution function of the absolute value of the aggregate return $|G_j|$.

The model given by Eq. (3.5) belongs to the well-known class of stochastic volatility models (see e.g. [12]) consisting of a noise term multiplied by a time-dependent volatility giving the magnitude of the fluctuations. In the present chapter, the model is based on a microscopic description of the price process, so that we can fit the microscopic quantities determining the aggregate return in order to estimate the parameters of the model. In this approach the model is parameter free in the sense that there are no parameters fitting the aggregate returns directly, though we fit the distributions of its determinants like the step width Δg_j and the number difference ΔN_j. We also discuss corrections to the model by including the tick return asymmetries according to Eq. (3.7).

We first analyze the distributions of Δg_j and ΔN_j. Figure 3.8(a) shows the cumulative distribution of Δg_j in a log-linear plot. The approximately straight line suggests that the tail follows an exponential distribution which can be well fitted with $P(x > \Delta g_j) = e^{-a(x-x_0)/\Delta\bar{g}}$ where $\Delta\bar{g} \approx 0.12$ is the average of all Δg_j and the parameters are $a = 3.6$ and $x_0 = 0.094$. In the region of the smallest values of $\Delta g_j \lesssim x_0$, the limited tick sizes of the different stocks lead to a plateau. In section 3.4 we already found evidence that ΔN_j behaves similarly to uncorrelated data since in Fig. 3.6 the shuffled data shows almost the same dependence on the rank of the corresponding $|G_j|$. Figure 3.8(b) shows that indeed ΔN_j can be well described by a Gaussian noise with mean 0.24 and standard deviation 9.0.

In order to analyze the accuracy of the approximation given in Eq. (3.5), we simulate two independent time series according to the fitted functions for Δg_j and ΔN_j and build the aggregate return G_j as the product of Δg_j and ΔN_j. In Figure 3.9 we can compare the empirically found cumulative distribution

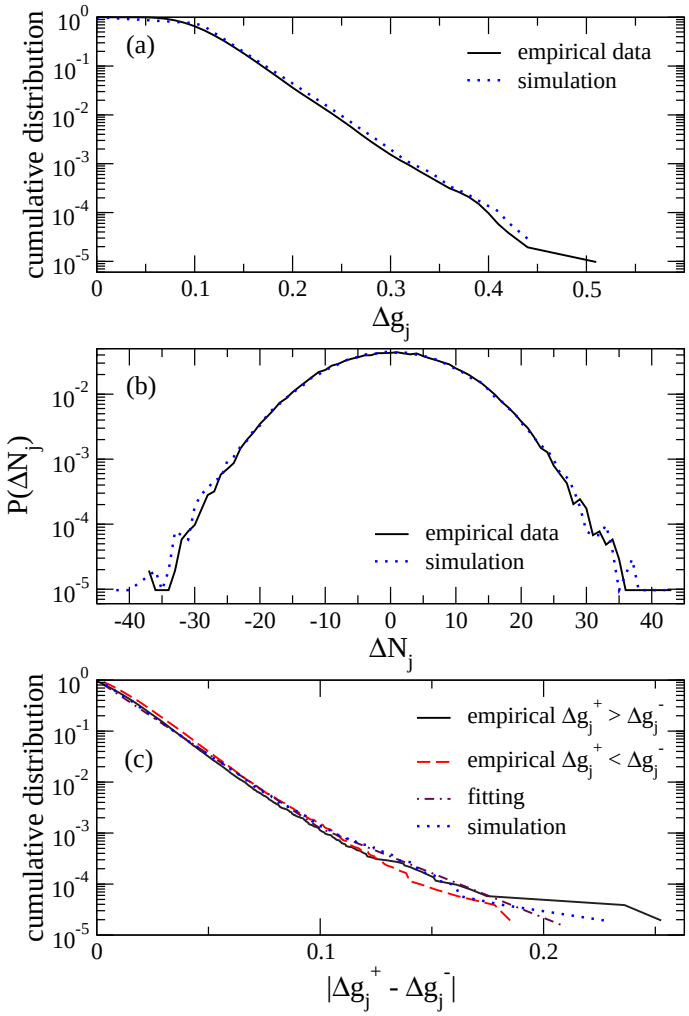

Figure 3.8: Estimation of the parameters for the simulation (results shown as dotted lines) from empirical data for ten Nasdaq stocks. (a) The tail of the cumulative distribution of Δg_j (line) can be well fitted with $P(x > \Delta g_j) = e^{-a(x-x_0)/\Delta\bar{g}}$ where $\Delta\bar{g} \approx 0.12$ is the average of all Δg_j and the parameters are $a = 3.6$ and $x_0 = 0.094$. For $\Delta g_j \lesssim x_0$ the limited tick size leads to a plateau. (b) The probability distribution of ΔN_j (line) follows in good approximation a normal distribution with mean 0.24 and standard deviation 9.0. (c) As a rough approximation, the average of the cumulative distribution of the positive (line) and negative (dashed line) values of $\Delta g_j^+ - \Delta g_j^-$ are parameterized proportional to two exponential functions $e^{-a_{1,2}x/\Delta\bar{g}}$ for $|\Delta g_j^+ - \Delta g_j^-| \lesseqgtr 0.1$, with $a_1 = 8.0$ and $a_2 = 4.8$ (dash-dotted line). The simulation (dotted line) uses the adapted $a_1 = 9.0$ and $a_2 = 2.0$ in order to compensate the change in the distribution after taking into account $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$.

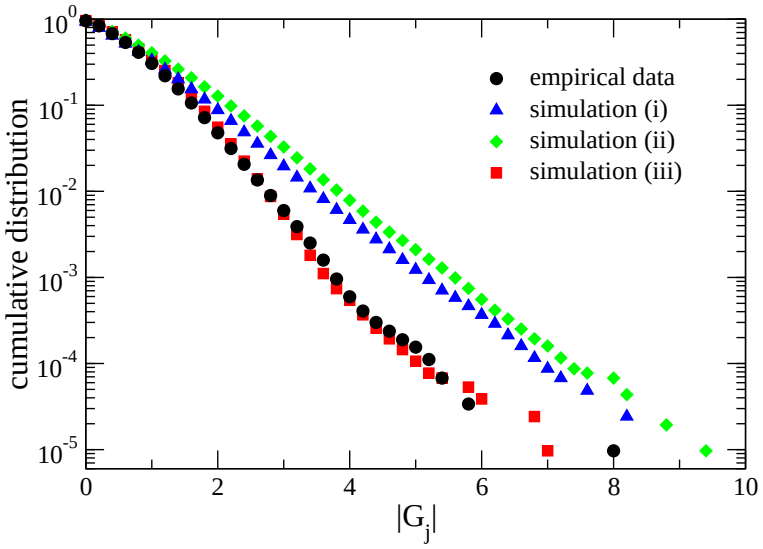

Figure 3.9: Cumulative distribution of the empirical aggregate return (circles) obtained from ten Nasdaq stocks in comparison with different simulations. (i) The simulation according to Eq. (3.5) (triangles) leads to a reasonable approximation of the empirical data, but it overestimates the probability of large returns. (ii) The distribution becomes a little broader if we add the tick return asymmetry $\Delta g_j^+ - \Delta g_j^-$ according to Eq. (3.7) and simulate independent quantities (diamonds). (iii) The simulation (squares) matches the empirical data very well if we incorporate correlations by generating $\Delta g_j^+ - \Delta g_j^-$ according to the conditional expectation value $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$.

of aggregate returns $|G_j|$ (circles) to the results of this simulation (triangles). The simulation of Eq. (3.5) leads to a reasonable agreement with the actual aggregate return, but it overestimates the probability of large aggregate returns. We note that the parameters of the simulation are completely determined by the empirically found distributions of Δg_j and ΔN_j, so that in this sense the simulation of $|G_j|$ has no free parameters.

In the following, we want to address the remaining deviations of the simulation from the empirical data. Eq. (3.7) gives an exact formula for G_j and provides a good parametrization for the error term which reads

$$G_j - \Delta g_j \Delta N_j = \frac{2 n_j^+ n_j^-}{n_j} (\Delta g_j^+ - \Delta g_j^-) \tag{3.10}$$

We find that the term $2 n_j^+ n_j^- / n_j$ has no systematic influence on the aggregate return since it shows almost no dependence on the rank according to the aggregate return. In the following, we thus approximate it by its average value

$\left\langle 2n_j^+ n_j^- / n_j \right\rangle = 28.7$, so that the error term is determined by the asymmetries $\Delta g_j^+ - \Delta g_j^-$ in the mean tick return size.

The cumulative distribution of $\Delta g_j^+ - \Delta g_j^-$ is shown in Fig. 3.8(c). The main part of the distribution could be well fitted by an exponential function, but in the tail the distribution becomes broader. Thus, we add the term with $\Delta g_j^+ - \Delta g_j^-$ to our simulation by creating a third independent time series according to the empirical distribution of $\Delta g_j^+ - \Delta g_j^-$. Figure 3.9 (diamonds) shows that this leads to an even broader distribution of the aggregate return. Since the difference to the distribution according to Eq. (3.5) is small, the tick return asymmetry seems to have only a small influence on the aggregate return.

A more accurate agreement with the empirical data can be obtained by taking into account correlations between the quantities involved in the process. The correlation coefficients between them are shown in the following table where the correlations between the absolute values are shown in brackets:

	ΔN_j	$\Delta g_j^+ - \Delta g_j^-$	$\Delta g_j \Delta N_j$
Δg_j	-0.02 (-0.07)	-0.01 (0.37)	-0.01 (0.34)
ΔN_j	1	-0.35 (0.01)	0.95 (0.87)
$\Delta g_j \Delta N_j$	0.95 (0.87)	-0.41 (0.02)	1

Δg_j and $|\Delta N_j|$ show slightly negative correlations which might suggest that people act more cautiously when large tick returns indicate a low liquidity. In these times, traders try not to place too many consecutive orders with the same sign because they know that it could lead to a large price change and increased trading costs. Furthermore, the strong anti-correlations between ΔN_j and $\Delta g_j^+ - \Delta g_j^-$ also indicate cautious traders: If there are large asymmetries, so that e.g. the positive tick returns are much larger than the negative ones, people tend to use the higher liquidity in negative direction so that in these times they sell more often than they buy. For an analysis of the relation between liquidity imbalance and market efficiency, see e.g. [102]. The large correlations between Δg_j and $|\Delta g_j^+ - \Delta g_j^-|$ show that we can expect large variations of the tick return in the positive and negative directions when the tick return is in general large.

We now want to incorporate correlations in our simulation. The strongest non-trivial correlations appear between $\Delta g_j \Delta N_j$ and $\Delta g_j^+ - \Delta g_j^-$ including also some of the correlations between $\Delta g_j^+ - \Delta g_j^-$ and Δg_j as well as ΔN_j. However, it turns out that the conditional expectation value $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$

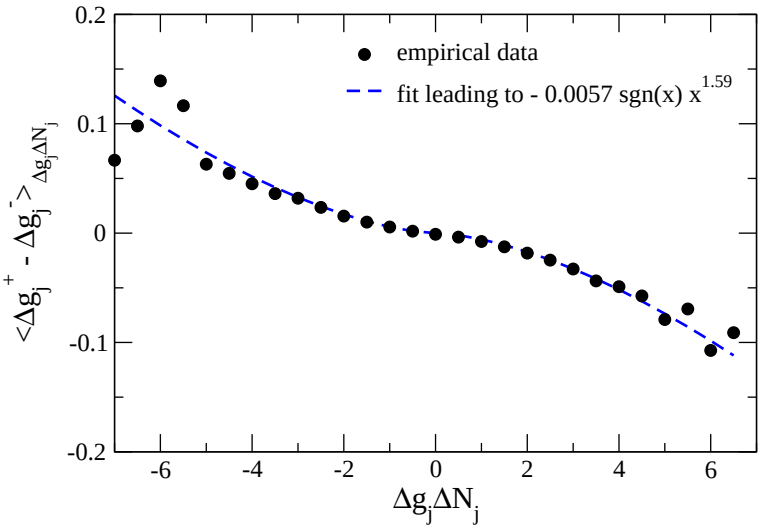

Figure 3.10: Conditional expectation value $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$ plotted against $\Delta g_j \Delta N_j$ (circles), obtained from the data of ten Nasdaq stocks. A fit leads to $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j} \approx -0.0057 \cdot \mathrm{sgn}(\Delta g_j \Delta N_j) \cdot (\Delta g_j \Delta N_j)^{1.59}$ (dashed line). The tick return asymmetry $\Delta g_j^+ - \Delta g_j^-$ is strongly correlated with the mean tick return size Δg_j and strongly anti-correlated with the number difference ΔN_j. Using the conditional expectation value in the simulation incorporates these correlations which allows the reproduction of the distribution of aggregate returns.

is nonlinear, as seen in Fig. 3.10 (circles) where it is plotted against $\Delta g_j \Delta N_j$. The function can be well fitted by $-\mathrm{sgn}(x)\alpha|x|^\beta$ with $\alpha = 0.0057$ and $\beta = 1.59$ (dashed line).

In order to incorporate this conditional expectation value into the simulation, we first create three independent time series for Δg_j, ΔN_j, and $\Delta g_j^+ - \Delta g_j^-$. Then, for each j we add the conditional expectation value $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$ to $\Delta g_j^+ - \Delta g_j^-$, according to the value of $\Delta g_j \Delta N_j$ for that j. This method leads to a different distribution for $\Delta g_j^+ - \Delta g_j^-$ than the initial one, so that we can not anymore generate $\Delta g_j^+ - \Delta g_j^-$ from the unconditional empirical distribution. As a rough approximation, we parameterize this distribution by two exponential functions $e^{-a_{1,2}x/\Delta\bar{g}}$ for $\Delta g_j^+ - \Delta g_j^- \lesseqgtr 0.1$. Then, we adapt the factors in the exponent in such a way that the resulting unconditional distribution fits the empirical one (a fit to the empirical distribution yields $a_1 = 8.0$ and $a_2 = 4.8$, for the simulation we use the adapted $a_1 = 9.0$ and $a_2 = 2.0$, compare Fig. 3.8(c)). The resulting distribution of G_j does not depend very much on the exact values of $a_{1,2}$.

The effect of the correlations represented by the conditional expectation value $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$ is very large and leads to a cumulative distribution of $|G_j|$ (squares in Fig. 3.9) very similar to the empirical one (circles). It is worth noting that now the largest events are not anymore necessarily the ones with the largest values of $\Delta g_j \Delta N_j$. Due to the anti-correlations expressed in $\left\langle \Delta g_j^+ - \Delta g_j^- \right\rangle_{\Delta g_j \Delta N_j}$, very large values of $\Delta g_j \Delta N_j$ can lead to relatively large values of $\Delta g_j^+ - \Delta g_j^-$ of the opposite sign reducing the aggregate return.

In addition to the distribution of the aggregate return, the simulation also agrees with other properties of the empirical data we found earlier in this chapter. In Fig. 3.3 and 3.6 we also plotted the data from the simulation against the rank according to the aggregate return $|G_j|$. For ΔN_j the simulation matches the empirical data very well, while in Figure 3.3 we see that the simulated Δg_j shows the same dependence on the rank as the empirical data, but it is generally a little larger than the real one except for the largest aggregate returns, which might be due to the cutoff around 0.094 we used in the simulation of the distribution of Δg_j. We also find that the role of $\Delta g_j^+ - \Delta g_j^-$ in determining large aggregate returns is a little overestimated by our simulation, but the simulation covers the main features of the empirical data although we neglected many of the subtle relations between the different quantities.

3.7 Discussion and Conclusion

Our findings presented in this chapter can be divided into two parts: First, we showed that the movement of stock prices in intervals with a constant number of trades can be understood as a diffusion process with a varying step width, similar to the findings of Plerou *et al.* for time intervals [91]. While Plerou *et al.* use the shape of the distribution of mean squared tick returns to explain the distribution of aggregate returns, we render this picture more precisely by specifically studying the intervals with the largest aggregate returns. By analyzing how each aggregate return is actually composed, we find that Gaussian fluctuations of the number difference determine the basic price movement, but the non-Gaussian large price changes occur only if a large number difference coincides with a large mean tick return size. Though the mean tick return size is not exclusively responsible for the occurrence of large returns, we confirm the result of [91] that the non-Gaussian shape of the mean tick return size is an important determinant of non-Gaussian aggregate returns. We also find that the large influence of the tick return size is caused by its autocorrelations assuring that in a 100 tick interval one can find many large tick returns so that the mean

value of the tick return can be large.

In the second part of this chapter, we found that the distribution of aggregate returns can be reasonably approximated by simulating the microscopic quantities mean tick return size and number difference according to their empirically found distributions. A more accurate agreement can be obtained by taking into account asymmetries in the tick return size in the positive and negative directions as well as correlations between the different quantities.

In summary, we found evidence that price fluctuations in intervals with a constant number of trades can be described by a diffusion process with a varying step width. The long-term autocorrelations in the tick return size make sure that periods, where the price change due to a trade is large, last long enough to cause large aggregate returns in intervals with many trades. Our results suggest that the power law distribution of aggregate returns might not be universal but rather depends on a more complicated mechanism which is a combination of the dynamics of the trading frequency, the dynamics of the step width and the Gaussian process of the step direction.

4 Price impact, liquidity, and large stock price changes

In the previous chapter, we performed a statistical analysis of the tick return time series, leading to a better understanding of the mechanisms behind large aggregate returns. We studied the fluctuations of the mean tick return size and the number difference in order to model the fluctuations of the aggregate return. In this way, we described how large price changes emerge from an interplay between these quantities together with asymmetries between positive and negative tick returns in the respective time interval.

When studying financial data sets, people often focus on the description of fluctuations, for instance when describing the stock return distribution [14, 15, 16, 17, 35, 36, 37, 38, 39, 40, 41]. In the present chapter, we use a different approach to further analyze the occurrence of large stock price changes. Instead of a *descriptive* study, we investigate the *reasons* that lead to large returns. Hence, we do not ask *how* large returns are composed, but *why* these large returns occur.

Besides the influence of news, a reasonable assumption is that price movements are caused by an imbalance between supply and demand: if there are more people who want to buy than to sell, prices will move up. This phenomenon can be quantified by the price impact function [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], which describes the price change as a conditional expectation value of volume imbalance, i.e. the difference between the volume of buy and sell market orders in a given time interval. Hence, it quantifies the price impact as it happened on average in response to a certain volume imbalance.

The price impact function is a response function, which in physics describes the response of a system to an external influence. For instance, the susceptibility quantifies how the magnetization changes in response to a magnetic field. If the system is cooled down and its temperature approaches the Curie temperature, the correlation length becomes very large so that the magnetizations of many subsystems are coupled. The resulting collective behavior of the subsystems lead to large global fluctuations as well as a strong response to an external

influence. Hence, resulting from the divergent correlation length, the susceptibility diverges, which for both quantities can be described by a power law. Hence, from physics one knows a mechanism that generates power laws and large fluctuations, so the study of response functions is a promising approach for the explanation of large price changes.

Indeed, the theory of Gabaix *et al.* [27] uses the price impact function to find a quantitative explanation for large price fluctuations. After approximating the price impact function with a time-independent square root function, these authors conclude that the power law distribution of returns with exponent three derives from the cumulative distribution of the order flow (i.e. the volume traded in a given time interval), which can be described by a power law with exponent 1.5 [106]. In this model, a large order flow leads to large volume imbalances that cause large price changes via the price impact function.

The model has been criticized by Farmer *et al.* [53], who questioned the square root fit of the price impact function, as this function varies for different assets and the test method used by Gabaix *et al.* to estimate the fitting might not be appropriate in the presence of correlations. There is also a discussion about the nature of price impact [31, 73, 102], asking whether it is fixed and temporary [31] or variable and permanent [73, 102]. This question becomes important if one tries to explain how uncorrelated returns can emerge from long range correlated orders [31, 73, 102], which will be discussed in the next chapter. If the order imbalance would permanently change the price via a fixed price impact function, the return would inherit the order correlations and would be long-term correlated as well. Bouchaud *et al.* [31] argue that a fixed price impact function can be reconciled with uncorrelated returns if the price impact is only temporary, so that the price change due to an order vanishes some time after due to market mechanisms. In contrast, the authors of [73, 102] show that a permanent price impact does not contradict uncorrelated returns if the price impact is variable and changes over time.

Here, we want to study price impact in detail to understand the mechanism leading to large price fluctuations [103, 104]. To this end, we use the same data as in chapter 3, containing all orders from the Island ECN order book in the year 2002 for the ten most frequently traded Nasdaq stocks [100]. In contrast to the previous chapter, we do not focus on the tick return time series obtained from this data, but use the additional information about the trade volume and complete information about all orders present in the market.

The remainder of this chapter is organized as follows. Section 4.1 studies the average price impact function calculated from market orders, which in section 4.2

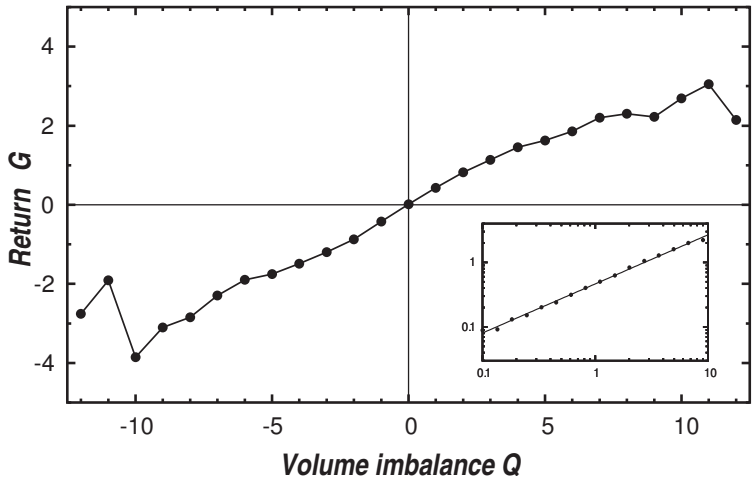

Figure 4.1: The price impact function $I_{\text{market}}(Q)$ describes returns between the beginning and end of five minute intervals in response to the volume imbalance in the same time interval. It is a monotonously increasing and concave function of the signed market order volume. A logarithmic plot (inset) shows that the function can be fitted by a power law. Adapted from [30].

is compared to different definitions of a virtual price impact function calculated from limit orders in the order book. In section 4.3, we use the average price impact function to show that a large volume imbalance alone cannot explain the occurrence of very large price changes. This explanation is provided in Section 4.4 where a time-varying price impact function is defined. We conclude with a discussion of the results in section 4.5.

4.1 Price impact of market orders

In order to describe how on average the price reacts to a traded volume, one defines the price impact function of market orders as the conditional expectation value

$$I_{\text{market}}(Q) = \langle G_{\Delta t}(t) \rangle_Q \quad . \tag{4.1}$$

It describes the average relation between the return $G_{\Delta t}(t) = \ln S(t + \Delta t) - \ln S(t)$ in a given interval of $\Delta t = 5\text{min}$ and the volume imbalance Q in the same time interval [1]. In contrast to the previous chapter, we now analyze time intervals of $\Delta t = 5\text{min}$ length. Though the analysis of intervals with a constant number of trades has the great advantage that one has not to deal with

[1] We do not include market orders executing "hidden" limit orders in the definition of $Q(t)$ as we want to compare our results with the order book that only contains "visible" orders.

the trading frequency so that the influence of other quantities is more clearly visible, it is also reasonable to study real time intervals since they agree with our "natural" experience. After all, people think and act in real time instead of ticks.

The volume imbalance Q in a time interval is the sum of all signed market order volumes executed between t and $t+\Delta t$. For the order book data, the sign of an order is stored in the data set indicating whether it is a buy or sell order. We also want to study the TAQ data base for the 44 most frequently traded Nasdaq stocks, which does not contain information about the direction of a trade. However, the sign of a transaction can be determined by the Lee and Ready algorithm [105], which compares the transaction price to the midquote price $S_{\mathrm{M}}(t) = \frac{1}{2}(S_{\mathrm{bid}}(t) + S_{\mathrm{ask}}(t))$. The sign is positive for buy orders (transaction price larger than midquote price) and negative for sell orders (transaction price smaller than midquote price). With the order book data, we tested the accuracy of the Lee and Ready algorithm by first computing the results using the algorithm and then performing the same analysis with respect to the buy and sell information contained in the order book data base. On the level of single events, the transaction directions from the Lee and Ready algorithm deviate from the exact ones, but upon averaging both methods yield a nearly identical price impact function.

For the analysis of TAQ data, we choose $S(t)$ as the price at which the last transaction before time t took place. For the analysis of order book data, $S(t)$ is chosen as the midquote price $S_{\mathrm{M}}(t)$ as we want to make comparisons to hypothetical price impacts calculated from the order book.

Similar to the previous chapters, returns G are normalized by their standard deviation σ_G which is well defined because the cumulative distribution function of returns follows a power law with exponent larger than two. Since trading volume is described by a cumulative distribution with power law exponent $\zeta_V = 1.5$ [106], its standard deviation is not well defined. Hence, the volume imbalance Q is normalized by its first centered moment $\sigma_Q = \langle |Q - \langle Q \rangle| \rangle$.

The functional form of $I_{\mathrm{market}}(Q)$ for the Island order book data is shown in Fig. 4.1. In order to get good statistics especially for large volume imbalances, we aggregate Eq. (4.1) over all ten stocks in our data set. The shape of the price impact function is in general agreement with the results [22, 23, 24, 26], we find that $I_{\mathrm{market}}(Q)$ is a concave function of volume imbalance [20], which can be well fitted by a power law $G = 0.48\, Q^{0.76}$.

Earlier studies show that the power law exponent characterizing the price impact function depends on the time horizon as well as on the market studied. Plerou *et*

al. [23] find that the exponent generally tends to increase for an increasing time horizon. On very small scales, i.e. on a tick by tick basis, the exponent is very small [26] or the price impact function can be characterized by a logarithm [28]. On an intermediate scale, the exponent was found to be 0.5 for 15 minute intervals [23, 27]. This value of 0.5 is also predicted by Zhang [107] using a simple market maker model.

Analyzing the TAQ data base for the year 1997 instead of the years 1994 and 1995 as in [23, 27] and for time intervals of five minutes as compared to the fifteen minute intervals in [23, 27], we find an exponent 0.58 for transaction price changes and 0.75 for midquote price changes. For the Island ECN data, the exponent is 0.76 for midquote returns and 0.73 for transaction returns, both calculated on a time scale of five minutes. The larger exponent for midquote prices compared to the exponent for transaction prices seems to contradict the intuition that the price impact for transaction prices should be larger than the one for midquote prices. However, most volume imbalances in a five minute interval are smaller than $Q = 1$, and on a logarithmic scale these values constitute a large part of the bins used for a fitting, so that values of $Q < 1$ contribute significantly to a logarithmic fit. For $Q < 1$, one has $|Q|^{\alpha} > |Q|^{\beta}$ for $\alpha < \beta$, and the price impact for transaction prices is indeed stronger than the price impact for midquote prices, in agreement with the intuition.

The concave shape of the function is very surprising: This type of price impact would theoretically be an incentive to make large trades as they would be less costly than many small ones. In contrast, a convex price impact would encourage a trader to brake up a large trade into several smaller ones, which is what actually happens.

4.2 Order book and virtual price impact

The above definition of the price impact function for market orders could be called an *ex-post* definition, since it calculates the price impact from information about how the price actually changed in the past. Next, we want to use order book information to find an *ex-ante* definition, allowing to forecast the expected price change from a virtual price impact function [30]. These results might help understand the counterintuitive convex shape of $I_{\text{market}}(Q)$.

To make the order book information amenable to a statistical analysis, we calculate at the beginning of each time interval and for each stock k the current order book as a density function $\rho^{k}_{\text{book}}(\gamma_i, t)$. Due to the complexity of this calculation, we use a discrete coordinate γ_i to obtain the order book from the data

structure by sorting orders with respect to their limit prices and aggregating the number of shares on a lattice with spacing $\Delta\gamma$. For each price S_{limit}, at which a limit order is placed, the coordinate γ_i is defined as

$$\gamma_i = \begin{cases} [(\ln(S_{\text{limit}}) - \ln(S_{\text{bid}}))/\Delta\gamma]\,\Delta\gamma & \text{limit buy order} \\ [(\ln(S_{\text{limit}}) - \ln(S_{\text{ask}}))/\Delta\gamma]\,\Delta\gamma & \text{limit sell order} \end{cases} . \tag{4.2}$$

Here, the function $[x]$ denotes the smallest integer larger than x. We define the density function such that $\rho^k_{\text{book}}(i\Delta\gamma, t)\Delta\gamma$ is the total volume in the price interval $[(i-1)\Delta\gamma, i\Delta\gamma]$ in the order book, where i is an integer. In our analysis, we chose $\Delta\gamma = 0.3\ \sigma_G$ as a compromise between computational speed and accuracy. We note that throughout the chapter γ is measured in units of σ_G.

Next, we want to compare the actual price impact $I_{\text{market}}(Q)$ to a virtual price impact function calculated from the average order book $\rho_{\langle\text{book}\rangle}(\gamma_i) = \langle\rho^k_{\text{book}}(\gamma_i, t)\rangle$, where $\langle ...\rangle$ denotes an average over both time t and different stocks k. The average order book is characterized by a flat maximum at $\gamma_i \approx 1$ and a slow decay for large γ_i. Its overall shape agrees with the results of [108, 109, 110].

In order to obtain a price impact function from the average order book, we calculate the market depth for a given return and invert this relation. Market depth is a liquidity measure that denotes the order flow innovation needed to change the price a given amount. We imagine a trader who wants to buy a volume Q of stocks and has only offers from the order book available. Beginning at the ask price, she executes as many limit orders as necessary to match her market order, and changes the ask price by an amount G. Traded volume (or market depth) $Q_{\langle\text{book}\rangle}(G)$ and return G are related by

$$Q_{\langle\text{book}\rangle}(G) = \sum_{\gamma_i \le G} \rho_{\langle\text{book}\rangle}(\gamma_i)\Delta\gamma\ . \tag{4.3}$$

By inverting Eq. (4.3), we define the virtual price impact $I_{\langle\text{book}\rangle}(Q)$ with respect to the average order book. Here, we assume that the bid-ask spread remains constant and that the midquote price changes by the same amount as the ask price. According to the above definition, the virtual price impact $I_{\langle\text{book}\rangle}(Q)$ describes the price change due to a single market order of arbitrary size. Now, we want to compare the virtual price impact with $I_{\text{market}}(Q)$, calculated as a function of the volume imbalance $Q(t)$ aggregated over a five minute interval. Predicting a return due to a time aggregated volume imbalance by using the virtual price impact function is an approximation that is only justified if (i) the order book is symmetric with respect to its buy and sell side and if (ii) the influence of its nonlinearities on the final result is small. The assumption of a buy-sell symmetry of the order book is justified for the average price impact, and Fig. 4.2 shows that the nonlinearity of the virtual price impact is weak.

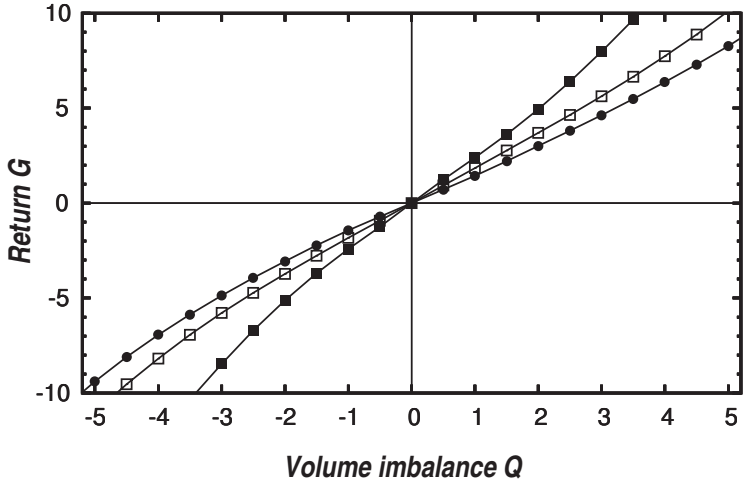

Figure 4.2: The average virtual price impact function $\langle I_{\text{book}}\rangle(Q)$ (full squares) is steeper than the typical virtual price impact $\langle I_{\text{book}}\rangle_{\text{median}}$ (open squares) calculated by taking the median instead of the average. The virtual price impact $I_{\langle\text{book}\rangle}(Q)$ calculated from the average order book (full circles) is weaker than the other two. Adapted from [30].

We find that the virtual price impact $I_{\langle\text{book}\rangle}(Q)$ is four times stronger than the price impact of actual market orders (see Fig. 4.1 and Fig. 4.2, as well as [30]), a volume imbalance of $5\sigma_Q$ causes a virtual price change of $8\sigma_G$ but only an actual price change of $2\sigma_G$. In addition, $I_{\langle\text{book}\rangle}(Q)$ is a convex function that can be fitted by a power law $I_{\langle\text{book}\rangle}(Q) = 1.22\ Q^{1.19}$, and not a concave function as $I_{\text{market}}(Q)$.

When we calculate the average order book in order to get the average virtual price impact by inversion, we do not get the "true" average virtual price impact. Instead, one should calculate the virtual price impact for each time interval and for each stock separately and average over these functions afterwards. To this end, we define a time resolved and per stock depth

$$Q_{\text{book}}(G,t,k) = \sum_{\gamma_i \le G} \rho^k_{\text{book}}(\gamma_i,t)\Delta\gamma \ . \tag{4.4}$$

By inverting this relation at each instant of time and for each stock, we obtain the virtual price impact $I_{\text{book}}(Q,t,k)$. We find that this function fluctuates strongly in time and that its average over time and different stocks $\langle I_{\text{book}}\rangle(Q)$ is dominated by rare events with low liquidity when only few orders are stored in the order book.

These strong fluctuations of $I_{\text{book}}(Q,t,k)$ make the calculation of $\langle I_{\text{book}}\rangle(Q)$ somewhat subtle. In time intervals with very low liquidity, the domain of

$I_{\text{book}}(Q,t,k)$ does not even extend up to $0.5\sigma_Q$ since the amount of limit orders stored in the order book is too small. In this case, the return caused by an order with signed volume $Q > 0.5\sigma_Q$ would be undefined and the average over all time intervals would be undefined as well. In order to expand the domain of $\langle I_{\text{book}}\rangle(Q)$ to at least $3.5\sigma_Q$, we extrapolate the depth linearly by connecting the last defined data point (with largest Q and G) with the origin. Since this procedure is necessary only for few time intervals, our extrapolation method does not disturb the final result. We checked this by using different methods, e.g. by continuing the depth function by a horizontal line instead of a linear extrapolation. The influence of the choice of a specific extrapolation method is clearly visible only for large volumes $Q > 4\sigma_G$. The average of $\langle I_{\text{book}}\rangle(Q)$ is calculated on an equidistant grid on the Q-axis, the values of the individual functions $I_{\text{book}}(Q,t,K)$ at these grid points are calculated by interpolation.

In doing so, one obtains $\langle I_{\text{book}}\rangle(Q)$ as a convex function of signed either buy or sell order volume which is much steeper than the average price impact, see Fig. 4.2. To reduce the influence of low liquidity periods on the virtual price impact, we have calculated a typical price impact $\langle I_{\text{book}}\rangle_{\text{median}}(Q)$ by replacing the average over time and different stocks by the median. For large trading volumes, $\langle I_{\text{book}}\rangle_{\text{median}}(Q)$ is considerably smaller than $\langle I_{\text{book}}\rangle(Q)$, see Fig. 4.2. $\langle I_{\text{book}}\rangle_{\text{median}}(Q)$ is also a convex function of signed volume and quite similar to $I_{\langle\text{book}\rangle}(Q)$.

All three virtual price impact functions studied here show a convex shape, in contrast to the concave shape of $I_{\text{market}}(Q)$. Hence, these virtual price impact functions correspond with reality in the sense that they encourage traders to place many small orders instead of a few large ones to reduce the costs due to the price impact, which is what actually happens. However, these functions describe a price impact that is much larger than the price impact $I_{\text{market}}(Q)$ actually measured on average in five minute intervals. If one wants to find an explanation for the difference between $I_{\text{market}}(Q)$ and the virtual price impact functions, one should use $I_{\langle\text{book}\rangle}(Q)$ as a starting point for the analysis, due to the influence of discretionary trading: with discretionary trading, large orders are placed only when there is enough liquidity present in the order book to match these orders. This behavior is best represented by the flattest curve $I_{\langle\text{book}\rangle}(Q)$, while $\langle I_{\text{book}}\rangle(Q)$ is dominated by few periods of low liquidity, where one expects little trading activity. The typical price impact $\langle I_{\text{book}}\rangle_{\text{median}}(Q)$ is less influenced by discretionary trading than $\langle I_{\text{book}}\rangle(Q)$, but it is still steeper than $I_{\langle\text{book}\rangle}(Q)$. Based on my diploma thesis, the empirical study [30] presented an explanation for the shape of the actual price impact function $I_{\text{market}}(Q)$.

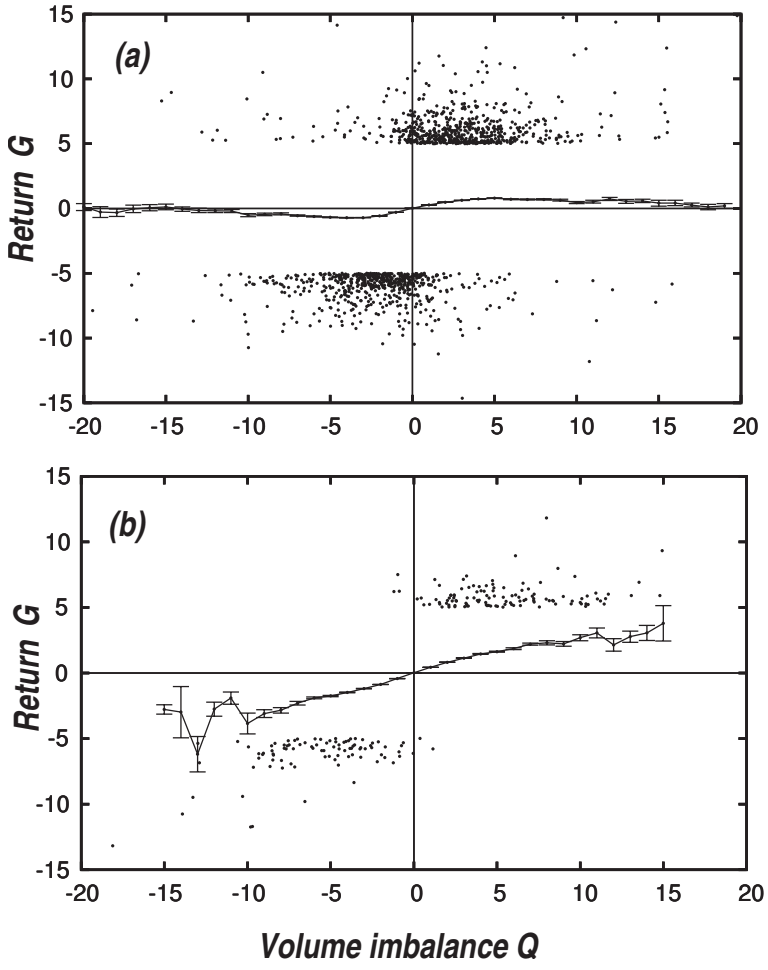

Figure 4.3: (a) Average price impact function for the 44 most frequently traded NASDAQ stocks in the year 1997 with standard deviation of the mean. Price changes larger than five standard deviations cluster in the region of small volume imbalance, all of them are clearly outside the error bars. (b) Same as (a) but for 2002 data from the Island ECN order book for the ten most frequently traded stocks.

Here, we find that one needs to take into account the additional limit order flow arriving in a five minute interval as well as a feedback mechanism: large price changes lead to an increased flow of limit orders that reduce the virtual price impact. These anticorrelations between returns and limit orders are part of the study in the next chapter.

4.3 Price impact and large price changes

After studying the time-independent price impact function in the first part of this chapter, we ask whether extremely large price changes can be described by $I_{\text{market}}(Q)$. To this end, we filter the time series for time intervals with returns $|G| \geq 5\sigma_G$. A detailed description of this procedure including the filtering of data errors is given in appendix A.

The events with price changes larger than five standard deviations are shown in Fig. 4.3, together with the price impact function $I_{\text{market}}(Q)$. We find 1198 such events for the TAQ data base and 210 for the Island ECN data. For some of these events the sign of Q and G do not agree. We believe that this disagreement is (i) caused by the inaccuracy of the Lee and Ready algorithm, as such situations are less frequent for the order book data, and (ii) due to the analysis of intervals with a fixed length rather than the analysis of individual transactions. From the shape of the price impact function, one would expect these events to appear at very large volume imbalance Q. However, Fig. 4.3 shows that these events have a broad distribution centered at intermediate values of Q, well outside the error bars of $I_{\text{market}}(Q)$, which is significantly below $G = 5\sigma_G$. One could argue that this result is not very surprising as the price impact function is an *average* so that large events scatter naturally around this average. Nevertheless, the result shows that (i) large returns occur at quite small volume imbalances and (ii) that the average price impact function cannot be used for a satisfactory prediction of the price change (especially large ones) from the volume imbalance, as it was proposed by Gabaix *et al.* [27].

We conclude that a large volume imbalance alone cannot be responsible for the occurrence of large price changes. An obvious reason for the inaccuracy of the forecast given by the average price impact is that the market does not behave like the average all the time. We already discussed in the first part of the chapter that it is difficult to calculate an average virtual price impact function since the price impact Eq. (4.4) fluctuates strongly over time. This suggests that the price impact might be stronger than average in times when large price changes occur, corresponding to a low liquidity. In the following, we will analyze the

liquidity in these time intervals with large returns and show that this is indeed the correct explanation.

Various transactional properties of markets can be described by the concept of liquidity [111]. Resiliency is the speed at which prices recover from a random uninformative shock. While the market depth that we already mentioned above denotes the order flow innovation needed to change the price a given amount, the tightness is the cost for a round trip, i.e. buying and selling a given amount of shares within a short time period.

Kyle [111] examines the liquidity characteristics of a speculative market within a dynamic model of insider trading and sequential auctions, finding that here both market depth and volatility are constant in time. Glosten [112] derives the equilibrium price schedule in an open limit order book and shows that the limit order book can well compete with other methods of exchanging securities. Madhavan *et al.* [113] study intraday patterns in volatility, bid-ask spreads and transaction costs empirically. Using a linear parametrization for the price impact of individual trades, they find that decreasing reliance on the signal content of order flow results in a sharp drop of price impact after the first half trading hour, while there is a slight increase at the end of the day. Chordia *et al.* [114] discover a weekly seasonality of liquidity and trading activity. These quantities also change depending on market trends or recent market volatility, or prior to major macroeconomic announcements. In an analysis of the limit order book of the Stockholm Stock Exchange [115], Sandas shows that the price impact calculated from the order book is significantly larger than what is expected from a regression model. As a possible explanation he suggests that price impact changes with time-varying market conditions. Similarly, a difference between hypothetical and actual price impact [116] is considered as evidence for discretionary trading, i.e. large trades are more likely to be executed in time periods with sufficient liquidity. Recently, Beltran *et al.* [117] studied the relation between volatility and liquidity for the Euronext trading platform. Using a two-state Markov switching process, they find that the liquidity is significantly higher in the high-volatility state, but their analysis based on a VAR model is not conclusive in whether volatility seriously impacts liquidity.

In order to have a theoretical framework in which we can discuss the mechanism underlying large stock price changes, we set up a price equation

$$S_i = S_{i-1} + c_i + \lambda_i Q_i + u_i \quad . \tag{4.5}$$

in the spirit of [118]. Here, the index i labels successive transactions at times t_i, S_i is the transaction price, c_i the transitory spread component, λ_t the slope of the virtual price impact at time t_i, and u_i is a white noise which describes

the fact that prices change not only due to trading but also due to the arrival of new public information. As we will mostly be concerned with the analysis of midquote price changes, we let $c_i \equiv 0$ in the following. For the price change in an interval with a fixed length Δt one finds [119]

$$S(t+\Delta t) - S(t) = \sum_{t_i \in [t,t+\Delta t]} \lambda_{t_i} Q_{t_i} + \sum_{t_i \in [t,t+\Delta t]} u_{t_i} \quad . \tag{4.6}$$

In this framework, the order book density is approximated as constant so that the depth defined in Eq. (4.4) would be $Q_{\text{book}}(G,t,k) = G/\lambda_t$. Here, the price impact of an order volume Q_{t_i} according to Eqs. (4.5), (4.6) is permanent but variable due to the t-dependence of λ_t. Hence, this framework agrees with the view of [73, 102] when one tries to reconcile uncorrelated returns and correlated order flow.

In the light of Eqs. (4.5), (4.6) there are three possible causes for large price changes: (i) large order flows Q_{t_i}, (ii) large price impacts (small liquidities) λ_{t_i}, and (iii) public information u_{t_i}. In this context, it can be confusing to precisely distinguish between volume imbalance and order flow, since Q_{t_i} can represent single orders, only buy or sell volume or the volume imbalance. For the sake of simplicity and readability, we use the expression "order flow" where its sometimes multiple meaning is clear from the context.

We saw that large order flow alone cannot explain large price changes, so that in the following we want to analyze the influence of a time-varying liquidity.

4.4 Time varying price impact

The time varying liquidity becomes manifest in the strong fluctuations of $\rho_{\text{book}}(Q)$. When we want to use these fluctuations to explain large price changes, we have to take into account that here the assumptions of (i) a symmetric order book and (ii) negligible nonlinearities are generally not satisfied, in contrast to the analysis of the average price impact function. For this reason, we will consider either the buy or the sell volume $\tilde{Q}$ in a given five minute interval, depending on the direction of the return in the respective interval. In this way, $\tilde{Q}$ is equal to the volume of buy market orders if $G_{\Delta t} > 0$ in that five minute interval. For $G_{\Delta t} < 0$ on the other hand, $\tilde{Q}$ is equal to the volume of sell market orders and has a negative sign. We have recalculated I_{market} as a function of $\tilde{Q}$ by averaging with respect to either the sell or the buy volume. This new $\tilde{I}_{\text{market}}$ is quite similar to the original one.

We try to find a quantitative explanation of extreme price changes by taking into account not only order flow but also market liquidity as described by market

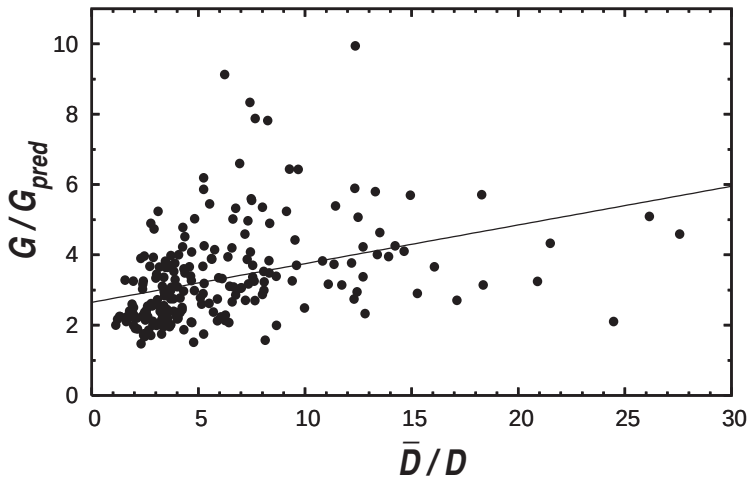

Figure 4.4: Ratio of actual price change to predicted price change plotted against the inverse market depth for large five minute returns contained in the 2002 Island data. A linear regression (line) yields a correlation coefficient $R^2 = 0.14$.

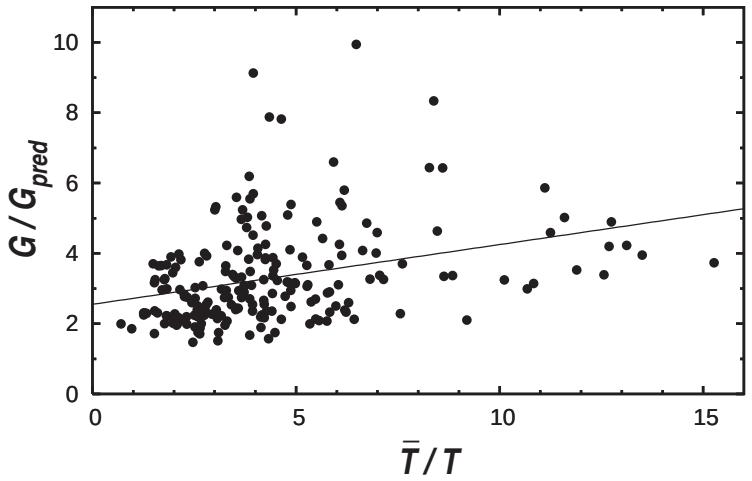

Figure 4.5: Ratio of actual price change to predicted price change plotted against the inverse market tightness for large five minute returns contained in the 2002 Island data. A linear regression (line) yields a correlation coefficient $R^2 = 0.11$.

depth and market tightness in the beginning of a given time interval. The depth D is the size of a market order required to change the price by a given amount $5\sigma_G$ and is obtained from Eq. (4.4). The tightness T is the cost of a round trip (buying and selling a volume of $2\sigma_Q$ within a short period of time). To determine the tightness for a given time interval, we calculated the virtual price impact $I_{\text{book}}(\tilde{Q})$ by inverting the relation Eq. 4.4 and define the tightness as

$$T = \frac{1}{|I_{\text{book}}(2\sigma_Q)| + |I_{\text{book}}(-2\sigma_Q)|} \quad . \tag{4.7}$$

In the framework of the model Eq. (4.5), the order book density is approximated as constant and the tightness would be just $T = 1/(4\sigma_Q\lambda_t)$.

We compare the ratio of the actual price change $G_{\Delta t}(t)$ and the predicted price change

$$G_{\text{pred}}(t) = \tilde{I}_{\text{market}}(\tilde{Q}(t)) \tag{4.8}$$

to the inverse liquidity as described by the inverse depth and the inverse tightness. Using the average depth $\overline{D}$ and the average tightness $\overline{T}$ calculated from the average order book, we normalize both liquidity measures depth and tightness. Since the statistics is insufficient for $|\tilde{Q}| > 18\sigma_Q$, we computed $\tilde{I}_{\text{market}}(\tilde{Q}(t))$ up to $|\tilde{Q}| = 18\sigma_Q$ to calculate G_{pred}. For this reason, we had to discard eleven events with $|\tilde{Q}| > 18\sigma_Q$ from this analysis. In addition, for eight events the tightness T could not be computed because the order book did not contain enough limit orders to trade a volume of $2\sigma_Q$. These events are excluded in the analysis of the inverse tightness as liquidity measure. For reasons of consistency, we also removed two events with $\overline{D}/D > 30$ in Figure 4.4. A scatter plot for events with $|G_{\Delta t}| > 5\sigma_G$ is shown in Figs. 4.4 and 4.5. In contrast to the expectation that small liquidity can explain the ratio of actual and predicted price change, there is only a moderate correlation between returns and liquidity for both depth and tightness. This visual impression is confirmed by correlation coefficients $R^2 = 0.14$ and $R^2 = 0.11$ for depth and tightness, respectively.

In the light of these results, the explanatory power of liquidity for large aggregate returns seems to be weak. However, the problem is that it is not sufficient to take into account only the order book density $\rho_{\text{book}}(\gamma_i, t)$ at one instant of time when one studies the price impact of order flow in a whole time interval. Earlier in this chapter we showed that the virtual price impact of a given order volume is roughly four times stronger than the actual one and pointed out that this difference is due to the additional orders placed in reaction to a price change [30]. Hence, one has to include dynamical effects, i.e. changes in the order book within the time interval, to correctly calculate the price impact and understand the occurrence of large price changes.

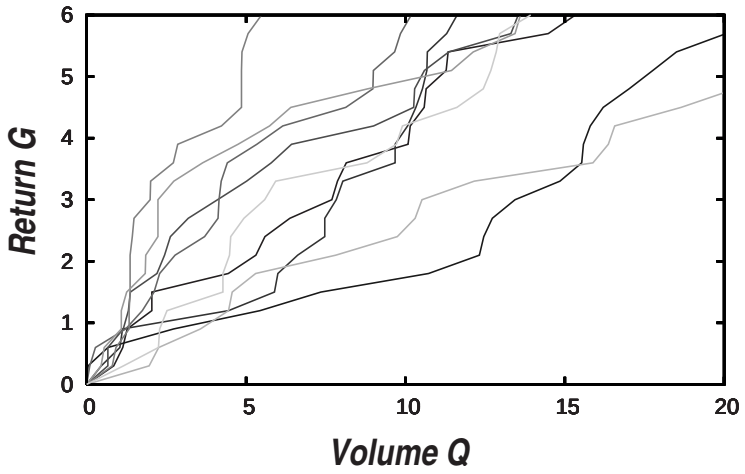

Figure 4.6: Price change as a function of buy or sell volume for ten of the largest price changes in the Island ECN data.

When we calculate the density of limit orders arriving in a given time interval, we want to do it in such a way that it is compatible with the density $\rho_{\text{book}}(\gamma_i, t)$ recorded in the beginning of the time interval. However, during five minutes bid and ask price can change significantly, so that an order placed close to the bid price at the end of the time interval can be far away from the bid price recorded in the beginning. In order to deal with this problem, we fix a reference frame by the bid and ask price in the beginning of the interval, so that arriving limit orders are not counted according to the current bid and ask price, but to this reference frame. Sell limit orders arriving at a price lower than the reference ask price are counted as if they were arriving at this ask price, vice versa for buy limit orders. Similar to the density $\rho_{\text{book}}(\gamma_i, t)$ of limit order volume at a depth γ_i recorded in the beginning of the time interval $[t, t + \Delta t]$, we define another density function $\rho_{\text{flow}}(\gamma_i, t, \Delta t)$ describing the density of limit order volume placed at a depth γ_i minus the limit order volume removed during this time interval with

$$\rho_{\text{flow}}(\gamma_i) = \langle Q^{\text{add}}_{\Delta t}(\gamma_i) - Q^{\text{canc}}_{\Delta t}(\gamma_i) \rangle \quad . \tag{4.9}$$

In Eq. (4.9), $Q^{\text{add}}_{\Delta t}(\gamma_i)$ is the volume of limit orders added to the book at a depth γ_i, and $Q^{\text{canc}}_{\Delta t}(\gamma_i)$ is the volume of orders canceled from the book. Thus, $\rho_{\text{flow}}(\gamma_i, t)\Delta\gamma$ is the net limit order volume arriving in the time interval $[t, t+\Delta t]$ and in the price interval $[(i-1)\Delta\gamma, i\Delta\gamma]$. The total density of limit orders available for transactions is then given by

$$\rho(\gamma_i, t) = \rho_{\text{book}}(\gamma_i, t) + \rho_{\text{flow}}(\gamma_i, t, \Delta t) \quad . \tag{4.10}$$

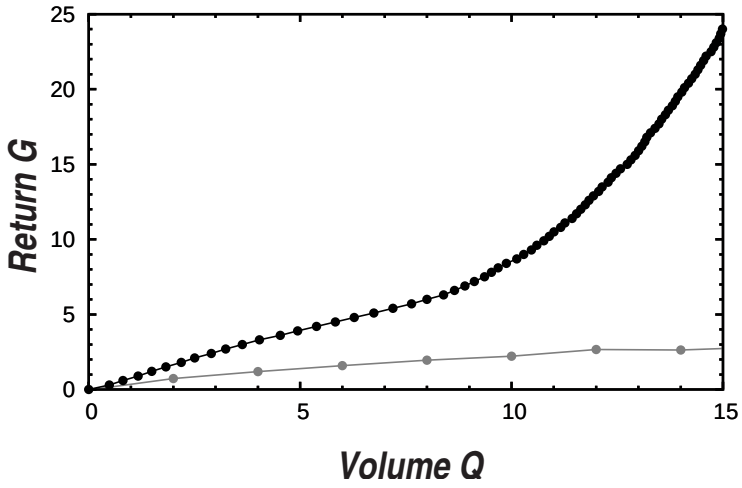

Figure 4.7: Price change as a function of buy or sell volume averaged over all time intervals with returns larger than $5\sigma_G$ (connected black circles). The price change averaged over all transactions (connected gray circles) is much smaller than that for the extreme events.

The density $\rho(\gamma_i, t)$ is related to the order flow $\tilde{Q}$ as

$$\tilde{Q}(G) = \sum_{\gamma_i \leq G} \rho(\gamma_i, t) \Delta\gamma \quad . \tag{4.11}$$

By inverting this relation we calculate a price impact function $I_{\text{actual}}(\tilde{Q})$. The sell order side of this function for ten events with price changes larger than $5\sigma_G$ is shown in Figure 4.6. In Figure 4.7, the average over all such events is compared to the average price impact function $\tilde{I}_{\text{market}}(\tilde{Q})$. Figure 4.7 shows that the slope of $I_{\text{actual}}(\tilde{Q})$ is much larger than the slope of $\tilde{I}_{\text{market}}(\tilde{Q})$. As a consequence, in these time intervals with large price changes there are less limit orders available than on average. Hence, we suggest to use the slope of the actual price impact function as a measure of market liquidity.

The price impact functions displayed in Figure 4.6 look quite linear, and the average of the I_{actual} for all large events (see Figure 4.7) is approximately linear as well [2]. Accordingly, we expect that the actual strength of the price impact can be well described by a linear fit to the actual price impact functions. Hence, for each time interval with $|G_{\Delta t}| > 5\sigma_G$, we define a susceptibility $\chi(t)$ by a

[2] We tested $I_{\text{actual}}(\tilde{Q})$ for each time interval with price change larger than $5\sigma_G$ for nonlinearities. As a simple descriptive method we fitted these curves with power laws. The exponents we found vary between 0.15 and 2.35 with a mean of 1.32 and they scatter with a standard deviation of 0.41. On the other hand, a power law fit to the average of I_{actual} for all such events yields an exponent of 1.03, which is approximately linear.

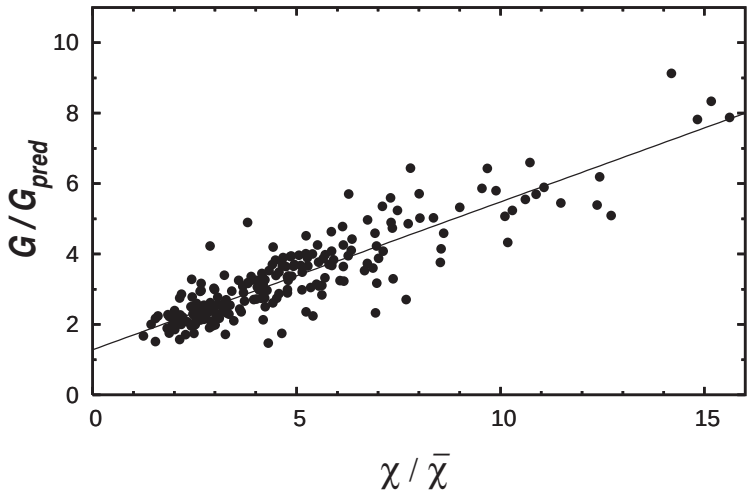

Figure 4.8: Ratio of actual price change to predicted price change plotted against the slope of the actual price impact function normalized by the slope of the average price impact function. The data points cluster in the vicinity of a linear fit with an $R^2 = 0.79$.

linear fit through the origin to the actual price impact function $I_{\text{actual}}(\tilde{Q})$ up to a return $G = 5\sigma_G$ or $G = -5\sigma_G$, depending on the sign of $G_{\Delta t}$. The susceptibility $\chi(t)$ can formally be identified with λ_t, though in the simple model Eq. (4.5) the order book density is approximated as constant and dynamical effects are not included. Liquidity is measured by the inverse $\frac{1}{\chi(t)}$. In this way, a large slope of the price impact function corresponds to a low liquidity.

In Figure 4.8 the ratio of G_{pred} and $G_{\Delta t}$ is plotted against the susceptibility $\chi/\overline{\chi}$ for all events with $|G_{\Delta t}| > 5\sigma_G$. The susceptibility χ is normalized by $\overline{\chi}$, the slope of a linear fit to the average price impact function $\tilde{I}_{\text{market}}$ up to $|G_{\Delta t}| = 5\sigma_G$. We removed one event with extremely small liquidity ($\chi/\overline{\chi} > 60$) to make this analysis consistent with the analysis of tightness and depth. The data points in Figure 4.8 cluster in the vicinity of a linear fit with an $R^2 = 0.79$. In comparison with the two liquidity measures studied above, this result is a considerable improvement. We believe that this improvement is due to the fact that the susceptibility χ takes into account the order book dynamics, which is important for describing liquidity. From this analysis, we conclude that liquidity defined as the time dependent slope of the price impact function has a large explanatory power for the occurrence of extreme price changes.

As additional evidence for the idea that the return in a given time interval is caused by a combination of the order flow and the time varying liquidity, we discuss returns as a function of both order flow $\tilde{Q}$ in the direction of the price

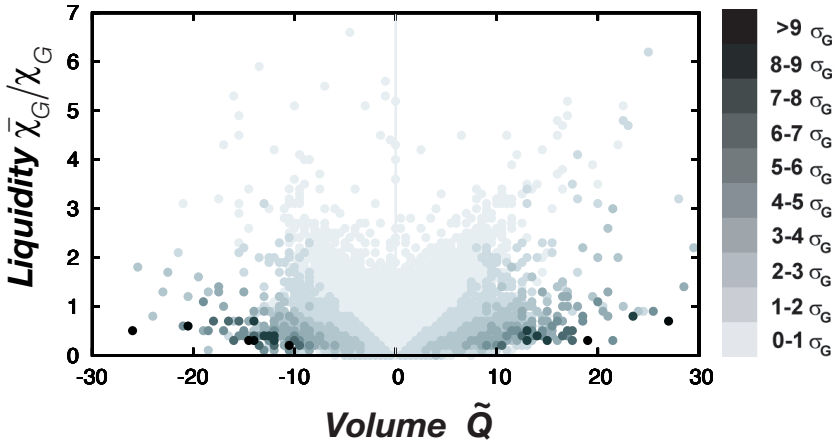

Figure 4.9: Expected return $\langle G_{\Delta t}(t)\rangle_{\tilde{Q},\overline{\chi}_G/\chi_G}$ as a function of order flow $\tilde{Q}$ and liquidity $\overline{\chi}_G/\chi_G$. For every combination of $\tilde{Q}$ and $\overline{\chi}_G/\chi_G$ we plotted (i) the average return if there is more than one matching time interval, (ii) the return if there is only one event or (iii) nothing if the combination never occurred. The magnitude of the return is coded from bright gray for small returns to black for the largest ones.

change and the susceptibility $\chi/\overline{\chi}$. Above, the susceptibility χ was defined by a linear fit to the actual price impact function $I_{\text{actual}}(\tilde{Q})$ up to a return $|G| = 5\sigma_G$. Now, we want to study also returns smaller than $G_{\Delta t} < 5\sigma_G$ where the order book density at a depth $\gamma_i > G_{\Delta t}$ does not affect the price dynamics. This effect would weaken the explanatory power of χ for these time intervals thus we define a new susceptibility χ_G by a linear fit through the origin to the actual price impact function $I_{actual}(\tilde{Q})$ up to the actual return $G_{\Delta t}$. In time intervals with $|G_{\Delta t}| < 1\sigma_G$, the linear fit extends up to $\sigma_G \,\text{sgn}(G_{\Delta t})$ in order to include enough data points for a reliable fit.

Figure 4.9 displays the average return plotted as a function of both market order flow and liquidity measured by $\overline{\chi}_G/\chi_G$. The magnitude of returns is coded in a gray scale from bright gray for small returns to black for the largest ones. One observes quite sharp borders between regimes of different expected returns demonstrating again that for a given order flow the magnitude of the return depends on liquidity. In addition, one sees that even very large volumes can lead to small returns, while large returns occur only if the liquidity is below average.

4.5 Discussion

We showed that fluctuations of the liquidity have a large influence on stock price changes, as large returns occur mostly in time periods with low liquidity. Together with the order flow, liquidity provides for a quantitative explanation of large price changes. However, so far we discussed only part of the terms in the standard pricing model Eqs. (4.5),(4.6). The white noise term u_t describes the influence of new public information, e.g. earnings announcements or monetary policy announcements, accounting for 35% to 46% of the volatility of transaction price movements, according to [113]. The experience is that such news can lead to price "jumps", and thus can influence also large stock price changes. However, although we did not include public information in our analysis explicitly, one can argue that the order book description contains this information. News announcements lead to reactions of the market participants, possibly prompting them to place or cancel orders. If there is good news, for instance, people would cancel their sell limit orders and place additional buy orders. In our liquidity measure $1/\chi(t)$, this would lead to a reduction of the liquidity, so that also in this case a large price change would correspond to a low liquidity. In this sense, our liquidity measure describes the combined influence of order book depth, resiliency, and public information.

In the previous chapter, we showed that in intervals with large aggregate returns the average tick return size is significantly larger than average. This can be seen as another manifestation of low liquidity: though liquidity measures like the market depth or the inverse slope of the actual price impact function involve the number of shares, it has been shown [52] that the traded volume usually matches the volume available at the bid or ask price. Thus, the return due to a single trade corresponds to the gap between the best price and the second best price [52], which in turn corresponds to the slope of the price impact function. In this sense, the average tick return size can be seen as an (inverse) liquidity measure, so that our results suggest that the diffusion process of stock returns depends largely on fluctuations in the liquidity.

We have argued that large stock price changes can be explained by periods of low liquidity. However, so far we did not discuss the possibility of an inverse causality, leading from large return to low liquidity. A recent study [120] of the market crash in October 1997 showed that the spread increased significantly on October 28th, the day after the market drop on October 27th. Hence, this study suggests that large returns can cause a low liquidity. We cannot exclude the possibility that the periods of low liquidity detected in our analysis are caused by large returns in previous time intervals, but our analysis shows that *in* the

intervals with the large price change the liquidity is low, causing an intermediate volume to create a large return.

In summary, we studied the mechanisms leading to large price fluctuations and showed that they cannot be related to only one single effect. A large trading volume changes the price, which can be quantified in the average price impact function. This effect might be analogous to the number imbalance studied in the previous chapter, which there accounted for the basic price movements. We find little evidence that this effect alone can be responsible for extreme price changes, which can only be explained by a combination of relatively large trading volume and low liquidity. The liquidity can be measured as the dynamically changing slope of the actual price impact function, but can be also related to the average tick return size from the previous chapter.

5 Trading strategies and uncorrelated stock returns

So far, we have analyzed extreme price changes in order to examine the underlying mechanisms. In the previous chapter, we saw that dynamic properties of the order book can be used to find an appropriate liquidity measure as well as to understand the difference between the actual price impact of market orders and the virtual price impact calculated from the limit order book [30]. While the average order book and thus the virtual price impact discussed in the previous chapter can be described by "zero intelligence models" [110, 121], in which orders are placed randomly, the order book dynamics might in fact be related to "intelligent" behavior of market participants. In the present chapter, we study this intelligent behavior in terms of trading strategies to explain further empirical findings about stock returns.

The behavior of traders directly determines the movement of stock prices via the trading process, where the buy and sell orders of many traders are matched against each other. Recent studies [31, 73] show that order signs, indicating buy or sell orders, are long-range correlated, so that for example a buy order leads to a prediction of many subsequent buy orders. From these results one would expect that through trading the correlations in the order signs would lead to similar long-range correlations in the returns, but surprisingly the returns are uncorrelated and thus not predictable. From a theoretical point of view, one can argue that this indicates market efficiency [122], stating that the market adjusts automatically so that easy opportunities for making profit ("arbitrage") are absent. However, this does not explain how the correlations disappear during the trading process.

In the previous chapter, we pointed out that a fixed price impact as it would be given by the average price impact function could not explain uncorrelated returns if the price impact would be permanent: if each market order changes the price permanently, the market order correlations lead to returns that are long-term correlated as well. Our finding of a strongly fluctuating price impact function presented above is compatible with the idea of a price impact that might be permanent, but not fixed.

Similarly, Lillo and Farmer [73] argue that the market compensates the correlations in the order flow by adjusting its properties such as liquidity. The authors show empirically that the probability of a market order to change the price decreases for larger predictability of the orders. In times of large predictability, market orders are smaller than average or the volume at bid or ask price is higher, lowering the probability of a price change.

Bouchaud *et al.* [31] present a different approach that explains uncorrelated returns with a fixed price impact, in contrast to the variable price impact used in the work of Lillo and Farmer. However, the price impact of Bouchaud *et al.* is only temporary, meaning that the price change caused by a market order vanishes after some time due to mechanisms of the market. The model is related to a trading strategy: liquidity providers give trading possibilities to both buyers and sellers. If the stock price stays constant, they can make profit by buying at the best bid and selling at the lowest ask price, profiting from the spread (i.e. the difference between these two). In order to keep prices constant, these liquidity providers try to mean revert the price. This mean reverting is supported by the empirically found anticorrelations between market orders and limit orders [30], so that limit orders are placed in response to market orders and thus compensate their price impact.

Based on statistical properties found empirically, Mike and Farmer [101] propose a model whereby different order types are represented by long-range correlated processes, which are then combined with a model for order cancelation due to asymmetries in supply and demand. In this model, returns exhibit no long-range correlations, though they do persist longer than in reality, as the autocorrelation function exhibits values of the order of 1% for about 50 time steps.

In this chapter, we model two different trading strategies and analyze their profit in the light of correlated orders. While in the model of Bouchaud *et al.* liquidity traders are afraid of price changes and try to act against the correlated order signs, we study a mechanism where traders use these correlations to increase their profit. Under the simplification of exponentially decaying correlations in the order signs, we show that correlations between returns vanish due to the studied trading strategy, and in addition, we qualitatively reproduce the cross-correlations between returns and both market orders and limit orders, which were presented in a previous work [30] and accounted for the connection between the actual price impact function of market orders and the virtual price impact calculated from the order book.

This chapter is organized as follows: in section 5.1, we describe the model we study, section 5.2 analyzes the trading strategy of a liquidity provider while in

section 5.3 we study a "front runner" strategy. Section 5.4 gives a summary and discussion of the results.

5.1 Description of the model

In order to keep our model as simple as possible, we simulate only one trader. This trader is acting in an environment with two basic properties summarizing the actions of all other traders: first, there is a flow of market orders $m(t)$ which is due to a stochastic process and independent of the price movement or the behavior of the trader. Second, there is a 'background liquidity', meaning that independent of the simulated trader there is always a certain amount of limit orders in the market, so that any market order m can be executed with a price impact λm proportional to the volume $|m|$ of the market order. The simulated trader can place market orders as well as limit orders.

If the trader in this model would not do anything, each market order would directly influence the price, so that the price $S(t)$ at time step t would change according to

$$S(t+1) = S(t) + \lambda m(t) \quad . \tag{5.1}$$

In this way, the time series of returns would be basically the same as the time series of market orders, so that it would also inherit the market order correlations, in striking contrast to reality. Hence, we suggest that the loss of correlations in stock price changes is due to the behavior of traders following certain strategies in order to optimize their profit.

The autocorrelation function of empirical market orders $c(m_{\mathrm{emp}}(t), m_{\mathrm{emp}}(t), \tau)$ (compare Eq. (2.8)) follows a power-law with an exponent smaller than one [31, 73]. This leads to the problem that the correlation function is not integrable, so that the number of predicted subsequent market orders is infinite. However, we can estimate the integral over the autocorrelation function of market orders for certain time intervals. Figure 5.1 shows this integral for data of Cisco (CSCO), a typical stock from the Island ECN in the year 2002. Here, a power law fit with $\frac{1}{1-\theta}\beta\tau^{1-\theta}$ yields $\beta = 0.26$ and $\theta = 0.73$. Measuring the time in such a way that each time step corresponds to a new market order, we find that (excluding the value 1 for time lag 0) this integral is around 4 after 250 time steps (corresponding to about 20 minutes), around 5.9 after 750 steps ($\approx$ one hour), and around 9.4 after 4875 steps ($\approx$ one trading day). Though this integral keeps growing infinitely for larger time windows, in this study we focus on a short time horizon and approximate the empirical market order time series by an $AR(1)$ process with exponentially decaying correlations. In this model, a

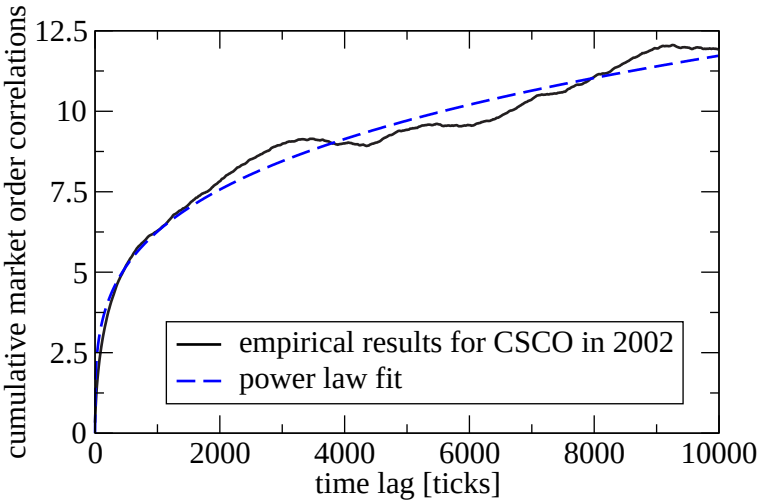

Figure 5.1: Cumulative correlations of market orders at a time lag τ (solid line). A power law fit with $\frac{1}{1-\theta}\beta\tau^{1-\theta}$ yields $\beta = 0.26$ and $\theta = 0.73$.

market order m at time t is given by

$$m(t) = \varrho m(t-1) + \varepsilon(t) = \sum_{j=0}^{\infty} \varrho^j \varepsilon(t-j) \quad . \tag{5.2}$$

Thus, a market order consists of an unpredictable part with the normally distributed random number $\varepsilon(t)$ with zero mean and variance one, and a predictable part $\varrho m(t-1)$ depending on the last order $m(t-1)$ and the parameter ϱ, which determines the strength of the correlation. For $\varrho = 0.8$, the overall prediction of market orders in a time interval of 20 minutes (corresponding to about 250 market orders) is the same as in the empirical data, whereas $\varrho = 0.9$ would correspond to the correlations of one trading day. This simplification gives us the opportunity to study the market in a very simple model in order to understand the influence of order strategies on the market.

The relevant parameters for this model are: (i) the spread s and the coefficient λ for the price impact, which together determine the scale on which the price and the trader's capital change. (ii) The inventory limit $I_{\max}$ defines the maximum number of shares hold by the trader, which can be positive as well as negative (meaning that the trader can sell borrowed shares hoping to buy them back later at a lower price). (iii) The correlation coefficient ϱ determines the correlations in the market orders.

5.2 Liquidity provider strategy

The first strategy we want to analyze is mentioned by Bouchaud *et al.* [31] supporting their explanation for the absence of correlations in stock returns. In this strategy, a trader places both buy and sell limit orders. Due to the spread, the trader makes profit at each round-trip when she buys shares at the bid price and sells them at the higher ask price. However, this strategy guarantees save profit only if the price stays constant. For instance, if the price rises after the trader sold some shares at the ask price, she cannot buy them back at the old bid price, but has to buy them at the new raised bid price. This price may be higher than the price the trader got for selling the stocks, so changing prices can cause losses for the trader who follows this strategy. Bouchaud *et al.* argue that for this reason such a trader would try to prevent the price from moving far from an estimated 'fair' price by placing limit orders that compensate the impact of market orders. In the following, we want to analyze the profit of this strategy under the assumption of uncorrelated as well as correlated market orders.

In the framework of our model, this strategy can be implemented as follows: in each time step t, there is a new market order $m(t)$. Instead of letting the trader place or cancel limit orders, we only give her the two choices either to match the market order (meaning that she placed some limit orders before) or to do nothing (meaning that she canceled all her limit orders before). If she matches the order, the price stays constant but her inventory changes. Otherwise, the market order changes the price according to Eq. (5.1).

If the trader would have an unlimited inventory, she could match all incoming market orders. When buying shares at the bid price and selling them in the next step at the higher ask price, she gains half the spread s for each traded share. Hence, if she has enough time to wait for her inventory to neutralize, after N time steps she gains

$$\frac{N}{2} \langle |m| \rangle \, s \tag{5.3}$$

with this strategy.

On the other hand, if we want to calculate the profit after a given number of time steps N, we assume that the trader has to close her positions at step N by placing market orders, leading to costs depending on the number of shares she is holding in her inventory. This is reasonable because many traders have a certain time horizon, so that e.g. day-traders close their positions at the end of a day avoiding bad surprises happening overnight. Assuming uncorrelated market orders, the number of shares in her inventory would grow with the number of time steps N like $\langle |m| \rangle \sqrt{N}$.

If we assume that the trader's market order m_{tr} (e.g. a sell order) executes several limit orders at different price levels until the entire order is fulfilled, the total cost C for the traded shares is given by

$$C = m_{\mathrm{tr}} S_{\mathrm{sell}} = m_{\mathrm{tr}}(S + \frac{1}{2}(sgn(m_{\mathrm{tr}})s + \lambda m_{\mathrm{tr}})) \tag{5.4}$$

The sign of the total cost depends on the sign of the order m_{tr} of the trader, so that negative costs indicate that the trader is selling shares and getting money, while she is buying with positive costs. Considering a round trip where the trader first buys a certain amount of shares at the bid price and then sells them again using market orders, we can calculate the loss resulting from this action. Before, she bought the shares using limit orders at the price $S_{\mathrm{buy}} = S + \frac{1}{2}\mathrm{sgn}(m_{\mathrm{tr}})s$, so now she realizes a loss L of

$$L = m_{\mathrm{tr}}(S_{\mathrm{sell}} - S_{\mathrm{buy}}) = \frac{\lambda}{2} m_{\mathrm{tr}}^2 \quad . \tag{5.5}$$

The Loss L is always positive, irrespective of the sign of m_{tr}.

Given the win the liquidity provider makes when buying and selling stocks at the bid and ask price and the loss when closing her positions, her expected profit after N time steps is given by

$$\langle pr_{\mathrm{LP}} \rangle = \frac{N}{2}\langle |m| \rangle s - \frac{\lambda}{2}\langle I^2 \rangle \quad . \tag{5.6}$$

Uncorrelated market orders

We can estimate for which parameters the strategy can be expected to be profitable, assuming uncorrelated Gaussian distributed market orders $m(t)$ with $\langle m \rangle = 0$, $\langle m^2 \rangle$, and thus $\langle |m| \rangle = \sqrt{\frac{2}{\pi}}$. After N time steps the square of the inventory $I^2(t)$ has the expectation value $\langle I^2(N) \rangle = N$, so that selling or buying these shares at time step N reduces the profit by $\lambda N/2$ according to Eq. (5.5). Thus, the total profit of the liquidity provider after N time steps is

$$\langle pr_{\mathrm{LP}}^{\mathrm{uncorr}} \rangle = \frac{N}{2} \langle |m| \rangle s - \frac{\lambda}{2}\langle I^2 \rangle = \frac{N}{2}\left(\sqrt{\frac{2}{\pi}} s - \lambda\right) \quad . \tag{5.7}$$

Hence, this strategy leads to profit if $0.8s \gtrsim \lambda$ so that the spread is of the order of the price impact. A similar argument was given by Wyart *et al.* [123] who in addition showed empirically that spread and price impact are usually of the same order. The difference to their model is that here we describe a model for "all" liquidity providers, represented by the one trader we simulate, while Wyart *et al.* model only a small fraction of the liquidity providers in the market. They

take into account the price impact due to the action of other traders and its temporal structure, whereas our model includes the price impact when a large position is closed.

Short-term correlated market orders

The liquidity provider's profit changes drastically if there are correlations present in $m(t)$. For instance, if $m(t)$ follows an $AR(1)$ process Eq. (5.2) with $\langle m^2 \rangle = \frac{1}{1-\varrho^2}$ and $\langle |m| \rangle = \sqrt{\frac{2}{\pi(1-\varrho^2)}}$, the profit Eq. (5.6) changes due to the change in $\langle |m| \rangle$ and $\langle I^2 \rangle$. Here, $\langle I^2 \rangle$ is given by

$$\begin{aligned} \langle I^2 \rangle &= \left(\sum_{i=1}^{N} m(i) \right)^2 = \sum_{i,j=1}^{N} \langle m(i)m(j) \rangle \\ &= N\langle m^2 \rangle + 2\sum_{i=1}^{N} (N-i)\langle m(t)m(t-i) \rangle \\ &= N\langle m^2 \rangle + 2\langle m^2 \rangle \sum_{i=1}^{N} (N-i)c(m(t), m(t), i) \end{aligned} \tag{5.8}$$

where $c(m(t), m(t), i) = \varrho^i$ is the correlation function for the $AR(1)$ process. Thus, if $N \gg 1$ we have for the inventory

$$\begin{aligned} \langle I^2 \rangle &= N\langle m^2 \rangle \left(1 + 2\frac{\varrho}{1-\varrho} \right) \\ &= N\frac{1}{1-\varrho^2} \left(\frac{1+\varrho}{1-\varrho} \right) = N\frac{1}{(1-\varrho)^2} \quad . \end{aligned} \tag{5.9}$$

Hence, the profit for the liquidity provider strategy with the exponentially decaying correlations of the $AR(1)$ process is given by

$$\begin{aligned} \langle pr_{\mathrm{LP}}^{\mathrm{exp}} \rangle &= \frac{N}{2}\langle |m| \rangle s - \frac{\lambda}{2}\langle I^2 \rangle \\ &= \frac{N}{2} s \sqrt{\frac{2}{\pi(1-\varrho^2)}} - \frac{\lambda}{2} N \frac{1}{(1-\varrho)^2} \\ &= \frac{N}{2} \left(s\sqrt{\frac{2}{\pi(1-\varrho^2)}} - \lambda \frac{1}{(1-\varrho)^2} \right) \quad . \end{aligned} \tag{5.10}$$

Figure 5.2 illustrates the dependence of the profit on the parameters of the model. The profit is displayed as the profit per normalized market order, calculated by dividing the profit by $N\langle |m| \rangle$. Each line shows this profit as a function of the market order correlation coefficient ϱ for different choices of the spread s, while $\lambda = 0.1$ for all curves. Figure 5.2(a) displays the average profit after

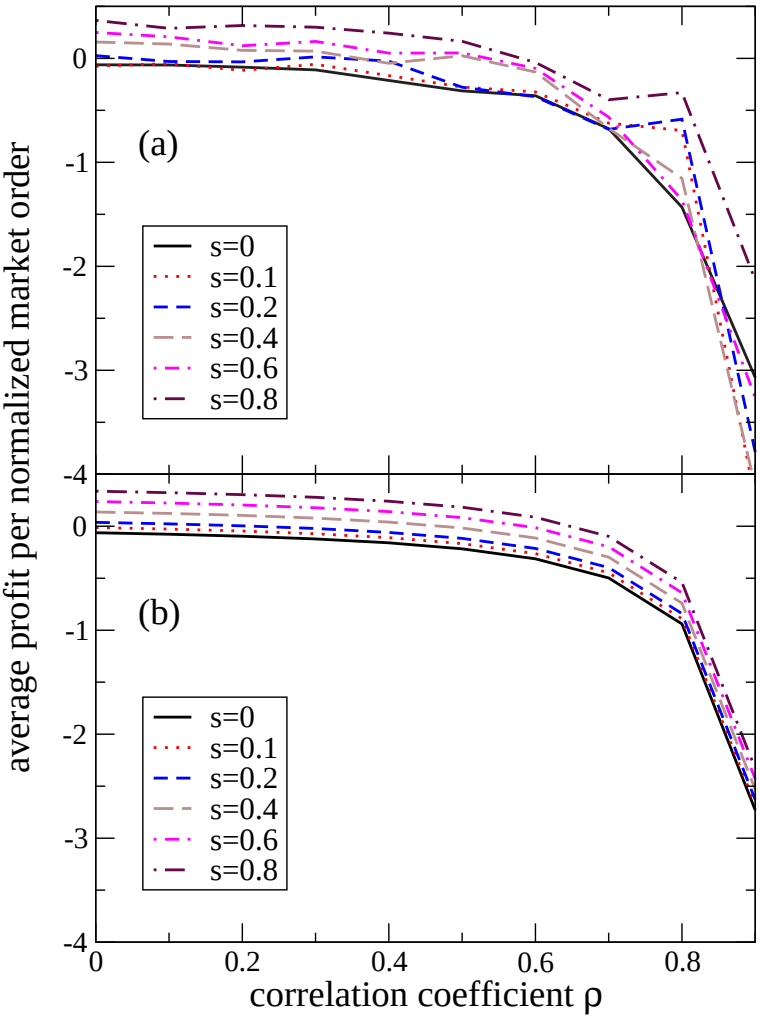

Figure 5.2: Profit per normalized market order in the liquidity provider strategy for different spreads. The price impact coefficient is hold constant at $\lambda = 0.1$. (a) From 100 simulations with $N = 100,000$ time steps each, the average profit at the end is divided by $N \langle |m| \rangle$ in order to obtain the profit per normalized market order. Figure (b) shows the normalized profit per market order obtained from the analytical result Eq. (5.10). The deviations of the simulations from the analytical result are due to strong fluctuations of the inventory after N time steps.

100 simulations with $N = 100,000$, whereas in Fig. 5.2(b) the analytical result Eq. (5.10) divided by $N\langle|m|\rangle$ is shown. The displayed deviations of the simulations from the analytical result are due to strong fluctuations of the inventory after N time steps.

The figure shows that the liquidity provider strategy works well for weak correlated market orders, but with increasing correlations the increasing inventory leads to losses, even for large spreads when the correlations are strong. From the analytical result Eq. (5.10), one can expect a positive profit if

$$\frac{s}{\lambda} > \frac{1}{1-\varrho}\sqrt{\frac{\pi(1+\varrho)}{2(1-\varrho)}} \quad . \tag{5.11}$$

For correlations with $\varrho = 0.8$ (corresponding to the empirical correlations within 20 minutes), the strategy is profitable if the spread s is about 19 times the price impact λ. With these strong correlations, this strategy can only be profitable for extremely large spreads. Though it is reasonable that with correlated market orders the spread is increased in order to compensate the risk of large price movements, this value is very high so that the strategy as it is presented here would usually not lead to profit without adaptions of the model.

Long-term correlated market orders

If we consider power law correlations instead of the exponentially decaying correlations of an $AR(1)$ process, the liquidity provider's profit for large N becomes even smaller. For power law correlated $m(t)$ with $\langle m(t)^2\rangle = 1$, $\langle m(t)\rangle = 0$, and correlation function $c(m(t), m(t), \tau) = \frac{1}{(1+\tau)^\theta}$, the expected inventory $I = \sum_{i=1}^{N} m(i)$ after N time steps is

$$\begin{aligned}
\langle I^2\rangle &= N + 2\sum_{i=1}^{N}(N-i)c(m(t), m(t), i) \\
&= N + 2\sum_{i=1}^{N}(N-i)\frac{1}{(1+i)^\theta} \\
&\approx N + 2\int_1^N dt(N-t)\frac{1}{(1+t)^\theta} \\
&= N + \frac{2}{(\theta-2)(\theta-1)}\Big((N+1)^{-\theta} + 2N(N+1)^{-\theta} + \\
&\quad + N^2(N+1)^{-\theta} - 2^{1-\theta}\theta + 2^{1-\theta}N\theta - 2^{2-\theta}N\Big) \quad .
\end{aligned} \tag{5.12}$$

For $N \gg 1$ and $\theta < 1$, we collect all terms with N^ζ for $\zeta > 0$ and replace $N+1$ by N, so that

$$\begin{aligned} \langle I^2 \rangle &\approx N + \frac{2}{(\theta-2)(\theta-1)}\left(2N^{1-\theta} + N^{2-\theta} + 2^{1-\theta}\theta N - 2^{2-\theta}N\right) \\ &= N + \frac{2}{(\theta-2)(\theta-1)}\left(N^{2-\theta} + 2^{2-\theta}(\frac{\theta}{2}-1)N + 2N^{1-\theta}\right) \end{aligned} \quad (5.13)$$

For normally distributed market orders with $\langle |m(t)| \rangle = \sqrt{\frac{2}{\pi}}$, the win due to the spread grows with N like $\frac{N}{2}s\sqrt{\frac{2}{\pi}}$, i.e. with a power of one. In contrast, the highest power of N in the expected inventory is $2-\theta > 1$ since $\theta < 1$ ($\theta \approx 0.73$ in Fig. 5.1). Hence, this strategy can lead to profit only for very short time horizons or very large spread $\frac{s}{\lambda} \gg 1$.

Cross-correlations

In the model described above, the liquidity provider matches every market order so that the price will not move at all. However, in reality a trader has to deal with certain limitations $I_{\max}$ of the inventory. So, if the inventory is full (meaning that the absolute difference between the number of bought and sold shares is larger than $I_{\max}$), she cannot match any more incoming market orders. In this case, she has to let the orders move the price while she has to wait for orders with the opposite direction in order to close her positions. Hence, the trader matches only very few market orders (if $I_{\max} << \sqrt{N}$) and waits for the right orders in the rest of the time. Thus, together with the inventory, the risk is limited by $I_{\max}$, but also the possible profit.

Although this is a parsimonious model, our simulations can lead to qualitative results which can be compared to empirical data if we include inventory limitations so that the price can move during the simulation. At this point, we want to recall some results from my diploma thesis, which were published in [30]. Figure 5.3 displays the cross-correlations between returns and the flow of (a) market orders and (b) limit orders, obtained as an average of ten Nasdaq stocks from the year 2002. The vanishing correlations for market order and limit order flow preceding returns for more than 50 seconds indicate market efficiency, meaning here that returns cannot be predicted over extended periods of time. For other time lacks, market orders (a) exhibit positive correlations with returns that are strongest when the time intervals for returns and order flow overlap (shaded region). In the non-overlapping region, the correlations decay slowly, which is probably due to the strong autocorrelations of market orders [20, 29, 73].

Surprisingly, the correlation function for limit orders with returns, shown in

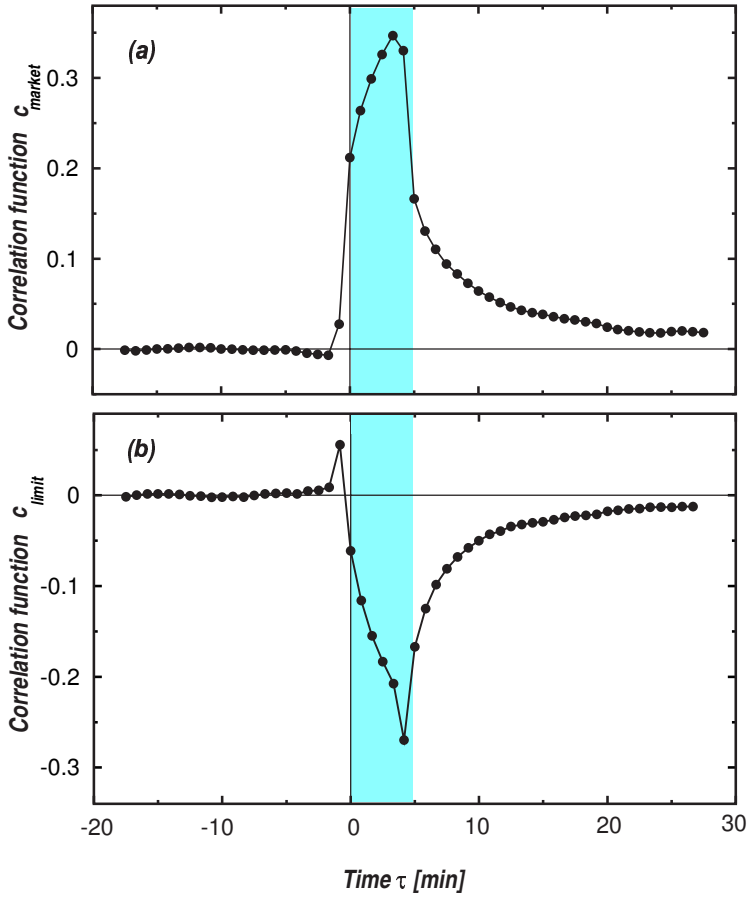

Figure 5.3: Correlation functions between return and signed order flow (buy minus sell orders), obtained as an average from order book data of ten Nasdaq stock from the year 2002. (a) Market orders and returns show strong positive equal time correlations (shaded region) decaying slowly to zero. (b) Limit orders preceding returns have weak positive correlations with them, while equal time correlations (shaded region) are strongly negative. Adapted from [30].

Fig. 5.3(b), exhibits strong equal time anticorrelations. These anticorrelations can be interpreted as an indication that rising prices cause an increased number of sell limit orders whereas falling prices induce additional buy limit orders. In this way, price changes seem to be counteracted by an orchestrated flow of limit orders.

The results for the cross-correlations obtained from the simulation of the liquidity provider strategy are in striking contrast to these empirical results. Figure 5.4 shows the cross-correlations between returns and market orders (circles) as well as limit orders (squares), and in addition the autocorrelations of returns (diamonds), for parameters $s = 0.1$, $\lambda = 0.1$, $I_{\max} = 30$. The smaller the inventory

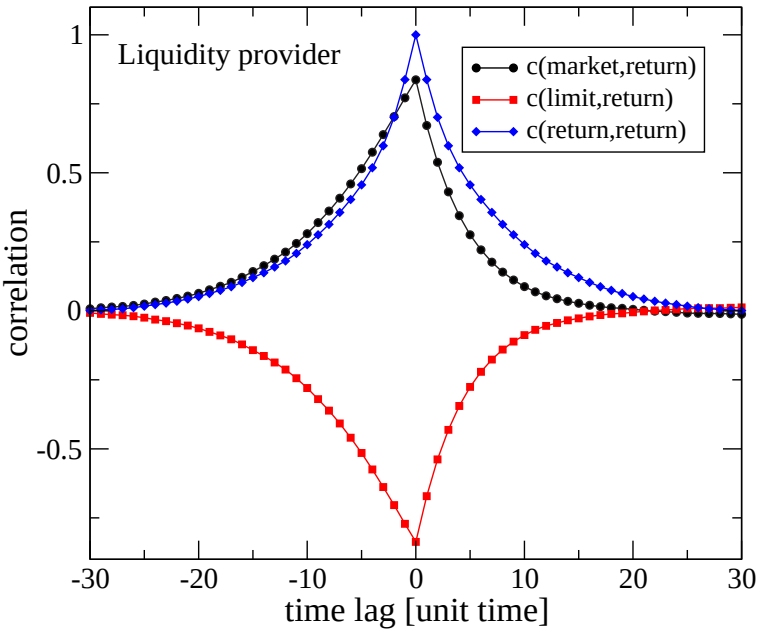

Figure 5.4: Correlations obtained from the simulation of the liquidity provider strategy with parameters $s = 0.1$, $\lambda = 0.1$, $\varrho = 0.8$ and $I_{\max} = 30$. Market orders and returns are highly correlated (circles), and the market order correlations show up in the returns (diamonds) almost unchanged, so that returns are still strongly predictable, in contrast to the empirical results shown in Fig. 5.3. The anticorrelations between returns and limit orders (squares) are somewhat artificial, because the limit order placement is calculated from the market orders by inversion.

limit $I_{\max}$, the more often the trader has to let the price move, so that with relatively small $I_{\max}$ the market order correlations show up in the returns almost unchanged. Hence, returns are still highly predictable, in contrast to reality. The anticorrelations between returns and limit orders are somewhat artificial, because the limit order placement is calculated from the market orders directly (the simulation assumes that there was a limit order placed before so that the market order can be matched).

In summary, the liquidity provider strategy leads to profit due to the spread, but the price impact when closing the positions can lead to losses. Especially with large correlations in the market orders the inventory can grow so large that severe losses are very likely with this strategy. A better strategy should involve a mechanism to limit the inventory without destroying the possibility of making profit. However, the liquidity provider strategy as it is presented here does not seem to explain the correlations and cross-correlations found empirically.

5.3 Front runner strategy

Instead of considering the correlations in market orders as a potential danger, the second strategy presented here uses these correlations for the prediction of future orders in order to optimize the profit. This idea is based on the basic strategy called 'front running' [124]: a front runner knows for some reason that someone wants to place a large market order, say a buy market order. Then, she himself buys this number of shares before the foreseen market order is placed. Due to the market order of the front runner, the price changes to a higher level according to Eq. (5.1). Now, she places sell limit orders at the increased price and waits for the expected buy market order. If this order is actually placed, the front runner can sell the just bought shares at a higher price.

In our model, the front runner uses the correlations between the market orders to predict future market orders. These correlations can be explained qualitatively by order splitting [73]: people split large orders into several smaller pieces and place them consecutively over a larger time period in order to affect the price not too much. Thus, the front runner can step between these split orders and perform her strategy.

According to Eq. (5.2) with $\langle \varepsilon(t+1) \rangle = 0$, the market order $m_{\text{pred}}(t+1)$ predicted for the next time step is

$$m_{\text{pred}}(t+1) = \varrho m(t) \tag{5.14}$$

which leads to the overall expected future order flow

$$M_{\text{pred}}(t) = \sum_{j=1}^{\infty} m_{\text{pred}}(t+j) = m(t) \sum_{j=1}^{\infty} \varrho^j = \frac{m(t)\varrho}{1-\varrho} \quad . \tag{5.15}$$

If the trader follows her front runner strategy, each incoming market order $m(t) = \varrho m(t-1) + \varepsilon(t)$ is split into two parts: the predicted part $\varrho m(t-1)$ due to the previous order is matched against limit orders placed by the trader in the previous time step and thus does not change the price, so that her inventory $I(t)$ is reduced to $I(t) - (m(t) - \varepsilon(t))$. On the other hand, the innovation part due to the random number $\varepsilon(t)$ leads to a new prediction of the future market order flow, and the trader adjusts her inventory by immediately buying (or selling) some shares in order to hold exactly the predicted market orders. Thus, the difference between the new inventory $I(t+1)$ and the inventory $I(t)$ before the

arrival of the current market order is

$$\Delta I(t) = I(t+1) - I(t) = \underbrace{\varepsilon(t)\sum_{j=1}^{\infty}\varrho^j}_{\text{new prediction}} - \underbrace{(m(t)-\varepsilon(t))}_{\substack{\text{executed via}\\ \text{limit orders}}} \quad . \tag{5.16}$$

In order to adapt her inventory, the front runner places market orders

$$m_{\mathrm{FR}}(t) = \varepsilon(t)\sum_{j=1}^{\infty}\varrho^j \quad . \tag{5.17}$$

At the same time, the trader also adapts her limit orders to match the predicted market orders for the next time step according to Eq. (5.14).

The price changes due to the innovation and the market orders of the trader:

$$S(t+1) - S(t) = \lambda(\varepsilon(t) + m_{\mathrm{FR}}) = \lambda\frac{1}{1-\varrho}\varepsilon(t) \quad . \tag{5.18}$$

Contrary to the liquidity provider strategy of the previous section, the spread constitutes a cost for the front runner instead of an opportunity. In order to adapt her portfolio to the predicted order flow, she constantly has to buy and sell shares using market orders. This leads to losses analogous to Eq. (5.3), depending on the average market order $\langle|m_{\mathrm{FR}}|\rangle$ she has to place.

Profit estimation

The front runner's profit can be approximated by dividing the process into two parts: The win the front runner would make if everything would match her prediction, and the loss due to the spread when trading m_{FR} in order to adapt to the innovation.

The win and loss are given by the difference between the buy and the sell price for all traded shares. As an example, we consider a positive (buy) innovation $\varepsilon(t) > 0$. The front runner buys the prediction $V_{\mathrm{buy}} = \frac{\varepsilon(t)\varrho}{1-\varrho}$ at price $S_{\mathrm{buy}} = S(t) + \frac{s}{2} + \frac{\lambda}{2}\frac{\varepsilon(t)\varrho}{1-\varrho}$, causing a price impact of $\Delta S = \lambda\frac{\varepsilon(t)\varrho}{1-\varrho}$. If we neglect future innovations, the front runner can sell her shares $V_{\mathrm{sell}} = V_{\mathrm{buy}}$ at $S_{\mathrm{sell}} = S_{\mathrm{buy}} + \frac{\Delta S}{2}$, leading to a win

$$w_{\mathrm{FR}} = S_{\mathrm{sell}}V_{\mathrm{sell}} - S_{\mathrm{buy}}V_{\mathrm{buy}} = \frac{\Delta S}{2}\frac{\varepsilon(t)\varrho}{1-\varrho} = \frac{\lambda}{2}\left(\frac{\varrho}{1-\varrho}\right)^2\varepsilon(t)^2 \quad . \tag{5.19}$$

Since $\langle\varepsilon^2(t)\rangle = 1$, the expected win after N time steps is

$$\langle w_{\mathrm{FR}}\rangle = N\frac{\lambda}{2}\left(\frac{\varrho}{1-\varrho}\right)^2 \quad . \tag{5.20}$$

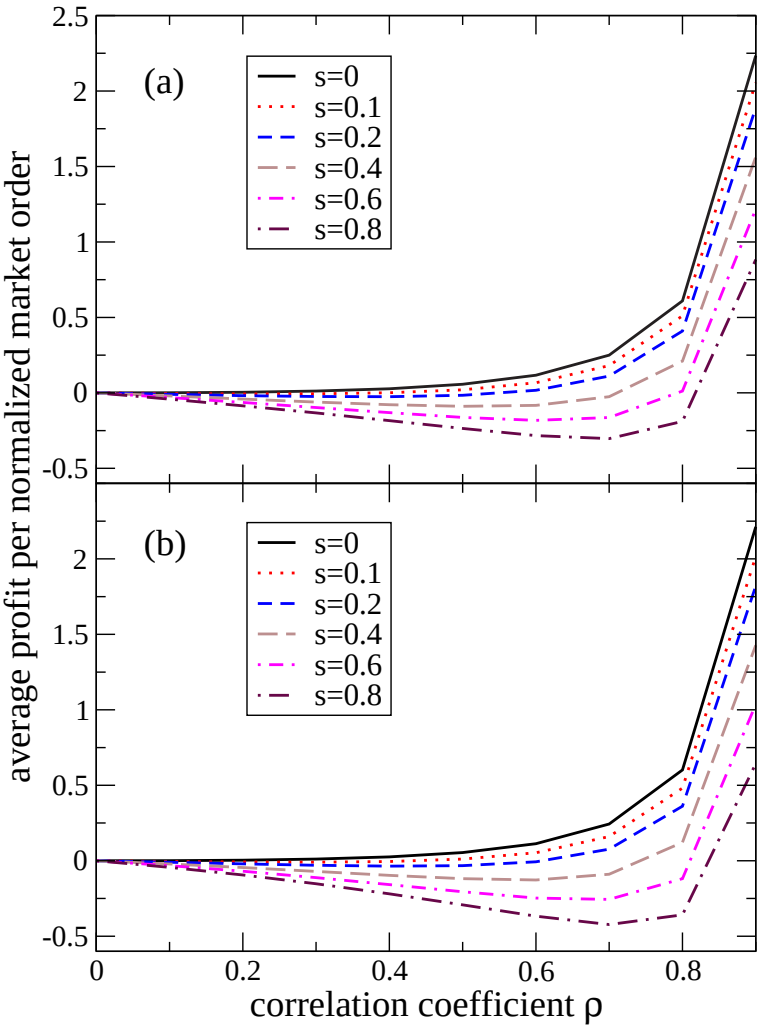

Figure 5.5: Profit per normalized market order in the front runner strategy for different spreads. The price impact coefficient is hold constant at $\lambda = 0.1$. (a) From 100 simulations with $N = 100,000$ time steps each, the average profit at the end is divided by $N \langle |m| \rangle$ in order to obtain the profit per normalized market order. Figure (b) shows the normalized profit per market order obtained from the analytical result Eq. (5.22).

On the other hand, the front runner has to face losses due to the spread when changing her strategy because of new innovations. In the worst case, she buys $V_{\text{buy}} = \frac{\varepsilon(t)\varrho}{1-\varrho}$ shares that she has to sell in the next time step because the innovation $\varepsilon(t+1)$ has a different sign. So, every two time steps she buys and sells V_{buy} loosing sV_{buy}. Hence, the expected loss after N time steps is given by

$$\langle l_{\text{FR}} \rangle = N \frac{s}{2} \frac{\varrho}{1-\varrho} \sqrt{\frac{2}{\pi}} \tag{5.21}$$

which leads to the expected profit

$$\langle pr_{\text{FR}} \rangle = \langle w_{\text{FR}} \rangle - \langle l_{\text{FR}} \rangle = \frac{N}{2} \frac{\varrho}{1-\varrho} \left(\lambda \frac{\varrho}{1-\varrho} - s\sqrt{\frac{2}{\pi}} \right) \quad . \tag{5.22}$$

Figure 5.5 illustrates the dependence of the front runner's profit on the parameters of the model, analogous to Fig. 5.2. Since the trader only acts on forecasts, she does not trade at all if the correlations (resp. ϱ) are zero, which always leads to no profit. However, when ϱ is increased, the front runner starts to take action in the market: the larger the correlations, the more she trades. Now, the front runner looses money due to the spread, but she also wins money from the correlated market orders. For small correlations, this usually leads to losses, but for $\varrho = 0.8$ (corresponding to the empirical correlation in a time horizon of 20 minutes) only very large spreads (eight times the price impact) prevent this strategy from being successful.

The agreement between the simulation results in Fig. 5.5(a) and the analytical result Eq. (5.22) shown in Fig. 5.5 is quite good. However, one sees deviations especially for large values of s, indicating that the approximation Eq. (5.22) overestimates the negative influence of the spread.

Cross-correlations

The front runner strategy agrees much better with empirical data than the strategy of a liquidity provider. Figure 5.6 shows that because of the action of the front runner following her strategy, the correlations between market orders do not show up in the returns (diamonds). This is not surprising because only the unpredictable part of the market order leads to a price change. For empirical data, the return shows a small anticorrelation in very short times, which is known as the 'bid-ask-bounce'. This phenomenon is not featured by our model. However, the empirically found cross-correlations [30] between the return and the market orders as well as the limit orders, which are shown in Fig. 5.3, can be reproduced qualitatively in our model. Figure 5.6 also shows that returns

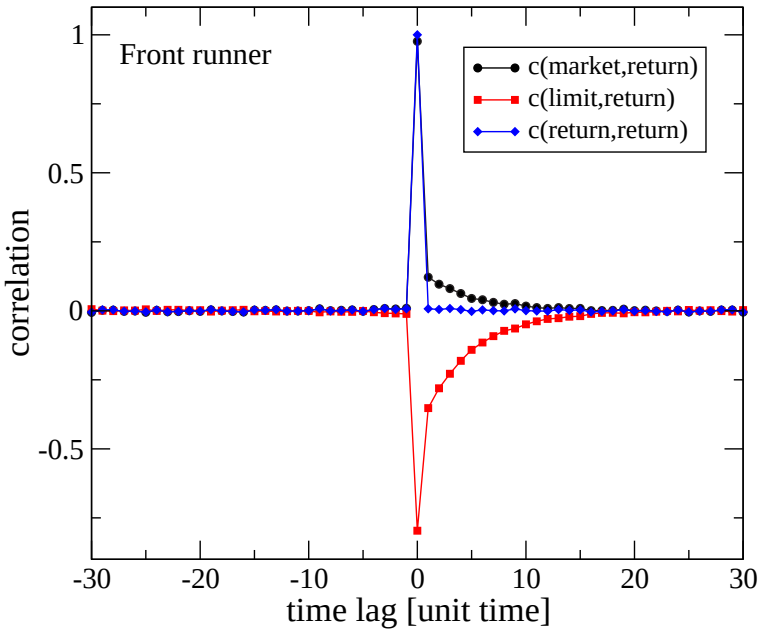

Figure 5.6: Correlations obtained from simulation of the front runner strategy with parameters $s = 0.1$, $\lambda = 0.1$, and $\varrho = 0.8$. Due to the action of the front runner, the correlations in the return (diamonds) disappear. There are no correlations between returns and market orders (circles) or limit orders (squares) for negative time lags, so that returns cannot be predicted from the order flow.

are not correlated with preceding market (circles) and limit orders (squares), so that a prediction of returns due to the order flow is not possible, in agreement with reality.

After the price change, there are positive correlations between returns and market orders, but we also see anticorrelations between the return and following limit orders. This phenomenon was described in [30] as a feedback mechanism reducing the price impact of an order. A possible explanation for this feedback mechanism is that traders hide their large orders because they do not want to provide liquidity to other traders. If the price changes due to a market order, the hidden orders are placed in the order book in expectation of many consecutive market orders. In this way, these market orders do not have a price impact anymore.

5.4 Discussion

We analyzed two different trading strategies that have a very different relation to correlations of market orders. The liquidity provider strategy works against these correlations, leading to losses due to the huge inventory after a certain number of time steps. Possibly, the model proposed here misses features of the market that are important for limiting the inventory. Figure 5.3 showed that a price change has a huge impact on the order flow, which has been shown to account for the quite flat shape of the price impact function of market orders [30]. Thus, rising prices should lead to reactions of the market participants that limit the price change due to an order. Hence, the model could be expanded in such a way that deviations from a reference price (e.g. given by a running average of past prices) lead to an increased flow of market orders that mean revert the price. However, this is not included in the model since we wanted to show that the front runner strategy can qualitatively generate the empirical correlations, which would be meaningless if we put these correlations in the model from the beginning.

The analysis in this chapter showed that by using the correlations in a successful strategy, their influence on returns is destroyed, which then exhibit the empirically observed lack of correlations. Though the presented front runner strategy is somewhat artificial, it illustrates how market efficiency can take place, so that arbitrage opportunities disappear as soon as someone takes advantage of them.

6 Test for statistical significance of empirical correlation matrices

So far, we have analyzed the autocorrelations in the return time series and studied the mechanisms underlying large stock returns. In this chapter, we do not focus on the risk inherent in single large price movements, but rather concentrate on the risk associated with the statistical uncertainty of the return of a portfolio [5]. This uncertainty is measured by the volatility, which is here defined as the variance of the portfolio.

Markowitz investigated the selection of an efficient portfolio that has minimal risk for a given expected return [71]. Such a portfolio can be obtained by appropriate diversification, meaning that one invests in a large number of stocks with weak cross-correlations. In this way, the loss of some of the stocks can be compensated by the gains of others. In order to be efficient, such diversification has to lead to a minimal variance and thus a minimal risk of the portfolio. Since the variance of the portfolio is determined by the cross-correlations between the return time series of all involved stock, one has to evaluate these correlations in order to calculate the risk.

The estimation of cross-correlations between a large number of time series can be very difficult due to the "curse of dimensionality": If the number of time series is of the same order as their length, a calculation of the correlation coefficients results in large errors, since in this case the number of calculated values is comparable to the number of data points available for their calculation. On the other hand, cross-correlations change over time so that one may have to study rather short time series. If one simulates a random market of N totally uncorrelated time series with a length $T \approx N$, the resulting correlation matrix will deviate significantly from the unit matrix. In the same way, the cross-correlations calculated from empirical time series contain spurious correlations that can make a precise risk estimation difficult.

In physics, correlations can be related to the interactions of particles. If a physical system is too complex so that the interactions cannot be estimated precisely, one uses random matrices to obtain information about the properties

of the system [72]. For instance, the excitation spectra of nuclei on low-energy levels can be explained by considering a model of independent particles with an average potential [126, 127]. For intermediate energies, the interactions between the nucleons cannot anymore be described correctly with an average potential, and it is impossible to explain the individual states [72]. Therefore, instead of trying to describe the individual states, one focuses on estimating average properties of the system by using a statistical theory where the Hamiltonian is described by a random matrix. Random matrix theory has also been successfully applied to economics to study correlations between financial time series [128, 129].

In this chapter, we want to use random matrices for studying correlation matrices of finite time series. We analyze a hypothesis test [125] that can distinguish spurious correlations from real correlations. Specifically, it tests whether the matrix contains statistically significant correlations or whether it is equivalent to the unit matrix. For covariance matrices, such tests were formulated by [130, 131] and later generalized to the degenerate case $N > T$, where the matrix dimension exceeds the time series length [132].

This chapter is organized as follows. In section 6.1 we define the test statistics and calculate its T–limiting distribution. The properties of the test for finite samples are analyzed in section 6.2, and a summary is given in section 6.3.

6.1 Definition of the test

We consider N time series $X_i(t)$ of length T, which are normally distributed with population mean μ_i and population correlation matrix C. For each time series $X_i(t)$, we define a new normalized time series $x_i(t)$ with zero mean as

$$x_i(t) = \frac{X_i(t) - \overline{X}_i}{s_i} \quad , \tag{6.1}$$

where $\overline{X}_i$ and $s_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_i(t) - \overline{X}_i)^2$ are the mean and the variance of the sample $X_i(t)$. Due to the normalization, the population covariance matrix of the $\{x_i\}$ is given by C as well.

We denote the sample correlation matrix of the time series $x_i(t)$ by

$$\tilde{C}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T} x_i(t)x_j(t) \ . \tag{6.2}$$

The correlation matrix $\tilde{\mathsf{C}}$ of the sample will generally be different from the true correlation matrix C if the considered time series have a finite length. The

question we want to answer in this chapter is whether it is possible to distinguish between spurious and real correlations if one only knows the sample correlation matrix, e.g. from an empirical data set. In particular, we want to test if a given sample correlation matrix is compatible with the null hypothesis of uncorrelated time series, i.e. a population correlation matrix C equivalent to the unit matrix.

A statistical test usually decides whether or not a given hypothesis is to be rejected, which then means that the hypothesis is false with a given probability. In our case, where we are looking for correlations in the time series, the null hypothesis corresponds to the absence of correlations. If now the test rejects this hypothesis, this can indicate that the alternative is true, i.e. that there are correlations. However, it is possible that both the hypothesis and the alternative are rejected, so that a rejection of the hypothesis does not necessarily mean that the alternative is true.

A test is characterized by size and power, which quantify how often the test correctly decides to reject or not reject the hypothesis. The size is the probability for rejecting the null hypothesis when it is true. Hence, if the size is zero, the test result is always correct if there are no real correlations in the data. A larger size indicates less accurate test results, so that sometimes the test rejects the null hypothesis although there are indeed no correlations. However, the size alone is insufficient for characterizing a test, since a test with arbitrary size could be easily created by simply rejecting every hypothesis with a given probability. Hence, the size is meaningful only in conjunction with the power, which characterizes the test in the presence of correlations: the power is the probability of rejecting the null hypothesis if there are correlations. Here, large values indicate that the test has a good capability of detecting correlations.

6.1.1 Test statistics

For the present test, we analyze the test statistics

$$\mathcal{R} = \frac{1}{N}\mathrm{tr}\left[\tilde{\mathsf{C}}^2\right] - 1 \quad , \tag{6.3}$$

allowing to distinguish true correlations from spurious correlations that result from a finite time series length.

The test is based on the knowledge about the expected distribution of $\mathcal{R}$ for many random matrices with a given N and T. Then, one can test if the value of $\mathcal{R}$ obtained from an empirical sample matrix agrees with the distribution. In order to get the binary result whether or not the hypothesis has to be rejected, one defines a significance level corresponding to a critical value $\mathcal{R}_{\mathrm{crit}}$ of the test

statistics. The empirical $\mathcal{R}$ can be compared to $\mathcal{R}_{\text{crit}}$, so that the hypothesis is rejected if $\mathcal{R} > \mathcal{R}_{\text{crit}}$.

If one wants to show analytically that the test statistics exhibits the desired properties, one first has to show that $\mathcal{R}$ is (N,T)–consistent, i.e. that the power of the test to separate the null hypothesis from the alternative converges to one as N and T go to infinity together. Specifically, one has to prove that the expectation value of the test statistics under the joint limit $T \to \infty$, $N \to \infty$, and $T/N = Q$ uniquely allows to decide whether the population correlation matrix differs from the unit matrix. Next, one calculates the T–limiting distribution and shows that its (N,T) asymptotics is indeed the true (N,T) asymptotics of $\mathcal{R}$.

For a similar test for covariance matrices, this procedure was successfully performed by Ledoit and Wolf [132]. For correlations matrices, the normalization Eq. (6.1) of the time series $x_i(t)$ leads to additional problems. Specifically, while the time series $X_i(t)$ are independent with $\langle X_i(t_1)X_i(t_2)\rangle = 0$ for $t_1 \neq t_2$, the normalization creates correlations in the $x_i(t)$. One has

$$\begin{aligned}\langle x_i(t_1)x_i(t_2)\rangle &= \langle \frac{(X_i(t_1)-\overline{X_i})(X_i(t_2)-\overline{X_i})}{\frac{1}{T-1}\sum_{k=1}^{T}(X_i(k)-\overline{X}_i)^2}\rangle \qquad (6.4)\\ &= \langle \frac{1}{\frac{1}{T-1}\sum_{k=1}^{T}X_i^2(k) - \frac{1}{T(T-1)}\sum_{p,q=1}^{T}X_i(p)X_i(q)}\Big(X_i(t_1)X_i(t_2)\\ &\quad -\frac{1}{T}\big[X_i(t_1)+X_i(t_2)\big]\sum_{l=1}^{T}X_i(l) + \frac{1}{T^2}\sum_{m,n=1}^{T}X_i(m)X_i(n)\Big)\rangle\end{aligned}$$

Since the sums run over all $t = 1 \ldots T$, there are products of $X_i(t)X_i(t)$ for the same time t, which lead to a finite contribution to the expectation value. One gets

$$\langle x_i(t_1)x_i(t_2)\rangle = -\frac{1}{T} \quad . \qquad (6.5)$$

6.1.2 T–limiting distribution

In this thesis, we will focus on the numerical properties of the test and not prove its consistency, which is done in [133]. However, in order to make the test useful for empirical studies, we first have to find the T–limiting distribution. For this calculation, we can neglect the correlations Eq. (6.5) between the $x_i(t)$ since they vanish in the limit of large T. Hence, we focus on uncorrelated $x_i(t)$ with $\langle x_i(t_1)x_i(t_2)\rangle \approx 0$ for $t_1 \neq t_2$ in order to derive a solution for the T–limiting distribution under the assumption that the true correlation matrix C is the identity.

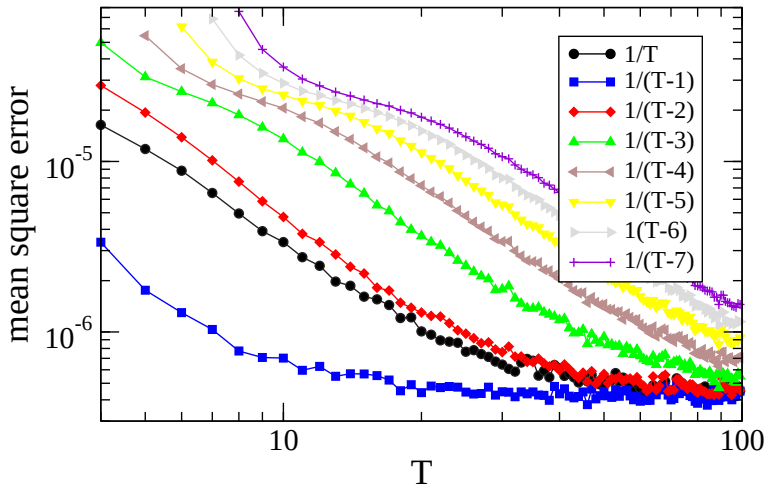

Figure 6.1: Comparison of different factors $\alpha(N,T) = \frac{1}{T-m}\frac{2}{N}$ (corresponding to $\beta = 0$ in Eq. (6.12)), $m = 0,\ldots,7$ and $N = 16$ for the T–limiting distribution of the test statistics $\mathcal{R}$. The mean square difference between the probability density function of $\frac{\mathcal{R}}{\alpha}$ and $\chi^2_{N(N-1)/2}$ decreases for increasing T. The best fit (i.e. smallest error) for small T is obtained with $\alpha = \frac{2}{N(T-1)}$. The curves are shown for 10,000 simulations.

We introduce the random variables $\eta_{ij}(t)$ by decomposing

$$\frac{1}{T-1}\sum_{t=1}^{T} x_i(t)x_j(t) = C_{ij} + \frac{1}{\sqrt{T}}\eta_{ij} \quad . \tag{6.6}$$

The newly defined random variables η_{ij} have expectation value zero and variance

$$\mathrm{Var}(\eta_{ij})^2 = \begin{cases} 1 & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \quad . \tag{6.7}$$

With these variables, we can rewrite

$$\frac{1}{N}\mathrm{Tr}[\tilde{\mathsf{C}}^2] = \frac{1}{N}\sum_{i,j=1}^{N}\tilde{C}_{ij}^2 = \frac{1}{N}\sum_{i,j=1}^{N}(C_{ij} + \frac{1}{\sqrt{T}}\eta_{ij})^2 \tag{6.8}$$

$$= \frac{1}{N}\sum_{i,j=1}^{N}(C_{ij}^2 + \frac{1}{\sqrt{T}}C_{ij}\eta_{ij} + \frac{1}{T}\eta_{ij}^2) \tag{6.9}$$

$$= 1 + \frac{2}{NT}\sum_{i<j}(\eta_{ij})^2 \quad . \tag{6.10}$$

The second term in Eq. (6.10) is the sum of $N(N-1)/2$ squares of standard normal distributed variables and hence follows a $\chi^2_{N(N-1)/2}$–distribution. We conclude that the T–limiting distribution of the test statistics $\mathcal{R}$ is given by

$$\mathcal{R} \xrightarrow{D} \frac{2}{TN}\,\chi^2_{N(N-1)/2} \quad . \tag{6.11}$$

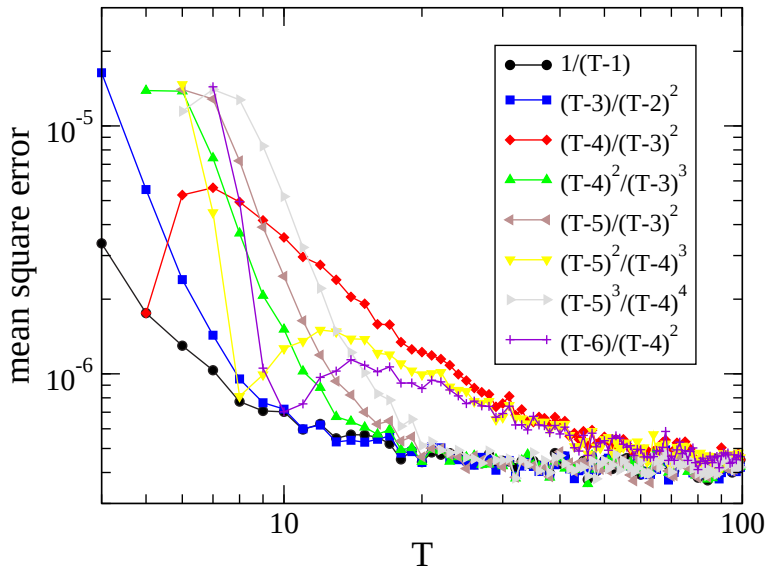

Figure 6.2: Comparison of different factors $\alpha(N,T) = \frac{(T-n)^\beta}{(T-m)^{\beta+1}} \frac{2}{N}$ for the T–limiting distribution of the test statistics $\mathcal{R}$. The figure shows the mean square difference (MSE) between the probability density function of $\frac{\mathcal{R}}{\alpha}$ and $\chi^2_{N(N-1)/2}$ depending on T. Shown are the eight curves with the lowest average error $\langle \text{MSE}(P_{\mathcal{R}}, P_{\chi^2}, T) \rangle$ for $T = max(1,m,n), \ldots, 100$. The best fit (i.e. smallest error) for small T is again obtained with $\alpha = \frac{2}{N(T-1)}$. The curves are shown for 10,000 simulations with $N = 16$.

6.2 Test properties for finite samples

The test proposed above is supposed to distinguish spurious correlations from real correlations in empirical time series. Such a test is of particular interest if the data sets are small so that one can only analyze short time series, since here the effect of spurious correlations is large. Therefore, it is important to analyze the properties of the test for finite samples, which we will do in the following using numerical simulations.

6.2.1 Finite size properties of T–limiting distribution

The T–limiting distribution is only exact for large T, so it is not clear whether the prefactor $\frac{2}{TN}$ in Eq. (6.11) is a good approximation for small and intermediate T where corrections of the order $\frac{1}{T}$ can be important. By simulating 10,000 random matrices in order to obtain good statistics for $\mathcal{R}$, we compare

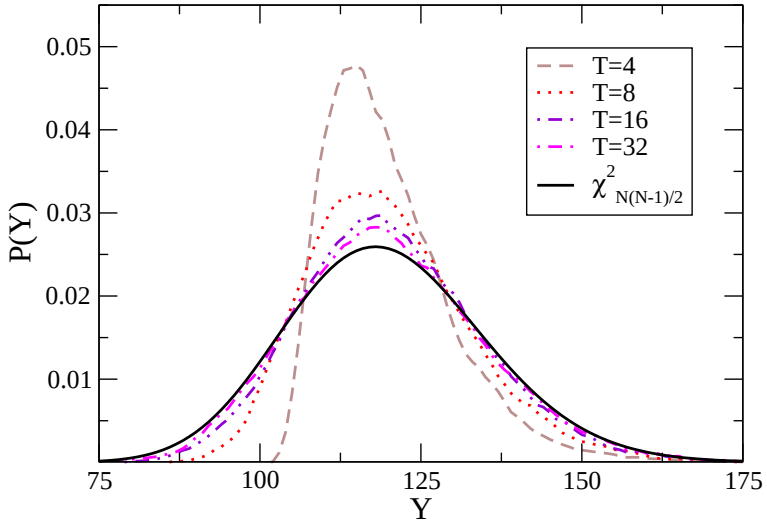

Figure 6.3: . A comparison of the PDF of $\frac{1}{2}\mathcal{R}N(T-1)$ for different T with the PDF of $\chi^2_{N(N-1)/2}$. The approximation is reasonable already for very small $T = 4$ and becomes very good for larger T. The curves are shown for 10,000 simulations with $N = 16$ and smoothed by a moving average.

the prefactor $\frac{2}{TN}$ to other factors α of the form

$$\alpha = \frac{(T-n)^\beta}{(T-m)^{\beta+1}}\frac{2}{N} \quad . \tag{6.12}$$

To this end, we calculate the probability density functions (PDF) of $\frac{\mathcal{R}}{\alpha}$ for different α and compare them with $\chi^2_{N(N-1)/2}$. The mean square error MSE between the PDF $P_\mathcal{R}$ of $\frac{\mathcal{R}}{\alpha}$ and the PDF P_{χ^2} of $\chi^2_{N(N-1)/2}$ is given by

$$\mathrm{MSE}(P_\mathcal{R}, P_{\chi^2}) = \langle (P_\mathcal{R}(y) - P_{\chi^2}(y))^2 \rangle \tag{6.13}$$

where y runs from 0 to $10N(N-1)$ in steps of 0.1. Since here many values are almost zero in both distributions, the MSE gets generally quite small. However, we only want to compare the distributions for different α so that only their relative MSEs are important. In Fig. 6.1 we show the results depending on T for $\alpha(N,T) = \frac{1}{T-m}\frac{2}{N}$ (corresponding to $\beta = 0$ in Eq. (6.12)), $m = 0,\ldots,7$ and $N = 16$. Although for large T all shown distributions converge to the χ^2–distribution, there are significant differences for small T. The curve with $m = 1$ is clearly the best one, while $m = 0$ and $m \geq 2$ lead to larger deviations. This indicates that a corrected prefactor $\frac{2}{(T-1)N}$ could lead to better test results for finite T. For very small T, where the T–limiting distribution should not be very accurate, even for the best prefactor $\frac{2}{(T-1)N}$ the MSE is indeed one magnitude larger than for large T, but the error becomes smaller very fast.

		T						
		4	8	16	32	64	128	256
	4	0.01	0.03	0.04	0.05	0.05	0.05	0.05
	8	0.02	0.03	0.04	0.05	0.05	0.05	0.05
	16	0.02	0.03	0.04	0.04	0.05	0.05	0.05
N	32	0.02	0.03	0.04	0.05	0.05	0.05	0.05
	64	0.02	0.03	0.04	0.05	0.05	0.05	0.05
	128	0.02	0.04	0.04	0.05	0.05	0.05	0.05
	256	0.02	0.03	0.04	0.04	0.04	0.05	0.05

Table 6.1: Size of the test $\mathcal{R}$ from simulation of 10,000 Monte Carlo simulations for each pair (N, T). The null hypothesis is rejected when the test statistics exceeds the cutoff $\mathcal{R}_{\text{crit}}$ obtained from the chi squared approximation with $\alpha = \frac{2}{N(T-1)}$. For $T \geq 32$, the actual size agrees well with the nominal size for all values of N.

We did not find any other combination of $m, n, \beta = 0, \ldots, 7$ that leads to a better agreement between the simulated distribution of $\mathcal{R}$ and the distribution of $\alpha\chi^2_{N(N-1)/2}$. Since we cannot plot all curves for all combinations of m, n, β, we show in Fig. 6.2 only the eight best fitting curves, meaning that for these curves the average error $\langle \text{MSE}(P_{\mathcal{R}}, P_{\chi^2}, T) \rangle$ for $T = max(1, m, n), \ldots, 100$ is minimal. Also in this figure, the factor $\alpha = \frac{2}{N(T-1)}$ clearly shows the best results, especially for small T.

Though there might be other pefactors of a different form fitting the real distribution of $\mathcal{R}$ better, we conclude that the prefactor $\alpha = \frac{2}{N(T-1)}$ leads to a good fit of that distribution. In the following, we will use this corrected prefactor instead of the one obtained above for the T–limiting distribution.

In order to illustrate the finite sample properties of the test statistics with the corrected $\alpha = \frac{2}{N(T-1)}$, we compare the PDF of $\frac{1}{2}\mathcal{R}N(T-1)$ for $N = 16$ and different T with the PDF of $\chi^2_{N(N-1)/2}$. The PDFs are shown in Fig. 6.3, where the curves are smoothed by a moving average. The approximation looks reasonable already for very small $T = 4$, becoming very good for larger T. For $T = 100$ (not shown here), the agreement is almost perfect.

		T						
		4	8	16	32	64	128	256
	4	0.02	0.08	0.20	0.44	0.77	0.98	1.00
	8	0.04	0.17	0.43	0.79	0.98	1.00	1.00
	16	0.09	0.35	0.74	0.97	1.00	1.00	1.00
N	32	0.18	0.57	0.92	1.00	1.00	1.00	1.00
	64	0.33	0.77	0.99	1.00	1.00	1.00	1.00
	128	0.48	0.90	1.00	1.00	1.00	1.00	1.00
	256	0.63	0.96	1.00	1.00	1.00	1.00	1.00

Table 6.2: Power of the test from simulation of 10,000 correlated matrices. The data is generated from a factor model with one factor, thus $g_i(t) = 0.5 * f(t) + \epsilon_i(t)$ for each time series. The null hypothesis is rejected when the test statistics exceeds the 95 % cutoff point obtained from the chi squared distribution with $\alpha = \frac{2}{N(T-1)}$ (approximated by a Gaussian for $N \leq 64$ with an error of ~ 0.1 percent).

6.2.2 Size and power of the test

In order to analyze the finite sample properties of the test statistics $\mathcal{R}$ with the adapted distribution $\alpha\chi^2_{N(N-1)/2}$ given by

$$\frac{2}{(T-1)N}\chi^2_{N(N-1)/2} \quad , \tag{6.14}$$

we study size and power, the standard measures characterizing a statistical test. First, we study the test size, i.e. the probability for rejecting the null hypothesis when it is true. We simulate N i.i.d. and normally distributed time series X_i of length T. For each simulation run, we compute the sample correlation matrix $\tilde{\mathsf{C}}$ and calculate the test statistics R from Eq. (6.3). The critical value $\mathcal{R}_{\text{crit}}$ is obtained from the upper 5% quantile of the T–limiting distribution, so that 5% of the R would be rejected even with no correlations in the data and with a perfect agreement of this distribution and the real distribution of $\mathcal{R}$. This value of 5% is called nominal size, because the size should approximate this value for T if the test works correctly.

Table 6.1 shows the results of 10,000 simulations for both N and T from the set $\{4, 8, 16, 32, 64, 128, 256\}$. For each (N, T), we compare the test statistics $\mathcal{R}$ with a critical value $\mathcal{R}_{\text{crit}}$ that is calculated for this (N, T) from the T–limiting distribution. For $T \geq 32$, the test works very well and the actual size agrees well with the nominal size, independent of N. For $T \leq 16$, the actual size is smaller

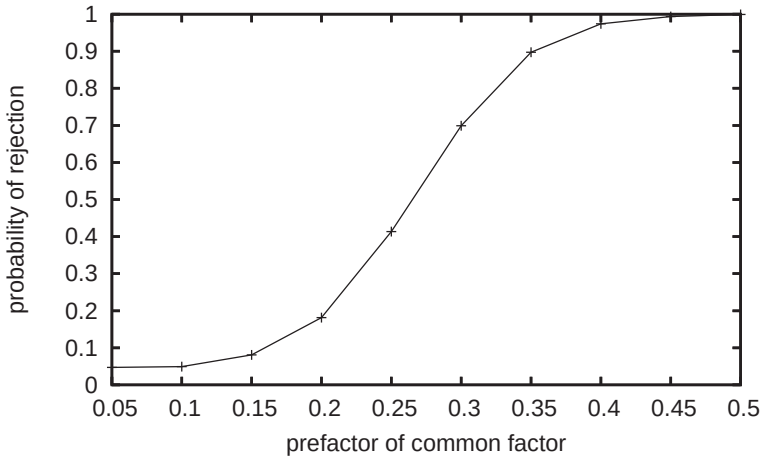

Figure 6.4: Power of the test with $\alpha = \frac{2}{N(T-1)}$ for different correlation strengths. The values are obtained from 10,000 simulations of time series according to Eq. (6.15) with γ varying from 0.05 to 0.5. The results are displayed for $N = T = 32$.

than the nominal size, i.e. $\mathcal{R}_{\text{crit}}$ is too large so that the test rejects more than 5% of the simulations. This suggests that for small T the test is too restrictive, which might also influence the power of the test to detect correlations.

In order to analyze the power of the test, we turn to simulate correlated time series so that the null hypothesis should be rejected. We start with a one factor model, where each time series X_i^1 is generated by

$$X_i^1(t) = \gamma * f(t) + \varepsilon_i(t) \tag{6.15}$$

with $i = 1, ..., N$, $t = 1, ..., T$, and $\gamma = 0.5$. While the normally distributed random numbers $\varepsilon_i(t)$ are generated for each time series, the also normally distributed random numbers $f(t)$ are the same for every time series in one time step. Thus, with this choice of $\gamma = 0.5$ the simulation generates strong correlations. We apply the test to the correlation matrices of 10,000 simulations for different combinations of N and T (Table 6.2). In contrast to the size, which mostly depends on T, the power increases with both increasing T and N since the eigenvalue due to the common factor is larger if there are more time series with this factor.

The power of the test to detect correlations depends largely on the strength of the correlations in the considered time series. By adjusting the prefactor γ of the common factor in Eq. (6.15), one can generate correlations of different strengths. The dependence of the power from these correlations is displayed in Fig. 6.4 for $N = T = 32$. One finds that the power is very good for $\gamma \gtrsim 0.4$,

		T						
		4	8	16	32	64	128	256
	4	0.03	0.12	0.45	0.89	1.00	1.00	1.00
	8	0.03	0.13	0.44	0.92	1.00	1.00	1.00
	16	0.04	0.14	0.45	0.94	1.00	1.00	1.00
N	32	0.04	0.14	0.46	0.95	1.00	1.00	1.00
	64	0.05	0.15	0.46	0.95	1.00	1.00	1.00
	128	0.04	0.15	0.47	0.95	1.00	1.00	1.00
	256	0.05	0.15	0.47	0.95	1.00	1.00	1.00

Table 6.3: Power of the test from simulation of 10,000 correlated matrices. The data is generated such that half of the eigenvalues of the correlation matrix are equal to 0.5 and the other half equal to 1.5. The null hypothesis is rejected when the test statistics exceeds the 95 % cutoff point obtained from the chi squared distribution with $\alpha = \frac{2}{N(T-1)}$ (approximated by a Gaussian for $N \leq 64$ with an error of ~ 0.1 percent).

but vanishes if the correlations are too small. Larger values of N and T allow the detection of smaller correlations.

In a further analysis using different correlations, the time series are generated such that half of the eigenvalues of the correlation matrix are 0.5 and the others are 1.5. The time series are pairwise given by

$$X_{2i-1}^{\text{corr}}(t) = \varepsilon_{2i-1}(t) \tag{6.16}$$

$$X_{2i}^{\text{corr}}(t) = \sqrt{3}X_{2i-1}(t) + \varepsilon_{2i}(t) \tag{6.17}$$

where $i = 1, ...N/2$. This definition generates rather small correlations, which are difficult to detect for the test. Hence, the power of the test for small T is quite weak, which is displayed in table 6.3. Reasonable power is only obtained for $T \geq 16$, but the test works perfectly if T is large enough ($T > 32$). The influence of N on the power is only weak, larger N only lead to a slight increase of the power.

In Fig. 6.3 we saw that for very small T there are significant deviations between the distribution of $\mathcal{R}$ and the χ^2–distribution Eq. (6.14). This suggests that the cut-off point R_{crit} derived from the T–limiting distribution might also not be appropriate for small T. This is also indicated by the results displayed in table 6.1, showing that for $T < 32$ the cutoff point obtained from the T–limiting distribution is too large since too few simulations are rejected. Thus, an adapted

		T						
		4	8	16	32	64	128	256
	4	0.10	0.20	0.50	0.90	1.00	1.00	1.00
	8	0.09	0.19	0.48	0.93	1.00	1.00	1.00
	16	0.09	0.18	0.48	0.94	1.00	1.00	1.00
N	32	0.09	0.18	0.50	0.95	1.00	1.00	1.00
	64	0.09	0.18	0.50	0.95	1.00	1.00	1.00
	128	0.09	0.19	0.48	0.96	1.00	1.00	1.00
	256	0.09	0.18	0.50	0.96	1.00	1.00	1.00

Table 6.4: Power of the test with from simulation of 10,000 correlated matrices. the data is generated such that the correlation matrix has half of the eigenvalues equal to 0.5 and the other half equal to 1.5. The null hypothesis is rejected when the test statistics exceeds the 95 % cutoff point, which is here not obtained from the chi squared distribution Eq. (6.14) but is adapted so that always 5% of the simulations of uncorrelated time series are rejected, even for small T.

cutoff point would be smaller and the test would reject the null hypothesis for more simulations, possibly resulting in a better power for correlated time series. In order to improve the power for small values of T, we adjust the cutoff point R_{crit} so that the size is 0.05 also for small T, i.e. that 5% of the simulations of uncorrelated time series are rejected.

The results with adapted R_{crit} can be seen in table 6.4. Though one finds a slight improvement of the power, the corrections are quite small. The effect is strongest for $T = 4$, where the power increases from values around 0.04 to values around 0.09, but the power of the test remains weak. For larger T, the improvement becomes quite small. This is not surprising because also the correction of R_{crit} is small for these T, since the distribution of R is close to the distribution Eq. (6.14) for large T, as shown in Fig. 6.3.

6.3 Summary

We developed and studied numerically a hypothesis test that is able to distinguish between spurious and real correlations. We calculated the T–limiting distribution analytically and used Monte Carlo simulations to adapt its prefactor in order to obtain a better agreement with the data for small T. This

prefactor leads to a significantly better agreement with the real distribution of the test statistics for small samples. Using the corrected prefactor, the studied significance test generally leads to very good results for numerous (large N) as well as long samples (large T). However, if the sample length is very short ($T \lesssim 32$), the power and the size of the test decrease. One reason might be that the obtained value of α is still not the best choice for small samples. Possibly, the corrections due to the correlations of order $\frac{1}{T}$ that we neglected in the limit of large T might be important at small T. However, even if we use the simulated distribution of the test statistics R to adapt R_{crit} so that the actual size matches the nominal size, the power for small T remains weak. Nevertheless, even with these limitations for very short time series the test could be a useful tool to analyze financial as well as physical time series.

In summary, this thesis studied various phenomena of stock returns. First, we found self-similar features in the time periods following large and intermediate crashes, and this self-similarity could be related to memory in the volatility. Then, we studied the mechanisms leading to large returns. We found that large returns are not due to one single effect, but rather depend on several quantities: In intervals with a fixed number of trades, the concurrence of a large number difference and a large mean tick return size leads to non-Gaussian returns. By analyzing the price impact function in time intervals, we showed that its time-varying slope can be a measure for liquidity, which together with the volume imbalance can explain large returns. In this picture, large price changes occur because in the respective time interval the liquidity is low, so that an intermediate volume can cause a very large return.

The notion that traded volume moves the price seemed to be paradoxical: the signs of orders are strongly correlated, but uncorrelated returns appear when these orders are executed. We showed that this lack of correlations in the returns can be explained by a trading strategy that uses the order sign correlations for an increased profit. Finally, we turned from the analysis of time correlations to studying cross-correlations between different time series. We presented a test that can distinguish spurious correlations from real correlations in the correlation matrix of finite samples. In a further study, this test could be applied to empirical correlation matrices in order to estimate their real correlations.

A Appendix: Data sets and analysis methods

In this appendix, we want to give an overview and more detailed information about the data sets studied in this thesis. We also want to explain the filtering methods used to remove recording errors and their influence on the results. Moreover, we describe the programming methods and how these programs can be used for future research.

A.1 Data sets and filtering

For this thesis, we studied four different financial data sets:

- (i) the one minute return time series of the S&P500 index from 1984 to 1989,
- (ii) the TAQ data base of the year 1997 for the 100 most frequently traded stocks,
- (iii) the one minute return series of General Electric (GE) stock from the TAQ data base of 2001, and
- (iv) complete order book information from the Island ECN for the ten most frequently traded stocks in 2002.

In the following, we will give detailed information about the data sets.

(i) The S&P500 index is composed of 500 stocks of mostly US-American companies that are traded at major US stock exchanges. Along with the Dow Jones Industrial Average and the Nasdaq Composite Index it is one of the most important indices in the US. The data set studied provides the index value in every minute of a trading day, so that a time series of one minute returns can be created. In the last hour of the trading day, the data set sometimes contains no records, so that these parts were excluded from the analysis (instead of filling them with zero returns). All in all, the whole data set contains about 500,000 data points for the time period 1984 to 1989.

(ii) The Trades And Quotes (TAQ) data base is provided by the New York Stock Exchange (NYSE) and contains data for all stocks traded at NYSE, Nasdaq, and American Stock Exchange (AMEX). In our data set for the year 1997, for each month the data set consists of two files: the first file contains all trades, i.e. the transaction price, the number of shares (volume) traded, and the time when the trade took place. The other file gives information about the quotes, which includes the bid and ask price, the volume present at bid and ask price, and a time stamp for the quote. Using the time stamps of the recorded trades and quotes, one can match all trades with the respective quotes which are chosen as the last quote before the trade happened. The quotes are used to calculate the midquote price but also for the Lee and Ready algorithm [105] that labels trades as buyer or seller initiated, depending on whether the transaction price is larger or smaller than the midquote price.

Unfortunately, this data set exhibits some recording errors: First, there are obvious misrecordings due to typos, so that sometimes the price jumps from 10 to 100 and immediately back to 10. On other occasions, in one trade the price changes by for example 20%, stays at this level for a few trades and jumps back exactly to the old price. When these events happen, the quotes do not change at all, which indicates that these jumps are artificial and do not reflect the natural behavior of price changes.

In order to remove these errors from the data set, we used the algorithm of Chordia et al. [114], which discards all trades for which the difference between trade price and midquote price is larger than four times the spread. After applying the filter algorithm, we checked visually the return and trading volume time series surrounding the largest price changes and found no evidence for remaining recording errors.

The filter algorithm is quite restrictive and removes about one percent of all transactions. This has a significant effect on the cumulative distribution function $P(G > x)$ of returns. We already pointed out that it is a common assumption that the tail of the return distribution follows a power law with $P(G > x) \sim x^{-\alpha}$, where α is around three. Here, we do not want to contribute to the discussion whether a power law is the best description of the return distribution, and which exponent is the correct one. However, one can use a power law fit to describe the changes caused by the filtering. For the raw data without any filtering, we find $\alpha = 2.1$, after applying the filter we find $\alpha = 3.9$ by fitting a straight line in a double logarithmic plot.

We note that the applied filtering algorithm removes not only the obviously erratic events described above, but also strong oscillations of several standard

deviations that are probably due to the combination of different ECNs: a large price change might not be reported by all ECNs at the same time. If the price changes on the leading ECN, there might still be limit orders at smaller ECNs providing opportunities to trade at the old price. Arbitrage traders exploit these remaining orders, so that some time after the "true" price change, there are still records of trades occurring at the old price. Though these oscillations are not due to recording errors, they are an artifact of the trading system and thus rightly removed by the filtering method.

(iii) The TAQ data for 2001 appears to have less recording errors than the data from 1997, so that we only needed to discard the first and last 15 minutes of a trading day to get rid of unusual (i.e. artificial) price fluctuations. Since we did not use this data set for studying single extreme price fluctuations, possible recording errors still contained in the data set have no negative influence on the results.

(iv) The entire market information of the Island ECN is stored in text files. We processed the 60GB of raw data in order to extract the data for the ten most frequently traded stocks [100] in one file for each stock and trading day. In these files, each line represents a message of one of the four major types: add limit order, cancel limit order, execute limit order, or trade message. The first three types allow a complete reconstruction of the order book at every instant of time, whereas "trade message" announces the execution of hidden orders which are not visible in the order book.

Each message contains all necessary information: The ticker symbol of the respective stock, the time past midnight in milliseconds, the number of shares, the limit price, an indicator whether it is a buy or sell order, and a unique order reference number. We use this number as a key to store and identify each order in our data structure and perform (partial) executions and cancelations until an order is completely executed or canceled. Since all open orders are purged from the book every evening, we can process the data day by day starting with an empty book each morning. The order book data contains only data about limit orders, but the placement of market orders is displayed indirectly by the execution of one or several limit orders. Limit order executions with the same time stamp are probably due to the same market order, which we take into account when studying tick returns.

For each stock, the database for the entire year contains about one to four million messages. We exclude "hidden" limit orders from our analysis because our data base contains no information about their placement. We have checked that on the level of the average price impact function $I_{\mathrm{market}}(Q)$ in Eq. (4.1)

our results do not change if we include hidden limit orders, since the additional order volume is accounted for by the change in the normalization σ_Q.

This data set is from an electronic market place, where all orders are given and thus recorded via computers, so that we find no evidence for recording errors. We checked this visually by looking at the largest price changes, but also by consistency tests, e.g. verifying that the bid price is always lower than the ask price.

A.2 Programming methods

Except for the Monte Carlo simulations presented in chapter 6, which were performed using Octave, most programs for this thesis have been written in Perl. A Perl program is (usually) not compiled but uses an interpreter, and it is much slower than e.g. a C++ program (about a factor of two to ten, depending on the task). However, for most purposes in this thesis the programs run less than one hour, mostly only a few seconds, so that the speed of the program does not play a major role. An exception are the simulations of the trading strategies in chapter 5, which are very time consuming because of the complicated trading process simulated.

Usually much more time consuming than the actual execution of the program is the programming part, where Perl is a very useful language. On the one hand, it provides helpful tools for manipulating strings, so that processing the various data files is rather simple. On the other hand, it has easily accessible and adaptable data structures, that allow a fast storing and manipulating of the data. For instance, in the Island order book data, each order has an order reference number that can be used as a unique key to identify the order in a hash table, together with its data such as the limit price or the amount of shares. Though such data structures can also be created in other languages like C++, Perl provides them with a minimal amount of code. In this way, programs can be written quite fast and can be quickly adapted to a new analysis.

For the documentation of the programs, we used POD (Plain Old Documentation). This documentation is included in the program files and can be extracted to various formats like HTML, PDF, or a Unix man page using commands such as pod2html or pod2pdf. In this way, the documentation can be maintained in conjunction with the program. We developed a program that runs recursively through the file system of this thesis and automatically creates an HTML documentation of all program files and self-written modules that can be displayed in a browser.

The file system for this thesis is organized in four main folders:

- **data** – In this folder, all the data sets are stored in one subfolder for each time period studied.
- **scripts** – Here one finds all the programs. For each chapter of this thesis, the programs are stored in a single subfolder. Many functions that are used in more than one program are collected in modules that are stored in a subfolder named *Modules*. There is also a subfolder for tools that are used in more than one chapter, e.g. the ones for processing the data files in order to create the return time series from the raw data. In this subfolder, one also finds the program for creating the documentation.
- **results** – Results from calculations are saved in this folder, again with one subfolder for each chapter. Also the data for plots is saved here in subfolders called *data*.
- **plots** – Here, we save the plots that are created by the programs using gnuplot. For each postscript file, there is also a text file that contains the commands for gnuplot to create the plot. However, many plots are created with xmgrace using the data stored in the result path.

Bibliography

[1] R.J. Gordon, Does The 'New Economy' Measure Up To The Great Inventions Of The Past?, J. Economic Perspectives **14**, 49 (2000).

[2] S.R. Bond and J.G. Cummins, The Stock Market and Investment in the New Economy: Some Tangible Facts and Intangible Fictions, Brookings Papers on Economic Activity **1**, 61 (2000).

[3] D.W. Jorgenson and K.J. Stiroh, Raising the Speed Limit: US Economic Growth in the Information Age, Brookings Papers on Economic Activity **1**, 125 (2000).

[4] B.P. Bosworth, J.E. Triplett, What's New about the New Economy? IT, Economic Growth and Productivity, Intern. Productivity Monitor **2**, 19 (2001).

[5] J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing - From Statistical Physics to Risk Management*, (Cambridge University Press, Cambridge, UK, Second Edition, 2003).

[6] J.K. Galbraith, *The Great Crash, 1929*, (Houghton Mifflin Co., Boston, 1997).

[7] P. Perron, The great crash, the oil price shock, and the unit root hypothesis, Econometrica **57**, 1361 (1989).

[8] C.D. Romer, The Great Crash and the Onset of the Great Depression, Quart. J. Economics **105**, 729 (1990).

[9] D. Ben-David and D.H. Papell, The Great Wars, the Great Crash and Steady State Growth, J. Monetary Economics, **36**, 453 (1995).

[10] G.W. Schwert, Stock volatility and the crash of '87, Rev. Financial Studies **3**, 77 (1990).

[11] D. Sornette, *Why Stock Markets Crash: Critical Events in Complex Financial Systems*, (Princeton University Press, Princeton, NJ, 2003).

[12] J.Y. Campbell, A.W. Lo, and A.C. MacKinlay, The Econometrics of Financial Markets, (Princeton University Press, Princeton, NJ, 1997).

[13] R.N. Mantegna and H.E. Stanley, *An Introduction to Econophysics*, (Cambridge University Press, Cambridge, UK, 1999).

[14] B.B. Mandelbrot, The variation of certain speculative prices, J. Business **36**, 394 (1963).

[15] A. Pagan, The econometrics of financial markets, J. Empirical Finance **3**, 15 (1996).

[16] T. Lux, The Stable Paretian Hypothesis and the Frequency of Large Returns: An Examination of Major German Stocks, Appl. Financial Economics **6**, 463 (1996).

[17] P. Gopikrishnan, M. Meyer, L.A.N. Amaral, and H.E. Stanley, Inverse Cubic Law for the Probability Distribution of Stock Price Variations, European Phys. J. B **3**, 139 (1998).

[18] L.D. Landau and E.M. Lifshitz, *Statistical Physics Part 1*, vol. 5 of *Course of Theoretical Physics*, (Pergamon, 3rd edition, 1994).

[19] D. Sornette, *Critical Phenomena in Natural Sciences, Chaos, Fractals, Self-organization and Dusorder: Concepts and Tools*, (Springer Series in Synergetics, Heidelberg, 2000).

[20] J. Hasbrouck, Measuring the Information Content of Stock Trades, J. Finance **46**, 179 (1991).

[21] J. Hausmann, A. Lo, and C. MacKinlay, An Ordered Probit Analysis of Transaction Stock Prices, J. Financial Economics **31**, 319 (1992).

[22] A. Kempf and O. Korn, Market Depth and Order Size, J. Financial Markets **2**, 29 (1999).

[23] V. Plerou, P. Gopikrishnan, X. Gabaix, and H.E. Stanley, Quantifying Stock Price Response to Demand Fluctuations, Phys. Rev. E **66**, 027104[1] (2002).

[24] B. Rosenow, Fluctuations and Market Friction in Financial Trading, Int. J. Mod. Phys. C **13**, 419 (2002).

[25] M.D. Evans and R.K. Lyons, Order Flow and Exchange Rate Dynamics, J. Political Economy **110**, 170 (2002).

[26] F. Lillo, J.D. Farmer, and R.N. Mantegna, Master curve for price-impact function, Nature **421**, 129 (2003).

[27] X. Gabaix, P. Gopikrishnan, V. Plerou, and H.E. Stanley, A theory of power-law distributions in financial market fluctuations, Nature **423**, 267 (2003).

[28] M. Potters, J.-P. Bouchaud, More statistical properties of order books and price impact, Physica A **324**, 133 (2003).

[29] C. Hopman, Are supply and demand driving stock prices?, MIT working paper, (Dec. 2002).

[30] P. Weber and B. Rosenow, Order Book Approach to Price Impact, Quant. Finance **5**, 357 (2005).

[31] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, Fluctuations and response in financial markets: the subtle nature of 'random' price changes, Quant. Finance **4**, 176 (2004).

[32] L. Bachelier, Théorie de la spéculation, [Ph.D. thesis in mathematics], Annales Scientifiques de l'Ecole Normale Supérieure *III-17*, 21 (1900).

[33] F. Black and M. Scholes, The pricing of options and corporate liabilities, J. Political Economy, **3**, 637 (1973).

[34] R.N. Mantegna and H.E. Stanley, Scaling behaviour in the dynamics of an economic index, Nature **376**, 46 (1995).

[35] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, M. Meyer, and H.E. Stanley, Scaling of the distribution of price fluctuations of individual companies, Phys. Rev. E **60**, 6519 (1999).

[36] U.A. Muller, M.M. Dacorogna, and O.V. Pictet, "Heavy Tails in High-Frequency Financial Data," in *A Practical Guide to Heavy Tails*, edited by R.J. Adler, R.E. Feldman, and M.S. Taqqu (Birkhäuser, Boston, 1998), p. 83.

[37] E.F. Fama, Mandelbrot and the Stable Paretian Hypothesis, J. Business **36**, 420 (1963).

[38] E.F. Fama, The Behavior of Stock-Market Prices, J. Business **38**, 34 (1965).

[39] R.N. Mantegna and H.E. Stanley, Turbulence and financial markets, Nature **383**, 587 (1996).

[40] R.N. Mantegna and H.E. Stanley, Stock market dynamics and turbulence: parallel analysis of fluctuation phenomena, Physica A **239**, 255 (1997).

[41] F.M. Longin, The asymptotic distribution of extreme stock market returns, J. Business, **69**, 383 (1996).

[42] V. Pareto, *Cours d'économie politique*, reprinted as a volume of *Oeuvres Complètes* (Droz, Geneva, 1896-1965).

[43] P. Lévy, *Théorie de l'Addition des Variables Aléatoires*, (Gauthier-Villars, Paris, 1937).

[44] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, (Addison-Wesley, Cambridge, MA, 1954).

[45] P. Gopikrishnan, V. Plerou, L.A.N. Amaral, M. Meyer, and H.E. Stanley, Scaling of the distribution of fluctuations of financial market indices, Phys. Rev. E **60**, 5305 (1999).

[46] R.R. Officer, The distribution of stock returns, J. American Statistical Association **67**, 807 (1972).

[47] P.D. Praetz, The Distribution of Share Price Changes, J. Business **45**, 49 (1972).

[48] R.C. Blattberg and N. Gonedes, A Comparison of the Stable and Student Distributions as Statistical Models for Stock Prices, J. Business **47**, 244 (1974).

[49] M. Loretan and P.C.B. Phillips, Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets, J. Empirical Finance **1**, 211 (1994).

[50] P.K. Clark, A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices, Econometrica **41**, 135 (1973).

[51] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, Institutional Investors and Stock Market Volatility, Quart. J. Econ. **121**, 461 (2006).

[52] J.D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, What really causes large price changes, Quant. Finance **4**, 383 (2004).

[53] J.D. Farmer, F. Lillo, On the origin of power-law tails in price fluctuations, Quant. Finance **4**, C7 (2004); V. Plerou, P. Gopikrishnan, X. Gabaix, and H.E. Stanley, On the origin of power-law fluctuations in stock prices, Quant. Finance **4**, C11 (2004).

[54] R. A. Wood, T.H. McInish, and J.K. Ord, An investigation of transactions data for NYSE stocks, J. Finance **40**, 723 (1985).

[55] L. Harris, A Transactions Data Study of Weekly and Intradaily Patterns in Stock Prices, J. Financial Economics **16**, 99 (1986).

[56] A. Admati and P. Pfleiderer, A theory of intraday patterns: volume and price variability, Rev. Financial Studies **1**, 3 (1988).

[57] K. Chan, K.C. Chan, and G.A. Karolyi, Intraday Volatility in the Stock Index and Stock Index Futures Markets, Rev. Financial Studies **4**, 657 (1991).

[58] G.W. Schwert, Why Does Stock Volatility Change Over Time?, J. Finance **44**, 1115 (1989).

[59] T. Bollerslev, R.Y. Chou, and K.F. Kroner, ARCH modeling in finance: A review of the theory and empirical evidence, J. Econometr. **52**, 5 (1992).

[60] A.R. Gallant, P.E. Rossi, and G. Tauchen, Stock Prices and Volume, Rev. Financial Studies **5**, 199 (1992).

[61] B. Le Baron, Some Relations between Volatility and Serial Correlations in Stock Market Returns, J. Business **65**, 199 (1992).

[62] Z. Ding, C.W.J. Granger, and R.F. Engle, A Long Memory Property of Stock Market Returns and a New Model, J. Empirical Finance **1**, 83 (1993).

[63] M.M. Dacorogna, U.A. Muller, R.J. Nagler, R.B. Olsen, and O.V. Pictet, A geographical model for the daily and weekly seasonal volatility in the foreign exchange market, J. Int. Money Finance **12**, 413 (1993).

[64] C.W.J. Granger and Z. Ding, Varieties of long memory models, J. Econometr. **73**, 61 (1996).

[65] Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, Correlations in economic time series, Physica A **245**, 437 (1997); P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H.E. Stanley, Volatility Distribution in the S&P500 Stock Index, Physica A **245**, 441 (1997).

[66] Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, Statistical properties of the volatility of price fluctuations, Phys. Rev. E **60**, 1390 (1999).

[67] R. Cont, Scaling and correlation in financial data, [Ph.D. thesis], Universite de Paris XI, 1998 (unpublished); see also e-print cond-mat/9705075.

[68] M. Pasquini and M. Serva, Multiscale behaviour of volatility autocorrelations in a financial market, Econ. Lett. **65**, 275 (1999).

[69] V. Plerou, P. Gopikrishnan, X. Gabaix, L. A. N. Amaral, and H. E. Stanley, Price fluctuations, market activity, and trading volume, Quant. Finance **1**, 262 (2001).

[70] V. Plerou, P. Gopikrishnan, and H. E. Stanley, Quantifying fluctuations in market liquidity: Analysis of the bid-ask spread, Phys. Rev. E **71**, 046131 (2005).

[71] H. Markowitz, *Portfolio selection: Efficient Diversification of Investments*, (Wiley, New York, 1959).

[72] M.L. Mehta, *Random Matrices*, (Academic Press, San Diego, 1991).

[73] F. Lillo and J.D. Farmer, The long memory of the efficient market, Studies in Nonlinear Dynamics & Econometrics, **8**(3), Article 1 (2004).

[74] F. Lillo and R.N. Mantegna, Power law relaxation in a complex system: Omori Law After a Financial Market Crash, Phys. Rev. E **68**, 016119 (2003).

[75] F. Omori, On the aftershocks of earthquakes, J. Coll. Sci. Imp. Univ. Tokyo 7, 111 (1894).

[76] P. Weber, F. Wang, I. Vodenska-Chitkushev, S. Havlin, and H.E. Stanley, Relation between volatility correlations in financial markets and Omori processes occurring on all scales, Phys. Rev. E, in press.

[77] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, and H.E. Stanley, Scaling and memory in volatility return intervals in financial markets, Proc. Natl. Acad. Sci. **102**, 9424 (2005).

[78] F. Wang, K. Yamasaki, S. Havlin, and H.E. Stanley, Scaling and memory of intraday volatility return intervals in stock markets, Phys. Rev. E **73**, 026117 (2006).

[79] I. Vodenska-Chitkushev, F. Wang, P. Weber, K. Yamasaki, S. Havlin, and H.E. Stanley, (unpublished).

[80] F. Wang, P. Weber, K. Yamasaki, S. Havlin, and H.E. Stanley, Statistical Regularities in the Return Intervals of Volatility, Eur. Phys. J. B **55**, 123 (2007).

[81] V. N. Livina, S. Havlin, and A. Bunde, Memory in the occurrence of earthquakes, Phys. Rev. Lett. **95**, 208501 (2005).

[82] A. Bunde, J. F. Eichner, S. Havlin, and J. W. Kantelhardt, Return intervals of rare events in records with long-term persistence, Physica A **342**, 308 (2004).

[83] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin, Long-Term Memory: A Natural Mechanism for the Clustering of Extreme Events and Anomalous Residual Times in Climate Records, Phys. Rev. Lett. **94**, 048701 (2005).

[84] D. Sornette, Y. Malevergne, J.F. Muzy, What causes crashes?, Risk **16**, 67 (2003).

[85] A. Corral, Long-Term Clustering, Scaling, and Universality in the Temporal Occurrence of Earthquakes, Phys. Rev. Lett. **92**, 108501 (2004).

[86] A.G. Zawadowski, J. Kertesz, G. Andor, Large price changes on small scales, Physica A **344**, 221 (2004).

[87] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, Mosaic Organization of DNA Nucleotides, Phys. Rev.

E **49**, 1685 (1994); C.-K. Peng, S. Havlin, H.E. Stanley, and A.L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos **5**, 82 (1995).

[88] A. Bunde, S. Havlin, J.W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, Correlated and Uncorrelated Regions in Heart-Rate Fluctuations during Sleep, Phys. Rev. Lett. **85**, 3736 (2000).

[89] K. Hu, P. Ch. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley, Effect of trends on detrended fluctuation analysis, Phys. Rev. E **64**, 011114 (2001).

[90] L. Borland and J.-P. Bouchaud, On a multi-timescale statistical feedback model for volatility fluctuations, eprint physics/0507073.

[91] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, X. Gabaix, and H.E. Stanley, Economic fluctuations and anomalous diffusion, Phys. Rev. E **62**, R3023 (2000).

[92] C.M. Jones, G. Kaul, M.L. Lipson, Transactions, Volume, and Volatility, Rev. Financial Studies, **7**, 631 (1994).

[93] D. Easley, N.M. Kiefer, M. O'Hara, One day in the Life of a Very Common Stock, Rev. Financial Studies, **10**, 805 (1997).

[94] L. Gillemot, J.D. Farmer, and F. Lillo, There is more to volatility than volume, eprint physics/0510007.

[95] P. Weber, Analysis of aggregated tick returns: Evidence for anomalous diffusion, Phys. Rev. E, **75**, 016105 (2007).

[96] C.W.J. Granger and Z. Ding, Some properties of absolute returns, an alternative measure of risk, Annales d'Economie et de Statistique **40**, 67 (1995).

[97] Z. Ding and C.W.J. Granger, Modelling volatility persistence of speculative returns: a new approach, J. Econometrics **50**, 987 (1996).

[98] T.G. Andersen and T. Bollerslev, Heterogeneous information arrivals of return volatility dynamics: increasing the long run in high frequency returns, J. Finance **52**, 975 (1997).

[99] P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H.E. Stanley, Volatility Distribution in the S&P500 Stock Index, Physica A **245**, 441 (1997).

[100] We analyzed the following companies (ticker symbols): AMAT, BRCD, BRCM, CSCO, INTC, KLAC, MSFT, ORCL, QLGC, SEBL.

[101] S. Mike and J.D. Farmer, An empirical behavioral model of price formation, eprint physics/0509194.

[102] J.D. Farmer, A. Gerig, F. Lillo, and S. Mike, Maket efficiency and the long-memory of supply and demand: Is price impact variable and permanent or fixed and temporary, eprint physics/0602015.

[103] P. Weber and B. Rosenow, in *Proceedings of the Third Nikkei Econophysics Research Workshop and Symposium, The Fruits of Econophysics, Tokyo, November 2004*, edited by H. Takayasu, (Springer, Berlin, 2005), p. 88.

[104] P. Weber and B. Rosenow, Large stock price changes: volume or liquidity? Quant. Finance **6**, 7 (2006).

[105] C.M. Lee and M.J. Ready, Inferring Trade Direction from Intraday Data, J. Finance **46**, 733 (1991).

[106] P. Gopikrishnan, V. Plerou, X. Gabaix, and H.E. Stanley, Statistical properties of share volume traded in financial markets, Phys. Rev. E **62**, 4493 (2000).

[107] Y.-C. Zhang, Toward a theory of marginally efficient markets, Physica A **269**, 30 (1999).

[108] S. Maslov and M. Mills, Price fluctuations from the order book perspective – empirical facts and a simple model, Physica A **299**, 234 (2001).

[109] D. Challet and R. Stinchcombe, Analyzing and modelling 1+1d markets, Physica A **300**, 287 (2001).

[110] J.-P. Bouchaud, M. Mézard, and M. Potters, Statistical properties of stock order books: empirical results and models, Quant. Finance, **2**, 251 (2002).

[111] A.S. Kyle, Continuous Auctions and Insider Trading, Econometrica **53**, 1315 (1985).

[112] L.R. Glosten, Is the Electronic Open Limit Order Book Inevitable?, J. Finance **49**, 1127 (1994).

[113] A. Madhavan, M. Richardson, and M. Roomans, Why Do Security Prices Change? A Transaction-Level Analysis of NYSE Stocks, Rev. Financial Studies **10**, 1035 (1997).

[114] T. Chordia, R. Roll, and A. Subrahmanyam, Market Liquidity And Trading Activity, J. Finance **56 (2)**, 501 (2001).

[115] P. Sandas, Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market, Rev. Financial Studies **14**, 705 (2001).

[116] M.T. Coppejans, I.H. Domowitz, A. Madhavan, Liquidity in an Automated Auction, AFA 2002 Atlanta Meetings (December 21, 2001).

[117] H. Beltran, A. Durré, and P. Giot, How does liquidity react to stress periods on a limit order market?, NBB working paper No. 49 (May 2004).

[118] L. Glosten and L. Harris, Estimating the Components of the Bid-Ask Spread, J. Financial Economics **21**, 123 (1988).

[119] F. Foster and S. Viswanathan, Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models, J. Finance **48**, 187 (1993).

[120] M. Goldstein and K. Kavajecz, Trading strategies during circuit breakers and extreme market movements, J. Financial Markets **7**, 301 (2004).

[121] M.G. Daniels, J.D. Farmer, L. Gillemot, G. Iori, and E. Smith, Quantitative Model of Price Diffusion and Market Friction Based on Trading as a Mechanistic Random Process, Phys. Rev. Lett. **90**, 108102[1] (2003).

[122] Efficient Capital Markets: A Review of Theory and Empirical Work, J. Finance **25**, 383 (1970).

[123] M. Wyart, J.-P. Bouchaud, J. Kockelkoren, M. Potters, M. Vettorazzo, Relation between Bid-Ask Spread, Impact and Volatility in Double Auction Markets, eprint physics/0603084.

[124] L. Harris, *Trading and Exchanges. Market Microstructure for Practitioners*, (Oxford University Press, New York, 2003).

[125] E.L. Lehmann, *Testing Statistical Hypotheses*, (Springer, Berlin, Third Edition, 2005).

[126] M.G. Meyer and J.H.D. Jensen, *Elementary Theory of Nuclear Shell Structure*, (Wiley, New York, 1955).

[127] L.S. Kisslinger and R.A. Sorenson, Kgl. Danske Videnskab. Selskab. Mat.-fys. Medd. **32**, 9 (1960).

[128] Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise Dressing of Financial Correlation Matrices, Phys. Rev. Lett. **83**, 1467 (1999); Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters, Random Matrix Theory, Risk **12**, 69 (1999).

[129] Plerou, V., P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley, Universal and non-universal properties of cross-correlations in financial time series, Phys. Rev. Lett. **83**, 1471 (1999).

[130] S. John, Some optimal multivariate tests, Biometrika **58**, 123 (1971).

[131] H. Nagao, On some test criteria for covariance matrix, Ann. Statistics **1**, 700 (1973).

[132] O. Ledoit and M. Wolf, Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size, Ann. Statistics **30**, 1081 (2002).

[133] B. Rosenow and P. Weber, unpublished.

Danksagung

Mein großer Dank gilt allen, die mich in den letzten Jahren unterstützt haben. Insbesondere möchte ich mich bedanken bei meinem Betreuer Bernd Rosenow für seine Anleitung, Geduld und die vielen anregenden Diskussionen, die wir geführt haben, und bei denen ich sehr viel lernen durfte. Außerdem danke ich Prof. Stanley und Prof. Havlin für die schöne und erfolgreiche Zusammenarbeit während meiner Zeit in Boston, zu der auch meine dortigen Kollegen Irena und Fengzhong beigetragen haben. Die moralische Unterstützung meiner Freunde Michael, Alex und Sascha war mir sehr wichtig, und auch meiner Familie danke ich sehr für den Rückhalt, den sie mir gegeben hat. Mein Dank gilt auch Andreas Glatz, der mir in technischen Fragen der Textgestaltung half.

Ganz besonders möchte ich aber meiner Verlobten Alexandra danken, die in den letzten Jahren sehr viel mitmachen und aushalten musste, und mich trotzdem die ganze Zeit unterstützt hat.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation is von Herrn Priv.-Doz. Dr. Bernd Rosenow betreut worden.

- P. Weber, F. Wang, I. Vodenska-Chitkushev, S. Havlin, and H. E. Stanley, Relation between volatility correlations in financial markets and Omori processes occurring on all scales, Physical Review E, in press.
- F. Wang, P. Weber, K. Yamasaki, S. Havlin, and H. E. Stanley, Statistical Regularities in the Return Intervals of Volatility, European Physical Journal B **55**, 123 (2007).
- P. Weber, Analysis of aggregated tick returns: Evidence for anomalous diffusion, Physical Review E 75, 016105 (2007).
- P. Weber and B. Rosenow, Large stock price changes: volume or liquidity?, Quantitative Finance 6, 7 (2006).
- P. Weber and B. Rosenow, Order book dynamics and price impact, in *Proceedings of the Third Nikkei Econophysics Research Workshop and Symposium, The Fruits of Econophysics*, Tokyo, November 2004, edited by H. Takayasu (Springer, Berlin, 2005), p. 88.