BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

STATISTICAL PHYSICS APPROACHES TO FINANCIAL FLUCTUATIONS

by

FENGZHONG WANG

B.S., Nanjing University, 1998
M.S., Nanjing University, 2001

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2009

Approved by

First Reader ____

H. Eugene Stanley, Ph.D.
University Professor and Professor of Physics

Second Reader ____

William Skocpol Ph.D.
Professor of Physics

ACKNOWLEDGMENTS

I would like to acknowledge everybody who has helped and inspired me during these years in Boston. First, I express my sincere gratitude to my wife, Weiwei, for her support, patience and encouragement. She helped me through the hardest times and taught me to be always optimistic. To her I dedicate my thesis.

I thank my thesis advisor, Professor H. Eugene Stanley, for making me part of his lively group and for his support, encouragement and guidance during the last six years. I am truly grateful to him for sharing with me his ingenious views and knowledge, for his inspiring energy, contagious enthusiasm and generosity.

I am thankful to Professor Shlomo Havlin, who is practically my second advisor, for initiating my thesis study and sharing his deep insights and knowledge in many scientific fields. I feel lucky and honored to have worked with him and learned from him.

I wish to thank Professor Kazuko Yamasaki, Jan Nagler, Shwu-Jane Shieh, Boris Podobnik, Fredrik Liljeros, and Plamen Ch. Ivanov for interesting and stimulating discussions. In particular, I thank them for teaching me everything I know on economics and finance making me aware of a new and exciting field of research on the front line between statistical physics and economics.

I am especially grateful to Professor William Skocpol for his magnificent help and critical comments on this dissertation. I also thank Dr. Woo-sung Jung, Dr. Philipp Weber, Dr. Dongfeng Fu, Dr. Matia Kaushik, Dr. Aaron Schweiger, Mrs. Irena Vodenska-Chitkushev, Mr. Alexander Petersen, Mr. Xuqing Huang, Mr. Chen Liu, Mr. Duan Wang, Miss Qian Li and Mr. Wei Li for interesting and stimulating discussions about science.

I must acknowledge all the friends who made my experience in Boston university fun and exciting. To mention a few, Zhenyu Yan, Alfonso Lam, Jose Borreguero, Zhonghua Ma, Yeming Wang, Jun Zhou, Xiang Liu, Daoxin Yao, Jia Shao, Sungho

Han, Ling Wang, Jiayuan Luo, Yongsheng Liu and many others.

I would like to acknowledge my thesis committee, Professor H. Eugene Stanley, Professor William Skocpol, Professor Robert Carey, Professor Karl Ludwig, Professor Shyamsunder Erramilli and Professor John Butler for their patience and constant help. I also thank Professor Claudio Rebbi, Professor James Stone, and Professor Bennett Goldberg for reading the prospectus and the abstract of this dissertation. I would like to specially than, Bob Tomposki, Jerry Morrow, Guoan Hu, and Mirtha Cabello for their help on this dissertation.

STATISTICAL PHYSICS APPROACHES
TO FINANCIAL FLUCTUATIONS

(Order No.)

FENGZHONG WANG
Boston University Graduate School of Arts and Sciences, 2009
Major Professor: H. Eugene Stanley, University Professor and Professor of Physics

ABSTRACT

Complex systems attract many researchers from various scientific fields. Financial markets are one of these widely studied complex systems. Statistical physics, which was originally developed to study large systems, provides novel ideas and powerful methods to analyze financial markets. The study of financial fluctuations characterizes market behavior, and helps to better understand the underlying market mechanism.

Our study focuses on volatility, a fundamental quantity to characterize financial fluctuations. We examine equity data of the entire U.S. stock market during 2001 and 2002. To analyze the volatility time series, we develop a new approach, called *return interval analysis*, which examines the time intervals between two successive volatilities exceeding a given value threshold. We find that the return interval distribution displays scaling over a wide range of thresholds. This scaling is valid for a range of time windows, from one minute up to one day. Moreover, our results are similar for commodities, interest rates, currencies, and for stocks of different countries. Further analysis shows some systematic deviations from a scaling law, which we can attribute to nonlinear correlations in the volatility time series. We also find a memory effect in return intervals for different time scales, which is related to the long-term correlations in the volatility.

To further characterize the mechanism of price movement, we simulate the volatility time series using two different models, fractionally integrated generalized autoregressive conditional heteroscedasticity (FIGARCH) and fractional Brownian motion (fBm), and test these models with the return interval analysis. We find that both models can mimic time memory but only fBm shows scaling in the return interval distribution.

In addition, we examine the volatility of daily opening to closing and of closing to opening. We find that each volatility distribution has a power law tail. Using the detrended fluctuation analysis (DFA) method, we show long-term auto-correlations in these volatility time series. We also analyze return, the actual price changes of stocks, and find that the returns over the two sessions are often anti-correlated.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMEX	American Stock Exchange
ARCH	AutoRegressive Conditional Heteroscedasticity
CDF	Cumulative Distribution Function
DFA	Detrended Fluctuation Analysis
DJIA	Dow Jones Industrial Average index
fBm	fractional Brownian motion
FIGARCH	Fractional Integrated GARCH
GARCH	Generalized ARCH
i.i.d.	independent and identically distributed
KS	Kolmogorov-Smirnov
MLE	Maximum Likelihood Estimator
NASDAQ	National Association of Securities Dealers Automated Quotations
NYSE	New York Stock Exchange
PDF	Probability Density Function
PL	Power Law
SE	Stretched Exponential
S&P 500	Standard & Poor's 500 index
TAQ	Trades And Quotes

Chapter 1

Introduction

1.1 Motivation of Thesis

Financial markets, which facilitate the trading of huge amounts of financial assets in a competitive global environment, are of great importance for the economy over the whole world. The ups and downs in the market not only affect traders and investors, but also influence the life of all of us. For instance, the present credit crisis accompanies with turmoil in the financial markets, and causes huge losses for many investors, initiating a worldwide recession. Moreover, significant risk can be inherent not only in market crashes, but also in less hazardous fluctuations if they are unexpected and investments are not well protected against them. Banks have to properly estimate the risk of their investments and make provisions in order to be able to withstand large fluctuations without going bankrupt. To reduce the risk ("hedging") or bet the chance ("speculation") of investment, we need to understand the underlying mechanisms of financial fluctuations, which are still an open question.

From a physicist's point of view, financial markets are a typical complex system, since a large number of participants with divergent anticipations and conflicting interests are simultaneously present in the markets. Moreover, financial markets are affected by external news, which is unexpected to a large degree for both in time and in nature. On the other hand, the stochastic nature of a system arises from its complexity, from the fact that causes are diverse, that tiny perturbations might result in large effects. Hence, financial markets offer ideal systems for studying complexity and understanding stochastic behaviors.

Statistical physics deals with systems comprising a very large number of interacting subunits, for which predicting the exact behavior of the individual subunit would be impossible. Hence, one is limited to making statistical predictions regarding the collective behavior of the subunits. Recently, it has come to be appreciated that many such systems consisting of a large number of interacting subunits obey universal laws, independent of the microscopic details. The finding, in physical systems, of universal properties that do not depend on the specific form of the interactions gives rise to the intriguing hypothesis that universal laws or behavior may also be present in economic and social systems [1–5].

An often-expressed concern regarding the application of physics methods to the social sciences is that physical laws are applied to systems with a very large number of subunits (at the order of Avogadro's number, 10^{23}), while social systems comprise a much smaller number of elements. Fortunately, due to the rapid development of electronic trading and data storing in the last few decades, financial data bases have become available with a huge number of data points (say 10^{8}, see Sec. 1.2 for some examples), enabling physicists to analyze them as dynamic systems. The data size becomes comparable to physical nano systems and the "thermodynamic limit" is reached so that methods from statistical physics can be applied. It is worth noting that there are only a small number of extremely large events even in very huge data bases. To understand these devastating events, it is of great importance to find laws describing the entire data set in order to approach extreme events by extensive analysis on small fluctuations.

Two important conceptual advances towards universal laws are *scaling* and *universality*. A system obeys a scaling law if its relation is characterized by the same functional form and exponent over a certain range of scales ("scale invariance"). The typical behavior for scaling is *data collapse*, all curves can be "collapsed" onto a single curve, after a certain scale transformation on the measure. The general principles of scale invariance used here have proved useful in interpreting a number of other phenomena, ranging from elementary particle physics and galaxy structure to finance [5–7]. At one time, many imagined that the "scale-free" phenomena are relevant only to a fairly narrow slice of physical phenomena [8, 9]. However, the range of systems that apparently display power law and scale-invariant correlations has increased dramatically in recent years, ranging from base pair correlations in noncoding DNA [10], lung inflation [11] and interbeat intervals of the human heart [12] to complex systems involving large numbers of interacting subunits that display "free will," such as city growth [13], university research budgets [14], and even bird populations [15]. In many of these diverse systems, the *same* scaling function exists for a significant range which is remarkable, apparently suggesting the universality of laws. Moreover, many systems share the same scaling functions and characteristic exponents and therefore belong to one universality class. This connection provides the people a comprehensive view over these diverse systems.

Scaling and universality are important properties of a data set describing the global behavior of the probability distribution. This usually does not fully characterize a sequence of data points, which also depends on the time organization of the sequence. If it is *uncorrelated*, the data points are independent of each other and the sequence is totally determined by the distribution. However, in most cases, the records are *correlated*, and affect the order in the data set. This behavior is called "memory", as the data points "remember" previous values. Trivially the memory decays with the time lag. The decay of memory, which could be characterized by the auto-correlation function, may follow different types of function. One typical function is exponential, and the existence of memory is described by a characteristic time scale. The memory almost disappears at the scales above the characteristic time and thus it only exists for a short-term. This kind of time series is called *short-term correlated*.

Another typical function for the auto-correlation is a power law. In this case there is no finite characteristic scale and the correlation exists for a much longer time. Therefore it is called *long-term correlated*. Note that short-term memory always exists in a long-term correlated time series. As for the study of financial markets, the temporal structure in a time series is of great importance since it influences the performance of any movement. Many studies show that price change ("return") does not exhibit any linear correlations extending over more than a couple of minutes, but their absolute value, which is a measure of volatility, exhibits long-term correlations (see Ref. [1–4] and references therein). This leads to long periods of high volatility as well as other periods where the volatility is low ("volatility clustering").

Extreme events do not only occur in economics, but also appear in very different fields like climate or earthquakes. For instance, Gutenberg and Richter related huge earthquakes to everyday tremors in one single power law curve [16, 17]. If one wants to prepare for a dangerous earthquake, it might be less important to exactly know how strong the next shock will be, but rather to know when a large shock will occur. A good approach is to study the time ("return interval") between two successive shocks larger than a threshold above which a shock would damage a building [18–37]. This way one can gather information on the temporal structure of the fluctuations. Recently Bunde et al. [18–22] studied the return intervals for climate records and found that the long-term memory leads to a stretched exponential distribution and clustering of extreme events. Analogous to earthquakes, we analyzed the return interval for volatility of financial data, and found striking scaling in the distribution and long-term memory in the time series [24–32].

1.2 Database Analyzed

One important reason that allows statistical physicists to explicitly study the financial markets is the availability of huge amounts of financial data, due to the rapid development of technology in past few decades. Below we list financial data sets that we analyzed, from stocks, indexes, currencies, and interest rates, to commodities. Note that some databases are high-frequency and the sampling interval could be as short as 1 second, others are recorded on a daily basis and the whole period could be as long as 82 years.

- Trades and Quotes (TAQ).

  This high-frequency database is for a 2-year period, from January 2, 2001 to December 31, 2002, totaling 500 trading days. It records *all* transactions ("tick") of *all* equities traded in the three major US stock exchanges, namely, (a) the New York Stock Exchange (NYSE), (b) the American Stock Exchange (AMEX), and (c) the National Association of Securities Dealers Automated Quotation (NASDAQ). In total more than 1.2×10^4 stocks and 1.8×10^9 records are included. From this database we extract the 500 component stocks for the Standard & Poor's 500 index (S&P 500) and 30 component stocks of the Dow Jones Industrial Average index (DJIA). For a typical stock from these two indices, the data set has 2.0×10^5 points with a sampling interval of 1 min. This database is analyzed in most of sections.

- S&P 500 index.

  This intraday index data set covers a 13-year period, from January 2, 1984 to December 31, 1996, with a sampling interval of 10 min. The S&P 500 index, which consists of 500 companies, is a benchmark of the U.S. stock market. Totally, 1.3×10^5 data points are included.

- Yahoo! finance.

  This historical daily database is obtained from website `http://finance.yahoo.com`. It includes more than 1.9×10^4 U.S, stocks from NYSE, AMEX, and NASDAQ markets, and covers a 46-year period, from January 2, 1962, to December 31, 2007. In total this database has 3.2×10^7 data points. For a typical stock such as General Electric (GE), the data size is 1.1×10^4 points. This database is analyzed in Sec. 2.3.

- Center for Research in Security Prices (CRSP).

  This is also a historical daily database. It records *all* U.S. equities for a 82-year period, from January 2, 1926 to December 31, 2007. In total, more than 2.5×10^4 stocks from NYSE, AMEX, and NASDAQ markets are included. The overall data size is 6.5×10^7 points. For the stock of GE, the data size is 2.2×10^4 points.

- Currency and interest rate.

  We collected daily exchange rates of 35 other currencies to United States Dollar (USD) and federal funds rate from website `http://www.federalreserve.gov`. The beginning and ending dates are different for those records. A typical exchange rate, USD vs. Japanese Yen (JPY), starts from January 4, 1971 to December 31, 2007. The USD/JPY rate has 1.0×10^4 points for the 37-year period. Federal funds rate starts from July 1, 1954 to December 31, 2007, which has 1.4×10^4 data points for the 54-year period.

- Oil and gold commodity.

  We also collected the daily spot price of west Texas intermediate (WTI) crude oil from website `http://www.eia.doe.gov` and the daily gold price (London P.M.) from website `http://www.onlygold.com`. The oil price starts from December 30, 1985 to December 31, 2007, totally 5.0×10^3 points for the 22-year period. The interval for gold prices is from January 2, 1985 to December 31, 2007, and the number of data points is 5.0×10^3 for the 23-year period.

1.3 Notations and Measurements of Fluctuation Fluctuations

1.3.1 Notations

In this subsection we describe those quantities that are widely-used in this dissertation. In the following x_i and y_i are two data sets, $i = 1, \ldots N$, and N is the size of the data set.

- Mean value

$$\langle x \rangle \equiv \sum_{i=1}^{N} x_i / N \tag{1.1}$$

  for variable x. $\langle \cdot \rangle$ stands for the average.

- Standard deviation

$$\sigma(x) \equiv \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \tag{1.2}$$

  for variable x. σ denotes the standard deviation and σ^2 is the corresponding variance.

- Probability density function (PDF) is denoted as $P(x)$ for variable x.

- Conditional distribution is denoted as $P(x|x_0)$, which measures the distribution conditional on that the previous value belongs to subset x_0.

- Cumulative distribution function (CDF) is denoted as $Q(x)$ for variable x, which is the integral of $P(x)$, i.e.,

$$Q(x) \equiv \int_x^\infty P(x)\, dx. \tag{1.3}$$

  Note that Eq. (1.3) precisely is for the complementary cumulative distribution function. In this dissertation we call it the cumulative distribution for simplicity.

- Auto-correlation function (ACF)

$$C(\Delta t) \equiv \frac{\langle x_i\, x_{i+\Delta t} \rangle - \langle x \rangle^2}{\sigma(x)^2} \tag{1.4}$$

  for variable x. Here Δt is the time lag.

- Cross-correlation (also called Pearson coefficient)

$$C(x,y) \equiv \frac{\langle x\, y\rangle - \langle x\rangle \langle y\rangle}{\sigma(x)\, \sigma(y)}. \tag{1.5}$$

  between variable x and y.

- Cross-correlation with time lag Δt

$$C_{\Delta t}(x,y) \equiv \frac{\langle x_i\, y_{i+\Delta t}\rangle - \langle x\rangle \langle y\rangle}{\sigma(x)\, \sigma(y)} \tag{1.6}$$

  between two variable x and y. Note that Eq. (1.6) is the generalization of Eq. (1.5). When $\Delta t = 0$, Eq. (1.6) reduces to Eq. (1.5).

1.3.2 Definitions of Return and Volatility

One basic quantity to measure financial fluctuations is return R, which is defined as the logarithmic price change,

$$R(t) \equiv \ln\left(\frac{S(t+\Delta t)}{S(t)}\right) \approx \frac{S(t+\Delta t) - S(t)}{S(t)}, \tag{1.7}$$

where $S(t)$ the price of stock or other financial assets at time t, and Δt is the time step. For small changes in S, the return is approximately the forward relative price change, as shown in Eq. (1.7).

Another basic quantity for financial fluctuations is volatility R, which characterizes the magnitude of fluctuations. For this purpose, volatility R is usually defined as the absolute value of return, i.e.,

$$V(t) \equiv |R(t)| \tag{1.8}$$

Note that the volatility is defined as the standard deviation of returns in many literature. However, Eq. (1.8) has one data point of volatility for every point of return, which will give us a time series of volatility, not a few points of volatility, to analyze. Therefore, in this dissertation we choose Eq. (1.8) as the volatility definition.

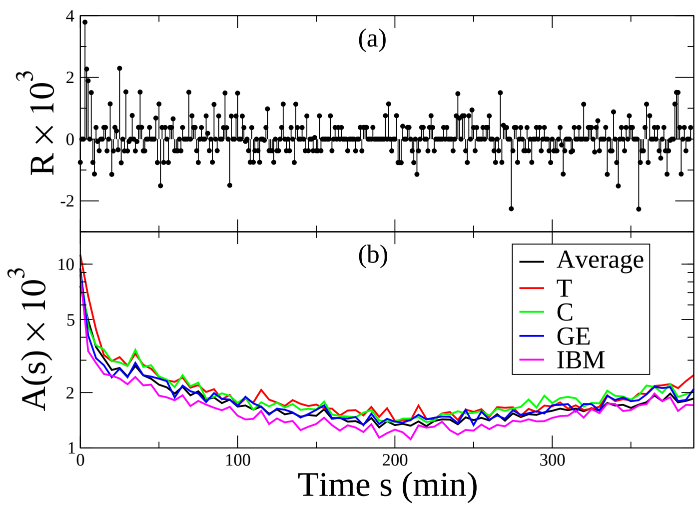

Figure 1.1: (a) Illustration of return R time series. The data points are for typical stock GE on October 18, 2002. (b) Intraday pattern $A(s)$ for stock T, C, GE, IBM, and the average over 30 DJIA stocks. The time s is the moment in one trading day, while $A(s)$ is the mean volatility over all trading days. Note that all curves have a similar pattern, such as a pronounced peak after the market opens and a minimum around noon ($s \simeq 200$ min).

The above definitions are widely-used for the daily equity data. However, the high-frequency data is known to have intraday patterns [26], due to different behaviors of traders at different periods during the trading day. For example, the market is very active immediately after the opening [38], due to information arriving while the market is closed. For example, we can see the return value is relative higher at the beginning, compared to other moments in that trading day [Fig. 1.1(a)]. To understand this tendency, we investigate the daily trend in the 30 component stocks of the DJIA index. A direct way to measure the intraday pattern is the averaged volatility, denoted as $A(s)$, which is defined as

$$A(s) \equiv \frac{\sum_{i=1}^{N} V^i(s)}{N}, \tag{1.9}$$

which is the average at a specific moment s of the day averaged over all N trading days, and $V^i(s)$ is the price change at time s in day i. As shown in Fig. 1(b), the intraday pattern $A(s)$ has similar "U" or "L" shape for the four stocks, namely, AT&T (T), Citigroup (C), General Electric (GE), International Business Machine (IBM), and the average over 30 DJIA stocks. The pattern is not uniformly distributed, exhibiting a pronounced peak at the opening hours and a minimum around $s = 200$ min, that may cause some artificial correlations. To avoid the effect of this daily oscillation, we remove the intraday pattern by studying

$$V'(t) \equiv V(t)/A(s). \tag{1.10}$$

In order to compare different stocks, we define the normalized volatility $v(t)$ by dividing $V'(t)$ by its standard deviation,

$$v(t) \equiv \frac{V'(t)}{\sigma(V')}, \tag{1.11}$$

Consequently, the normalized volatility v is in units of standard deviations.

In summary, for the daily data, we use V in Eq. (1.8) as the volatility definition; for the intraday data, we use v in Eq. (1.11) as the definition but call normalized volatility as volatility for simplicity.

1.3.3 Definition of Return Interval

In this dissertation we develop a new approach to analyze the volatility time series. As we known, the volatility has a power-law tail in the distribution which means their values are distributed in a wide range. However, compared to small volatilities, these large volatilities are usually more important since they corresponds to the big movement of the equity, implying a big gain or risk. To analyze them, we can choose a volatility value as threshold q and only pick up volatilities that are above q. According to this threshold, all these selected volatilities are large enough to be counted. Therefore we call them "events". From them we obtain the *return* *interval* , which is the time interval between consecutive events. By this definition, we know return intervals are strongly related to the temporal structure in volatilities. In other words, we can understand correlations in volatilities much further by studying their return intervals. We also notice that there is only one free parameter for return intervals, the threshold q. We use different q to examine events of different sizes. Note that the volatility (as defined in Eq. (1.11)) is in units of standard deviation for the intraday stock data. Thus the threshold q is also in units of the volatility standard deviation in the following. As an example, we show the generation of return intervals from volatilities of stock GE in Fig. 1.2.

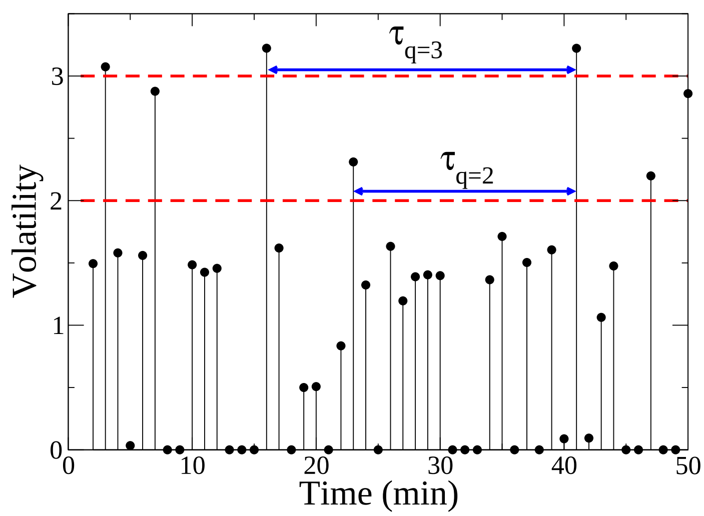

Figure 1.2: Illustration of volatility return intervals. The volatility is in units of its standard deviation. The solid circles are volatility values of the GE stock on January 8, 2001. Return intervals $\tau_{q=2}$ and $\tau_{q=3}$ for two typical thresholds q are displayed.

1.4 Organization of Thesis

The following part of this thesis is organized as below.

In Chapter 2, we first review the properties of returns and volatilities. Both returns and volatilities have power-law tails in the distribution. However, only volatilities have long-term correlations in the time series. Then we perform a comprehensive study of the two components of daily return and volatility–one covers the overnight session and another covers the daytime sessions. We show that both of them have similar features as the total one. Moreover, we find that the daytime component contributes more to the total. Further, we find that a significant anti-correlation exists between two the component returns.

In Chapter 3, we focus on the distribution of return intervals. A well-approximated scaling behavior is found in these distributions, and the scaling function is found to be well-approximated to the stretched exponential function. Remarkably, the scaling is persistent for various stocks, markets, thresholds, and sampling intervals. We also find that the stretched exponential scaling function is due to long-term correlations in volatilities.

In Chapter 4, we further analyze the distribution of return intervals and find a certain tendency to deviate from a pure scaling law. We study the moment of return intervals and suggest a new measurement for multiscaling. Moreover, we find that the multiscaling nature of return intervals is due to the non-linearity in volatility correlations. Further, we study the relation between multiscaling and several essential factors, which offers deep insights for understanding volatilities.

In Chapter 5, we study another interesting feature in return intervals, the memory effect in three types of time scales, short-range, medium-range, and long-range. We show a strong memory effects at all these three scales. We also find strong connection between memory effects in return intervals and correlations in volatilities.

In Chapter 6, we simulate volatilities and therefore return intervals with two classic models, FIGARCH and fractional Brownian motion (fBm). We find that both models exhibit memory effects in return intervals. However, only fBm shows scaling in return interval distributions. Moreover, we propose a new method to estimate the risk, based on the striking scaling and memory features of return intervals.

In the last, we describe the detrended fluctuation analysis (DFA) method in Appendix A, two classic methods to fit power-law functions and a classic method for goodness-of-fit in Appendix B.

Chapter 2

Returns and Volatilities

One basic measurement for changes in security prices, foreign exchange rates or other market quantities is *return*, giving the relative price change in a time interval. Returns show the speed and direction of the market movement. For instance, if most returns are positive during a few years, the stock market is called to be a "bull market". On the contrary, it is called a "bear market" if we observe mainly negative returns. The fluctuations in returns reflect the common behavior for financial markets, providing upside opportunities as well as downside risk to traders and investors. To characterize the market, *volatility* is introduced as another fundamental concept. It describes the magnitude of the market fluctuations, irrespective of the direction. In this chapter first we review the properties of returns and volatilities, then we study two special types of returns and volatilities, which cover the daytime period and the overnight period respectively.

2.1 Properties of Returns

2.1.1 Leptokurtic Distribution and Power-Law Tail

The nature of the distribution of price fluctuations in financial time series has been a topic of interest for over 100 years [39]. A reasonable *a priori* assumption, motivated by the central limit theorem, is that the returns are independent and identically distributed (*i.i.d.*) random variables, which results in a Gaussian random walk in the logarithm of price.

However, empirical studies [5, 6, 14, 40–48] show that the distribution of returns has pronounced tails, in striking contrast to that of a Gaussian distribution. In addition to being non-Gaussian, the process of returns shows another interesting property: "time scaling"—that is, the distributions of returns for various choices of Δt, ranging from one day up to even one year have similar functional forms [5]. These results together would suggest that the distribution of returns is consistent with a Lévy stable distribution [5, 45, 49–51], the rationale for which arises from a generalization of the central limit theorem to random variables which do not have a finite second moment. This type of distribution is referred to as leptokurtic, meaning that the distribution has a higher peak, lower shoulders, and fatter tails.

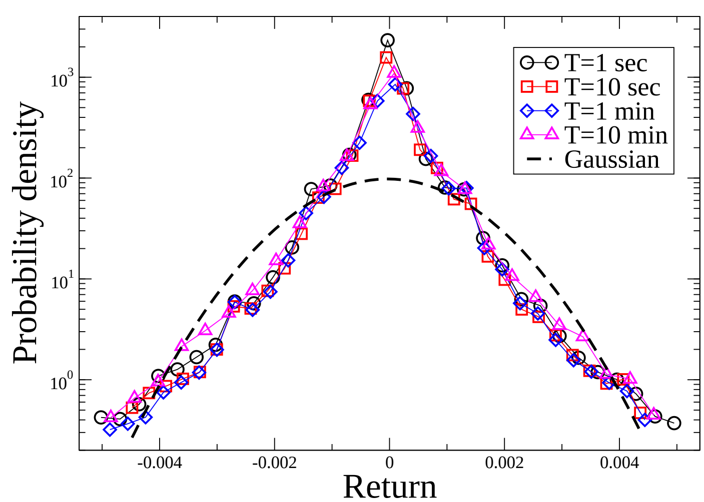

Figure 2.1: Distribution of returns for a typical stock, namely, General Electric (GE). Four sampling times, $T = 1$ sec, 10 sec, 1 min, and 10 min are plotted. Remarkably, all these distributions have leptokurtic shape which almost collapse onto a single one, indicating scaling in returns. As a reference, we fit the distribution with Gaussian function, as shown by the dashed line.

As an example, we plot the distribution of returns for stock GE with four time resolutions in Fig. 2.1. We can see that all distributions are leptokurtic, which clearly deviate from the Gaussian distribution. Moreover, in Fig. 2.1 we find that all curves almost collapse onto a single one, suggesting the time scaling.

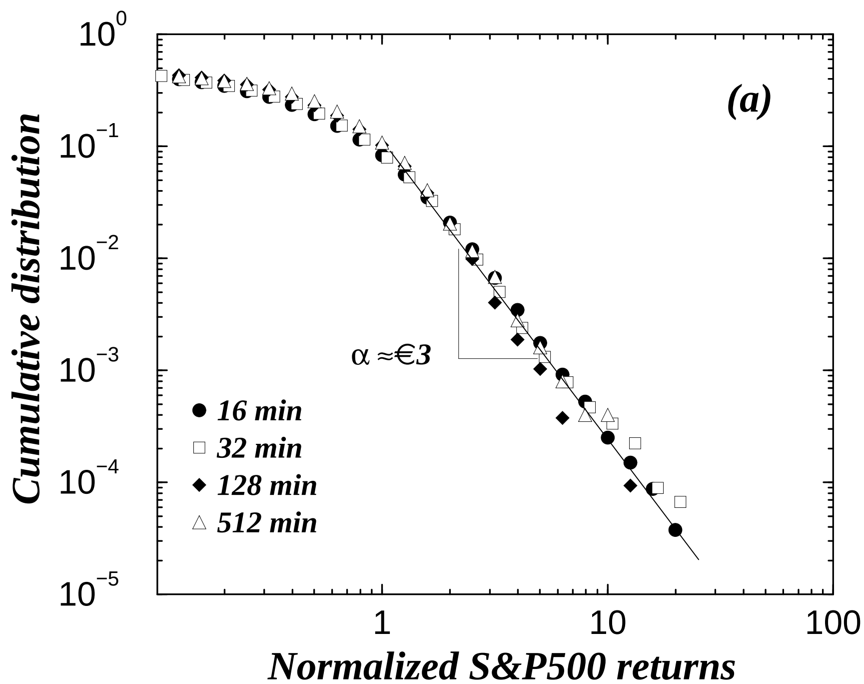

Figure 2.2: Cumulative distribution of normalized returns of the positive tails for $\Delta t = 16$, 32, 128, and 512 min. Power-law regression fits yield estimates of the tail exponent $\zeta = 2.69 \pm 0.04$, 2.53 ± 0.06, 2.83 ± 0.18 and 3.39 ± 0.03 for $\Delta t = 16$, 32, 128 and 512 min, respectively. (From Gopikrishnan et al. [52])

One of the most widely-used ways to analyze the tail of a distribution is the cumulative distribution function (CDF), the probability of a return larger than or equal to a threshold. Previous empirical studies found that the asymptotic behavior of the functional form of the cumulative distribution for returns R is consistent with a power-law,

$$P(R > x) \sim x^{-\zeta}\,, \tag{2.1}$$

where ζ, called "tail exponent", is the one to characterize power-law decay in the tail distribution. After normalizing the returns with their standard deviations in the 2-year period, which makes returns of different stocks comparable, Gopikrishnan et al. analyzed the 1000 U.S. stocks for both positive and negative tails of the distribution, as plotted in Fig. 2.2. They estimated the exponent ζ by a power-law regression and obtained the average ζ value, $\langle\zeta\rangle \simeq 3$, for all four time intervals, $\Delta t = 16$, 32, 128, and 256 min, supports the time scaling in returns. Similar results are found in the analysis of daily returns of 30 German stocks composing the DAX index [40], foreign exchange rates [43], and daily CRSP returns [42].

2.1.2 Correlations in Returns

In addition to the probability distribution, a complementary aspect for the characterization of any stochastic process is the quantification of correlations. A traditional way to measure the correlation is the auto-correlation function [Eq. (1.4)] Previous works found that the auto-correlation function of returns has an exponential decay with characteristic decay times τ of only 4 min [53, 54]. As is clear from Fig. 2.3, for time scales beyond 20 min the correlation is at the level of noise, in agreement with the *efficient market hypothesis* which states that is not possible to predict future stock prices from their previous values [55]. If price-correlations were not short-range, one could devise a way to make money from the market indefinitely.

It is important to note that lack of linear correlation does not imply an *i.i.d.* process for the returns, since there may exist higher-order correlations. Indeed, the amplitude of the returns, referred to in economics as the *volatility*, shows long-range time correlations that persist up to several months [53–60], and are characterized by an asymptotic power-law decay. This feature will be further discussed in the next section.

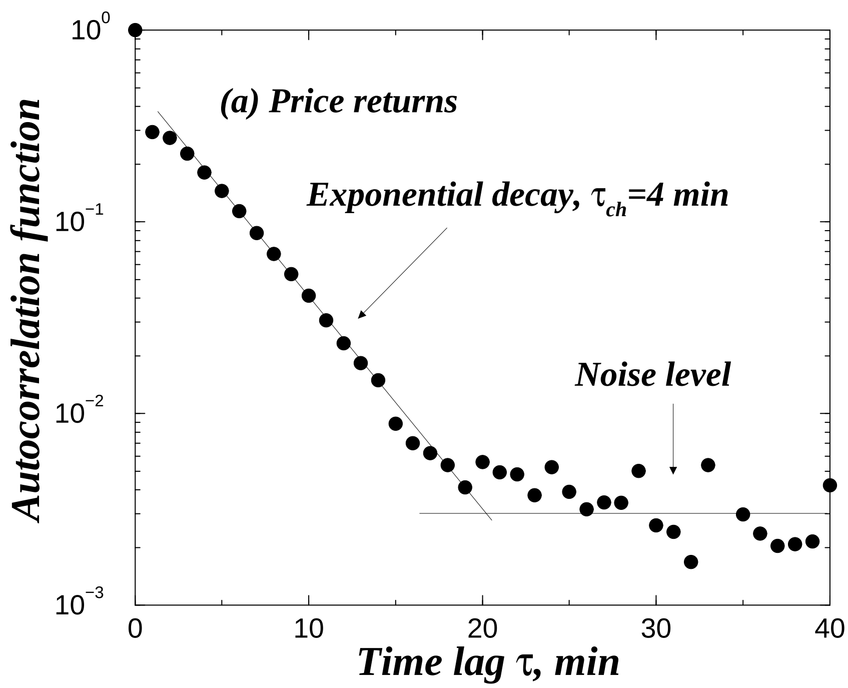

Figure 2.3: Auto-correlation function for the S&P 500 index returns sampled at a $\Delta t = 1$ min. The straight line corresponds to an exponential decay with a characteristic decay time $\tau = 4$ min. Note that after 20 min the correlations are at the noise level. (From Gopikrishnan et al. [52])

2.2 Properties of Volatilities

2.2.1 Log-Normal Distribution and Power-Law Tail

The volatility also has been extensively studied. Analogous to returns, we examine the distribution of volatilities in two regimes, center and tail. First we analyze the center regime of the distribution, which was found to be log-normal. To test this possibility, we rescale the volatility and plot its distribution In Fig. 2.4. The distributions of volatility for various choices of sampling interval T (from $T = 120$ min up to $T = 900$ min), collapse onto one curve and are well fit in the center by a quadratic function on a log-log scale. Since the central limit theorem holds also for correlated series [61], with a slower convergence than for non-correlated processes [59], in the limit of large values of T, one expects that $P(V)$ becomes Gaussian. However, a log-normal distribution fits the data better than a Gaussian. The apparent scaling behavior of volatility distribution could be attributed to the long persistence of its auto-correlation function [59].

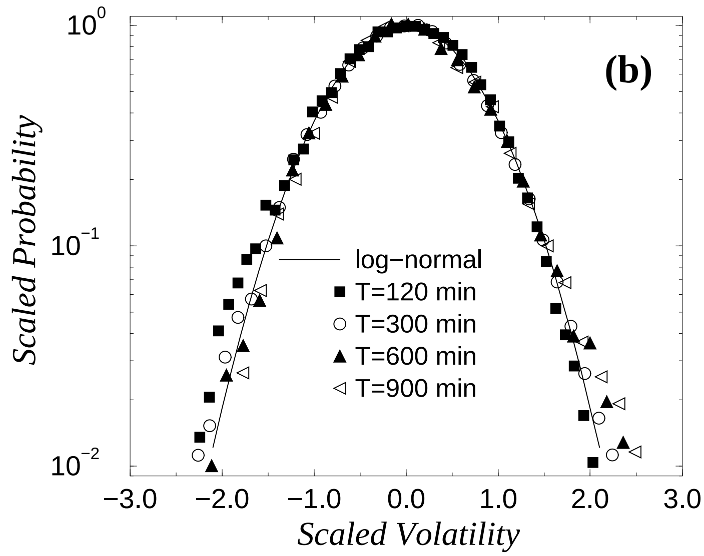

Figure 2.4: Center of volatility distribution for S&P 500 index: The volatility distribution for different window sizes T using the log-normal scaling form, $\sqrt{\nu}\,\exp(a+\nu/4)\,P(V)$ as a function of $(\ln(V)-a)/\sqrt{\pi\nu}$, where a and ν are the mean and the width on a logarithmic scale. By the scaling, all curves collapse to the log-normal form with $a=0$ and $\nu=1$, $\exp(-(\ln x)^2)$ (solid line). (From Liu et al. [60])

Although the log-normal function seems to describe well the center part of the volatility distribution, the distribution of the volatility might have quite different behavior in the tail, by the definition of volatility. Since our time window T for estimating volatility is quite large, it is difficult to obtain significant statistics for the tail. Previous studies of the distribution for price changes report power law asymptotic behavior [40–42]. From the relation in the definitions of return and volatility, it is possible that a similar power law asymptotic behavior might characterize the distribution of the volatility. Similarly, we compute the cumulative distribution , and find a consistent power law asymptotic behavior,

$$P(V > x) \sim x^{-\zeta} \,. \tag{2.2}$$

Regression fits yield estimates $\zeta = 3.10 \pm 0.08$ for $T = 32$ min, well outside the stable Lévy range $0 < \zeta < 2$. For larger time scales the asymptotic behavior is difficult to estimate because of poor statistics at the tails. In view of the power law asymptotic behavior for the volatility distribution, the drop-off of $P(V)$ for low values of the volatility could be regarded as a truncation to the power law behavior, as opposed to a log-normal. For a systematic study of the PDF dependence on company size, see [62–65], and references therein.

2.2.2 Long-Term Correlations in Volatilities

Numerous studies analyzed the correlations of volatilities [38, 42, 53, 54, 56–60, 66–69] which can be measured by the auto-correlation function [Eq. (1.4)]. The volatility turns out to be long-term correlated, meaning that the auto-correlation follows a power law,

$$C(\Delta t) \sim \Delta t^{-\gamma} \,. \tag{2.3}$$

with exponent $\gamma \simeq 0.3$ [60]. This is significantly different from the exponential decay in the auto-correlation for returns.

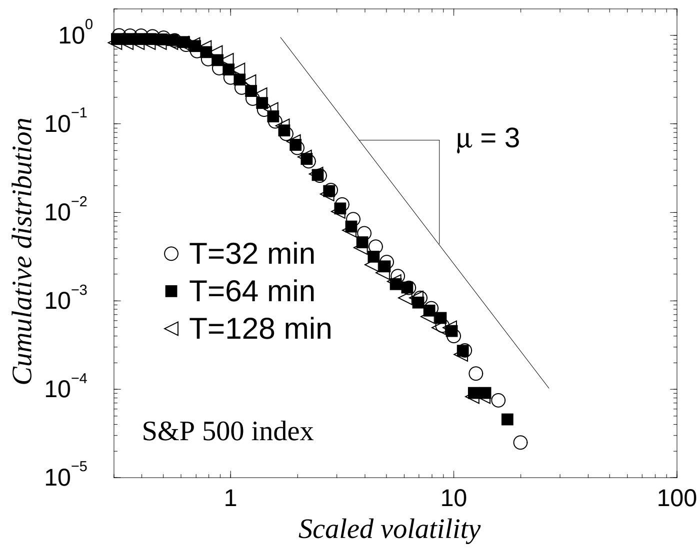

Figure 2.5: Cumulative distribution of the scaled volatility of S&P 500 index, for time scales $T = 32$, 64, 128 min with sampling interval $\Delta = 1$ min, using non-overlapping windows for the S&P 500 stock index. Regression lines yield estimates of the exponent $\mu = 3.10 \pm 0.08$ for $T = 32$ min, $\mu = 3.19 \pm 0.10$ for $T = 64$ min and $\mu = 3.30 \pm 0.15$ for $T = 128$ min. The fits were performed over the range of scaled volatility greater than 1 standard deviation. (From Liu et al. [60])

More accurate results for correlations are obtained by the detrended fluctuation analysis (DFA) method, see Appendix A. This method is based on the idea that a correlated time series can be mapped to a self-similar process by integration. Therefore, measuring the self-similar feature can indirectly tell us information about the correlation properties. The result for volatility is complicated, and shows some crossover in DFA curves, suggesting a multiscaling behavior. As an example, we plot the DFA curve for the S&P 500 index in Fig. 2.6. Two power law regimes with

$$\alpha_1 = 0.66 \pm 0.01, \qquad t < t_\times$$

$$\alpha_2 = 0.93 \pm 0.02, \qquad t > t_\times$$

were observed. The corresponding crossover is

$$t_\times \approx 600\,\mathrm{min},$$

which is approximately 1 trading day [60]. As expected, $\alpha = 0.5$ for the shuffled volatility time series.

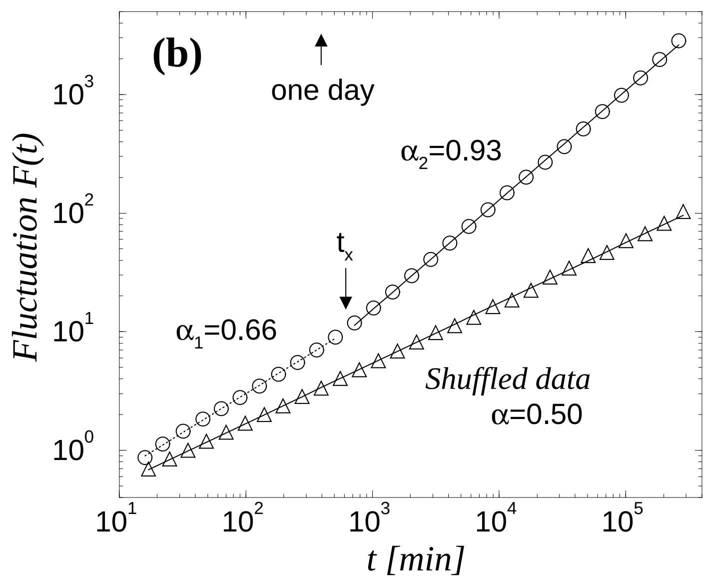

Figure 2.6: The detrended fluctuation analysis $F(t)$ of the S&P 500 index volatility with the sampling interval $\Delta t = 1$ min. The lines show the best power law fits (R values are better than 0.99) above and below the crossover time, $t_\times = 600$ min. The triangles show the DFA results for the "control", shuffled data. (From Liu et al. [60])

2.3 Two Components of Daily Return and Volatility: Overnight and Daytime

2.3.1 Introduction

Now we focus on a special analysis of return and volatility. As we know, usually financial markets are closed during the night, and all news or events in the night are reflected in the opening price of the next trading day. A day (from former day closing to current day closing) therefore can be decomposed into two sessions, overnight (from former day closing to current day opening) and daytime (from current day opening to closing) sessions. The study of the returns and the volatilities during these two sessions might provide new insights towards better understanding of the financial markets. Practically, this study can help traders to improve trading strategies at the market opening and closing. It also can help investors to analyze the dually-traded equiti

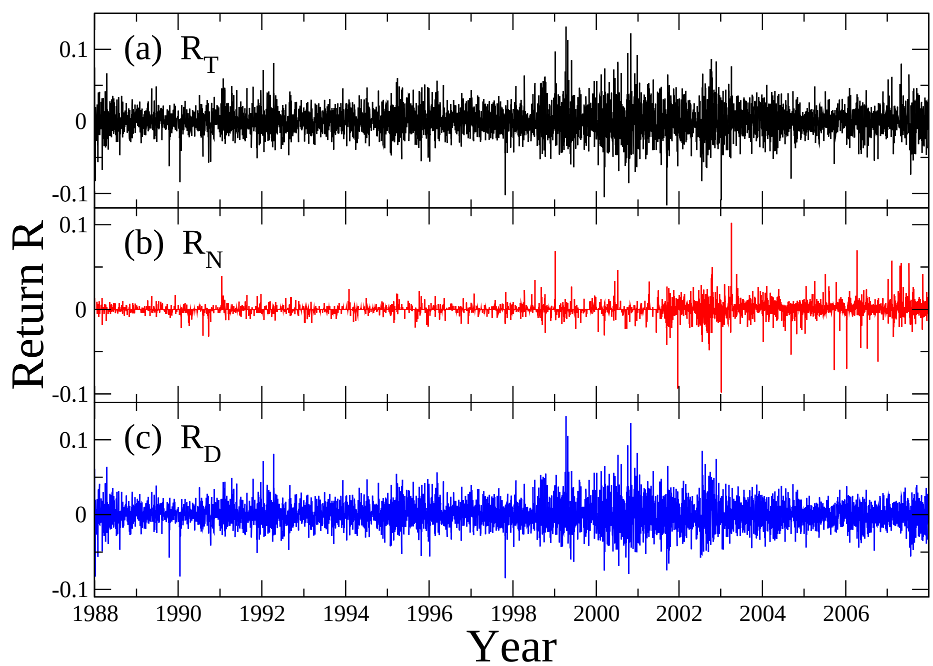

Figure 2.7: Three types of return, (a) the total return R_T, (b) the overnight return R_N, and (c) the daytime return R_D of a typical stock, AA, are shown. We can see that the fluctuations of R_N are relatively weaker, and the curve of R_D is more similar to that of R_T.

Here we present a comprehensive study on all equities listed on the NYSE on December 31, 2007, in total 2215 stocks. These historical stock data are available at http://finance.yahoo.com. To obtain good statistics, we only choose the stocks with more than 1000 records. The record starts from January 2, 1962, but many stocks have a much shorter history. We do not include the data before 1987 for two reasons. First, from 1962 to 1987 there exist only very little data, about 6.5% of all the data points for these 2215 stocks. Second and more important, there was a huge market crash on October 19, 1987 ("Black Monday"), and after that the market was adapted in a great extent. Thus, to reduce the complexity of market structure, we only examine the data from 1988 to 2007, in total 20 years. The length of the 2215 stocks ranges from $N = 1000$ to 5000 trading days. Note that many stocks have splits in the 20-year period, which causes significant change in the price. Therefore, we adjust all prices according to the historical splits. The 2215 stocks cover all industrial sectors, a wide range of market capitalization (from 6×10^6 to 5×10^{11} dollars), and a wide range of the average daily volume (from 500 to 2×10^7 shares a day).

In Chapter 1 we defined the daily return R, which is the logarithmic change of the successive daily closing prices ("total return" R_T in this section, to distinguish from its two component returns). Correspondingly, the return over the overnight session ("overnight return") is

$$R_N(t) \equiv \ln(p^{open}(t)/p^{close}(t-1)); \tag{2.4}$$

and the return over the daytime session ("daytime return") is

$$R_D(t) \equiv \ln(p^{close}(t)/p^{open}(t)). \tag{2.5}$$

Here $p^{close}(t)$ is the closing price and $p^{open}(t)$ is the opening price at day t. Note that $R_T(t) = R_N(t) + R_D(t)$, $R_D(t)$ and $R_N(t)$ are in the same day and $R_D(t)$ is after $R_N(t)$. As an example, Fig. 2.7 shows the three types of return for a typical stock AA (Alcoa, Inc.) from 1988 to 2007. Same as the definition in Chapter 1, the volatility V is defined as the absolute value of the return. Thus, corresponding to the three types of return, we have three types of volatility, the total volatility V_T, the overnight volatility V_N, and the daytime volatility V_D.

Table 2.1: Number of good fits of the volatility tail distribution for the 2215 NYSE stocks. Good fit refers to the cases where the null hypothesis is not ruled out for 1% significance level.

Volatility V	V_T	V_N	V_D
Power law	2066	1868	2066
Exponential	1693	644	1756
Power law with cutoff	1755	1772	1728

2.3.2 Tail of Volatility Distribution

The tail distribution accounts for large fluctuations and events which are very important for risk analysis. By the definition [Eq. (1.8)], the volatility aggregates both positive and negative returns and has better statistics. In addition, the distribution of the return is approximately symmetric in the two tails [60]. Therefore we focus on the tail distribution of the volatility. As shown by Eq. (2.2), the volatility has a power law tail, which might be also true for the overnight and daytime volatility. Here we employ the Maximum Likelihood Estimator (see Appendix B) to fit the tail distribution, using the Kolmogorov-Smirnov statistic D (see Appendix B also) as the goodness-of-fit.

To further test the volatility tail, we also try two other distribution functions in the same range and using the same method. One is the exponential distribution function,

$$P(x) \sim e^{-x/x^*}, \tag{2.6}$$

where x^* is a characteristic scale. The other is a power law function with an exponential cutoff,

$$P(x) \sim x^{-\zeta}\, e^{-x/x^*}. \tag{2.7}$$

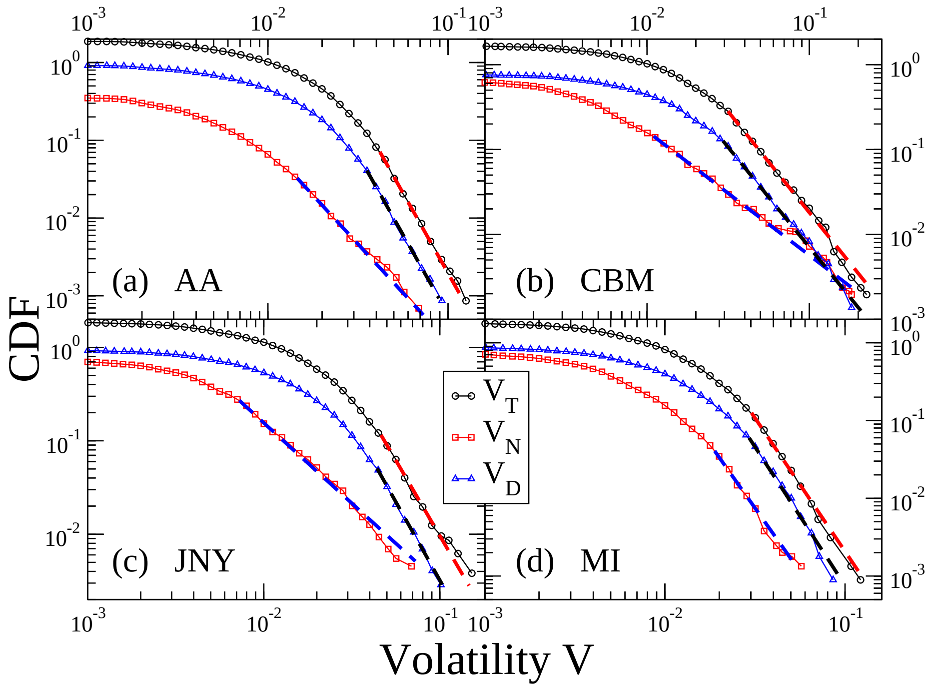

Figure 2.8: For four typical stocks, (a) AA, (b) CBM, (c) JNY, and (d) MI, three types of volatility, total volatility V_T (circles), overnight volatility V_N (squares), and daytime volatility V_D (triangles) are demonstrated. The dashed lines are power law fits to the distribution tails. Note that the curves for V_T (circles) almost coincide with those of V_D and thus they are vertically shifted for better visibility.

We examine the tail distribution of V_T, V_N and V_D for the 2215 NYSE stocks. The number of fits that the null hypothesis was valid for at a 1% significantance level ("good fit") is listed in Table 2.1. For the power law distribution, only a small portion (10%) of the three types of volatilities are ruled out, which manifests that the tail is well characterized by the power law function for the broad market. For the exponential hypothesis, almost half (38%) of all cases are ruled out. Moreover, for about 98% out of the good exponential fits, the power law hypothesis is not ruled out either. As a whole, the exponential function is poor for characterizing the tail, compared to the power law function. For the power law with an exponential cutoff, the percentage of good fit is 79% over the three volatilities, which is slightly lower than that for the power law. Besides, 99% of them do not reject the power law hypothesis either. Therefore, we conclude that the power law is the best among the three distributions.

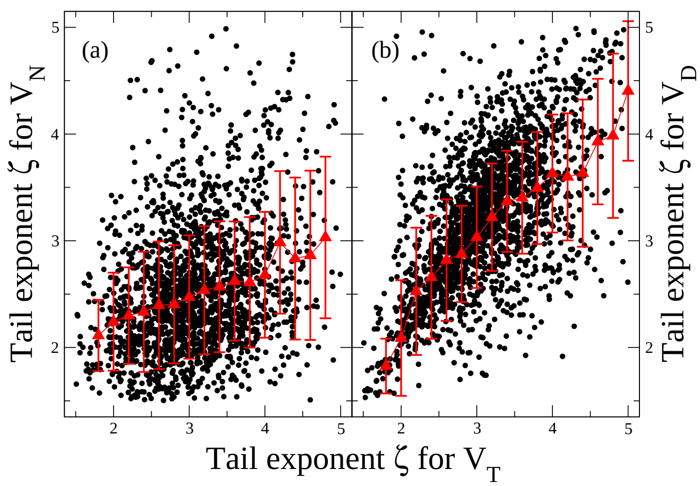

Figure 2.9: Relation between tail exponents of two component volatilities and that of total volatility V_T: (a) overnight volatility V_N and (b) daytime volatility V_D. A point represents a stock which has good power law fit to the tail for the corresponding two types of volatilities. 1812 out of the 2215 NYSE stocks are exhibited in panel (a) and 2001 stocks are exhibited in panel (b). To show the tendency, we divide the entire data set into equal-width subsets according the value of ζ for V_T and calculate the mean values and standard deviations in these subsets, as shown by the triangles and the error bars respectively. Both cases clearly show tendencies but that for the daytime volatility is stronger, indicating V_T is more influenced by V_D. Moreover, ζ for all three types of volatilities are distributed in a relatively narrow range and centered around 3.

In Fig. 2.8, we plot the CDF of V_T, V_N, and V_D for four typical stocks, namely, Alcoa, Inc. (AA), Cambrex Corp. (CBM), Jones Apparel Group, Inc. (JNY), and Marshall & Ilsley Corp. (MI). These stocks belong to diverse industrial sectors and their capitalization vary in a wide range, from 27 billion dollars for AA to 0.25 billion dollars for MI. As seen in Fig. 2.8, the tails are well fitted by power laws. Interestingly, the tails of V_D almost always decay faster than the tails of V_N, and V_T lies between the two component volatilities. Moreover, the log-log slope (tail exponent ζ) of V_T is closer to that of V_D, indicating that the daytime return contributes more to the total return.

To test this finding for the broad market, we plot in Fig. 2.9 the relation between the tail exponent ζ of V_T and ζ of the two component volatilities for the 2215 stocks. Both scatter plots show a certain dependence (as shown by the solid curves, which are averages over different bins of ζ of V_T), but the correlation between V_T and V_D is obviously stronger, which is consistent with Fig. 2.8. For all three types of volatilities, ζ is distributed in a certain range from 1.5 to 5, and centered around 3. The averages of ζ are: $\langle\zeta\rangle \approx 2.6$ for V_N is lower than $\langle\zeta\rangle \approx 3.2$ for V_D, while $\langle\zeta\rangle \approx 3.1$ for V_T is between the two component volatilities and it is slightly smaller than that for V_D. This behavior suggests that the daytime return influences the total return more than the overnight return.

2.3.3 Correlations in Component Returns and Volatilities

After analyzing the volatility distribution, a question naturally arises, how are these values organized in the time sequence? For the investors, the temporal structure is of special interest because it determines how and when to trade. The time organization in a time series can be characterized by the two-point correlation. It is known that the total return has only short-term correlations and the total volatility has long-term correlations [5, 40, 43, 60, 71, 72]. Now we examine the correlations in each of their two components (overnight and daytime).

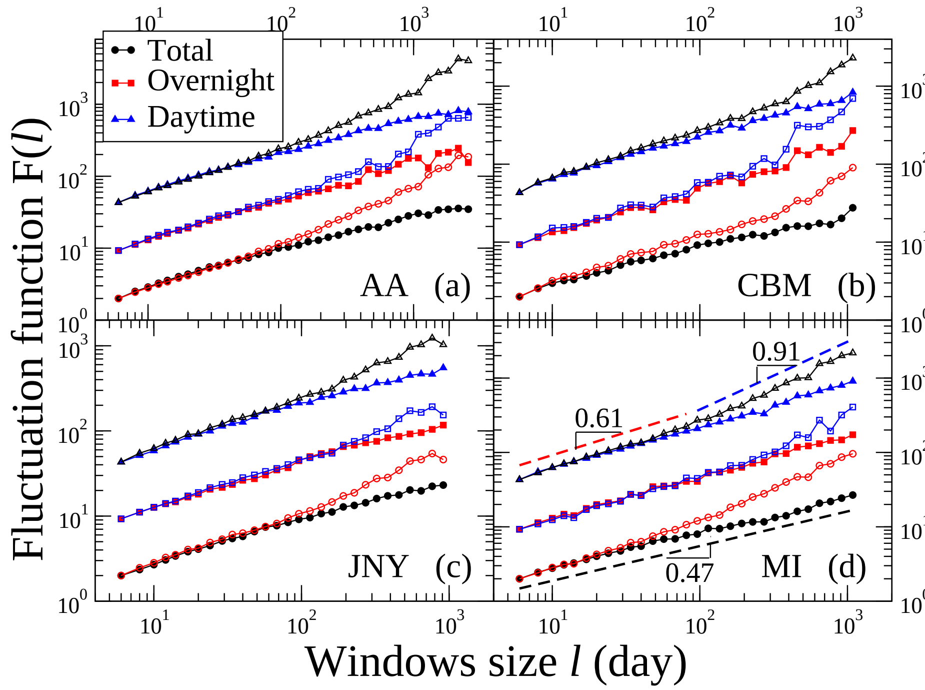

Figure 2.10: DFA fluctuation function F vs. windows size ℓ for the returns (filled symbols) and the volatilities (open symbols). The four panels are for stocks AA, CBM, JNY, and MI respectively. For each case, three types of data, total (circles), overnight (squares), and daytime (triangles) are shown. Note that the curves are vertically shifted for better visibility. To obtain the correlation exponent α, we fit all curves with power laws, as illustrated by the dashed lines in panel (d). For the volatility, the exponent α is significantly different for short and long time scales, thus we split the entire range into two regimes and fit them separately.

It is well known that financial time series are usually non-stationary. In such cases, the conventional methods for correlations such as auto-correlation and spectral analysis have spurious effects. To avoid the artifact correlations arising from non-stationarity, we employ the DFA method (see Appendix A). In Figure 2.10, we plot DFA curves for the returns and volatilities of the total, overnight, and daytime sequences for four typical stocks. The values of α are obtained by power law fits to the fluctuation function, as illustrated by the dashed lines in Fig. 2.10(d).

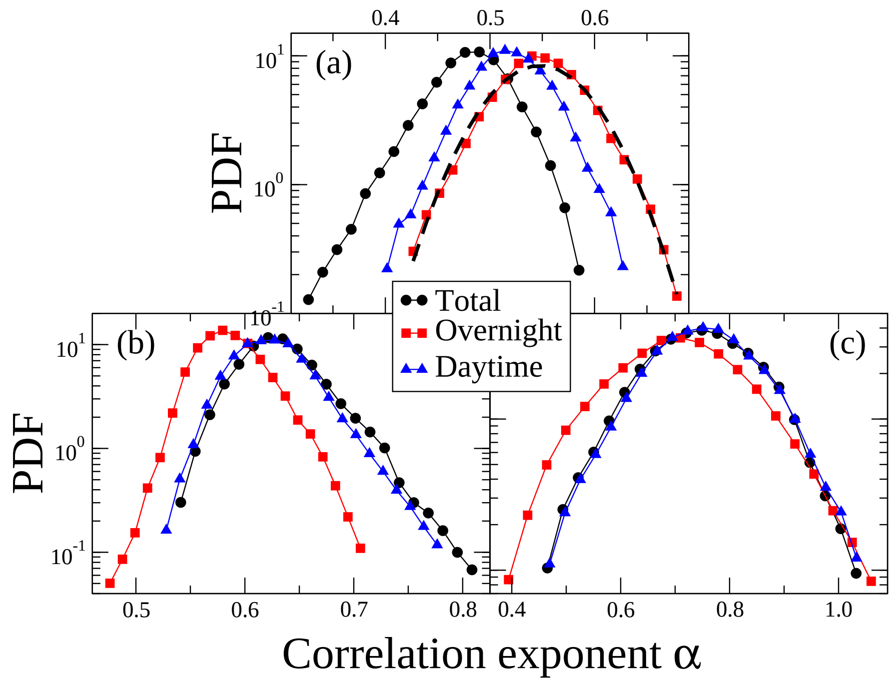

Figure 2.11: Distribution of the correlation exponent α, (a) the returns, (b) the volatilities of the short time scales, and (c) the volatilities of the long time scales. Three types of returns and volatilities, total (circles), overnight (squares), and daytime (triangles) are shown. All distributions approximately follow the normal distribution. For example, a normal distribution fit on the overnight return is shown by the dashed line in panel (a). For the volatilities, the curves for the total and the daytime almost collapse into a single curve in panels (b) and (c), suggesting that the total return is more influenced by the daytime return.

For all three types of returns, α is close to 0.5 and therefore there are no long-term correlations. For the volatilities, the fluctuation function is more complicated. The slopes (in log-log scale) of different regions are significantly different. Thus, we divide the whole curve into two equal-size regions in the logarithmic scale and fit them separately, as shown by the dashed lines in Fig. 2.10(d).

To test the universality of our findings, we plot in Fig. 2.11 the distribution of α for the three returns as well as for the short and long time scales of the volatilities. For the returns [Fig. 2.11(a)], the distributions are centered around 0.5, $\alpha = 0.48 \pm 0.04$ for the total, $\alpha = 0.55 \pm 0.05$ for the overnight and $\alpha = 0.52 \pm 0.04$ for the daytime. Here and in the following, the error bars are the standard deviations over all 2215 stocks. These error bars are quite small representing quite narrow distributions. This result is consistent with earlier studies, where no long-term correlations were found for the returns [60].

For the volatilities at short time scales [Fig. 2.11(b)], the distributions are centered around 0.6, $\alpha = 0.63 \pm 0.04$ for the total [60], $\alpha = 0.59 \pm 0.03$ for the overnight and $\alpha = 0.63 \pm 0.04$ for the daytime. For the volatilities at long time scales [Fig. 2.11(c)], $\alpha = 0.75 \pm 0.10$ for the total [60], $\alpha = 0.71 \pm 0.12$ for the overnight and $\alpha = 0.75 \pm 0.10$ for the daytime. For all time scales, the volatility α values are significantly larger than 0.5, suggesting long-term correlations in the volatility sequences. In addition, the α values of the long-term scales are systematically larger than that of the short-term scales. This multiscaling behavior indicates that the correlation becomes stronger for longer times. Moreover, all distributions are relatively narrow for both returns and volatilities, suggesting a universal feature over the entire market. We also notice that the curves of the total and daytime almost collapse onto a single curve, while the curve of the overnight departs away from them, supporting again that the daytime return contributes more than the overnight return to the total return.

Now we address the question whether there is a relation between the correlation exponents α of the two components of the return and volatility. If a certain stock has large (small) α for one component, does it have also large (small) α in the other component or in the total? To test this, we employ the cross-correlation function to quantitatively compare them. For our case, x and y are vectors representing the three

Table 2.2: Cross-correlation between the α values of the three types of returns and volatilities for the 2215 NYSE stocks. We divide the 2215 stocks into 10 equal-size subsets and calculate the cross-correlation for every subset. The error bar is the corresponding standard deviation of the 10 cross-correlations. The value in the parenthesis is the corresponding cross-correlation between two shuffled α records.

Cross-correlation	Total/Overnight	Total/Daytime	Overnight/Daytime
Return	0.25 ± 0.07 (-0.00)	0.51 ± 0.08 (-0.03)	0.48 ± 0.08 (0.02)
Volatility (short scale)	0.29 ± 0.09 (0.04)	0.80 ± 0.04 (-0.02)	0.24 ± 0.04 (0.03)
Volatility (long scale)	0.56 ± 0.06 (0.01)	0.90 ± 0.02 (0.02)	0.52 ± 0.05 (-0.02)

sequences of α (total, overnight, daytime) for the return, short time and long time volatilities for all companies. The companies are in the same order for all sequences. As shown in Table 2.2, all cross-correlations are significantly larger than that of shuffled records (values in the parenthesis), suggesting strong relations between the different returns or volatilities. Note again that the total-daytime pair is always the strongest one, which is in agreement with the assumption that the total return and volatility are significantly more influenced by the daytime return and volatility, than by the overnight return and volatility.

2.3.4 Relation between Component Returns

The overnight return and the daytime return are the price changes over different sessions of a trading day, and they make the total return. It is interesting to examine now if the three returns of the same stock are cross-correlated. This will test the question, e.g., how changes in the day time are related to those of night time or the total.

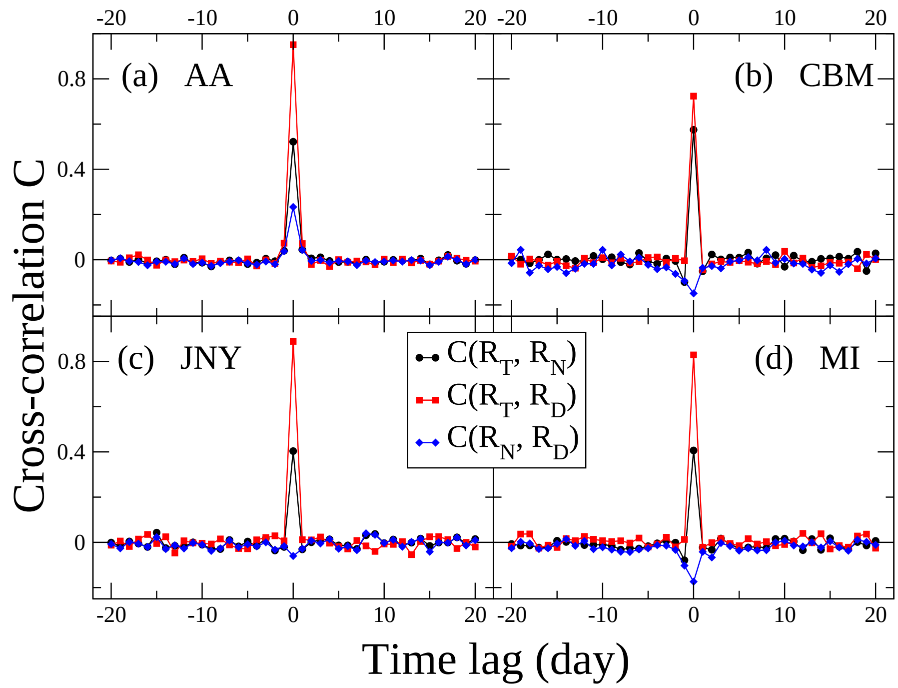

Figure 2.12: For all four stocks (a) AA, (b) CBM, (c) JNY, and (d) MI, the two cross-correlations with respect to the total return are significant larger than their cross-correlations with the time lags, which suggests both component returns are strongly positively correlated to the total return. However, the cross-correlation between the two component returns varies with the stocks, e.g., it is positive for AA and negative for MI.

The cross-correlation function [Eq. (1.5)] examines the two time series without any time lag. However, there might be some time delays between two time series, and therefore we shift the two sequences by time lag Δt to test this possibility. Moreover, the comparison between the cross-correlations with different lags allows us to examine the significance of a cross-correlation value. Therefore, we employ the cross-correlation with time lag Δt [Eq. (1.6)]. In general one tests the position of the maximum (minimum if it is anti-correlated) of $C_{\Delta t}$ which may occur at $\Delta t = \tau$ and τ is called the time delay [73]. Here we find that the maximum of $C_{\Delta t}$ is always for $\Delta t = 0$ (as shown in Figure 2.12).

In this section we use $C_{\Delta t}$ to test the significance of the cross-correlation at $\Delta t = 0$. If $C_{\Delta t=0}$ is significantly different (higher or lower) from $C_{\Delta t\neq 0}$, the cross-correlation can be regarded as reliable. Quantitatively, we use the standard deviation of $C_{\Delta t\neq 0}$ values over the range $-20 \leq \Delta t \leq 20$, $\sigma(C_{\Delta t\neq 0})$, to test the reliability of the cross-correlation [73]. As examples, we plot in Fig. 2.12 the cross-correlations of three pairs of returns for the four typical stocks, AA, CBM, JNY, and MI (other stocks have similar features). For both $C(R_T, R_N)$ and $C(R_T, R_D)$, the cross-correlations at $\Delta t = 0$ are more than 10 times higher than their $\sigma(C_{\Delta t\neq 0})$ so they are very robust. However, for $C(R_N, R_D)$, the cross-correlations vary with the stock. Some of them have significant cross-correlation values but some of them are in the range of their $\sigma(C_{\Delta t\neq 0})$. Since R_N and R_D covers different periods, there could be some strong correlations or almost independent, it is reasonable that the cross-correlation varies in a wide range. On the other hand, R_T always shares part of changes with its two component returns and deduce strong positive cross-correlations.

Next we examine the three pairs of cross-correlations $C_{\Delta t=0}$ for all the 2215 stocks (in the following, the function C refers to $C_{\Delta t=0}$ if the Δt subscript is missing). Their distributions are plotted in Fig. 2.13. For each pair, the cross-correlations are distributed in a certain range. The cross-correlation between the total return and the daytime return, $C(R_T, R_D) = 0.8 \pm 0.1$ (mean value and standard deviation over the 2215 stocks), is always the largest value in the three pairs. The cross-correlation between the total and the overnight, $C(R_T, R_N) = 0.4 \pm 0.1$, is a still high but significantly smaller than $C(R_T, R_D)$ values. The cross-correlation between day and night, $C(R_N, R_D) = -0.1 \pm 0.1$, is distributed around 0 with more tendency to have negative values. In summary, the total return is more synchronized with the daytime return. It is also interesting to note that there are significantly more stocks that have negative correlations between R_N and R_D. For example, 567 out of the 2215 stocks have values of $C(R_N, R_D) < -0.2$. This implies that the probability is relatively high for a large positive overnight return to be followed by a large negative daytime return. The overnight return and the daytime return tend to be slightly anti-correlated, and the total return usually moves in the same direction as the daytime return.

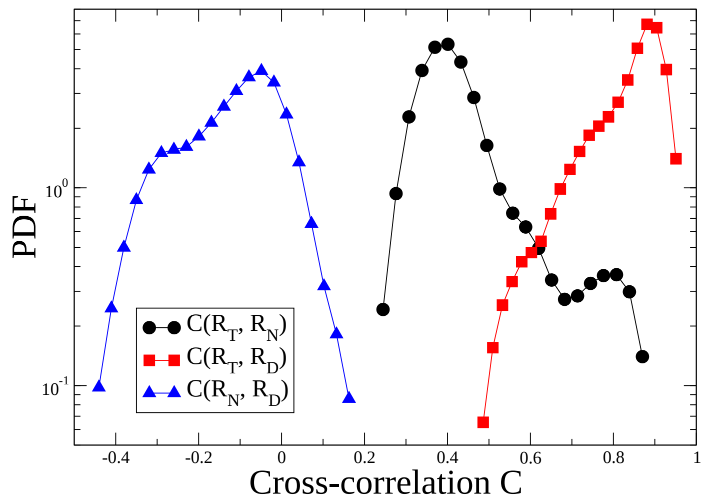

Figure 2.13: Distribution of the cross-correlations C between the returns R_T, R_N, and R_D for the 2215 NYSE stocks. Both cross-correlations with respect to the total return are significantly larger than 0 and that for the daytime return is stronger, suggesting that the total return is more correlated to the daytime return. The cross-correlation between the two component returns is relatively more distributed towards negative values, indicating the two component returns tend to be anti-correlated.

Due to many factors, such as changes in the regulations or new technologies, the markets evolve with time. An interesting question arises, whether the cross-correlation is stable in the sample years studied. To test the stability of the cross-correlations, we recalculate the cross-correlations year by year. The records in 1 year are enough to calculate the cross-correlation and more importantly, the equity market in such a short period can be assumed stable. In Fig. 2.14 we plot the averages and standard deviations (as error bars) of the cross-correlations over the 2215 stocks against the year. For the three types of cross-correlations, the curves only slightly vary with the years and all changes lie within the error bars. Moreover, the error bars are almost the same for all years, which clearly shows that the cross-correlation is quite stable over the 20 year period studied.

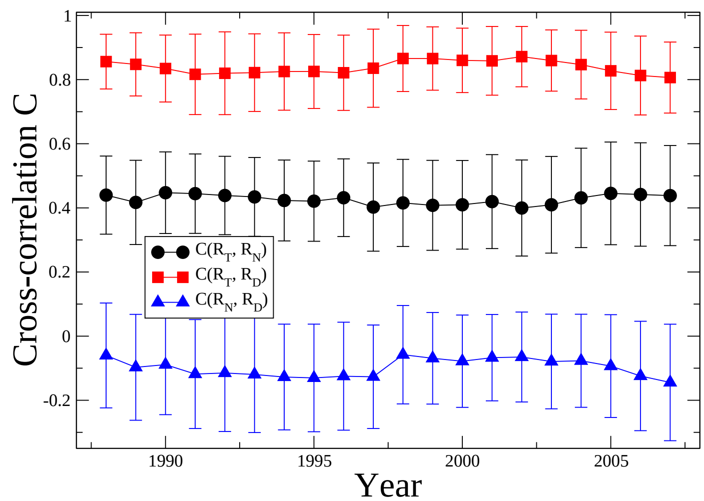

Figure 2.14: Evolution of the cross-correlations between the three returns, R_T, R_N, and R_D, from 1988 to 2007. Here a point represents the average over the cross-correlations of the 2215 NYSE stocks in a 1-year period, and the error bar is the corresponding standard deviation. Clearly, there are no significant changes for the cross-correlations over the 20 years studied.

2.3.5 Discussion

Returns and volatilities might be affected by some factors, such as the market capitalization and the mean volume [29, 30]. To test this for the entire stock market, we investigate the relation between the two factors, capitalization and mean volume, and the measures, such as the tail exponent ζ, the correlation exponent α, and the cross-correlations between the three returns. There are some tendencies between these factors and measures. However, most of these tendencies are within the range of the error bars, which suggests no significant dependence between the two factors and three measurements. The behavior of the three measures is quite universal over the entire market. To better understand the complexity of the equity market, the connection between different measurements and factors of stocks might need to be further analyzed.

In summary, we examined the distributions of the total, overnight and daytime volatility. Compared to the exponential and power law with cutoff, the power law distribution is found to be generally better. The tail exponent ζ is distributed among the different stocks between 1.5 and 5 for the three types of volatility. We also analyzed the correlations in returns and volatilities of the components, using the DFA method. For both returns, there are no long-term correlations. However, for both volatilities, there are long-term correlations at all time scales and the correlations are even stronger at long time scales. For the tail distribution and for the long-term correlations, the results of the two component returns and volatilities are similar to the total return and volatility. Moreover, the records of the daytime are more similar to the total of the same stock, suggesting that the daytime return contributes more to the total return. To better compare these similarities, we studied also the cross-correlations between the different types of return and found consistent behaviors, i.e., the daytime is more correlated to the total compared to the night time. Further, the cross-correlation between the overnight return and the daytime return varies for different stocks, and interestingly, for a significant fraction of the 2215 stocks is far below 0. This finding suggests that the daytime return has a considerable probability to strongly anti-correlate with the overnight return. Furthermore, we examined the cross-correlations year by year and found that the behavior is quite stable over the 20-year period.

Chapter 3

Scaling in Return Intervals

Now we focus on the new approach to financial fluctuations, return intervals, which reflects the temporal structure in the volatility time series. In this chapter we discuss the distribution, one of most important statistical properties for a time series. We show that return intervals have well-approximated scaling in the distribution. The scaling function follows a stretched exponential function. Interestingly, the shape parameter γ of the stretched exponential function is found to be related to correlations in volatilities. In addition, the tail distribution also obeys a power law function.

3.1 Return Interval Distribution and Scale Transformation

The distribution of a data set can be characterized by a probability density function (PDF) or a cumulative distribution function [Eq. (1.3)]. By the definition of return interval τ, we obtain one time series of τ for one threshold q. Thus first we examine the relation between the distribution of return intervals and the threshold. As an example, we plot the probability density $P_q(\tau)$ vs. threshold q for two typical stocks and a benchmark index, namely, Citigroup (C), General Electric (GE), and S&P 500 index in Fig. 3.1(a), (c), and (e). Note that $P_q(\tau)$ has high values at large τ for large q, in contrary to that of small q. This behavior is trivial. Since fewer volatilities are choose as "events" for a higher threshold, corresponding return intervals are larger and the probability of larger τ values will be higher. Now the question is, is there any relation for distributions at different thresholds?

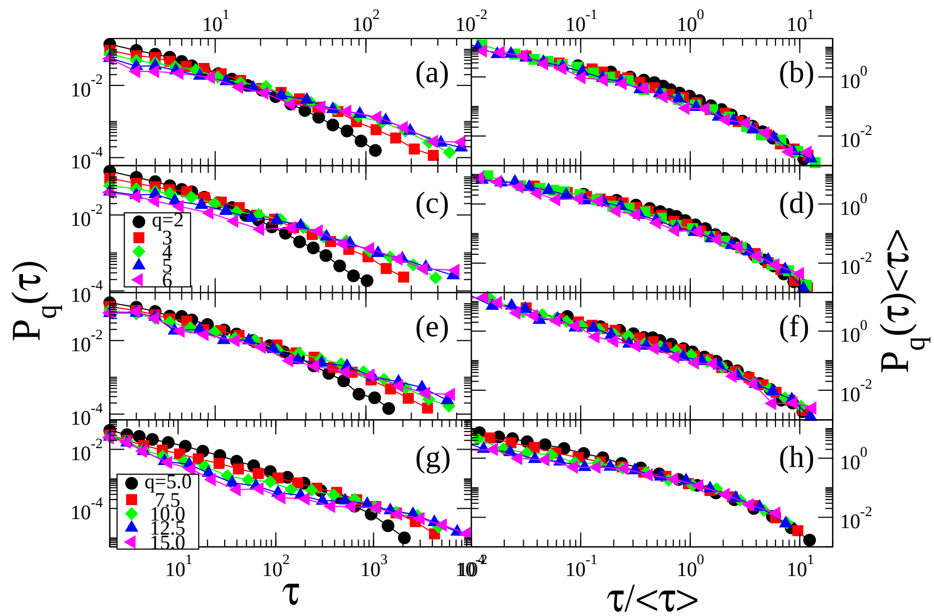

Figure 3.1: Distribution of return intervals and scaling. Panel (a) and (b) for stock C, (c) and (d) for stock GE, (e) and (f) for index S&P 500, and (g) and (h) for the ensemble of 30 DJIA component stocks. Symbols are for different threshold q, as shown in (c) for (a)-(f) and shown in (g) for (g) and (h). The sampling time is 10 min for S&P 500, and 1 min for stocks. For one data set, the distributions $P_q(\tau)$ are different with different q, but they collapse onto a single curve for $P_q(\tau)\,\langle\tau\rangle$ vs $\tau/\langle\tau\rangle$, which indicates a scaling relation. (g) and (h) show that the scaling can extend to very large thresholds.

Remarkably, if we have a scale transformation, i.e.,

$$\begin{aligned} &\textit{return interval } \tau \Rightarrow \textit{scaled return interval } \tau/\langle\tau\rangle, \\ &\textit{accordingly }, \textit{PDF } P_q(\tau) \Rightarrow \textit{scaled PDF } P_q(\tau)\,\langle\tau\rangle, \end{aligned} \tag{3.1}$$

all distributions of $q = 2$ to 6 (note that the volatility is in units of standard deviation) collapse onto a single one, as shown in Fig. 3.1(b), (d), and (f), i.e.,

$$P_q(\tau) = \frac{1}{\langle\tau\rangle} f(\frac{\tau}{\langle\tau\rangle}). \tag{3.2}$$

This behavior suggests that the distribution of return intervals has a well-approximated scaling [24, 26].

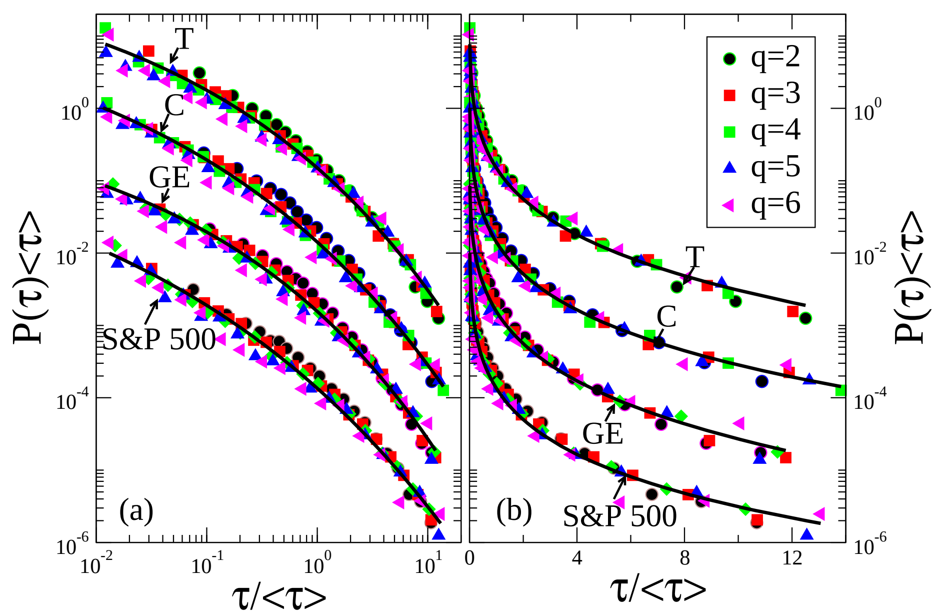

Figure 3.2: Stretched exponential fit of return interval distributions, (a) log-log scale, (b) semi-log scale. Scaled PDFs of three typical stocks, namely, AT&T (T), Citigroup (C), General Electric (GE), and S&P 500 index, are plotted. Each stock is well-approximated by stretched exponential for diverse thresholds, $q = 2, 3, 4, 5$, and 6, as shown by solid lines. For clarity, curves are vertically shifted by factors of $\times 10$.

The function f in Eq. (3.2) represents the scaling feature of the return interval distribution. This scaling function does not depend explicitly on q, but only through the mean interval $\langle\tau\rangle$ (Thus, to simplify the notation, we neglect the subscript q for

P in the following text). If $P(\tau)$ is known for one value of q, Eq. (3.2) can make predictions for other values of q—in particular for very large q (extreme events), which are difficult to study due to the lack of statistics.

We also examine all other 28 component stocks of DJIA and find consistent results. In Fig. 3.1 we show that the ensemble distribution of 30 DJIA has similar scaling transformation, over a wide range of thresholds, from $q = 5$ to 15 standard deviations. Further, we plot distributions of threshold $q = 2$ for the 30 DJIA component stocks and S&P 500 index in Fig. 3.2(a). Remarkably, we also find that all these distributions collapse onto a single curve. All these findings suggest that the scaling in return intervals is universal, which will be further discussed in later sections.

3.2 Form of Scaling Function

To fully describe the distribution of return intervals, we need to identify the form of the scaling function f in Eq. (3.2). First we examine the full range of the distribution, and then we specifically test the tail regime.

3.2.1 Full Range: Stretched Exponential

From Figs. 3.1 and 3.2(a), we can see that the distribution is close to a straight line on the log-log scale (which suggests a power law distribution) but not exactly. Clearly, in Fig. 3.2(b), we can see that all distribution curves have long tails in the semi-log plot, indicating that these distributions decay slower than an exponential function. The scaling function lies between a power law function and an exponential function. We test the distribution with the stretched exponential (SE) function find that there is a very good match, as shown by solid lines in Fig. 3.2. Therefore, the scaling function f can be written as,

$$f(x) \sim e^{-(\frac{x}{x^*})^{\gamma}}. \tag{3.3}$$

Here x^* is the characteristic scale and γ is the shape parameter, which is related to the correlations in the volatility sequence and thus called "correlation exponent" [19]. For an uncorrelated series, f reduces to a regular exponential function and $\gamma = 1$, as demonstrated in Fig. 3.2. From Eq. (3.3), the PDF function will be

$$P(\tau) \sim e^{-(\frac{\tau}{a})^{\gamma}}. \tag{3.4}$$

Then a is the characteristic scale. From the definition of PDF and $\langle\tau\rangle$, one may find that the parameter a depends exclusively on γ [23, 29],

$$a = \frac{\langle\tau\rangle\,\Gamma(\frac{1}{\gamma})}{\Gamma(\frac{2}{\gamma})}, \tag{3.5}$$

where $\Gamma(a) \equiv \int_0^\infty t^{a-1}e^{-t}dt$ is the Gamma function. However, due to the discreteness and finite size effects, there are some systematic deviations from the scaling law [22, 29]. To avoid them, we will also use a as a free parameter in the SE fit. To simplify the calculation and without loss of generality, we assume $\frac{\tau}{a}$ is continuous, then the corresponding CDF, $Q(\tau)$, follows

$$Q(\tau) \equiv \int_x^\infty P(\tau)\,d\tau \sim \Gamma(\frac{1}{\gamma},\,(\frac{\tau}{a})^{\gamma}). \tag{3.6}$$

where $\Gamma(a, x) \equiv \int_x^\infty t^{a-1}e^{-t}dt$ is the incomplete Gamma function. Note that in this dissertation the "cumulative distribution" is precisely the complementary cumulative distribution. Since CDF accumulates the information of the series and has better statistics than PDF, in the following we obtain the correlation exponent γ by fitting the CDF with Eq. (3.6).

As an example, we plot three CDFs (for q = 2, 4 and 6 respectively) of the GE stock in Fig. 3.3. The three curves are distant from the other, due to the difference in $\langle\tau\rangle$. The least-squares fits with Eq. (3.6) are illustrated by the solid lines, which agree with empirical points very well.

Table 3.1: Number of good fits on return interval CDF of the 1000 most traded U.S. stocks. If KS statistics D is smaller than the critical value CV, the corresponding distribution is not rejected (see Appendix B). Two types of distribution, stretched exponential (for the whole range) and power law (for the tail regime), are tested.

Threshold q	1	2	3	4	5	6
Stretched exponential fit	791	795	815	933	977	986
Power law fit	31	349	626	826	839	710

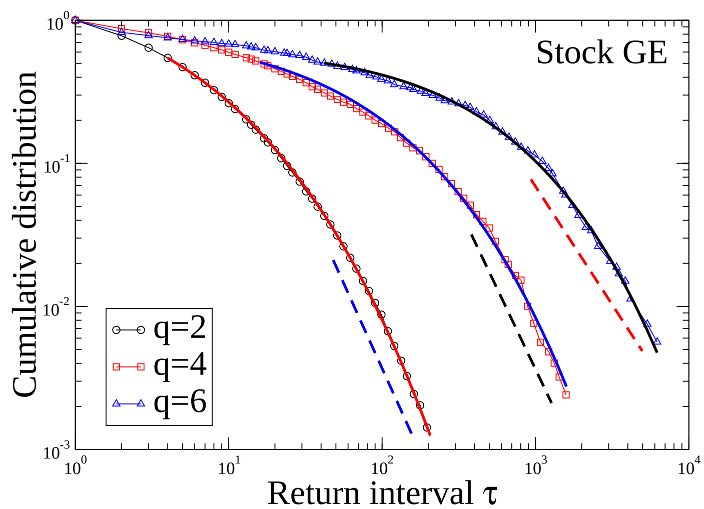

Figure 3.3: Cumulative distribution of return interval τ. CDF of three typical thresholds $q = 2$, 4 and 6 for the GE stock are plotted. Examples of two types of fit are displayed, dashed lines (left shifted for better visibility) are power law fits for the distribution tails and solid lines on symbols are stretched exponential fits for the whole range of distributions.

Further, we fit the CDF with a SE function for the 1000 most traded stocks in the U.S. markets for 2001-2002 period. To avoid the discreteness for small τ (Ref. [22]

suggested a power law function for this range) and large fluctuations for very large τ, we choose the range of $0.01 \leq Q(\tau) \leq 0.50$ to perform the fit. The Kolmogorov-Smirnov (KS) statistic D is employed to test the goodness-of-fit (see Appendix B). The range of threshold is from $q = 1$ to 6, and the number of fits that are not rejected ("good fit") is listed in Table 3.1. We can see that most of the cases have a good fit by a stretched exponential function. This remarkable universality will be discussed later.

3.2.2 Tail Regime: Power Law

For financial time series, the distribution tail usually is characterized by a power law function [1–5, 40, 41, 43, 44, 60, 71, 72]. As for the return interval, Yamasaki et al. suggested that the scaling function is also consistent with a power law tail for large intervals, where the tail exponent is around 1 for both stock and currency data [24]. Moreover, Bogachev and Bunde have shown that the distributions of return intervals are governed by power laws [34]. Then the CDF of return intervals would follow

$$Q(\tau) \sim \tau^{-\zeta}, \tag{3.7}$$

where ζ is the tail exponent.

To test this hypothesis we examine the distribution tail for the 1000 most traded U.S. stocks, using the maximum likelihood estimator (see Appendix B). Examples of power law fits are demonstrated by the dashed lines in Fig. 3.3. We still use KS statistics to test the goodness-of-fit. For threshold $q = 1$ to 6, the number of good fits are listed in Table 3.1. For return intervals of $q = 1$ and 2, the power law distribution is not ruled out for only a small portion of the 1000 stocks. However, for other thresholds, the power law distribution is not ruled out for a significant portion of stocks. In Fig. 3.4 we plot the distribution for tail exponent ζ. Interestingly, all PDFs are centered around a certain value which systematically shifts from large value to small, with increasing the threshold. For $q = 2$, ζ is centered around 2, and for

$q = 5$, ζ is centered around 1. The latter is consistent with Ref [24], which suggested that the difference may due to the limited number of data points. Ref [24] was using daily data, which are about 1/20 of the intraday data in this dissertation ($\sim 10,000$ points for the daily data vs. $\sim 195,000$ points for the intraday data). Similarly, the number of return intervals for $q = 5$ is only about 1/14 of that for $q = 2$ (average over the 1000 stocks, ~ 850 points for $q = 5$ vs. $\sim 11,800$ points for $q = 2$). We also must note that, for $q = 2$, only about 1/3 of the 1000 stocks have a good power law fit (Table 3.1).

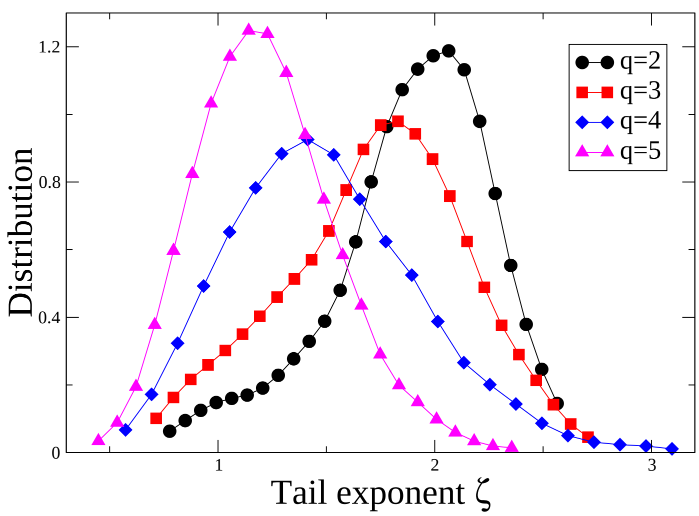

Figure 3.4: Probability density of tail exponent ζ from power law fit on the cumulative distribution of return intervals. The distribution systematically shifts from right to left, with increasing of the threshold.

In conclusion, we show that the distribution of return intervals is well-approximated by a stretched exponential. In addition, for many stocks, the power-law hypothesis on the tail distribution can not be ruled out.

3.3 Universality of Scaling

We already have shown some impressive universality of the scaling in return interval distributions, which is valid for a wide range of thresholds and many stocks. By the definition, return intervals are associated with several measures such as threshold and sampling time. Moreover, these time series could come from different stocks or different markets. To further explore the validity of scaling, we conduct a comprehensive examination of the scaling feature over all these additional "dimensions" in this section.

3.3.1 Thresholds

First we check the scaling with thresholds, which is the basic one for the return interval definition. We have found that the scaling on this dimension is strongly persistent. As shown in Fig. 3.1, the range of scaling can be extended to 15 standard deviations if we ensemble the 30 DJIA component stocks. Now we extend it to a much broader market, the 1000 most traded U.S. stocks from the TAQ database and examine the validity of scaling.

A direct way to test the scaling is the similarity of correlation exponent γ of different thresholds. From what we find for DJIA stocks, we expect that γ for different thresholds are strongly related, and their discrepancy should be small. To test this assumption we plot in Fig. 3.5 the dependence of the γ for typical thresholds $q = 3, 4, 5$ and 6 on the γ obtained for $q = 2$. Remarkably, all four cases show significant tendency and the slopes of linear fit are very close to 1, supporting the well-approximated scaling in the distribution of return intervals. Note that the fluctuation is larger for a higher q, and the slope slightly decreases, which may be due to the limited data size for large thresholds. We also test the dependence of other pairs of thresholds and observe similar behaviors. All these behaviors are consistent with Ref [30].

Moreover, we compare the value of the characteristic scale a with Eq. (3.5) and find that a from the fit is of the same order as that from Eq. (3.5), and usually the former is smaller. The ratio between two a values is centered from 0.4 (for $q = 1$) to 0.8 (for $q = 6$) for the 1000 stocks. It sounds that the ratio between two a values significantly deviates from 1. However, it is not that huge if we test the sensitivity on γ for Eq. (3.5). For instance, when γ changes from 0.30 to 0.31, the value a from Eq. (3.5) increases more than 30% with a constant $\langle\tau\rangle$.

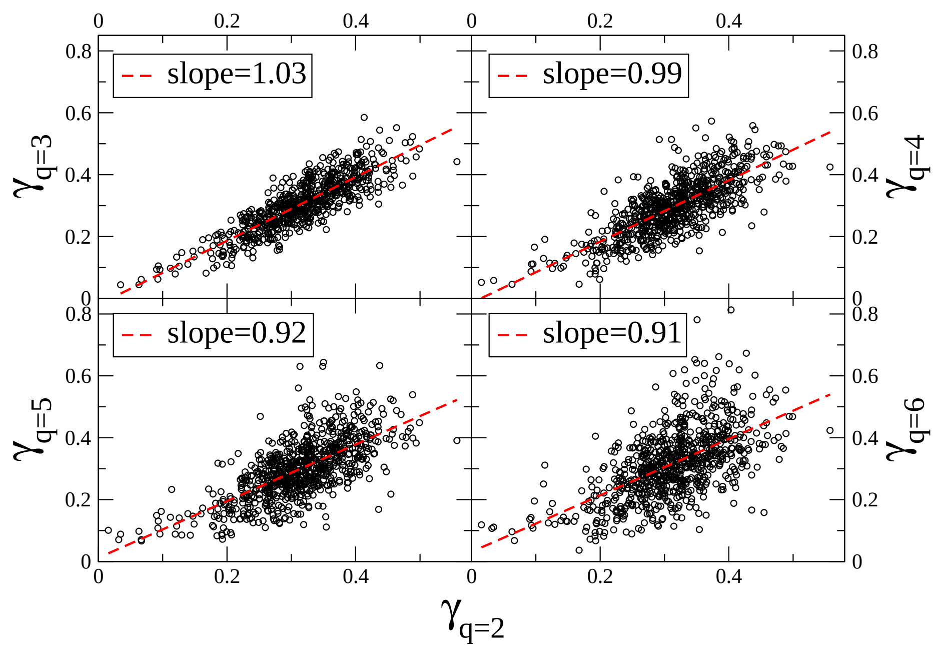

Figure 3.5: Relation between correlation exponent γ (Eq. (3.4)) of different thresholds. γ for four thresholds, $q = 3$ to 6 strongly depend on γ for $q = 2$, as indicated by dashed lines from the linear fit. All slopes of fit are quite close to 1, which suggests a good scaling in the distribution of return interval. Note that the fluctuation becomes stronger for a larger q, which relates to the smaller data size for the return interval with a larger q.

To now we have shown that the scaling holds for a quite large range, from $q = 1$ standard deviation to 6 standard deviations. The similarity for different stocks (see Fig. 3.1 for example) allows us to ensemble the data over many stocks and therefore extend the scaling examination to even larger thresholds. As shown in Fig. 3.1(g) and (h), we calculate the return intervals of each DJIA stock, and then aggregate all these points for certain thresholds. The scaling behavior is observed at even $q = 15$ standard deviations. We also have tried a large collection of data sets, 500 component stocks of S&P 500 index, and find that the scaling can be extended to 30 or even 40 standard deviations! Note that the probability of 15 standard deviations or more is $\sim 10^{-50}$ for a Gaussian distribution.

3.3.2 Sampling Intervals

For statistical analysis, the time resolution of records is an important aspect since the system may exhibit different behaviors in different time windows Δt. This is the second dimension for testing the universality of scaling. In Ref [27], Wang et al. have shown that the scaling law of return interval distributions is valid even to a sampling interval of 1 trading day. Here we change the volatility sampling interval from 1 min to 5, 10 and 30 min, and then examine its return interval distribution. For 1-day resolution, there are only 500 points for the 2-year intraday stock data set. Because the statistics are poor, we do not test the return interval here. Also for a good statistics we focus on return intervals of a typical threshold, $q = 2$. Similar to the 1-min resolution, most of cases can be well fit by Eq. (3.6). For instance, with 5-min resolution, the stretched exponential hypothesis for 812 of 1000 stocks are not rejected under 1% significance level. As examples, in Fig. 3.6 we show the return interval distribution of five time scales for four typical companies ($q = 2$): T, C, GE, and IBM. It is seen that for $\Delta t = 1$, 5, 10, 15, and 30 min, the $P_q(\tau) \langle \tau \rangle$ curves collapse onto a single curve, showing the persistence of the scaling for a broad range of time scales. Thus there seems to be universal structure for stocks not only in different companies, but also in each stock with various time resolutions. For a related study of persistence in different time scales of financial markets, see [74].

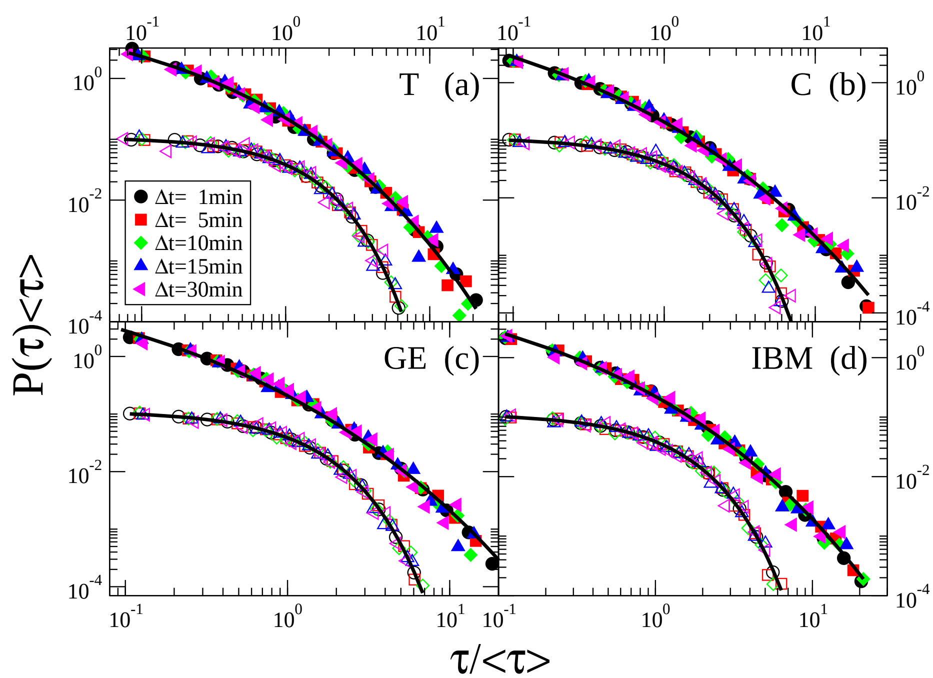

Figure 3.6: Scaling of return interval distributions for different time windows, $\Delta t = 1, 5, 10, 15$, and 30 min. Plots display scaled PDF $P(\tau)\,\langle\tau\rangle$ with threshold $q = 2$ for original (closed symbols) and shuffled volatility records (shifted by factor 10, open symbols) vs $\tau/\langle\tau\rangle$ of (a) T, (b) C, (c) GE, and (d) IBM. Each symbol represents one time scale Δt, as shown in panel (a). Curves fall onto a single one for actual return intervals and shuffled data, respectively, indicating the scaling. Also, the actual return intervals suggest a stretched exponential scaling function, demonstrated by the line fitting the solid symbols. The stretched exponential is a result of the long-term correlations in the volatility records. The shuffled records display no correlation, indicated by the good fit (solid line) to the Poisson distribution.

To examine the scaling on time resolutions over the broad market, we plot the PDF of γ for $\Delta t =$ 1, 5, 10 and 30 min in Fig. 3.7. The shape of the distribution systematically changes with increased sampling interval. The center shifts to the right slightly and the width increases. However, these behaviors are consistent with the change of dataset size. For a lower resolution, we have fewer data points and consequently stronger fluctuations of γ values. Therefore, these curves support the persistence of the scaling for a broad range of sampling intervals.

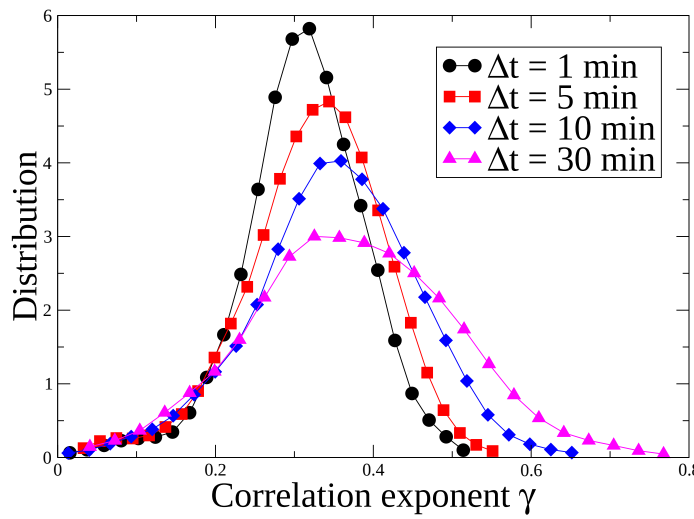

Figure 3.7: Distribution of correlation exponent γ for four representative sampling intervals, $\Delta t = 1$, 5, 10 and 30 min. With the increasing of sampling interval, the distribution tends to be wider. However, their centers are still close, changing from 0.31 for $\Delta t = 1$ min to 0.37 for $\Delta t = 30$ min, which suggests that scaling is valid for this range of sampling intervals. The broader distribution for lower resolution may be related to its smaller data size.

3.3.3 Stocks

Next, we investigate the similarity of scaling functions for different companies. Scaled PDFs $P(\tau)\langle\tau\rangle$ with $q = 2$ for return intervals (upper symbols) are plotted in Fig. 3.8, showing the S&P 500 index and 30 DJIA component stocks (one symbol represents one data set).

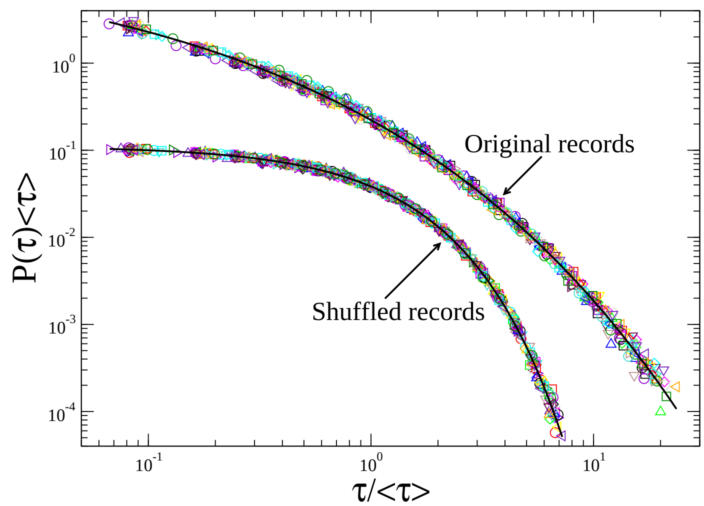

Figure 3.8: Distribution of return intervals of 30 DJIA component stocks (threshold $q = 2$). One kind of symbol in the plot represents one DJIA stock. Remarkably, all curves collapse onto a single one, suggesting that the scaling is persistent for many stocks. Note that the distribution can be well fit by a stretched exponential function, as shown by the solid line on the symbols. To find the origin of the form of scaling function, we also plot return interval distributions of shuffled records of 30 DJIA stocks, and find that they follow the exponential function of a Poisson distribution, as shown by the solid line on symbols. A Poisson distribution indicates no correlation in the shuffled data, however, the stretched exponential behavior for original records indicates strong correlations in original volatilities (see [19]).

We see that all PDF curves collapse, so their scaling functions are similar, consistent with a universal structure for $P(\tau)$. As suggested by the line on upper symbols in Fig. 3.8, as well as Figs. 3.1, 3.3, and 3.6, the scaling function follows a stretched exponential function as Eq. (3.4). Remarkably, we find that all 31 data sets in Fig. 3.8 have similar exponent values, and conclude that γ appears to be "universal", with

$$\gamma = 0.38 \pm 0.05. \tag{3.8}$$

The value a is found to be almost the same for these 31 data sets,

$$a = 3.9 \pm 0.5. \tag{3.9}$$

Note that the error bar is $\sigma(\gamma)$.

For a broad market, we can see Fig. 3.5 and find that γ values are distributed in a limited range for the 1000 most traded U.S. stocks from TAQ database, for example. The corresponding γ values (for $q = 2$ as instance)

$$\gamma = 0.31 \pm 0.07, \tag{3.10}$$

which is in agreement with the 30 DJIA component stocks. For other thresholds, we find similar results. In summary, we find that return interval distributions have well approximated scaling over stocks of a broad equity market.

3.3.4 Markets

Fig. 3.5 supports quite impressively the universality hypothesis of the correlation exponent γ since it holds for a broad market, the 1000 most traded stocks in the US markets, with a wide range of thresholds. We also show that the scaling holds for a wide range of sampling intervals. The following question is, can we find similar behaviors for other stock markets and financial markets? Here we check other financial markets, including currencies, federal funds rate (U.S. interest rate), oil and gold commodities. When we apply the scale transformation as Eq. (3.1), we find similar scaling feature for all these assets, as shown in Fig. 3.9.

Especially, in panel (e) we show that distributions of return intervals for typical assets of several total different financial markets, such as commodity, interest rate, stock, and currency can collapse onto a single curve. This remarkably consistency suggests that the scaling behavior is very robust over various financial markets, indicating that volatilities in financial markets share a certain universal mechanisms.

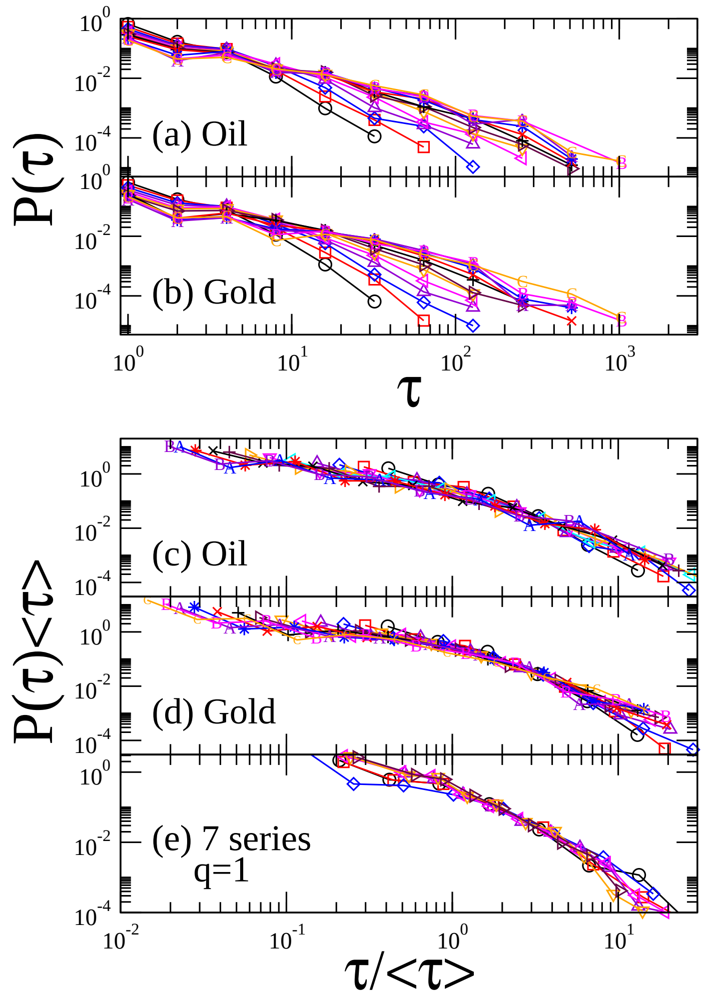

Figure 3.9: Distribution of the return intervals for volatility records of (a) oil commodity and (b) gold commodity, for 13 threshold values q from 0.6 to 3.0 ($\bigcirc$ 0.6, $\square$ 0.8, $\Diamond$ 1.0, $\triangle$ 1.2, $\triangleleft$ 1.4, $\bigtriangledown$ 1.6, $\triangleright$ 1.8, $+$ 2.0, $\times$ 2.2, $*$ 2.4, A 2.6, B 2.8 and C 3.0). Figures (c) and (d) show the scaled distribution $P(\tau)\langle\tau\rangle$ vs. scaled interval $\tau/\langle\tau\rangle$. Figure (e) shows the scaled plots of 7 financial time series of commodity, interest rate, stock and currency for $q = 1.0$ ($\bigcirc$ oil, $\square$ gold, $\Diamond$ federal funds rate, $\triangle$ IBM, $\triangleleft$ S&P500, $\bigtriangledown$ USD vs. JPY and $\triangleright$ UK Pound vs. Swiss Franc).

Recently many studies on return intervals were conducted. These studies confirmed that the scaling is also valid for other important markets, such as the Japanese market, a typical mature market, and the Chinese market, a prominent emerging market. Jung et al. analyzed the intraday data for 1817 stocks (1 year) and daily data for 3 typical companies (28 years) from the Japanese market [31]. They showed similar results as that of the US markets. For the Chinese market, 2 indices and 30 liquid stocks (both 2.5 years) were investigated, their behavior is also consistent with that we found for the US markets [35–37].

3.4 Origin of Stretched Exponential Distribution

We have shown that the impressive universality of the stretched exponential distribution of return intervals. Now we try to find the origin of this kind of distribution. From the generation of return intervals, their distributions are closely related to the temporal structure in volatilities. It is reasonable to assume the scaling in return interval distribution comes from correlations in volatilities. To test this assumption and find the origin of the scaling behavior, we shuffle the volatility time series which destroys the time organization in the time series, and then examine the corresponding return intervals. For uncorrelated data, we expect an exponential distribution from Poisson statistics.

In both Fig. 3.6 and Fig. 3.8, we plot return interval distributions for shuffled records. Indeed, we find that both of them follow exponential functions, as shown by the solid lines in these figures. We also note that all these curves collapse onto a single one for different sampling times or stocks, suggesting the scaling. In contrast to the distribution for uncorrelated records, the distribution of the actual return intervals (filled symbols in Fig. 3.6 and upper symbols in Fig. 3.8) is more frequent for both small and large intervals, and less frequent in intermediate intervals. The distinct difference between the distributions of return intervals in the original data and shuffled records suggests that the scaling behavior and the form in Eq. (3.2) must arise from long-term correlations in the volatility (see also [19]). Moreover, the relation between Hurst exponent α for correlations in volatilities and correlation exponent γ is suggested to be [19]

$$\alpha = 1 - \gamma/2. \tag{3.11}$$

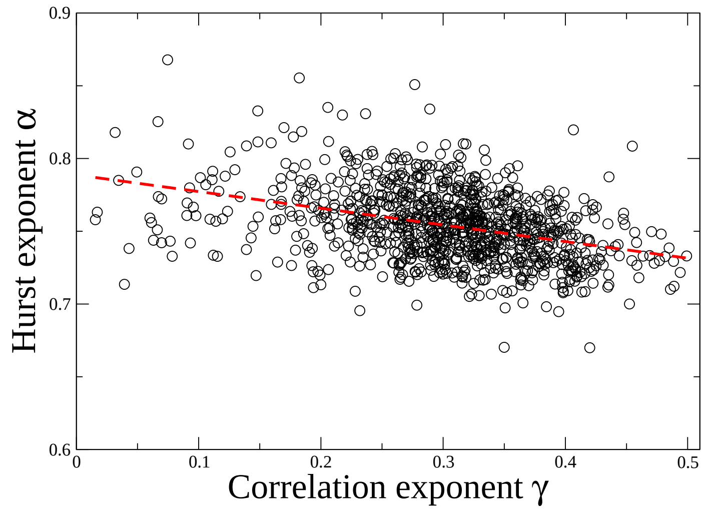

Figure 3.10: Dependence of Hurst exponent α on correlation exponent γ for the 1000 most traded U.S. stocks. α is obtained from the whole time scale of DFA curve, and γ is obtained from the stretched exponential fit on the cumulative distribution of return intervals. Clearly, there is a certain dependence between these two exponents, suggesting that the distribution shape of return intervals is related to the correlation in volatilities.

To test the validity of Eq. (3.11), we examine the 1000 most traded U.S. stocks and plot the scatter plot of α vs. γ in Fig. 3.10. We find a certain dependence between the two exponents,

$$\alpha = 0.78 - 0.11\,\gamma, \tag{3.12}$$

which does not perfectly satisfy Eq. (3.11). However, as we know, the correlation in volatilities is very complicated, and is not uniscaling but multiscaling (For example, the DFA curve for volatilities has a crossover but not a straight line on a log-log scale). On the other hand, Eq. (3.11) assumes that the correlation is uniscaling. Eq. (3.12) shows a significant dependence between α and γ which allows us to do detail analysis. Therefore, return intervals do provide a useful approach to analyze volatilities and the analysis on return intervals reveals underlined information.

Chapter 4

Further Analysis on Return Interval Distribution

In the previous chapter we showed that the return interval distribution has well-approximated scaling which is impressively universal for a wide range of thresholds and sampling intervals, as well as many stocks and financial markets. However, financial time series are known to show complex behavior and are not of uniscaling nature [75] and have non-linear features [76]. It is of great interest to test if there are some systematic deviations from the scaling for return interval distributions. If there are, what is the origin for such deviations? In this chapter we try to answer these questions.

4.1 Systematic Deviation from a Single Scaling Law

Though the scaling in the return intervals distribution is a good approximation, we find slight deviations. For example, in Fig. 3.1 we can see that the difference is relatively large in the head and tail regimes of the distribution. Compared to PDF, CDF accumulates data points and thus it integrates these deviations from the scaling law and provides a better statistics. By Eq. (3.2), we have

$$Q(\tau/\langle\tau\rangle) = \int_{\tau}^{\infty} P(\tau)\, d\tau = \int_{\tau/\langle\tau\rangle}^{\infty} f(x)\, dx. \tag{4.1}$$

If the scaling function $f(\tau/\langle\tau\rangle)$ is valid, the cumulative distributions should also collapse to a single curve. Otherwise, the cumulative distributions may show clearer deviations from scaling. Therefore we first examine the CDFs to see if there are any systematic deviations.

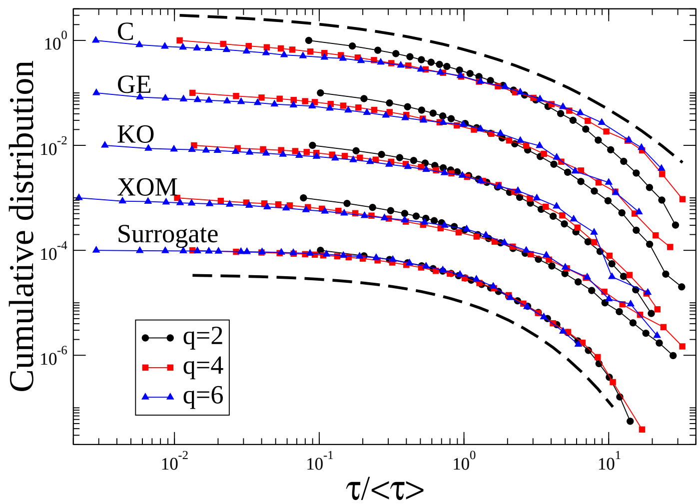

Figure 4.1: Cumulative distribution of the scaled intervals $\tau/\langle\tau\rangle$ for four typical stocks, C, GE, KO, XOM, and GE's surrogate. Symbols are for three representative thresholds $q = 2$ (circles), 4 (squares) and 6 (triangles) respectively. As two examples, we fit the cumulative distribution of stock C ($q = 2$) and the surrogate ($q = 2$) to a stretched exponential distribution [Eq. (3.4)] with exponent $\gamma = 0.25$ and $\gamma = 0.50$ correspondingly. Except that for stock C, all symbols and curves are vertically shifted for better visibility. For the surrogate records (see Sec. 4.4), symbols collapse almost perfectly onto one curve. However, all original data exhibit similar systematic deviations from the scaling. This suggests that non-linear correlations in the volatility series affect the scaling of its return intervals.

To explicitly explore the quality of the scaling in return interval distributions, we study all S&P 500 constituents (from TAQ database) and show the results of four representative stocks, namely, Citigroup (C), General Electric (GE), Coca Cola (KO) and Exxon Mobil (XOM). Most stocks studied here showed similar features. Indeed, in Fig. 4.1 we show cumulative distributions for three thresholds $q = 2$, 4 and 6. Note that the volatility is normalized by its standard deviation, the threshold q is in units of standard deviations and therefore $q = 6$ is a quite large volatility. It is clearly seen that those distributions are close to each other but do not collapse to a single curve. More important, they show apparent deviations from the scaling, which are systematic with the threshold. For small scaled intervals ($\tau/\langle\tau\rangle < 1$), the cumulative distribution decreases with q, while for large scaled intervals ($\tau/\langle\tau\rangle > 1$), it increases with q. Note that for very large scaled intervals, $\tau/\langle\tau\rangle \gg 1$, the curves have apparent fluctuations, which can not be trusted as much as that of smaller scaled intervals, due to poor statistics. In summary, the scaled interval prefers to be larger for higher threshold. This systematic trend suggests multiscaling in the return intervals.

4.2 Multiscaling in Return Interval Distribution

The systematic deviation of cumulative distributions indicates that the correlation exponent γ changes with the threshold q. To test the multiscaling behavior quantitatively and over the entire market, we compute γ for the 1137 most traded U.S. stocks and plot in Fig. 4.2 their averages and standard deviations (as error bars) as a function of q. The values of γ are obtained from the least-squares fit of the scaling function, Eq. (3.6), to the data for the range $\tau/\langle\tau\rangle \geq 0.1$. The range of q studied is from 1 to 6 with steps of 0.25. We consider a point as an outlier if its rms error is larger than 10%. In total 730 out of 22740 points or 3.8% of all points are removed. Figure 4.2 shows that the mean γ decreases with the increasing of q, from 0.49 for $q = 1$ to 0.28 for $q = 3$. For large thresholds (between $q = 3$ and 6) γ tends to be constant (around 0.26), where the distribution can be regarded as close to being of uniscaling nature. The difference in γ between small and large thresholds suggests multiscaling in the distribution for the whole range. The error bars shown in Fig. 4.2

are limited for all thresholds, indicating a consistent tendency for the entire market. Note that error bars for several of the largest thresholds such as $q = 6$ and 5.75 are slightly larger, probably due to the bad statistics of fewer events.

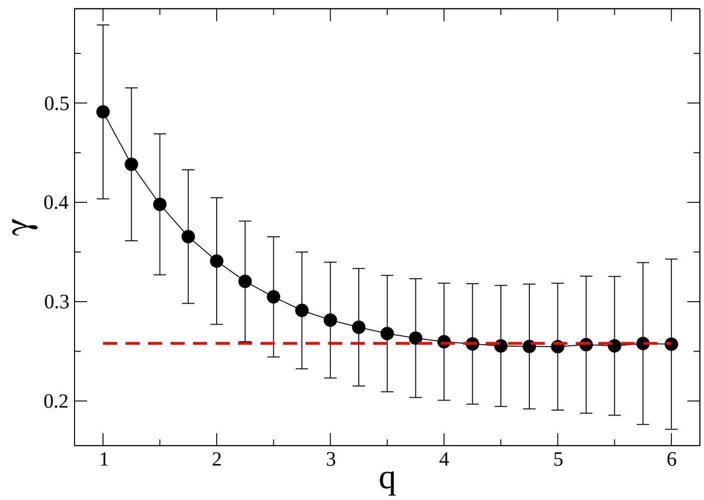

Figure 4.2: Correlation exponent γ vs. threshold q. Filled circles are values of γ averaged over 1137 most traded U.S. stocks and error bars are corresponding standard deviations. The average γ value converges to 0.26, as guided by the dash line.

The significant discrepancy between average γ values for small thresholds and large thresholds supports the multiscaling nature of return intervals. Another method for testing the multiscaling is using KS statistics (see Appendix B) to compare return interval distributions of two thresholds. If KS statistics $D > CV$, the null hypothesis that two distributions are same is rejected. For the Japanese market, Jung et al. have shown a good scaling by the KS test [31]. However, for the Chinese market, Ren and Zhou found that the null hypothesis is not rejected only for 12 of 30 liquid stocks. For other 18 stocks, the distributions are significantly different for different thresholds and therefore they don't obey a single scaling law [37].

4.3 Moment of Return Intervals and Multiscaling Characterization

4.3.1 Moment of Return Intervals

To further exhibit and characterize the multiscaling behavior, we utilize the moment to analyze return intervals. From Eq. (3.2), the variable for scaling function f is the scaled return interval $\tau/\langle\tau\rangle$ (assuming $x \equiv \tau/\langle\tau\rangle$ is continuous), thus we define the moment of return intervals, μ_m, as

$$\mu_m \equiv \left\langle \left(\frac{\tau}{\langle\tau\rangle} \right)^m \right\rangle^{1/m} \tag{4.2}$$

where m is the order of moment. From this definition, the moment accumulates the information over the entire data set and therefore obtains good statistics. As we know, a return interval time series is generated from a certain threshold q that we choose, and one q value corresponds to one $\langle\tau\rangle$ value for a volatility time series. Thus we can use $\langle\tau\rangle$ to characterize a return interval time series. Interestingly, the scaling function f is not directly related to $\langle\tau\rangle$. The relation between μ_m and $\langle\tau\rangle$ provides a good way to test the multiscaling in return intervals. If the scaling in Eq. 3.2 is valid, we have

$$\begin{aligned} \mu_m &= \left\{ \int_0^\infty \left(\frac{\tau}{\langle\tau\rangle} \right)^m \frac{1}{\langle\tau\rangle} f\left(\frac{\tau}{\langle\tau\rangle} \right) d\tau \right\}^{1/m} \\ &= \left\{ \int_0^\infty x^m f(x)\, dx \right\}^{1/m}, \end{aligned} \tag{4.3}$$

i.e., pure scaling of scaled return intervals yields that μ_m is independent of $\langle\tau\rangle$, regardless of the form of scaling function f. In other words, the significant dependence between μ_m and $\langle\tau\rangle$ (if exists) suggests multiscaling in the distribution of return intervals.

We analyze the moment for the 500 constituent stocks of the S&P 500 index, which allows us to examine the multiscaling for a broad market. In Fig. 4.3 we plot

μ_m vs. $\langle\tau\rangle$ of four typical stocks, C, GE, KO, and XOM, for four representative orders, $m = 0.25$, 0.5, 2, and 4.

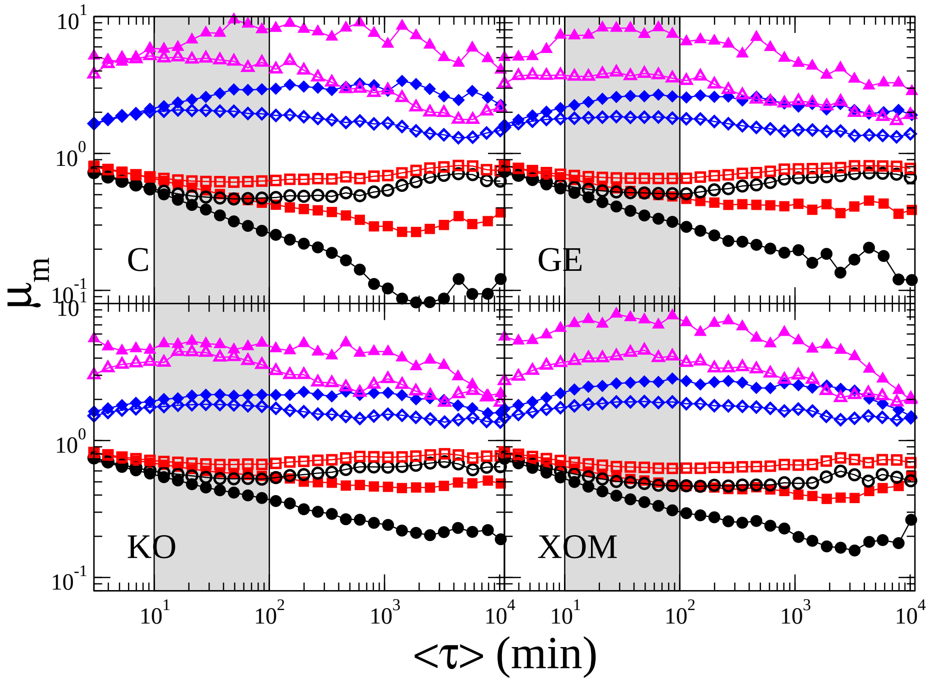

Figure 4.3: Moment μ_m for the scaled intervals of stock C, GE, KO and XOM. We show results from the original volatility series (filled symbols) and their surrogate (empty symbols, see Sec. 4.4). For each case, four moments, $m = 0.25$ (circles), 0.5 (squares), 2 (diamonds) and 4 (triangles) are demonstrated. Symbols for the original are clearly away from a horizontal line and their deviations are much larger than that for the surrogate, therefore the non-linear correlations in the original volatility are related to those deviations. To avoid effects from the resolution and size limit, we choose the shadow area, $10 < \langle\tau\rangle \leq 100$ (corresponds to $1.6 < q \leq 4.2$ for the 500 stocks that belongs to S&P 500 index), to study the multiscaling behavior.

Interestingly, we find that all curves in Fig. 4.3 are not exactly horizontal. More important, we show that these curves have consistent tendencies. For $m < 1$ cases, all of them have convex shapes, i.e., μ_m decreases with $\langle\tau\rangle$ first and then increases. On the contrary, all $m > 1$ cases have concave shapes. In the later parts of this section we show that these tendencies are far more than effects of discreteness and finite sizes, suggesting the multiscaling nature of return intervals. We find similar behaviors for other stocks that belongs to S&P 500 index.

4.3.2 Multiscaling Exponent

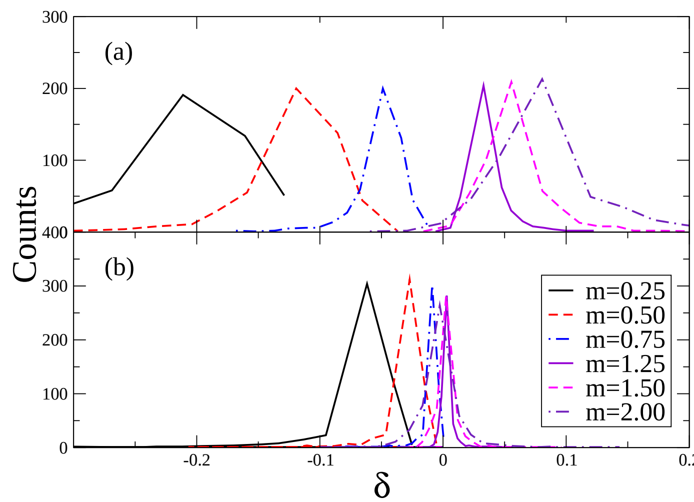

Figure 4.4: Distribution of multiscaling exponent δ for S&P 500 constituents. The exponent δ is obtained from the power-law fitting on moments in the medium range $10 < \langle\tau\rangle \leq 100$. (a) Histogram of δ for the original volatility and (b) for surrogate. The distributions have a systematic shift with m in (a) while all of them almost collapse in (b). This suggests a multiscaling behavior in the return intervals in the original records. The significant discrepancy between the original and the surrogate records manifests that non-linear correlations form the original volatility accounts for the multiscaling behavior.

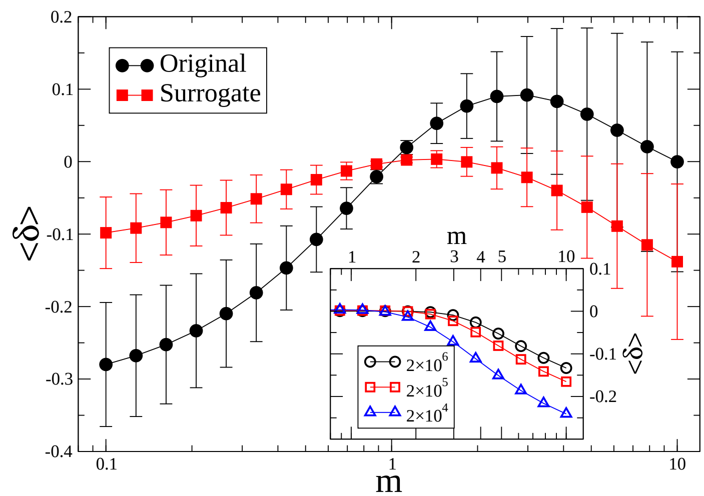

Figure 4.5: Dependence of average multiscaling exponent $\langle\delta\rangle$ on order m. The average $\langle\delta\rangle$ was taken over the 500 δ of the S&P 500 constituents. The error bars are standard deviations over the 500 δ. Results for the original (circles) and surrogate records (squares) are displayed. For large m, the two curves have the similar tendency which is attributed to the finite size effects. For small m, the two curves are significantly different which supports the multiscaling in the return intervals, due to the non-linear correlations in the original volatility. The inset demonstrates $\langle\delta\rangle$ averaged over 500 stretched exponential distributed i.i.d. simulations with $\gamma = 0.3$. Three sizes, 2×10^6, 2×10^5 and 2×10^4 are displayed, which clearly shows the finite size effect.

To quantify the tendency that we find in the moment of return intervals, we suggest to fit μ_m with a power law function in the medium range (to avoid the discreteness and finite size effects),

$$\mu_m \sim \langle\tau\rangle^{\delta}. \tag{4.4}$$

If the distribution of return intervals follows a scaling law, the exponent δ should be close to 0 (It does not exactly equal to 0 due to the discreteness and limited size of data sets, which will be discussed later in this section). In other words, a significant non-zero δ suggests multiscaling. Here we call δ the multiscaling exponent since it characterizes the multiscaling behavior. As demonstrated by the shadow regime in Fig. 4.3, we choose $10 < \langle \tau \rangle \leq 100$ as the medium range.

We calculate the δ value for each stock that belongs to S&P 500 index for orders between $m = 0.25$ and 2.00 and plot their histograms in Fig. 4.4. We can see that each histogram has a narrow distribution, which suggests that δ are similar for the 500 stocks. For the original records, almost all δ significantly differ from 0, thus the moments clearly depend on the mean interval. Moreover, the mean value of δ shifts with order m from $\langle \delta \rangle = -0.22 \pm 0.08$ for $m = 0.25$ to $\langle \delta \rangle = 0.08 \pm 0.05$ for $m = 2$ which means the dependence varies with the order m. This behavior supports the multiscaling nature in the return intervals distribution.

To remove fluctuations and show the tendency clearer, we plot the dependence of $\langle \delta \rangle$ on m, where $\langle \delta \rangle$ is the average δ over all 500 stocks belonging to S&P 500 index. In Fig. 4.5 we show this relation for a wide range of m, $0.1 \leq m \leq 10$. The plot shows two different behaviors. For small m (roughly $m \leq 2$), $\langle \delta \rangle$ clearly deviates from 0 and demonstrates the multiscaling behavior, while $\langle \delta \rangle$ for the surrogate is closer to 0. For large m ($m > 2$), the curve has a decreasing trend. Since large τ dominates high order moments, this decreasing trend may be due to finite size effects, which will be discussed in later subsection. Fig. 4.5 also shows error bars of $\langle \delta \rangle$, which is the standard deviation of δ values for the 500 stocks. Remarkably, we can see that the tendency is out of range of error bars, indicating the significance of the multiscaling behavior.

4.3.3 Relation between Moment and Order

The moment μ_m has systematic dependence on order m, as seen in Fig. 4.3 where moments are plotted against $\langle \tau \rangle$. It is of interest to explore the relation between μ_m and m directly. For a fixed $\langle \tau \rangle$, representing a given threshold q one can study, return intervals and their moments of various orders which exhibit information on different scales of τ. Moments of large m represent large τ and vice versa. If $\tau / \langle \tau \rangle$ follows a single distribution without corrections due to effects such as discreteness and finite size, curves of μ_m vs. m for different $\langle\tau\rangle$ should collapse to a single one, which only depends on the scaling function $f(x)$ as Eq. (4.3). We plot μ_m for m between 0.1 and 10 for three $\langle\tau\rangle$ values: 10, 80 and 400 min in Fig. 4.6(a) and find substantial deviations from a single curve. This supports our suggestion that the return intervals have multiscaling behavior.

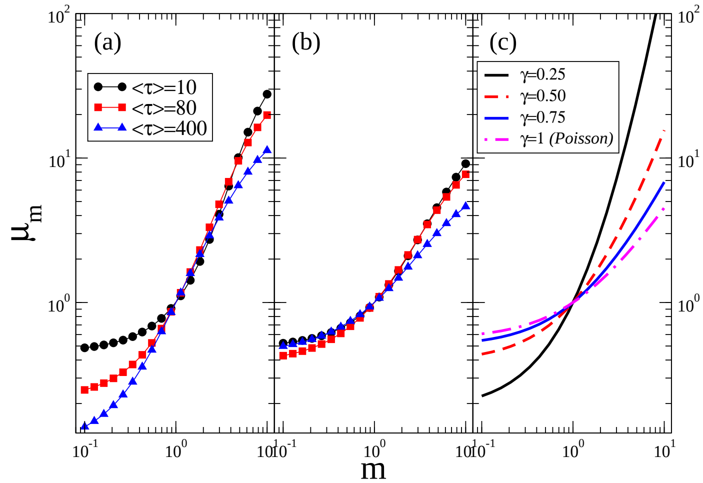

Figure 4.6: Dependence of the moment μ_m on the order m. (a) for the return intervals of the original volatility for stock GE. (b) for its surrogate. Three mean interval $\langle\tau\rangle = 10$ (circles), 80 (squares) and 400 min (triangles) are demonstrated in (a) and (b). (c) Analytical moments from stretched exponential distributions, $\gamma = 0.25$, 0.50, 0.75, and 1 (Poisson distribution), taken from Eq. (4.5). For large m, both (a) and (b) show discrepancies which is related to the finite size effect. For small m, the difference between (a) and (b) is due to the non-linear correlations in the original volatility. Compared to the analytical curves in (c), the return intervals for the original records shows multiscaling behavior.

As a reference, we also drive the analytical form for μ_m. Substituting Eq. (3.4) into Eq. (4.3), we obtain

$$\mu_m = \frac{1}{a} \left\{ \frac{\Gamma\left(\frac{m+1}{\gamma}\right)}{\Gamma\left(\frac{1}{\gamma}\right)} \right\}^{1/m} \tag{4.5}$$

for the stretched exponential scaling function. In Fig. 4.6(c) we show analytical curves for various correlation exponent γ. Clearly, these curves are similar to that in panel (a), confirming that return intervals have multiscaling behavior.

4.3.4 Discreteness and Finite Size Effects

Now we try to find the origins of the tendencies in the moments of return intervals. Beyond the multiscaling nature of return interval distribution, there might be certain common mechanisms for all time series. For very small and very large values of $\langle\tau\rangle$, we identify the discreteness and finite size effects respectively [29], which was also recognized for the general case by Eichner et al. [22].

First we look at the discreteness effect. Our data set is recorded in 1-min sampling intervals and the minimum difference in return intervals is 1 min. This discreteness causes a certain artificial tendency for small $\langle\tau\rangle$ ($\langle\tau\rangle \leq 10$). The relative errors in moments will be considerably larger for small $\langle\tau\rangle$ close to 1 min. By starting from $\langle\tau\rangle = 3$, we only partially avoid the discreteness in moments. To show this effect, we reduce the time resolution and compare the moments μ_m with 3 sampling intervals, 1, 5 and 10 min, since we can not increase the resolution. Fig. 4.7(a) shows this comparison for stock GE of $m = 0.5$ and $m = 2$. The three resolutions have a similar trend, showing that the curves become flatter for the higher resolution. For other stocks, we find similar behavior. This systematic tendency suggests that the recording limit (1 min) strongly affects the moments at short $\langle\tau\rangle$. To reduce it, we should raise the recording precision or study the moments of larger $\langle\tau\rangle$.

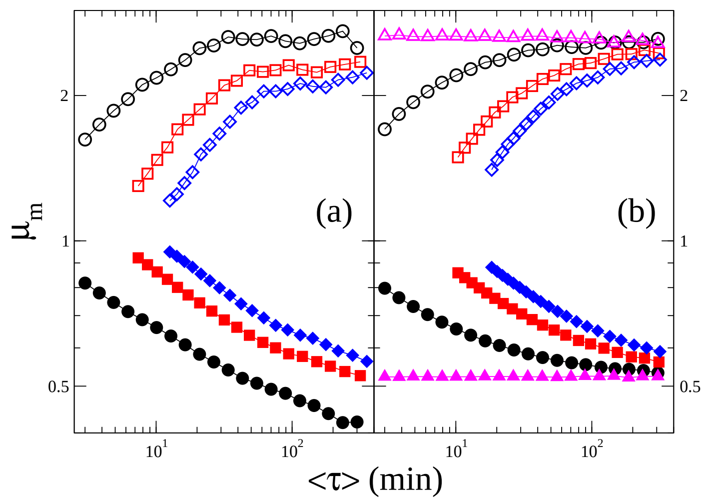

Figure 4.7: Discreteness effect in the moments. (a) Moments for stock GE with three resolutions, 1 (circles), 5 (squares) and 10 min (diamonds). Filled symbols are for $m = 0.5$ while empty symbols are for $m = 2$. (b) Moments of artificial records averaged over 100 simulations. For each trial, we simulate return intervals that follow a stretched exponential distribution with $\gamma = 0.3$ and a length of 200 thousands points. Symbols are similar to that in (a). Three resolutions, 1, 5 and 10 time units are displayed. To show the disappearing of the discreteness, we also plot simulation results for continuous return intervals (triangles), which exhibits a constant moment, independent of $\langle\tau\rangle$.

To further test this result, we simulate artificial return intervals with an i.i.d. process from stretched exponential distribution with $\gamma = 0.3$, same as the empirical γ for GE of $q = 2$. We examine moments of $m = 0.5$ and $m = 2$ with the same three resolutions (1, 5 and 10 time units) as in the empirical test done above. The simulated size is 200 thousands points for each trial and we use the average over 100 trials for each resolution. Fig. 4.7(b) shows curves similar to that of empirical [Fig. 4.7(a)]. For the higher resolution, the curve is closer to the horizontal line and finally may reach the line when we raise the resolution high enough. To show this, we also simulate continuous return intervals and find constant moments, as expected [Fig. 4.7(b)]. Therefore, the discreteness effects can be overcome if the resolution is improved enough. Note that for the empirical data, we expect the moment not to be constant for small $\langle\tau\rangle$ even if we have a much better resolution, since the return intervals has the multiscaling behavior, as shown for larger $\langle\tau\rangle$ in the range $10 < \langle\tau\rangle \leq 100$ which is not affected by discreteness.

Next we examine the finite size effect. To test this effect we simulate surrogate return intervals (see Sec. 4.4) by assuming a stretched exponential distribution i.i.d. process with 3 sizes (number of all return intervals in the time series), 2×10^6, 2×10^5 (the size of the empirical data set of stock GE), and 2×10^4. Without loss of generality, we choose $\gamma = 0.3$, which is the correlation exponent for GE of $q = 2$. To be consistent with the 500 S&P 500 stocks, we perform 500 realizations and plot their average exponent $\langle\delta\rangle$. As shown in the inset of Fig. 4.5, all $\langle\delta\rangle$ values show a similar decreasing trend as that of the empirical curves. However, it is seen that the trend starts earlier for smaller size, and thus the size limit has a strong influence on high order moments.

4.4 Origin of Multiscaling Behavior

To better understand the systematic trends and test if it is not due to finite size effect or discreteness of minutes, we also measure the cumulative distribution of return intervals for surrogate records of volatilities using the Schreiber method [77–79] where non-linearities are removed. For a given time series, we store the power spectrum and randomly shuffle the sequence, then we apply the following iterations. Each iteration consists of two consecutive steps:

(i) We perform the Fourier transform of the shuffled series, replace its power spectrum with the original one, then take the inverse Fourier transform to achieve a series. This step enforces the desired power spectrum to the series, while the distribution of volatilities usually is modified.

(ii) By ranking, we exchange the values of the resulting series from step (i) with that of the original record. The largest value in the resulting series is replaced by the largest one in the original series, the second largest value is replaced by the second largest one, and so on. This step restores the original distribution but now the power spectrum is changed.

To achieve the convergence to the desired power spectrum and distribution, we repeat these two steps 30 times. By this way, a "surrogate" series is generated. Because of the Wiener-Kinchine theorem [80], the surrogate record has the same linear correlations as the original, as well as the distribution. The only difference is that the original record has the non-linear correlations (if they exist) but the surrogate does not have any non-linear features.

In Fig. 4.1 we also plot the cumulative distribution for the surrogate with the same three thresholds as the original. Since the surrogate records lost the non-linear correlations, they are similar to each other, and we only show results for GE's surrogate. It is seen that the collapse of the surrogate for different q values is significantly better than that of the original and the deviation tendency with the threshold in the original records disappears. This indicates that the scaling deviations in the original are due to the non-linear correlations in the volatility. We also compare our results to the stretched exponential distribution (dashed lines in Fig. 4.1). This curve is very close to the empirical results, in particular for the surrogate records which contain only the linear correlations. This suggests that PDF of return intervals is well approximated by a stretched exponential.

To further test the non-linearity hypothesis, we analyze the moment for the surrogate records. As shown in Fig. 4.3, we find that their curves are more flat for most ranges. For the same order m, the moment of the surrogate obviously differs from that of the original, especially in the medium range of $\langle\tau\rangle$ ($10 < \langle\tau\rangle \leq 100$). This discrepancy suggests that the non-linear correlations exist in the original volatility and account for the scaling deviations. Nevertheless, all moments of surrogate show small curvature from a perfect straight line at both short and long $\langle\tau\rangle$, which are much weaker compared to the original records. The weak curvature suggests that some additional effects, not related to the non-linear correlations, affect the moments. For small $\langle\tau\rangle$, the resolution discrete limit seems to have some influence on the moments. For large $\langle\tau\rangle$, the moments are gradually approaching the horizontal line and are more fluctuating, the effect seems to be related to limited size of the record. On the other hand, for moments of high order such as $m = 4$, the trend in the medium range is not as obvious as for the lower order moments. This is probably to lack of statistics for high order moments.

Furthermore, in Fig. 4.4(b), the histograms of δ values for the surrogate records are more centered around values close to $\delta = 0$ (from $\langle\delta\rangle = -0.07 \pm 0.04$ for $m = 0.25$ to $\langle\delta\rangle = 0.00 \pm 0.02$ for $m = 2$). Moments for the surrogate records converge to a single curve for $m \leq 2$ but become diverse for high orders, which agrees with the strong influence of the finite size effects. The uniscaling behavior for the surrogate supports that the non-linear correlations in volatilities are responsible for the multiscaling behavior in their return intervals. We can find consistent behavior in Fig. 4.5. Note that both the original and surrogate records have similar decreasing tendency for large order m, as shown in Fig. 4.5. This similarity confirms the effect of finite size. In addition, Ren and Zhou also employed moment analysis on two Chinese indices [37] and confirmed the multiscaling behavior in the return interval distribution that we found.

4.5 Multifactor Analysis on Multiscaling

In the previous section, we showed that the multiscaling behavior in return intervals arises from the non-linearity in volatilities. The next question is, what kind of factors affect this behavior? As we known, financial markets are a typical kind of complex system which consists of various types of investors and activities. Therefore if we can relate the factor that characterizes the activity or fundamentals to the multiscaling behavior of return intervals, it will offer insights to deeply understand the volatility and price dynamics. Here we focus on four essential factors, market capitalization, return, risk, and number of trades. The data set used is the 1137 most traded stocks from TAQ database for 2001-2002 period.

Analysis with respect to several essential factors is widely used in economics studies. For instance, company size, market return, and book-to-market value are used to model asset pricing [81]. Volatilities and therefore return intervals may be affected by many factors. Here we study how the return interval distribution depends on a few essential measures which characterize different features of the stocks. The first one is the size of the company, which is a popular criterion for portfolio selection. Stocks of different scales are preferred by investors of different types. The size also limits the group of investors and market depth for a stock. On the other hand, the internal organization of a company might dramatically vary with its size. Thus, the volatility and its return interval may be strongly influenced by this factor. The size is usually characterized by the market capitalization, the product of the stock price and outstanding shares. Without loss of generality, we choose the price and outstanding shares on December 31, 2002 to calculate the capitalization. For the 1137 stocks, the range of capitalization is between 2×10^7 and 2×10^{11} dollars.

The reward and risk are basic concerns for any investment and we therefore choose them as the next two factors. The reward is usually measured as the average return of the price while risk is measured as the standard deviation of the return [82]. This traditional definition of the risk is based on the Gaussian distribution of the time series, which is not always adapted to the financial data [2]. Nevertheless, it characterizes the magnitude of fluctuations and therefore the risk. To avoid the intraday pattern [26, 60], we calculate the return on a daily basis. The return is the logarithmic daily price change averaged over the two-year period (2001-2002), which varies from -0.008 to 0.004 for the 1137 stocks. The risk, the standard deviation of daily returns in the two years, ranges from 0.012 to 0.12. The fourth factor we study is an activity measure, the number of trades per day. Note that the four factors reflect different aspects of a stock. The size represents the scale of the company. The return and risk are the historical price movement tendency and variation, which are helpful for the prediction of future price change. The number of trades shows the activity, i.e., how frequently a stock is traded.

4.5.1 Correlation Exponent γ and Factors

First we study the relations between γ (correlation exponent for volatilities, obtained from DFA method, see Appendix A) and the four essential factors market capitalization, risk, number of trades, and return. This tests the universality of γ over the entire market. If γ is sensitive to some factors, the market as one system is not of uniscaling. Furthermore, the dependence (if it exists) may indicate some origins for the multiscaling found in return interval distributions. In Fig. 4.8, we plot γ against the four factors for four thresholds $q = 2$, 3, 4, and 5. In each panel, the curves have a similar tendency and the value of γ decreases with increasing q. Note that the curves are closer to each other for large thresholds. This finding is consistent with Fig. 4.2, which shows that the mean γ decreases with q and reaches almost a constant value for large q. More important, Fig. 4.8 exhibits that γ for a given threshold is not uniformly distributed with the factors and thus the market has a multiscaling nature.

For the company size [Fig. 4.8(a)], γ increases for sizes between 5×10^7 and 2×10^{10} dollars and then shows a slight decrease. The market depth for small companies limits the number of investors and those companies usually attract some specific types of investors. Therefore corresponding strategies may be relatively similar and the volatility series tends to be strongly correlated, having a small γ [18]. With increasing size, more investors are involved, which may "randomize" the long-term correlations in volatilities. When the company size reaches a certain limit, the constitution of investor types may be relatively stable, some common modes might dominate the volatilities, and therefore the correlations become stronger and γ decreases with increasing size.

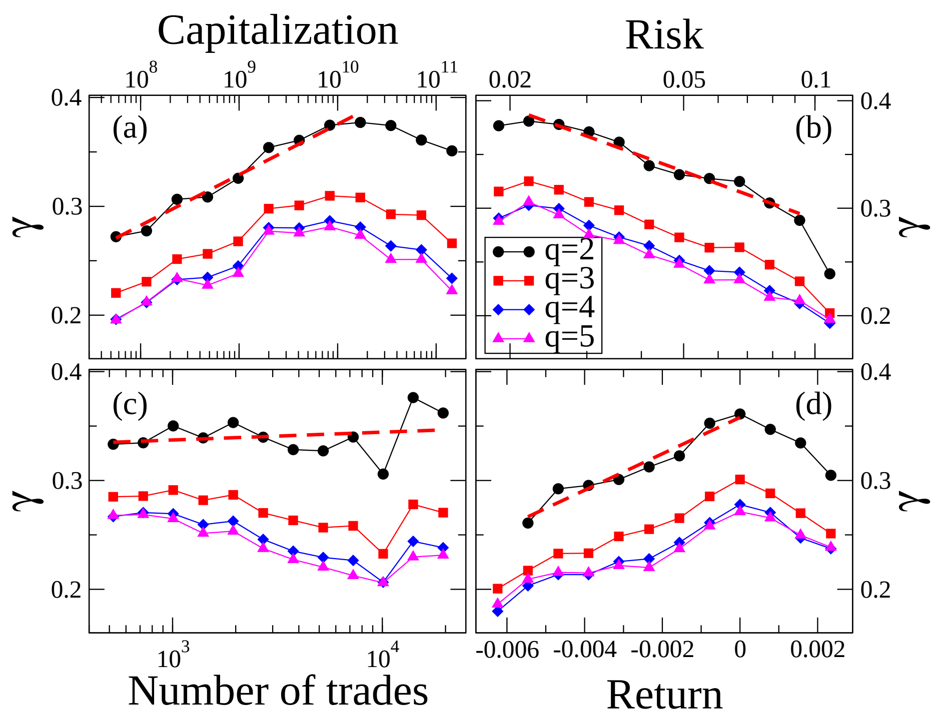

Figure 4.8: Relation between correlation exponent γ and four factors: (a) market capitalization, (b) risk, the standard deviation of daily return, (c) average daily number of trades, and (d) average daily return. Curves of four thresholds $q = 2$, 3, 4 and 5 are demonstrated. Dashed lines are logarithmic fits [except in (d) where the fit is linear] to the curve of $q = 2$.

Figure 4.8(b) shows that γ decreases with increasing risk except for very low risks. Figure 4.2 shows that larger volatilities tend to have smaller γ. Larger risk means that the probability of larger volatilities is higher. Therefore, Fig. 4.8(b) is consistent with Fig. 4.2. Price movement is realized by trades and the temporal structure of the volatility probably relates to the size of trades. Counter-intuitively, γ is almost insensitive to the market activity. Figure 4.8(c) suggests no apparent dependence between γ and the number of trades; see also [83]. A possible reason is that many investors do not change their strategies just because of a dramatic change of trading frequency. Note that there is a slight tendency in Fig. 4.8(c), which is negligible if we compare it with the trends in other panels. As an example, the logarithmic fit to the curve for $q = 2$ (shown by dashed lines in Fig. 4.8) has a slope of 0.06 for the capitalization, -0.19 for the risk, but only 0.01 for the number of trades. Similar results are shown in Fig. 4.9. Next we show in Fig. 4.8(d) the relation between γ and the return. For negative returns γ increases, and then decreases for positive returns. It has a maximum when the return is 0. This behavior suggests that the return is related to the size of risk. For returns with large magnitude representing high volatilities, the corresponding risk is relatively high and therefore γ is small [see Fig. 4.8(b)].

4.5.2 Multiscaling Exponent δ and Factors

To characterize the multiscaling strength in a single stock, we suggested a multiscaling exponent δ in previous section. Now we study its dependence on the four essential factors. We plot in Fig. 4.9 the curves for four orders $m = 2$, 4, 8, and 16. These curves have a similar tendency in each panel. The value of δ increases a little from $m = 2$ to 4, then decreases, which is consistent with the result in Ref [29]. As shown in Fig. 4.9(a), δ decreases with increasing capitalization until about 2×10^{10} dollars and then the curves increase. This suggests that δ is also related to the constitution of investors. A small company has few investors who have some specific strategies. However, if the company is very large, some types of investors finally dominate the price movement. In Fig. 4.9(b), δ increases almost monotonically with the risk, indicating that, if a stock has larger volatility values, its return interval distribution has a stronger multiscaling effect. Similar to γ, δ is almost independent of the number of trades as shown in Fig. 4.9(c). In Fig. 4.9(d), δ has a minimum at zero returns, which also agrees with the relation between δ and risk.

There are clear connections between Figs. 4.8 and 4.9, which indicate that γ and

δ are strongly related. From Fig. 4.2, γ decreases with q, one q value corresponds to one $\langle\tau\rangle$ value, and δ is the power law fit exponent for the moment vs $\langle\tau\rangle$.

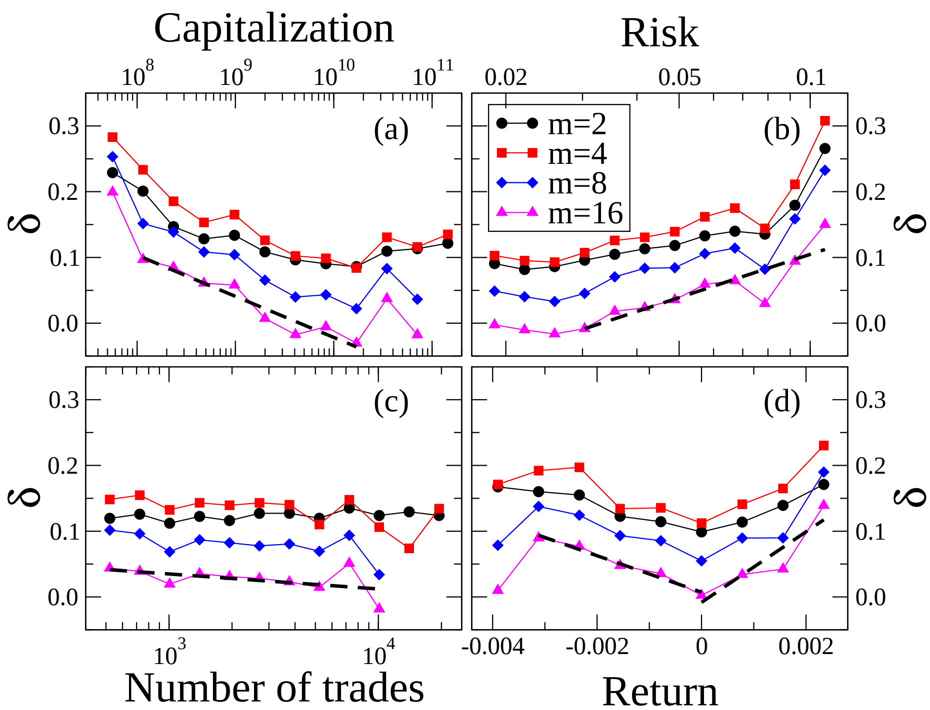

Figure 4.9: Relation between multiscaling exponent δ and four factors: (a) market capitalization, (b) risk, the standard deviation of daily return, (c) average daily number of trades, and (d) average daily return. Curves of four moments $m = 2$, 4, 8, and 16 are shown. Dashed lines are fits to the curve of $m = 16$, which demonstrate the tendency.

To examine the relation between the two exponents we plot δ against γ in Fig. 4.10. Our results suggest that δ decreases with increasing γ for all four thresholds $q = 2$, 3, 4, and 5 when $m = 2$. These curves approximately follow a linear function as shown by the dashed lines with slopes -0.63, -0.75, -0.74, and -0.62 respectively. Other thresholds q and orders m show similar results. The smaller the value of γ, the larger the deviation from a single scaling function observed for the return interval distribution.

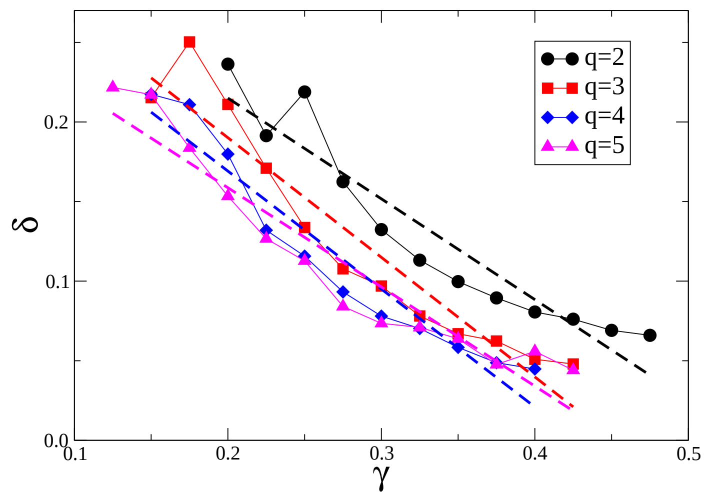

Figure 4.10: Multiscaling exponent δ vs correlation exponent γ. The values of δ are for order $m = 2$ and γ are for thresholds $q = 2$, 3, 4, and 5. Linear fits are shown by dashed lines.

Chapter 5

Memory in Return Intervals

Scaling and universality are important properties of a data set describing the global behavior of the probability distribution. This may, or may not, fully characterize a sequence of data points, depending on the time organization of the sequence. If it is *uncorrelated*, data points are independent of each other and totally determined by the probability distribution. On the other hand, if the points are *correlated*, it will also affect the order in the data set. This behavior is also called "memory", as the data points "remember" previous values.

Many studies showed that returns do not exhibit any linear correlations extending over more than a couple of minutes, but their absolute value, which is a measure for volatility, exhibits strong correlations. This leads to long periods of high volatility as well as other periods where the volatility is low. This effect is known as volatility clustering. We find similar effects also for return intervals, so that large (small) return intervals are more likely to be followed by large (small) return intervals. In this chapter we analyze in detail the memory effect in the return interval time series. We employ the conditional distribution to examine short-term memory, cluster analysis to test medium-term memory, and DFA method to analyze long-term memory. We find significant memory effects for all time scales. Moreover, we show a strong similarity between long-term correlations in return intervals and long-term correlations in volatilities.

5.1 Short-Term Memory

5.1.1 Conditional Distribution of Return Intervals

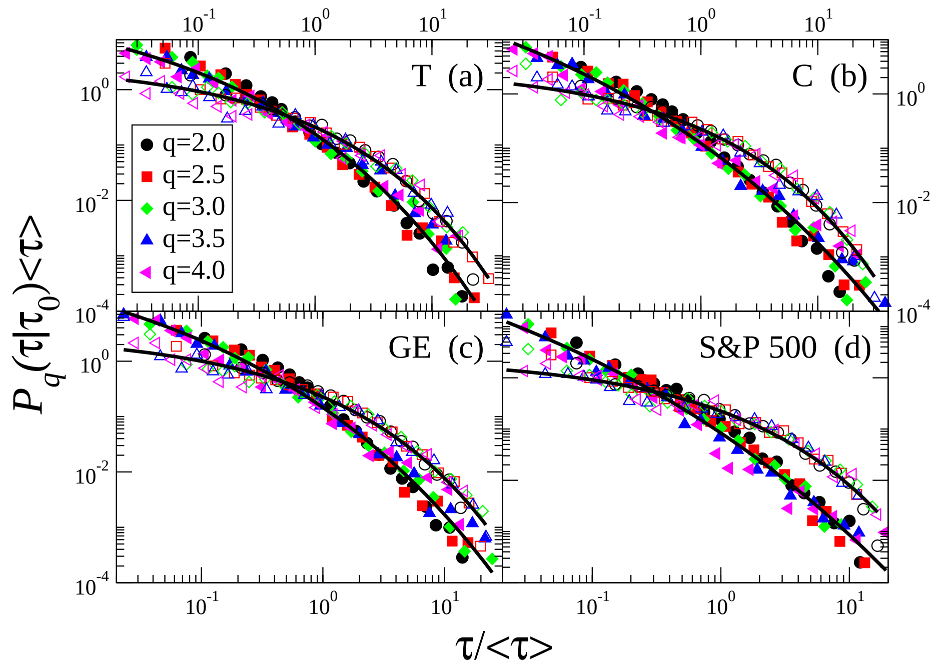

Figure 5.1: Scaled conditional distribution $P_q(\tau|\tau_0)\,\langle\tau\rangle$ vs. $\tau/\langle\tau\rangle$ for (a) T, (b) C, (c) GE, and (d) S&P 500. Here τ_0 represents a subset which contains 1/8 of the total number of return intervals in increasing order. The lowest 1/8 subset (filled symbols, Q_1) and the largest 1/8 subset (open symbols, Q_8) are displayed, which clearly have different tendencies, as guided by black curves. Symbols are plotted for different thresholds, denoted in panel (a). The significant difference between two types of conditional distributions suggests the existence of memory. Small intervals tend to follow small ones and large intervals tend to follow large ones. The solid curves on the symbols are stretched exponential fittings.

First we analyze the memory in a short-term time scale. This type of memory can be measured by the conditional distribution, $P(\tau|\tau_0)$, which is the probability of finding a return interval τ immediately after a return interval of size τ_0 [19, 20, 22, 24, 26, 27]. In records without memory, $P(\tau|\tau_0)$ should be identical to $P(\tau)$ and independent of τ_0. When memory exists, it should depends on the choice of τ_0.

Due to the poor statistics for a single value of return interval, a binning of τ_0 is needed. We split the entire database into 8 equal-size subsets ("octave"), Q_1, Q_2, ..., Q_8, with intervals in increasing length [24, 26, 27]. It is found that for τ_0 in Q_1, the probability is higher for small τ, while for τ_0 in Q_8, the probability is higher for large τ. Thus, large (small) τ_0 tends to be followed by large (small) τ ("clustering"), which indicates memory in the return interval time series. In Fig. 5.1 we show the results for three typical stocks and one benchmark index, namely, T, C, GE, and S&P 500, and found consistent behaviors for all four cases.

Remarkably, similar memory effects are found for other stocks in the US markets, currencies, interest rates, and commodities [24, 26, 27], as well as stocks for the Japanese market [31] and Chinese market [35, 36]. In conclusion, the short-term memory in return intervals is universal for financial fluctuations. In addition, we can see in Fig. 5.1 that $P(\tau|\tau_0)$ seems to collapse onto a single curve for each of the τ_0 subsets, and these curves can be well fit by a SE function according to Eq. (3.4), which further confirms features of return interval distributions in previous chapters.

5.1.2 Mean Conditional Interval

The conditional distribution shows significant deviations for cases of different bins. Though it clearly shows the short-term memory, the method is qualitative. To show the effect more directly and quantitatively, we examine the mean conditional return interval immediately after a given τ_0 subset, $\langle\tau|\tau_0\rangle$, which is the first moment of $P(\tau|\tau_0)$. A power law dependence of $\langle\tau|\tau_0\rangle$ on τ_0 for the GE stock is showed in Fig. 5.2, as an example. We can see that large (small) τ tend to follow large (small) τ_0, similar to the clustering in $P(\tau|\tau_0)$. Correspondingly, shuffled data (open symbols in Fig. 5.2) are almost constant as expected, demonstrating that the value of τ is independent of the previous interval τ_0. The "strength" of memory can be measured by the exponent of the power-law fitting. Clearly, we can expect that different stocks might have different exponents, and we can use this information to predict the the next move of price, which will be further discussed in Chapter 6.

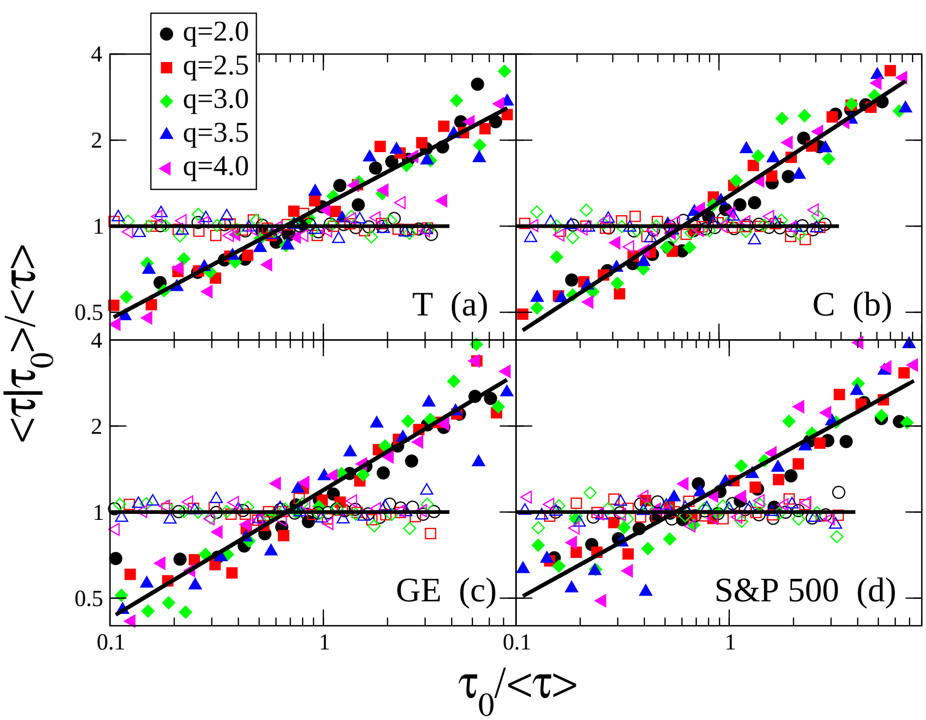

Figure 5.2: Scaled mean conditional return interval $\langle\tau|\tau_0\rangle/\langle\tau\rangle$ vs. $\tau_0/\langle\tau\rangle$ for (a) T, (b) C, (c) GE, and (d) S&P 500. That for original returns intervals (filled symbols) and shuffled records (open symbols) are plotted. Five thresholds, $q = 2.0, 2.5, 3.0, 3.5$, and 4.0 are represented by different symbols, as shown in panel (a). The distinct difference between actual intervals and shuffled records implies memory in the original return intervals.

5.2 Cluster Size Distribution

Clustering phenomena are displayed by $P(\tau|\tau_0)$ and $\langle\tau|\tau_0\rangle$, indicating the memory in return intervals. However, both functions measure the intervals that immediately follow an interval bin τ_0. In order to investigate longer clustering in a more straightforward way, we analyze "clusters" of return intervals, which are composed by successive intervals with similar size [26, 27, 31, 32, 35]. To obtain good statistics we divide return intervals into two groups, separated by the median of the entire time series. We denote intervals that are above the median by sign "+", and the ones below the median by "–". Accordingly, consecutive "+" or "–" intervals form a positive or negative cluster.

Obviously, these positive and negative clusters contain the memory information in return intervals. To quantitatively show the effect with good statistics, we examine the cumulative distribution of cluster size n. If the time series is in a random order, the cluster size distribution should follow an exponential function, by Poisson statistics. On the other hand, if there is a certain memory in the time series, similar intervals tend to be together and therefore the probability of large clusters is relatively high. Fig. 5.3 shows the cumulative distribution of n for a typical stock, namely GE. Both positive and negative clusters have quite long tails, compared to that for the shuffled records which follows an exponential function and shows a much faster decay. For the positive clusters, the distribution still has good statistics even for size $n = 18$, while the negative clusters extend to $n = 25$. Thus, the memory effects persist for quite long times (e.g., the average return interval for GE with threshold $q = 2$ is about 9 min, so there are still some clusters corresponding to even 200 min in the time scale). All these behaviors suggest that a strong memory effects exist in return intervals. Note that the distribution of positive clusters is very similar for different thresholds $q = 2, 3, 4$, while the negative clusters show the same effect only for $n \leq 10$. Similar clustering has been found also in earthquake and climate data [20, 21].

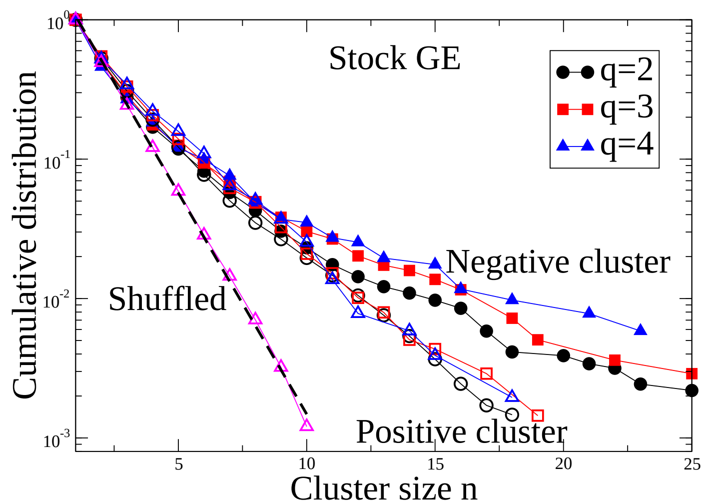

Figure 5.3: Cumulative distribution of size for return interval clusters. The cluster consists of consecutive return intervals that are all above ("positive cluster", open symbols) or below ("negative cluster", filled symbols) the median of time series. To reveal the memory effects, we plot the distribution for both original and shuffled records. For the shuffled records, the distribution follows an exponential function. However, for the original records, distributions for both positive and negative clusters have much longer tails, suggesting a significant memory in return intervals.

5.3 Long-Term Correlations

Now we focus on the long-term correlations in return intervals. Due to the similar reasons that we explained in previous chapters, we employ the DFA method (see Appendix A) to examine long-term correlations in return interval time series. Without loss of generality, we investigate the return interval for a typical threshold $q = 2$. The fluctuation function vs. window size ℓ is shown by open squares in Fig. 5.4. We can see a crossover in the curve. Thus we can divide the entire regime into two sub-regimes $\ell < \ell^*$ and $\ell > \ell^*$ (ℓ^* can be chosen to be 390 min or 1 trading day. Note that the mean interval for $q = 2$ is around 10 min.) [26]. The corresponding power-law fitting (solid lines in Fig. 5.4) and therefore α values are distinctly different in the two regimes. However, both α values are significantly larger than 0.5, $\alpha = 0.66$ for the short scale and $\alpha = 0.84$ for the long scale, suggesting long-term correlations in return intervals although the long-term correlations are multiscaling.

Interestingly, the above behaviors are universal for return intervals in different stocks and markets. We examine the 1000 most traded stocks in the U.S. markets between 2001 and 2002 and find that their α values are evenly distributed in a narrow range, $\alpha = 0.64 \pm 0.04$ for the short scale ($\ell < \ell^*$) and $\alpha = 0.80 \pm 0.06$ for the long scale ($\ell > \ell^*$). Here the error bar is the standard deviation of α values for the 1000 stocks. This narrow distribution of α values can be observed in both Fig. 5.5 and Fig. 5.6.

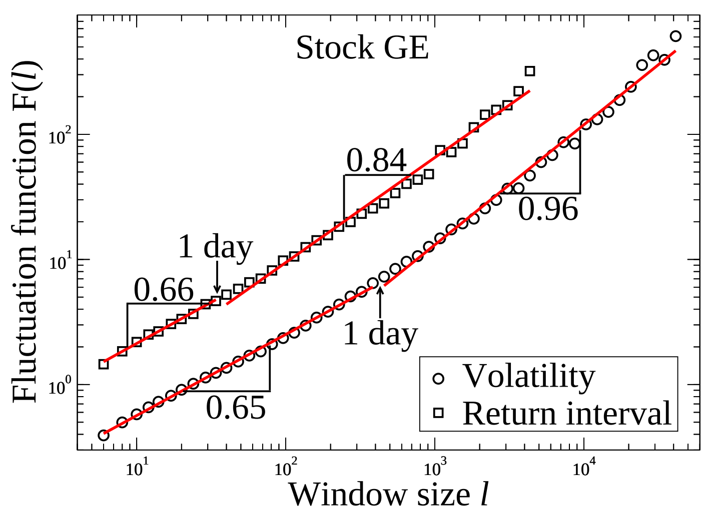

Figure 5.4: DFA curve of volatility and return interval ($q = 2$) for the GE stock. Two curves are similar and both of them have crossovers around 1 trading day. Solid lines are power-law fits on the two regimes. For short scale, the two α values (slopes in the plot) are almost the same. For the long scale, the two α values are different but they are strongly related.

5.4 Relation between Return Interval Correlations and Volatility Correlations

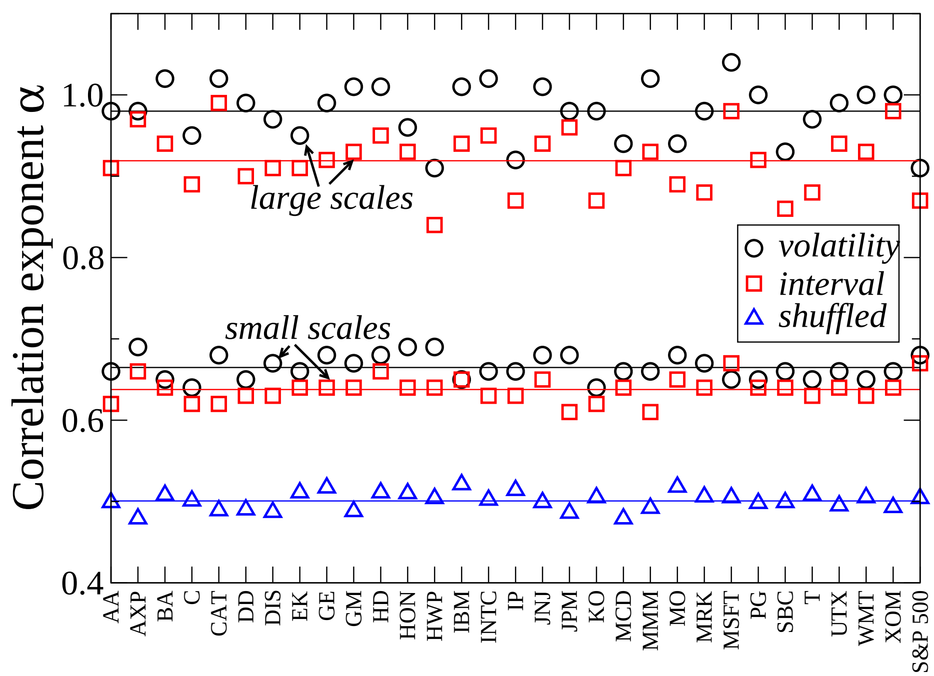

Figure 5.5: Correlation exponent α of volatilities and return intervals. The fluctuation $F(\ell)$ vs. ℓ curves are fitted in two regimes, small scales ($\lesssim 1$ day) and large scales ($\gtrsim 1$ day), thus we obtain two α values for each data set. The correlation exponent α for 30 DJIA stocks and for the S&P 500 index (stocks are identified by their tick symbols, e.g., AA denotes Alcoa Inc.) are shown. Note that most data sets have a smaller exponent for intervals than for volatilities, but their differences still are in the range of the error bars. Shuffled records (triangles) yield α values around 0.5 that indicate no correlation. Both small scales ($\alpha_1 = 0.66 \pm 0.01$ and $\alpha_1 = 0.64 \pm 0.02$ for volatilities and return intervals respectively) and large scales ($\alpha_2 = 0.98 \pm 0.04$ and $\alpha_2 = 0.92 \pm 0.04$ correspondingly) appear to show different correlations for different scales, since $\alpha_1 \neq \alpha_2$.

The next question is, what is the origin of memory in the return intervals? By definition, return interval reflects the time organization of volatilities, and it is well known that volatilities have long-term correlations. Thus we focus on the relation between long-term correlations in these two types of time series. To compare them directly, we also plot the DFA curve of volatilities of GE stock in Fig. 5.4, as shown by open circles. Remarkably, we can see strong similarity between the two DFA curves. Both of them have similar crossovers which are around 1 trading day, and more interestingly, their α values are similar for both sub-regimes. For the volatility, $\alpha = 0.65$ for the short scale and 0.96 for the long scale, which are close to that we obtained for the return intervals.

Then we test the universality of this similarity. We calculate α values of two sub-regimes for the 30 DJIA stocks and the S&P 500 index for both the volatility time series and return interval time series. The results are plotted in Fig. 5.5. For both types of records and both sub-regimes, α values are significantly larger than 0.5 for the shuffled records, which are shown by open triangles in Fig. 5.5. This finding suggests that all of them have long-term correlations. For one α value of the volatility, we can see that the corresponding α value of the return interval is similar, implying that the memory in return intervals is associated with that in volatilities. One might also note that almost all α values of the return interval are slightly smaller than that of the volatility, indicating a certain regularity in the memory mechanism.

To further show the relation between long-term correlations in volatilities and that in return intervals, we also compute α values for volatilities of the 1000 most traded U.S. stocks. The corresponding averages and standard deviations are $\alpha = 0.66 \pm 0.02$ for the short scale regime ($\ell < \ell^*$), and $\alpha = 0.88 \pm 0.06$ for the long scale regime ($\ell > \ell^*$). Both cases are similar to what we found for return intervals. Such behavior suggests a common origin for the strong persistence of correlations in both volatility and return interval records, and in fact the clustering in return intervals is related to the known effect of volatility clustering [84]. Moreover, we draw the scatter plot for the dependence of α values for the two types of records, as shown in Fig. 5.6. We can see a significant dependence for α in the long scale (open squares in Fig. 5.6). However, α for the short scale (open circles) are crowded together so that there is no strong tendency.

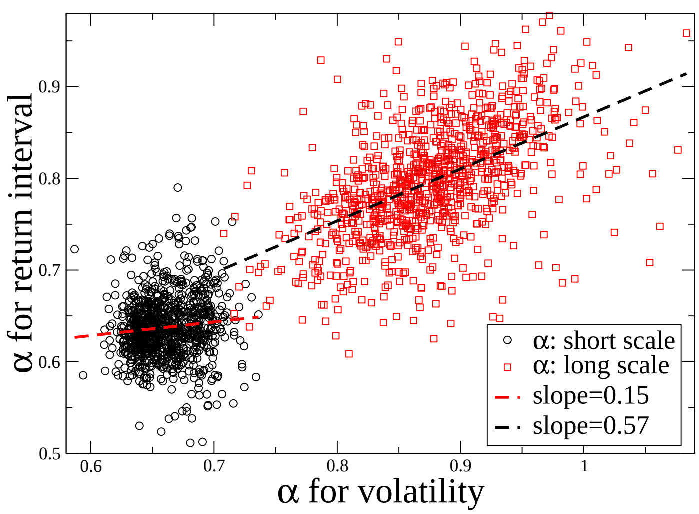

Figure 5.6: Dependence of long-term correlations in the volatility and the return interval ($q = 2$) time series. The results for two scale regimes are showed. As indicated by the two linear fits on the symbols (dashed lines), the dependence is not strong for the short scale but it is significant for the long scale. The weak relation for short scale is related to the limited range of α values.

In conclusion, we find systematical relation between volatilities and return intervals. Our results suggests that the long-term correlation in return intervals is associated with that in volatilities.

Chapter 6

Models for Financial Fluctuations

To understand financial fluctuations and market dynamics, economists, physicists, and mathematician have proposed many models and processes (for example, see references in[1]). It is well known that returns only has short-term correlations and volatilities has long-term correlations. To realize these features, we employ two popular long-term memory models, FIGARCH [85] and fractional Brownian motion (fBm) [86], to simulate returns and therefore volatilities and return intervals. Then we compare the simulation to the empirical results, which will help us to understand the nature of scaling and memory features in volatilities and return intervals. In addition, we propose a new method to forecast the price movement, based on the scaling and memory features of return intervals that we have shown in previous chapters.

6.1 FIGARCH

Fractional integrated generalized autoregressive conditional heteroscedasticity (FIGARCH) is a popular model for return simulation. It is derived from the well-known autoregressive conditional heteroscedasticity (ARCH) and generalized autoregressive conditional heteroscedasticity (GARCH) processes which are based on time-dependent stochastic volatility [87, 88]. For a univariate discrete time series x_t, we can express it as

$$x_t = E(x_t|\Omega_{t-1}) + \epsilon_t, \tag{6.1}$$

where t represents time, $E(.)$ stands for expectation value, and $E(. \mid .)$ denotes expectation value conditional on information set Ω_{t-1} which is one time step earlier. In Eq. (6.1), ϵ_t denotes the noise, which has zero mean, $E(\epsilon_t) = 0$, and no time-dependence, $E(\epsilon_t\, \epsilon_s) = 0$, for all $t \neq s$.

For Eq. (6.1), which is called the "mean equation", the first question is how to extract the influence of previous information Ω_{t-1} and count it in the prediction of x_t. This question has been studied and modeled in many ways. Here we use the classic ARMA process, which employs two of the most famous specifications, Autoregressive (AR) and Moving Average (MA). Thus x_t follows,

$$\Phi(L)(x_t - \mu) = \Theta(L)\epsilon_t, \tag{6.2}$$

where L is the lag operator, $\mu \equiv E(x_t)$, $\Phi(L) = 1 - \sum_{i=1}^{n} \phi_i L^i$ and $\Theta(L) = 1 + \sum_{j=1}^{s} \theta_j L^j$ are AR and MA coefficients respectively.

The second question of the mean equation is how to describe the noise item ϵ_t. Following the classic ARCH(q) and GARCH(p, q) models [87, 88], ϵ_t is defined as

$$\epsilon_t \equiv \sigma_t\, \eta_t, \tag{6.3}$$

where η_t is an *i.i.d.* process with zero mean and unit variance ("Gaussian noise"), and the conditional variance σ_t^2 has different expressions ("variance equation") for different ARCH-type models. In FIGARCH(p, d, q) model, the variance follows

$$\phi(L)(1 - L)^d \epsilon_t^2 = \omega + [1 - \beta(L)](\epsilon_t^2 - \sigma_t^2), \tag{6.4}$$

which is introduced by Baillie, Bollerslev and Mikkelsen (denoted as BBM) [85]. Here $\alpha(L) = \sum_{i=1}^{q} \alpha_i L^i$ and $\beta(L) = \sum_{j=1}^{p} \beta_j L^j$ are ARCH and GARCH parameters, $\phi(L) = [1 - \alpha(L) - \beta(L)](1 - L)^{-1}$ is of order $[max\{p, q\} - 1]$, and $d \in [0, 1]$ is the fractional differencing parameter.

From Eq. (6.4), we can formulate the conditional variance σ_t^2 as

$$\sigma_t^2 = \sigma^2 + \lambda(L)(\epsilon_t^2 - \sigma^2) \tag{6.5}$$

where σ^2 is the unconditional variance of ϵ_t and $\lambda(L) = 1 - [1 - \beta(L)]^{-1}\phi(L)(1-L)^d$. Due to the fractional differencing operator $(1-L)^d$, $\lambda(L)$ is an infinite polynomial of operator L, which, in practice, this summation has to be truncated. BBM suggested a truncated length with 1000 lags. Meanwhile, Chung [89] proposed a slightly different method for FIGARCH process. In Chung's method, the ϵ_t^2 in the left side of Eq. (6.4) is replaced by $\epsilon_t^2 - \sigma^2$, $\lambda(L)$ is truncated at the size of the information set $(t-1)$ and the unobserved $(\epsilon_t^2 - \sigma^2)$ is initialized as 0. For our simulation, we use the BBM method.

Compared to other ARCH/GARCH processes, a specific point for the FIGARCH model is the fractional differencing operator $(1-L)^d$. This operator includes in the information of many previous time steps and thus induces the long-term dependence in the time series. The memory effect is affected by the value of parameter d. When d increases, the effect will gradually vanish.

6.2 Fractional Brownian Motion

The second classic model we tested is fractional Brownian motion (fBm) [86], which is a generalization of Brownian motion. fBm is a centered Gaussian process with stationary increments, similar to the standard Brownian motion. However, the increments of fBm are dependent, while they are independent for the standard Brownian motion. The long-range dependence of the increments can be characterized by the Hurst parameter $H \in (0,1)$, which is the only parameter to index a fBm process $B_H(t)$. Correspondingly, the covariance of the fBm process follows,

$$E(B_H(t)\, B_H(s)) = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}), \tag{6.6}$$

where t and s represent certain times. Note that $B_H(t)$ reduces to a standard Brownian motion when $H = 1/2$ $(E(B_H(t)\, B_H(s)) = max(s,t))$. For $H > 1/2$ $(H < 1/2)$, the increments are positively (negatively) correlated. Another important feature for the fBm process is the scaling invariance,

$$B_H(c\,t) = c^H B_H(t) \tag{6.7}$$

for all $c > 0$.

Since stock returns only have short-term correlations while volatilities have long-term correlations, we simulate returns r_t by

$$r_t = e^{B_H(t)-B_H(t-1)} * \eta_t \tag{6.8}$$

where η_t is a Gaussian noise.

6.3 Comparison between Empirical Data and Simulation

Now we try to simulate returns therefore volatilities and return intervals and compare them with the empirical data. Without loss of generality, we choose the S&P 500 index data (see description in Chapter 1) as the empirical data in the following. The comparison between the empirical data and simulation records will demonstrate two essential features of return intervals, the scaling of their distribution and their size distribution of clusters that demonstrates the memory effect in the return interval time series.

First we test FIGARCH model. From the description, numbers of AR, MA, ARCH, and GARCH coefficients are flexible, thus FIGARCH more precisely is a family of models with different numbers of coefficients. To find the best one from this family, we employ the GARCH package [90] to compare all of them with the empirical data and choose the one with maximum likelihood. For the S&P 500 index, we find that the likelihood is maximized for the case of 1, 1, 1, 0 for the four numbers of coefficients respectively. Therefore in total we have two control parameters, the fractional differencing parameter d and the GARCH parameter β.

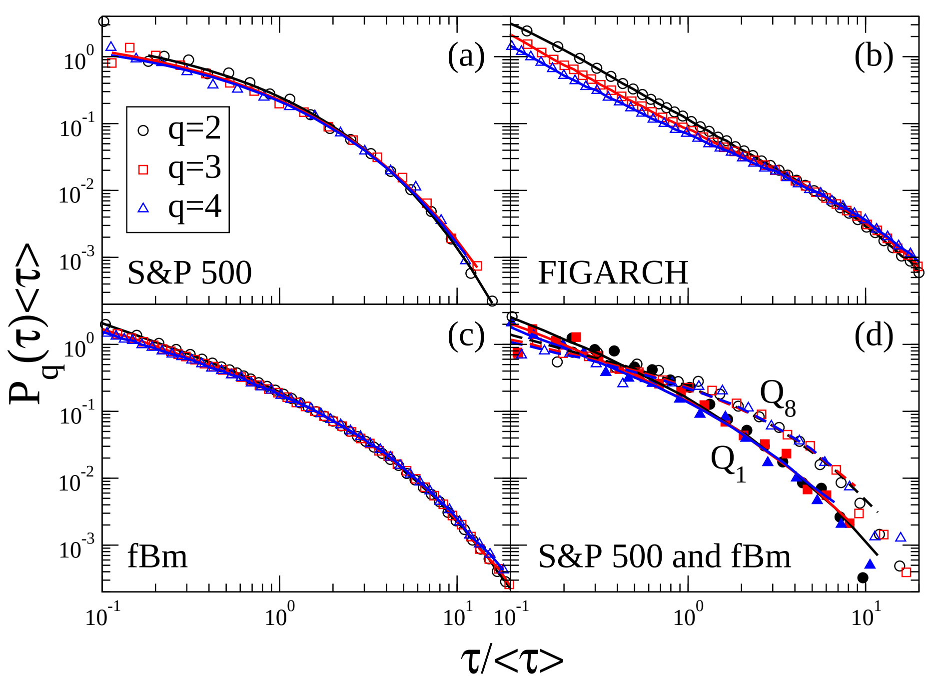

Figure 6.1: Comparison of scaling in scaled return interval distribution of empirical records and two models: FIGARCH and fBm. (a) S&P 500. (b) FIGARCH simulation with $d = 0.33$ and $\beta = 0.15$. (c) fBm simulation with $H = 0.86$. (d) conditional distribution $P_q(\tau|\tau_0)\langle\tau\rangle$ for both fBm simulation (represented by solid lines) and empirical data. Clearly, we can see that fBm shows scaling in the distribution while FIGARCH does not. In panel (d), the conditional distribution following two subset, Q_1 (filled symbols) which has the shortest octave of return intervals for τ_0 and Q_8 (open symbols) which has the longest octave for τ_0, are displayed. The distinct discrepancy between the two octaves suggests that there is a significant short-term memory in the time series. Note that the fBm process mimics the empirical data very well.

By GARCH package [90], we find that $d = 0.33$ and $\beta = 0.15$ for S&P 500 index. We also simulate returns with $d = 0.1, 0.2, ..., 0.8$ and $\beta = 0.1, 0.2, 0.3$ and 0.4 and find that the case of $d = 0.33$ and $\beta = 0.15$ best mimic the empirical data.

Then we derive volatilities and therefore return intervals from returns and test their properties. As shown in Fig. 6.1(b), we plot the return interval distribution for three typical thresholds, $q = 2, 3$, and 4. There are significant deviations from the scaling for both small and large intervals, compared the distributions of S&P 500 [Fig. 6.1(a)]. This result manifests that FIGARCH does not show good scaling in the distribution, which might be due to the stochastic nature of its volatility. Next we test the size distribution of return interval clusters, which is demonstrated in Fig. 6.2. Both positive and negative clusters are tested. We can see that FIGARCH (diamonds in Fig. 6.2) has a relative high probability for large clusters for both positive and negative cases, suggesting the model captures the memory effects for both positive and negative clusters. Their effects are slightly stronger than the empirical memory (circles in Fig. 6.2).

Next we test the fBm process and reiterate the various results previously shown. We employ the DFA method (see Appendix A) to extract the Hurst exponent H (which is equal to the correlation exponent α) for the volatility time series. However, we know the volatility has multiscaling in the long-term correlations [26]. To simplify the simulation, we choose the average α over the entire time scale. By this way, we get $H = 0.86$ for the S&P 500 index. Then we simulate returns and therefore volatilities and return intervals. The distributions of return intervals are shown in Fig. 6.1(c). Remarkably, all curves collapse onto a single curve, suggesting scaling in the distribution. We can see the behavior in panel (c) of Figure 6.1 is very close to that in in panel (a). Further, we also test the distribution conditional on first (Q_1 and last octave Q_8 (similar to Fig. 5.1) and plot them (curves in Fig. 6.1(d)) with that of S&P 500 (symbols in Fig. 6.1(d)) in Fig. 6.1. The two cases match very well. All these results suggesting that the fBm process mimics the empirical records in their scaling behavior. We also test the size distribution of return interval clusters, as shown by triangles in Fig. 6.2. Clearly, the distribution deviates from the shuffled case (squares in Fig. 6.2), suggesting a significant memory in the simulated time series. The curves from fBm process are quite close to that of S&P 500 index (circles in Fig. 6.2). They are only slightly smaller for both positive and negative clusters.

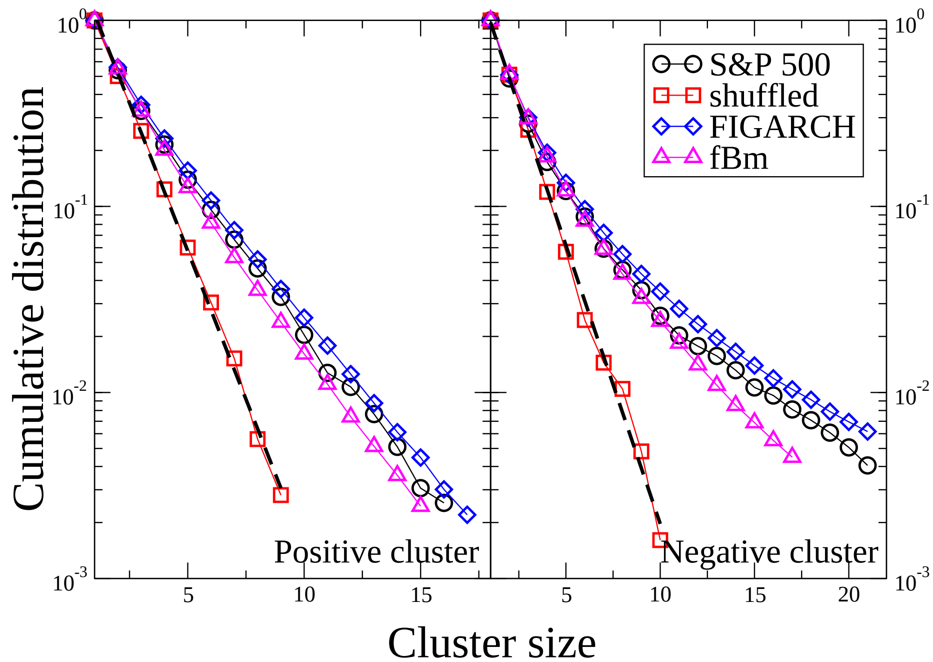

Figure 6.2: Comparison of cumulative distribution of cluster size for return intervals. Distribution for both positive and negative clusters for four cases, the ironical S&P 500 index (circles), shuffled index (squares), simulation for FIGARCH (diamonds), simulation from fBm (triangles), are plotted. The output of two models are very close to that of S&P 500 index and significantly away for the shuffled data, which suggests that the memory in the empirical data can be repeated by both FIGARCH and fBm. Note that FIGARCH slightly overestimates the memory while fBm slightly underestimates it.

In conclusion, we test two popular models, FIGARCH and fBm, for return intervals. For the scaling in the distribution, we find that only fBm shows this behavior. On the other hand, both models show significant memory effects. FIGARCH slightly overestimates the memory while fBm slightly underestimates it.

6.4 A New Method for Risk Estimation

In previous chapters we show many interesting features for returns, volatilities, and return intervals. It is of great interests to see if we can use these features to predict the price movement. Therefore we want to focus on a possible forecasting method. In particular, the scaling and memory properties of the return interval time series can be used for a new method of risk estimation [25]. The most common indicator of risk in the financial world is Value-at-Risk (VAR), which is defined by the risk at a "level of loss" Λ

$$\int_{-\infty}^{-\Lambda} P(r)dr = P^* \,, \tag{6.9}$$

where P^* is the probability of loss and $P(r)$ is the probability density function for returns $r(t)$.

We already analyzed return intervals between events, where the absolute value of the return exceeds a threshold q. Now we focus on losses below $-q$ in order to estimate their risk. Since

$$\langle \tau_q \rangle \equiv \frac{1}{N_q} \sum_{i=1}^{N_q} \tau_q(i), \tag{6.10}$$

where

$$\sum_{i=1}^{N_q} \tau_q(i) \approx \text{total number of returns} \,, \tag{6.11}$$

and

$$N_q + 1 = \text{number of returns with } r < -q \,, \tag{6.12}$$

the average return interval $\langle \tau_q \rangle$ can be related to the VAR via Eq. (6.9) with $\Lambda = q$,

$${\tau_q}^{-1} = \int_{-\infty}^{-q} P(r)dr = \frac{\text{number of returns } r < -q}{\text{total number of returns}} \,, \tag{6.13}$$

which means that $\bar{\tau}_q^{-1}$ gives the loss probability for a risk level $-q$.

In the following we use the additional memory information contained in the return interval time series $\{\tau_i\}$ to improve the estimation of the risk level of loss $-q$. First, we estimate the conditional mean return loss interval $\bar{\tau}_q(\tau_0)$ depending on the previous return interval τ_0. We also expect that, in analogy to Eq. (6.13)

$$\frac{1}{\langle \tau_q(\tau_0) \rangle} = \int_{-\infty}^{-q} P(r|\tau_0) dr \,. \tag{6.14}$$

Here $P(r|\tau_0)$ is the conditional probability that a return r will follow a return interval τ_0. The conditional mean return interval $\langle \tau_q(\tau_0) \rangle$ should be a straight line in a double logarithmic plot (see Fig. 5.2), so that

$$\log\left(\frac{\langle \tau_{\mathrm{q}}(\tau_0)\rangle}{\langle \tau_{\mathrm{q}}\rangle}\right) \propto \log\left(\frac{\tau_0}{\langle \tau_{\mathrm{q}}\rangle}\right) \,. \tag{6.15}$$

Furthermore, $\langle \tau_q \rangle^{-1}$ scales like a certain power of q, $\langle \tau_q \rangle \sim q^{\psi}$ (e.g., $\psi = 3.3$ for IBM). Thus, the scaling properties of the return interval time series enable us to estimate $\langle \tau_q(\tau_0) \rangle$ also for large values of q. The memory in the return intervals as shown in Fig. 5.2 gives us a more accurate estimation of the probability for a loss P^* in Eq. (6.9). Since the condition τ_0 changes every day, the risk level for the next day also changes according to the conditional mean $\langle \tau_q(\tau_0) \rangle$, which can be estimated with this method.

Appendix A

Detrended Fluctuation Analysis (DFA)

It is well known that financial time series are usually non-stationary. In such cases, the conventional methods for correlations such as auto-correlation and spectral analysis have spurious effects. To avoid the artifact correlations arising from non-stationarity, we employ the Detrended Fluctuation Analysis (DFA) method [91–94]. This method is based on the idea that a correlated time series can be mapped to a self-similar process by integration. Therefore, measuring the self-similar feature can indirectly tell us information about the correlation properties. The advantages of DFA over conventional methods are that it permits the detection of long-range correlations embedded in a non-stationary time series, and also avoids the spurious detection of apparent long-range correlations that are an artifact of non-stationarities. This method has been validated on control time series that consist of long-range correlations with the superposition of a non-stationary external trend [91].

As shown in Fig. A.1, the DFA method has following steps:

- (i) integrate the given time series $x(t)$ with N data points into a new time series $y(t)$ (as shown in Fig. A.1(a) and (b)),

$$y(t) \equiv \sum_{t'=1}^{t} x(t'). \tag{A.1}$$

- (ii) divide the integrated time series $y(t)$ into windows of equal length ℓ. In each box, a least squares line is fit to the data (representing the *local trend* in that window, as shown in Fig. A.1). The y coordinate of the straight line segments is denoted by $y'(t)$.

- (iii) de-trend the integrated time series, $y(t)$, by subtracting the local trend, $y'(t)$, in each window, then calculate the root-mean-square fluctuation F of this integrated and detrended time series

$$F(\ell) = \sqrt{\frac{1}{N}\sum_{t'=1}^{N}[y(t') - y_t(t')]^2}. \tag{A.2}$$

- (iv) fit the fluctuation function $F(\ell)$ with the following scaling function,

$$F(\ell) \sim \ell^{\alpha}, \tag{A.3}$$

and obtain the exponent α, which reflects correlations in the $x(t)$ time series. Therefore we call α the correlation exponent.

The α exponent can also be viewed as an indicator of the "roughness" of the original time series: the larger the value of α, the smoother the time series. For different kinds of correlations in a time series, the correlation exponent α has different values. To better understand this relationship, we also relate it to the direct method of measuring correlations, the auto-correlation function [Eq. (1.4)]. However, the ACF has some difficulties to estimate correlations: (*i*) ACF assumes stationarity of the time series. This criterion is not usually satisfied by real-world data. (*ii*) ACF is sensitive to the true average value, $\langle x(t)\rangle$, of the time series, which is difficult to calculate reliably in many cases. Thus ACF can sometimes provide only a qualitative estimation [61].

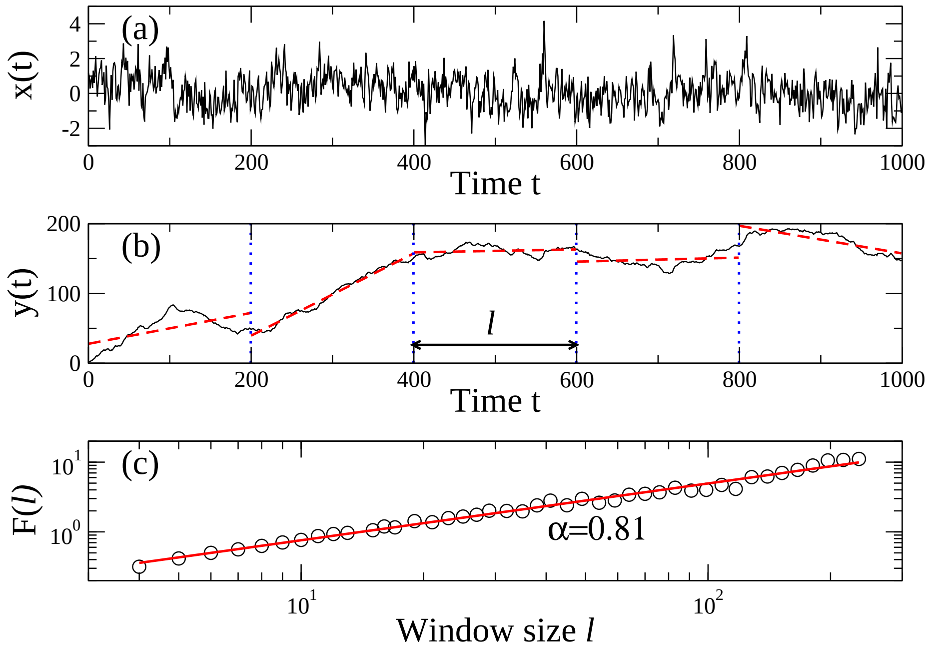

Figure A.1: Illustration of DFA method. (a) is the original time series $x(t)$, (b) is the integrated time series $y(t)$, (c) is the fluctuation function $F(\ell)$ vs. window size ℓ. For a given time series $x(t)$, we integrate it to time series $y(t)$ as Eq. (A.1). Then we divide the entire time scale into many windows with equal-size d, as demonstrated by vertical dot lines in panel (b). After this we fit the curve locally in each window, as shown by dash lines (for example, here we use linear fit). Those fits consist of the local trends $y'(t)$. Then we accumulate the squares of the difference between $y(t)$ and $y'(t)$ over the whole range and get the fluctuation function as described in Eq. (A.2). The last step is the power-law fitting of $F(\ell)$ vs. ℓ to obtain the correlation exponent α, as displayed in panel (c).

Now we summarize several correlation types of time series and the corresponding α and γ values.

- (i) White noise.

For this case, the value at one instant is completely uncorrelated with any previous values. The integrated time series, $y(t)$, corresponds to a random walk and therefore $\alpha = 0.5$, as expected from the central limit theorem [95–97]. Correspondingly, $C(t) = 0$ for any $t \neq 0$.

- (ii) Short-term correlated time series.

  Many natural phenomena are characterized by short-term correlations with a characteristic time scale, τ, and their ACF decays exponentially, [i.e., $C(t) \sim \exp(-t/\tau)$]. The initial slope of $\log F(\ell)$ vs. $\log \ell$ may be different from 0.5, nonetheless the asymptotic behavior for large window sizes ℓ with $\alpha = 0.5$ would be unchanged from the purely random case.

- (iii) Long-term correlated time series.

  $0.5 < \alpha < 1$ indicates persistent long-range power-law correlations, i.e., $C(t) \sim t^{-\gamma}$. The relation between α and γ is

  $$\gamma = 2 - 2\alpha \ . \tag{A.4}$$

- (iv) Long-term anti-correlated time series.

  When $0 < \alpha < 0.5$, power-law *anti-correlations* are present such that large values are more likely to be followed by small values and vice versa [61].

- (v) Out of range.

  When $\alpha > 1$, correlations exist but cease to be of a power-law form.

Appendix B

Power-Law Fitting

In this dissertation we have performed many fittings on various data sets, especially power-law fittings which mainly includes two cases, the fit on the (cumulative) distribution tail and the power-law scaling function. Two classic methods, least-squares fitting and maximum likelihood estimator, are employed. Now we describe and discuss these two methods. We also explain a classic way, Kolmogorov-Smirnov Statistic, to test the goodness-of-fit. Note that both methods, the least-squares fitting and maximum likelihood estimator, can also be applied to fittings other than the power-law.

B.1 Least-Squares Fitting

For a power-law function with scaling parameter ζ,

$$y(x) \sim x^{\zeta}, \tag{B.1}$$

the most common approach is to reduce it to a linear equation,

$$y' = \zeta\, x' + c, \tag{B.2}$$

then fit it by least-squares linear regression. Here $y' = ln(y)$, $x' = ln(x)$, and intercept c is a constant. Residuals, the sum of squares of differences between the empirical values y' and fitting values $\zeta\, x' + c$, are used to test the goodness of the least-squares fitting. When the residuals,

$$R^2 \equiv \sum_{i=1}^{N} (y' - \zeta\, x' + c)^2, \tag{B.3}$$

reaches the minimum, we find the best fit. Therefore, setting $\partial R^2/\partial\zeta = 0$ and $\partial R^2/\partial c = 0$, solving for ζ, we have the estimated scaling parameter

$$\hat{\zeta} = \frac{\langle y'\rangle \langle x'^2\rangle - \langle x'\rangle \langle x'\, y'\rangle}{\sigma^2(x')}. \tag{B.4}$$

Although least-squares fitting appears very frequently in the literature, it has some bias. For example, linear regression assumes a normal distribution for the errors, and thus errors for the power-law fitting should follow a log-normal distribution, which violates the central limit theorem (for more details, see Ref. [72]). To avoid these problems, we only employ least-squares fitting for the cases that the data points are limited and are almost uniformly (in log-log scale) distributed in the entire scale, such as the DFA fluctuation function.

B.2 Maximum Likelihood Estimator

The tail distribution, a key issue in this dissertation, accounts for large fluctuations and events which are very important for risk analysis. It is usually characterized by a power law, i.e. the probability density function

$$P(x) = \frac{\zeta - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\zeta}, \tag{B.5}$$

where ζ is the scaling parameter (called "tail exponent" for the tail distribution) and x_{min} is the minimum value at which power-law behavior holds. Due to the bias of the least-squares fitting, we employ another wide-used method, maximum likelihood estimator (MLE), a.k.a Hill estimator [71, 72, 98], to estimate the tail exponent. To simplify the expression, we assume that the distribution of variable x is continuous in the following. For a discrete distribution, the result is similar [72].

Given a data set containing N observations $x_i \geq x_{min}$, we would like to know the value of ζ for the power-law model that is most likely to have generated our data. The probability that the data were drawn from the model is proportional to

$$P(x \,|\, \zeta) = \prod_{i=1}^{N} \frac{\zeta - 1}{x_{min}} \left(\frac{x_i}{x_{min}} \right)^{-\zeta}. \tag{B.6}$$

This probability is called the *likelihood* of the data given the model. The data are most likely to have been generated by the model with scaling parameter ζ that maximizes this function. Actually we work with $\mathcal{L}$, the logarithm of likelihood, which has its maximum in the same position,

$$\begin{aligned} \mathcal{L} = \ln P(x \,|\, \zeta) &= \ln \prod_{i=1}^{N} \frac{\zeta - 1}{x_{min}} \left(\frac{x_i}{x_{min}} \right)^{-\zeta} \\ &= \sum_{i=1}^{N} \left[\ln(\zeta - 1) - \ln x_{min} - \zeta \ln \frac{x_i}{x_{min}} \right] \\ &= N \, \ln(\zeta - 1) - N \, \ln x_{min} - \zeta \sum_{i=1}^{N} \ln \frac{x_i}{x_{min}}. \end{aligned} \tag{B.7}$$

Setting $\partial \mathcal{L} / \partial \zeta = 0$ and solving for ζ, we obtain the MLE for the scaling parameter:

$$\hat{\zeta} = 1 + N \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}. \tag{B.8}$$

B.3 Kolmogorov-Smirnov Statistic

Though the tail distribution is usually found to be power-law, there is some probability for other distributions. Therefore we need to test the goodness-of-fit for power-law or other fittings. A classic method is the Kolmogorov-Smirnov (KS) statistic D [99, 100], which is the maximum absolute difference between the cumulative distribution of the measured distribution $P(x)$ and that of the fit $S(x)$ (or measured distribution of another data set), i.e.,

$$D \equiv \max(|P(x) - S(x)|), \tag{B.9}$$

for all values in the tail. We can see that KS statistic does not assume any distribution for the data set, which is an important advantage compared to many other statistics. When the statistic D is larger than a certain value, which is called the critical value (CV), the null hypothesis that the distribution follows a power law is rejected. The CV is determined by the significance level and data size N, which can be found is a statistics book. In this dissertation we choose a significance level of 1% and the corresponding $CV = 1.63/\sqrt{N}$.

LIST OF JOURNAL ABBREVIATIONS

Adv. Phys.	Advances in Physics
Ann. Stat.	Annals of Statistics
Appl. Finan. Econ.	Applied Financial Economics
Econ. Lett.	Economics Letters
Eur. Phys. J. B	The European Physics Journal B
Europhys. Lett.	Europhysics Letters
J. Am. Stat. Assoc.	Journal of the American Statistical Association
J. Appl. Econometrics	Journal of Applied Econometrics
J. Bus.	Journal of Business
J. Econ. Surveys	Journal of Economic Surveys
J. Econometrics	Journal of Econometrics
J. Empirical Finance	Journal of Empirical Finance
J. Financ. Econ.	Journal of Financial Economics
J. Finance	Journal of Finance
J. Int. Money Finance	Journal of International Money and Finance
Phys. Rev. E	Physics Review E
Phys. Rev. Lett.	Physics Review Letters
Proc. Natl. Acad. Sci. USA	Proceedings of the National Academy of Sciences of the United States of America
Quant. Finance	Quantitative Finance
Rev. Financ. Stud.	Review of Financial Studies
Rev. Mod. Phys.	Review of Modern Physics
SIAM Rev.	Siam Review

BIBLIOGRAPHY

[1] R. Mantegna and H. E. Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, England, 2000).

[2] J.-P Bouchaud and M. Potters, *Theory of Financial Risk: From Statistical Physics to Risk Management* (Cambridge University Press, Cambridge, England, 2000).

[3] J. Voit, *The statistical Mechanics of Financial Markets* (Springer-Verlag, Berlin, 2001).

[4] N. F. Johnson, P. Jefferies, and P. M. Hui, *Financial Market Complexity* (Oxford University Press, New York, 2003).

[5] B. B. Mandelbrot, J. Bus. **36**, 394 (1963).

[6] R. N. Mantegna and H. E. Stanley, Nature **376**, 46 (1995).

[7] *Fractals in Science*, edited by A. Bunde and S. Havlin (Springer, Heidelberg, 1994).

[8] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford, 1971).

[9] H. E. Stanley, Rev. Mod. Phys. **71**, 358 (1999).

[10] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, Nature **356**, 168 (1992); S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993); S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993); R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **52**, 2939 (1995).

[11] B. Suki, A.-L. Barabási, Z. Hantos, F. Peták, and H. E. Stanley, Nature **368**, 615 (1994).

[12] C. K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993).

[13] H. A. Makse, S. Havlin, and H. E. Stanley, Nature **377**, 608 (1995); H. A. Makse, J. S. Andrade, M. Batty, S. Havlin and H. E. Stanley, Phys. Rev. E **58**, 7054 (1998).

[14] V. Plerou, L. A. N. Amaral, P. Gopikrishnan, M. Meyer, and H. E. Stanley, Nature **400**, 433 (1999).

[15] T. H. Keitt and H. E. Stanley, Nature **393**, 257 (1998); T. H. Keitt, L. A. N. Amaral, S. V. Buldyrev, and H. E. Stanley, "Scaling in the Growth of Geographically Subdivided Populations: Scale-Invariant Patterns from a Continent-Wide Biological Survey" in [Focus issue: *The Biosphere as a Complex Adaptive System*] Phil. Trans. Royal Soc. B: Biological Sciences **357**, 627 (2002).

[16] B. Gutenberg and C. F. Richter, *Seismicity of the Earth and Associated Phenomenon, 2nd Edition* (Princeton University Press, Princeton, 1954).

[17] D. L. Turcotte, *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, Cambridge, 1992).

[18] A. Bunde, J. F. Eichner, S. Havlin, and J. W. Kantelhardt, Physica A **342**, 308 (2004).

[19] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin, Phys. Rev. Lett. **94**, 048701 (2005).

[20] V. N. Livina, S. Havlin, and A. Bunde, Phys. Rev. Lett. **95**, 208501 (2005).

[21] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin, Phys. Rev. E **73**, 016130 (2006).

[22] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin, Phys. Rev. E **75**, 011128 (2007).

[23] E. G. Altmann and H. Kantz, Phys. Rev. E **71**, 056106 (2005).

[24] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **102**, 9424 (2005).

[25] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, and H. E. Stanley, in *Proceedings of the Third Nikkei Econophysics Research Workshop and Symposium, The Fruits of Econophysics, Tokyo, November 2004*, edited by H. Takayasu (Springer-Verlag, Berlin, 2005), p. 43.

[26] F. Wang, K. Yamasaki, S. Havlin, and H. E. Stanley, Phys. Rev. E **73**, 026117 (2006).

[27] F. Wang, P. Weber, K. Yamasaki, S. Havlin, and H. E. Stanley, Eur. Phys. J. B [Proc. Special Issue on Econophysics] **55**, 123 (2007).

[28] P. Weber, F. Wang, I. Vodenska-Chitkushev, S. Havlin, and H. E. Stanley, Phys. Rev. E **76**, 016109 (2007).

[29] F. Wang, K. Yamasaki, S. Havlin and H. E. Stanley, Phys. Rev. E **77**, 016109 (2008).

[30] F. Wang, K. Yamasaki, S. Havlin and H. E. Stanley, Phys. Rev. E **79**, 016103 (2009).

[31] W.-S. Jung, F. Z. Wang, S. Havlin, T. Kaizoji, H.-T. Moon and H. E. Stanley, Eur. Phys. J. B **62**, 113 (2008).

[32] I. Vodenska-Chitkushev, F. Z. Wang, P. Weber, K. Yamasaki, S. Havlin, and H. E. Stanley, Eur. Phys. J. B **61**, 217 (2008).

[33] M. I. Bogachev, J. F. Eichner, and A. Bunde, Phys. Rev. Lett. **99**, 240601 (2007).

[34] M. I. Bogachev and A. Bunde, Phys. Rev. E **78**, 036114 (2008).

[35] T. Qiu, L. Guo and G. Chen, Physica A **387**, 6812 (2008).

[36] F. Ren, L. Guo and W.-X. Zhou, http://arxiv.org/abs/0807.1818v1.

[37] F. Ren and W.-X. Zhou, http://arxiv.org/abs/0809.0250v1.

[38] A. Admati and P. Pfleiderer, Rev. Financ. Stud. **1**, 3 (1988).

[39] L. Bachelier, "Théorie de la spéculation" [Ph.D. thesis in mathematics], Annales Scientifiques de l'Ecole Normale Supérieure *III-17*, 21 (1900). Translated and Reprinted in *The Random Character of Stock Market Prices*, edited by P. Cootner, (MIT Press, Cambridge, Mass, 1964), p. 17.

[40] T. Lux, Appl. Finan. Econ. **6**, 463 (1996).

[41] P. Gopikrishnan, M. Meyer, L. A. N. Amaral, and H. E. Stanley, Eur. Phys. J. B **3**, 139 (1998).

[42] A. Pagan, J. Empirical Finance **3**, 15 (1996).

[43] 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 Publishers, 1998), p. 83.

[44] V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 6519 (1999); V. Plerou, P. Gopikrishnan, L. A. N. Amaral, X. Gabaix, and H. E. Stanley, *ibid* **62**, 3023 (2000); P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley, *ibid* **62**, 4493 (2000).

[45] E. F. Fama, J. Bus. **36**, 420 (1963).

[46] E. F. Fama, J. Bus. **38**, 34 (1965).

[47] R. N. Mantegna and H. E. Stanley, Nature **383**, 587 (1996).

[48] R. N. Mantegna and H. E. Stanley, Physica A **239**, 255 (1997).

[49] V. Pareto, *Cours d'Economie Politique* (Lausanne and Paris, 1897).

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

[51] J. Klafter, G. Zumofen, and M.F. Shlesinger, in *Lévy Flights in Physical Sceinces and Related Phenomena*, edited by M.F. Shlesinger, G. M. Zaslavsky and U. Frisch, (Springer, Berlin, 1995), p. 196.

[52] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 5305 (1999).

[53] M. M. Dacorogna, U. A. Muller, R. J. Nagler, R. B. Olsen, and O. V. Pictet, J. Int. Money Finance **12**, 413 (1993).

[54] Z. Ding, C. W. J. Granger and R. F. Engle, J. Empirical Finance **1**, 83 (1993).

[55] E. F. Fama, J. Finance **25**, 383 (1970).

[56] C. W. J. Granger and Z. Ding, J. Econometrics **73**, 61 (1996).

[57] Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 437 (1997).

[58] P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 441 (1997).

[59] R. Cont, Ph.D. thesis, Universite de Paris XI, 1998 (unpublished); see also e-print cond-mat/9705075.

[60] Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Phys. Rev. E **60**, 1390 (1999); V. Plerou, P. Gopikrishnan, X. Gabaix, L. A. N. Amaral, and H. E. Stanley, Quant. Finance **1**, 262 (2001); V. Plerou, P. Gopikrishnan, and H. E. Stanley, Phys. Rev. E **71** , 046131 (2005). For application to heartbeat intervals, see Y. Ashkenazy, P. Ch. Ivanov, S. Havlin, C.-K. Peng, A. L. Goldberger, and H. E. Stanley, Phys. Rev. Lett. **86**, 1900 (2001).

[61] J. Beran, *Statistics for Long-Memory Processes* (Chapman & Hall, NY, 1994).

[62] F. Lillo and R. N. Mantegna, Phys. Rev. E **62**, 6126 (2000).

[63] Z. Eisler and J. Kertész, Phys. Rev. E **73**, 046109 (2006).

[64] Z. Eisler and J. Kertész, Eur. Phys. J. B **51**, 145 (2006).

[65] Z. Eisler, I. Bartos and J. Kertész, Adv. Phys. **57**, 89 (2008).

[66] L. Harris, J. Financ. Econ. **16**, 99 (1986).

[67] G. W. Schwert, J. Finance **44**, 1115 (1989); K. Chan, K. C. Chan, and G. A. Karolyi, Rev. Financ. Stud. **4**, 657 (1991); T. Bollerslev, R. Y. Chou, and K. F. Kroner, J. Econometrics **52**, 5 (1992); A. R. Gallant, P. E. Rossi, and G. Tauchen, Rev. Financ. Stud. **5**, 199 (1992); B. Le Baron, J. Bus. **65**, 199 (1992).

[68] M. Pasquini and M. Serva, Econ. Lett. **65**, 275 (1999).

[69] R. A. Wood, T. H. McInish, and J. K. Ord, J. Finance **40**, 723 (1985).

[70] S. S. Wang, O. M. Rui, M. Firth, J. Int. Money Finance **21**, 265 (2002).

[71] V. Plerou and H. E. Stanley, Phys. Rev. E **76**, 046109 (2007).

[72] A. Clauset, C. R. Shalizi and M. E. J. Newman, http://arxiv.org/abs/0706.1062v1.

[73] K. Yamasaki, A. Gozolchiani, and S. Havlin, Phys. Rev. Lett. **100**, 228501 (2008).

[74] M. Constantin and S. Das Sarma, Phys. Rev. E **72**, 051106 (2005).

[75] T. Di Matteo, Quant. Finance **7**, 21 (2007).

[76] C. Q. Cao and R. S. Tsay, J. Appl. Econometrics **7**, S165 (1992)..

[77] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).

[78] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[79] H. A. Makse, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. E **53**, 5445 (1996).

[80] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry (Revised and Enlarged Edition)* (North-Holland, Amsterdam, 1992).

[81] E. F. Fama and K. R. French, J. Finance **51**, 55 (1996).

[82] H. Markowitz, J. Finance **7**, 77 (1952).

[83] P. Ch. Ivanov, A. Yuen, B. Podobnik, and Y. Lee, Phys. Rev. E **69**, 056107 (2004).

[84] T. Lux and M. Marchesi, Int. J. Theor. Appl. Finance **3**, 675 (2000); I. Giardina and J.-P. Bouchaud, Physica A **299**, 28 (2001); T. Lux and M. Ausloos, in *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*, edited by A. Bunde, J. Kropp, and H. J. Schellnhuber (Springer, Berlin, 2002), p. 373.

[85] R. T. Baillie, T. Bollerslev and H.O. Mikkelsen, J. Econometrics **74**, 3 (1996).

[86] B. B. Mandelbrot and J. W. Van Ness, SIAM Rev. **10**, 422 (1968).

[87] R. F. Engle, Econometrica **50**, 987 (1982).

[88] T. Bollerslev, J. Econometrics **31**, 307 (1986).

[89] C.-F. Chung, *Estimating the Fractionally Integrated GARCH Model* (National Taiwan University Working Paper, 1999).

[90] S. Laurent and J.-P. Peters, J. Econ. Surveys **16**, 447 (2002).

[91] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994); C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, Chaos **5**, 82 (1995).

[92] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, Phys. Rev. Lett. **85**, 3736 (2000).

[93] K. Hu, P. Ch. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley, Phys. Rev. E **64**, 011114 (2001); Z. Chen, P. Ch. Ivanov, K. Hu, and H. E. Stanley, *ibid* **65**, 041107 (2002); L. Xu, P. Ch. Ivanov, K. Hu, Z. Chen, A. Carbone, and H. E. Stanley, *ibid* **71**, 051101 (2005); Z. Chen, K. Hu, P. Carpena, P. Bernaola-Galvan, H. E. Stanley, and P. Ch. Ivanov, *ibid* **71**, 011104 (2005); J. W. Kantelhardt, S. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, Physica A **316**, 87 (2002).

[94] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, Physica A **295**, 441 (2001).

[95] J. Feder, *Fractals*, (Plenum Press, New York 1988).

[96] E. W. Montroll and W. W. Badger, *Introduction to Quantitative Aspects of Social Phenomena*, (Gordon and Breach Science publishers, New York, 1974).

[97] A. Bunde and S. Havlin, in *Fractals and Disordered Systems*, A. Bunde and S. Havlin (editors), 2^{nd} ed. (Springer, Heidelberg 1996).

[98] B. M. Hill, Ann. Stat. **3**, 1163 (1975).

[99] M. A. Stephens, J. Am. Stat. Assoc. **69**, 730 (1974).

[100] R. Engle and J. Russel, Econometrica **66**, 1127 (1998).

CURRICULUM VITAE

FENGZHONG WANG

Boston University, Physics Department
Center For Polymer Studies
590 Commonwealth Avenue
Boston, Massachusetts 02215 USA

Telephone: 617/353-2701
Facsimile: 617/353-9393 or 617/353-3783
E-mail: fzwang@buphy.bu.edu
http://buphy.bu.edu/~fzwang

EDUCATION

- Ph.D., 2009 Physics/Econophysics; Boston University, Boston, MA.

  Academic advisor: Professor H. Eugene Stanley.

- M.S., 2001 - Physics; Nanjing University, Nanjing, P.R. China.
- B.S., 1998 - Physics; Nanjing University, Nanjing, P.R. China.

ACADEMIC APPOINTMENTS

- *Research Assistant* [2003.9 - present] Center for Polymer Studies and Physics Department, Boston University.
- *Teaching Assistant* [2001.9 - 2003.7] Physics Department, Boston University.
- *Research Assistant* [1998.2 - 2001.6] National Laboratory of Solid State Microstructures, Nanjing University.
- *Teaching Assistant* [1999.9 - 2000.6] Physics Department, Nanjing University.

PROGRAMMING EXPERIENCE

• **OS** - LINUX & Windows.

• **Languages** - C/C++, shell script, Perl, Matlab, Fortran 77/90.

SELECTED PUBLICATIONS

• **F. Wang**, K. Yamasaki, S. Havlin, and H. E. Stanley, "Multifactor Analysis of Multiscaling in Volatility Return Intervals", Phys. Rev. E **79**, 016103 (2009).

• W.-S. Jung, **F. Wang**, T. Kaizoji, S. Havlin and H. E. Stanley, "Volatility Return Interval Analysis of the Japanese Stock Market", Eur. Phys. J. B **62**, 113 (2008).

• B. Podobnik, D. Horvatic, F. Pammolli, **F. Wang**, H. E. Stanley and I. Grosse, "Size-Dependent Standard Deviation for Growth Rates: Empirical Results and Theoretical Modeling", Phys. Rev. E **77**, 056102 (2008).

• I. Vodenska-Chitkushev, **F. Wang**, P. Weber, K. Yamasaki, S. Havlin, and H. E. Stanley, "Scaling and Memory Features of Volatility Return Intervals: Comparison between Empirical Data and Models", Eur. Phys. J. B **61**, 217 (2008).

• W.-S. Jung, **F. Wang** and H. E. Stanley, "Gravity Model in the Korean Highway", Europhys. Lett. **81**, 48005 (2008).

• **F. Wang**, K. Yamasaki, S. Havlin, and H. E. Stanley, "Indication of Multiscaling in the Volatility Return Intervals of Stock Markets", Phys. Rev. E **77**, 016109 (2008).

• W.-S. Jung, O. Kwon, **F. Wang**, T. Kaizoji, H.-T. Moon, and H. E. Stanley, "Group Dynamics of the Japanese Market", Physica A **387**, 537 (2008).

• 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 **76**, 016109 (2007).

• **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).

- **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).