DIPARTIMENTO DI FISICA E ASTRONOMIA "GALILEO GALILEI"
Corso di Laurea Magistrale in Fisica

Università degli Studi di Padova

TESI DI LAUREA MAGISTRALE

Modelization of high frequency financial data based on ensembles and scaling

Relatore:
Ch.mo Prof. Attilio STELLA

Laureando:
Federico MUSCIOTTO

Correlatore:
Dott. Michele CARAGLIO

Anno accademico 2013/2014

Contents

Introduction

The current financial crisis highlights the crucial need of a change of mindset in economics and financial engineering, that should move away from dogmatic axioms and focus more on data, orders of magnitudes, and plausible, albeit non rigorous, arguments.

Economics needs a scientific revolution
J.P. Bouchaud

In the past decades statistical mechanics made significant steps in the development of the notions of self-similarity and scaling [1], [2], which are key concepts in the description of critical phenomena and complex systems. The advantages of an approach based on scaling lie in the fact that it makes possible to treat and analyze in a simple way data taken at different time or space scales. In this thesis we want to adopt this approach to obtain a model capable to describe financial data. In fact, the correct performance of such a modelization is fundamental to understand and try to forecast the dynamics of global markets. Among all the possible applications of this knowledge, one of the most relevant is the opportunity it can give to identify the upcoming of financial crises. Indeed, as an example, if we consider the volatility of a given financial asset, which by definition is a measure of the absolute value of the maximum (logarithmic) variation of its price within a defined time window, it can be easily understood how a correct theoretical description of its correlations can be useful to forecast the potential occurrence of market instability.

Actually, the first contributions given by physicists to the financial field date back to the beginning of the past century. Indeed, the first one to perform stochastic studies applied to stock market was L. Bachelier [3], a Poincaré's student who in 1900 proposed a model based on the random walk of independent identically distributed (i.i.d.) variables as the process which generates the price fluctuations of the assets. Although this model was later invalited for its inaccuracies, it gave birth to a modelization of financial

dynamics as a process with totally random logarithmic increments, normally distributed. This kind of modelization was not able to describe many empirical features of financial data, such as the distribution of rare, extreme events, which is fundamental in the determination of financial risks [4], but it took more than half a century to abandon it. After Mandelbrot's work of 1963 on cotton prices [5], which for the first time proposed for the financial stochastic process a PDF different from the Gaussian distribution, the validity of the previous approach was questioned, and various attempts to formulate a more accurate modelization of market dynamics soon followed. More recently, the huge amount of data collected among market indexes since 1993[1] and the computer analysis performed on them have opened the door to new proposals of modeling market's behaviour. This modern approach moves from the consideration of a number of robust statistical properties that one can recognize analyzing financial time series for different markets and different periods of time. These are called *stylized facts* and we refer to a recent review by Rama Cont [6], for a rather clear and complete account. Nowadays, the challenge is to find a model capable to reproduce as closely as possible the empirical features shown by these data sets. However, when analyzing financial data in order to look for a way to correctly model them, it is a good point to question about the way they must be treated. Indeed, treating them as an empirical realization of an underlying stochastic process is not an immediate step. When we record financial data for a certain period, we obtain a realization of a process whose dynamics cannot be replicated. Indeed, only a single, possibly long time series is available. This is a frequent issue in statistical mechanics, as similar time series occur when studying many processes, e.g. solar flares [7], heartbeats [8], or earthquakes [9]. We thus must face an epistemological problem, since we have to find the way to treat data which cannot be reproduced. The most common way to overcome this problem is to look at the increments on variable intervals of the time series: if they are stationary, and ergodicity can be assumed, we can take time averages as proxies to the expectation values. Indeed, stationarity implies that the distribution function of increments does not depend on time: time average is then a reliable tool to identify the expectation value of a certain quantity. However, stationarity and ergodicity are not granted features for such processes: the existence of a single realization of the process makes problematic either to prove or disprove these properties, while in some cases data can show evident non-stationary behaviour [10], [11]. In the financial field, a different approach was thus proposed for intraday analysis by Bassler et al. in 2007 [10]. Indeed, looking at some ensemble averages, such as the empirical

[1]Since this year, financial and market data have been recorded transaction by transaction, with a frequency up to few seconds.

second moment of returns or the volatility autocorrelations, a rather evident periodic behaviour emerges in well defined time windows. This feature suggests the idea that, within each window, we are dealing with a single realization of the same process. We are thus induced to rearrange the data in an ensemble, i.e. to divide the time series in equal parts (called histories) whose duration is determined by the frequency of the periodic behaviour (in our case, it is a day of trading activity). In this way we obtain a collection of hystories, each of which shows similar dynamics and time evolution. The most immediate difference in this change of approach is that now we use ensemble averages to approximate the expectation values.

We will make use of this ensemble approach to define a suitable intraday model for the returns of a financial asset. Then we will exploit it to implement a trading strategy, which should be seen as a test to study the effects of linear correlations on the structure of the model itself. We will work at various frequencies with the S&P 500 index and other assets to explore and study different frameworks for our model. Moreover, we will look for a way to include in the model the treatment of night and interday returns, which do not belong to intraday context, to improve its predictive power and connect the intraday and the interday frameworks within the same formalism.

The mathematical features of the model have been suggested by the symmetries and the statistical properties of the ensemble itself. Among these symmetries the main one is scaling, which gives powerful tools for analyzing a given system at different space or time scales. After the first studies on scaling, the birth of the theory of renormalization group (RG) [2], [12], allowed a more accurate analysis of its properties and its origins in statistical mechanics. The leading principle in RG approach is to postulate scaling relations between the variables which we choose to describe the system at different scales. In this way empirical data collected under different conditions can be unified and analyzed through general tools. This allows to find the same statistical properties at different scales, when analyzing the behaviour of fluctuations. The word renormalization in fact means the procedure of redefining the parameters which describe the system after changing its space or time scale. In the context of financial modeling, to postulate scaling invariance with time for the system is equivalent to impose, for the PDF of returns r aggregated over an interval τ, the validity of the scaling law

$$p(r,\tau) = \frac{1}{\tau^D}\ g\Big(\frac{r}{\tau^D}\Big)\ , \tag{1}$$

where g is the scaling function, and D the Hurst exponent of the process [13]. Eq. (1) means that if the probability density function of the sum of the returns (the aggregated

ones) is multiplied for a power D of the number of addends which occur in r, it converges asymptotically to the scaling function g. The particular features of the system condition the value of D and the shape of g. For example, when $D = 1/2$ and g is Gaussian, normal scaling occurs, which means that the system can be described by a stochatic process with i.i.d increments, such as in the Bachelier's model. Anomalous scaling, which occurs when at least one of the conditions for D and g is not verified, usually indicates non-Markovian features for the underlying stochastic process, i.e. a memory longer than one step in the past. Our work of modelization will pass through the determination of the correct Hurst exponent and the most suitable scaling function to best describe our data.

This thesis is structured as follows. The first chapter introduces the concept of stylized facts, the data sets which have been analyzed and their robust statistical features. In the second chapter a definition of a model suitable to describe data's features is provided, then its calibration protocol and the implementation of the trading strategy are described. The third and the fourth chapters present the result of the application of such a model to the S&P 500 index and to different market equities. Finally, in chapter five, we draw our conclusions.

1. Stylized facts and S&P 500 overview

1.1. Definition of stylized facts

Since the birth of mathematical finance, the statistical properties of many different financial data sets have been investigated to provide an empirical hint to formulate new models. The outcoming possibility of storing huge amounts of data at really high frequencies, up to few seconds, and the computer based methods for analyzing their features have accelerated this process, leading to the discovery of many empirical properties which are shared by a large variety of prices, indexes and exchange rates among different markets and periods. From one side, this plenty of discoveries has led to the birth of new models which tried to include as many of these properties as possible. On the other, their variety, together with their occasional inconsistencies, made impossible to conceive an unique model capable to contain them all. For this reason, when presenting them it can be useful to formulate them just as empirical evidence, without forcing any parameterization or modelization. Guided by this approach, in the first part of this thesis we refer to Rama Cont's work [6], who in 2001 presented an useful review in which all the so far discovered empirical features, called stylized facts, were enumerated and described. Their discovery represents an alternative to the classic event-based approach, which attempts to explain market fluctuations by considering them as the results of some political or social event. The empirical robustness of these properties, shared by very different kinds of assets, such as e.g. US/Yen exchange rate and corn futures, gives accurate tools to build more versatile models. However, we underline that their generality makes them qualitative properties which lacks of an exact and definite parameterization, since they are obtained by taking the common denominator among properties observed for different data sets. Before introducing them, we start by fixing some notation. With $S(t)$ we indicate the price of a given asset, while the logarithmic return is defined as

$$r_\tau(t) = \ln S(t+\tau) - \ln S(t) \ , \tag{1.1}$$

where τ is the time lag and t is a given instant of the time series.

1.2. Stylized statistical properties of asset returns

The most common and significant stylized facts are presented as follows:

- **Absence of autocorrelations**: the empirical linear autocorrelation function calculated on a discrete time series of returns, $r_\tau(0), r_\tau(\tau), \ldots, r_\tau(n\tau)$, and defined as

$$\langle r_\tau(t) r_\tau(t+T) \rangle_e \equiv \frac{1}{n-k+1} \sum_{i=0}^{n-k} r_\tau(i\tau) r_\tau((i+k)\tau) \ , \tag{1.2}$$

  with $T = k\tau$, vanishes fot T longer than approximately $10 \div 20$ minutes. For smaller intraday scales microstructure effects can be observed.

- **Presence of correlations on absolute values of returns and high order correlations**: the autocorrelation function calculated on absolute values

$$\langle |r_\tau(t)||r_\tau(t+T)| \rangle_e \equiv \frac{1}{n-k+1} \sum_{i=0}^{n-k} |r_\tau(i\tau)||r_\tau((i+k)\tau)| \tag{1.3}$$

  shows instead a positive value and a slow decay, as $\frac{A}{T^\eta}$, with A constant and $\eta \in [0.2, 0.4]$. Moreover, various measures of squared returns show similar significant correlation over several days. This effect, known as *volatility clustering*, quantifies the tendency of high-volatility events to cluster in time.

- **Heavy tails**: for long time series the probability density function (PDF) of the returns $p_{R_\tau}(r)$ can be estimated. This function shows fat tails ruled by an exponent in the range of $2 \div 5$, i.e. heavier than the light tail of the Gaussian distribution. Therefore Normal distribution is not proper to indentify the right PDF and rare events, with respect to it, show much lower probabilities. The first one to introduce in the financial context a version of these tails, the so called *Pareto* ones, was Mandelbrot [5], who took the idea from Levy's work about stable distributions [14].

- **Leverage effect**: usually volatility of an asset is negatively correlated with the returns of the same asset, i.e. the leverage correlation function, defined as $L_\tau(T) = \frac{1}{K} \langle |r_\tau(t+T)|^2 r_\tau(t) \rangle$, where K is a suitable renormalization factor, shows negative values, asymptotically increasing toward 0, for $T > 0$. This implies that the persistence of a negative trend for an asset can lead to a rise in volatility.

- **Volume/volatility correlation**: measures of volatility are correlated with the volume of trading activities.

1.3. Some issues about statistical estimation

1.3.1. Stationarity and ergodicity

In order to ensure validity to any statistical analysis of a long time series, data taken from various assets must satisfy the stationary hypothesis. Its rigorous formulation claims that the joint PDF for the returns $r_\tau(1), r_\tau(2), \ldots, r_\tau(k)$ must be the same as the one calculated on the time-translated returns $r_\tau(1+T), r_\tau(2+T), \ldots, r_\tau(k+T)$, with $T = n\tau$, $n \in \mathbb{N}$. This means that there are some statistical properties which remain stable over time, and that data collected in different periods can be analyzed together. Stationarity is not obvious, and it can be invalidated by seasonality effects such as intraday variability, week end effects... For intraday analysis at high frequency, stationary hypothesis is not valid, [11], as it will be shown soon in this thesis, and a different way of treating the data sets has to be chosen.

If stationarity is required to ensure independence from time, ergodicity makes the empirical averages converge to the quantities they are supposed to represent: in other word, if some ergodic property is present, the sample moment defined by

$$\langle f(r_\tau) \rangle = \frac{1}{N} \sum_{t=1}^{N} f(r_\tau(t)) \tag{1.4}$$

can be identified with the expectation value $\mathbb{E}[f(r_\tau)]$ [6]. It can be noticed that in eq. (1.4) any dependence on time is cancelled by the mean evaluation. Like stationarity, ergodicity is in general difficult either to prove or to disprove. In particular, failure of ergodicity is quite common when considering physical systems with long-range dependence.

1.3.2. Intraday seasonalities

The stationarity hypothesis is often not valid when dealing with financial assets, as previously discussed for exchange rate [10], stock markets and market indexes [11]. In particular, if concerned with intraday analysis at high frequency, non-stationarity is often an evident issue, because intraday effects become predominant and alter time homogeneity during the day. This feature makes the time averages of eq. (1.4) inappropriate to identificate the desired expectation values, since data taken at high frequency show clear

time inhomogeneities. Moreover, we are going to show that in our case each single day of trading activity show a similar pattern of nonlinear moments and correlations, and this allows us to consider it as an iterate realization of an underlying stochastic process. In this context, a possible way to overcome the homogeneity problem is to consider instead ensemble means. Following an ensemble approach means to arrange data in a collection, called ensemble, of separate histories, each of the same duration. If the number of histories is large enough, we can treat our data through the ensemble statistical tools. Then, if M is the total number of histories, the ensemble mean is defined as

$$\langle f(r_\tau(t))\rangle = \frac{1}{M}\sum_{l=1}^{M} f(r_\tau^{(l)}(t)) \,, \tag{1.5}$$

where t runs in the time window of the single history. It can be noticed that eq. (1.5) indicates the evaluation of a mean among all the histories, which allows us to calculate different values of the same average quantity at different moments of their time window, while the long series mean of eq. (1.4) does not make any practical distinction between data taken at different instants of the day. We remark that the ensemble approach is justified by the presence of some periodic behaviour for the data set. In the next sections we will show that our data set, such as many other financial ones, shows a clear periodicity in the volatility pattern, which is U-shaped if we take a trading day as the time window of a single history. Moreover, this periodicity in volatility behaviour is a confirmation of the stylized fact which states a correlation between volatility and the number of transactions: along the morning trading activities usually decrease until lunch time, and then they start growing again. These remarks thus suggest to arrange our data as an ensemble of all the trading days, such as has already been made in [10], [11]. In this way all the time averages considered in long time series analysis, evaluated according to eq. (1.4), are substituted the ensemble means defined in eq. (1.5), and each day becomes an indipendent realization of the same process.

1.4. S&P 500 index

The data set used in the first part of this work is made up of the prices of S&P 500, a stock market index based on the market capitalizations of the 500 largest USA companies. For each trading day, returns were collected from 09:40 to 16:00 (New York time) from September 30th 1985 to June 28th 2013, with a total number of $M = 6852$ trading days. As already mentioned, the chosen approach is a high frequency, intraday analysis. The

first step to perform is demonstrating non-stationarity condition for returns collected along a single day. Figure 1.1(a) shows the behaviour of the empirical second moment $m_2^e(t,\tau)$, defined as

$$m_2^e(t,\tau) \equiv \frac{1}{M}\sum_{l=1}^{M} r_\tau^l(t)^2 \,, \tag{1.6}$$

where t runs inside a single stock market day and τ is chosen equal to 1 minute.
If the underlying stochatics dynamics had stationary increments, $m_2^e(t,\tau)$ would be constant. It shows instead a clear U-shaped pattern, with a decay approximately in the first three hours of the day, and an increase in the second part of the day. The assumption of a well defined stochastic dynamics which evolves during the trading day is confirmed by performing a corresponding analysis of $m_2^e(t,\tau)$ throughout a whole week, which is shown in figure 1.1(b). Roughly considering just the first three hours of the day it is evident that the empirical second moment decreases according to a generalized power law of time, $m_2^e(t,\tau) \sim t^{2D} - (t-1)^{2D}$: we then fixed $t_m = 20$ as the point in which $m_2^e(t,\tau)$ changes its trend and starts to grow again. Figure 1.2(a) shows a plot of this relation, with $D = 0.360$ and $m_2^e(1,\tau) = 5.2 \cdot 10^{-7}$, with t up to t_m. An important fact verified by our data set in the ensemble approach is that, although considering high frequency intraday dynamics, linear correlations of returns defined as

$$c_{lin}^e(1,t) \equiv \frac{\frac{1}{M}\sum_{l=1}^{M} r_1^l r_t^l}{\sqrt{m_2(1,1)m_2(t,1)}} \,, \tag{1.7}$$

are negligible with respect to correlations calculated on the absulate values of the same returns, as evident in the comparison made in figure 1.3. The periodicity which emerges from the plot of $m_2^e(t,\tau)$ is a confirmation of the non-stationarity of the underlying process and a justification of our ensemble approach. Moreover, a more accurate analysis of high order moments suggests the presence of some scaling properties to be investigated. Indeed, if the definition of empirical moments is extended to aggregated returns, defined as $r_t(t) = \ln S(t) - \ln S(0) = \sum_{i=1}^{t} r_t(i)$, and to a generic index q by defining

$$m_q^e(t,t) \equiv \frac{1}{M}\sum_{i=1}^{M} |r_t(t)|^q \,, \tag{1.8}$$

the presence of a scaling simmetry becomes evident. Indeed, in this range of t and for $q \in [0,2]$ one finds a regime of simple scaling, i.e. $m_q^e(t,t) \sim t^{qD}$, with D not depending on q. The Hurst exponent D can be obtained by computing qD for every $m_q^e(t,t)$ at

different values of q and then making a linear fit of the couples (q, qD), as shown in figure 1.2(b), where $D = 0.356$.

This last result suggests the existence of a scaling collapse for the PDF of aggregated returns $r_t(t)$ in the form of

$$t^D p_{R_t(t)}(t^D r) = g(r) \, , \tag{1.9}$$

where we use the $R_t(t)$ subscript to indicate the process which generates the aggregated return $r_t(t)$. From now on we will adopt this notation, except where explicitly specified. The collapse of eq. (1.9) gets confirmed by empirical estimation, as shown in figure 1.2(c). Even if we have not given yet any definition for our model it is already evident that $g(r)$ cannot be a Gaussian: it clearly shows tails which are heavier than the expected ones for the Normal distribution. It can also be assumed to be even to a good approximation. This is an interesting evidence, since it implies that the occurence of a positive increment of the price of the index has the same probability of the opposite one. This is not an obvious feature: many financial assets show instead a clear gain/loss asymmetry [6].

In the next chapter we will provide an exact formulation for the model used in this thesis on the basis of the empirical results just presented.

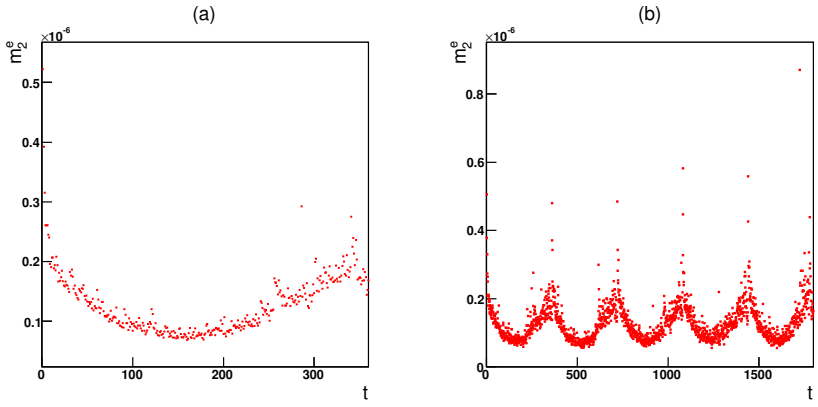

Figure 1.1: (a): The ensemble second moment $m_2^e(t, \tau)$ of the daily S&P 500 index as a function of the time of day, from 09:00 to 16:00. (b): The weekly behaviour of $m_2^e(t, \tau)$ for the same data.

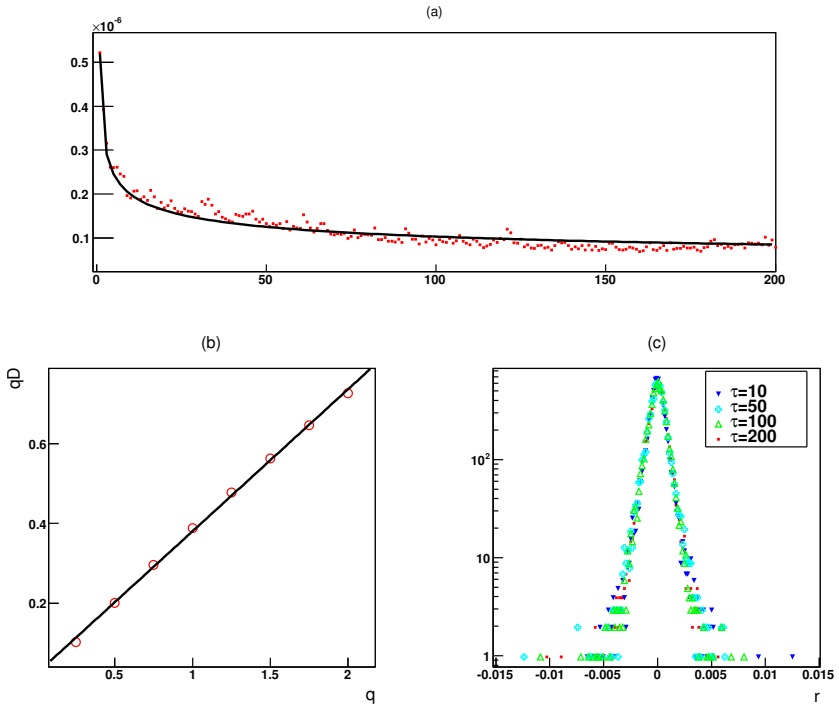

Figure 1.2: (a): Fit of $m_2^e(t,\tau) \sim t^{2D} - (t-1)^{2D}$ as a function of time. (b): Evidence of simple scaling behaviour for $m_q^e(t,t)$. The red circles are the couples (q, qD) evaluated from the logarithmic fits of $\langle |R_1 + \ldots + R_t| \rangle \sim t^{qD}$, with $q \in [0,2]$, $t \in [1,20]$. The black line is their linear fit, $D = 0.356$. (c): Scaling collapse for $p_{R_t(t)}$ according to eq. (1.9) of morning returns aggregated at $\tau = 10, 50, 100, 200$ minutes.

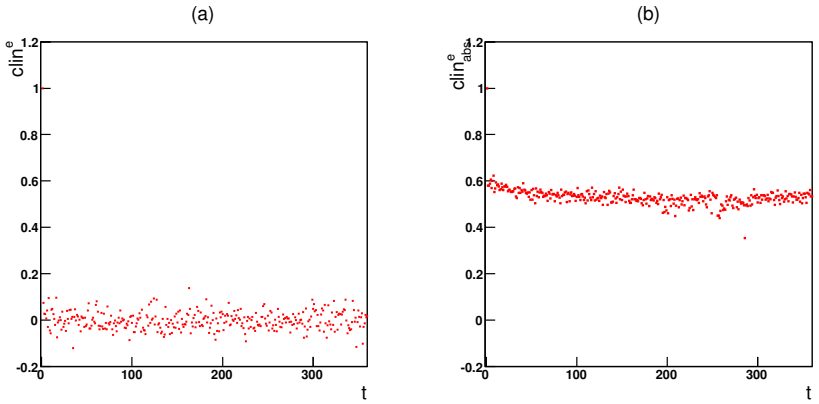

Figure 1.3: (a) shows the behaviour of the linear correlations $c_{lin}^e(1,t)$ during the day. (b) plots linear correlations computed from absolute values of the same data.

1.5. Night and interday returns

When looking for a model capable to describe the dynamics of a financial asset, to focus only on the intraday stock market fluctuations is not enough for a full description of the process that generates them: looking at interday features should enrich the formalism by considering the effects of exhogenous dynamics and/or long-range dependence. The most immediate implementation can be made by including night returns in our modelization: although New York stock exchange is not open all day long, indexes have not the same price at closure and opening. They show instead fluctuations which are the results of trading activity all over the world, and of political or social events happened during the night. In order to analyze the results of these influences two time series made up of all night and full day returns were isolated from our data, i.e. $r_n^l = \ln S^l(0) - \ln S^{l-1}(T)$, $r_d^l = \ln S^l(0) - \ln S^{l-1}(0)$, where $0, T$ stand for opening and closure time respectively, l for the considered day and n, d are the subscript indexes which indicate that we are evaluating returns aggregated respectively over a night and a full day interval. In this way we obtained two series of $N = 6851$ returns, i.e. one for each history of the ensemble, except the first one. Although we are now dealing with time lags longer than intraday ones, our aim is to look is their PDF, with a suitable rescaling parameter, can be collapsed to the same $g(r)$ of the intraday returns. This result would be very significant, since it would allow us to treat with the same model also interday returns, and to extend the process of aggregation beyond intraday framework. Then, among all the possible applications, we will exploit this feature to calculate conditioned PDFs which take into account what happened during the night. Here we report our first attempt to obtain the described result: figure 1.4 shows that, through the scaling parameters $\lambda_n = 8.0$, $\lambda_d = 13.8$, both fixed by hand, we obtain a good collapse of night and full day returns on intraday ones.

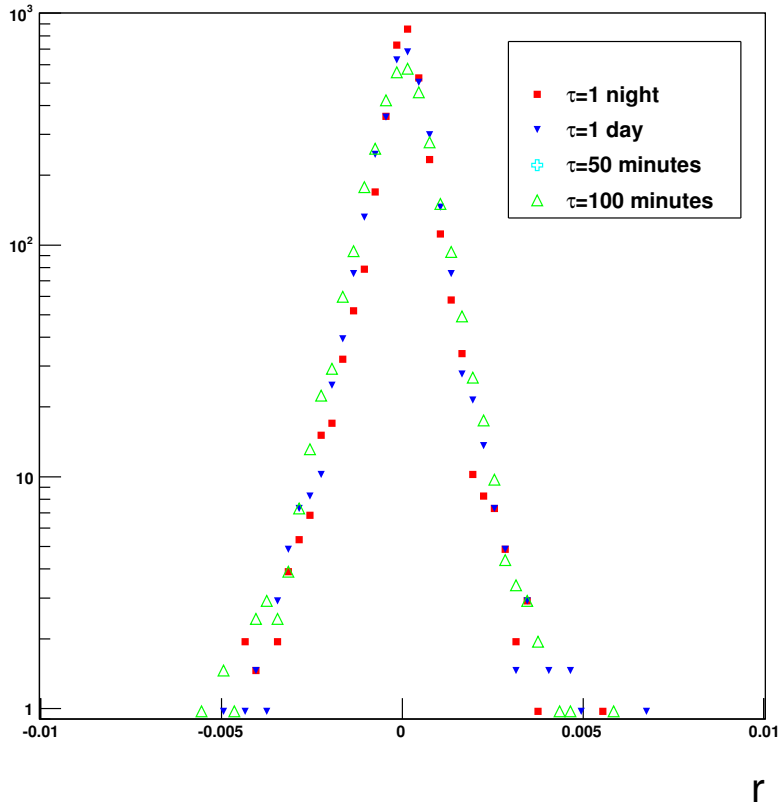

Figure 1.4: Scaling collapse of night (red squares) and full day (blue triangles) returns on intraday ones aggregated at $\tau = 50, 100$ minutes.

2. The model

In this chapter we will give the formulation for a model based on the empirical evidencies shown in the previous chapter, following the ideas developed in the works [11], [15]. The very first sections will be devoted to a brief introduction to stochastic processes, limit theorems and the historical evolution of financial modeling, in order to make this chapter as self contained as possible. In the central part we will introduce the mathematical structure of the model and we will describe its calibration protocol. Then we will present our result for the scaling of returns aggregated over night and full day intervals. Finally we will discuss the philosophy at the basis of the trading strategy, as a way to test the effect of the linear correlations on our model.

2.1. Stochastic processes

A stochastic process is a phenomenon whose dynamics is too complex to be analyzed in a deterministic way. Its complexity is usually due to the huge number of degrees of freedom involved in the process, so that an analytical approach would imply a terrific number of equations of motion to be solved. The simplest example that can be made is the toss of a coin: it is theoretically possible to study this process in a deterministic way and solve its equations of motion considering as degree of freedom the shape and the mass of the coin, the possible presence of wind, the magnitude and the direction of the force used to toss it, etcetera, but this approach practically involves an huge number of parameters and equations to be considered. Moreover, the evolution of the price of a financial asset is certainly a more complex process than the coin one: its variations (and hence its measures) depend on the combined effect of the actions of a large number of participants and the influences of external factors and, differently from the toss of a coin, we can expect the distribution of possible results to change in time. It is then clear that, for this kind of problems, it is much more convenient to treat the processes with a different approach. Therefore stochastic processes are studied through probabilistic

instruments, and they appear as random variables depending on time. Then, in order to describe a random process $X(t)$ whose outcomes are real numbers, we can use a probability density function (PDF) which, by definition, fixes the probability that $X(t)$ is within the small interval $[x, x+dx]$ at a certain time t as $P(x,t)dx$. Here we are considering the general case, in which stationarity is not assured and the PDF depends also on time. If we denote with $\mathcal{P}(\cdot)$ the probability measure of a certain event, we have that

$$\mathcal{P}_t(a < X(t) < b) = \int_a^b P(x,t)dx \tag{2.1}$$

is the probability that our process $X(t)$ is between a and b. In order to be well defined, a probability density function must satisfy the following conditions:

- It must be non negative, $P(x,t) \geq 0 \ \forall \ x \in \mathbb{R}, t \in \mathbb{R}^+$.
- It must be normalized, i.e.

$$\int_{x_m}^{x_M} P(x,t)dx = 1 \ , \tag{2.2}$$

  where x_m, X_M are respectively the smallest and the largest values which $X(t)$ can take. It is always possible to replace eq. (2.2) with $\int_{-\infty}^{+\infty} P(x,t)dx = 1$ just by setting to zero $P(x,t)$ in the intervals $]-\infty, x_m[$, $]x_M, +\infty[$.

2.2. Limit theorems

The Central Limit Theorem (CLT) [1] provides a fundamental background in statistical physics. It states that, given a set of independent, identically distributed (i.i.d.) random variables $\{X_i\}$, their sum $X = \sum_{i=1}^N X_i$ follows a Gaussian or Lévy distribution, with the two possibilities occurring respectively whether their variance is finite or not. Its demonstration is based on the stability of the limit distributions[1] and on the simple factorization which holds for the characteristic functions (CF)[2] of the sum of i.i.d. variables. One of the reasons of its importance lies in the fact that it allows to introduce for the variable X the scaling equation (1) in its simpler form ($D = 1/2$ and g Gaussian). When treating financial aggregated returns, which are by definition sums of elementary returns, $r_\tau(\tau) = \sum_{i=1}^\tau r_\tau(i)$, we would expect for them scaling behaviour according to $\tau^D p_{R_\tau}(\tau^D r) \to g(r)$ up to any intraday value of τ, since, as already shown

[1] A PDF f is called stable if, given two variables A and B f-distributed, their sum is still f-distributed.
[2] The characteristic function of a given PDF is defined as $\tilde{p}(k) = \int_{-\infty}^{+\infty} e^{ikx} p(x)dx$.

in figure 1.2(c), self-similarity is a property of the process itself. The problem here is that financial returns usually have not negligible nonlinear correlations, and cannot be considered as independent variables, as it is the case in many natural phenomena which show self-similarity features. Thus we need a PDF for dependent random variables capable to satisfy a scaling equation for every value of the number of summands τ. This implies the generalization of CLT with the reformulation of the necessary conditions to be fullfilled by the joint PDF, as developed in [16]. The most significant steps of this procedure will be reproduced when introducing the model.

2.3. Evolution of financial modeling

Physicists' contribution to mathematical finance has had a considerable modification along years, which matches with the evolution of the way in which stock market's dynamics has been considered and analyzed. The aim of this section is to give a brief introduction of past models and approaches in order to better understand the basis of financial modeling.

The most significant points on the basis of which a classification of the various models can be made are the shape of the PDF of returns, the presence of long memory effects and the behaviour of higher order moments. We will focus on the most common and historically significant models and show their approach to this matters.

2.3.1. Random walk in Finance

Mathematical finance has been dominated for decades by the random walk approach. The first step in this direction was made by Bachelier in 1900 [3]. His model associated the stock market's dynamics with a stochastic process based on the formulation of a random walk of infinite steps which advanced of five years Einstein's famous work on Brownian motion. Considered the price of a given asset $S(t)$, it focused on its increments $r_\tau(t) = S(t+\tau) - S(t)$ (without taking the logarithms), stating the following properties:

- The increments in different intervals of time are independent random variables.
- Their PDF $p_\tau(r,t)$ does not depend on time (stationarity).
- Each return calculated with a time lag τ follows a Gaussian distribution with zero mean and variance $\sigma^2\tau$.

The last two statement imply that the PDF for a single return with a time lag τ is

$$p_{R_\tau}(r) = \frac{1}{\sqrt{2\pi\sigma^2\tau}}\exp\Big(-\frac{r^2}{2\sigma^2\tau}\Big)\,, \tag{2.3}$$

and, exploiting the first property, we can write the joint PDF for n successive returns as

$$p_{R_\tau(1),\ldots,R_\tau(n)}(r_1, r_2, \ldots, r_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2\tau}}\exp\Big(-\frac{r_i^2}{2\sigma^2\tau}\Big)\,. \tag{2.4}$$

This model showed significant discrepancies with real market behaviour: in particular it allowed negative values for the prices and did not associate the behaviour of the returns with the value of the price itself, a feature contradicted by empirical evidence. In order to overcome these inaccuracies a new model, called Geometric Brownian motion, was defined by Black and Scholes [17]. In this model returns turn to be logarithmic, i.e. $r_\tau(t) = \ln S(t+\tau) - \ln S(t)$, but they still maintain the properties of Bachelier's model. This allows to overcome some of its problems without changing the structure of the PDF (which now has to be considered for logarithmic returns): the occurrence of negative prices is avoided and the size of fluctuations depends on the value of the price. The main point here, which is not touched by the reformulation of the model, is the randomness of the process. If we try to apply eq. (1) to this PDF it is straightforward to find $D = 1/2$ and g shaped as a Gaussian bell (as we would espect for the Central Limit theorem), due to the properties introduced to assure independent increments which follow a Normal distribution. This implies no memory for our process. The implications of this model became evident with the formulation of the concept of market efficiency, which was introduced by Fama in 1969 [18]. A market is called efficient in a semi-strong way if all the assets immediately adjust their price according to the information publicly known about them. This means that past prices are useless to forecast future behaviour for the asset, since all the relevant information is included in the current price. In order to make clearer the role played by an efficient market, we introduce the concept of *arbitrage*. In economy, an arbitrage is a set of financial operations which leads to a riskless profit, i.e, a profit without any initial investment. From a mathematical point of view it can be defined as a portfolio whose value $V(t)$ obeys the following conditions:

- $V(0) = 0$
- $V(t) \geq 0 \quad \forall\, t > 0$
- $V(T) > 0$ with a nonzero probability for some $T > 0$.

The first condition ensures the absence of any initial investment. The second and the third ones assert that the arbitrage cannot lead to a loss and that there is a certain prob-

ability to have a profit after a certain time T. It can be proved, within a mathematical framework, that the efficiency of a market implies no arbitrages and that, when adopting the random walk hypothesis, stock market trading becomes a fair game [19]. However, the storage of increasing amounts of data showed that the model of Geometric Brownian motion, although accounting for many stylized fact, provides just a first approximation for the underlying process: the real PDF of returns often shows leptokurtic shape and heavy tails which don't match with Gaussian distribution. Indeed, market efficiency is fully accomplished only for large time lags (τ of the order of months or years) and must be abandoned when dealing with the intraday dynamics of market stock.

2.3.2. Leptokurtosis and long range dependence

Mandelbrot, in 1963 [5], was the first one to introduce a different shape for the PDF for financial returns. Analyzing cotton prices along over a century, he proposed a model based on independent increments for the returns which followed a different series of distribution functions, the Lévy ones. Lévy distributions have fatter heavy tails, which are more useful to describe multiscale phenomena. Although there is no simple analytical expression for Lévy distributions, their asymptotic behaviour is

$$L_\mu(r) \sim \frac{\mu A^\mu}{|x|^{1+\mu}} \quad \text{for } x \to \pm\infty \,, \tag{2.5}$$

which shows the so called Pareto tails [14]. Thus Mandelbrot defined a stochastic process, known as Lévy flights, which is based on the following assumptions:

- The returns are independent random variables.
- Their PDF $p_\tau(r,t)$ does not depend on time (stationarity).
- Each return calculated with a time lag τ follows a Lévy distribution whose CF is $\tilde{p_1}(k) = \exp(-a|k|^\mu)$, with $a > 0$ and $0 \leq \mu \leq 2$.

If we recall the stylized fact introduced in section 1.2 about the statistical properties of rare events, it is now clearer what we meant when speaking of heavy tails: a Lévy distribution, which asymptotically follows a power law with exponent $-(1+\mu)$, decades faster than a Gaussian one, which is ruled by an exponential law. This feature makes extreme events become rarer when described by a Lévy distribution, and this matches better the empirical evidence of many different financial assets [6]. Lèvy distributions appear in the contest of Central Limit theorem since, because of their stability property under addition, a large class of distributions converge towards them under repeated

convolution. Indeed it can be proved that Mandelbrot's model describes a self-similar process with $D = 1/\mu$, and that, for $\mu \leq 2$, Levy variance diverge. This model still fails to represent multiscaling behaviour, higher order long range dependence and finiteness of the variance. The latter question can be corrected by introducing truncated Levy flights (TLF), stochastic processes whose returns follow a truncated version of Lèvy distributions, defined by

$$p(r) = \begin{cases} 0 \quad \text{for } r < -\alpha \\ bL_\mu(r) \quad \text{for } -\alpha \leq r \leq \alpha \\ 0 \quad \text{for } r > \alpha \, . \end{cases} \tag{2.6}$$

Although this PDF is not stable, due to its finite variance TLF asymptotically converge to a Gaussian process. In spite of this improvement, with TLF the problem of long range dependence for higher order moments still remains unaddressed.
A possible way to overcome this inaccuracy was proposed by Mandelbrot himself, who in 1968, together with Van Ness [20], introduced Fractional Brownian motion. This stochastic process was the first one which does not impose indepence for its increments, while leaving untouched stationarity and self-similarity properties. This allows the model to well explain such phenomena as volatility clustering and long range correlations, but showed also ergodicity problems, since linear correlations do not vanish.

As stated in section 1.1, volatility shows autocorrelations which make it increase after large returns and decrease in flatter periods. This behaviour is called heteroskedasticity and needs a model capable of considering long range dependence in order to be correctly described. The AutoRegressive Conditional Heteroskedasticity (ARCH) protocol has been introduced by Engle to take into account this phenomenon [21]. It includes a class of models which fix the returns as

$$r_k = \epsilon_k \sigma_k \, , \tag{2.7}$$

where ϵ_k are i.i.d. random variables with zero mean and unit variance and σ_k random processes linked to the size of fluctuations which determine the variance of r_k as σ_k^2. To properly describe heteroskedasticity $\{\sigma_i\}$ should show dependence on previous values: if σ_i is large we would expect that the probability of being followed by a large σ_{i+1} (although with an arbitrary sign) is high. Inside ARCH class there is a large variety of PDFs and different numbers of conditioning terms (called memory). The general definition of an ARCH model of memory q is

$$\sigma_k^2 = \alpha_0 + \alpha_1 r_{k-1}^2 + \cdots + \alpha_q r_{k-q}^2 \ , \tag{2.8}$$

where $\{\alpha_i\}$ is a set of parameters which shape the resulting PDF. The basic condition $\alpha_i \geq 0$ with $\alpha_0 > 0$ ensures the positiveness of the variance of returns, while more restrictive constraints can be found when imposing the finiteness of higher order moments of the process. From this formulation it is clear that past values of r_k, weighted by the parameters α_i's, influence the size of σ_k^2, as required by heteroskedasticity. ARCH models have been generalized to GARCH [22], in which a second set of parameters $\{\beta_i\}$ is introduced according to

$$\sigma_k^2 = \alpha_0 + \alpha_1 r_{k-1}^2 + \cdots + \alpha_q r_{k-q}^2 + \beta_1 \sigma_{k-1}^2 + \cdots + \beta_p \sigma_{k-p}^2 \ . \tag{2.9}$$

After this generalization also the past values of σ_k influence the process, through their squares σ_k^2. This quadratic structure, evident both in eq. (2.8) and (2.9), makes r_k react in the same way to either positive or negative returns, and this feature makes impossible a correpct description of the leverage effect, as presented in section 1.2. Moreover, although ARCH/GARCH model are still very common in finance, they are uncapable to show a well defined scaling behaviour.

2.4. An approach based on scaling

The model presented in this section has been implemented by Baldovin and Stella since 2007 to describe interday dynamics as a self-similar stochastic process within the framework of long time series [23]. However, its generality and independence on the value of time lag τ allow to make it suitable for intraday analysyis with the ensemble approach, as shown by the author themselves in [24]. Here we will present a formulation directly shaped for the latter approach.

As already stated, the basic statement in the development of the model is the validity of scaling behaviour according to eq. (1) in order to well describe the self-similarity of the process. As a starting point, we consider our series of returns $\{r_\tau^{(l)}(t)\}$, where t runs in the intraday interval and l refers to the day, i.e. the fixed history of the ensemble. We do not make any significant hypothesis on our set besides imposing self-similarity and absence of linear correlations, $\langle r_i r_j \rangle = 0 \ \forall i, j$. Moreover, since we want our model to show time inhomogeneity during the day to well describe the volatility pattern of figure 1.1, we can fix this general property for its PDF:

$$p_{R_\tau(1+(i-1)\tau)}(r) = p_{R_{a_i\tau}(1)}(r) \ , \tag{2.10}$$

where $R_\tau(1+(i-1)\tau)$ indicates a return of time lag τ generated at the time $t = 1+(i-1)\tau$. This property grants non-stationarity for our process by making the PDF's for different returns equal up to a rescaling factor of the time lag, a_i, which depends on the time interval $(i-1)\tau$ [3]. If now we take two successive returns $r_\tau(1)$, $r_\tau(2) = r_\tau(1+\tau)$ and write their joint PDF as ${p_{R_\tau(1),R_\tau(2)}}^{(2)}(r_1, r_2)$, we want it to satisfy the following conditions:

$$\int {p_{R_\tau(1),R_\tau(2)}}^{(2)}(r_1, r_2)\delta(r - r_1 - r_2)dr_1dr_2 = p_{R_{2\tau}(1)}(r) \ , \tag{2.11}$$

$$\int {p_{R_\tau(1),R_\tau(2)}}^{(2)}(r_1, r_2)dr_2 = p_{R_\tau(1)}(r_1) \ , \tag{2.12}$$

$$\int {p_{R_\tau(1),R_\tau(2)}}^{(2)}(r_1, r_2)dr_1 = p_{R_\tau(2)}(r_2) \ , \tag{2.13}$$

where the first equation produces the PDF for the aggregated return $r_{2\tau}(1)$, while the following two are required for causality. It is remarkable that, because of the inhomogeneity introduced with eq. (2.10), the right members of equations (2.12) and (2.13) are not equal, but differ by a scale factor. We are thus interested in a PDF capable to reproduce all the empirical features of our data: time inhomogeneity during the day through eq. (2.10), scaling behaviour according to eq. (1), absence of linear correlations (without cancelling higher order moments) and basic causality's condition (eq. (2.11)-(2.13)). A common solution, both in physics and in finance [24], to define a model capable to describe scaling features is to fix its $g(r)$ as a convex combination of Gaussian distributions, as proposed in [11]. This is a quite advantageous choice, since, starting from these convex combinations, it is possible to obtain a very large class of scaling functions. Moreover, another advantage of this choice is the fact that it generates a joint PDF which is a convolution of Gaussians with different widths, capable to reproduce the occurrence of not negligible and not stationary values of volatility for intervals of variable duration (volatility clustering), [4]. At this point Schoemberg's theorem [25] comes to help stating that the most general class of functions which satisfy the model

[3] It should be noticed that in eq. (2.10) we consider real values for the time lag. This is not theoretically wrong since, whatever finite set we define for the a_i's, time is still a discrete variable. Although, in this thesis we work only with data taken at τ minutes, with $\tau \in \mathbb{N}$, thus eq. (2.10) is just to be intended as an intermediate step to define a formalism practically consistent with our data.

conditions is of the form

$$g(r) = \int_0^\infty \rho(\sigma) \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} d\sigma \ , \tag{2.14}$$

$$\tag{2.15}$$

where ρ is a positive, normalized measure in $]0, +\infty[$. We now can make our first attempt to build from this $g(r)$ a joint PDF for n successive returns:

$$p_{R_\tau(1),\ldots,R_\tau(n)}(r_1, \ldots, r_n) = \int_0^\infty \rho(\sigma) \prod_{t=1}^{n} \frac{\exp\Big(-\frac{r_t^2}{2\sigma^2 a_t^2}\Big)}{\sqrt{2\pi\sigma^2 a_t^2}} d\sigma \ . \tag{2.16}$$

Here again we have the coefficient a_t's of eq. (2.10), which must be fixed in order to fit the $m_2^e(t,\tau)$ behaviour of figure 1.1(a). Moreover, from eq. (2.11), (2.14) and (2.16), it is straightforward to get, for the PDF of the aggregated return $r_t(t)$

$$p_{R_t(t)}(r) = \frac{g\left(r/\sqrt{a_1^2 + \ldots + a_t^2}\right)}{\sqrt{a_1^2 + \ldots + a_t^2}} \ , \tag{2.17}$$

while for a marginal return $r_\tau(t)$ the PDF $p_{R_\tau(t)}(r)$ is

$$p_{R_\tau(t)}(r) = \frac{1}{a_t} g\left(\frac{r}{a_t}\right) . \tag{2.18}$$

Now, if use eq. (2.18) to evaluate $m_2^e(t,\tau)$ we obtain

$$\begin{aligned} \langle r^2 \rangle_{p_{R_\tau(t)}} &= \int_{\mathbb{R}} r^2 p_{R_\tau(t)} dr \\ &= \int_{\mathbb{R}} r^2 \frac{1}{a_t} \, g\left(\frac{r}{a_t}\right) dr \\ &= a_t^2 \int_{\mathbb{R}} r'^2 \, g(r') dr' \\ &= a_t^2 \langle r^2 \rangle_g \ , \end{aligned} \tag{2.19}$$

where $\langle \cdot \rangle_{p_{R_\tau(t)}}$ and $\langle \cdot \rangle_g$ stand respectively for the averages evaluated with respect to $p_{R_\tau(t)}(r)$ and $g(r)$. If we compare eq. (2.17) with eq. (1) we obtain the condition

$$a_1^2 + a_2^2 + \cdots + a_t^2 = t^{2D} \quad \forall \, t \in [0, t_m] \ , \tag{2.20}$$

which leads to a definition of the a_t's as

$$a_t \equiv \sqrt{t^{2D} - (t-\tau)^{2D}} \, . \tag{2.21}$$

This is a good confirmation of the validity of our model, since, if we put eq. (2.21) in the second moment $m_2^e(t,\tau)$ evaluated in eq. (2.19), we reobtain what we alredy found empirically looking at figure 1.3(a), i.e. that $m_2^e(t,1) \sim (t^{2D} - (t-1)^{2D})$. Summarizing, we have so far developed a formalism to deal with general sets of non independent variables, allowing them to show strong nonlinear correlations while linear ones are put equal to zero. After introducing the scaling function $g(r)$ we have written the joint PDF for successive returns, using a set of scaling factors a_i's to grant time inhomogeneity. The function of the a_i's is to take into account the modifications in the distribution of returns during the system's evolution, after the starting point $t = 1$. A natural question here is whether the system has some restarting points t_r, in which $a_{t_r} = 1$, and how to detect them. In the long time series approach [23] the sequence of a_i's is stopped as an effect of exhogenous dynamics which make the system restart at random times with a certain probability. In this work we do not have to introduce such a mechanism or make particular hypotheses on the restarting procedure: since, guided by the evident periodicity of volatility, we are considering a collection of daily histories (each of which is a single realization of the general underlying process), the most natural choice is to make the system restart at the beginning of every single realization of the process, $a_1^{(l)} = 1 \; \forall \, l$.

The last step before defining the calibration procedure is identifying a proper parametrization for the scaling function through the choice of $\rho(\sigma)$. Here the choice is determined by the fitting of empirical data with the resulting g. It can also be noticed that the normalization of $\rho(\sigma)$ influences the correct normalization of g:

$$\int_{-\infty}^{+\infty} g(r)dr = \int_0^{+\infty} d\sigma\rho(\sigma) \int_{-\infty}^{+\infty} dr \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = \int_0^{+\infty} d\sigma\rho(\sigma) \, . \tag{2.22}$$

A convenient choice for $\rho(\sigma)$ that is capable at the same time to reproduce fat-tail behaviour in agreement with empirical data and to grant explicitly integration over σ is to use an inverse-gamma density function for σ^2. Thus $\rho(\sigma)$ becomes

$$\rho(\sigma) = \frac{2^{1-\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{\beta^\alpha}{\sigma^{\alpha+1}} \exp\Big(- \frac{\beta^2}{2\sigma^2} \Big) \, , \tag{2.23}$$

where $\alpha, \beta > 0$ are, respectively, a form and a scale parameter, whereas Γ is the Euler's gamma function. The resulting g after this choice of $\rho(\sigma)$, once performed integration

over σ, is then a Student's t-distribution:

$$g(r) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\beta}\left(1 + \frac{r^2}{\beta^2}\right)^{-\frac{\alpha+1}{2}} , \tag{2.24}$$

where $\alpha + 1$ shapes the decay of $g(r)$ at large $|r|$, while β sets the scale of its width.

This first formulation is valid in the morning time window, since when evaluating the Hurst exponent D we worked only with morning returns, $r_\tau(t)$ with $t \in [0, t_m]$. In the next sections we will define a proper extension of the model able to describe the data also in the afternoon session.

2.5. Calibration

Once the definition of the model is complete, its parameters can be calibrated in order to make a good fit of empirical data. From eq. (1) and (2.24) it is clear that the model is fully determined by its three parameters (D, α, β). As we already remarked, the daily plot in figure 1.1(a) shows that, at a certain moment t_m, volatility inverts its decreasing trend and starts growing. This implies that the set of parameters changes along the day, since, as shown in (2.19), the behaviour of volatility strictly depends at least on D. For this reason the calibration procedure is splitted in two stages, one for the morning and the other for the afternoon. In what follows we consider returns with $\tau = 1$ minute from 09:40 to 16:00 (New York time) and, since there is no reason left for ambiguity, we simplify our notation to $r_t = r_1(t)$ for the single returns and to $r_t^a = \sum_{i=1}^{t} r_i$ for the aggregated ones.

2.5.1. Morning calibration

The occurence of a defined scaling simmetry with a fixed D as in eq. (1) for our model implies a power-law behaviour also for the existing moments of the aggregated returns, according to

$$\langle |R_1 + \cdots + R_t|^q \rangle = \langle |R_1|^q \rangle \, t^{qD} . \tag{2.25}$$

The choice made for the scaling function g and the positive measure $\rho(\sigma)$ makes eq. (2.25) become

$$\langle |R_1 + \cdots + R_t|^q \rangle = \frac{\Gamma\left(\frac{q+1}{2}\right)\Gamma\left(\frac{\alpha-q}{2}\right)\beta^q}{\sqrt{\pi}\Gamma\left(\frac{\alpha}{2}\right)} \, t^{qD} . \tag{2.26}$$

We are now ready to start a multi-step calibration procedure structured as follows: we first determine the value of D for the first part of the day, D_m, then we use it to fit a collapse of all morning data according to eq. (1) and (2.18) and find α and β.

A straightforward way to find D_m is to use eq. (2.25): after calculating the empirical quantities corresponding to $\langle|R_1 + \cdots + R_t|^q\rangle$ and $\langle|R_1|^q\rangle$, defined as

$$\langle|R_1 + \cdots + R_t|^q\rangle_e \equiv \frac{1}{M}\sum_{l=1}^{M}|R_1^{(l)} + \cdots + R_t^{(l)}|^q \,, \tag{2.27}$$

we take their logarithms and fit them, expecting to find a linear behaviour with slope qD_m. Then we fit all the couples (q, qD_m) to determine D_m. We take q inside the interval $[0, 2]$ due to occurrence of multiscaling behaviour (i.e. $D(q)$ becomes a non-constant function of q for greater value of q). For our set of data taken from the S&P 500 with $\tau = 1$ minute, this procedure gives $D_m = 0.376$, which empirically confirms the occurrence of the anomalous scaling proposed for the morning window. The plot of the couples (q, qD_m) is shown in figure 2.2.

The next step is linking α_m and β_m. This can be made by evaluating $\langle|R_1|\rangle$ through a least squares fitting of eq. (2.25) with $q = 1$ and $D = D_m$, which gives $\langle|R_1|\rangle = 1.6\cdot10^{-3}$. Now, taking in consideration eq. (2.26) with $t, q = 1$, β_m can be written as a function of α_m. The last step is performing a full data-collapse of marginal and aggregated returns (according to eq. (2.18) and (1) respectively) and fitting it with the g proposed in (2.24), in order to find α_m, and thus β_m. The result of the data-collapse is shown in figure 2.1. Summarizing, for $\tau = 1$ minute the morning calibration gives the following parameter values: $D_m = 0.376$, $\alpha_m = 2.96$ and $\beta_m = 1 \cdot 10^{-3}$.

2.5.2. Calibration for the afternoon trading session

In order to deal with afternoon returns we have to extend the model by generalizing its structure with the introduction of different sets of parameters along the day. Before performing this action we rewrite the joint PDF for successive returns and the chosen form of $\rho(\sigma)$ after the transformation $\sigma \to \sigma/\beta$, which leads to

$$p_{R_1,\ldots,R_n}(r_1,\ldots,r_n) = \int_0^\infty \rho'(\sigma)\prod_{t=1}^{n}\frac{\exp\Big(-\frac{r_t^2}{2\sigma^2 a_t^2\beta^2}\Big)}{\sqrt{2\pi\sigma^2 a_t^2\beta^2}}d\sigma \,, \tag{2.28}$$

$$\rho'(\sigma) = \frac{2^{1-\frac{\alpha}{2}}}{\Gamma\big(\frac{\alpha}{2}\big)}\,\frac{1}{\sigma^{\alpha+1}}\exp\Big(-\frac{1}{2\sigma^2}\Big) \,, \tag{2.29}$$

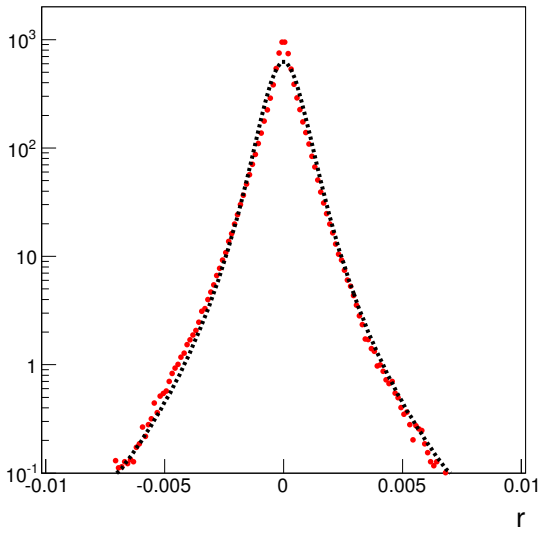

Figure 2.1: The scaling function for S&P 500 after collapse of marginal and aggregated returns (red dots) in the morning time window. The black dashed line is a fit of the points with the $g(r)$ of eq. (2.24).

and puts in evidence the role of β, together with a_t, as a rescaling parameter. Indeed, recalling eq. (2.19) it is evident that the increasing trend shown by volatility from t_m implies some $D > 1/2$ to be calibrated for this time window. The forms of eq. (2.28) makes clearer how to perform the extension of the model along the different phases of the day: the joint PDF for successive returns and the set of a_t's have to be generalized according to

$$a_t \equiv \sqrt{t^{2D_t} - (t-1)^{2D_t}} \ , \tag{2.30}$$

$$p_{R_1,\ldots,R_n}(r_1,\ldots,r_n) = \int_0^\infty \rho'(\sigma) \prod_{t=1}^{n} \frac{\exp\Big(-\frac{r_t^2}{2\sigma^2 a_t^2 \beta_t^2}\Big)}{\sqrt{2\pi\sigma^2 a_t^2 \beta_t^2}} d\sigma \ , \tag{2.31}$$

with

$$D_t \equiv \begin{cases} D_m & \text{if } \ 1 \le \tau \le t_m \\ D_a & \text{if } \ t_m < \tau \le t_f \ , \end{cases} \tag{2.32}$$

and

$$\beta_t \equiv \begin{cases} \beta_m & \text{if } \ 1 \le t \le t_m \\ \beta_a & \text{if } \ t_m < t \le t_f \ , \end{cases} \tag{2.33}$$

where $D_a > 1/2$. After explicit integration over σ we obtain

$$p_{R_1,\dots,R_n}(r_1,\dots,r_n) = \left(\prod_{t=1}^{n}\frac{1}{a_t\beta_t}\right)\frac{\Gamma\left(\frac{\alpha+n}{2}\right)}{\pi^{\frac{n}{2}}\Gamma\left(\frac{\alpha}{2}\right)}\left[1+\left(\frac{r_1}{a_1\beta_1}\right)^2+\cdots+\left(\frac{r_n}{a_n\beta_n}\right)^2\right]^{-\frac{\alpha+n}{2}} . \tag{2.34}$$

The just performed generalization of the model is conceived not to inhibit the occurrence of scaling simmetry, and it just modifies its parameters: the starting eq. (1) becomes now

$$p(r,\tau) = \frac{1}{\lambda(\tau,t_m)}\; g'\Big(\frac{r}{\lambda(\tau,t_m)}\Big) , \tag{2.35}$$

where

$$\lambda(\tau,t_m) \equiv \Big(\sum_{t=1}^{\tau} a_t^2\beta_t^2\Big)^{1/2} = \begin{cases} \beta_m\tau^{D_m} & \text{if } \; 1\le t\le t_m \\ [\beta_m^2(t_m)^{2D_m}+\beta_a^2[\tau^{2D_a}-(t_m)^{2D_a}]]^{1/2} & \text{if } \; t_m < t \le t_f \end{cases} , \tag{2.36}$$

and

$$g'(r) = \frac{\Gamma(\frac{\alpha+1}{2})}{\sqrt{\pi}\Gamma(\frac{\alpha}{2})}(1+r^2)^{-\frac{\alpha+1}{2}} . \tag{2.37}$$

This procedure is consistent with the basic formulation of the model: eq. (2.35) reduces to eq. (1) taking $\tau \le t_m$ in eq. (2.36), considering that $g(r) = \frac{1}{\beta_m}\, g'\big(\frac{r}{\beta_m}\big)$. Indeed also the form the PDF for the marginal returns modifies to

$$p_{R_t}(r) = \frac{1}{a_t\beta_t}\, g'\Big(\frac{r}{a_t\beta_t}\Big) . \tag{2.38}$$

Finally, eq. (2.26) for the moments of the aggregated returns becomes

$$\langle |R_1+\cdots+R_t|^q\rangle = \frac{\Gamma\Big(\frac{q+1}{2}\Big)\,\Gamma\Big(\frac{\alpha-q}{2}\Big)}{\sqrt{\pi}\Gamma\Big(\frac{\alpha}{2}\Big)}\,[\lambda(\tau,t_m)]^q . \tag{2.39}$$

We can now complete the calibration protocol by identifying β_a and D_a. From eq. (2.28) it is clear that the product $a_t\beta_t$ determines the width of the marginal PDF $p_{R_t}(r)$. Thus, we can fix β_a as a function of β_m, D_m and D_a by imposing the continuity constraint $\beta_m a_{t_m}(D_m) = \beta_a a_{t_m}(D_a)$, which can be explicitly written as

$$\beta_a = \beta_m\frac{[(t_m)^{2D_m}-(t_m-1)^{2D_m}]^{1/2}}{[(t_m)^{2D_a}-(t_m-1)^{2D_a}]^{1/2}} . \tag{2.40}$$

The last parameter to find is D_a. To this aim we use the same approach applied to

the determination of D_m, although referring to eq. (2.39) with $t \geq t_m$, and we make a least square fit of $\langle |R_1 + \cdots + R_t|^q \rangle_e$ with $q = 1$. The result is $D_a = 1.23$ and thus $\beta_a = 6 \cdot 10^{-6}$.

The complete calibration procedure thus gives the complete set of parameters $(\alpha, D_m, \beta_m, D_a, \beta_a) = (2.96, 0.38, 1 \cdot 10^{-3}, 1.23, 6 \cdot 10^{-6})$. Figure 2.3 shows how the model well reproduces the first empirical moments of the aggregated returns along the whole day.

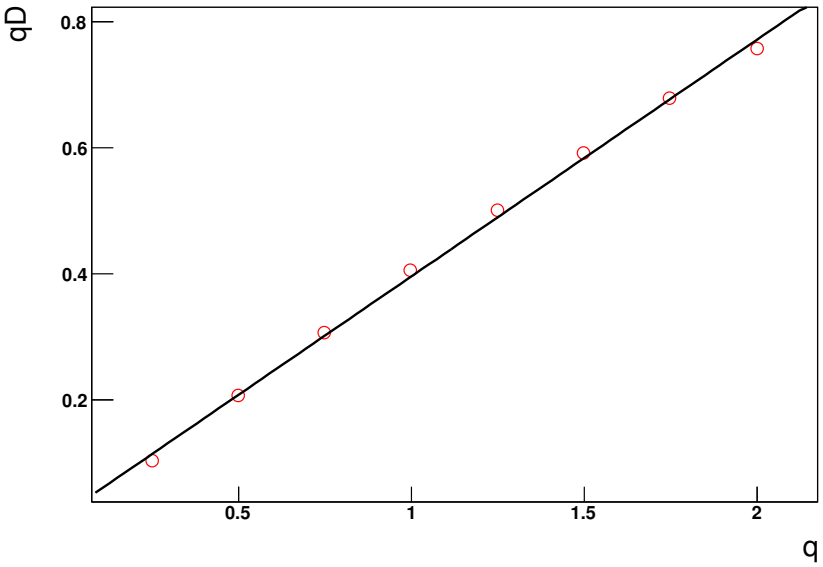

Figure 2.2: Scaling behaviour of the non-linear moments according to the model. The red circles are the couples (q, qD) evaluated from the logarithmic fits of $\langle |R_1 + \ldots + R_t| \rangle \sim t^{qD}$, with $q \in [0, 2]$, $t \in [1, t_m]$. The black line is their linear fit, $D = 0.376$.

2.6. Scaling of night and day returns

The so far implemented model works only in the intraday framework, and studies the scaling properties of returns taken in this framework. An interesting direction which can be explored is the study of a data collapse made up of night and full day returns, in order to bridge intraday and interday data treatment. In fact, if interday returns turn out to be described by the same scaling function of our model, we will obtain a way to describe a really large class of financial quantities exploiting the same formalism.

The most immediate way to explore this possibility is to study the collapse of interday

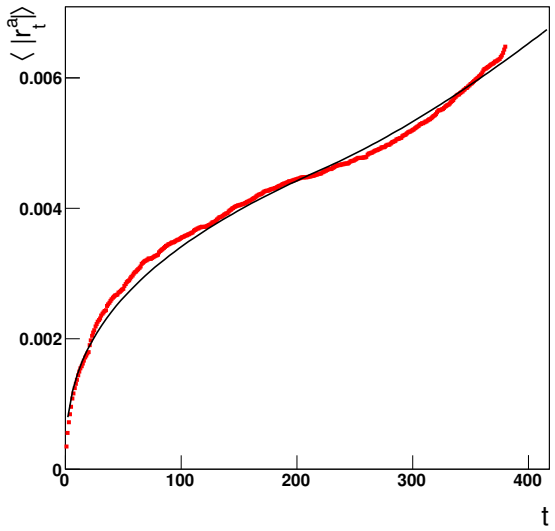

Figure 2.3: Data fit of the first empirical moments of the absolute value of aggregated returns of the S&P 500 index, $\langle |R_1 + \ldots + R_t| \rangle$, for the whole day, $t \in [1, t_f]$.

returns on the extended $g'(r)$ of the model, from eq. (2.37), trying to fix by hand a suitable scaling parameter, according to

$$p_{R_n}(r) = \frac{1}{\lambda_n} \, g'\Big(\frac{r}{\lambda_n}\Big) \, , \tag{2.41}$$

which recalls the structure of eq. (2.38). This attempt gave good results: figure 2.4 shows the collapse of returns aggregated over nights (r_n) and full days intervals (taken as $r_d = \ln(t_{i_2}) - \ln(t_{i_1})$, where t_{i_1}, t_{i_2} are the opening times of two successive trading days) on the intraday returns with $\tau = 1$ minute. It is rather clear that these new collapsed returns still follow the $g'(r)$. The found scaling parameters for night (λ_n) and day (λ_d) collapse are $\lambda_n = 8.0 \cdot 10^{-3}$ and $\lambda_d = 1.40 \cdot 10^{-2}$. This is an interesting result, because it allows to bridge in a totally new way intraday and interday tractation through the same formalism. Indeed, although previous works have already studied how to use intraday returns to obtain better prediction of volatility on an interday scale [26], this is the first successful attempt to describe with the same scaling function, calibrated through intraday data, also returns aggregated over time scales of different orders of magnitude. In the next chapter we will look for a way to exploit this result, extending the aggregation of returns beyond intraday limits, in order to condition our trading strategy.

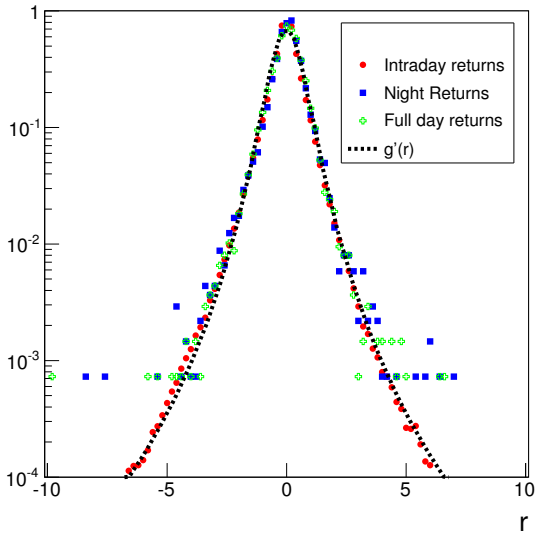

Figure 2.4

2.7. Density forecasts

In this section we show that our model holds implicit forecasting capabilities: once specified the frequency through which we are looking at the system and its calibration parameters, it provides a PDF for the evolution of marginal returns which allow us to calculate the expected volatility along the whole day. Moreover, according to eq. (1), this procedure can be extended to any aggregation of returns within the day. Combining these tools with the possibility of conditioning the PDF to the returns which have come out gives us the opportunity to implement a trading strategy in which PDF's forecasts are used as signals to open or close market positions. To this aim two different approaches are available. The first one uses the properly calibrated PDF and the first index value available for the day to calculate the density forecasts for the whole trading session, and we will refer to it as *unconditioned trading* since it does not take into account any daily return (i.e. $t_p = 0$). The other one instead modifies the density forecasts by conditioning the PDF to more recent returns (up to $t_p > 0$).

However, both approaches are based on the same idea: given a certain value $0 < Q < 1/2$, the quantile function $Q(1-Q)$ of the expected returns, extracted from their unconditioned PDF at a certain time of the trading session, is taken as a lower(upper) threshold for the real index value. When this threshold is passed, we have a sell(buy) signal. The same procedure applies in the case of conditioned PDFs, in which the first

returns to the day contribute to shape the quantile barriers. In other words, we look at this signal as the potential occurrence of a trend which overcomes random fluctuations and we try to exploit this trend. It is remarkable that this procedure is expected to give defined positive values of trading gain only if the returns are somehow linearly correlated: actually, our empirical data show small, non-vanishing linear correlations, as already shown in figure 1.3. Indeed, when this trading strategy is applied to histories numerically generated on the basis of the model (which is conceived with $\langle r_i r_j \rangle = 0 \, \forall i, j$) profit values vanish. Thus, this strategy can also be seen as a test of the effect, on our data, of the presence of non zero linear correlations, whithout forcing our formalism to include them applying complex modifications to the model. Indeed, the profit obtained through our trading strategy is a quantifier of the effect of linear correlations, which in our model are set equal to zero, on the behaviour of the data: we will see that the occurrence of positive, negative or negligible values of profit is directly related to the pattern of these correlations.

2.7.1. Unconditioned trading

In this approach, for each day l, we calculate the quantile function from the model PDF at time t, considering the opening value $S_0^{(l)}$. It can be noticed that, when computing density forecasts, we do not take into account any information on the previous asset's behaviour: all these pieces of information are assumed to be included in the parameters $(\alpha, D_m, \beta_m, D_a, \beta_a)$ after the the calibration procedure. Actually, two different approaches can be followed when implementing a trading strategy. We can work in sample, which means that we calibrate the model on all the M =6852 available trading days, and then we apply the strategy on the whole data set. In this way, the returns of each day in which we apply the strategy have actively been considered in the calibration procedure. Another solution is to operate out of sample, which means that we calibrate the model on a fixed subset of the data set and then we apply the strategy on the following days. In this case, in the model there is no information about the days in which the strategy is applied. Since we are more interested in testing the effects of linear correlations rather than looking at the predictive power of the model, in this thesis we worked only in sample.

After the extension of the model performed in last section, the PDF for aggregated

returns at a time t is

$$p_{R_1+\cdots+R_t}(r,t) = \frac{\Gamma(\frac{\alpha+1}{2})}{\sqrt{\pi}\Gamma(\frac{\alpha}{2})} \frac{1}{\lambda(t,t_m)} \left[1 + \left(\frac{r}{\lambda(t,t_m)}\right)^2\right]^{-\frac{\alpha+1}{2}} . \tag{2.42}$$

From this PDF we can find the lower threshold for the expected aggregated returns for a given value of Q by numerically solving the following equation with respect to $r_{min,t}(Q)$:

$$Q = \int_{-\infty}^{r_{min,t}(Q)} p_{R_1+\cdots+R_t}(r,t)dr . \tag{2.43}$$

Due to the parity of the PDF of the model the corresponding upper threshold values $r_{max,t}(Q)$ can be obtained just through a sign flip $r_{max,t}(Q) = -r_{min,t}(Q)$. Once the lower and upper thresholds for returns are obtained, the corresponding price values are easily computed: the asset price at the time t of the day l, $S_t^{(l)}$, is a monotonic function of the aggregated returns, and it can be calculated through

$$S^{(l)}(t) = S_0^{(l)} \exp\left(\sum_{i=1}^{t} r_i\right) . \tag{2.44}$$

It can be noticed that the distribution of the expected returns calculated at the quantile level Q throug eq. (2.43) is totally independent of the particular day we are taking into account, since only when evaluating the price $S_t^{(l)}$ we need to refer to $S_0^{(l)}$.

We have so obtained, for every value of $0 < Q < 1/2$, two barriers $S_{min,t}(Q), S_{max,t}(Q)$, within which the asset price $S^{(l)}(t)$ has a probability $1-2Q$ of being confined at any time t of the daily range. The trading strategy which will be explicitly implemented in the next sections is based on the comparison between these barriers and the real price, which leads to the definition of a buy/sell signal. In figure 2.5 is shown, as an example, the upper and lower thresholds at different quantile levels, together with the daily evolution of the price of the index, for the day October 11th, 1985.

2.7.2. Conditioned trading

We can implement the procedure presented in the last section by using a conditioned PDF when evaluating the price barriers. In this way we fully exploit all the features of the model, taking advantage of its non-Markovian behaviour due to its general formulation. Thus, in what follows, beside of the opening value $S_0^{(l)}$ we use also the first t_p returns to condition the subsequent expected evolution of the index. According to probability

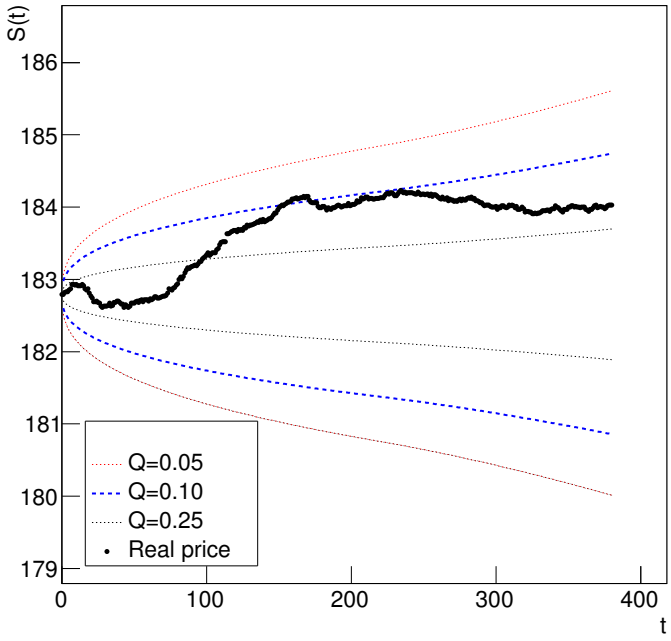

Figure 2.5: Upper and lower expected index values (dashed lines) for October 11th 1985, confronted with real prices (black dots).

theory, the conditioned distribution function of the sum of returns $r_{t_p+1} + \cdots + r_{t_p+t}$ given the previous ones $r_1, \ldots, r_{t_p}$ is given by

$$p(r, t|r_1, \ldots, r_{t_p}) = \frac{p_{R_t(t),R_1,\ldots,R_{t_p}}}{p_{R_1,\ldots,R_{t_p}}} \,. \tag{2.45}$$

Rewriting eq. (2.45) through eq. (2.34) gives

$$\begin{aligned} p(r, t|r_1, \ldots, r_{t_p}) =& \frac{\Gamma\left(\frac{\alpha+t_p+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha+t_p}{2}\right)} \left(\sum_{i=t_p+1}^{t_p+t} a_i^2\beta_i^2\right)^{-1/2} \cdot \\ & \cdot \frac{\left[1 + \frac{r^2}{\sum_{i=t_p+1}^{t_p+t} a_i^2\beta_i^2} + \left(\frac{r_1}{a_1\beta_1}\right)^2 + \cdots + \left(\frac{r_n}{a_n\beta_n}\right)^2\right]^{-\frac{\alpha+t_p+1}{2}}}{\left[1 + \left(\frac{r_1}{a_1\beta_1}\right)^2 + \cdots + \left(\frac{r_{t_p}}{a_{t_p}\beta_{t_p}}\right)^2\right]^{-\frac{\alpha+t_p}{2}}} \,. \end{aligned} \tag{2.46}$$

Thus we can write the equation defining the conditional quantile function, after a simple change of variable, as

$$\begin{aligned} Q &= \int_{-\infty}^{r_{min,t}(Q)} p(r, t|r_1, \ldots, r_{t_p}) dr \\ &= \frac{\Gamma\left(\frac{\alpha+t_p+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha+t_p}{2}\right)} \int_{-\infty}^{z_{min,t}(Q)} [1 + z^2]^{-\frac{\alpha+t_p+1}{2}} \,, \end{aligned} \tag{2.47}$$

where

$$z_{min,t}(Q) \equiv \frac{r_{min,t}(Q)}{\left(\sum_{i=t_p+1}^{t_p+t} a_i^2\beta_i^2\right)^{1/2} \left[1 + \left(\frac{r_1}{a_1\beta_1}\right)^2 + \cdots + \left(\frac{r_n}{a_n\beta_n}\right)^2\right]^{1/2}} \,. \tag{2.48}$$

Once eq. (2.47) is numerically solved, $r_{min,t}(Q)$ and $r_{max,t}(Q)$, can be used as in the past section to obtain $S^{(l)}_{min,t}(Q)$ and $S^{(l)}_{max,t}(Q)$ from $S^{(l)}_0$, and the first t_p returns. An example is shown in figure 2.6, in which the price barriers at different t_p's are plotted for the day November 6th, 1989. It can be noticed that the horizontal bells of price barriers show different widths at different t_p's: this means that conditioned trading is characterized by bounds which may vary due to the influence of the first returns of the day. In particular, if during the conditioning interval t_p the volatility is high the quantile barriers are larger, since the PDF has been conditioned with a set of returns whose absolute values are big. On the opposite, when the volatility is small in the first

part of the day, the resulting quantile barriers become closer. We recall that in this approach all the conditioning information are taken within the same day in which the trading strategy is implemented, and that the effects of global dynamics are supposed to be contained in the model itself.

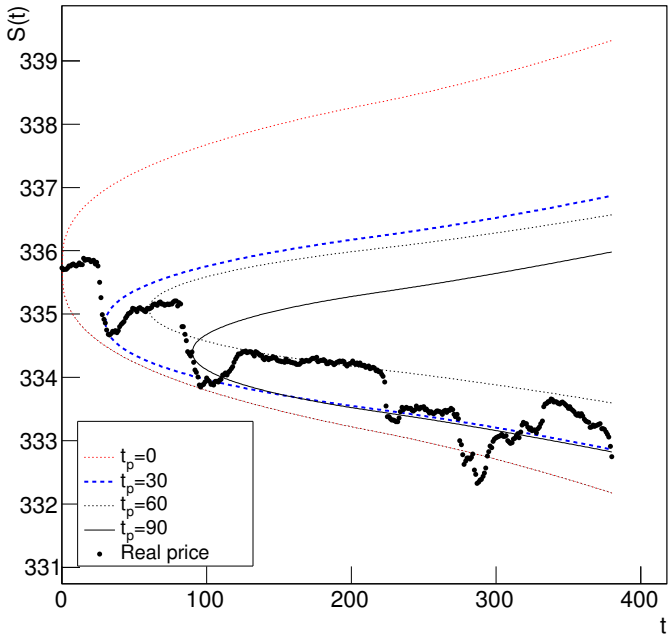

Figure 2.6: Upper and lower expected asset value [Q=10%] (dashed lines), for November 6th 1989, at different numbers of conditioning returns. The black dots plot the real prices.

2.7.3. Defining a trading strategy

Once the $S^{(l)}_{min,t}(Q)$ and $S^{(l)}_{max,t}(Q)$ barriers are calculated for every t of the intraday range $[t_p + 1, t_f]$, they can be used to create trading signals based on the real evolution of the market. To this purpose we are interested in the periods in which the index value breaks through the expected upper or lower quantile. The considered time interval for each day goes from 09:40 a.m to 16:00 p.m. New York time, and the trading session is active from 09:40 a.m.$+(t_p + 1) \times \tau$. If at market closing the trade is still open we close it anyway, in order to avoid the external influences which may act in the close-to-open time frame. In what follows, we will distinguish between long and short positions, which

are two different classes of financial operation. An opened position is called long if its holder invests on a certain security, i.e. he buys it expecting a positive trend. It is closed when the security is sold, and there is a profit if its value has been increasing in time. Differently, a short position is opened when the holder sells securities which he doesn't own, and subsequently repurchases them. In this case there is a profit if their price has decreased.

Thus, given a value of Q, within a certain day l the trading signals and the trading activity are as follows:

1. If there are no open positions:
   (a) Buy if $S_t^{(l)} > S_{max,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} < S_{max,t-1}^{(l)}(Q)$ (open a long position)
   (b) Sell if $S_t^{(l)} < S_{min,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} > S_{min,t-1}^{(l)}(Q)$ (open a short position)
2. If there are open positions:
   (a) Sell if $S_t^{(l)} < S_{max,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} > S_{max,t-1}^{(l)}(Q)$ (close a long position)
   (b) Buy if $S_t^{(l)} > S_{min,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} < S_{min,t-1}^{(l)}(Q)$ (close a short position)
   (c) Close any long/short position still open at the end of the day.

The just defined trading strategy exploits the potential occurrence of positive(negative) trends signaled by the breaking through of the price barriers according to the scheme of long(short) positions. The trading session during a single day can show different patterns: many operations can be opened and closed during high volatility days, single operations take place during trending days, while no operations are opened in stable days.

3. Trading on S&P 500

In this chapter we exploit the tools of the model so far developed to study more deeply the data set we introduced in section 1.4. We chose to analyze the S&P 500 returns $r_\tau(t)$ at different time lags, $\tau = 1, 2, 5, 10$ minutes. This means that we took from the same data set four different ensembles with the same number of daily realizations. Obviously the number of returns in each day depends on the chosen frequency. These are 380, 190, 76 and 38 respectively with a time lag of 1, 2, 5 and 10 minutes. We followed the calibration procedure for the various ensembles obtaining different sets of parameters, and then we checked the fitting capability of the model for the different values of frequency. Finally we evaluated the quantile barriers at different quantile levels and we applied the trading strategy previously defined. We will present our results in three different sections, the first one of which is devoted to the different behaviour of correlations of the various ensembles, the second one shows the products of the calibration protocol and the scaling properties of nights and full days and the last one analyzes the outcomes of our trading strategy.

3.1. Linear correlations and volatility

The most immediate difference among the S&P 500 returns at different frequencies can be found by looking at theirs linear correlations. As one could expect when thinking at the behaviour of the logarithmic returns defined in eq. (1.1), the larger the time lag is taken, the more the linear correlations between successive returns will be small. Indeed, as already stated when introducing the stylized facts in section 1.2, linear correlations of the form $\langle r_\tau(t+T)r_\tau(t)\rangle$ show a decaying behaviour when T increases and became irrelevant when T breaks through the value of about 10 minutes. Then, since T must be a multiple integer of τ in order to make $\langle r_\tau(t+T)r_\tau(t)\rangle$ well defined, when we evaluate linear correlations between successive returns at different time lags, $\langle r_\tau(t+\tau)r_\tau(t)\rangle$, $T = 1 \times \tau$, we would expect to find decreasing values when τ grows. In figure 3.1 it can

be observed that the empirical linear correlations evaluated from our data set according to eq. (1.7) well follow this theoretical behaviour. We have already mentioned that the size of linear correlations plays an important role when applying the trading strategy: the larger they are, the higher the probability of indentifying a real trend becomes. We will see in the next sections how this process interferes when applying our trading strategy.

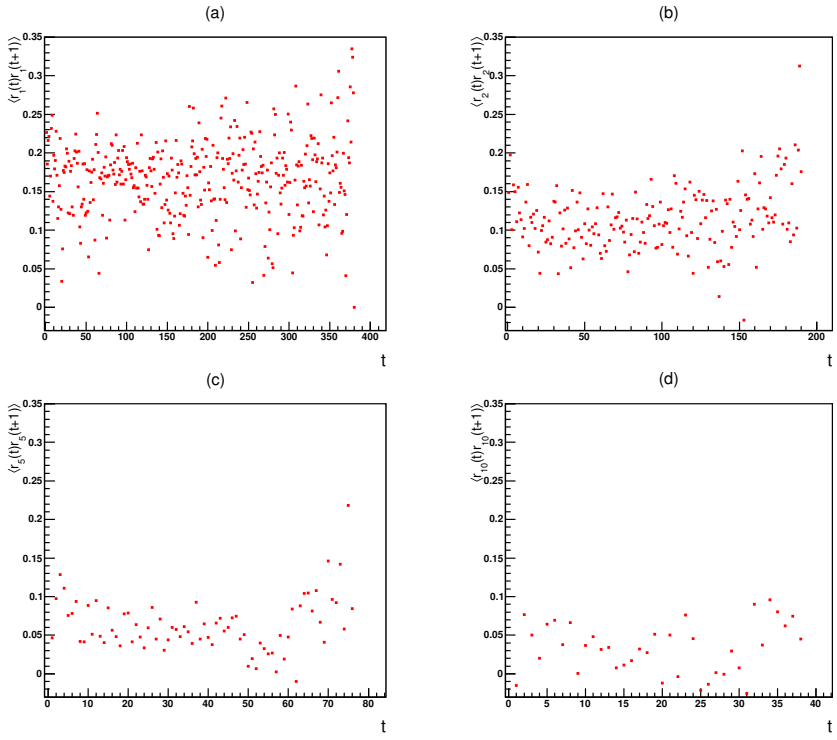

Figure 3.1: Empirical linear correlations between successive returns, evaluated according to $c^{e}_{lin,\tau}(t,t+1) \equiv \frac{\frac{1}{M}\sum_{l=1}^{M} r_\tau(t)t^l r_\tau(t+1)^l}{\sqrt{m_2(t,\tau)m_2(t+1,\tau)}}$, at τ =1, 2, 5, 10 of the S&P 500 index are respectively shown in the panels (a), (b), (c), (d).

We already stated in section 1.4 that along intraday evolution volatility shows a clear U-shaped pattern linked to the volume of trading activities during the day. This empirical evidence, together with the determination of the power law of eq. (2.19) which describes the behaviour of volatility, led us to distinguish between morning and afternoon trading and to develope a different calibration protocol for each time window. Thus, a graphical plot of volatility along the day, such as the one showed in figure 1.1(a), is useful to determine the t_m which separates mornings from afternoons. Here we introduce the

precise expression of volatility autocorrelation, defined as

$$c_{vol,\tau}(t) \equiv \frac{\frac{1}{M}\sum_{l=1}^{M}|r_\tau(1)^{(l)}||r_\tau(t)^{(l)}| - \left(\frac{1}{M}\sum_{l=1}^{M}|r_\tau(1)^{(l)}|\right)\left(\frac{1}{M}\sum_{l=1}^{M}|r_\tau(t)^{(l)}|\right)}{\frac{1}{M}\sum_{l=1}^{M}|r_\tau(1)^{(l)}|^2 - \left(\frac{1}{M}\sum_{l=1}^{M}|r_\tau(1)^{(l)}|\right)^2} \,. \quad (3.1)$$

In figure 3.2 we plot the four different sets of $c_{vol,\tau}(t)$ evaluated at different time lags. It can be noticed that we obtained totally overlapping patterns, where $t_m^\tau = \frac{220}{\tau}$ is the point in which volatility autocorrelations start to grow again.

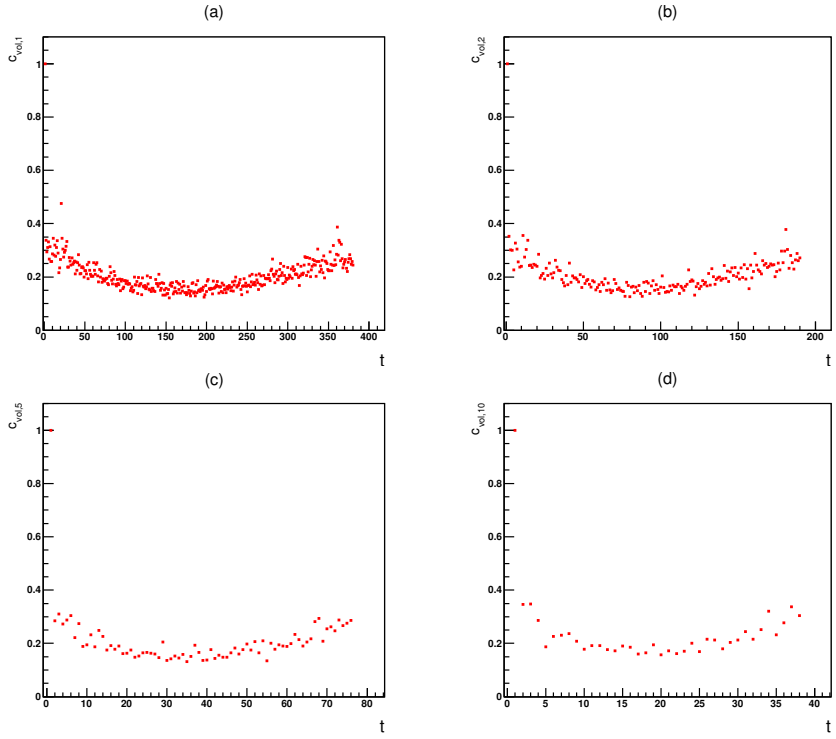

Figure 3.2: Volatility autocorrelation at τ =1, 2, 5, 10 of the S&P 500 index are respectively shown in the panels (a), (b), (c), (d).

3.2. Calibration and scaling results

Once the four different collections of daily realizations were obtained from our main data set, they were separately calibrated by performing on each of them the procedure described in section 2.5. In table 3.1 we report the resulting calibration parameters. The errors on α, D_m, D_a have been obtained through their fitting procedure, while for β_m, β_a

we propagated the errors of the others parameters through eq. (2.26),(2.40).

	$\tau = 1$	$\tau = 2$	$\tau = 5$	$\tau = 10$
α	3.68 ± 0.05	3.68 ± 0.06	3.67 ± 0.04	3.58 ± 0.06
D_m	0.376 ± 0.006	0.369 ± 0.006	0.358 ± 0.007	0.359 ± 0.006
β_m	$1.07 \pm 0.01 \cdot 10^{-3}$	$1.44 \pm 0.01 \cdot 10^{-3}$	$2.10 \pm 0.01 \cdot 10^{-3}$	$2.84 \pm 0.02 \cdot 10^{-3}$
D_a	1.23 ± 0.2	1.18 ± 0.03	1.28 ± 0.04	1.20 ± 0.03
β_a	$5.9 \pm 0.6 \cdot 10^{-6}$	$1.8 \pm 0.3 \cdot 10^{-5}$	$3.4 \pm 0.4 \cdot 10^{-5}$	$1.09 \pm 0.08 \cdot 10^{-4}$

Table 3.1: Sets of calibration parameters at different time lags.

We can observe that the parameters which show the most consistent variations of value are β_m and β_a. In fact, through eq. (2.25),(2.39), they are arithmetically proportional to $\langle |R_\tau(1)| \rangle$, whose size shows an increasing behaviour as a function of τ. The parameters α, D_m and D_a maintain themselves almost constant when the considered time lag changes. To verify the goodness of the calibration parameters just presented, we used them to perform a fitting of the empirical first moments of the S&P 500 returns. The resulting fits are showed in figure 3.3. An additional check to verify the goodness of the four different sets of parameters can be performed by calculating the scaling parameter $\lambda(t_f, t_m)$ of eq. (2.36), where t_f is the closure time of stock market. In fact the generic $\lambda(\tau, t_m)$ quantity is the scaling parameter of the aggregated returns of time τ, $r_\tau(\tau)$, according to eq. (2.35). Therefore, if we take $\tau = t_f$ we obtain a return aggregated over a whole daily trading session, $r_{t_f}(t_f) = \ln S(t_f) - \ln S(0)$, which does not depend on the particular τ. The corresponding scale parameter, $\lambda(t_f, t_m)$, should then be the same for every value of τ, independent of the set of parameters which we are using to evaluate it. The calculation of its value at different frequencies confirms our expectations: table 3.2 shows the four values found, whose relative variations are below 1%. In this case the errors have been evaluated by propagating the parameters' ones through eq. (2.36).

	1 minute	2 minutes	5 minutes	10 minutes
$\lambda(t_f, t_m)$	0.0111 ± 0.0008	0.0111 ± 0.0009	0.0112 ± 0.0010	0.0111 ± 0.0010

Table 3.2

In the previous chapter we introduced the possibility of collapsing interday returns on the same $g(r)$ of intraday ones. Here we analyze and comment the results found at all frequencies. Figure 3.4 shows the four corresponding night collapses: the scaling parameter is the same for all figures, $\lambda_n = 8.0 \pm 0.2 \cdot 10^{-3}$. It can also be noticed that the various collapses do not show any difference when the frequency changes. From

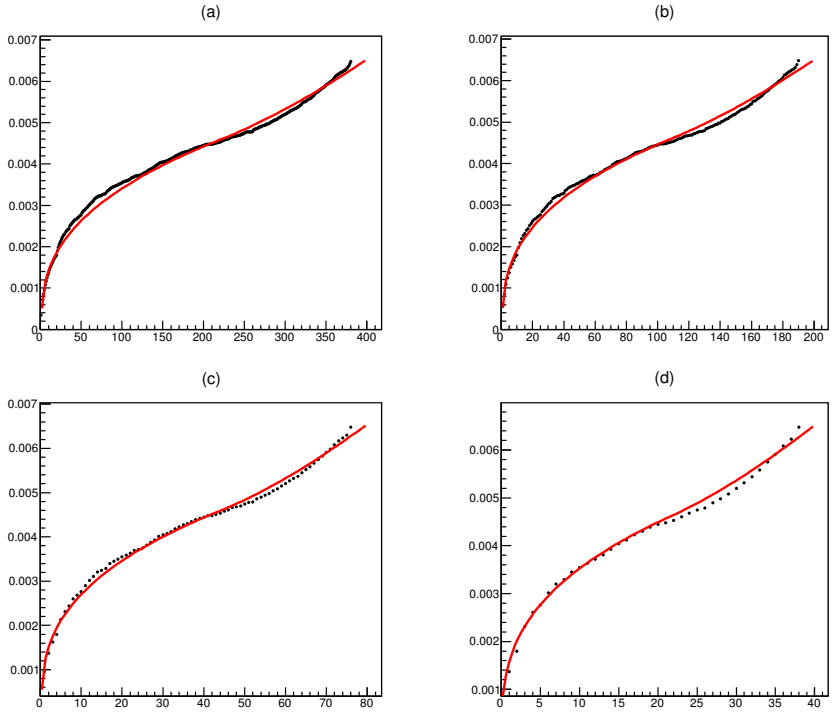

Figure 3.3: Data fit of the first empirical moments of the absolute value of aggregated returns, $\langle |R_1 + \ldots + R_t| \rangle$, for the whole day, $t \in [1, t_f]$, at $\tau = 1, 2, 5, 10$ minutes are respectively shown in panels (a),(b),(c),(d).

eq. (2.35),(2.38) we know that the collapse of intraday returns is lead by the model's parameters: so, although the structure of the scaling function does not depends on τ, in general we could expect some variations of the shape and the scale of the four different $g(r)$'s, since their parameters change at the various frequencies. Thus, the fact that night returns, whose values and temporal length obviously do not depend on the frequency chosen for the intraday collapse, have a constant scaling parameter for all frequencies is another confirmation of the goodness of the calibration results. At this point, we can confront night and daily dynamics on the basis of the acquired results: if we take the mean value upon frequency of $\lambda(t_f, t_m)$, $\hat{\lambda}(t_f, t_m) = 1.11 \pm 0.04 \cdot 10^{-2}$, we can calculate the ratio $\frac{\lambda_n}{\hat{\lambda}(t_f,t_m)} = 0.73 \pm 0.03$, which tells us that during the whole night, although stock market is closed, the S&P 500 index evolves with a rate which is approximatively 3/4 of the daily one. The whole procedure can be iterated for the full day returns, as shown in figure 3.5. Once again, we have that interday returns well collapse on intraday ones with a scaling parameter equal for all frequencies, fixed by hand at the value $\lambda_d = 1.40 \pm 0.02 \cdot 10^{-2}$. Moreover, since a full day is clearly the sum of a diurnal trading session and of the remaining time when the market is close (i.e. the night), we are interested in a relation which links the scaling parameter of its return, λ_d, with the ones of its summands, λ_n and $\lambda(t_f, t_m)$. To this purpose, it can be noticed that our empirical scaling parameter, from now on denoted as λ_d^e, is very near to the one obtained through

$$\lambda_d = \sqrt{\lambda^2(t_f, t_m) + \lambda_n^2}\,, \tag{3.2}$$

i.e $\lambda_d = 1.37 \pm 0.03 \cdot 10^{-2}$. This is a significant result, because, besides extending the procedure of aggregation of returns beyond intraday limits, eq. (3.2) generalizes the structure of eq. (2.36): the scaling parameter of an aggregated return is obtained through the square root of the sum of the squared λ_i's of the summands. Following this idea, we built larger, interday returns made up of multiple days in order to check if they still collapse to the same PDF, with the scaling parameter given by $\lambda_{kd} = \sqrt{k}\lambda_d$, according to the idea suggested by eq. (3.2), where $k \in \mathbb{N}$ is the number of aggregated days. The results are shown in figure 3.6, where we plot the collapse of aggregated returns of 5, 10, 20 and 50 days on the intraday ones, with $\tau = 10$ minutes. It can be noticed that, when k grows, the collapse ceases to follow the model's $g'(r)$ and its peak shows an evident skewness. A possible way to enlarge the domain of validity for the model is to look for a more general class of PDF's which reduces to the one of eq. (2.37) for aggregated returns $r_\tau(\tau)$ with τ up to few days.

Summarizing, eq. (3.2) is a result which makes a bridge between the existing intraday

[24], and interday [23], formalisms based on scaling properties. Indeed, while the first one exploits the ensemble approach that we used in this thesis, the other one works with time series made up of interday returns, $r^l = \ln S^l(0) - S^{l-1}(0)$. Then, although the two models have a similar structure and mathematical tools, until this result it was not possible to treat interday data with the intraday formalism, and viceversa. Eq. (3.2) is the first step in this direction, and its usefulness lies in the fact that it creates new possibilities of data analysis and further enhances the application possibilities. As an example, a possible application of this result, which we will exploit in the next section, is to verify the effect of the linear correlations between intraday and interday returns on the occurrence of volatility clustering or trends by conditioning the intraday PDF with interday returns.

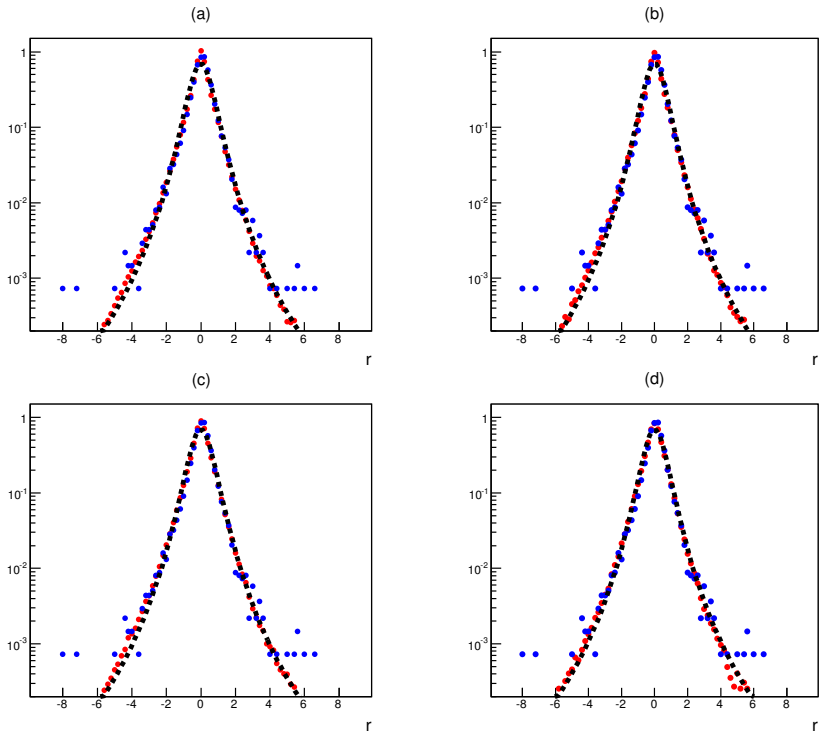

Figure 3.4: Data collapse of night returns (blue spots) on intraday ones (red spots) at $\tau = 1, 2, 5, 10$ minutes are respectively shown in panels (a),(b),(c),(d). The black dashed line is a fit to the points with the $g'(r)$ of eq. (2.37).

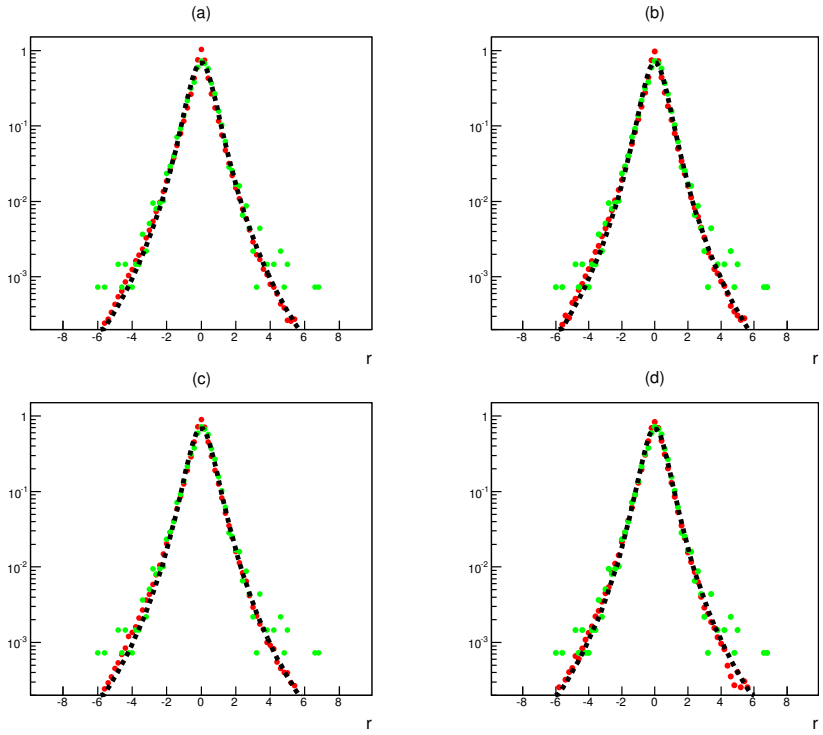

Figure 3.5: Data collapse of full day returns (green spots) on intraday ones (red spots) at the various frequencies. The black dashed line is a fit to the points with the $g'(r)$ of eq. (2.37).

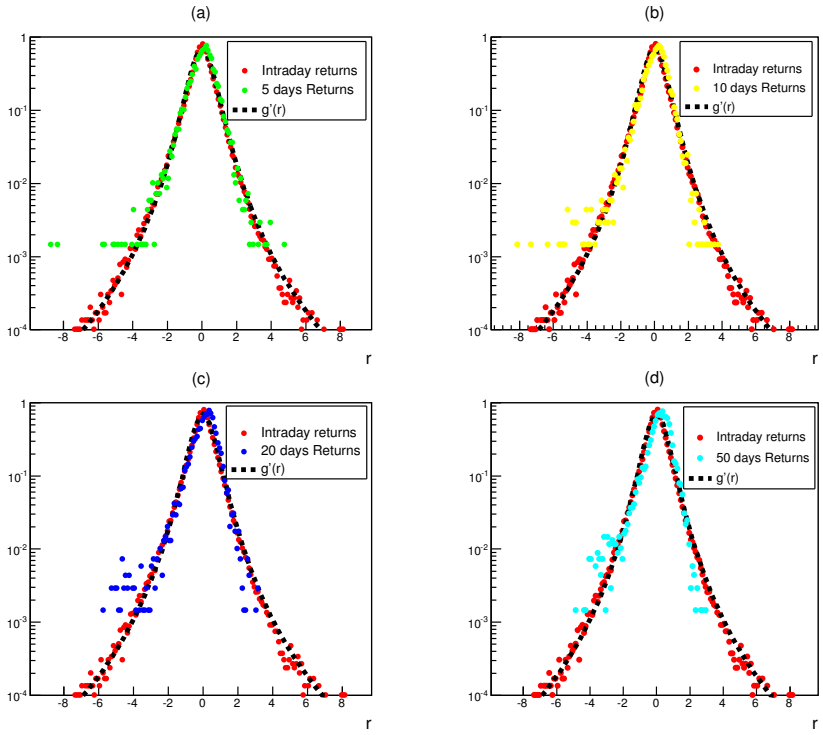

Figure 3.6: Data collapse of returns aggregated at 5,10,20,50 days on the intraday ones at $\tau = 10$ minutes (red spots) are respectively shown in panels (a),(b),(c),(d).

3.3. Trading strategy

Once the calibration procedure was completed, we applied on the four different ensembles the trading strategy defined in section 2.7, both in the unconditioned and in the conditioned case. To directly confront the results at different frequencies we chose the numbers of conditiong returns, t_p^τ's, in a such way that they all refer to the same time of the day, by taking $t_p^\tau = \frac{t_p^1[\text{minutes}]}{\tau[\text{minutes}]}$ for $\tau > 1$ minute. We chose as conditioning time 10:10, 10:40 and 11:10 (New York time), which correspond to $t_p^1 = 30, 60$ and 90. We chose to not consider higher values of t_p^1's for two reasons: their introduction leads to lower results and, since a single trading session lasts six hours and twenty minutes, we didn't want to restrict too much the time window for trading activity. Since we are interested in the determination of the most convenient framework for our trading strategy, we chose six different quantile levels, Q=2.5%, 5%, 10%, 15%, 20% and 25%, in order to identify as accurately as possible the behaviour of trading profit as a function both of τ and Q. In table 3.3 we show the average profit for trade at every τ, Q and t_p^τ combination. In paranthesis the total number of trades is reported.

It can be noticed that the number of trades is maximum at the highest frequency and quantile level, and decreases when τ increases and/or Q gets lower. We could expect this result: when Q is high, since the quantile thresholds are nearer to the real price values, they are more likely to be trespassed and thus to generate a buy/sell signal. On the other hand, when τ decreases the number of recorded prices in a day increases, so that we have more possibilities to get all the actual trading signals. It is maybe more useful to present the same results in a graphical way: in figure 3.7 we show the density plot for every couple (τ, Q) at different t_p's. From this graphical approach it is easier to analyze the behaviour of profit, following the color pattern. Thus figure 3.7 tells us that, once t_p^τ is fixed, profit tends to grow with Q for higher values of τ, while it shows the opposite trend when τ decreases. We can try to explain this behaviour in the following way: we already remarked how, for lower values of τ, we have a greater number of recorded prices inside a single day, and we can follow their dynamics at a finer scale. We are then more likely to identify, as soon as they appear, all the signals for new potential trends. Moreover, the higher values of linear autocorrelation shown in figure 3.1 tells us that this potential signals are more likely to actually become real trends, at least on short time windows. Thus, when the quantile level Q decreases, we get less trading signals (as the lower number of trades confirms), which are all likely to indentify actual trends: the average profit grows. When instead we consider higher values of Q, the probability of taking as trading signals simple fluctuations of the recorded prices increases, and this

	25%	20%	15%	10%	5%	2.5%
1 minute						
0	4.13(39216)	4.24(34314)	4.22(28934)	4.37(22033)	4.28(13843)	4.47(7906)
30	3.94(42124)	4.02(37944)	4.13(32929)	4.39(26501)	5.04(18322)	5.45(13368)
60	3.76(41271)	3.86(36503)	3.97(31328)	4.35(24642)	4.85(16936)	5.43(12135)
90	3.59(40577)	3.61(35710)	3.81(29676)	4.36(22685)	5.02(15290)	5.82(10786)
2 minutes						
0	3.99(30725)	4.07(26809)	3.94(22618)	4.01(17278)	3.70(10801)	3.78(6071)
15	4.09(31666)	4.15(28045)	4.24(23781)	4.30(18848)	4.96(12422)	5.13(8526)
30	3.99(30624)	4.02(26790)	4.10(22669)	4.59(17100)	4.99(11298)	5.56(7699)
45	3.81(29857)	3.86(25807)	4.02(21253)	4.58(15841)	5.26(10118)	5.86(7001)
5 minutes						
0	3.64(21867)	3.74(19137)	3.60(15982)	3.49(12171)	2.95(7484)	2.64(4255)
6	4.26(21701)	4.11(19096)	4.17(16033)	3.75(12566)	3.59(7977)	3.03(4983)
9	4.20(20788)	4.20(17949)	4.24(14773)	4.18(11235)	4.01(7111)	4.10(4609)
12	3.98(20170)	4.10(17038)	4.12(13861)	4.21(10274)	4.50(6394)	4.25(4214)
10 minutes						
0	3.17(16314)	3.22(14177)	3.04(11539)	2.92(8684)	2.38(5137)	1.76(2745)
3	4.09(16319)	4.07(14132)	3.71(11736)	3.20(9114)	2.96(5535)	1.55(3345)
6	4.15(15436)	4.27(13137)	4.09(10744)	3.59(8140)	3.45(4790)	2.31(2987)
9	4.12(14734)	4.13(12411)	4.09(9886)	3.95(7167)	3.52(4342)	2.78(2724)

Table 3.3: Average profit per trade in basis point in the interval (1985-2013). In parenthesis the total number of trades is reported.

process makes the trades become more, but less convenient: the average profit decreases. On the other side, figure 3.7 tells us also that when the time lag increases, the average profit shows the opposite behaviour: it grows together with the quantile level. We can try to explain this feature in the following way: at lower frequencies we have a loss of profit due to the delay with which new potential trends are detected. In fact, the later the trading position is opened, the lower will be its profit. In this context, the difference of profit between signals detected at lower or higher quantile levels can become negligible, and a greater number of trades is more likely to increase the average profit.

It is not possible to detect a clear, well determined trend for the average profit as a function of the number of conditioning returns, t_p. However, the most convenient framework for our trading strategy is determined by a low quantile level (Q=2.5%), an high frequency ($\tau = 1$ minute) and the highest number of conditioning returns, ($t_p^1 = 90$).

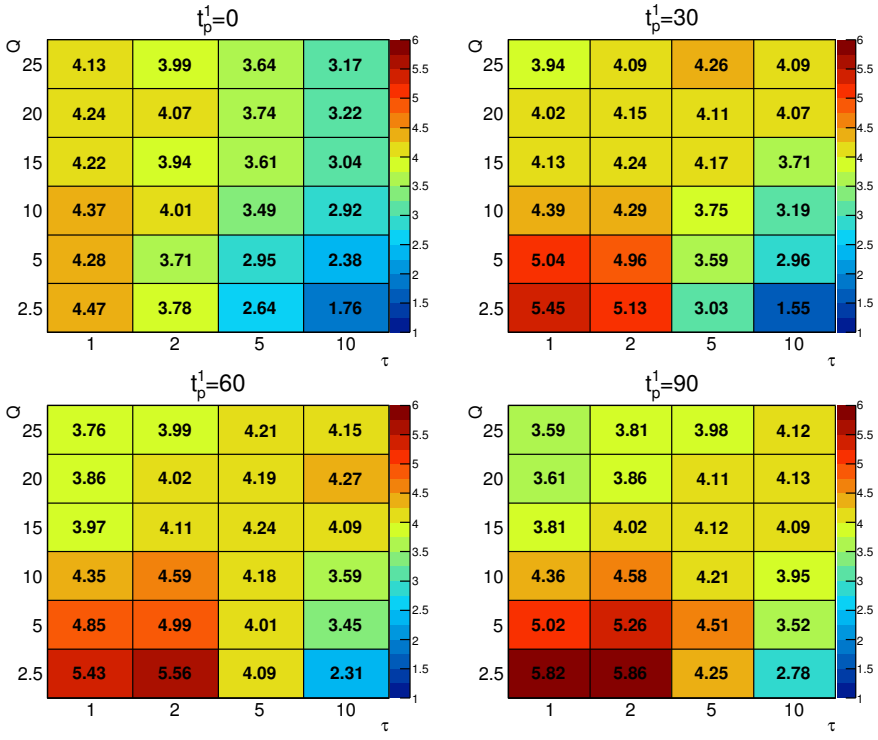

Figure 3.7: Density plot of the average profit at different t_p^1. On the x axis we report the time lag τ, on the y one the quantile level Q.

3.3.1. Night conditioning

In section 2.7.2 we introduced the possibility of conditioning our PDF to evaluate quantile barriers which take into account past returns. This procedure, described by eq. (2.47),(2.48), implies only that the returns which are used for the conditioning collapse on the $g'(r)$ with a suitable scaling parameter, without imposing any other condition. Thus, once we verified that night returns well collapse on our model's PDF, we can use them to evaluate conditioned quantile barriers, just by adding the corrispondent night return with the corresponding λ_n, and to apply our trading strategy. Table 3.4 reports our results: it can be noticed that the average profit does not have any significant modification for the higher numbers of conditioning returns, while it shows an evident decrease for $t_p^1 = 0,30$. To explain this behaviour we looked at the linear correlations between the aggregated returns and the night ones. In fact, as we explained in the past sections, linear correlations play a crucial role in determining the average profit, since they influence the probability of identifying a real trend from the trading signals of our strategy.

As it is shown in figure 3.8, we can see that the correlations have negative values in the first part of the day, and keep growing until the end of the trading session. The plot is made at $\tau = 1$ minute, but it is clearly the same at all frequencies, beside of the number of points. Thus the effect of night conditioning on our average profit can be fully explained: the presence of negative linear correlations between night returns and intraday, aggregated ones tells us that during the first part of the day returns will contrast the night trend. In this scenario, our trading strategy becomes ineffective. A possible way to overcome this problem is the development of strategies capable to exploit these negative linear correlations.

Night conditioning						
	25%	20%	15%	10%	5%	2.5%
1 minute						
0	3.83(40449)	3.98(35272)	4.10(29518)	4.18(22771)	3.88(14366)	3.79(8656)
30	3.92(42010)	3.96(38083)	4.12(32879)	4.38(26459)	5.01(18172)	5.36(13278)
60	3.77(40975)	3.85(36515)	3.96(31329)	4.32(24583)	4.84(16717)	5.40(11990)
90	3.60(40157)	3.61(35670)	3.82(29595)	4.37(22632)	5.03(15072)	5.83(10632)
2 minutes						
0	3.66(31743)	3.76(27743)	3.83(23026)	3.93(17700)	3.23(11274)	3.12(6603)
15	4.06(31783)	4.12(28199)	4.19(24038)	4.29(18827)	4.82(12478)	4.98(8507)
30	3.99(30640)	4.02(26785)	4.09(22655)	4.55(17049)	4.82(11431)	5.51(7697)
45	3.81(29861)	3.86(25838)	4.02(21249)	4.61(15735)	5.17(10120)	5.88(6917)
5 minutes						
0	3.35(22539)	3.29(19621)	3.14(16499)	3.11(12588)	2.36(7885)	1.69(4639)
6	4.10(21941)	4.04(19247)	4.10(16247)	3.72(12718)	3.50(8099)	2.94(5135)
9	4.15(20882)	4.19(18000)	4.15(14927)	4.05(11372)	3.82(7240)	4.22(4606)
12	3.97(20210)	4.05(17185)	4.11(13873)	4.23(10269)	4.40(6465)	4.07(4251)
10 minutes						
0	2.84(16808)	2.60(14679)	2.60(12015)	2.39(9057)	1.44(5469)	0.76(3112)
3	3.83(16554)	3.90(14381)	3.55(11922)	3.00(9251)	2.54(5721)	1.34(3501)
6	4.17(15513)	4.25(13279)	3.95(10873)	3.54(8279)	3.15(4919)	2.29(3095)
9	4.13(14827)	4.15(12481)	4.09(9970)	3.97(7233)	3.28(4451)	2.71(2808)

Table 3.4: Average profit per trade in basis point in the interval (1985-2013). All the quantile barriers have been evaluated conditioning the PDF (also) on night returns.

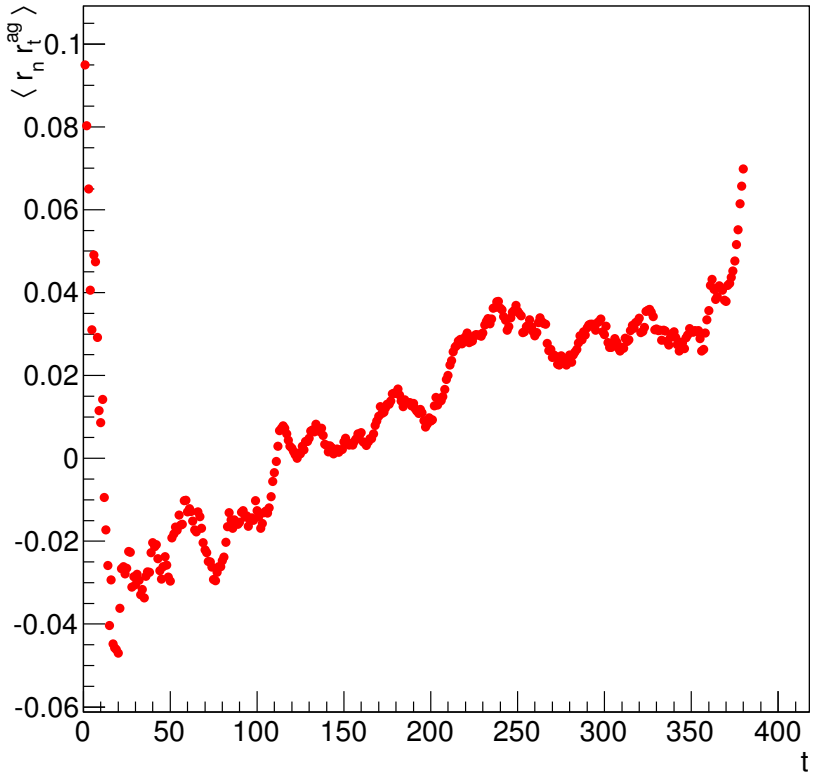

Figure 3.8: Linear autocorrelation between night returns and intraday, aggregated ones, with $\tau = 1$ minute.

4. Working on different assets

In the previous chapters we worked exclusively on the S&P 500 index, since the model itself has been shaped on the empirical properties of its data set. However, our mathematical framework can be formally applied to any financial data set with a similar structure. We thus applied the model to various different assets, to analyze its fitting properties and the possibility of implementing a profitable trading strategy. The assets we used for our analysis are fifteen: BA, BAC, C, FDX, HON, HPQ, IBM, JPM, MDLZ, PEP, PG, T, TXN, TWX and WFC. All the data sets taken from the history of these assets cover the same time interval, from January 1_{st} 2003 to June 28_{th} 2013, for a total of M=2640 days. As for the S&P 500 index, each days goes from 09:40 to 16:00 (New York time). We proceeded in the usual way: we first looked at the linear correlations and the volatility pattern to identify a suitable t_m, then we went on with the calibration procedure and the fitting of the first empirical moment, after verifying the presence of scaling features. We worked with a time lag of $\tau = 10$ minutes. We then looked at the linear correlations of the same assets at $\tau = 1$ minute and eventually applied the trading strategy of section 2.7 to the assets with the more significant correlation patterns.

4.1. Linear correlations, volatility and scaling

If we look at the linear correlation pattern of each asset, evaluated according to eq. (1.7), we find that they are below a low barrier of about 0.1, such as in the S&P 500 case at the same frequency. In figure 4.1(a) we plot, for $\tau = 10$ minutes, the linear correlations of the IBM asset, as an example. The next step was to look at the pattern of the second empirical moment $m_2^e(t, 10)$, to verify the presence of a periodic behaviour to justify the ensemble approach. In figure 4.2 we confront the plot of $m_2^e(t, 10)$ in a day (a) and in a week (b) window. As we got for the S&P 500 index an evidence of periodicity is found. We then looked to the plot of the volatility, evaluated according to eq. (3.1), to identify the proper t_m for the calibration procedure. As shown in figure 4.1(b), the

choice of $t_m^{10} = 22$ made for the S&P 500 index is still valid for the IBM (and all the other) asset. This means that the U-shaped pattern for volatility, linked to the different volumes of trading during each day, is a feature shared by many of the single assets in the stock market. The last step to verify the full compatibility of the assets' data sets with our model was to look for its scaling properties. We then looked for the proper Hurst exponent through the linear fit of the couple (q, qD) obtained by linearizing eq. (2.25) at various q. In figure 4.3 we show the result for the FDX asset, with $D_{FDX} = 0.29$. Thus we built a data collapse as suggested by eq. (1), putting the just found value of D, as we did in chapter 1 for the S&P 500 index. Figure 4.4 shows the data collapse for the MDLZ asset.

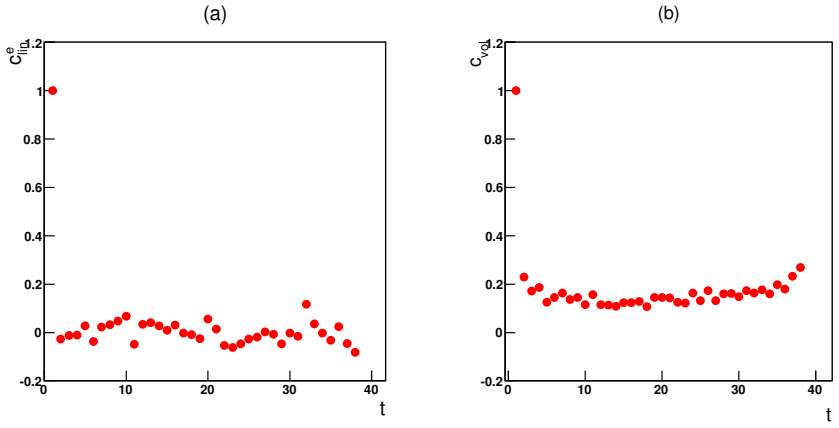

Figure 4.1: Empirical linear autocorrelation $c_{lin}(1, t)$ for the IBM asset, $\tau = 10$ minutes.

4.2. Calibration and fitting properties

Once the requested empirical features were verified in our set of assets, we proceeded with the calibration protocol, as described in section 2.5. In table 4.1 we present the fifteen sets of parameters that we found performing this operation. If we look at the mean value and at the standard deviation σ for each parameter, it can be noticed that β_m, β_a and D_a show the largest relative variations. We already remarked that the β's are strictly connected to the value of the first empirical moment of the absolute value of the first return of the day, $\langle |R_1| \rangle$, while D_a is evaluated through eq. (2.39), whose right member depends on β_m, β_a through the scaling parameter $\lambda(\tau, t_m)$. Finally we verified the goodness of the calibration procedure by performing the fit of the absolute values of the first empirical moment of the aggregated returns along the day, as shown in figure

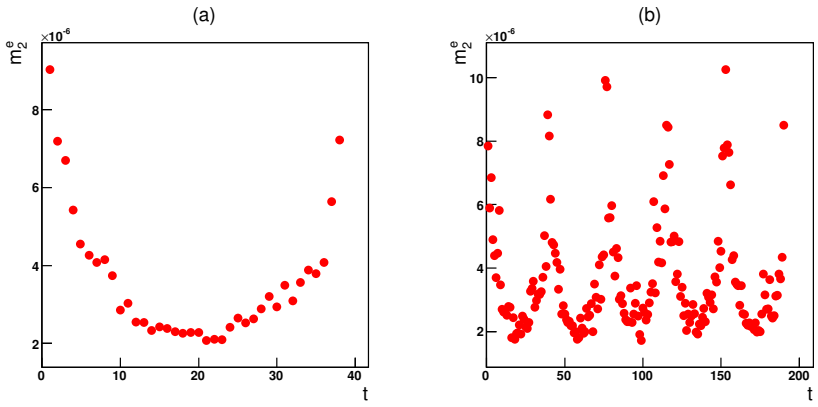

Figure 4.2: (a): The ensemble second moment $m_2^e(t, 10)$ of the daily IBM asset as a function of the time of day, from 09:40 to 16:00. (b): The weekly behaviour of $m_2^e(t, 10)$ for the same data.

4.5 for the HON asset.

4.3. Structure of correlations and trading strategy

Since the various assets' data sets are well parameterized by our model, we tried to apply to them the trading strategy conceived for the S&P 500 index. However, before performing this operation, guided by what we learned in the previous chapter, we looked at the structure of the linear correlations between successive returns $c_{lin,\tau}(t, t+1)$. In figure 4.6 we confront the values of $c_{lin,10}(t, t+1)$ for the PEP asset, (a), and for the S&P 500 index, (b). It can be noticed that the asset's successive returns are more likely to be anticorrelated, and that the positive entries of $c_{lin,10}(t, t+1)$ show smaller values than the S&P 500's ones. Since we already underlined the importance of the linear correlations $c_{lin,\tau}(t, t+1)$ for the profit values of the trading strategy, we can expect lower results. We looked then, among all the different assets, at the $c_{lin,\tau}(t, t+1)$ at various frequencies to find the framework in which this difference is more evident. In figure 4.7 we show what we found: as it can be seen, the linear correlations between successive returns of the asset BAC, evaluated at $\tau = 1$ minute, (a), are almost symmetrical to the S&P 500's ones, (b). We then completed the calibration procedure for the returns of the BAC asset at $\tau = 1$ minute and we applied the trading strategy of section 2.7, expecting to find negative values. Table 4.2 shows that the results we found confirmed our expectations. It is remarkable, however, that if we completely invert our strategy actions, the profit values become positive. This procedure of inversion means that we

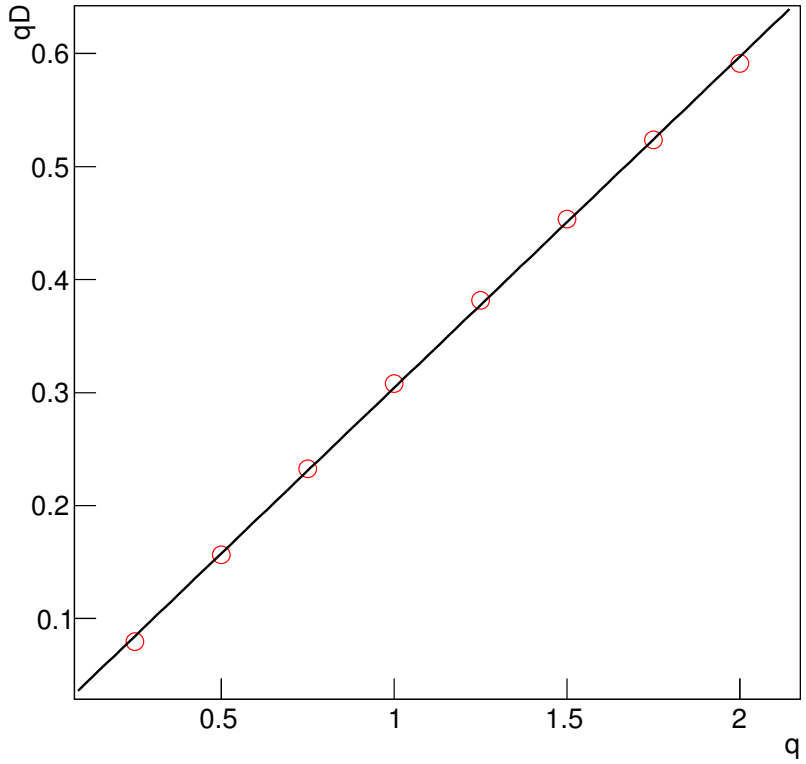

Figure 4.3: Scaling behaviour of the non-linear moments of returns of the FDX asset. The red circles are the couples (q, qD) evaluated from the logarithmic fits of $\langle |R_1 + \ldots + R_t| \rangle \sim t^{qD}$, with $q \in [0, 2]$, $t \in [1, t_m]$. The black line is their linear fit, $D = 0.29$.

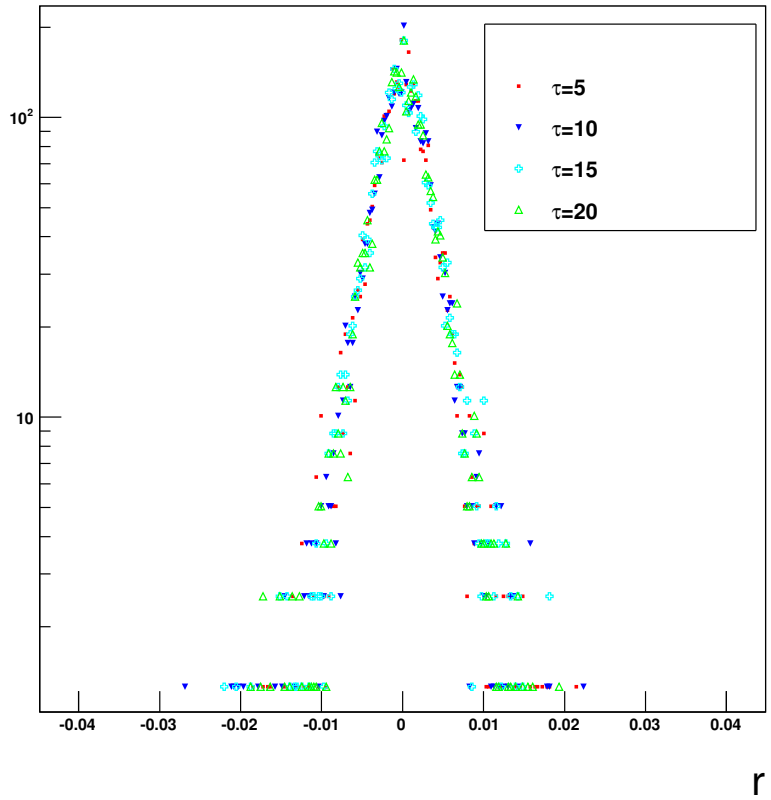

Figure 4.4: Scaling collapse of the returns aggregated at $\tau = 5, 10, 15, 20$ minutes of the MDLZ asset according to eq. (1).

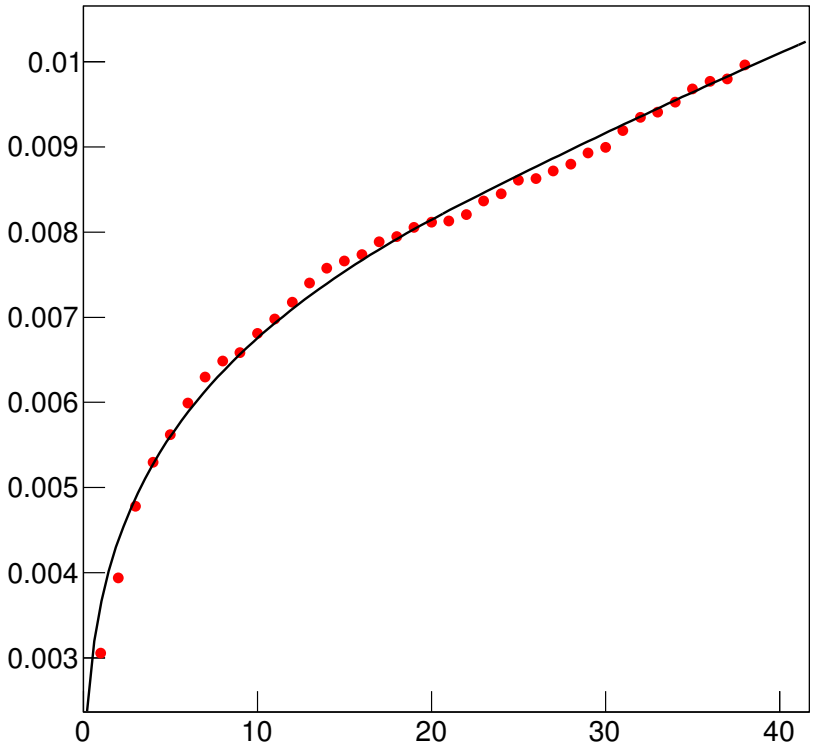

Figure 4.5: Data fit of the first empirical moments of the absolute value of aggregated returns of the HON asset, $\langle |R_1 + \ldots + R_t| \rangle$, for the whole day, $t \in [1, t_f]$.

	α	D_m	β_m	D_a	β_a
BA	3.70	0.30	$5.97 \cdot 10^{-3}$	0.51	$2.40 \cdot 10^{-3}$
BAC	2.27	0.32	$7.04 \cdot 10^{-3}$	1.15	$0.29 \cdot 10^{-3}$
C	2.75	0.35	$6.90 \cdot 10^{-3}$	0.91	$0.79 \cdot 10^{-3}$
FDX	3.48	0.29	$6.20 \cdot 10^{-3}$	0.55	$2.03 \cdot 10^{-3}$
HON	3.65	0.27	$6.43 \cdot 10^{-3}$	0.54	$2.12 \cdot 10^{-3}$
HPQ	3.96	0.30	$6.00 \cdot 10^{-3}$	1.30	$0.13 \cdot 10^{-3}$
IBM	3.36	0.31	$4.00 \cdot 10^{-3}$	0.83	$0.50 \cdot 10^{-3}$
JPM	2.98	0.29	$6.90 \cdot 10^{-3}$	1.13	$0.27 \cdot 10^{-3}$
MDLZ	3.90	0.25	$4.96 \cdot 10^{-3}$	0.82	$0.47 \cdot 10^{-3}$
PEP	3.81	0.29	$3.89 \cdot 10^{-3}$	0.50	$1.97 \cdot 10^{-3}$
PG	3.73	0.29	$3.60 \cdot 10^{-3}$	0.78	$0.48 \cdot 10^{-3}$
T	3.23	0.30	$4.90 \cdot 10^{-3}$	0.50	$2.05 \cdot 10^{-3}$
TWX	3.37	0.31	$5.67 \cdot 10^{-3}$	1.04	$0.33 \cdot 10^{-3}$
TXN	3.85	0.26	$7.90 \cdot 10^{-3}$	0.65	$1.51 \cdot 10^{-3}$
WFC	2.95	0.27	$7.00 \cdot 10^{-3}$	1.34	$0.11 \cdot 10^{-3}$
Mean	3.40	0.29	$5.82 \cdot 10^{-3}$	0.84	$1.03 \cdot 10^{-3}$
σ	0.47	0.02	$1.26 \cdot 10^{-3}$	0.29	$0.84 \cdot 10^{-3}$

Table 4.1: Sets of calibration parameters of the various assets, $\tau = 10$ minutes.

have to give to the trading signals the opposite meaning to the one we gave to them in the previous definition:

1. If there are no open positions:
   (a) Sell if $S_t^{(l)} > S_{max,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} < S_{max,t-1}^{(l)}(Q)$ (open a short position)
   (b) Buy if $S_t^{(l)} < S_{min,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} > S_{min,t-1}^{(l)}(Q)$ (open a long position)
2. If there are open positions:
   (a) Buy if $S_t^{(l)} < S_{max,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} > S_{max,t-1}^{(l)}(Q)$ (close a short position)
   (b) Sell if $S_t^{(l)} > S_{min,t}^{(l)}(Q)$ & $S_{t-1}^{(l)} < S_{min,t-1}^{(l)}(Q)$ (close a long position)
   (c) Close any long/short position still open at the end of the day.

This suggest different ways to make trades on the various assets or indexes, on the basis of the linear correlations $c_{lin,\tau}(t, t+1)$ of the data set of returns used for the calibration. It can be noticed that for the BAC asset with $\tau = 1$ minute a well defined behaviour for the profit values as a function of the quantile level Q does not emerge: looking again at figure 4.7 this can be explained by observing that the $c_{lin,1}(t, t+1)$ shows some little fluctuations among positive and negative which were not present in the S&P 500's framework.

BAC asset						
	25%	20%	15%	10%	5%	2.5%
1 minute						
0	-1.43(16672)	-1.47(13312)	-1.56(10485)	-1.33(7290)	-0.49(4224)	-0.95(2372)
30	-2.11(21285)	-2.06(17266)	-2.00(13307)	-2.50(9316)	-3.62(5221)	-2.99(2790)
60	-2.10(20156)	-1.96(16299)	-2.02(12485)	-1.92(8231)	-2.85(4487)	-3.36(2603)
90	-2.39(19530)	-2.83(15895)	-3.01(11715)	-3.02(7647)	-3.41(3795)	-4.17(2118)

Table 4.2

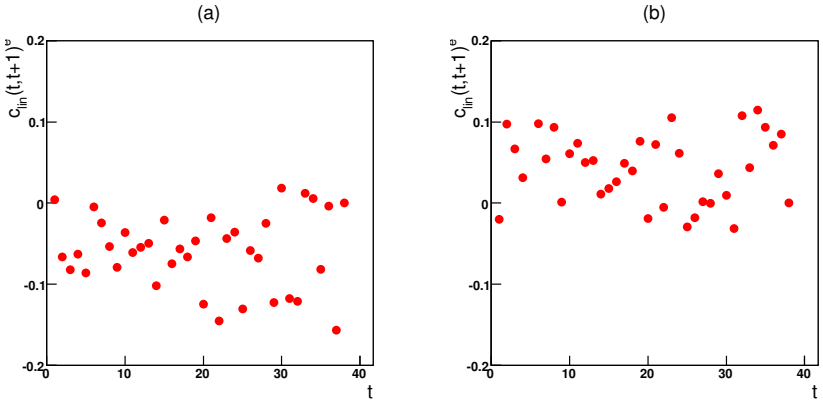

Figure 4.6: Empirical linear correlations between successive returns of the PEP asset, (a), and the S&P 500 index (b), $\tau = 10$ minutes.

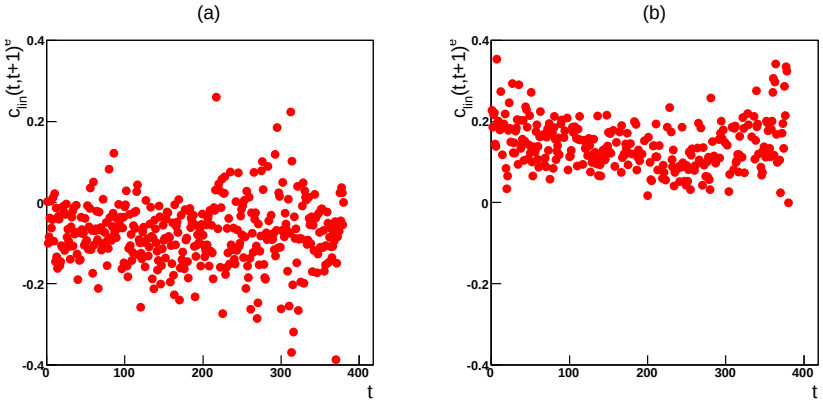

Figure 4.7: Empirical linear correlations between successive returns of the BAC asset, (a), and the S&P 500 index (b), $\tau = 1$ minute.

5. Conclusions

In this thesis we provided a complete modelization of intraday financial returns, built by focusing on the empirical features of the S&P 500 index but capable to describe also different assets. We firstly looked at the pattern of high order moments and correlations within different time windows, and we found the evidence of a cyclical behaviour of these quantities, whose period has the duration of a day of trading session. We further found that the second moment of the returns and the volatility autocorrelations show an U-shaped pattern, which is due to the changing volume of trading activity during the day, and makes a clear separation between the morning and the afternoon time windows. We thus decided to adopt an ensemble approach in order to give a better description of the intraday dynamics, which shows these strong non stationary features during daily evolution [10], [11]. Once we fixed the correct approach, the major guiding principle that we followed in the model construction was scaling simmetry. Indeed, our data showed evident self-similarity features, which confirmed the idea that the statistical averages of different aggregated returns can be compared and analyzed together by exploiting proper rescaling parameters, in our case a power of the interval of aggregation to the Hurst exponent D. Starting from the empirical evidence of this scaling simmetry and from the presence of strong, non linear correlations we built the model, defining its PDF and the general structure of the scaling function $g(r)$ as a convex combination of Gaussian distributions. We fixed the precise form of the $g(r)$ by imposing the fitting of some empirical features of our data distribution, such as the absence of skewness and the occurrence of heavy tails [24]. We thus obtained a Student's t-distribution, fully determined by fixing its form and scale parameters, respectively α and β, which well describes the statistics of morning returns. In order to obtain a correct modelization of the market dynamics of the whole trading session, we performed an extension of the model to the afternoon time window by introducing a time dependence in some of its parameters. In the end, we found a model capable to well fit the moments at various orders of the absolute values of returns after calibrating its five parameters:

$D_m, D_a, \alpha, \beta_m, \beta_a$, where the m, a pedixes indicate whether they occurr in the morning or in the afternoon time window. However, the model fails when it has to reproduce linear correlations: although its martingale structure fixes them equal to zero, $\langle r_i r_j \rangle = 0 \; \forall \; i, j$, empirical evaluation contradicts this prescription to a moderate extent. Thus, in order to measure the effects of this discrepancy from the model, we implemented a trading strategy. In fact, by evaluating the (un)conditioned expected returns distributions at different quantile levels and checking how many times recorded prices break through their values, we monitored the occurrence of long lasting trends, which would be absent if linear correlations were equal to zero. More precisely, we expected to find positive values of profit when successive returns are positively correlated, and viceversa. Indeed, the occurrence of positive linear correlations is a signal that trends are more likely to last, and the strategy is conceived in order to obtain larger values of profit when the trends maintains themselves stable during the whole trading session. We performed this operation, for various assets, at different values of frequency and quantile level, and we found a clear confirmation of our expectations. More precisely, we found that the strong correlations between successive returns found for the S&P 500 index at various values of frequency, conditioning returns and quantile level, lead to positive values of profit included between 1.55 and 5.86 basis points of the initial cash investment, while, for example, a similar trading strategy performed on the same data set exploiting a strategy based on quantile barriers evaluated through a GARCH process [22] lead to results of about an order of magnitude smaller [24]. As far as the other assets are concerned, we found a general persistence of negative linear correlations, which in the case of the BAC asset produced negative results ranging between -4.17 and -0.49 basis points.

Furthermore, we extended for the first time the domain of validity of the model to interday returns. We isolated from our data two time series, made up of returns aggregated respectively over a night and a full day. We then successfully collapsed them, through suitable scaling parameters, together with intraday returns, demonstrating that they follow the same scaling function $g(r)$. We also found that the scaling parameters of these interday returns can be evaluated from the ones of their summands, through an additive formula which is an extension of the one already implemented in the intraday context. This result is the first evidence of the possibility of unifying the treatment of returns aggregated over interday and intraday intervals. In fact, previous works have developed different models to treat data in the two different contexts [24], [23], while others have studied how to model intraday data in order to obtain a better measure of volatility on a daily scale [26], but we have introduced for the first time the possibility

of using the same PDF, with the introduction of just an additional scaling parameter, to describe both interday and intraday aggregated returns. This allows us to add interday dynamics to our treatment, conditioning the PDF of the model also on these returns. As an application, we then looked at the effects of the conditioning on night returns in the daily trading strategy. We found poorer profits, and we explained the results looking at the structure of the linear correlations between night returns and intraday, aggregated ones, which show negative values in the first part of the day. An open question left by this thesis is to look at a more general class of scaling functions capable to describe returns aggregated at larger interday intervals, since we observed that, when looking at aggregation times of the order of about 20 days, their distributions show an evident skewness which cannot be reproduced by our $g(r)$.

Bibliography

[1] B. Gnedenko and A. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*. Reading: Addison Wesley, 1954.

[2] L. Kadanoff, *Statistical Physics, Statics, Dynamics and Renormalization*. Singapore: World Scientific, 2005.

[3] L. Bachelier, "Théorie de la spéculation," *Annales Scientifiques de l'Ecole Normale Supérieure*, 1900.

[4] J. Bouchaud and M. Potters, *Theory of Financial Risks*. Cambridge: Cambridge University Press, 200.

[5] B. Mandelbrot, "The variation of certain speculative prices," *Journal of Business*, pp. 394–419, 1963.

[6] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issue," *Quantitative Finance*, pp. 223–236, 2001.

[7] H. Hudson, "Solar flares, microflares, nanoflares, and coronal heating," *Solar Physics*, no. 133, 1990.

[8] S. Havlin, L. Amaral, Y. Ashkenazy, A. Goldberger, P. Ivanov, C. Peng, and H. Stanley, "Application of statistical physics to heartbeat diagnosis," *Physica A*, no. 274, 1999.

[9] J. B. Rundle, S. Gross, W. Klein, C. Ferguson, and D. L. Turcotte, "The statistical mechanics of earthquakes," *Tectonophysics*, no. 277, 1997.

[10] K. Bassler, J. McCauley, and G. Gunaratne, "Nonstationary increments, scaling distributions, and variable diffusion processes in financial markets," *PNAS*, pp. 17287–17290, 2007.

[11] F. Baldovin, D. Bovina, F. Camana, and A. Stella, "Modeling the non-markovian, non-stationary scaling dynamics of financial markets," *Econophysics of of order-driven markets*, pp. 239–252, 2011.

[12] J. Cardy, *Scaling and Renormalization in Statistical Physics.* Cambridge: Cambridge University Press, 1996.

[13] H. Hurst, "Long term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, no. 116, 1951.

[14] P. Lévy, *Calcul des probabilités.* Paris: Gauthier-Villars, 1925.

[15] F. Baldovin and A. Stella, "Scaling and efficiency determine the irreversible evolution of a market," *PNAS*, pp. 19741–19744, 2007.

[16] F. Baldovin and A. Stella, "Anomalous scaling due to correlations: Limit theorems and self-similar processes," *Journal of Statistical Mechanics*, 2010.

[17] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, pp. 637–654, 1973.

[18] F. Fama, "Efficient capital markets: a review of theory and empirical work," *PNAS*, pp. 383–417, 2007.

[19] G. L. Vasconcelos, "A guided walk down wall street: an introduction to econophysics," *Brazilian Journal of Physics*, 2004.

[20] B. Mandelbrot and J. van Nesse, "Fractional brownian motions, fractional noises and applications," *The SIAM Review*, pp. 422–437, 1968.

[21] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, pp. 987–1007, 1982.

[22] T. Bollerslev, "Generalized conditional heteroskedasticity," *Journal of Econometrics*, pp. 307–327, 1986.

[23] F. Baldovin, M. Caraglio, A. Stella, and M. Zamparo, "Scaling symmetry, renormalization, and time series modeling," *Physycal Review E*, no. 88, 2013.

[24] F. Baldovin, F. Camana, M. Caporin, M. Caraglio, and A. Stella, "Ensemble properties of high frequency data and intraday rules," *Quantitative Finance*, 2014.

[25] I. Schoenberg, "Metric spaces and completely monotone functions," *Annals of Mathematics*, no. 30, 1938.

[26] F. Corsi, G. Zumbach, U. Müller, and M. Dacorogna, "Consistent high-precision volatility from high-frequency data," *Economic Notes*, no. 30, pp. 183–204, 2001.