UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

HIGH-FREQUENCY DYNAMICS OF THE MICROSCOPICAL STRUCTURE IN FINANCIAL MARKETS

Autor:
JAVIER VICENTE

Director/es:
DOYNE FARMER / MIKEL TAPIA

DEPARTAMENTO DE MATEMÁTICAS

Leganés, mayo 2.011

TESIS DOCTORAL

HIGH-FREQUENCY DYNAMICS OF THE MICROSCOPICAL STRUCTURE IN FINANCIAL MARKETS

Autor: JAVIER VICENTE

Director/es: DOYNE FARMER / MIKEL TAPIA

Firma del Tribunal Calificador:

		Firma
Presidente:	ANGEL SANCHEZ SANCHEZ	
Vocal:	MIGUEL ANGEL MARTINEZ SEDANO	
Vocal:	JESÚS GUILLERMO LLORENTE ALVAREZ	
Vocal:	JOSEP PERELLO PALOU	
Secretario:	JESÚS DAVID MORENO MUÑOZ	

Calificación:

Leganés, de de

Abstract

In the first part of this thesis, I address the classical problem of asset price dynamics based on a new theoretical framework developed for nonequilibrium physical systems. This problem is mainly relevant for two reasons. First, because understanding the true distribution of returns is important for asset allocation, risk management, and option pricing. Second, because in spite of all the effort in determining the origin of non-Gaussian returns no conclusive result has been achieved yet. The most important result of this part is the demostration that the non-Gaussian shape and stable scaling of the returns distribution are due to slow, but significant, fluctuations in volatility. Futhermore, this result suggests that stock price fluctuations are universal, and that return distributions can be described by one functional form.

In the second part, I present an empirical study about the execution of large orders in two stock exchanges: the London Stock Exchange, and the Spanish Stock Exchange. This type of orders can cause a tremendous impact because they are larger than the available liquidity in the order book at a time. For this reason, they are split to minimize transaction costs. Market price impact is the basic factor of these costs, so an accurate description of its functional form is necessary to any optimal execution. The most important result in this part is the empirical determination of this functional form in two markets and the finding of a common behavior in both markets that can be summarized into a concave temporary impact, roughly described by a square root function of the hidden order size, and a price reversion after the completion of the hidden order making permanent impact equal to roughly half of the temporary impact.

Acknowledgements

A dissertation is the final expression of some individual work, many discussions and much moral support. I want to deeply thank to all people who made possible this project. To my advisors, Doyne Farmer and Mikel Tapia, thank you for your time and your wise advises. I hope you feel this work as a reward to your effort. To my collaborators, Miguel Fuentes and Austin Gerig, thank you because working with you has been a great experience for me. You showed me that scientific work can be funny and challenging at the same time. To my fellow econophysics researchers at the Santa Fe Institute, thank you for your stimulating discussions: Adlar Kim, Anuj K. Purwar, Arturo Rodríguez, Yonathan Schwarzkopf, Mike Twardos, and Neda Zamani. To the research group at the University Carlos III of Madrid, and very especially to Esteban Moro for your thoughtful suggestions.

To all people who gave me moral support for finishing this work. To all nice people that María José and I met in Santa Fe: José Manuel, Janelle, Jane and Sandford Brickner, Lillian, Jane, and James Baumbach, Ale, Kelle, Marisa, Sonia, Jo and Rich Franzen, Tim Taylor, Orlando, Patrisia. To my dear friends: Manolo, Ana, Javier and Clara.

Finally, to my family. To my mother who was always there whenever I needed some help. To Pedro, my brother, thank you for showing me the way since I was a child. To Emilia, thank you for your generosity and comprehension. To my wife María José and my daughter Carlota, you are the most important people in my life. You make me feel a very lucky person.

Contents

0 Introduction **1**
0.1 Science and Finance . . . 1
0.2 Complexity in Financial Markets . . . 2
0.3 Stock Price Dynamics . . . 3
0.4 Non-Gaussian Returns Distribution . . . 4
0.5 Statistical Mechanics and Finance . . . 5
0.6 Statistical Properties of Empirical Returns . . . 6
0.6.1 Heavy Tails . . . 7
0.6.2 Aggregational Gaussianity . . . 9
0.6.3 Absence of Linear Autocorrelation . . . 11
0.6.4 Volatility Clustering . . . 12
0.7 The Problem of Large Orders Execution . . . 14

1 1 Gaussian Models **21**
1.1 Introduction . . . 21
1.2 A First Gaussian Model . . . 22
1.2.1 Theoretical Assumptions on Market and Prices . . . 23
1.2.2 Mathematical Description of Price Dynamics . . . 24
1.2.3 Empirical Returns Distributions . . . 27
1.2.4 Stochastic Processes in Finance . . . 28
1.3 Standard Gaussian Model . . . 32
1.3.1 Log-Normal Distributions of Stock Prices. . . . 34
1.4 Conclusions . . . 34

2 Non-Gaussian Models **39**
2.1 Introduction . . . 39
2.2 Non-Gaussian Models . . . 40
2.2.1 Stable Distributions Models . . . 41
2.2.2 Mixture Distribution Hypothesis Models . . . 44
2.3 Stochastic Volatility Models . . . 51
2.3.1 The Standard Stochastic Volatility Model . . . 54
2.4 Conclusions . . . 56

3 A Superstatistical Stochastic Volatility Model **59**
3.1 Introduction . . . 59
3.2 Superstatistics . . . 60
3.2.1 Superstatistics in Finance . . . 62
3.2.2 Superstatistical Classes . . . 64
3.3 High-Frequency Financial Time Series . . . 65

3.4 A Superstatistical Stochastic Volatility Model 66
3.4.1 Theoretical Description of The Model 67
3.4.2 Data 70
3.4.3 Empirical results 72
3.5 Conclusions 80

4 Study of Hidden Orders in LSE and SSE 85
4.1 Introduction 85
4.2 Hidden Orders 86
4.2.1 Data 87
4.2.2 Detection Algorithm 88
4.2.3 Classification of Hidden Orders 89
4.2.4 Statistical Properties of Hidden Orders 90
4.3 Market Impact 94
4.3.1 Definition 94
4.3.2 The Noisy Nature of Market Impact 95
4.3.3 Impact of Limit Orders vs. Market Orders 96
4.3.4 Impact vs. N 97
4.3.5 Temporary vs. Permanent Impact 99
4.4 Trading Profile 100
4.5 Conclusions 102

5 Summary 109

A Financial Preliminaries 115
A.1 Electronic Markets 115
A.2 Order Book 116
A.2.1 Basic Dynamics of the Order Book 117
A.3 Financial Products 120
A.3.1 Bonds 120
A.3.2 Commodities 120
A.3.3 Stocks and Stock Indices 120
A.3.4 Derivatives 121

B Volatility 123
B.1 Volatility Modelling 123

C Resumen 127

List of Figures

1 **Probability density function of standardized returns**, $P(r')$, **for the stock AZN.** The pdf is shown for time scales $t = 1$ hour. The solid blue line is the pdf for a normal distribution with mean zero and unit variance. 8
2 **Cumulative distribution function of absolute standardized returns**, $C(|r'|)$, **for the stock AZN.** The cdf is shown for time scales $t = 1$ hour. The solid blue line is the cdf for a normal distribution with mean zero and unit variance. 9
3 **Probability density function of standardized returns**, $P(r')$, **for the stock AZN.** The pdf is shown for time scales from $t = 1$ day to $t = 100$ days. The solid blue line is the pdf for a normal distribution with mean zero and unit variance. 10
4 **Autocorrelation function of returns for AZN for time scales of minutes.** 11
5 **Autocorrelation function of absolute returns for AZN for time scales of minutes.** 13
6 **Autocorrelation function of squared returns for AZN for time scales of minutes.** 13

2.1 **Probability density function of a Lévy distribution compared to a normal distribution.** 42
2.2 **Probability density function of standardized returns**, $P(r')$, **for AZN compared to a fitted Lévy distribution and a normal.** The solid blue line is the pdf for a normal distribution with mean zero and unit variance. 43
2.3 **Probability density function of Student's t-distribution for different values of** ν **and a normal.** The solid blue line is the pdf for a normal distribution with mean zero and unit variance, Student's t-distribution is plotted for ν values from 2 to 30. 49
2.4 **Probability density function of standardized returns for AZN stock fitted by a Student's and a normal distribution.** 50

3.1 **AZN. The probability density of daily** β **fit by a gamma distribution.** . . 73
3.2 **AZN. The probability density of** ξ^* **for different** τ **compared to** $\mathcal{N}(0,1)$**.** 73
3.3 **AZN. The probability density of returns for different** τ **compared to theory.** 74
3.4 **AZN. The cumulative distribution of returns for different** τ **compared to theory.** 74

3.5 **The cumulative distribution of β compared to the cumulative distribution from the best fit to a gamma distribution for all stocks in the LSE group.** 76

3.6 **The normalized probability density of returns with $\tau = 80$ compared to theory for all stocks in the LSE group.** 76

3.7 **Collapse of the complementary cumulative distribution (ccd) of absolute scaled returns, $C(|r'|)$, for the stock IBM.** The ccd is shown for times scales $\tau = 10$ to $\tau = 640$. The solid black line is the theoretical ccd using $\beta_0 = 1.4 \times 10^7$ and $n = 3.89$ from fitting β to a gamma distribution. Inset: ccd of the slow fluctuating variable β for IBM, the red curve is the empirical ccd and the solid black line is a fit to a gamma distribution. 78

3.8 **Collapse of the return distribution on the function $f(r')$, for the stocks studied in this group.** For each stock, the return distribution for $\tau = 80$ is shown in logarithmic coordinates. Inset: The same plot in regular coordinates. 79

3.9 **Predicted vs. empirical tail exponent for the stocks under study.** The tail exponent is the asymptotic slope of the tail of the ccd when measured in logarithmic coordinates. The dashed line shows $y = x$ for comparison only. 79

4.1 **Exponents g_i ($i = 1,2,3$) of the allometric relations of Eq. 4.5**. For each of the stocks considered in our LSE and BME databases and for hidden orders with $N \geq 10$ and $T < 1$ day, as a function of the number of detected hidden orders per year. Error bars are 95% confidence intervals obtained by bootstrapping the data. In the analysis of market impact we consider only stocks with at least 250 hidden orders per year (those in the white area of the figure). 91

4.2 **Ensemble statistics of the fraction of market orders f_{mo} and participation rate α of hidden orders in both the BME and LSE.** Left panels show the probability distribution function of both parameters, while the right panel shows the conditional average of the participation rate conditioned on a given value of f_{mo}. 93

4.3 **Conditional average $\langle R|T \rangle$ of the rescaled impact of hidden orders** (Eq. (5.11)) as a function of their time duration T (symbols) compared to the average return of the stock market index over random periods of the same time duration (solid lines). The inset shows the price of the FTSE100 and IBEX35 indices over the period of study. In this figure we are using all detected hidden orders without any conditioning on T or f_{mo} values but with $N > 10$. 96

4.4 **Average rescaled market impact R for hidden orders shorter than 1 day as a function of N for the BME (left) and LSE (right).** Circles are the results for all hidden orders, while squares are the results when there is a low fraction of market orders ($f_{mo} < 0.2$) and diamonds are for when there is a large fraction of market orders ($f_{mo} > 0.8$). Dashed lines are power law fits $R \sim N^{\gamma}$. Values of γ are reported in Table (4.2). 98

4.5 **Market impact versus time.** 100

4.6 **Trading profile inside the hidden order.** Average volume of the transactions within the hidden order divided by the average volume in the hidden order as a function of the normalized time t/T. Circles are the results for all hidden orders, while squares are the volume traded in the market (in the same stock) concurrently with the hidden order. Data is only for hidden orders with $f_{mo} > 0.8$. 101

4.7 **Initial and final times of the hidden orders.** Probability distributions of the initial time t_i and final time t_f of the hidden orders, measured with respect of the time of the day. Data is only for hidden orders with $f_{mo} > 0.8$ 102

A.1 **A schematic of the order book used in modern electronic markets.** . 118

List of Tables

3.1 Table of parameters for five stocks studied in this section. 77

4.1 Statistics of the hidden order ensembles used in the paper. Only hidden orders with $T < 1$ day and $N > 10$ transactions are used. 92

4.2 Parameters of the fitting of the market impact with Eq. 4.15. 99

Chapter 0

Introduction

0.1 Science and Finance

Louis Bachelier's thesis entitled "Théorie de la Spéculation" is a pioneer work in financial mathematics [19]. In his dissertation, Bachelier developed a model for explaining the price variations of French government bonds, and also performed an empirical study to check his theory. Moreover, he presented the theory of the Random Walk for the first time - he predated Einstein's work on Brownian motion by five years. Although it is considered as one the first attempts of applying mathematical methodology to a financial problem, there were earlier approaches from mathematics to finance, e.g. Carl Friedrich Gauss studied the pensions fund for widows of the professors of the University of Göttingen (1845-1851). This is a seminal application of probability theory to finance.

In 1908, Vinzenz Bronzin, a professor of mathematics at the Accademia di Commercio e Nautica in Trieste, published a booklet in German entitled Theorie der Prämiengeschäfte (Theory of Premium Contracts) which is an old type of option contract [48]. Almost like Bachelier's dissertation (1900), the work seems to have been forgotten shortly after it was published. However, almost every element of modern option pricing can be found in Bronzin's book. In particular, he uses the normal distribution to derive a pricing equation which comes surprisingly close to the Black-Scholes-Merton formula.

These two authors, Bachelier and Bronzin, did not have much influence on any of his contemporaries. It was sixty years later when financial community began to be interested in stochastic processes as a mean to model price variations.

A physicist, M.F.M. Osborne (1959), rediscovered the Brownian motion of stock markets [40]. He was among the influential advocates of using this model to describe asset returns. Another physicist, Fisher Black, together with Myron Scholes solved the option pricing problem by reducing it to a diffusion equation [6]. Gaussian random walk would be assumed to underlie asset price dynamics when such basic financial economics concepts as the Black-Scholes-Merton model [6, 39], and the Capital Asset Pricing Model (CAPM) would be developed.

It was in 1963 when Mandelbrot postulated that the stochastic process describing financial time series would deviate from geometric Brownian motion in fundamental and essential aspects [36]. Although tests performed by Bachelier and Roberts [19, 44] seemed to be in agreement with theory, the empirical study of cotton prices by Mandelbrot showed that returns of this commodity were not normally distributed.

Given the major importance of finding a plausible description and understanding of the true distribution of returns for asset allocation, risk management, and option pricing - in addition to the scientific challenge,- a large number of recent papers on the subject have been written by physicists [1, 9, 25, 26, 28, 29, 38, 42, 47]. Much effort has been done from a theoretical and empirical point of view. Several new models have been produced for attempting to explain new empirical facts found in the study of data sets recorded in modern electronic markets.

0.2 Complexity in Financial Markets

A taxonomy of market paticipants is really depending on the set of criteria employed to classify them. We can talk of enterpreneurs, people who need funding, and investors, people who have money to invest. Another possible classification is that composed of hedgers, brokers-dealers, and speculators. Hedgers are people willing to protect themselves of market risks, broker-dealers are people who provide liquidity to the market by buying or selling at any given moment, and speculators are investors in the short to medium term. Speculators are commonly used in contrast to investors, these last ones are assumed to take positions in the long term.

Another possible classification of the market participants may be made by considering if they are buying or selling exchange services, and liquidity is the service taken into account. Liquidity is the ability to trade when you are willing to trade. Based on the consumption of this service, we distinguish two sides: buy side, and sell side. Thus, there are market participants buying liquidity, while others are selling it. The buy side includes individuals, funds, firms who buy financial products, e.g. bonds and stocks, to move their income from the present to the future. Funds are not the only institutions in the buy side, we can also include in this side to trusts, endowments, and foundations. The sell side includes dealers and brokers.

A relevant conclusion of any possible taxonomy of market participants is that financial markets are systems which contain multiple agents, of different types (producers and consumers; risk averse and risk takers; firms and individuals, etc.), all competing for finite resources of some kind or another, and interacting in such a way as to generate the properties and dynamics of economic systems and subsystems. Therefore, financial markets are good candidates for being considered complex systems. Econophysicists agree that these properties and the dynamics fit the complex system requirements: scaling and universality, criticality, fractal patterns, and (candidates for) emergent properties. All attributes that a good complex system should possess.

It is a broadly tested fact that financial market time series display statistical regularities [18]. These regularities have similar characteristics to those observed in other complex systems in the physics of critical phenomena. In particular, one can interpret the stylized facts [43] as scaling laws. In other words, given that financial markets have a physical composition like that of systems dealt with in statistical physics (large numbers of interacting individuals) and given, furthermore, that the time series exhibit statistical regularities similar to that of systems dealt with in statistical physics [31], it follows that a good modeling strategy is to apply statistical physics to financial markets. The huge amount of data collected about financial markets makes them an excellent field of research from the point of view of complexity. There are few areas which have as much data recorded so accurately and at so many time scales.

0.3 Stock Price Dynamics

The aim of the first part of this thesis is to find an explanation to stocks price dynamics. This explanation will let us obtain a theoretical returns distribution that matches empirical data. The theoretical results on price dynamics are relevant from a scientific point of view because they mean the understanding and accurate representation of a complex system, and from an economical perspective they are important due to the repercussion of these results in several financial areas such as option pricing and portfolio management. In spite of its relevance, the asset price dynamics problem has been broadly studied for more than a century without achieving definitive and conclusive results.

Gaussian models were a first attempt to describe and explain price dynamics. This theoretical approach was originally developed from first principles. For this, Bachelier [19] postulated a set of conditions that a theoretical market and price series should fulfill in order to avoid prices were predictable. The conditions about the market were three: perfect market, efficient market, and complete market, but they can be reduced to only two conditions: efficient market, and complete market, given that efficient market is a more flexible version of perfect market. The condition about prices stated that successive prices should be statistically independent. These conditions about market and prices are not enough for deriving the exact shape of the probability distribution of prices. For doing this, Bachelier employed three different mathematical reasonings to reach the theoretical distribution that matched all the conditions and solved the problem. Although he made a mistake and considered normal distribution as the only possible solution to the problem when others were also correct, the mathematical methods employed in the deduction of the solution have been broadly employed later in mathematical finance, i.e. the Random-Walk hypothesis, martingale methods, and the use of the Chapman-Kolmogorov-Smoluchowski equation in finance.

Bachelier's work is a milestone in mathematical finance because it provided a solid support to much of the modern financial framework. In spite of this achievement, some results obtained by his model such as the acceptance of negative prices, or considering returns as absolute price variations instead of relative variations were against basic economic principles. All these problems were fixed by the standard Gaussian model which is commonly known as Black-Scholes-Merton (BSM) model [6, 39]. This theoretical approach constitutes the basis to modern derivatives pricing.

Black-Scholes-Merton model and Bachelier's model may be classified as members of the family of Gaussian models. The former assumes a normal distribution for returns, whereas the latter assumes it for prices. Although Gaussian models meant a major advance in the understanding and modelling of price dynamics, it was demostrated that they could not reproduce certain characteristics of empirical distributions [36, 43] such as the frequency of large returns or the change of shape of returns distribution.

0.4 Non-Gaussian Returns Distribution

Empirical results, like these shown in Section(0.6), not only demonstrated that Gaussian models were not in agreement with empirical data, but these models were not theoretically valid to describe empirical observations because they could not produce heavy-tailed distributions. Two new frameworks were conceived as an explanation to empirical data. The first of them was postulated by Mandelbrot [36] after finding that time series of cotton prices were fitted for different time intervals by a Lévy distribution which is a heavy-tailed stable distribution. Given that the shape of the empirical distributions was nearly constant for the studied intervals, it was assumed they were stable. Although this model represented a solution to the findings against Gaussian models, it also presented a serious theoretical inconvenient: variance was not finite. The acceptance of a solution with that characteristic created important inconsistency problems to other financial models, e.g. models based on the paradigm of mean-variance. In addition to this theoretical problem, a new finding showed that empirical returns distributions converged to a Gaussian when they were aggregated at larger time scales. Stable Distribution (SD) models were not able to justify the convergence because heavy-tailed stable distributions keep unchanged their shape under aggregation up to rescaling. As a possible solution, it was postulated that returns distributions followed a Truncated Lévy Distribution (TLD) [38] which is a crossover between a Gaussian and a pure Lévy distribution. These two regions with different behavior are delimited by an additional parameter.

Clark proposed a different framework: the Mixture Distribution Hypothesis (MDH) [15], which is the basis for many different models. His original work assumed that returns distribution was subordinated to Gaussian, consequently another distribution was necessary to explain the observed properties of the time series. That distribution was related by Clark to different information arrival rates, hence these changing rates were the cause of the non-Gaussian properties of empirical data. The main problem of this model is that information is not directly observable, then it must inferred from another variable or this another variable must be taken as a proxy. From a mathematical point of view, the subordination is the result of compounding two distributions: a Gaussian, and another distribution called the directing process. This directing process was originally considered by Clark as a clock measuring the rate of evolution, but this is not the only possibility. There are other candidates that justify different implementations with different distributional shapes, being the most relevant of them for this thesis the one assuming a Gamma distribution as the directing process because it generates a Student's t-distribution [7] for the returns, and this is exactly the shape of the solution found in the new model presented here. This solution is appealing because it is heavy-tailed and converges to a Gaussian. Hence it seemed that a simple explanation to heavy tails for non aggregated returns and the convergence to a normal for longer time intervals was possible. However this solution was theoretically deduced instead of empirically observed. So if the Gamma were empirically explained, the Student's t-distribution of the returns would be theoretically justified.

Stochastic Volatility (SV) models [46] may be classified into the MDH family due to they meet all the assumptions of this framework. SV models employ volatility as the directing process, making necessary the modelling of the returns volatility which is a latent variable as it is information. Therefore volatility estimation is crucial for the properties of the returns generated by this model, but volatility can be computed and even defined in several different ways, e.g. realized volatility, implied volatility, and conditional volatility. The problem with the different models within MDH framework, e.g. SV models, is not theoretical as it was with SD and Gaussian models, but it arises from the specific implementations which are not able to explain the entire shape of the empirical distribution. Being more evident the problems in the tails of the distribution where theoretical models usually fail in matching empirical data. Hence the question to be answered is wether based on MDH assumptions it is possible to find a new model which accurately matches empirical findings.

0.5 Statistical Mechanics and Finance

As I mentioned in Section(0.2), financial markets are complex systems with a similar behavior to this observed in the physics of critical phenomena. The Superstatistics [2] is a branch of statistical mechanics which is devoted to the study of non-linear and non-equilibrium systems such as turbulences, cosmic ray statistics, and solar flares. Superstatistical systems must hold some conditions [3] to be treated like that. The first condition is the existence of an intensive parameter called β. The second condition is the existence of two dynamics: a fast dynamics, and a slow dynamics. And the third condition is that when the system is divided into cells which may be spatial or temporal the value of β within each cell has to be constant or nearly constant but may be different from cell to cell.

Superstatistical systems are described in a similar mathematical way than SV models are, but they are not absolutely equivalent. Superstatistical systems are doubly stochastic where the first stochastic element is normally distributed and represents the distribution of the physical magnitude conditioned to the intensive parameter, and the second stochastic element is a normalizable distribution representing the distribution of the intensive parameter. These mathematical elements are common to both models, but there are important differences between them such as the non explicit existence of a slow and a fast dynamics in SV models, or the constant value of the intensive parameter in a limited region as it is required in superstatistical systems.

I have developed a new model where superstatistical concepts are employed in the modelling of a high-frequency financial time series, in order to test if financial systems can be treated as a superstatistical complex system. For this, time series is divided into pieces of length a day which is taken as the dimension of the cell, volatility is considered the intensive parameter and assumed to be constant along each day, and finally slow dynamics is this of the volatility and fast dynamics that of the returns. This makes a difference with a common SV model where usually returns and volatility are sampled at the same frequency. The latter is drawn from the volatility distribution at a time step and its composition with another value pulled from a normal produces the return at that time step. In the new model all returns in a day share the same value for the volatility, hence volatility is not sampled at the same frequency than returns are.

0.6 Statistical Properties of Empirical Returns

Empirical returns distributions are characterized for a set of properties common to most of the financial assets: bonds, equities, commodities, indices... These statistical properties are usually described in a qualitative way by mean of what Nicholas Kaldor [32] called stylized facts instead of using a very detailed mathematical description. The reason is that by losing in individual quantitative details we gain in generality when dealing with a broad variety of assets and markets. Kaldor suggested that theorists should be free to start off with a stylized view of the facts, i.e. concentrate on broad tendencies, ignoring individual details. The existence of stylized facts gives us a proof about the fact that different assets affected by different news share common statistical properties. Thus it seems plausible the idea of a universal behavior for price dynamics independently of the specific asset. Moreover, this is a proof against a common point of view shared by many practitioners in financial markets: the event-based approach which tries to explain price dynamics based on political and economic announcements.

It has been reported a vast diversity of stylized facts [8, 11, 12, 30, 35, 41], some of them related to the shape of returns distribution, e.g. heavy tails, aggregational gaussianity; to the volatility of returns, e.g. volatility clustering; to the correlations, e.g. absence of linear autocorrelations of returns, slow decay of autocorrelation in absolute returns, volume/volatility correlation. There are others not included into these categories, e.g. asymmetry in time scales, and intermittency. Finally, there are some stylized facts which specifically appear at high frequencies, e.g. those ones related to price formation: negative first-order autocorrelation of returns, discreteness of quoted spreads, or short-term triangular arbitrage. I am mainly focused on the stylized facts related to the shape of returns distribution and those ones that cause it, because the more accurately are reproduced empirical distributions the better is the theoretical price dynamics process.

0.6.1 Heavy Tails

This stylized fact is related to the shape of emprical returns distribution, and it makes reference to the insufficiency of the normal distribution for modelling the marginal distribution of stock returns. The main difference between a heavy-tailed distribution and normal is observed in the tails. This is the reason because the comparison of the tails of empirical and theoretical distributions is that important, and the extensive use of log-log plots to emphasize this part of the distributions.

In Fig.(1) I show the probability density function (pdf) of the standardized empirical returns for AZN with prices sampled at hourly frequency in log-linear coordinates. This is compared to a normal distribution with zero mean and unit variance for showing the differences between them. Empirical returns distribution is leptokurtic this means that it is fat-tailed and large returns happen more often than we would expect in a normal distribution. This is quite evident at the tails of the distributions where we find more density of events in the AZN distribution than in the same region of the normal. In addition to this, the peak at $|r'| = 0$ is more pronounced in the AZN distribution. A similar distribution of returns has been observed in the other assets studied in this thesis. Empirical distributions is leptokurtic at different time scales, only being more pronounced this behavior at high frequencies [8].

In Fig.(2) I show the cumulative distribution function (cdf) of the standardized empirical returns distribution for AZN with prices sampled at hourly frequency in log-log coordinates. This is compared to the cdf of a normal distribution with zero mean and unit variance. The behavior of the distributions is clearly different at the tails, more specifically in this last plot we can see that

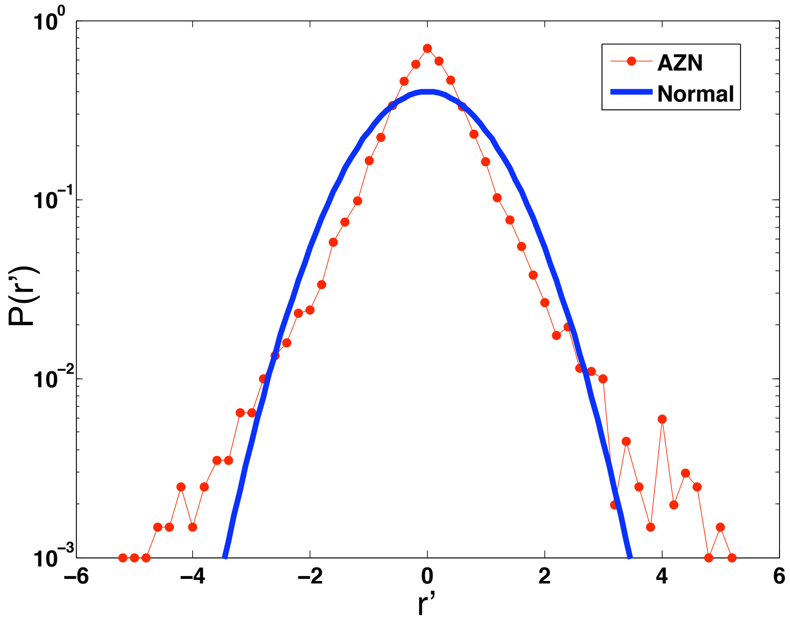

Figure 1: **Probability density function of standardized returns**, $P(r')$, **for the stock AZN.** The pdf is shown for time scales $t = 1$ hour. The solid blue line is the pdf for a normal distribution with mean zero and unit variance.

the tail of the empirical distribution slowly decays whereas normal distribution shows an abrupt decay.

The deviation from normality may be also expressed by mean of the kurtosis of the distribution defined as

$$\kappa = \frac{\langle (r(t,T) - \langle r(t,T)\rangle)^4\rangle}{\sigma^4(T)} - 3, \tag{1}$$

where $\sigma^2(T)$ is the variance of the log returns[1]. The kurtosis as it has been defined in Eq.(1) takes value 0 for a Gaussian distribution. A positive value means the distribution is fat-tailed. This implies that probability density function slowly decays in the tails.

A problem when studying heavy-tailed distributions is that standard deviation is not enough to measure the variability of return distributions. Then we need to take into account higher-order moments of the distribution. However, these moments can be not well-defined. I use the term well-defined if they take a finite value. An alternative method is the tail index of the distribution, k, and it represents the highest well-defined absolute moment. For a Gaussian distribution all the moments exist, then $k = \infty$. As a general rule, the lower the index the fatter the tail.

[1] I have computed returns in the usual way. Given the price of the asset, $p(t)$, at time t. $X(t)$ represents its logarithm, and $r(t,T) = X(t+T) - X(t)$ represents the return of the asset in the time interval Δt.

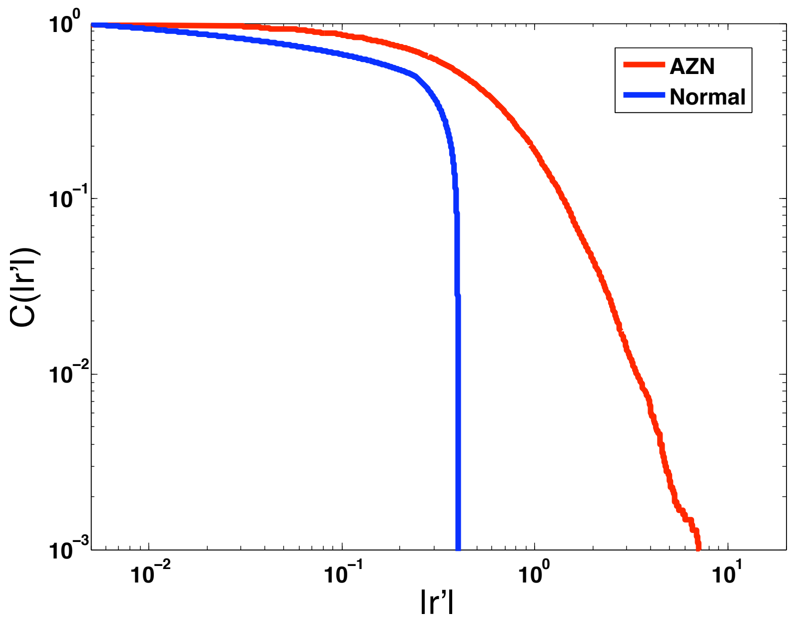

Figure 2: **Cumulative distribution function of absolute standardized returns**, $C(|r'|)$, **for the stock AZN.** The cdf is shown for time scales $t = 1$ hour. The solid blue line is the cdf for a normal distribution with mean zero and unit variance.

Another method, suggested by Mandelbrot [36, 37] , is based on the representation of moments as a function of the sample size. The idea behind the method is that if we have a finite theoretical moment, the empirical one will converge to the theoretical value when increasing the size of the sample. Otherwise, theoretical moment is not well-defined for that case. The larger the sample size the larger the value of the empirical moment. Then, it does not converge to a specific value.

0.6.2 Aggregational Gaussianity

This is another stylized fact about the shape of returns distribution, and shows us how empirical distributions come closer to a normal when prices are sampled at lower frequencies. This statement can be rephrased by considering disaggregation of returns, so the more disaggregated the returns are the more leptokurtic the empirical distributions are.

In Fig.(3) I show the probability density function (pdf) of the standardized empirical returns for AZN for time scales $t = 1$ day, $t = 10$ days, and $t = 100$ days. This is compared to a normal distribution with zero mean and unit variance for showing how when time interval of the sampled data increases, the pdf of the empirical returns comes closer to a Gaussian. We can see that when time interval increases the frequency of events in the tails decreases and empirical distributions tend to overlay normal distribution. This is a slow process making

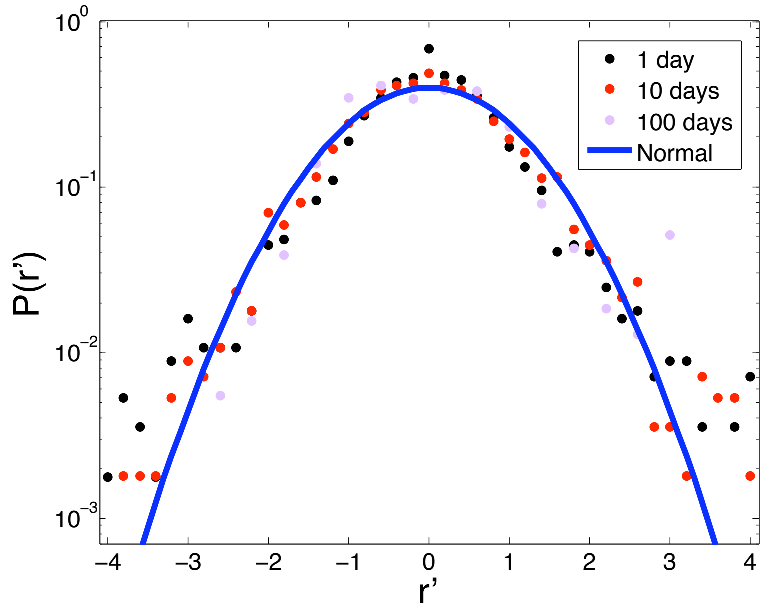

Figure 3: **Probability density function of standardized returns**, $P(r')$, **for the stock AZN.** The pdf is shown for time scales from $t = 1$ day to $t = 100$ days. The solid blue line is the pdf for a normal distribution with mean zero and unit variance.

necessary to consider long time intervals for being more evident.

Apparently, the distribution changes as a function of time interval. Then, we could expect that distribution shape for high-frequency returns should be different from that sampled at lower frequencies. For testing this hypothesis, we check if these distributions may be collapsed onto a master curve. For collapsing the distributions we only need to find a certain translation and dilation of the returns, such as

$$P(r_t, t)dr_t = P_1(r_1)dr_1, \tag{2}$$

where $r_t = a_t r_1 + b_t$ is the equation for translation and dilation. If we find the solution for Eq.(2) we may transform a distribution into the other one, and distributions are defined as scale invariant.

It is shown in Chapter 3 that returns distributions at different Δt collapse for high-frequency returns, but they do not collapse onto a Gaussian distribution. This is a proof against Gaussian models because if they were right, then empirical distributions should perfectly collpase onto a Gaussian distribution no matter of the observed time interval. The apparent contradiction can be solved if we take into account that the Gaussian is a particular case of stable distribution. This family of distributions is characterized for acting as attractors. This means that if we sum a large number of random variables we get as limit distribution a stable distribution. The Central Limit Theorem (CLT) is only its precise

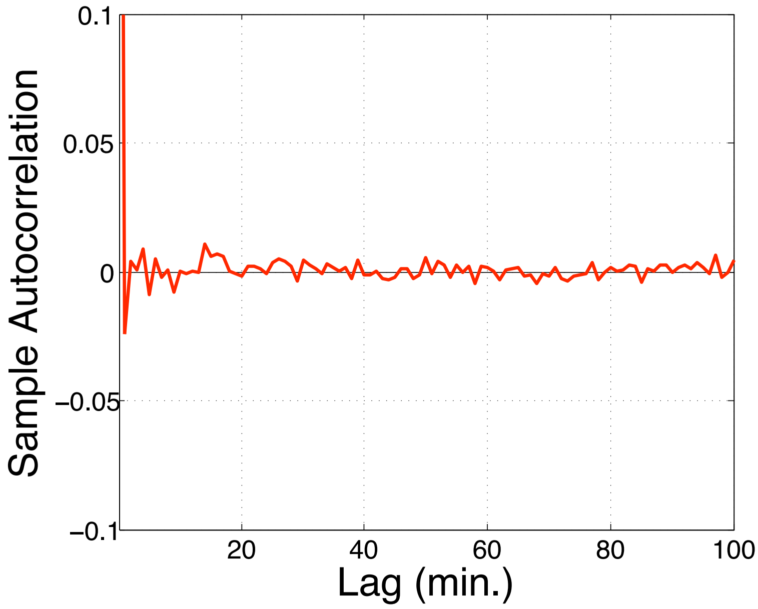

Figure 4: **Autocorrelation function of returns for AZN for time scales of minutes.**

formulation. A consequence of this theorem is that if random variables with finite variance are added, we finally get a Gaussian distribution. Then, a possible solution would be that empirical distributions were drawn from a non stable distributions with finite variance. The sum of random variables would approach to a Gaussian as a consequence of CLT.

0.6.3 Absence of Linear Autocorrelation

This stylized fact has been extensively studied and reported [24, 41] due to the importance that it has for theoretical models. Given that if returns were strongly correlated they would be predictible and this information could be used by trading strategies to make net profit based on a statistical forecast of the next few returns, being the number of forecasted returns related to the persistence of the autocorrelation. The absence of autocorrelations is generally taken as a support of the Efficient Market Hypothesis (EMH) which is one of the basic assumptions made by Bachelier in his model. Empirical studies show that autocorrelation decays rapidly and in a few minutes it can be assumed to be zero or negligible. Correlation is commonly expressed as

$$C(\tau) = \mathrm{corr}(r(t, \Delta t), r(t+\tau, \Delta t)), \tag{3}$$

where corr is the sample correlation.

In Fig(4) I show the autocorrelation function of returns for AZN for time scales of minutes. We can see that autocorrelation decays very fast being negligible in a few minutes, we can also observe in the first minute a certain negative autocorrelation due to microstructural effects wich may be attributed to the action of market makers.

0.6.4 Volatility Clustering

It has been shown that returns are almost linearly uncorrelated, but this is not the case for nonlinear functions of returns such as absolute or squared returns which clearly exhibit positive autocorrelation, also known as persistence, this is an evidence of a well-known phenomenon called volatility clustering which may be expressed in a general form by stating that large price variations tend to be followed by large price variations. Based on the volatility clustering is possible to make forecasts about the magnitude of the next returns. Moreover, volatility clustering is against the independence of the returns because independence implies that any nonlinear function must show no autocorrelation [13, 17]. For the first part of this thesis volatility clustering is important because it gives support to the assumption about the slow dynamics of the volatility.

The autocorrelation of the squared returns is a common method for measuring volatility clustering, such as

$$C(\tau) = \mathrm{corr}(|r(t,\Delta t)|^2, |r(t+\tau,\Delta t)|^2). \tag{4}$$

In Fig(5) I show the autocorrelation function of absolute returns for AZN for time scales of minutes. We can see how the autocorrelation function for these returns decays slowly and stays significant for long time intervals. It has been indeed reported that it remains significantly positive over several days [8, 17, 20, 21, 22, 16].

In Fig(6) I show the autocorrelation function of squared returns for AZN for time scales of minutes. We can observe that the autocorrelation of squared returns also decays slowly but faster than this of absolute returns, then absolute returns are more predictable than squared returns. This has been already observed by Ding and Granger [20, 21], their empirical results show that the autocorrelation function is highest for the absolute returns than for other powers of returns. Independently of how slowly decays the autocorrelation function, this is taken as a proof of long-range dependence in volatility.

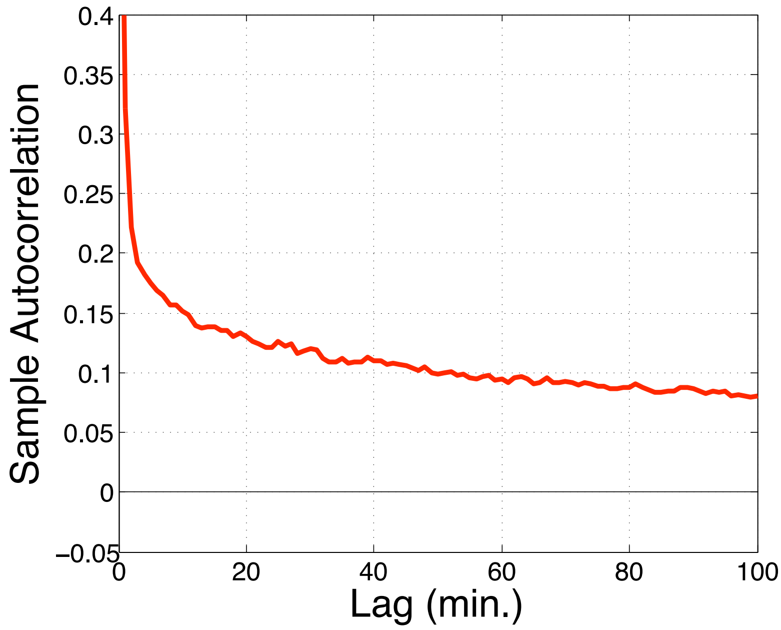

Figure 5: **Autocorrelation function of absolute returns for AZN for time scales of minutes.**

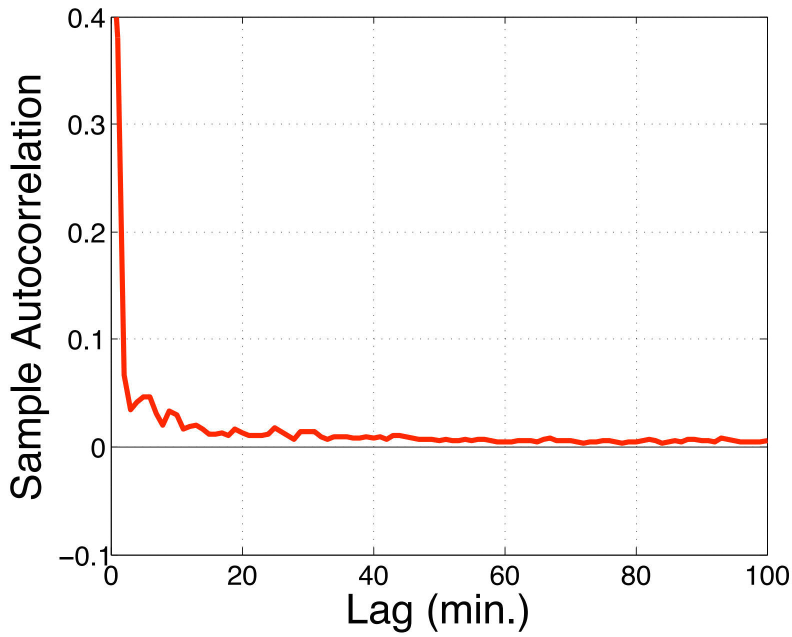

Figure 6: **Autocorrelation function of squared returns for AZN for time scales of minutes.**

0.7 The Problem of Large Orders Execution

Markets are places, which can be physical or an electronic system, where market participants gather to make transactions. These transactions can be studied as if they were made by anonymous agents, or taking into consideration the identities of the different participants. In the first part of this thesis, I have studied price changes disregarding the participants who caused them being only interested in the magnitude of the variation. In the second part, I differentiate transactions depending on the members involved in them. In doing this, I may trace the transactions made by every single market member and based on this information I can study the execution strategies implemented by the different participants. These strategies are especially relevant in the execution of a large order because this may cause important transaction costs due to the lack of available liquidity in the order book at a given time. Therefore this type of execution turns into an optimization problem for minimizing transaction costs. Large orders are usually split into pieces and executed incrementally for taking advantage of the available liquidity at any time. I call hidden order to any order executed following this procedure. Although there exist orders like iceberg orders [23] that let participants show only a certain fraction of the total volume at a time, hidden orders constitute a broader concept of execution because they are neither constrained to a single price of execution nor to the exclusive use of limit orders. Iceberg orders may be considered a particular and restrictive case of hidden orders.

The study of hidden orders brings us the opportunity of understanding a complex system where the participants fit their strategies to optimize the transaction costs. However the main problem with hidden orders is that they are not explicitly submitted as one only order, and their pieces are not identified as making part of a larger order. Consequently they must be inferred from transactions data. Otherwise they only would be available for research if we had the explicit information of a market participant [14] who had submitted them. Although we can find several methods [5, 27, 34] in the literature for detecting hidden orders, the underlying problem is that we can not assure we have accurately classified all of them.

The aim of the empirical research presented in the second part of this thesis is the determination of the functional form of the impact of transactions on stock price, also known as market price impact, of hidden orders. The functional form is important [33] to quantify total market impact of a large order, and market impact is the main factor of transaction costs. Therefore the functional form of price impact is a crucial element of any optimized execution. Moreover since impact is a cost of trading, it exerts selection pressure against a fund becoming too large, and therefore is potentially important in determining the size distribution of funds [4, 45]. Finally, market impact reflects the shape of excess demand, which is of central importance in economics. Despite its conceptual and practical importance, a proper empirical characterization and theoretical understanding of market impact is still lacking [10]. In Chapter 4, I show the results of the empirical study for the hidden orders obtained from the London Stock Exchange (LSE) and the Spanish Stock Exchange (SSE). These results are relevant because they show for the first time a comparative study of this type of execution in two markets. Although the development of a theoretical explanation to the empirical findings it is out of the scope of this thesis, the fact that we observe similar behavior in both the London and Spanish stock exchanges, and that others have also observed this in the New York Stock Exchange, suggests the possibility of a “law" for market impact.

Bibliography

[1] K. E. Bassler, J. L. McCauley, and G. H. Gunaratne, Proc. Natl. Acad. Sci. U.S.A. **104**, 17287 (2007).

[2] C. Beck, E. G. D. Cohen, Physica A **322**, 267 (2003).

[3] C. Beck, E. G. D. Cohen, and H. L. Swinney, Phys. Rev. E **72**, 056133 (2005).

[4] J. B. Berk, and R. C. Greene, Journal of Political Economy **112**, 1269 (2004).

[5] P. Bernaola-Galvan, P. Ch. Ivanov, L.A.N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).

[6] F. Black, and M. Scholes, Journal of Political Economy **81**, 637 (1973).

[7] R. C. Blattberg, and N. J. Gonedes, J. Business **47**, 244 (1974).

[8] T. Bollerslev, R. C. Chou, and K. F. Kroner, J. Econometrics **52**, 5 (1992).

[9] J.-P. Bouchaud, and M. Potters, *Theory of Financial Risks and Derivative Pricing* (Cambridge Univ. Press, Cambridge, U.K., 2003).

[10] J.-P. Bouchaud, J. D. Farmer, and F. Lillo, How markets slowly digest changes in supply and demand, *Handbook of Financial Markets: Dynamics and Evolution*, T. Hens, and K.R. Schenk-Hoppé (Editors), (North-Holland, NY, 2009).

[11] W. A. Brock, and P. J. F. de Lima, Nonlinear time series, complexity theory and finance, *Handbook of Statistics Volume 14: Statistical Methods in Finance*, G. Maddaland, and C. Rao (Editors) (North-Holland, NY, 1995).

[12] J. Campbell, A. H. Lo, and C. McKinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, NJ, 1997).

[13] J. Campbell, A. H. Lo, and C. McKinlay, *A Non-Random Walk Down Wall Street* (Princeton University Press, Princeton, NJ, 1999).

[14] L. Chan, and J. Lakonishok, Journal of Finance **50**, 1147 (1995).

[15] P. K. Clark, Econometrica **41**, 135 (1973).

[16] F. Comte, and E. Renault, J. Econometrics **73**, 101 (1996).

[17] R. Cont, M. Potters, and J.-P. Bouchaud, Scaling in stock market data: stable laws and beyond, *Proc. CNRS Workshop on Scale Invariance, Les Houches, 1997*, Dubrulle, Graner, and Sornette (Editors) (Springer, Berlin, 1997).

[18] A. C. C. Coolen, *The Mathematical Theory of Minority Games: Statistichal Mechanics of Interacting Agents* (Oxford University Press, Oxford, 2004).

[19] P. H. Cootner (ed.), *The Random Character of Stock Market Prices* (MIT Press, Cambridge, MA, 1964).

[20] Z. Ding, C. W. J. Granger, and R. F. Engle, J. Empirical Finance **1**, 83 (1983).

[21] Z. Ding, and C. W. J. Granger, Working Paper, University of California, San Diego, (1994).

[22] R. F. Engle, *ARCH: Selected Readings* (Oxford University Press, Oxford, 1995).

[23] A. Esser, and B. Mönch, Finance Research Letters **2**, 68 (2007).

[24] E. F. Fama, J. Business **38**, 34 (1965).

[25] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, Quant. Finance **4**, 383 (2004).

[26] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, Nature **423**, 267 (2003).

[27] A. A. Gerig, *Theory for Market Impact: How Order Flow Affects Stock Price.* PhD thesis, University of Illinois, (2007).

[28] L. Gillemot, J. D. Farmer, and F. Lillo, Quant. Finance **6**, 371 (2006).

[29] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 6519 (1999).

[30] Ch. Gourieroux, and J. Jasiak, *The Econometrics of Individual Risk: Credit, Insurance, and Marketing* (Princeton University Press, Princeton, NJ, 2007).

[31] N. F. Johnson, P. Jefferies, and Pak Ming Hui, *Financial Market Complexity: What Physicists can Tell us About Market Behaviour* (Oxford University Press, Oxford, 2003).

[32] N. Kaldor, Capital accumulation and economic growth, *The Theory of Capital*, F. A. Lutz, and D. C. Hague (Editors) (Macmillan, London, 1961).

[33] R. Kissell, and M. Glantz, *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk* (American Management Association, 2003).

[34] B. Tóth, F. Lillo, and J. Doyne Farmer, Eur. Phys. J. B **78**, 235 (2010).

[35] G. S. Maddala, and C. R. Rao (ed), *Handbook of Statistics: Statistical Methods in Finance, vol 14*, (North-Holland, Amsterdam, 1997).

[36] B. B. Mandelbrot: J. Business **36**, 394 (1963).

[37] B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration and Risk* (Springer, NY, 1997).

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

[39] R. C. Merton, and J. Bell, Econ. Manag. Sci. **4**, 141 (1973).

[40] M. F. M. Osborne, Operations Research **7**, 145 (1959), reprinted in [19].

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

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

[43] R. Cont, Quantitative Finance **1**, 223 (2001).

[44] K. V. Roberts: in Cootner [19]

[45] Y. Schwarzkopf, and J. D. Farmer, preprint at arXiv:0807.3800 (2008).

[46] N. Sephard, *Stochastic Volatility: Selected Readings*, (Oxford Univ. Press, USA, 2005).

[47] P. Weber, and B. Rosenow, Quantitative Finance **6**, 7 (2006).

[48] H. Zimmerman, and W. Hafner. Amazing Discovery - Professor Bronzin's option pricing models (1908), Manuscript, long version, Universität Basel, (2004).

Chapter 1

1 Gaussian Models

1.1 Introduction

Louis Bachelier, in his dissertation, presented the first theoretical model on asset price dynamics [6]. This was deduced from first principles and based on a series of assumptions about market and prices which made the latter were not predictible by any market participant, because if they were the market itself would be endangered. In addition to these theoretical conditions, Bachelier employed three different mathematical reasonings: the Chapman-Kolmogorov-Smoluchowski (CKS) equation, the Random Walk (RW) hypothesis [11, 14], and the diffusion equation to derive the probability distribution of price variations. These reasonings are currently used in mathematical finance, and stochastic processes - as it was also considered by Bachelier - are the usual probabilistic model for price dynamics. In spite of he made a mistake by considering only one solution, normal distribution, was compatible with his reasonings when in reality others were also valid, his work is not only relevant from a historical point of view but from a theoretical perspective because further models use many of his methods and postulates. Although Bachelier's dissertation may be considered as a purely theoretical work, he not only developed a new model but he indeed performed tests on his theoretical predictions against empirical data. This data set was a price series of the French government bond and the futures corresponding to this asset, surprisingly he found that theoretical predictions produced more large returns than these observed in empirical series when in reality it is the opposite.

An important aspect of Bachelier's theory is the differentiation between two components in the overall dynamics of bond prices: a regular component related to bond's coupon, and a pure random component related to price fluctuations of bond's principal. This distinction is important because it made possible to employ the same mathematical description for apparently different price series by only subtracting the regular component, which was included into the equations as a drift term. The combination of both components may be described by mean of a generalized Wiener process, as it was shown for the first time by Bachelier. This stochastic process applied to prices produces results such as negative prices which are unacceptable from an economical point of view.

Black-Scholes-Merton (BSM) [3] model adopted the assumptions and reasonings of Bachelier's model, but this new solution fixed the results against the economic theory by considering returns instead of prices as the subject to be represented by the stochastic process. Therefore the BSM can be considered more a refinement than a new approach. The most remarkable characteristic shared for both models is that the shape of the distribution of their solutions is the same: a normal Gaussian distribution. This probability distribution has strong implications about the occurrence of large events which were called into question for the empirical observations [9, 18] obtained in the 1960s. As it was presented in the Introduction, empirical returns distribution is leptokurtic and this characteristic is not possible if a Gaussian distribution is held as solution because a normal cannot produce heavy tails. However Gaussian models are a reasonable approach to empirical results at a very aggregated level, that's for returns computed taking prices sampled at a very low frequency, because the convergence to a normal is another characteristic of the empirical distributions. Finally, a very important consequence of the BSM is that gave theoretical support to option pricing [7, 14, 23].

1.2 A First Gaussian Model

In this section I present how Bachelier derived his model from first principles, I also show the mathematical reasonings for deducing the shape of the probability distribution function of prices or returns depending on the specific model. These first principles continue to be held by theoretical models, and the probabilistic methodology employed by Bachelier is broadly studied and employed in mathematical finance. We can consider this first attempt to describe price dynamics as a good approach to many theoretical tools of modern finance.

Although Bachelier's model is highly theoretical, his author was also concerned with empirical results. Indeed, he developed his model for describing the price dynamics of a future of that time, this future had as underlying asset a French government bond. There are important differences between the future studied by Bachelier and a current future traded in any exchange market. The future traded in 1900 in Paris was settled in cash, whereas current bond futures are physical delivery. This difference means that when an old future contract expired the parties involved settled by paying the gains or losses related to the contract in cash, whereas when a modern future expires parties settled by delivering the amount specified of the underlying asset. The second important difference is about expiration date. In the old futures, expiration date could be extended to the end of the next month, and this is not allowed in modern futures. This feature of 1900s futures is assimilated to an embedded optionality because it gives the right to postpone the expiration date. Although futures were Bachelier's main subject of research, he also studied other finacial instruments similar to current options, exotic options, and combinations of options.

The empirical time series employed by Bachelier in his dissertation spanned five years of prices of the future mentioned above and its underlying asset, the French government bond. In spite of this data set, as it is mentioned in Section(1.2.3), Bachelier had not enough data points to determine the validity of his theory considering discrepancies were due to a sample size problem. It is striking that he believed his theory produced too many extreme events when it is exactly the opposite.

Large databases are necessary for testing and validating new financial models which must deal with extreme events. This type of events are very infrequent, then only a small fraction of the entire data set can be classified into this category. Thus, the oiriginal size of the sample is critical to check empirical distributions against theoretical predictions. One of the aims of the first part of this thesis is to find an explanation for the tails of returns distribution. For this, I have studied high-frequency data from different stock exchanges. Only for having an idea of the magnitude of this data set, I may mention that a liquid stock in an active trading day can generate more data points from trades and quotes than the total number of points in Bachelier's database.

1.2.1 Theoretical Assumptions on Market and Prices

In his theory Bachelier postulated that at any instant of time market participants, taken as an ensemble of individuals, did not believe either in falling or in rising prices. This postulate does not mean that single participants have any expectation about future market movements, but when all these individual expectations are aggregated net result is zero. Hence the expected future price variation is zero, or in other words current price is the expected price. This postulate was formulated as a series of assumptions about market and prices that can be summarized into one: prices are not predictable by any participant, otherwise the market itself would be endangered. Market practitioners usually deny the validity of these assumptions because if they were correct markets participants would not be able to beat the market in a consistent way.

Bachelier conditions about markets are three: perfect market, efficient market, and complete market. A market is called a perfect market when all the information available up to present time is completely accounted for by the current price. This condition can be relaxed and reformulated in a slightly different way for obtaining the efficient market condition. This condition allows small irregularities, but these irregularities do not produce net profit if we take into account transaction costs. Thus, the Efficient Market Hypothesis (EMH) [10] rules out the possibility of net profit for trading systems based only on currently available information. The third condition on a market is the complete market condition, this means that we can not deduce any net price movement from outstanding positions. In reality, this condition is only fulfilled in an ideal market because it requires to have outstanding positions in any quoted price, in both sides: buy, and sell. Although this is not true all the time in any asset what it can be assumed as correct is that a net movement can not be deduced from the quotes information. The fourth condition is about prices and it says that successive prices are statistically independent. This last assumption is a basic condition for the Random Walk (RW) hypothesis which is treated in more depth below in Section 1.2.2. This is probably the most tested condition of the all four [1, 9, 10, 17, 19, 21, 22], and it holds for any kind of asset: bonds, indices, commodities, etc. Practitioners by mean of technical analysis - or any other forecasting technique - believe these conditions can be violated, but there is no conclusive proof of it.

1.2.2 Mathematical Description of Price Dynamics

Bachelier focused on the study of bond futures with underlying French government bonds. This underlying asset can be separated into two different parts: coupon, and principal. The coupon can be considered deterministic and accurately described by a linear equation with no uncertainty. The principal follows a random movement depending on the yield dynamics. Therefore, this component makes necessary the use of random processes for describing it. A key concept in financial applications and mathematical finance in general is the martingale, which is a special class of stochastic processes. Martingale is fully understood in the context of betting and gambling. A sequence of random variables $\{X_1, X_2, ..., X_n\}$, is called absolutely fair when for all $n \in \mathbb{N}$ we have

$$\langle X_1 \rangle = 0 \text{ and } \langle X_{n+1} | X_1, ..., X_n \rangle = 0. \tag{1.1}$$

Then we can define another sequence of random variables $\{Y_1, Y_2, ..., Y_n\}$ with $n \in \mathbb{N}$ by $Y_n = \langle Y_1 \rangle + X_1 + ... + X_n$, so we have

$$\langle Y_{n+1} | Y_1, ..., Y_n \rangle = \langle Y_{n+1} | X_1, ..., X_n \rangle = Y_n. \tag{1.2}$$

Then a sequence is a martingale iff $\langle Y_{n+1} | Y_1, ..., Y_n \rangle = Y_n$. We can state that the conditional expected value[1] of an observation at time $t+1$, given all the observations up to some earlier time t, is equal to the observation at that earlier time t,

$$\langle Y_{t+1} | Y_t, Y_{t-1}, ..., Y_1 \rangle = Y_t. \tag{1.3}$$

In general, financial time series on any asset follow an equivalent martingale process, which is a martingale process where future asset prices are discounted by the risk-free rate, or short interest rate, as the interest rate of the discount

[1] In this thesis, I use the common notation in Physics. Thus, $\langle x \rangle$ represents the expectation of x such as $\langle x \rangle = \int_{-\infty}^{\infty} x p(x) dx$.

factor[2]. This property of asset prices is a consequence of perfect and complete market assumptions. Bachelier assumed a martingale process as the random process followed by the price variations of the underlying asset.

Based on the other conditions, Bachelier deduced some properties of the probability distribution of prices. From the complete market assumption, he deduced the symmetry of the distribution about 0. From the fair game condition, he deduced that the maximum of the distribution should be located at $x = 0$, at any time. Finally, he also considered that the probability distribution should be normalizable. This last consideration implies that probability must be integrable in the interval $[-\infty, \infty]$, this implies that probability distribution must decay sufficiently quick. Therefore, some distributions were ruled out, i.e. most of the stable distributions, but other heavy tailed distributions were allowed. Only with these results, it was not possible to derive the exact shape of distribution.

The first mathematical reasoning employed for deriving the probability distribution of price changes is that of the law of multiplication of probabilities. Let $p(x_1, t_1)dx_1$ be the probability of observing a price change from x_1 to $x_1 + dx_1$ at time t_1, and let $p(x_2 - x_1, t_2)dx_2$ be the probability of a price change from x_1 to x_2 in time t_2. The joint probability of a price change from x_1 at time t_1 and to x_2 at time $t_1 + t_2$ is

$$p(x_2, t_1 + t_2)dx_2 = \left[\int_{-\infty}^{\infty} p(x_1, t_1)p(x_2 - x_1, t_2)dx_1\right] dx_2. \tag{1.4}$$

This equation, Eq.(1.4), is known in Physics as the Chapman-Kolmogorov-Smoluchowski (CKS) equation, and it is a convolution equation for the probabilities of statistically independent random processes. The solution to Eq.(1.4) obtained by Bachelier was a normal Gaussian distribution

$$p(x, t) = p_0(t)exp\left[-\pi p_0^2(t)x^2\right]. \tag{1.5}$$

Substituting Eq.(1.5) into Eq.(1.4), we have the condition

$$p_0^2(t_1 + t_2) = \frac{p_0^2(t_1)p_0^2(t_2)}{p_0^2(t_1) + p_0^2(t_2)}, \tag{1.6}$$

which determines the time evolution of $p(t)$ as

$$p_0(t) = \frac{K}{\sqrt{t}}, \tag{1.7}$$

where K is a constant. Then, by substituting

$$\sigma^2 = \frac{t}{2\pi K^2}, \tag{1.8}$$

---

[2]Discount factor, $d(T)$, is the factor by which a future price must be multiplied in order to obtain the present price, $d(T) = \frac{1}{(1+r)^T}$, where r is the interest rate, and T is the time interval.

we recover the usual expression for a Gaussian distribution

$$p(x,t) = \frac{1}{\sqrt{2\pi}\sigma(t)} exp\left(-\frac{x^2}{2\sigma^2(t)}\right). \tag{1.9}$$

This probability distribution meets all the conditions of Bachelier's model. First, at time $t = 0$, $p(x) = \delta(x)$, this means that current price of the asset is known with certainty because of the properties of Dirac delta function. Second, peak and mean of the ditribution are constant. I show later that this is true when we detrend the empirical time series. Therefore, martingale property of the prices holds. And third, $\sigma \propto \sqrt{t}$. This last property of the solution means that probability distribution decays quickly enough for excluding the possibility of large price movements in finite time intervals. The third property is not necessarily against empirical distributions, because it only rules out the possibility of distributions with infinite variance. There are distributions, as it is mentioned in Chapter 2, with fat tails and finite variance. The shape of the tails of the distributions is one of the main arguments against Gaussian models. Although Bachelier obtained a Gaussian distribution as the only possible solution to Eq.(1.4), this is not correct. Other solutions were also corret and compatible with Eq.(1.4).

The Random Walk Hypothesis

We can describe qualitatively a Random Walk (RW) as the resulting trajectory traced by taking successive random steps. This trajectory can be traced by a molecule in a liquid, by the price of a financial asset, or by other object which moves randomly. RW has been applied to many different fields: physics [12], economics [16], finance [17], biology [2] ... Bachelier employed the RW in his dissertation for describing the asset price dynamics. By doing this he predated by five years to Einstein [8], who reached the same solution when he was trying to demonstrate that statistical theory of heat required the motion of particles in suspension.

The classical formulation of the problem related to RW, also known as drunkard's walk, is that a walker can take random steps along a line with equal probability of taking a step either to the left or to the right. The length of step, l, is constant no matter the sense of it, and the number of steps, n, is a natural number. The problem to be solved is the probability associated to find the walker at a certain distance nl, measured from origin, in a certain time interval. We can easily translate the terms of the classical formulation into the description of price movements. We consider the price of a certain asset at time t, time is taken as a discrete variable. At that point the price can move up with an associated probability p, or can move down with another associated probability q. These two events are defined as mutually exclusive and they happen one at a time. Then, we can state that the sum of both probabilities equals 1, $p + q = 1$. The magnitude of price change l, as it happened with step length, is constant for the two possible movements.

For a certain amount of events n, which can be assimilated to steps for the walker's case and to price movements for the asset. The probability of having α up movements of the price, and consequently $n-\alpha$ down movements is binomially distributed

$$p_{u,d}(\alpha, n-\alpha) = \frac{n!}{\alpha!(n-\alpha)!} p^{\alpha}(1-p)^{n-\alpha}. \tag{1.10}$$

When the number of events increases, in the limit $n \to \infty$, $\alpha \to \infty$, and with $k = \alpha - np$ constrained to take a finite value. This follows the expression

$$p(k) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{k^2}{2npq}\right). \tag{1.11}$$

Solving for the particular case of a fair game. This is when p and q are equal to $1/2$. We set $k \to x$, and $t = n\Delta t$. Where t is the total time interval, and Δt is the unit of time for a step. We have a Gaussian distribution

$$p(x) = \frac{\sqrt{2\Delta t/\pi}}{\sqrt{t}} \exp\left(-\frac{2\Delta t x^2}{t}\right), \tag{1.12}$$

we can simplify this equation by grouping all the constant terms into one only constant value

$$c = \sqrt{\frac{2\Delta t}{\pi}}, \tag{1.13}$$

by doing this we recover the usual expression of a normal Gaussian distribution

$$p(x) = \frac{c}{\sqrt{t}} \exp\left(-\frac{\pi c^2 x^2}{t}\right). \tag{1.14}$$

This method for solving the RW problem gives a Gaussian solution by construction. It is known that if the number of events n is large enough, that's assuming $n >> 1$, the skew of the Binomial distribution takes a small value, and a Gaussian distribution is a good approach to the final distribution. There are several rules of thumb for giving a minimum value of n, from which Binomial distribution can be taken as a Gaussian. This value of n is small and when $n > 20$ the goodness of fit shows reasonable values. But it is important to mention that the Gaussian distribution is not the only limiting distribution for a Binomial. If the constraints applied were $n \to \infty$, $p \to 0$, and with $n\alpha$ = constant, then the limiting distribution would be a Poisson distribution.

1.2.3 Empirical Returns Distributions

In his thesis, Bachelier tested his theoretical predictions against empirical data. For this, he employed a data series of the French government bond, and the futures corresponding to that underlying asset. This data set spanned five years, from 1894 to 1898, and contained 1452 data points. Bachelier calculated the drift and the standard deviation of empirical distributions. Surprisingly, he found a smaller number of large returns than he expected according to his theory. This is a striking result because one of the main criticisms to his theory [18] is exactly the opposite as it was shown in Chapter 0, but it was possible to reconcile theoretical and empirical results based on an argument about finite sample size. In addition to this, because of the low frequency of the data set it is plausible the idea that the tails of the distribution were not fully taken into consideration. It is known that the higher the sample frequency the more leptokurtic are empirical distributions [4].

Although Gaussian models produce consistent theoretical results, and these results are a good approach to emprirical data sampled at very low frequencies. It has been consistently shown [20] that the number of discrepancies with real data make necessary a different solution. This new solution should be based on a distribution with different properties. Gaussian models produce fewer extreme events than we observe in financial markets. Indeed, an extreme event as Black Monday, market crash happened on October 19th 1987 shouldn't be possible if this theorical framework were right [15].

1.2.4 Stochastic Processes in Finance

So far I have presented Bachelier's assumptions and reasonings leading to the solution for the probability distribution of asset prices. According to his model prices are drawn from a normal and their time evolution is unpredictable because it is probabilistic. Therefore his model on prices follows a probabilistic model called a stochastic process. Bachelier formulated indeed for the first time the Wiener process for describing asset price dynamics. As I said earlier, the aim of the first part of this thesis is the explanation of price dynamics and this must be expressed as a stochastic process. For this reason, in this section I give some basic definitions about these processes and present some of them broadly employed in mathematical finance.

A variable with unpredictable time evolution follows a stochastic process. Thus, this type of processes are employed to describe mathematically systems which evolve probabilistically in time. In these systems it exists a time-dependent variable $X(t)$, and we can measure its values $\{x_1, x_2, ..., x_n\}$ at times $\{t_1, t_2, ..., t_n\}$. It is assumed that a set of joint probability densities exists, this joint probability can be expressed as

$$p(x_1, t_1; ...; x_n, t_n), \tag{1.15}$$

where $t_1 \geq t_2 \geq ... \geq t_n$, and system can be described completely based on Eq.(1.15). Another important definition broadly used in mathematical finance is that of conditional probabilities densities, such as

$$p(x_1, t_1; ...; x_n, t_n | y_1, \tau_1; ...; y_n, \tau_n) = \frac{p(x_1, t_1; ...; x_n, t_n; y_1, \tau_1; ...; y_n, \tau_n)}{p(y_1, \tau_1; ...; y_n, \tau_n)}, \tag{1.16}$$

where $t_1 \geq t_2 \geq ... \geq t_n \geq \tau_1 \geq \tau_2 \geq ... \geq \tau_n$, time index is usually sorted incrementally. In an evolution equation, conditional probabilities are commonly taken as predictions of the future because current values are considered related to past values - to some extent - in a causal way which is not conclusively correct. For defining the stochastic process it is necessary to know at least all possible joint probabilities of the type decribed in Eq.(1.15). If this information is enough to define the process, it is called a separable stochastic process. The most simple of this kind is that of complete independence

$$p(x_1, t_1; ...; x_n, t_n) = \prod_{i=1}^{n} p(x_i, t_i). \tag{1.17}$$

This equation means that the value of X at time t is completely independent of its values in the past. Then, Bachelier's assumption about prices would hold if Eq.(1.17) fitted to price dynamics.

Several distinctions of stochastic processes can be made. One of them is based on the way the time variable t is measured. If time is considered a continuous variable then the stochastic variable is a continuous one, but if time is measured as a discrete variable the stochastic variable is a discrete one. Another distinction is based on the way the noise term is acting on the stochastic variable, which can be additive or multiplicative, but we need some more explanations for understanding the difference between these two types of stochatic processes.

For describing a stochastic process, it is necessary to specify its dynamics and the probability distribution function of the random variable. The dynamics is commonly given by a stochastic difference equation, such as

$$x(t+1) = x(t) + \varepsilon(t), \tag{1.18}$$

where $x(t)$ is the stochastic variable and $\varepsilon(t)$ is a random variable. The probability distribution of $\varepsilon(t)$ is necessary for the full description of the stochastic process, because several dynamics are possible depending on this distribution. A possible probability distribution function of $\varepsilon(t)$ might be

$$p(\varepsilon, t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} exp\left(-\frac{\varepsilon^2}{2\sigma^2 t}\right). \tag{1.19}$$

An alternative way to describe the stochastic process is by mean of a differential equations, such as

$$dx(t) = ax(t) + b\varepsilon(t), \tag{1.20}$$

$$dx(t) = ax(t) + bx(t)\varepsilon(t). \tag{1.21}$$

Given that in the Eq.(1.20) the random variable is added to the stochastic variable, it is describing an additive noise. Following with the same reasoning, Eq.(1.21) is describing a multiplicative noise. Another important aspect in the description of a stochastic process is the correlations. These may be expressed by an equation which shows the dependence of the time series, e.g. the ARCH models use autoregressive processes equations; or may be expressed by the conditional probabilities. In this last case, the current value of the variable is conditioned to its past values.

Markov Processes

Markov processes are extensively used in finance and economics, e.g. asset prices, market crashes, volatility [13]. Dynamic macroeconomics also uses markov processes, e.g. to exogenously model prices of equity in a general equilibrium setting [5].

The Markov assumption can be formulated in terms of conditional probabilities. For a Markov process the conditional probability is determined entirely by the knowledge of the most recent condition. This is because the defining property of this type of process is that it has no memory.

$$p(x_1,t_1;...;x_n,t_n|y_1,\tau_1;...;y_n,\tau_n) = p(x_1,t_1;...;x_n,t_n|y_1,\tau_1), \tag{1.22}$$

where $t_1 \geq t_2 \geq ... \geq t_n \geq \tau_1 \geq \tau_2 \geq ... \geq \tau_n$. This means that everything can be defined in terms of the simple conditional probabilities $p(x_1,t_1|y_1,\tau_1)$. By the definition of the conditional probability density

$$p(x_1,t_1;x_2,t_2|y_1,\tau_1) = p(x_1,t_1|x_2,t_2;y_1,\tau_1)p(x_2,t_2|y_1,\tau_1), \tag{1.23}$$

and using the Markov assumption (1.22), we have

$$p(x_1,t_1;x_2,t_2;y_1,\tau_1) = p(x_1,t_1|x_2,t_2)p(x_2,t_2|y_1,\tau_1), \tag{1.24}$$

therefore, any arbitrary joint probability can be expressed as

$$p(x_1,t_1;...;x_n,t_n) = \prod_{l=2}^{n} p(x_{l-1},t_{l-1}|x_l,t_l)p(x_n,t_n), \tag{1.25}$$

where $t_1 \geq t_2 \geq ... \geq t_n$.

Wiener Processes

Bachelier formulated the Wiener process for the first time, but it was Norbert Wiener who studied it extensively. Wiener process is a particular Markov process but in continuous time and with continuous variable.

Let's consider a stochastic variable z with two properties. The first one is that consecutive increments of variable z, Δz, are statistically independent. The second property is about making a distinction between the variation of z for finite time, and for an infitesimal time interval dt, which may be expressed as

$$\Delta z = \varepsilon\sqrt{\Delta t}, \tag{1.26}$$

$$dz = \varepsilon\sqrt{dt}, \tag{1.27}$$

where ε is distributed as a Gaussian normal distribution with zero mean and unit variance, $\mathcal{N}(0,1)$. Then, ε can be written as

$$p(\varepsilon) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{\varepsilon^2}{2}\right). \tag{1.28}$$

The main difference between a Markov process and a Wiener process is about the conditions the latter must fulfill. The Wiener process is more restrictive because random variables must be independent and identically distributed (iid). Therefore, the correlations of the random values taken by ε are

$$\langle \varepsilon(t)\varepsilon(t')\rangle = \sigma^2 \delta(t-t'), \tag{1.29}$$

where σ^2 is the variance of the normal distribution.

A Wiener process is characterized by two distributional properties, which meet two conditions of Bachelier's model. These conditions are about the mean value and the variance of the distribution. The expectation value of the stochastic variable takes value 0 for small time intervals

$$\langle \Delta z\rangle = \int_{-\infty}^{\infty} d(\Delta z)\Delta z p(\Delta z) = 0, \tag{1.30}$$

and the variance which grows linearly in the time interval ΔT

$$\mathrm{var}(\Delta z) = \int_{-\infty}^{\infty} d(\Delta z)(\Delta z)^2 p(\Delta z) = \Delta t, \tag{1.31}$$

when dividing time intervals of length T into smaller subintervals, every subinterval may also be considered a Wiener process of length one time step. Final result for a sum of random quantities drawn from a normal distribution is also normally distributed because of the properties of this particular stable distribution. Mean value and variance are additive, in this case; and their final values are

$$\langle z(T) - z(0)\rangle = 0, \tag{1.32}$$

$$\mathrm{var}\,[z(T) - z(0)] = \mathrm{T}. \tag{1.33}$$

In the modelling of asset prices it appears a drift term representing the mean value about which prices fluctuate. It is important to differentiate between a deterministic component, e.g. the accrued interest of coupon's payment, and the drift which is part of the random component. This drift term is added to the Wiener process by mean of a linear equation making possible the description of price variations with one only equation, such as

$$dx = \mathrm{a}dt + \mathrm{b}dz, \tag{1.34}$$

where a and b are constant terms. These terms are not necessarily constant, but they must be deterministic functions and would be represented as $a(x,t)$ and

$b(x,t)$. This new formulation is called a generalized Wiener process. We may calculate the mean and the variance of the new process, which are

$$\langle x(T) - x(0) \rangle = \mathrm{a}T, \tag{1.35}$$

$$\mathrm{var}\,[x(T) - x(0)] = \mathrm{b}^2 T. \tag{1.36}$$

These last two equations, Eq.(1.35) and (1.36), can be compared with those obtained for the Wiener process. The differences are caused by the drift term in both cases. A last remark about the drift term is that it can be eliminated by considering the equivalent martingale process as the fundamental variable. In this case we would have a Wiener process with zero mean and variance would grow linearly in time.

1.3 Standard Gaussian Model

In the previous sections I have presented a Gaussian model from first principles. We have seen the assumptions and the mathematical reasonings that let us deduce the specific shape of the prices distribution. Moreover, I have mentioned the stochastic process for price dynamics which gave support and explanation to that probability distribution of prices. From a mathematical point of view there was no important mistake, only the problem about considering one possible solution when in reality others were allowed. But Bachelier's model reached certain results which were in clear contradiction with basic economical principles. The first of these problems was the possibility of having negative prices. Given that the model considered prices were normally distributed about zero, negative prices were obviously allowed. The second problem was about returns. When an individual is making an investment decision, she is concerned about return on invested capital and will try to maximize the expected return. A result of the model was that returns were related to absolute changes in prices, instead of being related to relative changes. Therefore, the invidual investor avoids assets with high initial price because return is lower in comparison to low initial price.

For better understanding the problem of Bachelier's model with returns I will show an example. Let's take two assets, a_1 and a_2, with different initial prices, p_1 and p_2, at time t_0. If we have the same price variation for both assets, Δp, independently of the initial prices, and this variation happens in the same time interval for both assets. The final price for a_1 is $p_1 + \Delta p$, and $p_2 + \Delta p$ for a_2. Returns for a_1 and a_2 are

$$r(a_1) = 1 + \frac{\Delta p}{p_1}, \tag{1.37}$$

$$r(a_2) = 1 + \frac{\Delta p}{p_2}, \tag{1.38}$$

where $r(a_1)$ and $r(a_2)$ represent the returns for a_1 and a_2. Therefore, return on invested capital would depend on the initial price of the asset. This is in contradiction with economical reasoning because prices - nominal prices - are arbitrary quantities, the relevant magnitude for a traded stock is market capitalization.

In Bachelier's model, profit of an investment into a certain stock with price S in time T is

$$\langle S(T) - S(0) \rangle = \frac{dS}{dt} T, \tag{1.39}$$

where dS/dt is the drif term, and it is independent of S. This problem can be solved with the following corrected equation

$$dS = \mu S dt, \tag{1.40}$$

where μ represents the rate of return which is independent of the initial price of the asset. $\mu \Delta t$ is the return over a time interval Δt. The solution to Eq.(1.40) is

$$S(t) = S(0) e^{\mu t}, \tag{1.41}$$

where $S(0)$ is the price of asset S at initial time $t = 0$. The reasoning about the independence of returns and initial prices can be also applied to volatility and initial prices. Moreover, given that volatility can be taken as a measure of uncertainty about future returns. We could deduce that uncertainty about future returns is related to absolute price variations. This problem can be fixed by assuming the independence of these two quantities, such as

$$\sigma^2 \Delta t = \text{var}\left(\frac{\Delta S}{S}\right). \tag{1.42}$$

We can sort the terms of Eq.(1.42) for having an equation on the variance of asset price, expressed as

$$\text{var}(S) = \sigma^2 S^2 \Delta t. \tag{1.43}$$

All the corrections to Bachelier's model are fulfilled by an stochastic process: a generalized Wiener process. As this shown below

$$dS = \mu S dt + \sigma S dz, \tag{1.44}$$

and substituting $dz = \varepsilon \sqrt{dt}$, we have

$$\frac{dS}{S} = \mu dt + \sigma \varepsilon \sqrt{dt}. \tag{1.45}$$

This means that returns dS/S are pulled from a normal distribution with mean μdt and standard deviation $\sigma \sqrt{dt}$. Therefore, stock price S is described by stochastic process with a multiplicative noise. This new theoretical description fixes the mentioned problems in Bachelier's model.

1.3.1 Log-Normal Distributions of Stock Prices.

We have seen how the standard Gaussian model fixed the problems of Bachelier's model. The solutions were mainly based on considering returns as the subject to be modeled, instead of prices themselves. In doing so, problems related to absolute variations of prices disappeared because returns are relative price variations. Moreover, the distributional shape of the solution could be kept untouched because negative returns are possible and logical, but not negative prices. This new model is more a refinement of Bachelier's model than a new theoretical framework.

In the previous section we have shown a stochastic process, which described returns dynamics. This process met the requirements expressed by Bachelier, and avoided some economical problems. Here I show the probability distribution associated to price dynamics described by Eq.(1.45).

When applying the Itô lemma[3] to Equation(1.45) with $G(S,t) = \ln S(t)$, we have

$$dG = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma dz, \tag{1.47}$$

where $\ln S$ follows a generalized Wiener process with a drift term $\mu - \sigma^2/2$, and standard deviation σ. Given that probability distribution of $\ln S$ is a normal distribution. The mean and variance for a time interval of lenght T, and initial time $t = 0$ are

$$\langle \ln S \rangle = \left(\mu - \frac{\sigma^2}{2}\right)(T - t), \tag{1.48}$$

$$\mathrm{var}(\ln S) = \sigma^2 (T - t), \tag{1.49}$$

where T is a future time, and t is a past time refered to T. When comparing these last Equations, Eq.(1.48) and (1.49), with those obtained as solution to the generalized Wiener process for the Bachelier's model, Eq.(1.35) and (1.36), in reality, they only differ on the described subject: returns, and prices.

1.4 Conclusions

In this Chapter I have presented the basic principles and mathematical tools that are commonly employed in the different theoretical approaches to asset price dynamics. These theoretical elements were originally developed by Bachelier and presented in his dissertation. The first of these elements is a series of assumptions on prices and markets in order to avoid prices are predictible. The

[3]Let $x(t)$ follow an Itô process $dx = a(x,t)dt + b(x,t)\varepsilon\sqrt{dt}$. Then, a function G(x,t) is an Itô process such as

$$dG = \left(\frac{\partial G}{\partial x} a + \frac{\partial G}{\partial t} + \frac{1}{2} b^2 \frac{\partial^2 G}{\partial x^2}\right) dt + b \frac{\partial G}{\partial x} dz, \tag{1.46}$$

where $\left(\frac{\partial G}{\partial x} a + \frac{\partial G}{\partial t} + \frac{1}{2} b^2 \frac{\partial^2 G}{\partial x^2}\right)$ is the drift of the Itô process followed by G, and $b\frac{\partial G}{\partial x}$ is the standard deviation rate.

second element is about the random behavior of price dynamics and the necessity of stochastic processes to model the probability distribution of prices. These elements are taken as valid and assumed as correct for all models presented in this thesis.

Bachelier's model considered prices were normally distributed and were uncorrelated, but it was demonstrated the convenience of studying returns instead of prices for avoiding results in clear contradiction with economic principles such as negative prices. The BSM approach is the result of incorporating certain financial constraints to Gaussian model for making it more realistic from an economic perspective. This new theoretical approach was able to give a plausible explanation of price dynamics from a simple stochastic process: generalized Wiener process, and based on this proces it was posible to generate a probability distribution of prices which reconciled Gaussian models and economic principles. In spite of the theoretical validity of the solution, empirical findings on returns distributions showed that normal distributions are only a good approach to real dynamics of prices at very low frequencies. Moreover, models based on normal distributions failed to explain several stylized facts common to most of the financial assets, making necessary a different explanation for matching empirical observations. A final remark is that Bachelier made a mistake when deducing the shape of the probability distribution of prices and considered normal as the only right solution when others were also valid keeping unchanged all the assumptions and methods presented in his dissertation, so a new model for asset price dynamics is not necessarily against the first principles of Gaussian models.

Bibliography

[1] S. Basu, Journal of Finance **32**, 663 (1977).

[2] C. H. Berg, *Random walks in biology* (Princeton Unniversity Press, Princeton, NJ, 1983).

[3] F. Black, and M. Scholes, Journal of Political Economy **81**, 637 (1973).

[4] T. Bollerslev, and I. Domowitz, Journal of Finance, **48**, 1421 (1993).

[5] M. Brennan, and Y. Xia, Journal of Monetary Economics **47**, 249 (2001).

[6] P. H. Cootner (ed.), *The Random Character of Stock Market Prices* (MIT Press, Cambridge, MA, 1964).

[7] N. A. Chriss, *Black-Scholes and Beyond* (Irwin Professional Publishing, Chicago, 1997).

[8] A. Einstein, Ann. Phys. (Leipzig) **17**, 549 (1905).

[9] E. F. Fama, J. Business **38**, 34 (1965).

[10] E. F. Fama, Journal of Finance **25**, 383 (1970).

[11] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1968).

[12] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences* (Springer, Berlin, Heidelberg, 1998).

[13] J. D. Hamilton, Econometrica **57**, 357 (1989).

[14] J. C. Hull, *Options, Futures, and Other Derivatives* (Prentice Hall, Upper Saddle River, 1997).

[15] C. J. Jacklin, A. W. Kleidon, and P. Pfeiderer, Rev. Financ. Stud. **5**, 35 (1992).

[16] A. Jaeger, The Review of Economics and Statistics **74**, 607 (1992).

[17] B. G. Malkiel, *A Random Walk Down Wall Street* (W.W. Norton & Company, New York, NY, 2003).

[18] B. B. Mandelbrot, J. Business **36**, 394 (1963).

[19] F. Nicholson, Financial Analysts Journal **1**, 105 (1968).

[20] R. Cont, Quantitative Finance **1**, 223 (2001).

[21] B. Rosenberg, K. Reid, and R. Lanstein, Journal of Portfolio Management **13**, 9 (1985).

[22] P. Samuelson, Industrial Management Review **6**, 41 (1965).

[23] P. Wilmott, *Derivatives* (Wiley, Chichester, 1998).

Chapter 2

Non-Gaussian Models

2.1 Introduction

It has been shown that Gaussian models are a reasonable approach to modeling asset price dynamics at very low frequencies, but they fail in explaining the whole shape of empirical returns distribution because this is leptokurtic and Gaussian models systematically underestimate the tails, which represent the probability of large returns. Non-Gaussian models [5, 14] came up as an attempt to solve the problem with the tails, and reconcile theoretical models with empirical observations [6, 14]. Mandelbrot [14] developed the first non-Gaussian model and since that moment a vast collection of possible solutions has been produced by researchers.

In this thesis, I have classified non-Gaussian models into two families: Stable Distributions (SD), and Mixture Distribution Hypothesis (MDH). Although these two solutions make different assumptions, a common aspect is that both assume non-Gaussian distributions for describing the empirical data. In spite of the differences, it has been demostrated [2] that the SD framework can be derived as a particular case of the MDH being this last approach a more general solution to the problem. The SD is able to explain the occurrence of large events and the apparent stability of returns distribtuions up to a certain time interval. However, SD presents two serious problems: it does not converge to a Gaussian distribution for long time intervals, and it assumes variance of empirical distribution is infinite. As an attempt to solve these problems, it was proposed the Truncated Lévy Distribution (TLD) [15] which is a slightly modified version of the SD. The specific modification consists in truncating the tails of a pure Lévy distribution to let it converge to a Gaussian. In doing this, variance is well-defined and TLD slowly converges to a normal. The main criticism to TLD is that it needs an additional empirical parameter for defining the two regions: normal and Lévy.

MDH framework was developed by Clark [5] as an alternative theory that avoided distributions with infinite variance for explaining empirical data. MDH does not try to fit empirical returns with a different distribution, but it states that when returns are conditioned to another variable - Clark originally considered informational arrival rate was that variable - these become normally distributed. So MDH model is the result of a doubly stochastic process with a normal distribution for conditioned returns, and another distribution called directing process representing the variable to which returns are conditioned. MDH framework is not very restictive and let consider different variables such as volume, trading activity, and volatility for the directing process. For this thesis, the most relevant directing processes is the volatility because the new model presented in Chapter 3 may be considered within the Stochastic Volatility (SV) [23, 24] framework. SV models are good candidates for solving several stylized facts because MDH gives theoretical support to their explanation and the former is only a specific case of the latter. This is the main difference with SD and Gaussian models which were in theoretical contradiction with empirical findings. On the other hand, a problem with SV is that volatility is a latent variable so it is not directly observable, then it must be inferred as it happened with informational arrival rate in Clark's model. Given that returns distribution characteristics in SV models are driven by these of the volatility, returns are directly affected by the accuracy of volatility estimation and this depends on the specific definition of volatility and the available number of data points. Modern financial markets are usually electronic, so a huge amount of data is recorded every trading day reducing dramatically the problems related to sample size.

2.2 Non-Gaussian Models

Empirical distributions are not Gaussian, they are leptokurtic. Therefore, the probability of very large returns, also called extreme events, must be fitted by a heavy-tailed distribution. Another important characteristic of the empirical distributions is that they do not scale as it would be expected from a Gaussian. Emprical distributions slowly converge to a Gaussian when returns time intervals increase. Due to that slow rate of convergence these distributions are apparently stable, but they are not. These differences are relevant and made necessary a new theoretical approach.

The first non-Gaussian model was published by Madelbrot in 1963 [14]. He realized, when studying the prices of cotton, that the cumulative distribution function (cdf) of returns did not fit the expected result based on a Gaussian model. Mandelbrot plotted the cdf on double logarithmic scale, and he found its tail followed a straight line. This was in good agreement with a stable Lévy distribution, and it was taken as a conclusive proof against Gaussian models. The explanation to his finding, according to the assumption of a Lévy distribution, is that Lévy distributions asymptotically decay with power laws of their variables. Thus, when plotting a power law in double log scale, also called log-log scale, gives as result a straight line. This finding was also supported by Fama in 1965 [6] in his investigations on analyzing prices in the New York Stock Exchange. Since that moment, Gaussian models were abandoned as a plausible explanation and new theoretical approaches were all non-Gaussian.

2.2.1 Stable Distributions Models

Gaussian models were developed from first principles taking into account only theoretical constraints about an idealistic market and statistical properties of consecutive prices. However non-Gaussian models in general, and more specifically SD Models, are developed as an attempt to explain empirical findings, and since that moment theoretical models are generated for explaining the stylized facts of empirical distributions. The empirical findings addressed by Mandelbrot were the heavy tails, and the scalability of empirical data. The challenging solution proposed by him to the problem of non-gaussianity was a Lévy dristribution. This let solve the problem of heavy tails, because Lévy distribution has much more weight in its tails than Gaussian as it is illustrated in Fig.(2.1). Moreover, as a Lévy is a stable distribution[1] the apparent stability of empirical distributions was also explained. However, this solution brought a new theoretical problem: its second moment is not well-defined, this means that variance is non finite. Moreover, the stochastic process generating the distribution of returns - Lévy flight - is different from the geometric Brownian motion because in the Lévy flight the length of the random steps is drawn from a heavy-tailed distribution and the distance from the origin tends to a stable distribution.

A stable Lévy distribution is defined by its characteristic function

$$\hat{L}_{a,\beta,m,\mu}(z) = \exp\left\{-a|z|^{\mu}\left[1 + i\beta sign(t)\tan\left(\frac{\pi\mu}{2}\right)\right] + imz\right\}, \tag{2.1}$$

where β is a skewness parameter, $\beta = 0$ means a symmetric distribution. μ is the index of the distribution which gives the exponent of the asymptotic power-law tail. a is a scale factor which characterizes the width of the distribution, and m gives the peak's location.

Taking into consideration Bachelier's assumptions, we need a symmetric function. Then, $\beta = 0$ is the necessary value for this parameter. We also know that the maximum of the distribution must be located at $x = 0$, leading to $m = 0$. Then, the characteristic function may be simplified and expressed as

$$\hat{L}_{\mu}(z) = \exp\left(-a|z|^{\mu}\right). \tag{2.2}$$

Depending on the specific value taken by μ in Eq.(2.2) we have different distributions, e.g. Gaussian for $\mu = 2$, and Lorentz-Cauchy for $\mu = 1$. Given that μ measures the exponent of the tails, its value indicates the higher well-defined moment, e.g. for $\mu < 2$ the variance is infinite and the mean absolute value is

[1] A distribution is classified as stable distribution if it has the property that a linear combination of two independent copies of the distribution gives as result the same distribution, up to location and scale parameters.

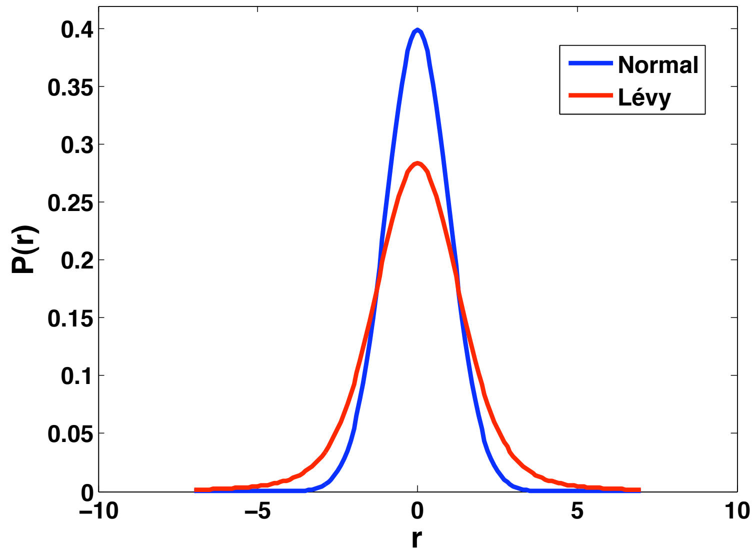

Figure 2.1: **Probability density function of a Lévy distribution compared to a normal distribution.**

finite if $\mu > 1$. In general, the moments with orders $n < \mu$ are finite. About the stability, only Lévy distributions with $\mu \leq 2$ are stable meaning that all of them act as attractors, whereas for $\mu > 2$ are not stable. An important property of the stable distributions is that they keep unchanged their shape under aggregation up to rescaling, so this type of distribution is defined as self-similar. This is an important issue for describing empirical distributions because stable distributions do not tend to a Gaussian when they are aggregated except normal itself. Aggregational gaussianity therefore can not be explained by stable distributions making this solution inadequate for explaining empirical distributions. Although CLT may not be applied to Lévy distributions, there is a generalized version due to Gnedenko and Kolmogorov [8] that may be employed. This states that if many independent random variables are added whose probability distributions have power-law tails $p_i(x_i) \sim |x_i|^{-(1+\mu)}$, with an index $0 < \mu < 2$, their sum will be distributed according to a stable Lévy distribution $L_\mu(x)$.

There are several references on the good agreement of empirical distributions with Lévy processes, e.g. Fama [6] studied single stocks listed in the Dow Jones Industrial Average (DJIA) in the 1960s and he found that Mandelbrot's hypothesis was supported by his empirical research, Mantegna and Stanley [16] studied the S&P500 index and were able to roughly collapse the distributions onto a master curve that had an index $\mu = 1.4$.

In Fig.(2.2) I show a comparison of the probability density function (pdf) of the standardized empirical returns for AZN with a fitted Lévy and a normal with zero mean and unit variance. We see that Lévy distribution fits better than normal the body of the empirical distribution, but both of them fail to fit the

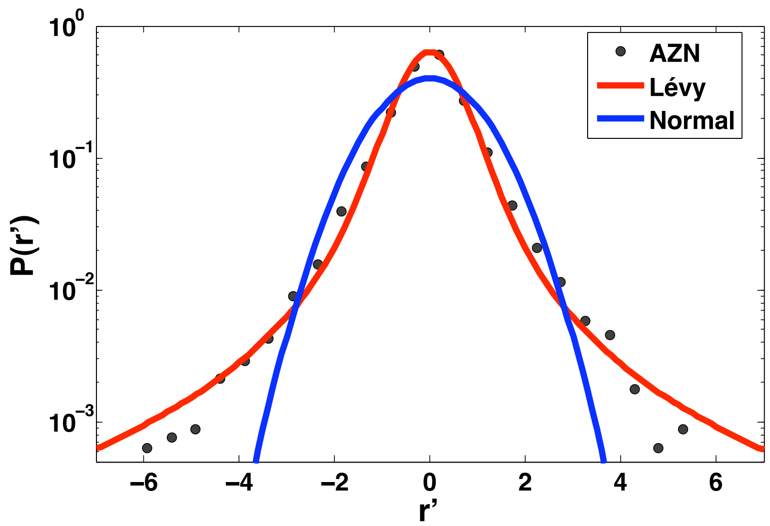

Figure 2.2: **Probability density function of standardized returns**, $P(r')$, **for AZN compared to a fitted Lévy distribution and a normal.** The solid blue line is the pdf for a normal distribution with mean zero and unit variance.

tails. Lévy clearly overestimates the occurrence of large returns, whereas normal underestimates them.

The behavior of Lévy distributions for large values of x is

$$L_\mu(x) \sim \frac{\mu A_\pm^\mu}{|x|^{1+\mu}} \text{ for } x \to \infty, \tag{2.3}$$

where $0 < \mu < 2$ is an exponent, sometimes called α, and $A_\pm^\mu$ are two constants called tail amplitudes because they give the order of magnitude of the positive and negative large fluctuations of x. Lévy distributions have been tested against empirical data and they systematically overestimate [3] the tails of empirical distributions, so SD model as it was formulated by Mandelbrot does not fit financial data in spite of being heavy tailed. As an attempt to solve the overestimation of the tails within the SD framework, it was introduced the Truncated Lévy distribution (TLD) [15] which is a way to reduce the weight in the tails of stable distributions. The main advantage of the TLD is that it has finite variance and it slowly converges to a Gaussian distribution. Moreover, financial models related to mean-variance paradigm would continue to be meaningful. A truncated Lévy distribution (TLD) is defined by its characteristic function

$$\hat{L}_\mu^{(t)}(z) = \exp\left[-a_\mu \frac{(\alpha^2+z^2)^{\frac{\mu}{2}}\cos(\mu \arctan(|z|/\alpha)) - \alpha^\mu}{\cos(\pi\mu/2)}\right], \tag{2.4}$$

where α is the parameter used for delimiting the regions with different behavior. The main difference between TLD and a pure Lévy distribution is shown when they are added. To make it more explicit let's consider the sum of TLD distributions. Let X be the sum of N random variables distributed as TLD distribution. Considering that a TLD behaves like a Lévy distribution for $x << \alpha^{-1}$. More specifically, it behaves as a power-law of exponent μ and tail amplitude A^{μ}. If N is not too large, most values of x are in the Lévy region. However when x reaches the cutoff value α^{-1} and if N is a large number X converges to a Gaussian. The region of the tails - outside central region - of the sum decays as an exponential. This solution is a crossover between Gaussian and Lévy behaviors strongly depending on the value of α. An evident criticism to both models is that they only try to fit the returns distribution without any further theoretical explanation. This is more obvious in the TLD model where an additional parameter is introduced to match empirical distributions due to the lack of accuracy in the tails.

2.2.2 Mixture Distribution Hypothesis Models

In SD models the price variations are independent and their distribution does not have finite variance, then all higher moments of the distribution are not well defined. The acceptance of these properties made necessary to abandon well established results in finance, e.g. those derived from the mean-variance framework. Clark[5] developed a new theory which was compatible with other models in finance where the existence of a well defined variance was a crucial element.

The basic assumption of Clark's theory was that the distribution of price variations was subordinated to a Gaussian. This assumption made necessary the existence of another distribution causing the observed heavy tails of empirical data. He postulated that the different number of events in a certain fixed calendar time interval was that cause. So prices evolved at different rates at this fixed time interval as a consequence of the different number of events. Clark assumed that the origin of the different number of events was differences on information arrival rate at different times, on days when there were more news trading activity was high and prices evolved fast, and the opposite, on the days when there were not many news prices evolved slowly. However this assumption brings a new problem: information was not strictly defined, and different models within this framework try to derive it by measuring another observable magnitude.

A common feature to SD and MDH frameworks is that CLT can not be applied to any of them. In Mandelbrot's model because individual transactions are drawn from a distribution with infinite variance. In Clark's model because it assumes different rates of change in prices for equal calendar time intervals.

Subordinated Stochastic Processes

Subordination is the key element in MDH models, and Clark related subordination to information arrival. This is an inconvenient for model's implementation because information arrival is taken as a latent variable which must be inferred, but from a mathematical point of view a formal definition of subordination is possible, and some results theoretically derived from that formal definition are necessary for understanding further models developed within this framework.

Let's consider a stochastic discrete variable: $X(0), X(1), ..., X(t)$, where $X(i)$ is the particular realization of a stochastic process at time i. And, let's consider the indexes of this stochastic process realizations of another stochastic process where ($t_1 \leq t_2 \leq t_3 \leq ...$). That is, $T(t)$ is a positive stochastic process, a new process $X(T(t))$ can be generated. This new process $X(T(t))$ is said to be subordinated to $X(t)$; and $T(t)$ is called the directing process. Finally, the distribution of $\Delta X(T(t))$ is said to be subordinated to the distribution of $\Delta X(t)$. This last term assumes the role of the individual effects in the evolution of the price process, while $T(t)$ is a clock measuring the rate of evolution. Although, change of time is common in subordination it can be found in the literature on Lévy processes too [1, 4]. There are two different strategies for modelling $T(t)$, the first strategy is more related to find a process with distributional properties compatible with those observed in empirical data, the second strategy is related to find a variable with a similar behavior to that of the information as a mean for capturing information.

In [5] we find several theorems related to subordinated processes. For the aim of this thesis, the next theorem which holds for general classes of subordinated stochastic processes with independent increments is the most relevant.

Theorem: Let X(t) and T(t) be processes with stationary independent increments; that is,

- $X(t_{k+1}) - X(t_k)$ (k=1,2,...,n-1) are mutually independent for any finite set $t_1 \leq t_2 \leq ... \leq t_n$, and similarly for $T(t)$;
- X(s+t)-X(s) depends on t but not on s for all s, and similarly for $T(t)$.

Let the increments of $X(t)$ be drawn from a a distribution with mean 0 and finite variance σ^2. And let the increments of $T(t)$ be drawn from a positive distribution with mean α, independent of the increments of $X(t)$. Then the subordinated stochastic process $X(T(t))$ has stationary independent increments with mean 0 and variance $\alpha\sigma^2$. Thus, if the steps $\Delta X(t)$ are independent with mean 0 and variance σ^2, then ν steps have mean 0 and variance $\nu\sigma^2$. Therefore, the variance of $\Delta X(T(t))$ conditional on $\Delta T(t)$ is

$$\mathrm{var}(\Delta X(T(t))|\Delta T(t) = \nu) = \nu\sigma^2. \tag{2.5}$$

It is important to realize that if the directing process has a finite mean, then $\Delta X(T(t))$ will have a finite variance unless $\Delta X(t)$ does not have it. This implies that given the parameters α, and σ^2, we can change the distribution of $\Delta(t)$, and a family of distributions with mean zero and same variance can be obtained.

It is also demonstrated in [5] that the introduction of any directing process makes the distribution of $X(T(t))$ only more leptokurtic. Thus, it turns crucial for the MDH the selection of the directing process with the only condition that it must be an increasing process, this a very little restrictive condition and offers many possible solutions compatible with it. A common solution is to employ a log-normal distribution, that's Clark's solution. This choice creates an implementation's problem because for calculating the pdf of the returns it is necessary to compute a numerical integration that turns out to be relatively unstable.

Clark in his seminal paper not only developed a new theoretical framework. He also compared the lognormal-normal distribution with the stable distribution. For this comparaison, he studied the time series of prices of cotton futures for two periods: 1947-50 and 1951-55. The data series used by Clark makes it difficult to compare models by several reasons. The first reason is because cotton contracts do not have four year lives, then Clark had to splice series across contract lives. Consequently, additional noise may have been added to his data set. Moreover, the time period from 1951 to 1955 was preceded by a suspension of trading due to existing price controls, this could affect the amount of variation in the time series at the beginning of the period, just as trading began, and finally the open interest[2] is not constant. Volume and price fluctuations may be influenced by changes in open interest.

Directing Processes

MDH models can be classiffied employing the directing process as criterion. It has been mentioned that Clark assumed this process related to information arrival, but this is only a postulate. There are several processes compatible with the subordination model.

Informational arrival rate is an unobservable variable, but information affects observable variable as for example volume, or trading activity. This is a very general statement that can be broadly accepted, but the problem is how information modifies the obsevable variable and the returns. It may be postulated that information and the observable variable are driven in a similar way, or we can employ the observable variable as a proxy of the information arrival rate. Following this last approach, many possible proxies [12] have been studied, e.g. the volume measured as the number of shares, the volumen measured in money, the number of transactions, the average trade size... Another strategy to address the problem is to measure the volatity of the returns, and study its changes over time as a different stochastic process. This method gives as result the family of solutions called Stochastic Volatility (SV) models [20]. This family of models is the most relevant for this thesis because the new model developed in Chapter 3 may be classiffied as a member of it.

Many different directing processes hold Clark's assumptions, and all of them

[2]Open interest is the total number of derivative contracts that have not been settled in the immediately previous time period for a specific underlying security.

share a common mathematical behavior that may expressed such as

$$f(r_t) = \int_{I_t \in R^+} f(r_t | I_t) g(I_t) dI_t, \tag{2.6}$$

where $f(r_t)$ represents the density of returns, the distribution of increments of the directing process I_t follows a distribution with density $g(I_t)$, and $f(r_t|I_t)$ means that conditional on information arrival flow I_t, the density of returns is assumed to be given by the normal distribution. Then, it turns clear that any different behavior of $f(r_t)$ is a consequence of the distributional properties of $g(I_t)$, given that $f(r_t|I_t)$ is always represented by a Gaussian. Now, it becomes obvious that independently of the theoretical reasoning behind the choice of a certain variable as a proxy of the information. It is its shape, that's how it is distributed, the driving factor for several stylized facts, e.g. the heavy tails of the returns distribution.

Clark in his model assumed that I_t was log-normallly distributed, so

$$log(I_t) \sim \mathcal{N}(\mu, m_2), \tag{2.7}$$

where, $\mathcal{N}(\mu, m_2)$ represents a common normal distribution with mean μ and variance m_2. Integrating, as it is shown in Eq.(2.6). We obtain the solution to his model, such as

$$f(r_t) = \int_{I_t \in R^+} \frac{1}{\sqrt{2\pi\sigma_r^2 I_t}} exp\left(-\frac{1}{2}\frac{\left(r_t - \mu_r I_t\right)^2}{\sigma_r^2 I_t}\right) \frac{1}{\sqrt{2\pi m_2} I_t} exp\left(-\frac{1}{2}\frac{\left(log(I_t) - \mu\right)^2}{m_2}\right) dI_t, \tag{2.8}$$

where μ_r and σ_r^2 are the parameters of the normal distribution conditional on information arrival flow.

We can find several different distributions for the information arrival in the literature. Log-normal was not only postulated by Clark[5], but also by Tauchen and Pitts [21], and by Foster and Viswanathan [7]. Richardson and Smith [18] considered in addition a uniform arrival rate, a Poisson distribution, and the inverted-gamma distribution. Blattberg and Gonedes [2] used the inverse gamma too. Madan and Seneta [13] proposed the gamma distribution, generating the called variance-gamma model.

So far, I have shown different attempts of modelling information, and how the probability distribution associated to the information affected the returns distribution, but all these results are theoretical results and much reasearch has focused on testing the MDH models. Tauchen and Pitts [21] provided an economic model, yielding testable implications concerning returns and volume jointly. Harris [9, 10] extended the predictions of that model by considering a model where price increments, r_t, and volume, v_t, of a given day, are conditionally normal. Also, conditionally on the information arrival, I_t, the covariance between returns and volume is equal to zero. Harris model can be mathematically written, such as

$$r_t \sim \mathcal{N}(m_r I_t, \sigma_r^2 I_t), \tag{2.9}$$
$$v_t \sim \mathcal{N}(m_v I_t, \sigma_v^2 I_t), \tag{2.10}$$
$$Cov[r_t, v_t | I_t] = 0. \tag{2.11}$$

The first equation corresponds to Clark's initial model. If we consider a restriction of the bivariate model, we obtain Tauchen and Pitts [21]. Harris [10] derives a large set of conditional and unconditional moments that he uses to extend the set of stylized facts concerning asset returns and volume, already obtained by Tauchen and Pitts [21]. Richardson and Smith [18] further built on these conditional and unconditional moments to construct a formal test of the MDH based on the method of moments.

Student's t-Model

In the previous section, I have shown that for MDH models the classifying factor is the directing process, $T(t)$. This is the explanatory variable for the properties of the different models within the MDH framework, because all the other elements are common to compute returns distributions. I have also mentioned that the selection of the directing process may be based on different theoretical assumptions: a proxy of the volatility, a model of the volatility, traded volume... but from a mathematical perspective the only important characteristic is the shape of its distribution.

The original model by Blattberg and Gonedes [2] is very relevant for this thesis because it gives a theoretical foundation to the Student's t-distribution for the returns, and that's the shape empirically demonstrated in Chapter 3. They state that if $[T(t)]^{-1}]$ follows a Gamma-2 distribution[3], which is asymmetric and strictly positive, $X[T(t)]$ will follow a Student's t-distribution. On the other hand, they also state that if $T(t)$ follows a strictly positive asymmetric stable distribution with[4] $\alpha \in (0,1)$ then $X[T(t)]; t \geq 0$ will follow a symmetric-stable distribution with $\alpha < 2$. Based on this result we can state that Stable Distributions model is a particular case of the Mixture Distributions Hypothesis, demonstrating that MDH is a more general approach than SD.

The Student's t density function with location parameter m, scale parameter $H > 0$, and degrees of freedom parameter, $v > 0$, such as

$$f(x|m,H,v) = \frac{v^{(1/2)v}}{B\left(\frac{1}{2},\frac{1}{2}v\right)} \left[v + H(x-m)^2\right]^{-1/2(v+1)} \sqrt{H}, \tag{2.12}$$

where $B(.,.)$ is the beta function, that is, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, and $\Gamma(.)$ is the gamma function. The Student's t function has the following properties:

[3]Let $X > 0$ be a real random variable with a Gamma distribution parametrized by a shape $\frac{v}{2} > 0$ and scale $\frac{s}{2} > 0$. We will denote $X \sim G_2(v,s) \equiv G\left(\frac{v}{2},\frac{s}{2}\right)$ and say that X has a Gamma-2 distribution

[4]α is the exponent of the characteristic function of the symmetric-stable distribution.

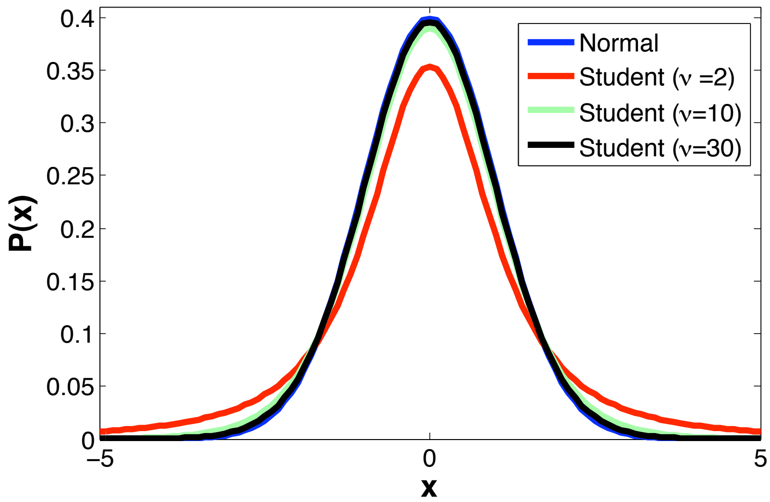

Figure 2.3: **Probability density function of Student's t-distribution for different values of ν and a normal.** The solid blue line is the pdf for a normal distribution with mean zero and unit variance, Student's t-distribution is plotted for ν values from 2 to 30.

(i)$\langle x\rangle = m$, for $\nu > 1$ this means that mean is defined for $\nu > 1$, (ii) var(x) = $H^{-1}\nu/(\nu-2)$, for $\nu > 2$ this means that variance is defined for $\nu > 2$; (iii) in general, all the moments of order $r < \nu$ are finite; (iv) when $\nu = 1$, the Student's density function is the Cauchy density function, (v) As $\nu \to \infty$, the Student's t-distribution converges to the normal disrtibution as it is shown in Fig.(2.3). This property of the Student's t-distribution is essential for understanding the aggregational gaussianity.

When a Student's random variable with $\nu > 2$, x, is standardized by taking normalized returns such as

$$x^* = \frac{x - \langle x\rangle}{\sqrt{\mathrm{var(x)}}}, \tag{2.13}$$

then the density function of x^* has the following properties: (i) it has fatter tails than the density function of a Gaussian distribution with mean zero and variance equal to one, (ii) the peak about zero, the mean value, is higher than in the Gaussian distribution about the mean. If we assume Student's t-distribution is standardized, in that case $m = 0$ and $H = 1$, and we have

$$f(x|m,H,\nu) = \frac{\nu^{(1/2)\nu}}{B\left(\frac{1}{2},\frac{1}{2}\nu\right)}\left[\nu + x^2\right]^{-1/2(\nu+1)}, \tag{2.14}$$

Student's t-distribution has two properties which make it especially interesting for describing empirical returns. The first property is that it is heavy-tailed with more weight in its tails than Gaussian. The second property is its convergence to

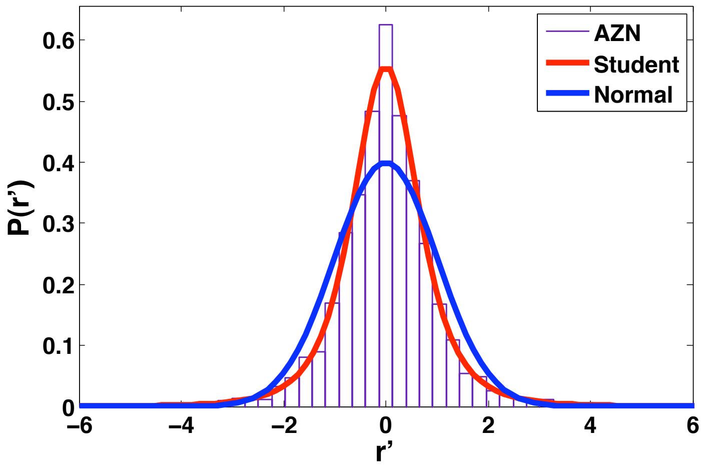

Figure 2.4: **Probability density function of standardized returns for AZN stock fitted by a Student's and a normal distribution.**

a Gaussian without the necessity of adding extra parameters, as it was the case for TLD.

As I said in Chapter 1, the aggregational gaussianity can be observed by reducing the sample frequency of the prices. We can rephrase this last statement, when we agregate longer intervals returns their distribution comes closer to a Gaussian. Let's consider an example, let $x_t, t = 1, 2, ...,$ the rate of return - continuous compounding - of a certain asset for day t, $t = 1, 2, ...,$ its rate of returns is

$$S^T = \sum_{t=1}^{T} x_t, \tag{2.15}$$

under continuous compunding, S^T, is the rate of return over a period of T days. Now, suppose that $(x_1, x_2, ..., x_T)$ is a sequence of independent random variables. Under the Student's t-model with $\nu > 2$, the distribution of S^T converges to a normal distribution as $T \to \infty$. This convergence result is a consequence of the Central Limit Theorem because Student's t-distribution is heavier tailed than a Gaussian but with finite variance. Student's t-model for daily returns implies that there is a time period such that rates of return may be described by a Gaussian distribution, and this is in good agreement with empirical returns. If we had a stable distribution with $\alpha < 2$, then the distribution of S^T wouldn't converge to a normal distribution because they are stable and tend to a distribution with the same shape.

In Fig.(2.4), I show the pdf of empirical returns for AZN fitted by a Student's and a normal distribution. We can see that Student's t-distribution does estimate better the peak of empirical distribution and it also shows a more leptokurtic behavior with a slower decay in the tails than normal.

2.3 Stochastic Volatility Models

MDH models rely their explanatory capability on finding a variable which is distributed in such a way that when is marginalised with a Gaussian reproduces the empirical returns distribution, that variable is not strongly constrained by the definition of the MDH framework. We have also seen that many attempts have been made to find theoretical models more fitted to empirical data, these attempts have taken two different directions: to find new variables, and to redefine previously employed variables. SV models [20, 23] as a family of solutions can be included into the first direction because it uses volatility as a directing process. This family assumes that volatility changes through time in a random way, so it may be modeled as a random process. Moreover, SV family has produced a collection of different particular models depending on the distributional properties of the volatility process.

Given that SV models follow MDH and the volatility is the only variable to be estimated, we might wrongly conclude that SV models should reduce to one only model, but this is not true because of the nature of volatility. Volatility can not be observed directly from returns time series, it is a latent variable, so it must be estimated based on past and current returns. As it happens with any other estimation the first question is about the method employed for it. The second question is about the available time series and the frequency at which it has been recorded. A final remark is that there are financial instruments: VIX, VDAX, volatility swaps, which have the volatility as underlying asset. Although these financial products are traded in financial markets, volatility as a variable itself continues to be a latent variable and must be modeled through its influence in the magnitude of returns.

SV models had its origin in the work of several authors, e.g. Rosenberg [19], Clark [5], Taylor [22], and Tauchen and Pitts[21] and have slowly grown because their difficult estimation. This is also the cause of the broad use of ARCH models[5] in financial industry.

All SV models share some basic assumptions as market efficiency, and the ones derived from MDH. In addition to these, there are some other features common to most of the models in the family. These features are summarized in the next example. Let's consider a model which assumes the volatility on a given time interval t - for the most general form of SV model we do not consider any specific length for the time interval - denoted by σ_t, which is partially

[5]An ARCH model in its most general form makes conditional variance at time t a function of exogenous and lagged endogenous variables, time, parameters, and past residuals

$$\begin{aligned} e_t &= \sigma_t Z_t, & (2.16)\\ \sigma_t^2 &= \sigma^2(\sigma_t^2 - 1Z_t - 1, \sigma_t^2 - 2Z_t - 2, ..., x_t, t, b), & (2.17) \end{aligned}$$

where $Z_t \sim iid$ with $\langle Z_t \rangle = 0$ and $\mathrm{var}(Z_t) = 1$, e_t are prediction errors, b a vector of parameters, x_t a vector of exogenous and lagged endogenous variables, and σ_t^2 the variance of e_t given information at time t.

determined by unpredictible events on the same time interval. Following with Clark's theory, volatility would be related to a stochastic number of price revisions in that interval. The total number of news items on the period t would be represented by a random variable, denoted by N_t. When news item i reaches the market , the logarithm of the price changes in a certain amount, ε, and these changes are added, such as

$$r_t = \sum_{i=1}^{N_t} \varepsilon_{i,t}, \tag{2.18}$$

where $\varepsilon_{i,t}$ are random variables normally distributed with mean 0 and variance θ^2, and independent of N_t. Then, the distribution of the return conditional on n_t news items is a normal with variance

$$\mathrm{var}(r_t|N_t = n_t) = n_t\theta^2. \tag{2.19}$$

Then, the stochastic volatility process can be described such as

$$\sigma_t^2 = N\theta^2, \tag{2.20}$$

and returns are

$$r_t = \sigma_t u_t, \tag{2.21}$$

where u_t is a standard normal random variable that is independent of the random variable σ_t. From Eq.(2.19) it can be seen that volatility changes when the amount of news N_t changes. Volatility clustering will then occur if there is sufficient positive autocorrelation. If we observe that positive autocorrelation periods of high volatility will be followed by more high volatility periods, and the other way low volatility periods will be followed by low volatility periods.

So far, we have shown a general outline common to most of the different implementations of the SV family. A long list of specific implementations can be cited. Tauchen and Pitts [21] assumed expected trading volume is proportional to the number of news items and hence volatility and volume are positively correlated variables. Harris [10] considers empirical transaction counts.

SV models are relevant from a theoretical point of view, because based on the general features presented above and assuming that volatility is autocorrelated - this has been empirically observed [17] and it is generally accepted, - three important stylized facts of the returns distribution can be explained. Let's suppose daily returns described by the equation

$$r_t = \mu + \sigma_t u_t, \tag{2.22}$$

this equation is equivalent to Eq.(2.21), but we have added explicitly the term μ for gaining in generality. For explaining those three stylized facts, we need six more assumptions: (i) the expected return μ is constant , (ii) σ_t is a positive random variable, (iii) the stochastic process $\{\sigma_t\}$ is stationary[6], $\langle\sigma^4\rangle$ is finite and

[6] A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time. As a result, parameters such as the mean and variance, if they exist, also do not change over time.

all the autocorrelations of $\{\sigma_t^2\}$ are positive, (iv) u_t is a standard random variable, so $u_t \sim \mathcal{N}(0,1)$, (v) the u_t are iid variables, (vi) the process $\{\sigma_t\}$ and $\{u_t\}$ are stochastically independent.

The first explained stylized fact is that the distribution of returns is not normal. Based on Eq.(2.21) the distribution of returns is a mixture of normal distributions, with the mixture determined by the distribution of volatility. This mixture distribution has higher kurtosis than that of a normal distribution, since

$$
\begin{aligned}
\mathrm{var}(r_t) &= \langle (r_t-\mu)^2\rangle = \langle \sigma_t^2 u_t^2\rangle = \langle \sigma_t^2\rangle \langle u_t^2\rangle = \langle \sigma_t^2\rangle, && (2.23)\\
\langle (r_t-\mu)^4\rangle &= \langle \sigma_t^4 u_t^4\rangle = \langle \sigma_t^4\rangle \langle u_t^4\rangle = 3\langle \sigma_t^4\rangle, && (2.24)\\
\mathrm{kurtosis}(r_t) &= \frac{3\langle \sigma_t^4\rangle \langle u_t^4\rangle}{\langle \sigma_t^2\rangle \langle u_t^2\rangle} = 3\left(1+\frac{\mathrm{var}(\sigma_t^2)}{\langle \sigma_t^2\rangle^2}\right) > 3. && (2.25)
\end{aligned}
$$

The second explained stylized fact is that returns are almost uncorrelated. The autocorrelations are zero at all positive lags τ when the assumptions apply, because

$$
\begin{aligned}
\mathrm{cov}(r_t, r_{t+\tau}) &= \mathrm{cov}(\sigma_t u_t, \sigma_{t+\tau} u_{t+\tau}),\\
&= \langle \sigma_t u_t \sigma_{t+\tau} u_{t+\tau}\rangle - \langle \sigma_t u_t\rangle \langle \sigma_{t+\tau} u_{t+\tau}\rangle,\\
&= \langle \sigma_t \sigma_{t+\tau}\rangle \langle u_t\rangle \langle u_{t+\tau}\rangle - \langle \sigma_t\rangle \langle \sigma_{t+\tau}\rangle \langle u_t\rangle \langle u_{t+\tau}\rangle = 0.
\end{aligned}
$$

The third explained stylized fact is that both absolute returns and squared returns are positively autocorrelated. Let $s_t = \left(r_t-\mu\right)^2$. Then, for all positive lags τ,

$$
\begin{aligned}
\mathrm{cov}(s_t, s_{t+\tau}) &= \mathrm{cov}\left(\sigma_t^2 u_t^2, \sigma_{t+\tau}^2 u_{t+\tau}^2\right),\\
&= \langle \sigma_t^2 u_t^2 \sigma_{t+\tau}^2 u_{t+\tau}^2\rangle - \langle \sigma_t^2 u_t^2\rangle \langle \sigma_{t+\tau}^2 u_{t+\tau}^2\rangle,\\
&= \langle \sigma_t^2 \sigma_{t+\tau}^2\rangle \langle u_t^2\rangle \langle u_{t+\tau}^2\rangle - \langle \sigma_t^2\rangle \langle \sigma_{t+\tau}^2\rangle \langle u_t^2\rangle \langle u_{t+\tau}^2\rangle,\\
&= \mathrm{cov}\left(\sigma_t^2 \sigma_{t+\tau}^2\right) > 0.
\end{aligned}
$$

Consequently, positive dependence in the volatility process implies positive dependence in squared excess returns. This result can be extended to positive dependence in absolute excess returns, $a_t = |r_t-\mu|$.

The six assumptions held by Eq.(2.22) are enough to provide a framework within which volatility changes explain the three above mentioned stylized facts for returns. At this point, we have not taken into consideration any specific process for volatility, but any process holding the assumptions before mentioned is valid for explaining those stylized facts. A last remark is that the six assumptions are not necessarily met by all SV models, i.e. there are certain implementations where σ_t and u_t are not statistically independent [11].

Although SV and ARCH models are often taken as equivalent with the only difference that the former is described in continuous time and the latter in discrete time, this is not absolutely corrrect and it is convenient to clarify the subtle but important difference between them. A general ARCH model for the excess return can be expressed such as

$$e_t = h_t^{1/2} z_t, \tag{2.26}$$

where e_t is the excess return over a certain mean value μ. $z_t \sim iid$ with zero mean and unit variance, and the conditional variance h_t is formulated as

$$h_t = \omega + \alpha (r_{t-1} - \mu)^2, \tag{2.27}$$

where the volatility parameters $\omega > 0$ and $\alpha \geq 0$. The volatility of the returns in period t then depends solely on the previous return. From here it is not inmediately deduced that ARCH models are equivalent to SV models, by only substituting $\sigma_t^2 = h_t$ because there is no unpredictible volatility component in h_t since var $(h_t|I_{t-1}) = 0$, where I_{t-1} means the information up to time $t-1$. On the other hand, in a SV model as it is shown in Eq.(2.29) conditional variance σ_t depends on an additional noise process η_t and so is itself an unobservable variable. This makes SV models much more difficult to estimate than ARCH models as the likelihood function cannot be written down directly. However, when z_t is a mixture of normal distributions, we can write $z_t = m_t^{1/2} u_t$ with $\langle m_t \rangle = 1, \mathrm{var}(m_t) > 0$, and m_t independent of both $u_t \sim \mathcal{N}(0,1)$ and I_{t-1}. Then Eq.(2.22) and Eq.(2.26) are equivalent with $\langle \sigma_t^2 | I_{t-1} \rangle = h_t$ and $\mathrm{var}(\sigma_t^2 | I_{t-1}) = h_t^2\ \mathrm{var}(m_t) > 0$. So ARCH models with appropriate fat-tailed conditional distributions are SV models.

2.3.1 The Standard Stochastic Volatility Model

So far I have described a general SV model which can be taken as a theoretical representation of the broad SV family, but no specific model has been presented yet. This particular representation is necessary for computing numerical properties of the solutions. A final remark about the differences between models in this framework is that are mainly based on two factors: the modelling of the volatility process, and if random variables are correlated or not.

Let's consider a simple case where the stochastic volatility, σ_t, for daily returns is log-normally distributed. Then, $log(\sigma_t) \sim \mathcal{N}(\alpha, \beta^2)$, with α and β the usual parameters of normal distribution. There is no theoretical reason behind the common use of daily frequency as sampling frequency for returns. It is plausible the idea that this frequency was chosen because it was the highest available frequency for financial time series up to very recently. Although the lognormal distribution is the standard choice when a continuous distribution is used for volatility - and it was Clark's choice, - there are other distributions: gamma distribution, inverse gamma distribution, which are in good agreement with empirical data. Other important aspect for volatility modelling is the presence of autocorrelations, which are proportional to those of absolute excess returns. This indicates that the autocorrelations of volatility must decrease slowly, because this is the observed behavior for absolute excess returns.

The simplest stationary stochastic process for volatility is a Gaussian AR(1) process[7] for its logarithm such as

$$log(\sigma_t) - \alpha = \phi(log(\sigma_{t-1}) - \alpha) + \eta_t. \tag{2.29}$$

The parameter ϕ represents volatility persistence, with $-1 < \phi < 1$. The iid volatility residuals η_t have distribution $\mathcal{N}(0, \sigma_\eta^2)$. This process is a simple modelling of volatility where autocorrelations and slow decay are taken into account.

The standard SV model - based on Taylor [23] - is given by Eq.(2.22) and two further assumptions. First, the iid variables u_t are distributed as $\mathcal{N}(0,1)$, and second processes σ_t and u_t are stochastically independent. The returns process is strictly stationary, since it is the product of independent strictly stationary processes. The main properties of this model are summarized as (i) all the moments of returns are finite, (ii) the kurtosis of returns equals $3\exp(4\beta^2)$, (iii) the correlation between returns r_t and $r_{t+\tau}$ is zero for all $\tau > 0$, (iv) the correlation between the squared excess returns $s_t = (r_t - \mu)^2$ and $s_{t+\tau}$ is positive for all $\tau > 0$, and (v) the autocorrelation function of $a^p = |r_t - \mu|^p$ has approximately the same shape as that of s_t for all positive p.

In the standard SV model the unconditional density function of returns is symmetric about its mean μ, and volatility is modeled by a log-normal density function. Thus, when we integrate these distributions we get a solution equivalent to that previously obtained by Clark, Eq.(2.6). Reinforcing the idea that a SV model is a particular case of MDH model. A result of this model is that all its moments are finite. This is a very important difference with SD models that makes SV paradigm very attractive, given that it makes possible to keep untouched all the results related to a finite variance. Moreover, the moments of the distribution can be easily calculated. So, for any positive number p,

$$\langle |r_t - \mu|^p \rangle = \langle \sigma^p \rangle \langle |u_t|^p \rangle, \tag{2.30}$$

as $\log(\sigma_t^p) = p\log(\sigma_t)$, the distribution of $\log(\sigma_t)$ is $\mathcal{N}(p\alpha, p^2\beta^2)$ and thus

$$\langle \sigma_t^p \rangle = \exp(p\alpha + \frac{1}{2}p^2\beta^2), \tag{2.31}$$

then first three moments of the distribution are

$$\begin{aligned} \langle |r_t - \mu| \rangle &= \sqrt{2/\pi}\exp\left(\alpha + \frac{1}{2}\beta^2\right), && (2.32)\\ \text{var}(r_t) &= \exp\left(2\alpha + 2\beta^2\right), && (2.33)\\ \text{kurtosis}(r_t) &= 3\exp(4\beta^2). && (2.34) \end{aligned}$$

[7] An AR(1) process is a first-order one process, meaning that only the immediately previous value has a direct effect on the current value. It is employed for describing a stochastic process that can be expressed as a weighted sum of its previous values and a white noise error,

$$e_t = re_{t-1} + u_t, \tag{2.28}$$

where r is a constant that has absolute value less than one, and u_t is drawn from a distribution with mean zero and finite variance, often a normal distribution.

The most important aspect of this solution is not its specific numerical values, but all of them are finite. The main advantage of SV models over SD models is that the former are compatible with other well established paradigms in mathematical finance. Moreover, from a theoretical point of view, empirical distributions might be accurately fitted by SV models. However, the main problem with these models in general is that they do not fully match empirical distributions and as it happened with SD models empirical tails are not accurately represented, then it seems logical that some further research is necessary but without discarding the SV assumptions.

2.4 Conclusions

In this chapter, I have presented the two main families of models: SD and MDH, which were developed as an attempt to describe empirical returns distribution. Both families assume non-Gaussian solutions which are represented by heavy-tailed distributions with more density of probability associated to large events. On the other hand, the main difference between them is about their convergence to a Gaussian. In the SD family, solution is based on stable distributions which are self-similar and are attractors of the family of distributions with non-finite variance, they therefore do not converge to a Gaussian. The family of solutions based on the Mixture Distributions Hypothesis is characterized by producing solutions with all their moments well-defined, so this family converges to a Gaussian when they are aggregated as a consequence of the Central Limit Theorem. Another relevant difference between the two families is the shape of the tails of the returns distribution according to each model: SD models only produce tails which decay as a power-law, whereas MDH can generate more different types of decay such as exponential.

MDH solution is the theoretical framework of Stochastic Volatility models which constrained by a few theoretical assumptions are able to explain three sytilized facts: non-gaussianity of returns distribution, absence of linear correlations, and positive autocorrelation of squared returns - also known as volatility clustering. This explanatory capability of SV models makes them very attractive for researchers, but not for practitioners because of their difficult implementation. The difficulty comes from the modelling of volatility which is a latent variable. In general, practitioners implement a discrete-time version of SV models which are ARCH models and that let produce distributions with similar characteristics to these generated by SV, while they are more easily implemented.

The accuracy of any specific solution is not only depending on the theoretical approach, but on the available database that lets researchers to employ a certain estimation of the volatility. Modern high frequency databases bring a great opportunity to researchers for making models that explain prices at the finest time level: event time.

Bibliography

[1] J. Bertoin, *Lévy Processes* (Cambridge University Press, Cambridge, 1996).

[2] R. C. Blattberg, and N. J. Gonedes, J Business **47**, 244 (1974).

[3] J.-Ph. Bouchaud, Physica A **285**, 18 (2000).

[4] P. Carr, and L. Wu, Journal of Financial Economics **71**, 113 (2004).

[5] P. K. Clark, Econometrica **41**, 135 (1973).

[6] E. F. Fama, J. Business **38**, 34 (1965).

[7] F. D. Foster, and S. Viswanathan, Journal of Business and Economic Statistics **13**, 379 (1995).

[8] B. V. Gnedenko, and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables* (Addison-Wesley, Reading, Massachusetts, 1968).

[9] L. Harris, Journal of Financial and Quantitative Analysis **21**, 39 (1986).

[10] L. Harris, Journal of Financial and Quantitative Analysis **22**, 127 (1987).

[11] S. Heston, Review of Financial Studies **6**, 327 (1993).

[12] A. W. Lo, and J. Wang, Review of Financial Studies **13**, 257 (2000).

[13] D. B. Madan, and E. Seneta, Journal of Business **63**, 511 (1990).

[14] B. B. Mandelbrot: J. Business **36**, 394 (1963).

[15] R. N. Mantegna, and H. E. Stanley, Phys. Rev. Lett. **73**, 2946 (1994).

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

[17] R. Cont, Quantitative Finance **1**, 223 (2001).

[18] M. Richardson, and T. Smith, Journal of Financial and Quantitative Analysis **29**, 101 (1994) .

[19] B. Rosenberg, Working paper 11, Graduate School of Business Administration, University of California, Berkeley, (1972), reprinted in [20].

[20] N. Sephard, *Stochastic Volatility: Selected Readings* (Oxford Univ. Press, USA, 2005).

[21] G. Tauchen, and M. Pitts, Econometrica **51**, 485 (1983).

[22] S. J. Taylor, Financial returns modelled by the product of two stochastic processes, a study of daily sugar prices, 1961-1979, *Time series analysis: theory and practice*, O. E. Anderson (Editor) (North-Holland, Amsterdam, 1982).

[23] S. J. Taylor, *Modelling Financial Time Series*, (Wiley, Chichester, 1986).

[24] S. J. Taylor, Mathematical Finance **4**, 183 (1994).

Chapter 3

A Superstatistical Stochastic Volatility Model

3.1 Introduction

In the previous chapters I have shown that empirical price dynamics can be characterized by a series of stylized facts which are common to different financial assets. I have also presented different attempts to explain and describe the asset price dynamics. First models - Gaussian models - tried a theoretical explanation from first principles. Although it was demostrated in the 1960s that these models were not able to reproduce empirical distributions, different methods employed in the deduction of the solution are currently used in mathematical finance and the postulates for markets and prices are commonly assumed by any theoretical framework. Moreover, Gaussian distribution is broadly used as a benchmark for further theoretical models.

Non-Gaussian models - SD and MDH - were conceived as an attempt to mainly explain the heavy tails observed in empirical distributions, due to this all new solutions were logically heavy-tailed. Although SD and MHD are commonly presented as different solutions, SD are inded a particular solution of the more general MDH. This last framework is not very restrictive and may be described as the result of a doubly stochastic modelling with a conditioned probability which is always a Gaussian, and another distribution which is the driving factor of the non-Gaussian behavior of the returns distribution. Many models fitting this simple description of the framework may be produced by only considering different non-Gaussian distributions. SV models made this by modelling the volatility and taking it as the cause of the observed non gaussianity. It has been presented in Chapter 2 that SV models give theoretical support to the explanation of several stylized facts, but the problem with SV is the accuracy of the specific implementations.

It was shown in the Introduction that financial systems exhibit several properties which can be also found in physical complex systems, that is the reason because mathematical tools previously employed in the modelling of complex systems are tested in financial systems. A branch of statistical mechanics called Superstatistics [6] which has been succesfully used in the study of atmospheric turbulences [22, 23], cosmic ray statistics [7], solar flares [5], and random networks [1, 15] is mathematically expressed in a similar way SV models are, but with some important differences. Superstatistics assumptions state the existence of two dynamics, and the possibility of dividing the system into cells where an intensive parameter takes a constant value. These assumptions are more restrictive than those of the SV models then it must be checked if they are compatible with financial systems.

In this chapter, I present a new theoretical approach which may be understood as a SV model in which superstatistical assumptions are implemented. As a result of this, I distinguish between a fast dynamics for returns and a slow dynamics for volatility considering this last factor constant at intraday scales. I have tested the theoretical predictions of the model against an empirical series recorded at the finest time level: event time. Athough stylized facts about the shape of returns distribution at high and low frequencies are equivalent, at this time level microstructural effects related to price formation must be taken into account. Theoretical predictions of the model presented in this Chapter and their excellent agreement with empirical data let me conjecture a universal behavior of the empirical returns distribution.

3.2 Superstatistics

As I said in the Introdution of this Chapter, SV models give a solid theoretical support to price dynamics understanding because there is no contradiction between empirical findings and the theoretical results which may be produced within this framework. The lack of accuracy of the theoretical predictions may be regarded as due to the specific implementations. In this Section, I present a theoretical framework previously employed in the modelling of physical systems with similar properties to these of financial systems. The aim is to develop a superstatistical SV model to improve the accuracy of theoretical results.

The superstatistics is a branch of statistical mechanics devoted to the study of non-linear and non-equilibrium systems. It is characterized by using the superposistion of multiple differing statistical models to explain the non-linearity, so in terms of common statistical ideas this is equivalent to compounding the distributions of random variables, and it may be considered a case of a doubly stochastic model. Complex non-equilibrium systems may be described by a superposition of different dynamics on different time scales. This framework assumes the existence of a fast dynamics which is represented by a given stochastic process and a slow dynamics which is the responsible for the parameters of that process.

Superstatistics was first employed in the study of physical systems [6, 8], then usual superstatistical systems are physical, e.g. a Brownian particle moving through a changing environment. In this problem, the fast dynamics is that of the velocity of the Brownian particle, and the slow dynamics is that related to the changes in the environment. The two effects are associated with two well separated time scales, as a result we have the superposition of two statistics. The aim of this thesis is the description of price dynamics where fast dynamics is that of intraday price changes, and slow dynamics is the one related to volatility. As it is shown below time scales are different for these two processes, and their marginalization gives as result the distribution of returns.

The stationary distributions of superstatistical systems typically exhibit non-Gaussian behavior with fat tails, which may decay with a power law, or as an exponential, or even in a more complicated way. This is a common element with financial systems where we can often find distributions with slow decay. In addition to the two different time scales, a key ingredient of superstatistic models is the existence of an intensive parameter β that fluctuates on a large spatio-temporal scale T. In the case of the brownian particle, β is the fluctuating inverse temperature of the environment, but β might represent an effective friction constant, a changing mass parameter, a changing amplitude of Gaussian white noise, the fluctuating energy dissipation in turbulent flows, or simply a local variance parameter extracted from a signal. In the present study, intensive parameter β is related to daily variance of returns and is considered constant for daily time scales. According with previous definition β gives the parameter for understanding and describing intraday returns.

A usual superstatistical system is a non-equilibrium system which is inhomogeneous and consists of many cells with different values of the intensive parameter β. Cells may be spatial, but they may be also temporal as it is the case in financial systems because the subject of study is time series. A necessary condition for the cells is that they must be clearly differentiated. The cell size is conditioned on the behavior of β, and the value taken by this parameter must be constant or roughly constant in each cell during the time interval T. For quantifying the cell size we can employ the correlation length of β series in the case of time series, or of β-field in a spatial system. The condition is that the magnitude of β stays constant[1] in the entire cell, then it changes in a certain amount.

I have presented the basic concepts for describing a superstatistical system: the existence of two dynamics with two two time scales, an intensive parameter β, and a set of cells with different values of β from cell to cell but with constant value within themselves.

All these concepts and definitions are common to every superstatistical system, but it is convenient to see how they were originally applied to a physical system. Following with the system of a Brownian particle moving through a

[1] β, in the specific case of the model developed in this thesis is related to the variance. Thus, the condition for the cell size is that variance can be considered nearly constant in that interval of time.

changing environment, as it was firstly described in the work of Beck and Cohen [6]. Let's consider the Brownian particle, in the long term run ($t >> T$), the stationary distributions of the inhomogeneous system arise as a superposition of Boltzmann factors $e^{-\beta E}$ weighted with the probability density $f(\beta)$ to observe some value β in an arbitrary cell:

$$p(E) = \int_0^\infty f(\beta) \frac{1}{Z(\beta)} \rho(E) e^{-\beta E} d\beta, \tag{3.1}$$

where E is an effective energy for each cell, $\rho(E)$ is the density of states, and $Z(\beta)$ is the normalization constant of $\rho(E)e^{-\beta E}$ for a given β. When these concepts are applied to a Brownian particle of mass m moving through a changing environment in d dimensions. For its velocity v one has the local Langevin equation

$$dv = -\gamma v + \sigma L(t), \tag{3.2}$$

where $L(t)$ is a d-dimensional Gaussian white noise which becomes superstatistical because for a fluctutating environment the parameter β becomes a random variable as well, it varies from cell to cell on the large spatio temporal scale T. In this case $E = \frac{1}{2}mv^2$, and while on the time scale T the local stationary distribution in each cell is Gaussian with variance $1/\beta$,

$$p(v|\beta) = \left(\frac{\beta}{2\pi}\right)^{d/2} e^{-\frac{1}{2}\beta m v^2}, \tag{3.3}$$

the marginal distribution describing the long-time behavior of the particle for $t >> T$,

$$p(v) = \int_0^\infty f(\beta) p(v|\beta) d\beta, \tag{3.4}$$

exhibits non-trivial behavior. The distribution of $|v|$ is heavy tailed and its distribution depend on the behavior of $f(\beta)$. This pattern is similar to that found in returns distribution, where we also observe heavy tailed distributions depending on the behavior of the volatility. It seems that financial systems may be modeled as superstatistical systems, but it is necessary to find an equivalence of the physical magnitudes into financial variables.

3.2.1 Superstatistics in Finance

Superstatistical framework has been applied to physical systems with similar distributional characteristics to these that we can find in financial systems, being the paralellism between thermodynamics and finance not new [10, 19, 24]. The main challenge is to find a plausible description of the financial distributions in terms of superstatistical concepts. Otherwise, we only would be able to state that are systems with similar distributions but essentially different from this point of view.

It is important to realize that Eq.(3.4) is equivalent to that found for SV models where returns distributions depended on the behavior of the volatility, so if it were possible to find a temporal cell with a constant value for the volatility in a financial time series all the methods and results of the Superstatistics framework would be valid for these systems, and we could define a financial system as a superstatistical system.

Let's consider a financial time series r and the probability distribution, $P(r)$, of this random variable r. The time series is divided into N equal slices of lenght l, and the subject of study is the distributions of the sliced series. According to Superstatistics, we assume two different time scales τ and T such that $\tau/T << 1$. We also assume that time scale T is the length of the slices , $l = T$, and study local probability distribution for every time slice. It is a good approach to consider that local distributions[2] are Gaussian distributed

$$p_{T,i}(r) = \sqrt{\frac{\beta_{T,i}}{2\pi}} \exp^{-\frac{1}{2}\beta_{T,i} r^2}, \tag{3.5}$$

where $\beta_{T,i} = \frac{1}{\langle r^2 \rangle_{T,i}}$ is β factor computed for slice i of length T. Following with this assumption, the distribution $P(r)$ is approximated by

$$P(r) \approx p_T(r) = \frac{1}{N} \sum_{i=1}^{N} p_{T,i}(r), \tag{3.6}$$

where N is the number of points in a slice of length T. Therefore, there are N values for $\beta_{T,i}$. When N is large enough, Eq.(3.6) may be replaced by

$$P(r) \approx p_T(r) \approx p_{T,f}(r) = \int_0^\infty d\beta f_T(\beta) \sqrt{\frac{\beta}{2\pi}} \exp^{-\frac{1}{2}\beta r^2}, \tag{3.7}$$

where $f_T(\beta)$ is the probability density of β in a randomly chosen time slice of length T equals β. The distribution of $f_T(\beta)$ depends on T because this parameter determines the length of the time slices. Therefore, it also affects the value of β.

Given the above theoretical description it seems plausible to consider financial systems as a good candidate to superstatistical system. Moreover SV models are expressed by mean of equations similar to these employed in the physical example presented in Section(3.2), giving a formal support from a mathematical perspective. However there is an additional condition which must be fulfilled by financial systems for being considered as superstatistical: we must find the set of cells with constant β, and then we can perform empirical tests to check the adequacy of the model. A final remark is that the superstatistics concept is quite general and has recently been applied to a broad variety of physical systems, including atmospheric turbulence [22, 23], cosmic ray statistics [7], solar flares [5], random networks [1, 15].

[2] I consider local distribution to the distribution of length T of the random variable r.

3.2.2 Superstatistical Classes

Superstatistical classes are related to the probabilistic description of the parameter β and discern all possible candidates for describing $f\left(\beta\right)$, which must fulfill some constraints: (i) $f\left(\beta\right)$ must be a normalized probability density, (ii) the integral shown in Eq.(3.1) must exist, (iii) in the case that $f\left(\beta\right)$ is constant the new statistics should reduce to a Boltzmann-Gibbs statistics in a physical case, or in the returns distributions case to a normal distribution giving the standard Gaussian model. It has been shown [8] that many experimental data are well described by three major universality classes: Gamma, Inverse Gamma, and log-normal superstatistics. These superstatistics represent a universal limit statistics for large classes of dynamical systems and all of them fulfill the conditions. The distribution $f(\beta)$ is determined by the spatio-temporal dynamics of the entire system under consideration and by construction β is positive, so $f(\beta)$ cannot be Gaussian.

Let's consider three examples. First, there may be independent, or weakly correlated, microscopic random variables $\xi_n, n = 1,...,N$, contributing to β in an additive way. For large N their rescaled sum $\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\xi_n$ will approach to a Gaussian random variable X_1 due to Central Limit Theorem. There may be many different random variables consisting of microscopic random variables, i.e. we have n Gaussian random variables $X_1,...,X_n$ due to various degrees of freedom in the system. As mentioned before, β needs to be positive; a positive β is obtained by squaring these Gaussian random variables. The resulting $\beta = \sum_{i=1}^{n} X_i^2$ is Gamma distributed,

$$f(\beta) = \frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{n}{2\beta_0}\right)^{n/2}\beta^{n/2-1}e^{-\frac{n\beta}{2\beta_0}}, \tag{3.8}$$

where β_0 is the average of β. When marginalised with normal distribution the results exhibit power-law tails. This distribution was expected by construction, because the Gamma distribution naturally arises when n independent Gaussian random variables X_k with average 0 are squared and added, then β is Gamma distributed.

In the second example we make the same assumptions for the intensive parameter β^{-1} rather than for β is the sum of several squared Gaussian random variables ξ_n. The resulting $f(\beta)$ is the inverse Gamma distribution given by

$$f(\beta) = \frac{\beta_0}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{n\beta_0}{2}\right)^{n/2}\beta^{-n/2-2}e^{-\frac{n\beta_0}{2\beta}}, \tag{3.9}$$

this generates final distributions - after being marginalised - that have exponential decays.

Third example, in this case instead of β being a sum of many contributions, for other systems - in particular, turbulent ones - the random variable β may be generated by multiplicative random processes. We may have a local random variable $X_1 = \prod_{n=1}^{N} \xi_n$, where N is the number of steps and the ξ_n are random variables. By the Central Limit Theorem, for large N the random variable $\frac{1}{\sqrt{N}} \log X_1 = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \log \xi_n$ becomes Gaussian for large N, then X_1 is log-normally distributed. In general, there may be n such product contributions to β, i.e. $\beta = \prod_{i=1}^{n} X_i$. Then $\log \beta = \sum_{i=1}^{n} \log X_i$ is a sum of Gaussian random variables; therefore, it is Gaussian too. Thus β is log-normally distributed,

$$f(\beta) = \frac{1}{\sqrt{2\pi} s \beta} \exp\left\{ \frac{-\left(ln\frac{\beta}{\mu}\right)^2}{2s^2} \right\}, \tag{3.10}$$

where μ and s^2 are the common mean and variance parameters of a log-normal distribution, then log-normal superstatistics is universal too and as in the previous two cases this distribution was expected by construction.

Although more complicated classes are possible, most experimental cases fall into one of these three universality classes or into simple combinations of them. I show below that financial time series are in good agreement with inverse Gamma superstatistical class.

3.3 High-Frequency Financial Time Series

I already mentioned in the Introduction of this Chapter that data set employed to test the accuracy of the new theoretical approach is recorded at high-frequency. This implies that we can study the finest time level of financial series being possible to obtain higher levels by simple aggregation of this one. On the other hand, it also implies that new effects only observed at this level must be taken into account, e.g. bid-ask bounce effect, and the negative first-order autocorrelation of returns. Sampling frequency of the data set is a relevant factor due to the large number of data points that let me test properly theoretical predictions against the tails of empirical distributions. We can not forget that Bachelier did not realize his model was wrong because the deviations from the theoretical model were attributed to the size of the sample.

High-frequency term is used to indicate higher frequency than daily. Authors like Engle [13] use the expression ultra high-frequency for datasets containing prices and/or quotes at intraday intervals. There is a problem with data recorded at this frequency because data points are not uniformly distributed, whereas common research employs regularly sampled data at five minutes interval for two reasons: at higher frequencies we have to deal with microstructural effects, and working with non homogeneous time intervals is unusual for econometric models. In addition to these factors, we must consider that five-minute returns require a new price every five-minute interval and this can be difficult when dealing with illiquid stocks because it is possible not to have any quote or transaction in that interval. A usual method to solve this is the use of interpolated prices between the last price in an interval and the first price in the next interval [2], but this method can create spurious predictability. In addition to these reasons, to obtain reliable high-frequency databases is not cheap, compared to daily prices which are free for many stocks. There were few high-frequency studies before 1990s [17, 18, 25] for the US equity market. Anyway the most studied high-frequency equity data are probably those of the Trade and Quotation database of NYSE, AMEX, and NASDAQ prices, which are available from the New York Stock Exchange. Olsen & Associates (O&A) gave away a year of their ultra high-frequency exchange rate data, leading to several studies that were presented at the O&A conference in 1995.

Stylized facts for high-frequency returns are similar to those at lower frequencies. The aim of this thesis is to explain the shape of returns distribution, and this is leptokurtic at high frequencies presenting a sharp spike at zero as it happened at lower frequencies. This can simply reflect the feasible set of discrete prices rather than few trades per interval, e.g. some 22% of the five-minute returns [4] are zero but less than 3% of their five-minute intervals contain no transactions. Autocorrelation of intraday returns is the stylized fact where microstructural effects are more pronounced, then more dependence may be assumed in intraday returns than in daily or lower frequency returns for two reasons. First, bid-ask bounce in transaction prices caused by dealers with order imbalances is more evident at higher frequencies and the negative autocorrelation created by bouncing prices is proportional to the variance of the spread divided by the variance of the returns; the former is constant while the latter decreases as the frequency of returns increases. Second, net profit based on any dependence is more difficult because expected profits per trade decline as data frequency increases and costs stay constant. Finally, autocorrelation of intraday absolute returns is also present in intraday returns. As I mentioned in Section(3.2), the autocorrelation is an important element for the variable related to the slow dynamics because it is linked to the size of the cell.

3.4 A Superstatistical Stochastic Volatility Model

In this section, I present a new theoretical model for high-frequency returns distribution. The model can be considerd within the SV framework because returns are functionally determined by the marginalisation of a Gaussian with the volatility distribution. However this new model holds an important and remarkable particularity: volatility is taken as constant for intraday time intervals. This is a new approach to volatility modelling because it assumes volatility as a slow variable which is less frequently shocked than returns. Moreover, it implicitly assumes two stylized facts: the absence of linear autocorrelation of returns, and the positive autocorrelation of absolute returns, as empirical support to the existence of a double dynamics responsible of the final shape of the returns distribution. This feature is related to Superstatistics where we have two different time scales, and two different dynamics. Slow dynamics would be given by this of volatility, which is positively autocorrelated. Fast dynamics would be given by that of intraday returns, which are almost uncorrelated. Then the model from a statistical mechanical point of view is superstatistical.

This model works at the finest time scale dynamics: event time dynamics. The specific definition for event, as it is employed in the present work, is any action in the order book that leads to a change in any of the two best prices: bid, or ask. Hence microstructure effects and high-frequency details must be taken into account. As with any other theoretical approach to model returns the aim is to explain their empirical distribution. For this, I not only present a theoretical model, but I have also performed empirical tests with data from different stocks wich are traded in different exchanges showing an excellent agreement with theoretical predictions. In addition to the different exchanges, the time periods under study are also different. In spite of all this diversity of economic sectors, different countries, and periods of time results are similar, giving support to a universal behavior of price dynamics.

3.4.1 Theoretical Description of The Model

In this Section, I present the new model for high-frequency returns from a theoretical perspective. From a physical point of view it may be classified as superstatistical, but according to mathematical finance may be assimilated to a SV model, and indepently of the classifying criterion is doubly stochastic. It is necessary to explain the specific meaning of common terms of the model, in the way they are employed in the present study, and a mathematical formulation of these terms for making them more precise. I have already explained that the term event is applied to any action in the order book which causes a change in one of the best prices: bid, and ask. I make no distinction based on the origin of the variation being only interested in the change of the price itself.

The price of a stock at any time is the midpoint price, such as

$$p_t = \frac{\left(p_{b,t} + p_{a,t}\right)}{2}, \tag{3.11}$$

where p_t is the midpoint price at time t. $p_{b,t}$ is the bid price at time t, and $p_{a,t}$ is the ask price at time t. There are two reasons for using midpoint price. First, because midpoint price takes into account bid and ask prices which are quotes. In doing so, any variation in these values is immediately reflected in price, so I may take into account the information contained in quotes and not only in transactions. Second reason is about a microstructural effect at high-frequency, by calculating midpoint price we are able to avoid the bid-ask bounce effect which could induce some spurious statistics.

Returns are defined as the diference in two logarithmic prices from time t to time $t+\tau$

$$r_t(\tau) = \ln(p_{t+\tau}) - \ln(p_t), \tag{3.12}$$

where $r_t(\tau)$ is the t^{th} return, $p_{t+\tau}$ is the midpoint price at time $t+\tau$, and p_t is the midpoint price at time t. When aggregating returns over longer time periods, I only use non-overlapping intervals. Overlapping time intervals increases the number of data points, so the distance between the starting points of two subsequently analyzed intervals becomes smaller than the interval size. Although intuitively we can see that adding overlapping intervals to the sampling scheme might increase the precision of the result, the resulting data series certainly exhibits serial dependence [16]. Therefore we cannot consider the observations as independent. Another important aspect is the time unit. I have set the the unit of time index, t, as midpoint time. This means that t is updated whenever an event causes a change in the midpoint between the prevailing best quotes. Although the results presented in this section are calculated in event time, tests performed in calendar time show similar results.

In the present model, individual returns, $r_t(\tau=1)$ are decomposed into two components, a volatility term, σ_t, and a Gaussian , $\mathcal{N}(0,1)$ noise term, ξ_t, such as

$$r_t = \sigma_t \xi_t. \tag{3.13}$$

It is assumed that σ_t is sufficiently slow varying, such we can treat it as a constant over intraday time scales. Replacing σ_t with its local constant value, σ. The individual returns are expressed by following equation,

$$r_t \approx \sigma \xi_t. \tag{3.14}$$

For the computation of the constant daily volatility σ. I calculate the returns, using Eq.(3.12), at all possible time intervals for each day, that is from $\tau=1$ to $\tau=N$, where N is the total number of events in a day[3], then I calculate the volatility of this returns series. For this, it is necessary to take into account that we are considering different time intervals; therefore, individual volatilities at any time horizon σ_τ must be scaled as

$$\sigma^*_\tau = \frac{\sigma_\tau}{\sqrt{\tau}}, \tag{3.15}$$

where σ_τ is the volatility for returns of time interval τ at a specific day, and σ^*_τ is the scaled individual volatility at time τ. This scaling is appropriate because we have a process with normal diffusion, following

$$\langle r^2(\tau)\rangle = \sigma^2\tau. \tag{3.16}$$

Daily σ is the average for all σ^*_τ from $\tau=1$ to $\tau=N$.

$$\sigma = \sqrt{\frac{1}{N_T}\sum_{j=1}^{N}\sum_{k=1}^{N_\tau}\sigma^{*2}_{j,k}}, \tag{3.17}$$

where N_T is th total number of individual volatilities, N is the maximum value of τ, and N_τ is the total number of scaled individual volatilities of length τ. Thus,

[3] N is not a constant value, and it takes a specific value for every single day.

daily volatility is the result of averaging all scaled individual volatilies. Taking into account all these intraday volatilities leads to a large number of observations within a single day. This technique is not new in high-frequency finance [12], and the reason of its use is because we can reduce the stochastic error as a consequence of increasing the available statistics.

Following with the Superstatistics framework, we define the variable β as the inverse squared volatility,

$$\beta \equiv \frac{1}{\sigma^2}, \tag{3.18}$$

where σ is the value of daily volatility, then β is a daily value too. In Eq.(3.13), I have assumed a Gaussian process for the dynamics of the returns. So, for a certain value τ in a given day characterized for a certain value β. I obtain that the probability distribution of returns is

$$p(r,\tau|\beta) = \sqrt{\frac{\beta}{2\pi\tau}} \exp\left(-\frac{\beta r^2}{2\tau}\right), \tag{3.19}$$

I consider β varies slowly, because I have assumed that slow dynamics is that of the volatility. By slowly variation, I mean that β fluctuations are negligible compared to price fluctuations when observed over the time scales studied here, this is up to one trading. This is not inconsistent with shocks to volatility, such as might occur during significant news events, as long as these shocks are relatively infrequent, i.e. not a daily occurrence.

So far, I have shown that daily returns can be explained by a Gaussian distribution and volatility can be considered constant during a day, but it is necessary to model the behavior of β in a diferent time scale. Fluctuations of β over longer time scales can be characterized by a probability distribution $g(\beta)$, there are several studies which stated different functional forms for the distribution of volatility [9, 11, 20]. We have also seen in the section about Superstatistics three possible solutions to it: lognormal distribution, gamma distribution, and inverse gamma distribution. It is out of the scope of this thesis to compare the different possible functional forms. Instead of that, I assume - empirical results support my assumption - that $g(\cdot)$ is similar across stocks and close to a gamma distribution[4]

$$g_{n,\beta_0}(\beta) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right). \tag{3.20}$$

There are several simple explanations to why the inverse variance might have this distribution [9, 21]. The probability distribution obtained in the Eq.(3.19) is that for returns on any single day in our model. Because β can vary at longer time scales, the returns distribution observed with data pulled from many different days that span a long period of time is obtained by marginalising over β.

[4]It is also a common expression for the gamma distribution an equation, such as $f_{a,b}(\beta) = \frac{b^a}{\Gamma(a)} \beta^{a-1} e^{-b\beta}$. Both forms are obviously equivalent.

A straightforward integration of the conditonal probability of returns, $p(r,\tau|\beta)$, and the distribution $g(\beta)$ yields the following for the return distribution

$$P(r,\tau) = \frac{\Gamma\left[\frac{(n+1)}{2}\right]}{\Gamma\left[\frac{n}{2}\right]}\sqrt{\frac{\beta_0}{\pi n\tau}}\left(1+\frac{\beta_0 r^2}{n\tau}\right)^{-\frac{n+1}{2}}, \tag{3.21}$$

which is a variant of the Student's t-distribution. The non-Gaussian shape of the distribution results from collecting returns from time periods separated by long intervals where β is different. The stability of this shape for short to intermediate τ results from negligible fluctuations of β over these time scales.

Although it is well known that a gamma distributed inverse variance leads to Student's t-distribution for returns [12,13], this result does not explain why the return distribution retains its non-Gaussian shape for longer time scales. To explain the persistence of the non-Gausssian shape - as I showed in Chapter 2,- a possible model was the Lévy stable distributions. But, Eq.(**??**) gives an explanation to both problems: the non-Gaussian shape of returns distribution, and its apparent stability. It is the properties of volatility the cause for these two empirical findings. It was implicitly assumed that in a SV model, volatility should be the cause. However, I demosntrate this by the excellent agreement of theoretical predictions and empirical observations.

3.4.2 Data

I have studied a large set of empirical data, of the order of 10^7 data points. They were from three different stock markets: the London Stock Exchange (LSE), the New York Stock Exchange (NYSE), and the Spanish Stock Exchange (SSE). Time periods are also different: from May 2, 2000 to December 31, 2002 for LSE data, from January 2, 2001 to December 31, 2002 NYSE data, and from January 2, 2004 to December 29, 2006 for SSE. Moreover, stocks are from different economic sectors: telecommunications (TEF, VOD), banking (SAN, LLOY), financial services (PRU, RTR), pharmaceuticals and biotechnology (AZN), information technology (IBM), automobile (GM). This variety of stock markets, time periods, and economic sectors provides a more solid support to theoretical results. Therefore, theoretical model can be considered a good candidate for a universal behavior.

The three stock markets are not only located in different countries, but they rule trading sessions in a slightly different way. The London Stock Exchange consists of two parts: the completely automated electronic downstairs market (SETS), and the upstairs market (SEAQ). The trading volume, for the time period under study, is split roughly equally between the two markets. I only study the downstairs market because in its dataset there is a record of each action taken by each market member as it occurs, and timestamps are precise. In contrast, trades in the upstairs market are arranged informally between agents, and are printed later. The Spanish Stock Exchange also consists of two parts, as in the LSE it has an automated electronic downstairs market (SIBE), and an upstairs market -corros,- but in the SSE this upstairs market is very illiquid, and traded volume is a small fraction of total traded volume in both markets, less than 10%.

About market makers, there are no designated market makers either for SETS or for SIBE. However, any member of the exchange is free to act as a market maker by posting simultaneous bids and offers. This should be contrasted with the NYSE, for example, which has a designated specialist to trade each stock. Another diffrence between the markets is that clearing in the LSE and SSE is fully automated and instantaneous; in contrast, in the NYSE, clearing is done manually, creating an uncertainty in response times. During the period under study the book of the LSE is fully transparent, i.e. all orders in the book are fully revealed. In 2003 the LSE began to allow "iceberg orders", which contain a hidden component that is only revealed as the exposed part of the order is removed. This type of orders is also available in the SSE.

Trading session begins each day with an opening auction. There is a period leading up to the opening auction in which orders are placed but no transactions take place. The market is then cleared and for the remainder of the day, except for occasional periods, there is a continuous auction. I removed all data associated with the opening auction, and only analyse orders placed during the continuous auction. An analysis of the limit order placement showed [14] that in the LSE dataset approximately 35% of the effective limit orders are placed inside the book, this means that the prices of selling orders are higher than ask price and the prices of buying orders are lower than bid price. Thirty-three percent are placed at the best prices this means at bid or ask prices, and 32% are placed inside the spread this is at prices higher than bid price but lower than ask price, these findings are similar for all the LSE stocks.

The stocks studied in this Chapter are from different economic sectors. From an economic point of view, and more specifically from an event-based approach, this means that when having different news for different sectors and different markets it would be logical to expect for different behaviors. I show that this is not correct and all the stocks collapse onto a single curve, and this may be taken as a conclusive proof of the inaccuracy of the event-based foundations.

For all the stocks, we have updated information at any time of best prices: bid and ask. This information is available because quotes are taken into account, and not only transactions. Thus, if we have a new limit order inside the spread, or a cancellation of all the outstanding shares at one of the best prices, or a transaction which matches all the available liquidity at bid or ask; best prices change at that right instant. This is the reason because the present research deals with the finest possible level of price changes because if I had studied changes caused only by transactions I would be missing all the information related to order placement and cancellations at best prices. Independently if the position is finally transacted or not, I assume that any outstanding position at best prices in the order book is informative.

About filtering of databases, I have discarded data from opening and closing auctions and I have only taken into consideration price changes from trading hours. I have also taken out the first half an hour for the LSE, and the NYSE, but not for the SSE - because timestamps were not available in the database format I employed. The reason for discarding this first half an hour of trading is that it is generally considered a period of time when price formation is taking place, there is therefore a high volatility that causes a certain unstability in prices. I have studied time series with and without this first half an hour, and results are very similar, concluding that this period of price formation - at least for the present study - is not relevant. A possible explanation for this result is that price formation process happens entirely during opening auction.

3.4.3 Empirical results

In this section, I show the results of the performed tests to check theoretical results with empirical distributions. I have divided the tests into two groups, the first group of results shows the comparison of the theoretical results with a set of stocks which are traded in the same stock exchange: the LSE. Time series for all stocks span the same period of time from May 2, 2000 to December 31, 2002. These stocks are from different economic sectors, and different microstructural characteristics: liquidity, tick size, traded volume, etc. The second group of results shows the comparison of the theoretical results with six liquid and highly traded stocks from three different stock exchanges, and different periods of time. Results shown below are similar for the first and second group, this means that theoretical model is able to accurately describe empirical distributions independently of the time period taken into account, the stock exchange, the economic sector, and the specific microstructural characteristics - liquidity, tick size, traded volume - of the stocks under study.

First Group: LSE Stocks.

The stocks studied in this group are AstraZeneca (AZN), Lloyds TSB Group (LLOY), Prudential Plc (PRU), Reuters Group (RTR), and Vodafone Group (VOD). Dataset contains information about the complete on-book market (called SETS), this information includes all on-book transactions, order placement, and cancellations.

I have truncated the first 30 minutes of market activity to remove the effects of price discovery at the beginning of the day. Although I do not show below, I have reproduced same plots but considering these first 30 minutes and results are similar.

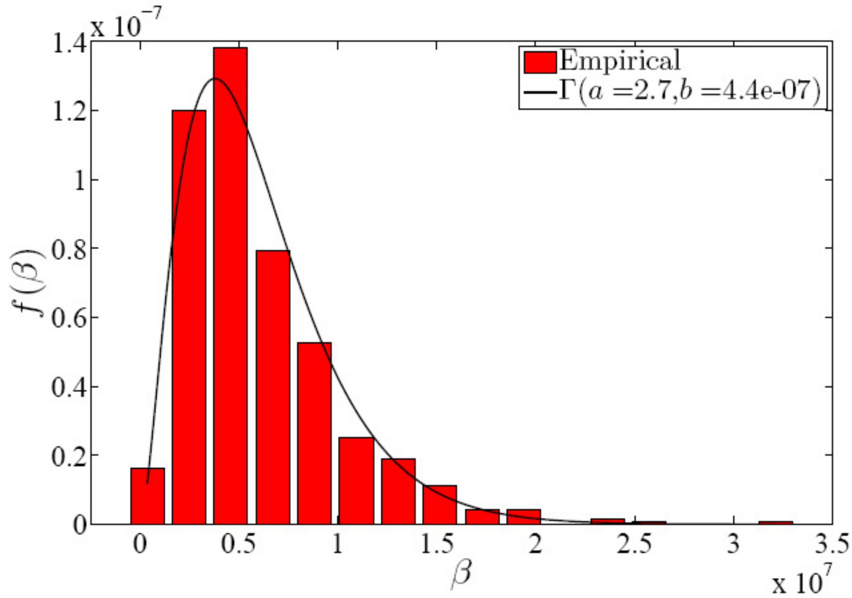

Figure 3.1: **AZN. The probability density of daily β fit by a gamma distribution.**

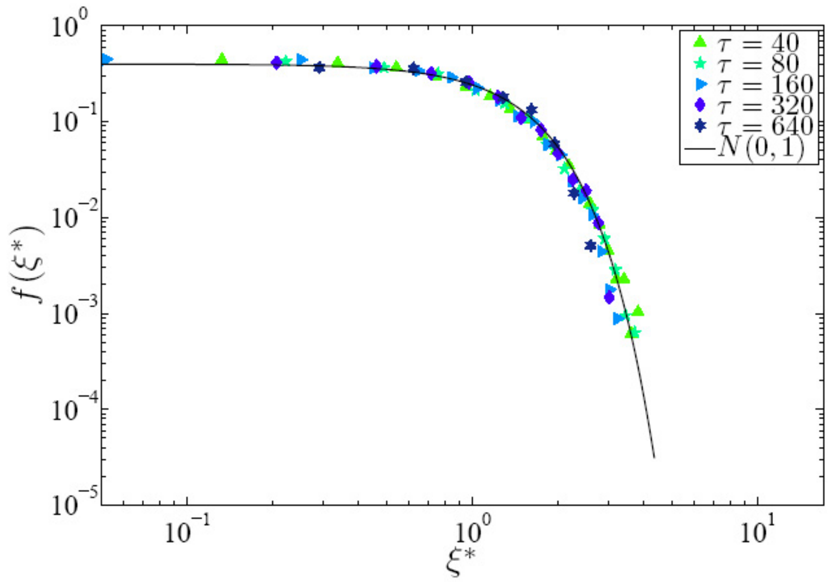

Figure 3.2: **AZN. The probability density of ξ^* for different τ compared to $\mathcal{N}(0,1)$.**

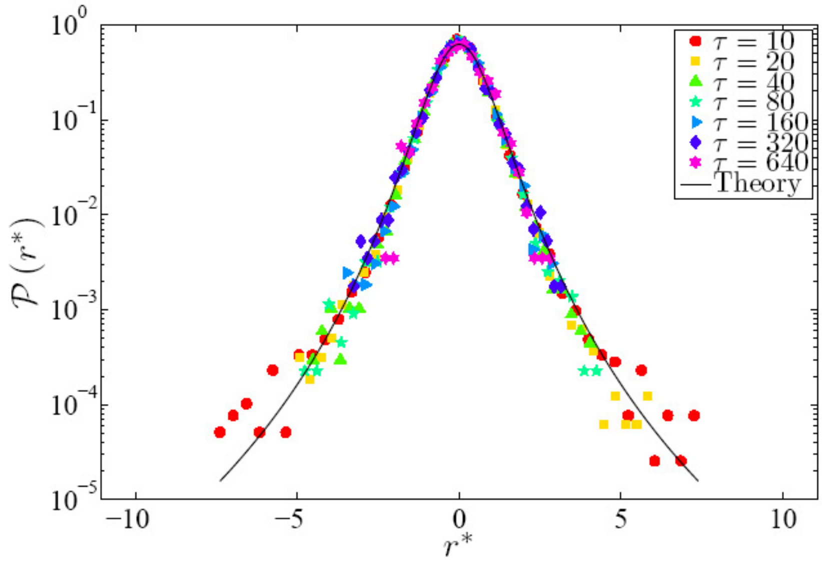

Figure 3.3: **AZN. The probability density of returns for different τ compared to theory.**

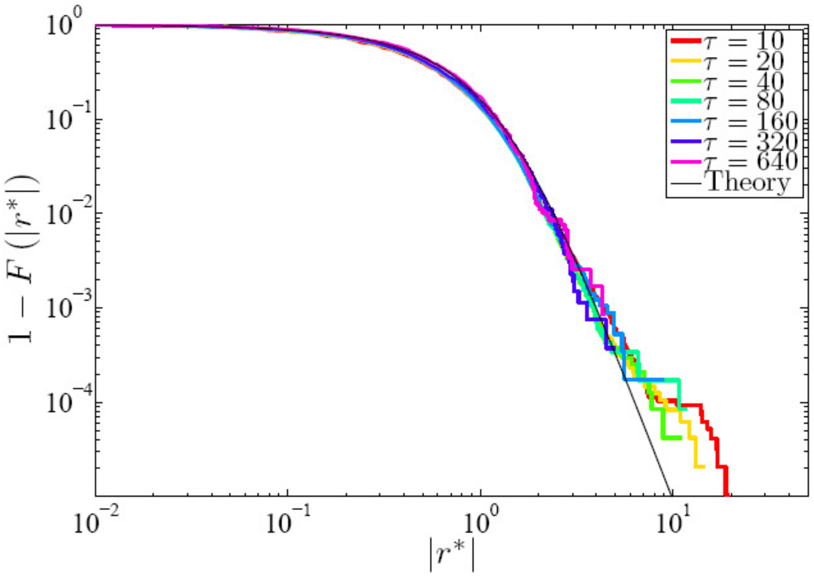

Figure 3.4: **AZN. The cumulative distribution of returns for different τ compared to theory.**

These four plots present the results obtained for one stock: AZN. Results for the other stocks in the first group, which are studied in this section are similar in appearence.

In Fig.(3.1), I plot the probability density function of β. I overlay the plot with the best fit gamma distribution and I report the parameters for this fit in the figure legend. These parameters for all the stocks of this group are reported in Table(3.1). A remarkable detail about the model is that this is the only fitted distribution. The other distributions and theoretical results are all analytically derived from the theoretical model. Consequently the implementation problems are notoriously simplified.

In Fig.(3.2), I show the probability density of ξ^* - which represents the normalized variable[5],- for $\tau = 40$ to $\tau = 640$ in loglog coordinates. This is compared to a normal distribution with zero mean and unit variance - which is assumed in the theoretical model. At time scales shorter than $\tau = 40$, which are not shown in the plot, the distribution of ξ^* is leptokurtic but with finite variance. As seen in the figure, the distribution approaches a Gaussian for time scales, $\tau > 40$. That ξ^* is Gaussian distributed was also reported for daily time scales in [3].

In Fig.(3.3), I plot the scaled return probability density for $\tau = 10$ to $\tau = 640$ in semi-log coordinates. Using the parameters obtained from the fit of the gamma distribution, I predict the full probability distribution of returns as derived in Eq.(3.21) and overlay this prediction on the plot.

In Fig.(3.4), I focus on the tails of the distribution by plotting the scaled cumulative distribution for the unsigned returns $F(|r^*|)$, where r^* represents normalized returns[6] in loglog coordinates. As seen in these two last plots, the distribution collapse both in central region and in the tails and are well described by the theoretical curve.

---

[5] $\xi^* = \frac{r_t(\tau)}{\sqrt{\frac{\tau}{\beta}}}$, where τ is the time interval.

[6] $r^* = \frac{r_t(\tau)}{\sqrt{b\tau}}$, where b is one of the parameters of the Gamma distribution of β.

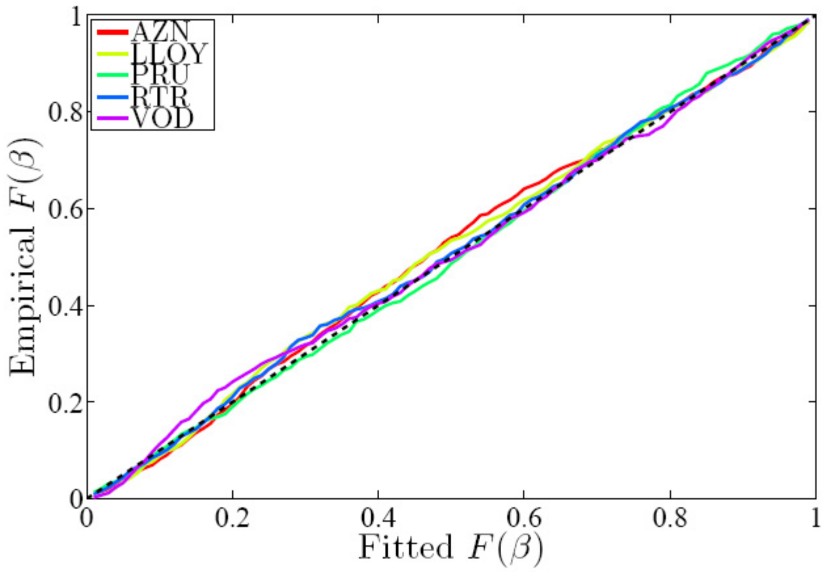

Figure 3.5: **The cumulative distribution of β compared to the cumulative distribution from the best fit to a gamma distribution for all stocks in the LSE group.**

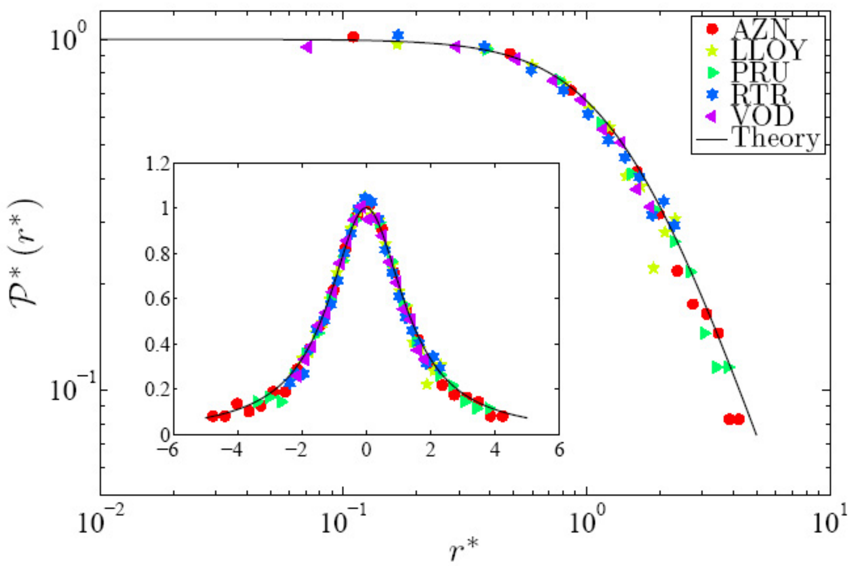

Figure 3.6: **The normalized probability density of returns with τ = 80 compared to theory for all stocks in the LSE group.**

In Fig.(3.5), I plot the empirical cumulative distribution of β versus the fitted cumulative distribution for all stocks: AZN, LLOY, PRU, RTR, VOD. This plot is created by first fixing the value of the fitted $F(\cdot)$, calculating β at this point, and then plotting the value of the empirical $F(\cdot)$ for this β. The plot is similar to a Q-Q plot, when the empirical distibution follows the fitted distribution exactly, the curve will lie on the 45° line.

In Fig.(3.6), I plot the normalized probability density $P^*(r^*)$, which is the normalized probability of the normalized returns[7] with $\tau = 80$ for the five stocks taken into account in this group. The data from all five stocks collapse on the theoretical curve.

Taken together, all these empirical results suggest that slow, but significants fluctuations in volatility produce the interesting features of the intraday return distributions, giving a strong support to the model presented in this thesis.

Security	Events	Events/Min	a	b
AZN	962516	3.0	2.7	.44e-6
LLOY	746845	2.3	3.4	1.1e-6
PRU	583792	1.8	2.6	1.5e-6
RTR	653915	2.0	3.9	3.6e-6
VOD	770352	2.4	3.9	2.1e-6

Table 3.1: Table of parameters for five stocks studied in this section.

Table(3.1) shows the different characteristics: number of events in the same time interval, number of events per minute, and the parameters of the fitted distribution of daily β's for the five stocks under study.

7

$P^* = \left(P \frac{\Gamma[a]}{\Gamma\left[\frac{2a+1}{2}\right]} \frac{\sqrt{2\pi}}{1} \right)^{\frac{2}{2a+1}}$, where P is the distribution of returns $P(r,\tau)$, and the parameters a, and b are those obtained from fitting the gamma distribution.

Second Group: LSE, NYSE, SSE Stocks.

The stocks studied in this group are AZN and VOD from the LSE, GM and IBM from the NYSE, SAN and TEF from the SSE. Periods of time are from May 2, 2000 to December 31, 2002 for the LSE data, from January 2, 2001 to December 31, 2002 for the NYSE data, and from January 2, 2004 to December 29, 2006 for the SSE. Thus, for each market I have studied two highly traded stocks which are from different economic sectors.

I only consider the electronic markets for these stocks. I have eliminated the first 30 minutes of the trading day to remove the effects of price discovery - except for the SSE because time stamps were unavailable in the database format I employed. As it happened in the previous section with stocks from the LSE, although I do not show them below, I have reproduced same plots but considering these first 30 minutes and results are similar.

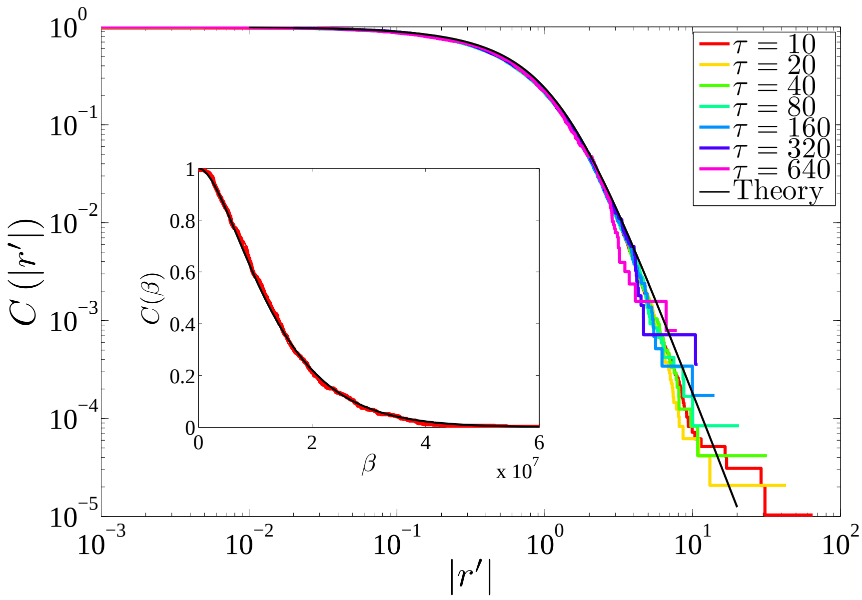

Figure 3.7: **Collapse of the complementary cumulative distribution (ccd) of absolute scaled returns, $C(|r'|)$, for the stock IBM.** The ccd is shown for times scales $\tau = 10$ to $\tau = 640$. The solid black line is the theoretical ccd using $\beta_0 = 1.4 \times 10^7$ and $n = 3.89$ from fitting β to a gamma distribution. Inset: ccd of the slow fluctuating variable β for IBM, the red curve is the empirical ccd and the solid black line is a fit to a gamma distribution.

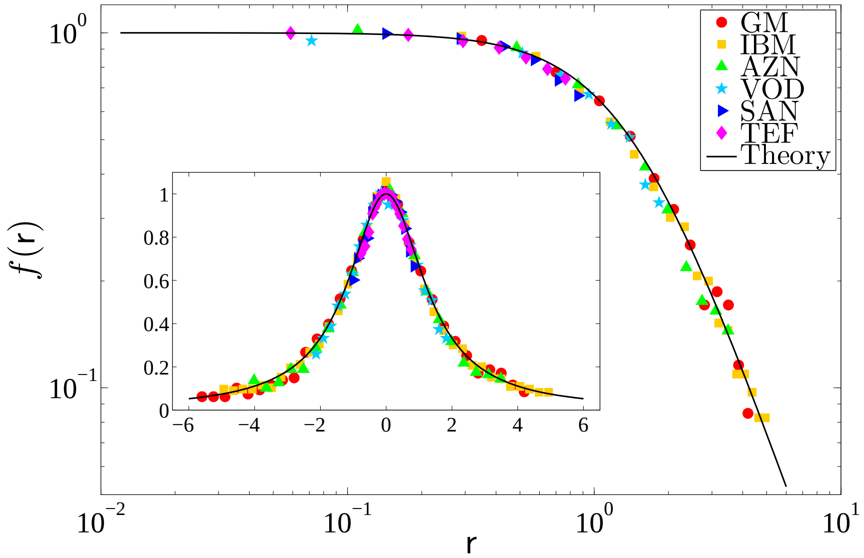

Figure 3.8: **Collapse of the return distribution on the function $f(r')$, for the stocks studied in this group.** For each stock, the return distribution for $\tau = 80$ is shown in logarithmic coordinates. Inset: The same plot in regular coordinates.

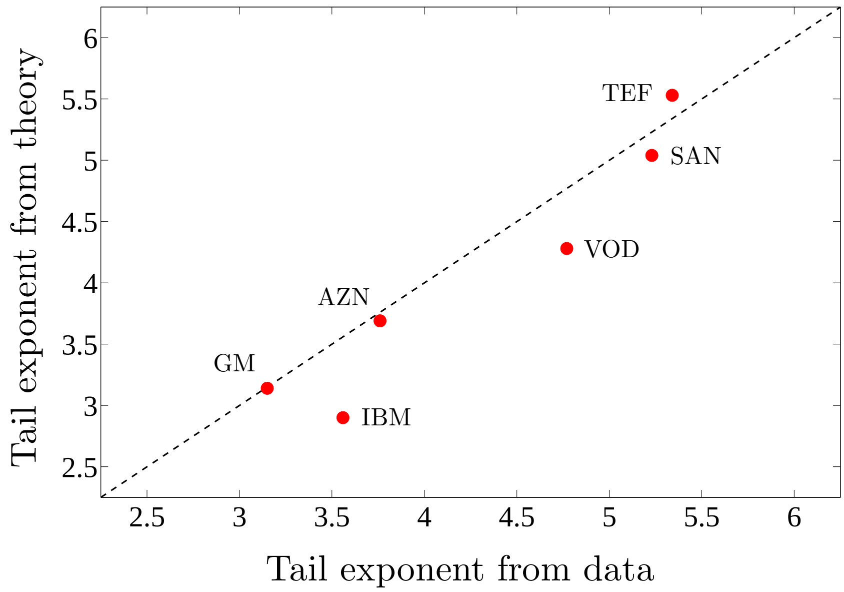

Figure 3.9: **Predicted vs. empirical tail exponent for the stocks under study.** The tail exponent is the asymptotic slope of the tail of the ccd when measured in logarithmic coordinates. The dashed line shows $y = x$ for comparison only.

In Fig.(3.7), I show the time collapse of the complementary cumulative distribution (ccd) of absolute scaled returns, $C(|r'|)$ with $r' = r\sqrt{2\beta_0/(n\tau)}$, where n and β_0 are the parameters of the fitted gama distribution, for the stock IBM (the ccd is the integral of the probability function). The ccd is plotted for $\tau = 10$ to $\tau = 640$, which is up to one trading day for the stocks studied in this section. I show this plot in logarithmic coordinates to focus on the tails of the distribution, and I overlay the plot with the ccd of the theoretical distribution. It is demonstrated that the model matches the data well and the shape of the distribution is stable over these time scales. The parameters n and β_0 are determined using a maximum likelihood fit of β to a gamma distribution, where β is measured once per day. In the inset of this figure, I show the ccd of β compared to the fit. Although not shown, these plots are very similar for the others stocks in this section.

The theoretical model presented in this Chapter suggests that the functional form of the return distribution for stocks is universal, and that the differences are due to the particular properties of volatility for each stock. This is verified in Fig.(3.8), where I show the collapse for all stocks using the following functional transformation, derived from the analytical results presented above

$$f(r') = \left[\Lambda P(r', \tau)\right]^{\frac{2}{n+1}}, \tag{3.22}$$

where $\Lambda = \sqrt{2\pi}\frac{\Gamma[n/2]}{\Gamma[(n+1)/2]}$. Notice that Fig.(3.8) shows not only the universality of the shape of the distribution but also the normal transport explicitly suggested by the theory and observed in Fig.(3.7).

Finally, in Fig.(3.9), I focus on the probability of large returns and compare the tail of the observed distribution to that of the predicted distribution for each stock. For this figure, I measure the slope of the tail of the empirical ccd (in logarithmic coordinates) for $\tau = 80$ using the Hill estimator on the largest five percent of the data. This is compared with the slope of the tail from the predicted distribution in the same region. The measured values are in good agreement with theoretical predictions, showing a pronounced variation across stocks that is explained by the model. This indicates that the likelihood of extreme price movements is determined by the parameters obtained from fitting β to a gamma distribution for each stock.

3.5 Conclusions

I have presented a new theoretical explanation to returns distribution based on Superstatistics which distinguishes between two different dynamics, a slow dynamics and a fast dynamics, and two time scales. Slow dynamics is this describing the volatility which is positively autocorrelated and taken as constant at intraday time scales, fast dynamics is that of the prices which are almost uncorrelated. This differentiation of dynamics with different time scales has been succesfully employed in the description of non-linear and non-equilibrium systems.

The presented model implicitly takes into account two very well-known stylized facts of financial time series: the absence of autocorrelation of returns, and the positive autocorrelation of absolute returns. The former was postulated by Bachelier and it empirically holds, while the latter means that volatility is a long-memory process. By considering that volatility fluctuations are slow, but significant, it is possible to explain two important characteristics of emprical returns distribution: its non-Gaussian shape, and its apparent stability. These two properties are modeled and understood as a consequence of the properties of volatility distribution which is fitted by a gamma distribution. This is relevant because previous solutions to these problems were not able to explain both simultaneously. It was already shown that SD models were able to describe the apparent stability of the distributions but paying a serious price: stable distributions have infinite variance, and this is against some other paradigms in mathematical finance. Moreover, these models did not converge to a Gaussian for long time intervals as it is observed in empirical distributions. On the other hand, MDH models were theoretically able to explain several stylized facts but specific implementations failed to fully explain the entire empirical distribution of returns. However, MDH framework is not very restrictive, and based on its theoretical foundations different implementations are allowed. The model presented here may be classified into MDH family, but the novelties introduced in this specific theoretical approach let produce very accurate predictions.

The solution derived from the mentioned assumptions and the fitted volatility is a variant of the Student's t-distribution which is fat-tailed but with finite variance. This lets us explain the apparent stability and heavy-tailed behavior of empirical returns distribution. Moreover, the convergence to a Gaussian is an inmediate consequence of CLT. I have also performed tests to check the accuracy of the model, showing an excellent agreement with empirical data. The presented results suggest that stock price fluctuations are universal, and that return distributions for stocks from different exchanges, time periods, over different time scales, and from different economic sectors can be described by one functional form, since the universal behavior of price fluctuations is rooted in the characteristics of the volatility.

Bibliography

[1] S. Abe, and S. Thurner, preprint at cond-mat/0501429.

[2] T. G. Andersen, and T. Bollerslev, Journal of Finance **52**, 975 (1997).

[3] T. G. Andersen, T. Bollerslev, F. X. Diebold, and H. J. Ebens, Financial Economics **61**, 43 (2001).

[4] N. M. P. C. Areal, and S. J. Taylor, Journal of Futures Markets **22**, 627 (2002).

[5] M. Baiesi, M. Paczuski, and A. L. Stella, preprint at cond-mat/0411342.

[6] C. Beck, and E. G. D. Cohen, Physica A **322**, 267 (2003).

[7] C. Beck, Physica A **331**, 173 (2004).

[8] C. Beck, E. G. D. Cohen, and H. L. Swinney, Phys. Rev. E **72**, 056133 (2005).

[9] J.-P. Bouchaud, and M. Potters, *Theory of Financial Risks and Derivative Pricing* (Cambridge Univ. Press, Cambridge, U.K., 2003).

[10] J. Bryant, *A Thermodynamic Approach to Economics* (Butterworthand Co., London, 1982).

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

[12] F. Corsi, G. O. Zumbach, U. A. Müller, and M. M. Dacorogna, Review of Banking, Finance and Monetary Economics **30**, 183 (2001).

[13] R. F. Engle, Econometrica **68**, 1 (2000).

[14] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, Quantitative Finance **4**, 383, (2004).

[15] H. Hasegawa, preprint at cond-mat/0501429.

[16] L. P. Hansen, and R. J. Hodrick, Journal of Political Economy **88**, 829 (1980).

[17] L. Harris, Journal of Financial Economics **16**, 99 (1986).

[18] I. G. Kawaller, P. D. Koch, and T. W. Koch, Journal of Finance **42**, 1309 (1987).

[19] J. H. C. Lisman, *Econometrics, Statistics and Thermodynamics* (The Netherlands Postal and Telecommunications Services, TheHague, Holland, 1949).

[20] S. Micciché, G. Bonanno, F. Lillo, and R. N. Mantegna, Physica A **314**, 756 (2002).

[21] E. Platen, and D. Heath, *A benchmark Approach to Quantitative Finance* (Spriger, Berlin, 2006).

[22] S. Rizzo, and A. Rapisarda, Proceedings of the 8th Experimental Chaos Conference, Florence, AIP Conf. Proc. **742**, 176 (2004).

[23] S. Rizzo, and A. Rapisarda, preprint at cond-mat/0502305.

[24] P. A. Samuelson, Structure of a minimum equilibrium system, *The Collected ScientiÞc Papers of Paul A. Samuelson*, R. W. Pfouts (Editor), (MIT Press, Cambridge, MA, 1960).

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

Chapter 4

Empirical Study of Hidden Orders in Two Markets: LSE, SSE

4.1 Introduction

In this Chapter, I present an empirical study of hidden orders in two markets: the LSE (London Stock Exchange), and the SSE (Spanish Stock Exchange). I call hidden order to a large trading order that is split into pieces and executed incrementally. The reason for executing an order in this way is for reducing transaction costs, because of the tremendous market price impact [1] that an order of that size would cause. Moreover, many times the splitting is simply necessary because there is not enough available liquidity in the order book to fully match a large order. The main problem when studying hidden orders is that they are not specified as a single order, and individual pieces are not reported as a part of a larger order, hidden orders therefore must be inferred from transactions data. For this, it is necessary to know the specific participants involved in a transaction, that's the participant who originally placed the limit order and the participant who transacted the available shares in the order book. This is an additional problem because it is required a data set with the identities of participants and their transactions, and databases with this type of information are not often available to researchers. Data sets employed in this thesis included the corresponding codes of market members involved in every transaction. This information does not let us compute hidden orders in a simple or direct way, because a market member, e.g. a broker, manages orders from several market participants included her own orders. So, we would need a lower disaggregation level to accurately classify hidden orders, that's trading accounts which are not publicly available. Therefore, in the case that we were able to accurately detect a hidden order executed by a market member, we can not certainly know if this is a single hidden order, or the final result of more than one hidden order acting concurrently. Although all the problems related to hidden orders detection make impossible to discern them conclusively, there are statistical algorithms

[22, 27] which make plausible classifications. I employ one of these methods for detecting the hidden orders in both markets. It is out of the scope of this thesis to compare the different algorithms used in the detection of hidden orders.

Statitiscal properties of hidden orders obtained from both markets let us compare the way large orders are executed in two different markets. Consequently, we can extract commonalities about some of their properties, e.g. functional form of their price impact, or trading profile. Moreover, we classify the hidden orders by using a set of criteria related to price impact for studying this conditioned to different factors: total number of single orders, temporal impact, permanent impact, types of single orders composing the hidden orders. These results are relevant because they show for the first time a comparative study of this type of execution in two markets. Although the developement of a theoretical explanation to the empirical findings it is further work, the fact that we observe similar behavior in both the London and Spanish stock exchanges, and that others have also observed this in the New York Stock Exchange, suggests the possibility of a “law" for market impact.

4.2 Hidden Orders

In the Introduction, I have defined hidden orders as large trading orders that are split into pieces and executed incrementally. It is important to make a distinction between iceberg orders [36] that are large orders divided into smaller pieces so that only a small fraction is shown at a time, and hidden orders. An iceberg order is explicitly submitted as a whole order at a price, although only a fraction is shown at a time. So, there is no doubt about classifying this type of orders. However a hidden order is executed incrementally, this means that may be transacted at several prices and time of execution can span days or even weeks. Finally, an important characteristic for our study is that it is not revealed as a single order at any time, nor when placed in the order book neither when reported after being fully executed. This is relevant because it arises the question of the existence of hidden orders, and the answer is based on strategic reasons [15]. Let’s consider an example, a trader wants to execute a large buy order - size can be several times daily transacted volume,- she attempts to keep the true size of the order in secret to minimize transaction cost. She of course wishes to buy her shares at the lowest price possible. For minimizing the final price she has to take into account two factors. First, anytime she buys she will push the price up. Second, if other market participants know beforehand the size and sign (buy or sell) of the transaction they can modify current and future orders for taking advantage of this information. In addition to strategic reasons, there are studies of hidden orders executed by a specific financial institution [37] who made the data available to the authors of the paper.

Hidden orders are employed by market participants and placed through market members for minimizing price impact and not making public the intention of executing a large transaction. The problem for studying this type of orders comes from their detection, because they are not specified at any time and must be inferred. The aim of the empirical research presented here is the determination of the functional form of the impact of transactions on stock price, also known as market price impact, of hidden orders. The functional form is important [1] to quantify total market impact of a large order, and market impact is the main factor of transaction costs. Therefore the functional form of price impact is a crucial element of any optimized execution. Moreover since impact is a cost of trading, it exerts selection pressure against a fund becoming too large, and therefore is potentially important in determining the size distribution of funds [2, 3]. Finally, market impact reflects the shape of excess demand, which is of central importance in economics. Despite its conceptual and practical importance, a proper empirical characterization and theoretical understanding of market impact is still lacking [4].

4.2.1 Data

Databases employed in this part of the thesis contain the on-book (SETS) market transactions of the London Stock Exchange (LSE) from January 2002 to December 2004 and the electronic open-book market (SIBE) of the Spanish Stock Exchange (BME, Bolsas y Mercados Españoles) from January 2001 to December 2004. Roughly 62% of the transactions at the LSE are executed in the open book market and roughly 90% of the transactions at the BME are executed in the electronic market.

We have initially considered a subset consisting of the most heavily traded stocks in the two markets, 74 stocks traded in the LSE and 23 stocks traded in the BME. For both markets we have considered exchange members who made at least one trade per day for at least 200 trading days per year and with a minimum of 1000 transactions per year. This filter yielded approximately 60 exchange member firms per stock. We then applied the algorithm for detecting hidden orders described in [27], which we have already discussed, to identify hidden orders that consist of at least ten transactions. It is worth noting that the detected patches are not necessarily composed of the same type of trades (buy or sell) but that at least 75% of the transacted volume in the patch must have the same sign. The algorithm detected 90,393 hidden orders in the LSE and 55,309 in the BME.

This study is based entirely on trades that take place through a continuous double auction. “Continuous" refers to the fact that trading takes places continuously and asynchronously, and “double" to the fact that both buyers and sellers are allowed to place and cancel orders at any time. There are two fundamentally different ways to execute an order in such a market. One is to use a limit order, in which an order is placed inside the order book, which is essentially a list of unexecuted orders at different prices. The other is to place a market order, which we define as any order that results in an immediate transaction. Every transaction involves a market order transacting against a limit order. A given real order might act as both, e.g. part of it might result in an immediate transaction and part of it might be left in the order book. We only consider transactions, so in the example above we would treat the first part as a market order and treat the second part as a limit order, but the second part will enter our analysis only if it eventually results in a transaction. The LSE database allows us to identify whether the initiator of the transaction was the buyer or the seller. For BME this information is not available and we infer it with the Lee and Ready algorithm [29]

4.2.2 Detection Algorithm

There are several algorithms in the literature which have been employed in the detection of hidden orders [22, 28]. All these algorithms share the idea of detecting the splitting of large orders based on the signs - buy or sell - of the transacted orders from each brokerage firm. The algorithm employed in this thesis has been previously applied to probe the temporal organization of heteroginieties in human heartbeat interval time series [28]. Although, the aim of the present study and the former one are obviously different in many aspects, the series of the inventory of different broker firms in the market are a perfect field for this segmentation algorithm. The following procedure is applied to temporal time series of the inventory of the broker firms, market members, who are actively trading a stock and the same procedure is applied to all stocks in both markets.

The computational description of the algorithm can give us a deeper understanding of it. The first step is to compute a inventory series for each member. For this we take into account all the transactions with their corresponding sign. Then, we move a sliding pointer from left to right along the inventory series measured in the currency stocks are quoted, brithish pounds for the LSE stocks, and euros for the SSE stocks. At each position of the pointer, we compute the mean of the subset of the signal to the left of the pointer (μ_{left}) and to the right (μ_{right}). To measure the significance of the difference between μ_{left} and μ_{right}, we compute the statistic $t = |(\mu_{left} - \mu_{right})/s_D|$, where s_D is the pooled variance for unequal sample sizes. We next determine the position of the pointer for which t reaches its maximum value, t_{max}, and compute the statistical significance of t_{max}. When we segment an inventory time series we are computing the partition of a non stationary time series, which is composed of many segments with different mean value, in such a way as to maximize the difference in the mean values between adjacent segments.

After segmenting the series we only take into account directional patches, this is because we are only interested in patches which show a clear buy or sell pattern. Therefore, when in a patch the traded volume of a sign is larger than a certain threshold - this magnitude is arbitrary, in this case the value is 75% - the patch is considered of that sign, e.g. $V_b/V > \theta$ is a buy patch, where V_b is positive volume (buy), V is the total traded volume, and θ is the threshold. If $V_s/V > \theta$ is a sell patch, where V_s is negative volume (sell). Total traded volume must hold $V = V_b + V_s$. I have performed several tests for different values of the threshold and results of the algorithm were not dramatically sensitive to its variations.

4.2.3 Classification of Hidden Orders

In the study of hidden orders there is no single variable which can employed for understanding all their characteristics and behavior. For this reason, we define a set of variables which are useful for discovering their extensive properties, and for undesrtanding the interplay between hidden orders and the other transacted orders in the market. Thus, we characterize hidden orders as a function of several variables. These are

- The execution time T (in seconds) of the hidden order, measured as the trading time interval between the first and the last transaction of the hidden order.

- The number N of transactions of the hidden order. We consider hidden orders of length $N > 10$.

- The volume V of the hidden order defined as

$$V = \sum_{j=1}^{N} v_j, \tag{4.1}$$

  where v_j is the signed volume of each transaction of the hidden order. For buy trades $v_i > 0$ and for sell trades $v_i < 0$. We consider the hidden order to be a buy order if $V > 0$ and a sell order if $V < 0$. The buying/selling nature of a hidden order is thus encoded in its sign, $\epsilon = \text{sign}(V)$. The volume is the product of the number of shares times the price and is measured in Pounds (LSE) or in Euro (BME).

- The volume fraction of market orders f_{mo}. A hidden order can be implemented with very different liquidity strategies, i.e. with different compositions of market and limit orders. In order to quantify this we define the fraction (in volume) of market orders within a hidden order as

$$f_{mo} = \frac{\sum_{j=1}^{N} |v_{j,mo}|}{\sum_{j=1}^{N} |v_j|}, \tag{4.2}$$

  where $v_{j,mo}$ is the traded volume at each transaction done through market orders. Values of f_{mo} close to zero mean that the broker completed the hidden order by using mainly limit orders, while values of f_{mo} close to one imply the broker used mainly market orders during the execution of the hidden order.

- The participation rate α of a hidden order defined as

$$\alpha = \frac{\sum_{i=1}^{N} |v_i|}{V_M}, \tag{4.3}$$

where V_M is the unsigned volume of the stock traded in the market concurrently with the hidden order. Values of α close to zero imply the hidden order was negligible compared to the activity in the market, while values of α close to one mean that most of the activity in the market came from the transactions of the hidden order.

First three variables are extensive, and are strongly depending on the original size of the order. Last two variables are not strictly depending on the size of the order, but on the style of execution. Trader can make a choice on the aggressiveness of the execution by playing with f_{mo} and α variables. A f_{mo} close to 1 means that hidden order is taking the available liquidity in the order book, and it is considered an aggressive execution. An explanation why it is considered an aggressive strategy is that market orders are related to immediacy, if we are not concerned with immediate execution it makes sense to place limit orders wich give the opportunity to improve best prices: bid, and ask. The other factor related to aggressiveness is α, when this is close to 1 means that we are present in most of the transactions at that time in the market. Given that the algorithm employed for detecting hidden orders has a threshold - mentioned above,- the hidden order is clearly pushing prices in one direction.

These variables have been chosen because it is expected that the market impact of a hidden order can be described as a function

$$r = f(N, V, T, f_{mo}), \tag{4.4}$$

plus possibly other variables specific of the stock, such as the participation rate, the capitalization, the volatility, or the spread.

Note that in all the analyses and figures we compute error bars as standard errors. It should be born in mind that this procedure underestimates the errors due to the heavy tails of the fluctuations and due to possible long-memory properties of the data.

4.2.4 Statistical Properties of Hidden Orders

I investigate the statistical properties of the variables characterizing hidden orders. In [27] it was considered a set of 3 most capitalized stocks traded at the BME and studied the probability distribution of the variables characterizing the hidden orders and the scaling relations between these variables.

In [27] no restriction on the length or on the fraction of market orders was set on the hidden orders and the authors found that the distribution of hidden order size is fat tailed and consistent with a distribution with infinite variance. They also showed that this broad distribution is due to an heterogeneity of

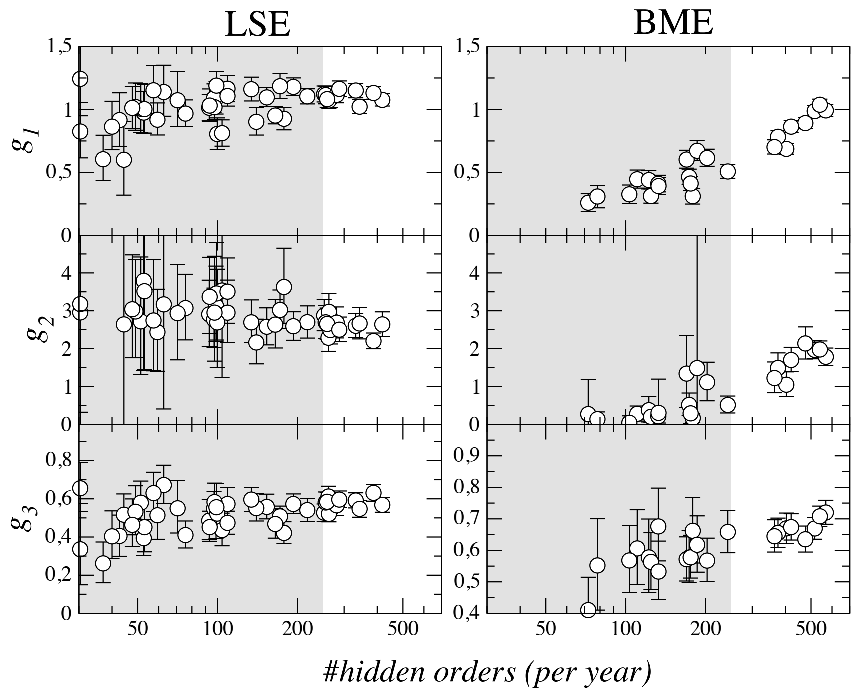

Figure 4.1: **Exponents g_i ($i = 1,2,3$) of the allometric relations of Eq. 4.5**. For each of the stocks considered in our LSE and BME databases and for hidden orders with $N \geq 10$ and $T < 1$ day, as a function of the number of detected hidden orders per year. Error bars are 95% confidence intervals obtained by bootstrapping the data. In the analysis of market impact we consider only stocks with at least 250 hidden orders per year (those in the white area of the figure).

scales among different brokerage firms rather than to the heterogeneity of scales within the hidden orders of each brokerage firm. By using Principal Component Analysis (PCA) on the logarithm of the variables characterizing the hidden orders, it was found that N, V and T are related through scaling relationships

$$N \sim V^{g_1}, \quad T \sim V^{g_2}, \quad N \sim T^{g_3}, \tag{4.5}$$

where $g_1 \simeq 1$, $g_2 \simeq 2$ and $g_3 \simeq 0.66$ for 3 highly capitalized stocks in the BME and including all hidden orders. I repeat the two dimensional PCA analysis of [27] on our much larger data set.

Fig. (4.1) shows the value of the three exponents for all the stocks as a function of the number of hidden orders per year. We observe that for stocks with a small number of hidden orders the heterogeneity in the value of the exponents is pretty large, while, as the number of hidden orders detected by the algorithm increases, the exponent estimations become less noisy and tend to converge to similar values. Moreover for BME stocks there is a clear trend of the exponents as a function of the number of hidden orders.

In order to measure market impact in a statistically reliable way, we pool together data from different stocks. We need therefore an homogeneous sample of stocks. To this end in the following analysis we restrict our dataset to those stocks for which our algorithm detects at least 250 orders per year.

These stocks are TEF, SAN, BBVA (as in [27]) but also REP, ELE, IBE, POP and ALT for the BME market and AZN, BSY, CCH, DVR, GUS, KEL, PO, PSON, SIG, TATE and TSCO for the LSE market. Moreover, in this thesis we will focus mainly on short hidden orders, considering the set of hidden orders of time duration T smaller than one trading day. The reason for this choice, detailed below, is to obtain stable statistical averages for the market impact. Applying these two restrictions, we obtain a final dataset that contains 14,655 hidden orders in the BME and 11,165 orders for the LSE (see Table 4.1).

Table 4.1: Statistics of the hidden order ensembles used in the paper. Only hidden orders with $T < 1$ day and $N > 10$ transactions are used.

Market	# orders	$\langle N\rangle$	$\langle f_{mo}\rangle$	$\langle\alpha\rangle$	$\langle R\rangle$	$\langle R\rangle_{f_{mo}>0.8}$
BME	14,655	95.58	0.52	0.17	1.127	3.983
LSE	11,165	97.53	0.53	0.34	0.587	2.156

We repeat the two-dimensional PCA analysis of [27] on the pooled set of hidden orders from different stocks. We find for the BME market the following exponents

$$g_1 = 0.81 \ (0.79; 0.82), \tag{4.6}$$
$$g_2 = 1.57 \ (1.43; 1.72), \tag{4.7}$$
$$g_3 = 0.67 \ (0.65; 0.68), \tag{4.8}$$

where quantities in parenthesis are 95% confidence intervals obtained through bootstrapping the data. These relations explains 83%, 61% and 80%, respectively, of the variance observed in the data. For the LSE dataset we get

$$g_1 = 0.99 \ (0.98; 1.01), \tag{4.9}$$
$$g_2 = 2.41 \ (2.29; 2.52), \tag{4.10}$$
$$g_3 = 0.58 \ (0.57; 0.59), \tag{4.11}$$

and these relations explain 88%, 75% and 86%, respectively, of the variance. These allometric relations are roughly consistent with those obtained in [27].

The left panels of Fig. (4.2) show the probability density function of f_{mo} and of the participation rate α. We observe that the distribution of the fraction of market orders is rather broad and is roughly centered around $f_{mo} = 0.5$.

In addition two peaks are observed for $f_{mo} \simeq 0$ and $f_{mo} \simeq 1$. For the BME the participation rate has a peak around $\alpha = 5\%$, while for the LSE the distribution of α is broader, and peaks at a value closer to 20%. This value is pretty large and we

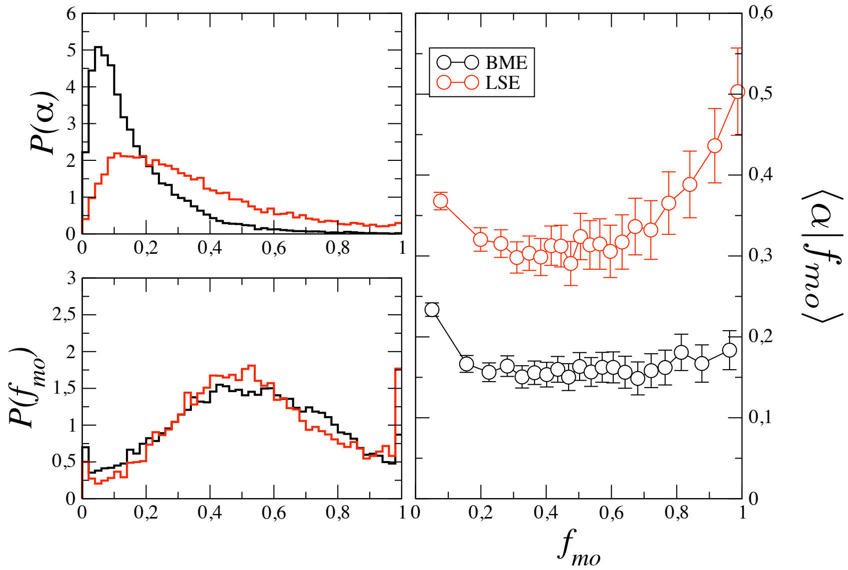

Figure 4.2: **Ensemble statistics of the fraction of market orders f_{mo} and participation rate α of hidden orders in both the BME and LSE.** Left panels show the probability distribution function of both parameters, while the right panel shows the conditional average of the participation rate conditioned on a given value of f_{mo}.

do not have an explanation for the difference in the participation rate between the two markets.

Finally, the two parameters α and f_{mo} are not independent. Fig. (4.2) shows the expected value of α conditioned on f_{mo}. For the BME the expected value of α is almost constant except for very small values of f_{mo}. In contrast, for the LSE the dependence is much stronger. The participation rate is higher when f_{mo} is at either of its extremes.

The bottom left panel of Fig.(4.2) shows that f_{mo} has a broad distribution. Hidden orders can therefore differ a lot in terms of the fraction of market orders used to complete them.

In the investigation of the market impact of hidden orders we will consider hidden orders characterized by a restricted set of values of f_{mo} to better characterize their profile with respect to the fraction of market orders used to complete the hidden order. Specifically we will use $f_{mo} > 0.8$ (large fraction of market orders used) and $f_{mo} < 0.2$ (large fraction of limit orders used). The reasons why we expect this distinction to be critically important will be described in the next section.

4.3 Market Impact

Market impact is the expected price change conditioned on initiating a trade of a given size and a given sign. One naturally expects that initiating a buy order should drive the price up, and initiating a sell order should drive it down. This has roots in standard economic theory: An increase in demand should increase prices, while an increase in supply should decrease prices.

Market impact is important for theoretical and practical reasons. First of all, in order to be able to estimate transaction costs, and in order to optimize a trading strategy to minimize such costs, it is necessary to understand the functional form of market impact [1].

Moreover since impact is a cost of trading, it exerts selection pressure against a fund becoming too large, and therefore is potentially important in determining the size distribution of funds [2, 3]. Finally, market impact reflects the shape of excess demand, which is of central importance in economics. Despite its conceptual and practical importance, a proper empirical characterization and theoretical understanding of market impact is still lacking [4].

The functional forms of market impact vary from study to study. This is in part because there are several different types of market impact that must be distinguished. Studies of individual transaction impact yield a strongly concave functional form which appears to vary from market to market [5, 6, 7, 8]. Other studies have looked at market impact under aggregation, in which the impact is conditioned on the sum of the signed transaction volume associated with a given number of trades or a given interval of time. These studies have tended to observe a somewhat less strongly concave shape [4, 11, 12, 13, 9, 10]. Other research has focused on orders executed through specific mechanisms, e.g. block markets [14].

In this thesis, I instead focus on the impact of hidden orders. In a model by Kyle [15], a linear dependence of impact on trading volume is theoretically predicted. More recent theoretical approaches[16, 17, 18, 13, 19, 20, 21, 22, 4, 23] have proposed different functional forms for the market impact of hidden orders.

4.3.1 Definition

The main focus of this Chapter is the empirical measurement of the market impact of hidden orders. Given a hidden order traded on stock i between times t and $t+T$, we measure the market impact by considering the change in the log price of the stock between time t and time $t+T$, i.e.

$$r_i(t,T) = \log p_{i,t+T} - \log p_{i,t}, \tag{4.12}$$

where $p_{i,t}$ is the price of stock i at time t. We have used for p_i the midprice, but our results do not depend on this. Our objective is to study how $r_i(t,T)$ changes as a function of the main properties of the hidden order.

Different stocks have different scales of their price fluctuations. In order to be able to take the average of market impact across different stocks, we rescale it by dividing by the mean value of the spread s_i of the stock during the year, where the spread is the difference between the lowest selling price (ask) and the highest buying price (bid).

Specifically, we define the rescaled market impact as

$$R_i(t,T) = \epsilon_i r_i(t,T)/s_i, \tag{4.13}$$

where, as before, $\epsilon_i = +1$ for a buy hidden order and $\epsilon_i = -1$ for a sell hidden order.

Although we observe a small asymmetry between the market impact of buy vs. sell orders, similar to that observed elsewhere [31], for the purpose of our study here we lump together buy and sell hidden orders in order to obtain better statistics.

4.3.2 The Noisy Nature of Market Impact

While a given hidden order is trading there are typically many other orders trading at the same time, as well as news arrival, and thus there is a considerable amount of noise in the price change associated with any particular hidden order. The price change associated with a hidden order functionally depends on several factors, which can be written

$$r_i(t,T) = \mathscr{R}[r_M(t,T), \rho_i(t,T), \eta_i(t,T)], \tag{4.14}$$

where r_M corresponds to market-wide movements [25], ρ is the average market impact of the hidden order, and η_i is the background uncorrelated noise coming from the trading of the rest of the market [12].

While the background noise can be controlled by taking averages over different orders with the same properties or restricting our analysis to very small values of T, market-wide movements remain large, especially for large values of T. During the years 2001-2004 stock markets were in a substantial decline for more than two years, only recovering at the end of 2003 and 2004 (see the inset of Fig.(4.3)).

Fig.(4.3) shows the conditional average $\langle R|T\rangle$ of the rescaled market impact of the hidden orders as a function of their time duration T. We observe that for T larger than one day, rescaled impact is on average negative, irrespectively of the sign of the hidden order. The reason for this phenomenon is that market-wide movements were mostly negative for values of T larger than one day. Only for hidden orders of duration close to or below one day do we observe negligible changes in market indexes when compared to price changes during hidden order completion. This is the motivation of our choice of restricting our study to hidden orders of duration T less or equal to one day[1].

[1]Other authors [25] have proposed to use industrial sector indexes as proxies for market-wide

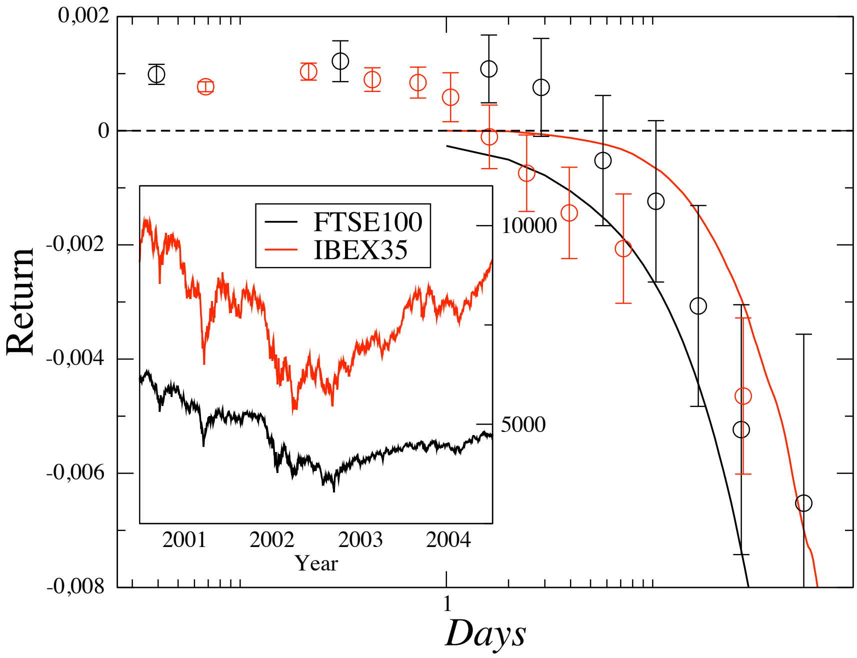

Figure 4.3: **Conditional average $\langle R|T\rangle$ of the rescaled impact of hidden orders** (Eq. (4.13)) as a function of their time duration T (symbols) compared to the average return of the stock market index over random periods of the same time duration (solid lines). The inset shows the price of the FTSE100 and IBEX35 indices over the period of study. In this figure we are using all detected hidden orders without any conditioning on T or f_{mo} values but with $N > 10$.

4.3.3 Impact of Limit Orders vs. Market Orders

It is important to stress that market impact comes about through changes in supply and demand, and that this causes a strong *a priori* difference in the impact one expects to observe in the execution of a limit order vs. a market order. For example consider buy orders. A buy market order reflects an increase in demand at the current price. If sufficiently large it will cause a positive price change. Since in a continuous double auction market orders always execute against limit orders, this implies that the sell limit order that the buy market order executes against will generate a positive market impact. We therefore expect that executed limit orders have the opposite impact of market orders: Buying drives the price down and selling drives it up.

The problem with this line of reasoning is that we are considering only *executed* limit orders, which creates a strong selection bias. To measure the impact

movements of a given stock and thus the study can be extended to larger values of T. We do not follow this procedure due to lack of that information.

of limit orders correctly we need to condition on all orders that are placed, rather than only on those that are executed. When this is done the impacts for limit orders should be roughly the same as for market orders, as otherwise it would be possible to make a profit by simply using limit orders instead of market orders.

If a buy limit order is placed below the current price it is executed only if the price drops. The probability of execution of a limit order depends on future price movements: under an adverse price movement the probability of execution is higher than for a favorable price movement. This is caused in part by the mechanical dynamics of a random walk, but also by asymmetric information: Placing a limit order gives others the option of executing at their will, when they have information that indicates it is favorable to do so. This phenomenon is called *adverse information*. When this is properly taken into account, limit orders have impact in the direction one would expect, i.e. buying has positive impact and selling has negative impact [32, 33]. Furthermore the magnitude of the impact of limit orders when the selection effects are properly taken into account is comparable to that of market orders.

For the BME we have a record of transactions but not of orders. Thus to measure market impact and avoid the selection bias associated with executed limit orders we are forced to use only those hidden orders that are predominantly built out of market orders. For consistency we analyze both the BME and the LSE data in the same way.

In Table (4.1) we show the mean value of the rescaled market impact R of Eq. (4.13) for hidden orders of duration less than one day. We also show the mean value $\langle R\rangle_{f_{mo}>0.8}$ of the rescaled market impact computed over the set of hidden orders with a large fraction of market orders ($f_{mo} > 0.8$). $\langle R\rangle_{f_{mo}>0.8}$ is significantly larger than $\langle R\rangle$ indicating that hidden orders mainly composed by market orders have on average a larger market impact than hidden orders composed of both limit and market orders.

4.3.4 Impact vs. N

Fig.(4.4) shows the average over all hidden orders of the rescaled market impact $\langle R|N\rangle$ as a function of the conditioning variable N. This grows slightly as a function of N, but one must keep in mind that the meaning of this is difficult to interpret in view of the discussion above, since we are averaging together a roughly equal number of market orders and executed limit orders.

To investigate the average market impact and minimize the effect of the selection bias, we divide the data into two groups: liquidity providing hidden orders, with $f_{mo} < 0.2$, and liquidity demanding hidden orders, with $f_{mo} > 0.8$. As expected, for the former group the market impact is on average negative, while for the latter it is positive. Using ordinary least squares, we find that for both groups the dependence of $\langle R|N\rangle$ on N is well described by the power law

$$|\langle R|N\rangle| = A\, N^{\gamma}. \tag{4.15}$$

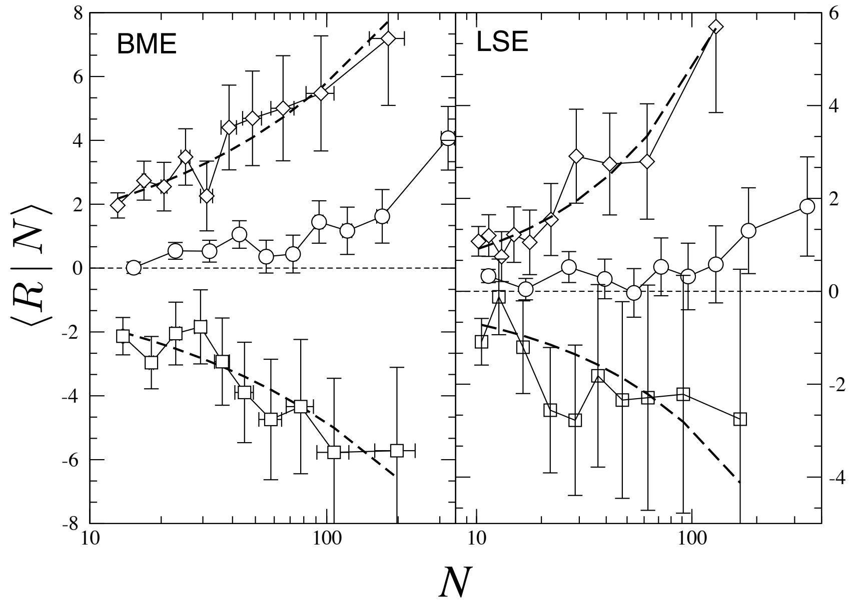

Figure 4.4: **Average rescaled market impact R for hidden orders shorter than 1 day as a function of N for the BME (left) and LSE (right).** Circles are the results for all hidden orders, while squares are the results when there is a low fraction of market orders ($f_{mo} < 0.2$) and diamonds are for when there is a large fraction of market orders ($f_{mo} > 0.8$). Dashed lines are power law fits $R \sim N^{\gamma}$. Values of γ are reported in Table (4.2).

The estimated parameters are in Table 4.2.

In summary, we find that the market impact of hidden orders dominated by market orders is consistent with

$$\langle r|N\rangle \propto \epsilon s N^{\gamma} \tag{4.16}$$

where ϵ is the sign of the order and s is the spread. For hidden orders dominated by limit orders the market impact is very similar to minus the impact of hidden orders dominated by market orders.

For the BME the exponent is consistent with a square root function while for the LSE the exponent of the impact is slightly larger than 0.5. The square root dependence is consistent with other studies and with the predictions of some models. The BARRA model [24] uses an exponent 0.5 for estimating market impact. Almgren et al. [26] found an exponent approximately equal to 0.6 for the temporary impact of hidden orders. The theories of references [13] and [23] predict that the exponent of the impact should be roughly 0.5, with the exact value depending on the heavy tail of the volume distribution.

In Eq.(4.16) the spread gives the proportionality constant, i.e. the global scale of the impact. By using the results of Wyart et al. [34], who derive a propor-

Table 4.2: Parameters of the fitting of the market impact with Eq. 4.15.

Market	$A_{f_{mo}>0.8}$	$\gamma_{f_{mo}>0.8}$	$A_{f_{mo}<0.2}$	$\gamma_{f_{mo}<0.2}$
BME	0.63 ± 0.17	0.48 ± 0.07	-0.63 ± 0.22	0.44 ± 0.09
LSE	0.17 ± 0.05	0.72 ± 0.10	-0.16 ± 0.14	0.64 ± 0.30

tionality between the spread and the volatility per trade, it is possible to rewrite Eq.(4.16) in such a way that the proportionality constant is the volatility per trade.

4.3.5 Temporary vs. Permanent Impact

Finally we study how market impact builds as hidden orders are executed and how the price reverts when the execution is completed.

Here we show the impact as a function of time rather than of executed trades. In order to consider hidden orders of different length we normalize the time by dividing it by the execution time T. With our normalized time, the initial time of the order corresponds to $t/T = 0$ while the final time is $t/T = 1$. We consider only orders with $f_{mo} > 0.8$.

The results are shown in Figure 4.5 where we see that earlier transactions within the hidden order have more impact than later ones. In fact we observe that

$$R \sim (4.28 \pm 0.21) \times \left(\frac{t}{T}\right)^{0.71 \pm 0.03} \qquad (BME) \tag{4.17}$$

$$R \sim (2.13 \pm 0.05) \times \left(\frac{t}{T}\right)^{0.62 \pm 0.02} \qquad (LSE) \tag{4.18}$$

We also observe that after the completion of the hidden order price drops, suggesting that not all of the market impact is permanent. The tendency for reversion has also been observed previously [14, 22, 25]. We compare the impact at its peak when the order has just finished, $R_{temp} = R(t = T)$ (see Table 1), to the permanent impact $R_{perm} = R(t \gg T)$ estimated by averaging over a period $1.5 \leq t/T \leq 3$. The drop in impact is $R_{perm}/R_{temp} \simeq 0.51 \pm 0.22$ (BME) and $R_{perm}/R_{temp} \simeq 0.73 \pm 0.18$ (LSE).

In Fig.(4.5), the symbols are the average value of the market impact of the hidden order as a function of the normalized time to completion t/T. The rescaled time $t/T = 0$ corresponds to the starting point of the hidden order, while $t/T = 1$ corresponds to the end of the hidden order. The rescaled time t/T is extended up to $t/T = 3$ to study the permanent and temporal impact of the hidden order. Solid lines are power-law fits (see text) while dashed lines correspond to temporary (upper) and permanent (lower) market impact. Temporary impact R_{temp} is measured at the end of the order $t = T$ (see Table 4.1), while permanent impact R_{perm} is obtained through an average of the $R(t/T)$ with $1.5 \leq t/T \leq 3$ obtaining $R_{perm} = 2.03 \pm 0.68$ for the BME and $R_{perm} = 1.48 \pm 0.06$ for the LSE. Data is only for hidden orders with $f_{mo} > 0.8$.

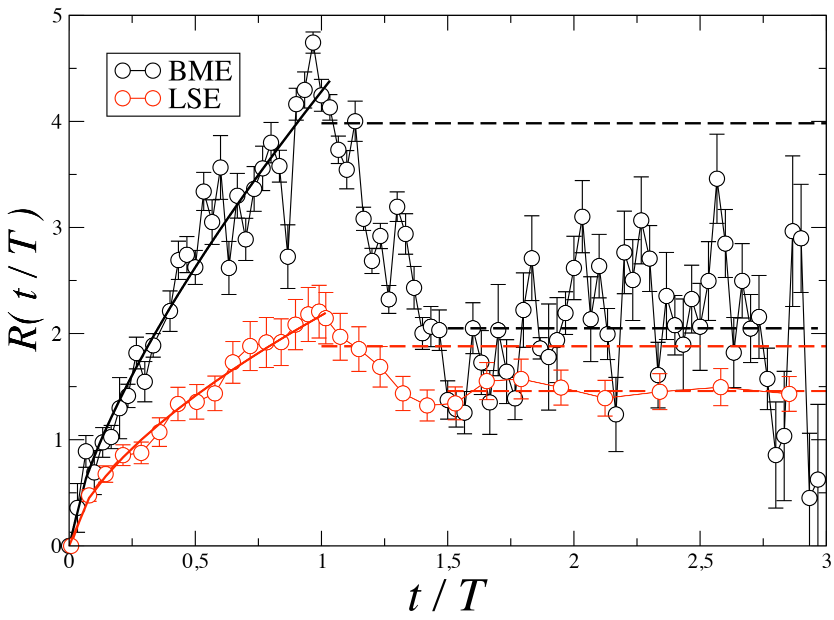

Figure 4.5: **Market impact versus time.**

It is interesting to note that the square root impact and the reversion can be predicted through a simple argument based on the hypothesis that the price after reversion is equal to the average price paid during execution [23]. If during execution price impact grows like $A \times (t/T)^{\beta}$ then the average price payed by the agent who executes the order is

$$\langle p \rangle = p_t + A \int_0^1 (t/T)^{\beta} dt = p_t + \frac{A}{1+\beta}, \tag{4.19}$$

i.e. the permanent impact is $1/(\beta+1)$ of the peak impact. In our case and using the exponents β obtained in figure 4.5 we get $1/(\beta+1) \simeq 0.58 \pm 0.01$ for the BME and $1/(\beta+1) \simeq 0.62 \pm 0.02$ for the LSE which are statistically similar to the ratios R_{perm}/R_{temp} for each market.

4.4 Trading Profile

In this section we investigate how hidden orders are executed as a function of time, which we call the trading profile. By this we mean the traded volume of the hidden order as a function of time elapsed from the time of the first trade.

As before, in order to average across orders of different length we use the normalized time t/T. We measure the normalized average volume of each transaction $v_i/\langle v_i \rangle$ traded at time t inside the hidden order as a function of the normalized time. Here $\langle v_i \rangle$ is the average volume exchanged in the individual transactions used to execute the hidden order.

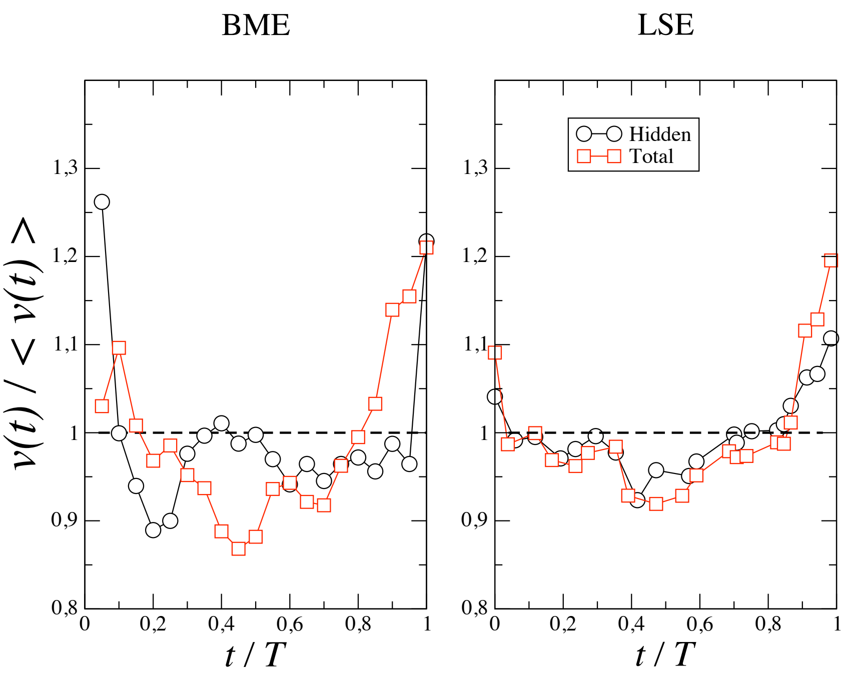

Figure 4.6: **Trading profile inside the hidden order.** Average volume of the transactions within the hidden order divided by the average volume in the hidden order as a function of the normalized time t/T. Circles are the results for all hidden orders, while squares are the volume traded in the market (in the same stock) concurrently with the hidden order. Data is only for hidden orders with $f_{mo} > 0.8$.

Fig.(4.6) shows that trading within the hidden order is fairly homogeneous except for the initial and final times of the order, for which there is a small increase in the traded volume. This can be understood if we look at the concurrent trading in the market.

In Fig.(4.6) we see that the profile of the hidden order substantially matches the concurrent trading in the market. In fact, the rise and fall of concurrent trading is a bit stronger than it is for hidden orders. The cause and effect of this phenomenon is not clear: Does trading rise and fall because of the pattern of hidden order placement, or do people placing hidden orders try to match trading volume, e.g. through VWAP (volume weighted average price) strategies.

As shown in Fig.(4.7), the starting and ending of hidden orders is substantially correlated with overall volume of trading in the market. In particular we see that a significant fraction of hidden orders start at the beginning of the day and finish at the end of the day. These are the times of day in which the volume traded is larger, corresponding to the well-known U-shaped volatility and trading volume profile.

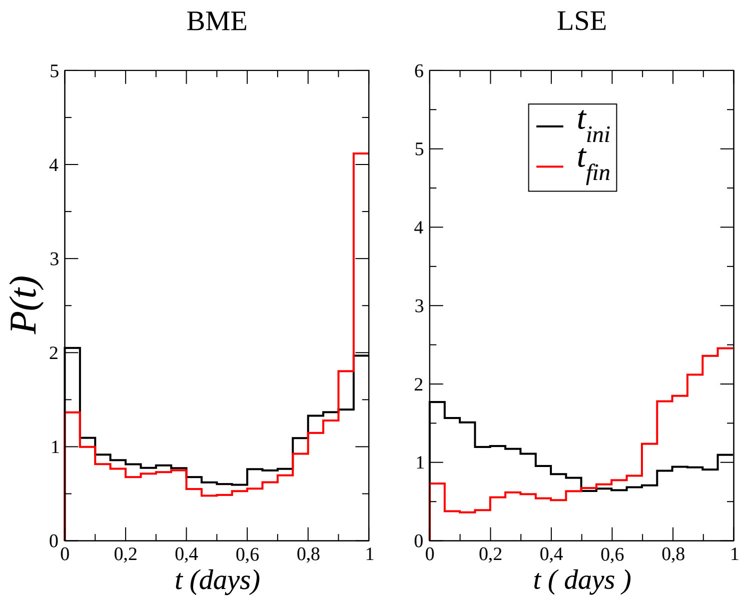

Figure 4.7: **Initial and final times of the hidden orders.** Probability distributions of the initial time t_i and final time t_f of the hidden orders, measured with respect of the time of the day. Data is only for hidden orders with $f_{mo} > 0.8$

Once again the cause and effect is not clear. It has been shown that when impact of a transaction decays gradually with time, the optimal trading strategy has a turnpike shape [35, 21, 4]. Thus this might drive the end of day increase in trading, or alternatively, the daily variation in trading may drive the profile of hidden orders simply due to the desire to match traded volume.

4.5 Conclusions

Large orders are typically not executed in a single transaction because the market impact of such an order would be huge. Beside destabilizing the market, such a large order would leak information about the intention of the agent who placed the order. Therefore it is customary to split the order and to trade it incrementally taking advantage of the available liquidity at every moment in the market. Splitting orders may increase the cost payed to complete the order and, generally, an agent tries to split the order in such a way of minimizing the cost for the execution. Different agents may have different ways to implement optimal trading strategies to minimize their impact and the price paid to process a large order.

In this chapter I have empirically studied the main properties of the impact and of the trading protocol of intradaily hidden orders using a large fraction of either market orders or limit orders. We have found that the temporary impact of hidden orders is concave and roughly described by a square root function of the hidden order size. Moreover the price reverts after the completion of the hidden order in such a way that the permanent impact is equal to roughly 0.5 – 0.7 of the temporary impact.

We have also studied how the order is completed in time and we have shown that more volume of the hidden order is traded at the beginning and at the end of the hidden order. When we take into account that hidden orders are more likely to start at the beginning of a day and are more likely to end near the end of the day, this roughly matches the volume traded in the market.

The fact that we observe similar behavior in both the London and Spanish stock exchanges, and that others have also observed this in the New York Stock Exchange, suggests the possibility of a “law" for market impact. It will be very interesting to see whether this hypothesized law continues to hold up under future studies.

Bibliography

[1] R. Kissell, and M. Glantz, *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk* (American Management Association, 2003).

[2] J. B. Berk, and R. C. Greene, Journal of Political Economy **112**, 1269 (2004).

[3] Y. Schwarzkopf, and J. D. Farmer, preprint at arXiv:0807.3800 (2008).

[4] J.-P. Bouchaud, J. D. Farmer, and F. Lillo, How markets slowly digest changes in supply and demand, *Handbook of Financial Markets: Dynamics and Evolution*, T. Hens, and K.R. Schenk-Hoppé (Editors), (North-Holland, NY, 2009).

[5] F. Lillo, J. D. Farmer, and R. N. Mantegna, Nature **421**, 129 (2003).

[6] M. Potters, J.-P. Bouchaud, Physica A **324**, 133 (2003).

[7] J. D. Farmer, and F. Lillo, Quantitative Finance **314**, 7 (2004).

[8] J. D. Farmer, P. Patelli, and I. Zovko, Proceedings of the National Academy of Sciences of the United States of America **102**, 2254 (2005).

[9] C. Hopman, Quantitative Finance **7**, 37 (2007).

[10] F. Lillo, J.D. Farmer, and A. Gerig. A theory of aggregate market impact (in preparation).

[11] A. Kempf, O. Korn, Journal of Financial Markets **2**, 29 (1999).

[12] V. Plerou, P. Gopikrishnan, X. Gabaix, and H.E. Stanley, Phys. Rev. E **66**, 027104 (2002).

[13] X. Gabaix, P. Gopikrishnan, V. Plerou, and H.E. Stanley, Nature **423**, 267 (2003).

[14] D. B. Keim, and A. Madhavan, Review of Financial Studies **9**, 1 (1996).

[15] A. S. Kyle, Econometrica **53**, 1315 (1985).

[16] D. Bertsimas, and A. W. Lo, Journal of Financial Markets **1**, 1 (1998).

[17] R. Almgren, and N. Chriss, Journal of Risk **3**, 5 (2000).

[18] R. Almgren, Applied Mathematical Finance **10**, 1 (2003).

[19] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, Quantitative Finance **4**, 176 (2004).

[20] G. Huberman, and W. Stanzl, Review of Finance **9**, 165 (2005).

[21] A. A. Obizhaeva, and J. Wang, Optimal Trading Strategy and Supply/Demand Dynamics, AFA 2006 Boston Meetings Paper (2005).

[22] A. A. Gerig, *Theory for Market Impact: How Order Flow Affects Stock Price.* PhD thesis, University of Illinois, 2007.

[23] J.D. Farmer, A. Gerig, F. Lillo, and H. Waelbroeck, Information arbitrage and the rational pricing of market impact, in preparation (2009).

[24] N. Torre. BARRA Market Impact Model Handbook. BARRA Inc., Berkeley (1997).

[25] L. Chan, and J. Lakonishok, Journal of Finance **50**, 1147 (1997).

[26] R. Almgren, C. Thum, H. L. Hauptmann, and H. Li, Risk, July 2005, 57 (2005).

[27] G. Vaglica, F. Lillo, E. Moro, and R. N. Mantegna, Physical Review E **77**, 036110 (2008).

[28] P. Bernaola-Galvan, P. Ch. Ivanov, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).

[29] C. M.C. Lee, and M. J. Ready, J. Finance **46**, 733 (1991).

[30] F. Lillo, E. Moro, G. Vaglica, and R. Mantegna, New Journal of Physics **10**, 043019 (2008)

[31] M. J. Aitken and A. Frino. Asymmetry in Stock Returns Following Block Trades on the ASX: A Note, Abacus, **32** (1996).

[32] N. Zamani, and J.D. Farmer (in preparation).

[33] Z. Eisler, J.-P. Bouchaud, and J. Kockelkoren, The price impact of order book events: market orders, limit orders and cancellations. Preprint at arxiv.org/0904.0900

[34] M. Wyart, J.-P. Bouchaud, J. Kockelkoren, M. Potters, and M. Vettorazzo. Relation between bid-ask spread, impact and volatility in double auction markets. preprint at arXiv:physics/0603084 (2006).

[35] J. Hasbrouck, *Empirical market microstructure: The institutions, economics and econometrics of securities trading,* (Oxford University Press, Oxford, 2007).

[36] A. Esser, and B. Mönch, Finance Research Letters **2**, 68 (2007).

[37] L. Chan, and J. Lakonishok, Journal of Finance **50**, 1147 (1995).

Chapter 5

Summary

In Chapter 0, I introduce the two main problems addressed in this dissertation from a Complexity Science perspective. The first problem is related to asset price dynamics and return distributions. Theoretical explanation and understanding of the underlying process driving price dynamics which finally generates return distributions is the key factor in several financial areas such as asset allocation, option pricing, and risk management because all of them derive their conclusions from the distributional properties of returns. It is broadly known that financial time series exhibit statistical regularities which are similar to those observed in the physics of critical phenomena, more specifically stylized facts may be understood as scaling laws. In spite of the effort made by physicists and mathematicians for more than a century to achieve a theoretical explanation to these empirical findings, no conclusive theoretical model has been yet produced. In this thesis, I am mainly focused on the comprehension of two stylized facts: apparent stability and heavy-tailed behavior of return distributions. For this, I employ a new theoretical approach which has been developed for the description of physical systems: Superstatistics.

The second main problem addressed in this thesis is the experimental determination of the functional form of market price impact of large orders. Large orders are difficult to be transacted at a time because of the impact of such a volume would cause in prices, for this reason they are split and incrementally transacted. This type of orders has not been often studied because to properly classify them it is necessary to know the identity of market participants involved in transactions, and this information is not publicly available. On the other hand, the mere empirical description of the statistical properties of these orders and their impact in asset prices is relevant because market participants employ them strategically to reduce transaction costs which are a key factor in any optimized execution.

In Chapter 1, I have presented the first theoretical attempt to describe asset price dynamics. This model was developed by Bachelier from first principles by setting a series of conditions that an idealistic market and price series should fulfill in order to avoid prices were predictable. In spite of these conditions have been broadly tested, there is no conclusive evidence against them and continue to be assumed by modern theoretical approaches. Based on these postulates and on the random behavior of prices, Bachelier derived with three different mathematical reasonings the shape of price distributions: a Gaussian. Nowadays, we know that he made a mistake in his dissertation and only considered a possible solution when there were others compatible with the theoretical conditions. In addition to this theoretical mistake, Bachelier's solution was not compatible with some basic economic concepts such as negative prices, or returns as absolute price variations instead of relative.

Standard Gaussian model is a refinement of Bachelier's model that solved the problems with economic theory by modelling returns instead of prices. The mathematical solution to returns dynamics was a generalized Wiener process

$$\frac{dS}{S} = \mu dt + \sigma\varepsilon\sqrt{dt}, \tag{5.1}$$

with prices lognormally distributed. Although this last model gave theoretical support to important financial problems such as option pricing, it was not compatible with certain aspects of empirical distributions such as heavy tails.

In Chapter 2, I have presented two families of non-Gaussian models: SD and MDH. Although they are apparently different, SD is a particular case of the most general MDH framework. Both families solved the problem of heavy-tailed behavior with a different set of theoretical assumptions. SD considered that empirical distributions were accurately described by a stable distribution, this challenging solution originally proposed by Mandelbrot created a serious problem because it assumes that empirical return distributions have infinite variance, and this is against economic paradigms, i.e. mean-variance framework. Moreover, stable distributions are self-similar and they keep unchanged under aggregation up to rescaling. This model therefore could not explain the slow convergence to a Gaussian when returns are aggregated.

MDH family was developed as an attempt to explain heavy tails without assuming variance was inifinite. In this case, the basic assumption of the model is that prices evolve at different rates at a fixed time interval. This was formally expressed through the subordination of price variations to a Gaussian. So the density of returns was caused by the distribution of increments of a directing process - originally related to informational arrival flow, - and the Gaussian. This means that returns are normally distributed when conditioned to the directing process, expressed as

$$f(r_t) = \int_{I_t \in R^+} f(r_t|I_t) g(I_t) dI_t. \tag{5.2}$$

MDH framework is not very restrictive about the directing process, and this has caused the development of several models by considering different financial variables for this process, or even by taking alternative distributional shapes for the same financial magnitude. The most relevant directing process for this thesis is that of the volatility because the new model presented in Chapter 3 may be classified within this family. In a SV model, returns are modelled with a simple equation

$$r_t = \mu + \sigma_t u_t. \tag{5.3}$$

The main advantage of SV models is that they give theoretical support for the explanation of several stylized facts, and the main problem is the accuracy of every specific implementation.

In Chapter 3, I have presented a new model for asset return distributions which is within SV family but with certain features common to physical systems studied from a superstatistical perspective. Superstatistics is a branch of statistical mechanics devoted to the study of non-linear and non-equilibrium systems which is characterized by using the superposistion of multiple differing statistical models to explain the non-linearity, so in terms of common statistical ideas this is equivalent to compounding the distributions of random variables, and it may be considered a case of a doubly stochastic model. This framework assumes the existence of a fast dynamics represented by a given stochastic process and a slow dynamics which is the responsible for the parameters of that process. In addition to the two different dynamics, superstatistical models assume the existence of an intensive parameter β and that the system may be described as consisting of many cells with different values of β from cell to cell but with constant value within each cell.

Model explains prices at the finest time level: event time, and for this I have taken midpoint price

$$p_t = \frac{\left(p_{b,t} + p_{a,t}\right)}{2}, \tag{5.4}$$

time is measured in midpoint time, this means that time, t, is updated whenever p_t changes.

As a difference with a standard SV model, volatility is taken as constant within cells of length a day. In other words, we are assuming that volatility can be taken as constant at intraday time intervals and returns corresponding to this time scale are expressed by

$$r_t = \sigma \xi_t. \tag{5.5}$$

We define the intensive parameter β as the inverse squared volatility,

$$\beta \equiv \frac{1}{\sigma^2}, \tag{5.6}$$

and following with basics assumptions of MDH and SV, I have assumed a Gaussian process for conditioned returns. So, for a certain value τ which represents a time interval in a given day characterized for a certain value β, probability distribution of returns is

$$p(r, \tau | \beta) = \sqrt{\frac{\beta}{2\pi\tau}} \exp\left(-\frac{\beta r^2}{2\tau}\right). \tag{5.7}$$

This last equation describes daily returns which can be explained by a Gaussian distribution with constant volatility, but β fluctuates over longer time scales and can be characterized by the probability distribution which is similar across stocks and close to a gamma distribution such as

$$g_{n,\beta_0}(\beta) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right). \tag{5.8}$$

Then a straightforward integration of the conditional probability of returns, $p(r,\tau|\beta)$, and the distribution $g(\beta)$ yield the following for the return distribution

$$P(r,\tau) = \frac{\Gamma\left[\frac{(n+1)}{2}\right]}{\Gamma\left[\frac{n}{2}\right]} \sqrt{\frac{\beta_0}{\pi n \tau}} \left(1 + \frac{\beta_0 r^2}{n\tau}\right)^{-\frac{n+1}{2}}, \tag{5.9}$$

which is a variant of the Student's t-distribution. The non-Gaussian shape of the distribution results from collecting returns from time periods separated by long intervals where β is different. The stability of this shape for short to intermediate τ results from negligible fluctuations of β over these time scales.

This new model gives an explanation to non-Gaussian shape of returns distribution, and its apparent stability. It also shows that the properties of the volatility are the cause for these two empirical findings.

In Chapter 4, I have presented an empirical study about the functional form of market price impact of hidden orders in two different markets: the LSE, and the SSE. I call hidden order to a large trading order that is split into pieces and executed incrementally for reducing transaction costs, otherwise they would cause a tremendous price variation and therefore increase costs. Market impact is the expected price change conditioned on initiating a trade of a given size and a given sign, and I measure it for each stock as

$$r_i(t,T) = \log p_{i,t+T} - \log p_{i,t}, \tag{5.10}$$

given that we have studied a set of stocks and different stocks have different scales of their prices prices fluctuations, I have defined a rescaled market impact as

$$R_i(t,T) = \epsilon_i r_i(t,T)/s_i. \tag{5.11}$$

I have also defined a set of variables for understanding their characteristics and behavior. These variables are the execution time, T, the number of transactions, N, the transacted volume, V, the volume fraction of market orders, f_{mo}, and the participation rate of a hidden order, α. In addition to this, it is expected that market impact of a hidden order can be described as a function

$$r = f(N,V,T,f_{mo}), \tag{5.12}$$

plus possibly other variables specific of the stock, such as the participation rate, the capitalization, the volatility, or the spread.

While a hidden order is trading there are typically more orders trading at the same time, there is therefore a considerable amount of noise in the price change associated with any particular hidden order. The price change associated with a hidden order functionally depends on several factors, which can be written as

$$r_i(t,T) = \mathscr{R}[r_M(t,T), \rho_i(t,T), \eta_i(t,T)], \tag{5.13}$$

background noise can be controlled by taking averages but market-wide movements remain large for large values of T. For this reason, I have limited the analysis to hidden orders of duration close to or below one day where changes in market indexes are negligible compared price changes during hidden order completion.

We find that the market price impact of hidden orders dominated by market orders, for avoiding the problem of adverse selection, is consistent with

$$\langle r|N\rangle \propto \epsilon s N^{\gamma}, \tag{5.14}$$

and for hidden orders dominated by limit orders the market impact is very similar to minus the impact of these dominated by market orders. The exponent is close to 0.5 in both markets what it is consistent with a square root function, as other empirical studies and theoretical models previously had shown.

I also study how market impact builds as hidden orders are executed and how the price reverts when the execution is completed. For this, I continue to consider only hidden orders spanning at most one day and mainly executed by mean of market orders. In both markets, we observe a similar behavior where earlier transactions within the hidden order have more impact than the latter ones and after the completion of the hidden order price drops, suggesting that not all impact is permanent. This tendency to reversion had also been observed previously.

Although a theoretical explanation of the empirical findings is considered further work, it is interesting to note that the square root and the reversion can be predicted through a simple argument based on the hypothesis that the price after reversion is equal to the avaerage price paid during execution. If during execution price impact grows like price impact grows like $A \times (t/T)^{\beta}$ then the average price payed by the agent who executes the order is

$$\langle p\rangle = p_t + A\int_0^1 (t/T)^{\beta} dt = p_t + \frac{A}{1+\beta}. \tag{5.15}$$

The fact that we observe similar behavior in both the London and Spanish stock exchanges, and that others have also observed this in the New York Stock Exchange, suggests the possibility of a law for market impact what it is part of my further work.

Appendix A

Financial Preliminaries

A.1 Electronic Markets

In 1973 currencies began to be traded in in the foreign exchange market (forex, FX, or currecy market), this means a financial market open and active 24 hours a day with the exception of weekends. Average daily turnover in global foreign exchange markets is estimated at $3.98 trillion as of April 2010, a growth of approximately 20% over the $3.21 trillion daily volume as of April 2007. Currency market is a typical example of decentralized, over-the-counter, market where market participants trade directly between two parties.

Another example of financial market is derivatives market where derivatives intruments such as futures, and options are traded. This market can be exchange-traded and over-the-counter. A derivatives exchange is a market where individuals trade standardized contracts that have been defined by the exchange. The derivatives exchange acts as an intermediary to all related transactions. According to the Bank for International Settlements (BIS), the combined turnover in the world's derivatives exchanges totaled $344 trillion during Q4 2005. Reporting of OTC amounts are difficult because trades are private, without activity being visible on any exchange. According to BIS, the total outstanding notional amount of OTC derivatives is $684 trillion (as of June 2008).

The most relevant financial market for this thesis is stock market (equity market) which is a market for the trading of shares at an agreed price. The size of the world stock market was estimated at about $36.6 trillion at the start of October 2008. The largest stock market in the United States, by market cap, is the New York Stock Exchange, NYSE. The London Stock Exchange (LSE) is currently the fifth-largest stock exchange in the world, by market cap[1], and the second largest in Europe after Euronext. Although, for the period of time under study in this thesis the LSE was the third stock exchange in the world, and the largest in Europe. Finally, the Spanish Stock Exchange (SSE) was the eighth largest market

[1] Market capitalization, or market cap, is a measurement of size of a corporation equal to the share price times the number of shares outstanding.

in the world for the studied time interval in this thesis.

All figures mentioned above are only for showing the magnitude of current financial markets, and their economical importance. But from a scientific point of view, it is more relevant the change which happened in the 1980s in financial markets: electronic trading. This meant a huge amount of data about quotes and trades recorded at high-frequency - this term is commonly used for a frequency higher than daily. Liquid markets generate thousands of ticks - one logical unit of information, like a quote or a transaction price - per business day. Data vendors like Reuters transmit more than 275,000 prices per day for foreign exchange spot rates alone.

High-frequency data are the finest level of financial time series. Therefore, it seems the logical object of research for describing and understanding financial markets. Moreover, because many practitioners make investment decisions at that time scale, e.g. high-frequency trading. However there are two reasons which make difficult the research of this type of data sets. First, these databases are costly, whereas low-frequency data are free. Second, statistical tools have been developed for data equally spaced intime, whereas high-frequency data are not homogeneously spaced.

In this thesis, all empirical results have been obtained based on high-frequency data sets. This gives a strong support to theoretical results presented in the previous chapters.

A.2 Order Book

The definition of electronic markets only takes into account the media to bring together buyers and sellers that must be electronic, no matter if they are OTC or exchange-traded. I assume most of the current markets are electronic makets. I classify them, based on the way orders are placed and executed, into two different types of markets: Quote-driven markets, and Order-driven markets.

In a quote-driven market, a market participant, called a dealer posts a price at which she is willing to buy or sell a certain quantity of a given asset. When an individual investor wishes to buy or sell the asset, must get in contact with a broker, who will call the dealer and finally the transaction will be executed. The quotes posted by a dealer can be indicative or firm. In the last case, any other dealer can trade at the posted price up to a certain amount. If the price was indicative, the dealer who is willing to initiate the transaction will confirm the price with the dealer who originally posted the quote.

In an order-driven market, investors can trade each other through an electronic trading system. There are several types of orders, but we can consider only two types which are common to all the markets: market orders, and limit orders. A market order is an order to purchase or sell inmediately a certain number of shares, and is executed at the best available price. A limit order indicates the willingness of trading a certain number of shares at a given price. It is the cumulation of unexecuted limit orders what generates the order book.

The order book can be defined as a sorted list of unexecuted limit orders waiting for a matching order to generate a transaction. In reality, there are two separate lists, one for buy orders, and another for sell orders. Both lists are sorted by two criteria: price, and time. Limit sell orders are sorted from lowest price to higher. Lowest price is called ask, and it is the best available price on the sell side of the book at a given moment. Shares at the ask price will be the first matched shares by a buy market order. Limit buy orders are sorted from highest price to lower. Highest price is called bid, and it is the best available price on the buy side of the book. Shares at the bid price are the first to be transacted when a sell market order is posted.

I have presented a simplistic view of the order book. There are many papers related to a more in depth study of the order book and its statistical properties. There are other magnitudes which should be defined to fully understand its dynamics. Although this is out of the scope of the research presented here, there are two additional necessary definitions: bid-ask spread, and midpoint price.

Bid-ask spread is the difference between bid and ask. This magnitude is important because gives us the order of magnitude of transaction costs, which are a proxy of the liquidity. This cost is measured by what is called a round trip. This round trip consists in buying a share at the ask price and selling it at the bid price. The net of these two transactions is the cost, and the lower is the cost the more liquid is considered a stock.

Finally, midpoint price is the midpoint between bid and ask, and can be taken as a “fair price" because is not biased by the willing of purchasing or selling. Another advantage - and the reason why I have employed it in this thesis - is because the bid-ask bounce effect[2] of high-frequency series can be ignored when prices are measured in this way.

A.2.1 Basic Dynamics of the Order Book

In Fig.(A.1) I show a schematic of the order book. The best price willing to buy is called the best bid price. The best price willing to sell is called the best offer price or the best ask price. The midpoint between these two prices is called the midpoint price and is a standard reference for the current price of the stock. The difference between the best bid and best offer price is called the spread. Prices for electronic markets are discrete, meaning that limit orders must specify prices in increments, the minimum sized increment of price for a stock is called the tick size.

Let’s assume that we have a set of outstanding orders conforming the order book. Both sides of the book are containing orders, so we have well-defined bid and ask prices. This is the usual aspect of the order book of a liquid asset during

[2] This effect makes reference to the observed bounce of transactions prices between the bid and ask price.

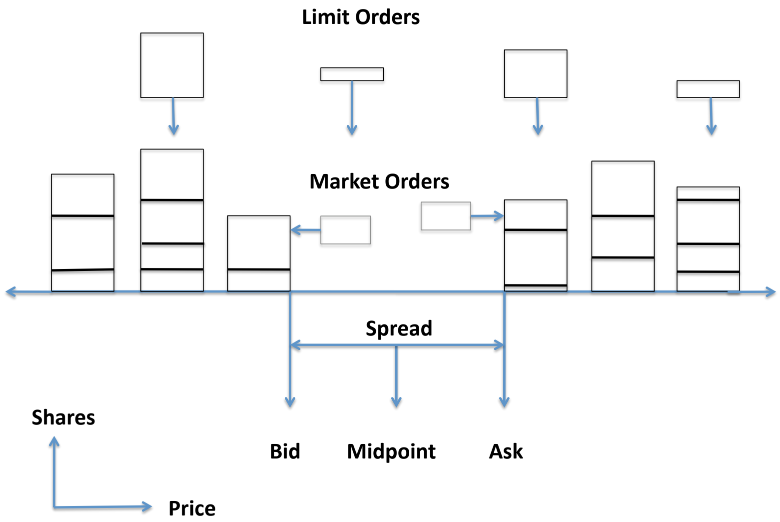

Figure A.1: **A schematic of the order book used in modern electronic markets.**

a trading session after excluding the period of price formation when one of the two sides of the book could be empty.

Let's see the dynamics of the book when a new market order is posted. This order will immediately cause a transaction at the best price of the opposite side of the book. If the number of shares available at the best price is enough to fill the market order, then will be fully matched at that only price. But if market order exceeds the number of shares at the best price -bid or ask,- the unmatched part of the market order will transact with the second best price, which is now the new best price, and so on until the entire market order is matched. It is unusual to see market orders larger than the liquidity available at the best price in the order book at a given time.

Now I consider the case when a new limit order is posted. The relevant parameter for this type of order is price. If there is already an outstanding order at that price, the new order will be added to the queue, with lower priority than shares previously posted. If there is no order at that price, a new queue is generated and these shares will be first. Limit orders are also classified depending on their relative price to the best price. If price is better than best prices this limit order is considered inside spread. Whereas if the price is higher than ask or lower than bid this order is called on the book. So, spread is itself a dynamical quantity.

Limit orders can be cancelled at any point of time if they have not been previously matched. Depending on the specific stock exchange limit orders can be conditional on some events. Moreover, limit orders can be specified as valid for a day, and if they have not been matched at the end of the day are automatically cancelled. They also can be specified as valid until the end of the month and then cancelled. Every order that is not traded inmediately is called a booked order. Booked orders are also known as outstanding orders. Market orders tend to deplete the order book, are liquidity takers, and increase the spread; whereas limit orders increase the order book, are liquidity providers and tend to decrease the spread. It can be said that the dynamics of the order book is an interplay between the flow of limit and market orders.

So far, I have described order book dynamics during continuous trading. But, in many stock exchanges there is a previous phase: the pretrading phase. During pretrading phase the operator can enter, change, and delete orders in the order book. The traders can not access any information on the order book. Then, the continuous trading starts with an opening auction. During opening auction indicative auction prices are displayed, and the present orders in the order book are the sum of the orders left over from the preceding day, more those ones entered in the pretrading phase, and finally the orders entered during the auction. The price is determined following a set of rules which try to execute a maximal volume of orders with a minimal residual of unexecuted order volume consistent with the order limits.

After continuous trading there is another auction: the closing auction. Then, we have the post-trading period. During this last period, as it happened in the pretrading, operators can modify their orders to prepare next trading session.

The Bid-Ask Spread

Bid-ask spread is a consequence of the way order book is built by the submission of limit orders, and the execution of market orders. In general, we can talk about three components of the bid-ask spread. The first one is the inventory component, which is related to adverse information. If a participant has a certain number of shares, any adverse information will affect the price of the asset and the inventory will lose value. For this reason, markets with larger price volatility have a larger inventory component. The second one is the transaction costs component. Finally, the third component is the asymmetric information. Based on this last component, investors have been classified into various classes: noise traders, informed traders. Noise traders are these one who trade for random reasons, and not based on news concerning the asset. Informed trader, as for example an institutional trader, is that one who trades based on the idea that a certain asset is mispriced. As a consequence of this factor when relevant news about a certain asset are expected, spread usually widens to avoid getting disadvantaged because of some informational asymmetry.

A.3 Financial Products

Financial markets currently offer the possibility to invest in a vast variety of financial products, it is not the aim of this section to provide a taxonomy of all these instruments. I only describe some of the characteristics of the products mentioned and employed in the next chapters.

A.3.1 Bonds

A bond is a debt instrument issued for a period of more than one year with the purpose of raising capital by borrowing. Generally, a bond is a promise to repay the principal along with interest (coupons) on a specified date (maturity). Some bonds do not pay interest - zero-coupon bonds,- but all bonds require a repayment of principal.

The price of a bond depends on the current level of interest rate. The yield, y of a zero-coupon is priced as

$$B(t) = \frac{P}{(1+y(t))^{T-t}} \approx Pe^{-y(t)(T-t)}, \tag{A.1}$$

where $B(t)$ is the current price of the bond, and T is the time to maturity. The yield is the compounded rate of interest computed as if the owner of bond held it to maturity. In general, bonds pay coupons before the maturity day, when principal is paid. This type of bonds is called coupon bearing bond. In this case, bond is price as

$$B(t) = \sum_{i=1}^{N} \frac{P_i}{(1+y)^{T_i-t}} \approx \sum_{i=1}^{N} P_i e^{-y(T_i-t)}, \tag{A.2}$$

where the set of $\{P_i\}$ are the coupons paid by the bond at times $\{T_i\}$.

A.3.2 Commodities

Commodities are usually raw materials, which are available in standard conditions of quality and format. Commodities market led to the development of derivatives products such as forwards and options, which were finally broadly traded in the financial markets. The most common types of commodities are crude oil, natural gas, gold, silver, copper, soy, wheat corn, sugar, coffee, cocoa, and cotton.

The price of a commodity is subject to supply and demand. Risk is actually the reason exchange trading of the basic agricultural products began. For example, a farmer risks the cost of producing a product ready for market at sometime in the future because he doesn't know what the selling price will be.

A.3.3 Stocks and Stock Indices

A stock is an instrument that signifies an ownership position (called equity) in a corporation, and represents a claim on its proportional share in the corporation's assets and profits. In general, the terms stocks and shares are subsumed under the term equity. Stocks can pay to its owners dividends, this is a regular payment as it was coupons in bonds case.

Stock indices are indices that reflect the price movements of equity markets. They are computed and published by stock exchanges. Indices are an attempt to reflect the performance of a certain group of stocks. The criterion employed for creating a group can be based on geographical reasons, on market capitalization, and on economical sector. Most stock indices are price averages weighted by market capitalization, which represents the total market value of a company. This value is usually computed as current price of a share multiplied by the number of outstanding shares.

A.3.4 Derivatives

A derivative is a financial instrument whose characteristics and value depend upon the characteristics and value of an underlier, typically a commodity, bond, equity or currency. Examples of derivatives include forwards, futures, and options.

A forward is a contract obligating one party to buy and another other party to sell a financial instrument, equity, commodity or currency at a specific future date. A future is a standardized, transferable, exchange-traded contract that requires delivery of a commodity, bond, currency, or stock index, at a specified price, on a specified future date. Unlike options, futures convey an obligation to buy. Futures contracts are forward contracts, meaning they represent a pledge to make a certain transaction at a future date. The exchange of assets occurs on the date specified in the contract. Futures are distinguished from generic forward contracts in that they contain standardized terms, trade on a formal exchange, are regulated by overseeing agencies, and are guaranteed by clearinghouses. Also, in order to insure that payment will occur, futures have a margin requirement that must be settled daily.

An option is the right, but not the obligation, to buy (for a call option) or sell (for a put option) a specific amount of a given stock, commodity, currency, index, or debt, at a specified price (the strike price) during a specified period of time.

Appendix B

Volatility

B.1 Volatility Modelling

We have shown the importance of the volatility for SV models that can be concisely summarized into a sentence: the non-Gaussian features of returns distribution are caused by the characteristics of volatility distribution. Thus, given the importance of this variable it seems necessary to define it as accurately as possible. However, there is no one only possible definition for it. Moreover, there is an interrelation between the available database and the most adequate definition for that specific collection of data.

As a general definition, the volatility is a measure of price variability over some period of time. It typically describes the standard deviation of returns. Alternatively, we can say that volatility is the standard deviation of the change in the logarithm of a price during a period of time. This general definition must be implemented following a more specific definition, and there are several realized volatility, conditional volatility, and implied volatility. These are the most common.

Realized volatility, also called historical volatility, is the standard deviation of a set of previous returns. For n trading periods, and returns $r_{t-n},...,r_{t-1}$ whose average is $\langle r \rangle$, the historical standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_{t-i} - \langle r \rangle)^2}, \tag{B.1}$$

this definition provides a simple estimate of the standard deviation of the return for the period t, usually a day. This volatility measure can be also expressed in annual units as $s\sqrt{N}$, where$N$ is the number of trading days in one year.

Conditional volatility is the standard deviation of a future return that is conditional on known information, i.e., previous returns. Unlike realized volatility, the expectation for the next period is calculated using a time series model. Then, for this measure of the volatility there is an additional factor: the model for the estimation. ARCH models provide equations for volatility expectations. The autoregressive conditional heteroskedastic (ARCH) models specify the conditional variance h_t of the return in period t using prior information I_{t-1}. A very well known example is the weighted sum of squared excess returns, defined by a recursive equation such as

$$h_t = \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{B.2}$$

where α, β, μ, and ω are parameters to be estimated for a returns series.

Implied volatility is a value calculated from an option price. It equals the volatility parameter σ for which an option's market price equals its theoretical price according to a pricing formula. The Black-Scholes pricing formula provides theoretical prices for Europen call options, say $c(\sigma)$, and assumes that the asset price process is described by a geometric Brownian motion with annual variance rate σ^2. As $c(\sigma)$is an increasing function of σ, for any market price c_M between the lower and upper bounds that exclude arbitrage profits there is a unique solution to the equation

$$c_M = c(\sigma), \tag{B.3}$$

that defines the implicit volatility. These volatility measures depend on the time until expiry and the exercise price of the option. Option markets are competitive and prices must incoprporate the market's expectations about future volatility. It is therefore reasonable to conjecture that implied volatilities are the best source of information when forecasting volatility. At any time the values of realized volatility, conditional volatility, and implied volatility will usually be all different, because different data and assumptions are employed when these values are calculated.

Realized and conditional volatility can be implemented on low and high frequencies time series. This is not the case of the implicit volatility which is calculated from the quoted prices of the options. The question when studying the volatility at ultra-high frequencies, that's intra-day time intervals, is what of all the available prices is taken as the magnitude for estimating the volatility. I have taken the finest possible price which was compatible with a reliable statistics: the midpoint price. This price is the midpoint between bid and ask for avoiding the bid-ask bounce effect. Thus, volatility is affected by the definition and the data set employed in its estimation.

As I have mentioned in the Introduction, volatility process is autoccorrelated. This is a very important characteristic because it makes possible to consider volatility as a slow process compared to returns, which may be taken as a fast process. This property of the volatility process is a key factor for the new model presented in Chapter 3.

In addition to the definition of volatility and the specific prices taken for the computation of the volatility, there is another factor: the probability distribution chosen as candidate. Given a certain empirical series representing the volatility, we may consider different candidates to fit that series. There are several examples in the literature where these possible distributions are compared, i.e. it has been shown that the probability density of the high-frequency volatility is fitted by two probability distributions: a log-normal and an inverse gamma, the two fits were performed using maximum likelihood. Results for inverse gamma were slightly better. The log-normal distribution tends to underestimate the tail of the distribution, whereas the inverse gamma tends to overestimate it. Same conclusions are reached for individual stocks. Therefore, it seems that inverse gamma is a slightly better fit to empirical distributions.

Appendix C

Resumen

Los dos problemas principales abordados en esta tesis son la dinámica de los precios de los activos financieros y la forma funcional del impacto en el precio de mercado de las órdenes ocultas. Estos problemas son relevantes tanto por su repercusión en diversas áreas de la Economía Financiera tales como: la valoración de opciones, la gestión del riesgo, la alocación de activos, o la ejecución óptima, como por el desafío científico que supone la comprensión de un sistema complejo como el que constituyen los mercados financieros actuales, y más específicamente los mercados electrónicos. Por tanto la motivación de este estudio es doble, por un lado la resolución de un problema de carácter fundamental que tiene importantes implicaciones en la resolución de problemas de índole aplicada, y por otro lado supone el intento de resolución de un problema que aun siendo de un área aparentemente lejano a las Ciencias Naturales ha sido abordado en repetidas ocasiones por físicos y matemáticos en un intento de demostrar que las Finanzas pueden ser también entendidas como un sistema complejo, y por tanto abordadas mediante el uso de herramientas desarrolladas para la explicación de sistemas físicos con los que comparten una serie de propiedades.

Ambos problemas son estudiados al nivel de detalle más fino posible, el de la alta frecuencia, esto es posible por la oportunidad que nos brindan los mercados financieros electrónicos en los que se guarda toda la información que se transmite a los sistemas de contratación. De esta forma hemos sido capaces de tener en cuenta no sólo las transacciones, sino también aquellas órdenes que no finalizan en transacción pero que sí transmiten información y condicionan las decisiones posteriores de los participantes en el mercado. También nos ha sido posible poder clasificar las órdenes en función de los miembros involucrados en las mismas, lo cual nos ha permitido entender las estrategias implementadas por los diversos participantes para reducir los costes de transacción de las órdenes grandes, que son aquellas cuyo volumen supera la liquidez disponible en el libro de órdenes en un momento dado.

El tipo de estudio realizado ha sido doble. En el caso del problema de la dinámica de precios en alta frecuencia, el estudio ha sido predominantemente teórico y analítico, aunque posteriormente se han realizado los pertinentes experimentos para ver el grado de adecuación de la nueva teoría con los datos empíricos. En el caso de la determinación de la forma funcional del impacto en el precio de mercado de las órdenes ocultas, el estudio ha sido de un marcado carácter experimental ya que el problema se haya aún en una fase de recopilación de resultados empíricos previa a un posterior desarrollo teórico, sobre el que de hecho realizo un esbozo al final del Capítulo y que tomo como base de partida en mi investigación futura.

La estrategia seguida en el estudio del primer problema ha sido la de realizar un modelo teórico de la dinámica de los precios en alta frecuencia en la que hemos definido el proceso estocástico seguido por los mismos siguiendo las premisas que deben cumplir los modelos superstadísticos, entendiendo que caso de ser cierta la asunción los resultados experimentales corroborarían la validez de la misma. Siguiendo esta estrategia hemos considerado que los precios siguen una dinámica rápida mientras que la correspondiente a la volatilidad es lenta, también hemos considerado que la distribución condicionada a la volatilidad de los retornos en escalas temporales iguales o inferiores al día sigue una normal, y por último también hemos considerado que la volatilidad cuando es estimada en la misma escala que los retornos condicionados puede ser tomada como constante. Por tanto el modelo es totalmente analítico a excepción de la estimación de la distribución de la volatilidad, para lo cual consideramos que sigue una de las superclases que son comunes en Superestadística. La resolución de la dinámica de los precios es analíticamente abordable, solventando problemas de implementación comunes a la dinámica en alta frecuencia de precios y retornos, y sin tener que recurrir a soluciones numéricas.

La estrategia seguida en el estudio del segundo problema es diferente a la del primero, dado que la naturaleza y fundamentalmente el estado de resolución del mismo es distinto. En este caso, la investigación es marcadamente experiemental y quiere servir de base a futuros desarrollos teóricos. El primer paso ha sido la clasificación de las órdenes ocultas mediante un mismo algoritmo de detección en los dos mercados estudiados. Posteriormente hemos definido una serie de variables que hemos considerado necesarias para poder caracterizar la forma funcional del impacto en precio de mercado de dichas órdenes. Estas variables nos han permitido considerar subgrupos del total de órdenes ocultas, donde era posible identificar el impacto causado por éstas del causado por el resto de órdenes presentes durante la ejecución de la primera. Finalmente, hemos restringido nuestro universo de activos a los que mostraban una actividad frecuente a nivel de órdenes ocultas para poder establecer conclusiones dentro de un grupo homogéneo. Tras todo este proceso experimental, hemos llegado a resultados que muestran un comportamiento similar para este tipo de órdenes en ambos mercados, y que está también en consonancia con estudios previos realizados sobre este tipo de órdenes en otros mercados. Esta similitud de resultados es la que me lleva a conjeturar que el impacto en precio sigue probablemente una ley, y que con la base proporcionada de resultados experimentales se va a poder llegar a entender y generalizar en trabajos próximos.

En la Introducción de la tesis, he descrito el objetivo de la misma y además he mostrado los resultados experimentales sobre los que he buscado una explicación teórica en la primera parte de la tesis. Si bien estos resultados experimentales son conocidos desde hace décadas y son comunes de forma general a todos los activos financieros, en este caso concreto he mostrado que los activos que después nos han servido para realizar el trabajo experimental estaban de acuerdo con resultados previos mostrados por otros autores. Además de esto, he introducido esta vez de manera más cualitativa el significado que se preseguía con la realización de los estudios experimentales de la segunda parte de la tesis. En los Capítulos 1 y 2, la principal aportación es la de presentar de forma sistemática los fundamentos teóricos necesarios para la comprensión del nuevo modelo teórico desarrollado en el capítulo siguiente. Además de dichos fundamentos, entre los que destaco los postulados que de forma general son asumidos sin ninguna otra consideración por los modelos modernos, lo que he pretendido es dar una justificación teórica de dichos postulados. Asímismo una de las aportaciones de esta parte es la de la clasificación de los modelos en cuanto a su capacidad teórica para la explicación de los eventos extremos, y que entiendo puede ser tomada como un baremo para medir la capacidad explicativa de los modelos, ya que, es precisamente en la predicción de estos eventos raros donde radica una de las mayores fortalezas del modelo nuevo introducido en el capítulo siguiente.

Los resultados presentados en los Capítulos 3 y 4 son originales por completo. Siendo los mostrados en el Capítulo 3 los correspondientes al nuevo modelo de Volatilidad Estocástica en alta frecuencia que tiene como base teórica los fundamentos de la Superestadística. En dicho modelo consideramos que los retornos son el resultado de dos dinámicas que actúan en distintas escalas temporales. En primer lugar tenemos una dinámica rápida relacionada con la variación de los precios en escalas temporales inferiores a un día. En segundo lugar tenemos la dinámica lenta relacionada con la volatilidad y que a escalas inferiores a un día puede ser considerada como constante, aunque en definitiva sus variaciones a escalas más extensas de tiempo son las causantes de las propiedades no gausianas de la distribución de los retornos. El modelo es totalmente analítico a excepción de la distribución de la volatilidad que es previamente asumida, aunque los resultados experimentales corroboran la bondad de la distribución elegida de antemano. Finalmente, vemos que la distribución teórica de los retornos se ajusta excelentemente a la obtenida experimentalmente lo cual demuestra que la dinámica de los retornos y por ende de los precios puede ser interpretada como propia de un sistema superestadístico y que las asunciones previas de la doble dinámica, dos escalas temporales y el doble carácter estocástico son acertados. Dada la variedad de acciones de renta variable sobre las que hemos efectuado los ensayos experimentales, pertenecientes a períodos temporales distintos, mercados diferentes y sectores económicos diversos, así como con distintas características de liquidez, actividad, capitalización... conjeturamos que el comportamiento de los activos financieros es universal y adecuadamente descrito por el esquema teórico propuesto.

Los resultados presentados en el Capítulo 4 también son totalmente originales, y constituyen un estudio empírico sobre las llamadas órdenes ocultas en dos mercados: LSE y SSE, que poseen características diferentes en varios aspectos tales como volumen transaccionado, número y tipo de participantes, así como la forma en que se ejecutan las transacciones, con mayor relevancia de las transacciones bilaterales en el caso del LSE. Previamente se habían estudiado dicho tipo de órdenes en mercados aislados, pero no se había realizado un estudio comparado de las propiedades de dichas órdenes en dos mercados. La principal dificultad que presenta el estudio de las órdenes ocultas es que están compuestas de una colección de transacciones de menor volumen que no son identificadas como pertenecientes a una orden mayor, y por tanto han de ser inferidas a partir de las transacciones, de las órdenes de tipo limit y de los miembros de mercado que las han enviado. El principal objetivo de este Capítulo es caracterizar la forma funcional del impacto en precio de mercado de las órdenes ocultas en ambos mercados y comprobar si esta forma funcional era particular de cada mercado o si bien se puede encontrar algún paralelismo entre ambos, y con los estudios previamente publicados de otros mercados. He encontrado que el impacto se puede dividir en un impacto temporal y otro permanente, que el impacto temporal es cóncavo y se puede describir aproximadamente como la raíz cuadrada del volumen transaccionado de la orden oculta y que el impacto permanente supone alrededor del impacto temporal, estando estos resultados en consonancia y desde luego nunca en desacuerdo con los estudios anteriores, entiendo que sugiere la existencia de una ley para el impacto en precio de mercado.

Entre las conclusiones más importantes alcanzadas en esta tesis, podemos citar en primer lugar que el modelo de subordinación supone un marco teórico suficientemente amplio y flexible dentro del cual se puede hallar una solución al problema de la distribución de los retornos de los activos financieros. En segundo lugar, que la consideración de la volatilidad como proceso director nos permite explicar la no gausianidad hallada en las distribuciones empíricas, y que incluso es válido para la comprensión de la dinámica en alta frecuencia. En tercer lugar, que la dinámica de los precios y por tanto de los retornos puede ser considerada un sistema complejo y modelizada como un sistema superestadístico con todas las características de este tipo de modelos, entre las que destaco la existencia de una doble dinámica, con dos escalas temporales distintas, y la existencia de intervalos en los que la volatilidad puede ser tomada como constante y los retornos condicionados a dicha volatilidad normalmemte distribuidos. En cuarto lugar, y como solución al problema original planteado de la explicación de las colas gruesas de las distribuciones de los retornos así como su aparente estabilidad, se ha llegado a a la conclusión de que ambos son una consecuencia de la dinámica de la volatilidad. Dando un soporte teórico y experimental fuerte a los modelos de volatilidad estocástica. Sobre las conclusiones fundamentales a las que se ha llegado en la tesis sobre el segundo de los problemas planteados, destacaría que los resultados experimentales encontrados en los dos mercados estudiados demuestran la existencia de un impacto temporal en precio de mercado que es compatible con la raíz cuadrada del volumen transaccionado, y un impacto permanente que supone una reversión del precio alcanzado durante el impacto temporal. Estos resultados son compatibles con otros resultados experimentales previos encontrados en otros mercados, y que no entran en contradicción con modelos teóricos desarrollados para otro tipo de órdenes, ademĞs de sugerir como en el caso del problema primero que una ley universal es compatible con los hechos experimentales.

Los resultados de la tesis han sido publicados en revistas internacionales de reconocido prestigio (todas ellas se encuentran en el primer cuartil de la clasificación de JCR). Los resultados que aún o han sido publicados no han sido expuestos pero me permito mencionarlos como preprints dado que constituyen parte de la investigación actual, y porque son conclusiones basadas en el trabajo previamente mostrado.

1. Fuentes, M. A., A. Gerig, and J. Vicente. "Universal behavior of extreme price movements in stock markets." 2009, PLoS ONE, 4(12), e8243.

2. Gerig, A., J. Vicente, and M. A. Fuentes. "Model for non-Gaussian intraday stock returns." 2009, Phys. Rev. E, 80, 065102(R).

3. Moro, E., J. Vicente, L. G. Moyano, A. Gerig, J. D. Farmer, G. Vaglica, F. Lillo, and R. N. Mantegna. "Market impact and trading profile of hidden orders in stock markets." 2009, Phys. Rev. E, 80, 066102.

4. J. Vicente, Fuentes, M. A., A. Gerig, and Mikel Tapia. "Universal behavior of price dynamics under clock changes", preprint (2011).

5. Mikel Tapia, and J. Vicente. "Tick size and the probability of extreme events", preprint (2011).

Además de los los citados artículos, los principales resultados de la presente tesis van a ser publicados en el libro: "Derivative Securities Pricing and Modelling" perteneciente a la serie Contemporary Studies in Economics and Financial Analysis, de la editorial Elsevier.