Time Series Models for Measuring Market Risk

Technical Report

José Miguel Hernández Lobato, Daniel Hernández Lobato and Alberto Suárez

Departamento de Ingeniería Informática,
Universidad Autónoma de Madrid,
C/ Francisco Tomás y Valiente, 11,
Madrid 28049 Spain.

July 18, 2007

Abstract

The task of measuring market risk requires to make use of a probabilistic model that captures the statistical properties of price variations in financial assets. The most important of these properties are autocorrelations, time-dependent volatility and extreme events. GARCH processes are financial models that can successfully account for the time-dependent volatility. However, they assume Gaussian errors whereas empirical studies generally lead to residuals which exhibit more extreme events than those implied by a Gaussian distribution. In this document we analyze the performance of different models which try to solve this deficiency of standard GARCH processes. The first group of models is based on mixtures of autoregressive experts which work together following three possible strategies: collaboration, soft competition or hard competition. Mixtures of soft competitive experts produce the best estimates of risk because their hypothesis space is a mixture of Gaussian distribution which can account for extreme events. Finally, we study a model which improves standard GARCH processes by means of modelling the innovations in a non-parametric way. The resulting model turns out to provide very precise measurements of market risk and outperforms soft competitive mixtures with 2 experts.

Contents

Chapter 1

Introduction

Market risk is caused by exposure to uncertainty in the market price of an investment portfolio [Holton, 2003]. Any financial institution which holds a portfolio of financial assets is exposed to this kind of risk and consequently should implement risk measurement and management methods in order to optimize the *manner* in which risk is taken. Doing so will reduce the probability of incurring big economic losses or even bankruptcy and will make the institution more competitive [Jorion, 1997].

The process of measuring market risk can be described as summarizing in a single number the risk involved by holding a portfolio, a task which requires an accurate modeling of the statistical properties of future price variations in financial assets [Dowd, 2005]. The approach usually followed consists in making use of machine learning techniques [Bishop, 2006] in order to infer the distribution of future price variations from historical data. The steps involve suggesting a model for price changes, fitting its parameters to historical data and then using the model to make inference about the future. Once an estimate of the distribution of future price changes is available, it is necessary to employ a *risk measure* to quantify risk.

There are several risk measures available, some of the most commonly used are the standard deviation, Value at Risk, and Expected Shortfall [Dowd, 2005]. The standard deviation has the inconvenient that it is a low (second) order moment and therefore is not very sensitive to the behavior at the tails, which is crucial for the adequate characterization of risk. Furthermore, it is a symmetric measure which means that it would be affected by profits as well as by losses. Value at Risk is a percentile at a high probability level (usually 95% or 99%) of the distribution of losses. It can be interpreted as an estimate of the lower bound for large losses that occur with a low probability. Finally, Expected Shortfall is the expected value of the loss conditioned to the loss being larger than the Value at Risk. Expected Shortfall has the advantage that, unlike Value at Risk, it is a coherent risk measure [Artzner et al., 1999], and that it provides an expected value for the magnitude of large losses.

After choosing a particular risk measure and a model for price variations it is necessary to validate the model for estimating risk by means of the selected risk measure. This process is called *Backtesting* [Dowd, 2005] and generally consists in testing the hypothesis that the price changes observed in a certain period are consistent with the level of risk estimated immediately before that period. The process of backtesting requires to make use of advanced statistical tests [Kerkhof and Melenberg, 2004] and can also be used to

compare different models in order to determine which one leads to the best risk forecasts.

Daily price variations within various types of financial markets and a large set of different financial assets present common empirical properties [Cont, 2001]. These properties can be seen as constraints that a probabilistic model should fulfill in order to accurately capture process of price changing. Here, we review some of the most important of these properties. The first characteristic is that price variations show no significant autocorrelations at lags longer than one, although sometimes a small (but significant) positive autocorrelation appears at the first lag. The second property is that the unconditional distribution of price changes shows tails heavier than those of a Gaussian distribution. Finally, the last characteristic is that the standard deviation or volatility of price variations is time-dependent. GARCH processes [Bollerslev, 1986] are financial models that can successfully account for this last property of price changes. However, in their classical formulation they assume Gaussian innovations. Empirical studies show that after fitting a GARCH process to empirical data its residuals still exhibit tails heavier than those of a Gaussian distribution [Bollerslev, 1987]. In order to address this problem several extensions of classical GARCH processes with non-Gaussian heavy-tailed innovations were proposed, [Bollerslev, 1987], [Forsberg and Bollerslev, 2002] and [Mittnik and Paolella, 2003] are some examples. In this document we analyze two alternative solutions, the first based on the mixture of experts paradigm [Bishop, 2006] and the second based on non-parametric kernel density estimates [Silverman, 1986].

We first study mixtures of up to three autoregressive experts [Jacobs et al., 1991] whose outputs are combined using different strategies: soft competition, hard competition and collaboration [Hernández-Lobato and Suárez, 2006]. Soft competition implies that, for a fixed input, any expert is stochastically selected for generating the output of the mixture. A hard competitive strategy requires that, for a fixed input, the output of the mixture is always generated by a single expert which is deterministically selected. On the other hand, collaboration implies that the output of the mixture is a weighted average of the output of each expert. It turns out that mixtures which employ a soft competitive strategy outperform the other models. This is due to the fact that soft competitive mixtures predict a future distribution for price variations which is a mixture of Gaussians (one for each expert), a paradigm which can effectively account for the heavy tails of financial time series (note that a mixture of an unlimited number of Gaussians can approximate, up to any degree of precision, any density function). However, one drawback of mixture models is that the number of experts (Gaussians) is limited due to overfitting and training cost.

Finally, an extension of GARCH processes is given that involves modeling innovations in a non-parametric way [Hernández-Lobato et al., 2007]. The distribution of innovations is approximated in terms of kernel density estimates defined in a transformed space [Wand et al., 1991] to better account for the heavy tails of financial time series. The mentioned kernel estimates can also be regarded as constrained mixtures of Gaussians. However, the difference with respect to soft competitive mixtures is that, in this case, it is feasible to employ thousands of Gaussians without causing overfitting. The experiments performed demonstrate the superiority of GARCH processes with non-parametric innovations for performing market risk estimation.

The document is organized as follows.

- Chapter 2 is an introduction to the whole process of measuring market risk. It gives a description of risk, market risk and risk measures and indicates how a probabilistic

model for price changes has to be employed in order to successfully estimate market risk. Next, we review the most characteristic properties of price variations as well as the the most popular financial models for price changes. Finally, the process of validating market risk models (backtesting) is outlined.

- In Chapter 3 we present a study where mixtures of competitive and collaborative autoregressive experts with Gaussian innovations are analyzed for estimating market risk. The output generated by a collaborative mixture is an average of the predictions of the experts. In a competitive mixture the output is generated by a single expert which is selected either deterministically (hard competition) or at random with a certain probability (soft competition). The different strategies are compared in a sliding window experiment over the series of price variations of the Spanish index IBEX 35 which is preprocessed to account for its time-dependent volatility. The backtesting process indicates that the best performance is obtained by soft competitive mixtures.

- Chapter 4 presents a procedure to estimate the parameters of GARCH processes with non-parametric innovations by maximum likelihood. An improved technique to estimate the density of heavy-tailed distributions from empirical data is also given. The performance of GARCH processes with non-parametric innovations is evaluated in a series of experiments on the daily price variations of IBM stocks. These experiments demonstrate the capacity of the improved processes to yield a precise quantification of market risk. Furthermore, GARCH processes with non-parametric innovations outperform soft competitive mixtures with 2 experts.

- Finally, Chapter 5 contains a brief summary of the conclusions reached throughout this document and gives some ideas for a possible extension of the research performed.

Chapter 2

Measuring Market Risk

This chapter provides the reader with a brief description of market risk, the process of market risk measurement and the process of validating market risk models. Such models must accurately capture the statistical properties of price variations in financial assets. Because of this, we also describe those properties and review the most popular models which are currently used in the field.

2.1 Risk, Market Risk and Risk Measures

The term *'risk'* denotes an abstract concept whose meaning is rather difficult to describe. Informally, someone faces risk when they are exposed to a situation which might have a detrimental result (the word *'might'* is quite important as will be seen later). For example, if I bet 1.000€ on number 7 in a roulette of a casino I am facing risk because I can loose money if the ball lands on a number different from 7.

A more formal definition of risk is given in [Holton, 2004], where it is argued that risk has two components: *uncertainty* and *exposure*. Uncertainty is the state of not knowing whether a proposition is true or false and exposure appears when we do care about the proposition actually being true or false. Both uncertainty and exposure must be present, otherwise there is no risk. Going back to the example of the roulette we notice that there is uncertainty and exposure in that situation. There is uncertainty because I do not know which of the numbers the ball is going to land on and there is exposure because I could lose 1.000€. As soon as the ball stops moving uncertainty vanishes and so does risk: I have either won 35.000€ or lost 1.000€. On the other hand, if I were the owner of the casino I would neither win nor lose money by betting. The result of the spin of the roulette would still be uncertain. However, there would be no exposure and as a result no risk.

The level of risk is monotonically related to the levels of uncertainty and exposure. The lower the uncertainty or the exposure, the lower the risk. This can be illustrated in the example of the roulette. If I knew that the wheel of the roulette is biased and that the probability of the ball landing on number 7 is 1/2 instead of 1/37, I would be less uncertain about the possible outcomes and I would face less risk. On the other hand, if instead of betting 1.000€ I bet 10€, I would be less exposed to the outcome of the game (I do not care much if I lose 10€) and the level of risk would diminish.

2.1.1 Market Risk

Market risk is caused by exposure to uncertainty in the market value of an investment portfolio [Holton, 2003]. If I am holding a portfolio with stocks from several companies I know what the market value of the portfolio is today, but I am uncertain about what it will be at some time horizon τ in the future. If the price of the portfolio decreases I will be exposed to economic loss: I am facing market risk. Uncertainty in price and market risk generally increase with τ. Because of this, we must fix a value for τ if we want to quantify the risk exposure for holding a portfolio. The appropriate value of τ depends on the longest period needed for an orderly portfolio liquidation (in case it is necessary to reallocate investments to reduce risk for example) [Jorion, 1997]. For the trading portfolio of a bank composed of highly liquid assets a time horizon of one day may be acceptable. However, for a pension fund which is rebalanced every month a time horizon of 30 days would be more suitable.

The uncertainty in the future market value of a portfolio stems from the efficient market hypothesis [Fama, 1970]. This hypothesis states that current market prices reflect the collective beliefs of all investors about future prospects. As a result, price movements correspond to the arrival of new unexpected information whose impact on the future expected evolution is rapidly incorporated into the asset prices. In consequence, if the market is efficient, there should be no arbitrage opportunities. That is, it should not be possible to make a profit without being exposed to some amount of risk. Because new information is by nature unpredictable (otherwise it would not be new) so are the variations in market prices. Under the efficient market hypothesis investors can only outperform the market through luck.

Any financial institution that holds an investment portfolio is exposed to market risk, a risk which has rapidly augmented during the last decades due to the increased trade and volatility in financial markets [Dowd, 2005]. To cope with such markets with increasing levels of risk, financial institutions must implement risk measurement and management methods. Failure to do so will only increase the list of financial disasters with billions of dollars in losses that have taken place since the early 1990s and that could have been avoided with adequate risk management systems [Jorion, 1997].

2.1.2 Risk Measures

The discipline of *financial risk management* consists in the design and implementation of procedures for controlling financial risk (any risk associated with the loss of money, like market risk) [Jorion, 1997]. By means of adequate risk management practices a financial institution can protect itself from situations which involve an unnecessary high level of risk or take action if its current level of risk is too high. However, this requires that the institution is able to measure the amount of risk which it is or will be facing.

From a statistical perspective, the risk profile of a given situation can be analyzed in terms of the probability distribution of its possible outcomes. This distribution contains all the information about the uncertainty and the exposure that the situation involves. For instance, in the case of market risk, the distribution of profits at the selected time horizon τ would be sufficient to completely characterize the risk exposure of our investment. As an example, Figure 2.1 shows the profit distributions for holding 1.000 € of two imaginary financial assets for a time horizon $\tau = 1$ year. These distributions completely characterize the uncertainty and exposure implied by holding one asset or the other. However, at a

first glance, it is not obvious how to determine which position involves more risk and how much difference in risk there is between the two positions. The reason is that distribution functions do not provide a quantitative measurement of risk, something that risk measures actually do.

A risk measure is a procedure that summarizes in a single number the risk which a particular situation involves. Because risk is completely determined by the probability distribution of possible outcomes a risk measure can be seen as a map from a distribution function to a scalar. More formally, a risk measure is a functional $\rho : \mathcal{D} \to \mathbb{R} \cup \pm\infty$, where $\mathcal{D}$ is the set of all distribution functions. It is not necessary for a risk measure to make use of the whole distribution function. As a matter of fact, it can use only a segment of the distribution, for instance a fixed fraction of the worst or best outcomes.

In finance the most common general risk measures are the standard deviation, Value at Risk and the more recently proposed Expected Shortfall [Dowd, 2005].

- **Standard Deviation**. This is the risk measure used in modern portfolio theory [Markowitz, 1991]. For a given distribution $\mathcal{P}$ its functional form is

$$\rho_{sd}(\mathcal{P}) = \sqrt{\int_{-\infty}^{\infty} x^2 \, d\mathcal{P}(x) - \left(\int_{-\infty}^{\infty} x \, d\mathcal{P}(x) \right)^2} \,. \tag{2.1}$$

  The main disadvantage of the standard deviation as a risk measure is that it will generally fail to accurately quantify risk when $\mathcal{P}$ is not Gaussian. Interestingly, this is precisely the case in many of the situations found in finance.

- **Value at Risk (VaR)**. It is defined as the worst result within the α fraction of best results where α is usually high, 0.95 or 0.99 for example. Intuitively, the Value at Risk can be considered as the worst expected result within a probability level α. Its functional form is defined as

$$\rho_{VaR}(\mathcal{P}) = -\mathcal{P}^{-1}(1-\alpha) \,. \tag{2.2}$$

  The minus sign is included because $\mathcal{P}^{-1}(1-\alpha)$ is usually negative. Thus, the Value at Risk is a positive number that corresponds to a loss. The main limitation of VaR is that it provides no information about how bad things can get outside the α fraction of best results. In this way, two situations might have the same VaR, so that apparently they have the same risk exposure, and yet one could actually be riskier than the other.

- **Expected Shortfall (ES)**. Also called conditional VaR, it attempts to address some of the deficiencies of Value at Risk. The Expected Shortfall for a level α (0.95 or 0.99 for example) can be defined as the average result obtained when the result is worse than the Value at Risk for the α fraction of best results. Its functional form is

$$\rho_{ES}(\mathcal{P}) = -\frac{1}{1-\alpha} \int_{-\infty}^{\mathcal{P}^{-1}(1-\alpha)} x \, d\mathcal{P}(x) \tag{2.3}$$

  which, likewise Value at Risk, is a positive quantity representing a loss.

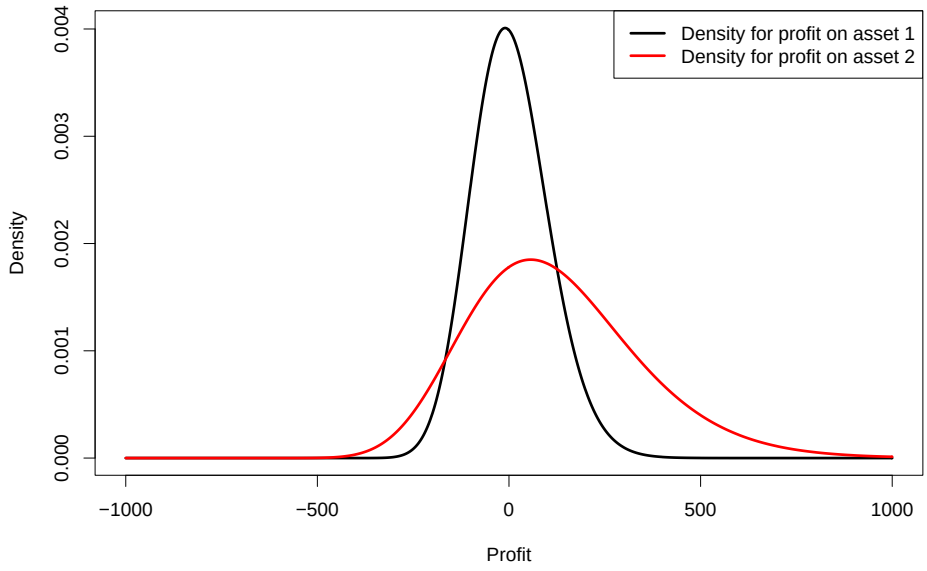

Figure 2.1: Density functions for one-year-profits obtained after investing 1.000 € in each of two imaginary assets. The distribution for the profits is lognormal so that the highest possible loss is 1.000 €.

With the help of a risk measure it is possible to quantify the risk associated with investments in the assets displayed in Figure 2.1. For example, if we fix the level $\alpha = 0.95$ and compute the VaR for holding 1.000 € of asset 1 and for holding 1.000 € of asset 2 we obtain the values 151 € and 208 €. Informally, we can say that asset 2 is 57 € riskier than asset 1 and that if things get really bad we can expect to loose 151 € if we invest in asset 1 and 208 € if we invest in asset 2. More formally, this means that there is a 0.05 probability of loosing more than 151 € if we invest in asset 1 and at least 208 € if we invest in asset 2. Finally, we can conclude that asset 2 is riskier than asset 1 according to the employed risk measure.

2.1.3 Coherent Risk Measures

In [Artzner et al., 1999] it is postulated a set of axioms which a risk measure ρ should fulfill in order to be considered coherent for quantifying financial risk (any risk associated with the loss of money, like market risk). The axioms are common-sense rules designed to avoid awkward outcomes when dealing with risk in finance. For example, if the profit for portfolio A is always bigger than the profit for portfolio B we would expect B to be riskier than A. We would also expect diversification to decrease risk, remember the age-old saying *"do not put all your eggs in only one basket"*.

The set of coherent axioms is

1. **Translation Invariance**. If $X \sim \mathcal{P}$ denotes the profit of a portfolio at some horizon in the future, and $Y = X + \alpha$, $Y \sim \mathcal{Q}$ where α is some fixed amount (positive or negative) of monetary units, then $\rho(\mathcal{Q}) = \rho(\mathcal{P}) + \alpha$.

2. **Positive Homogeneity**. If $X \sim \mathcal{P}$ denotes the profit of a portfolio at some horizon in the future, and $Y = \alpha X$, $Y \sim \mathcal{Q}$ where $\alpha > 0$, then $\rho(\mathcal{Q}) = \alpha\rho(\mathcal{P})$.

3. **Monotonicity**. If $\mathcal{P}$ and $\mathcal{Q}$ are the profit distributions for two portfolios A and B at some horizon in the future and the profits for portfolio A are always higher than those for portfolio B, then $\rho(\mathcal{P}) \leq \rho(\mathcal{Q})$.

4. **Subadditivity** . If $\mathcal{P}$ and $\mathcal{Q}$ are the profit distributions for two portfolios A and B at some horizon and $\mathcal{U}$ is the profit distribution for the portfolio $A \cup B$ at the same horizon, then $\rho(\mathcal{U}) \leq \rho(\mathcal{P}) + \rho(\mathcal{Q})$. It reflects the fact that due to diversification effects the risk for holding two portfolios is less (or at least the same) than the sum of the risks for holding each portfolio separately.

From the three risk measures previously seen only Expected Shortfall is a coherent risk measure. Specifically, Standard Deviation clearly fails axioms 1 and 3 and Value at Risk fails axiom 4. It is quite obvious to see why Standard Deviation is not coherent. For Value at Risk, we can prove it is not subadditive with a *counter-example* where it violates that condition [Dowd, 2005]. Let us consider the following example:

> We have two financial assets A and B. After one year, each of them can yield a loss of 100€ with probability 0.04 and a loss of 0€ otherwise. The 0.95 VaR for each asset is therefore 0, so that VaR(A) = VaR(B) = VaR(A) + VaR(B) = 0€. Now suppose that losses are independent and consider the portfolio that results from holding asset A and B. Then, we obtain a loss of 0€ with probability $0.96^2 = 0.9216$, a loss of 200€ with probability $0.04^2 = 0.0016$, and a loss of 100€ with probability $1-0.9216-0.0016 = 0.0768$. Hence, VaR($A \cup B$) = 100€ > 0 = VaR(A) + VaR(B), and the VaR violates subadditivity.

Standard Deviation and Value at Risk are therefore undesirable risk measures for quantifying market risk. However, these two risk measures are still being widely used because of their simplicity and the success of many portfolio optimization applications and risk management systems (e.g. RiskMetrics™ [Morgan, 1996]) which employ them.

2.1.4 Estimating Market Risk Measures

We have already seen in Section 2.1.2 that the exposure to market risk of a portfolio can be easily quantified by means of a risk measure. The only requirement is that the future profit distribution of the portfolio for a given time horizon must be available. However, if the efficient market hypothesis [Fama, 1970] is taken to be true, such distribution will not generally be known. This is so because the source of future price changes is future information whose statistical properties might be time-dependent, having some degree of randomness. The usual approach to tackle the problem consists in assuming that the distribution of future price changes will be related to that of recent past price changes. In this way, machine learning and pattern recognition techniques [Bishop, 2006, MacKay, 2003, Duda et al., 2000] can be used to make inference about the statistical properties of future prices. The general process involves choosing a probabilistic model that captures the properties of financial time series (series of market prices). The parameters of such model are fixed by a fit to recent historical data (e.g. by maximum likelihood) and then the model is used to infer the distribution of future price changes. However, financial time series show complex properties and obtaining a good model that can accurately describe the process of price changing in financial assets is generally a very challenging task.

2.2 Properties of Price Variations and Time Series Models

This section introduces a set of empirical properties observed in price variations within various types of financial markets and a large set of different financial assets. These properties can be seen as constraints that a probabilistic model should fulfill in order to accurately capture the process of price variation. It turns out that such properties are very demanding and most of the currently existing models fail to reproduce them [Cont, 2001].

2.2.1 Financial Returns

Let P_t denote the price for a financial asset at date t. In general, financial models focus on returns instead of prices for several reasons [Campbell et al., 1997, Dowd, 2005]. First, returns are a complete and scale-free free summary of profits and losses. Second, returns have theoretical and empirical properties which make them more attractive than prices. Third, if we have a probabilistic model for returns it is straightforward to obtain a probabilistic model for prices or profits. The *net return*, R_t, on the asset between dates $t-1$ and t is defined as

$$R_t = \frac{P_t}{P_{t-1}} - 1\,. \tag{2.4}$$

The net return over the most recent k periods from date $t-k$ to date t is simply

$$R_t(k) = (R_{t-1}+1)\cdot(R_{t-2}+1)\cdot(R_{t-3}+1)\cdots(R_{t-k+1}+1) - 1\,, \tag{2.5}$$

where this multiperiod return is called *compound* return. The difficulty manipulating series of products like 2.5 motivates another approach to calculate compound returns. This is the concept of continuous compounding. The *continuously compounded return* or *logarithmic return* r_t between dates $t-1$ and t is defined as

$$r_t = log(1+R_t) = log\frac{P_t}{P_{t-1}} = log(P_t) - log(P_{t-1}) \tag{2.6}$$

and is very similar in value to the corresponding net return R_t because $r_t = log(1+R_t) \simeq R_t$ for small values of R_t in absolute value. Continuously compounded returns implicitly assume that the temporary profits are continuously reinvested. Furthermore, they are more economically meaningful than net returns because they ensure that the asset price can never become negative no matter how negative the return could be. However, the main advantage of logarithmic returns becomes clear when we calculate multiperiod returns

$$r_t(k) = log(1+R_t(k)) = r_{t-1} + r_{t-2} + r_{t-3} + \cdots + r_{t-k+1}\,, \tag{2.7}$$

which allows to model the change in price as an additive process. As an example, we show on Figure 2.2 the time series of daily prices (properly adjusted for dividends and splits) and daily percentage logarithmic returns (100 times the logarithmic returns) for the stocks of Microsoft since 4th of January 2000 to 18th of April 2007, a total of 1831 measurements. Because of their multiple advantages we will focus on the statistical properties of continuously compounded returns and from now on the term *'return'* will always refer to logarithmic return.

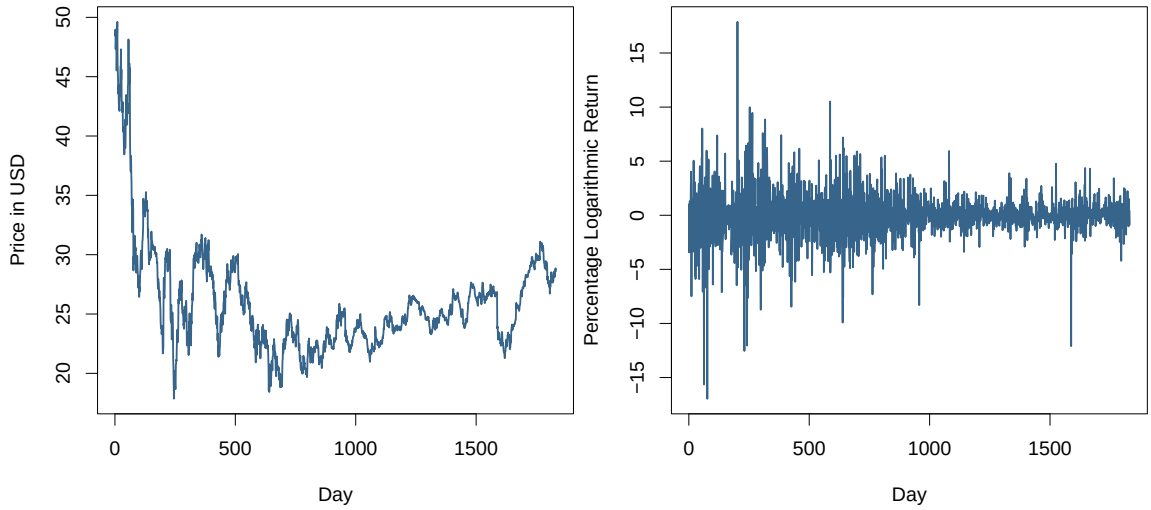

Figure 2.2: Daily prices (left) and daily percentage logarithmic returns (right) for the stocks of Microsoft since 4th of January 2000 to 18th of April 2007, a total of 1831 measurements. The prices were properly adjusted for dividends and splits.

2.2.2 Statistical Properties of Asset Returns

Intuitively, it would seem that returns obtained from different financial assets should exhibit different statistical properties. After all, why should there be any relationship between the properties of cotton futures and the properties of Microsoft stocks? The truth is that empirical studies over financial time series indicate that such is the case and that returns of different assets share similar statistical properties. Below, we list the most important of these properties, for a thorough description see [Cont, 2001].

1. **Linear autocorrelations**. Generally, returns for individual assets seem to show no significant correlations at any lag [Fama, 1970], although in the case of some stock indexes or portfolios, it can usually be appreciated a small (but significant) positive autocorrelation at the first lag [Campbell et al., 1997]. However, these results should be taken cautiously because sample autocorrelations can be unreliable estimates if the data is heavy-tailed distributed [Davis and Mikosch, 1998] (which is the case of asset returns).

2. **Heavy tails**. The empirical unconditional distribution of returns shows heavier tails and higher kurtosis than those implied by Gaussianity. Studies based on extreme value theory suggest a tail index (the highest finite moment) for the distribution of returns which is between five and three [Longin, 1996, Jansen and de Vries, 1991]. However, the precise shape for the tails is difficult to determine, a task which is vital for accurate risk measurement because risk measures like Value at Risk and Expected Shortfall will provide misleading results if the loss tail of the distribution for returns is wrongly modeled. As an example, we show on Figure 2.3 the histogram and normal quantile-quantile plot for the standardized daily returns of Microsoft stocks.

3. **Volatility clustering**. The time series of financial asset returns are usually heteroskedastic. That is, the volatility or standard deviation of the returns exhibits a

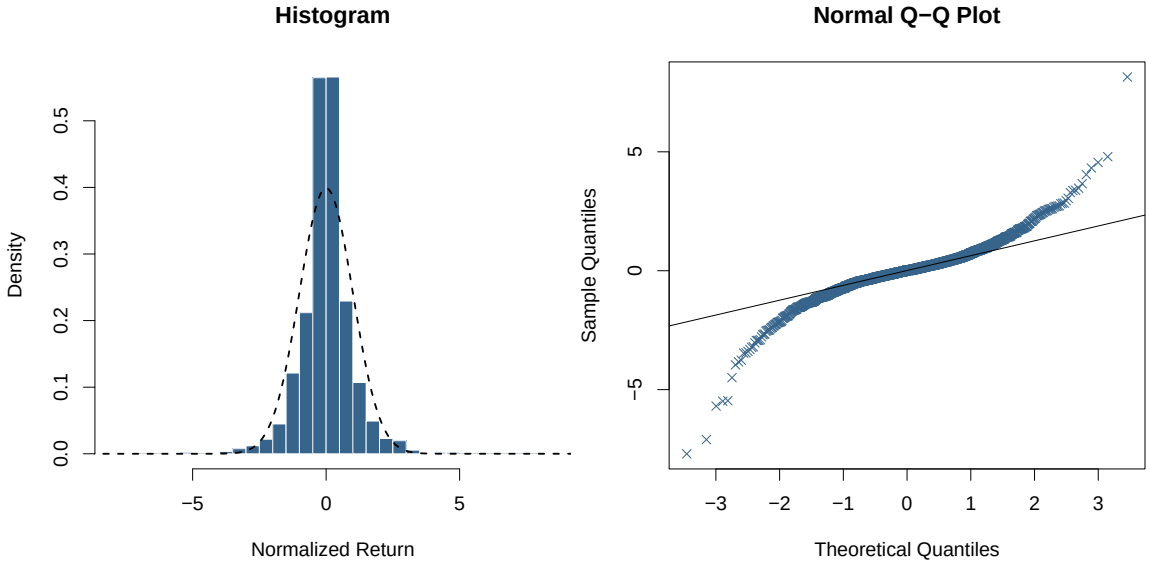

Figure 2.3: Left, histogram for the standardized daily returns of Microsoft stocks and standard Gaussian density. It is noticeable the higher kurtosis of the empirical density for the returns. Right, normal quantile-quantile plot for the standardized daily returns of Microsoft stocks. It can be appreciated how the tails of the empirical density are heavier than those of a standard normal distribution.

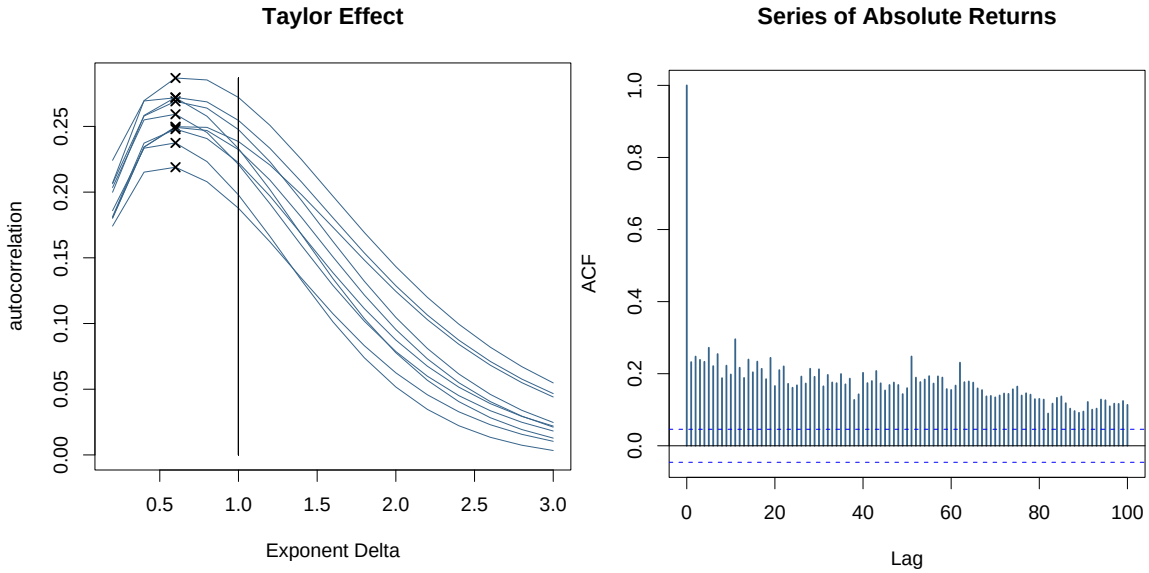

Figure 2.4: Left, Taylor effect plot for the returns of Microsoft stocks where it is shown the autocorrelation function between $|r_t|^\delta$ and $|r_{t+h}|^\delta$ for $h \in \{1, ..., 10\}$. Each curve corresponds to a different value of $h = 1, 2, \ldots, 10$ and the maximum of each function is shown with a cross. Right, autocorrelations between $|r_t|$ and $|r_{t+h}|$ where $h = 1, 2, \ldots, 100$ for Microsoft stocks.

time-dependent structure. Volatility clustering implies that large price variations (either positive or negative) are likely to be followed by price variations that are also large. This phenomenon is evidenced by a plot of the autocorrelations in the powers of the absolute values of returns

$$C_\delta = corr(|r_{t+h}|^\delta, |r_t|^\delta),\ h = 1, 2, \ldots \tag{2.8}$$

These correlations are positive for various delays in the range of weeks to months (see Figure 2.4) and the highest values are usually achieved for δ around 1 [Ding et al., 1993]. This behavior is known as the Taylor effect and is displayed on the left of Figure 2.4 for the returns of Microsoft stocks.

2.2.3 Time Series Models for Asset Returns

Probably, the most successful models to account for the time-dependent volatility in financial time series are GARCH processes [Bollerslev, 1986, Hamilton, 1994] and particularly the GARCH(1,1) process. We say that a time series $\{r_t\}_{t=1}^T$ follows a GARCH(1,1) process with normal innovations if

$$r_t = \sigma_t \varepsilon_t \tag{2.9}$$
$$\sigma_t^2 = \gamma + \alpha |r_{t-1}|^2 + \beta \sigma_{t-1}^2 \,, \tag{2.10}$$

where $\gamma > 0$, $\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta < 1$ and the innovations $\varepsilon_t \sim \mathcal{N}(0, 1)$ are distributed according to a standard normal distribution. The condition $\alpha + \beta < 1$ guarantees that the time series generated by the model has finite variance and the parameters α and β represent respectively the degree of surprise and the degree of correlation in the volatility process. The parameter γ allows to model the volatility as a mean-reverting process with expected value $\gamma/(1 - \alpha - \beta)$. All the parameters are usually fixed by performing a constrained maximum likelihood optimization conditioning to r_0 and σ_0. Finally, because of their simplicity, there exist closed form expressions for the forecasts of GARCH(1,1) processes at any time horizon [Dowd, 2005]. However, Monte Carlo methods could be used to obtain forecasts if the complexity of the model was increased.

On the left of Figure 2.5 we show the returns for Microsoft stocks and two times the volatility estimated by a GARCH(1,1) process. It can be appreciated how the GARCH process captures quite accurately the clusters of high and low volatility. On the right of Figure 2.5 it is also displayed a plot with the autocorrelations of the absolute values of the standardized returns. It is noticeable how the correlation present on the right of Figure 2.4 has now vanished.

Even though in a GARCH(1,1) process the conditional distribution for r_t given σ_t is Gaussian as noticed from (2.9), the marginal distribution for r_t will generally show heavy tails [Mikosch and Starica, 2000] due to the stochastic process followed by the time-dependent volatility (2.10). In this way, GARCH processes with normal innovations could, in principle, account for both the time-dependent volatility and the heavy tails of financial returns. However, empirical studies show that after correcting the returns for volatility clustering (standardizing the returns with the volatility estimated by a GARCH process) the residual time series still exhibits heavy tails [Bollerslev, 1987]. This fact, which can be appreciated on Figure 2.6 for the returns of Microsoft stocks, is due to the failure of GARCH models to describe the volatility process accurately enough [Andersen et al., 2003,

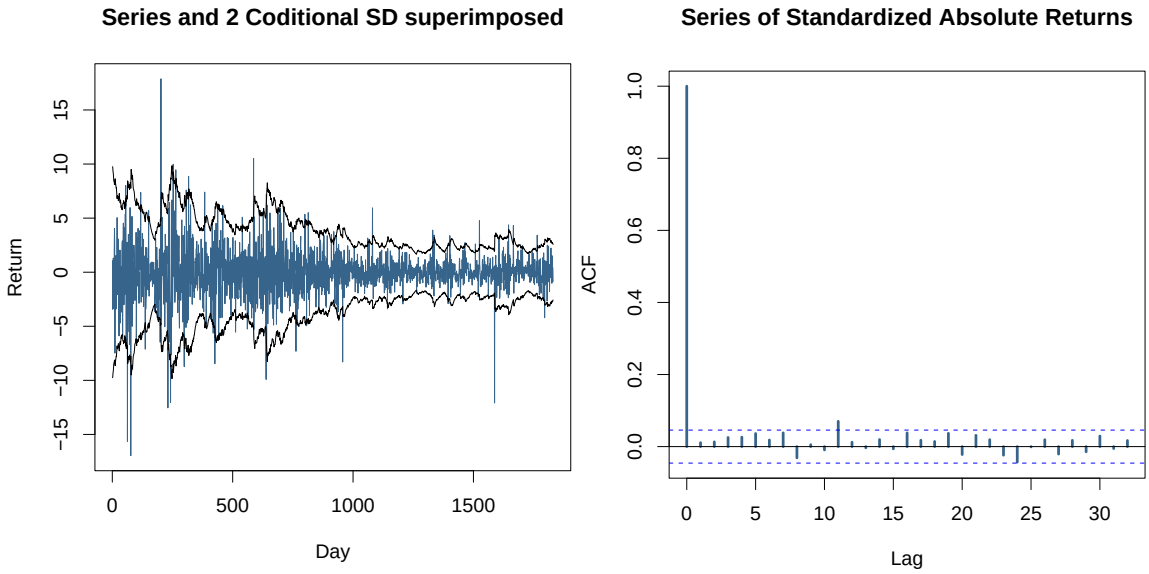

Figure 2.5: Left, Time series with the returns of Microsoft stocks and two times the standard deviation estimated by a GARCH(1,1) model. Right, autocorrelations between $|r_t|/\sigma_t$ and $|r_{t+h}|/\sigma_{t+h}$ where $h = 1, 2, \ldots, 20$ for Microsoft stocks. A GARCH(1,1) process was used to estimate the volatility process σ_t.

Forsberg, 2002] (in part because $|r_{t-1}|^2$ is a bad proxy for estimating past volatility). To solve the problem a wide variety of volatility models similar to GARCH processes but with non-Gaussian heavy-tailed innovations were suggested. Some examples are models with innovations that follow Student distributions [Bollerslev, 1987], stable distributions [Panorska et al., 1995] [Mittnik and Paolella, 2003], normal inverse Gaussian distributions [Forsberg, 2002, Forsberg and Bollerslev, 2002] or the generalized hyperbolic distribution [Prause, 1999].

Other models which show more complexity than GARCH processes have also been used to describe financial time series. They are generally based on different paradigms of machine learning techniques [Bishop, 2006] like input output hidden Markov models [Bengio et al., 2001], mixtures of experts [Suarez, 2002, Vidal and Suarez, 2003] or neural networks [Franke and Diagne, 2006, Weigend and Huberman, 1990]. Nonetheless, because of their higher complexity, the training process for such models is more difficult.

2.3 Backtesting Market Risk Models

Once a probabilistic model for the price of an asset is available it is necessary to validate the model for market risk estimation before putting it into practical use. We might as well want to compare different models to determine which is the best for estimating market risk. Those two tasks can be achieved by means of *backtesting*, which is the application of quantitative methods to determine whether the risk forecasts of a model are consistent with empirical observations [Dowd, 2005].

The general backtesting process is based on a standard hypothesis testing paradigm where the null hypothesis states that the model is consistent with the data. A *statistic*

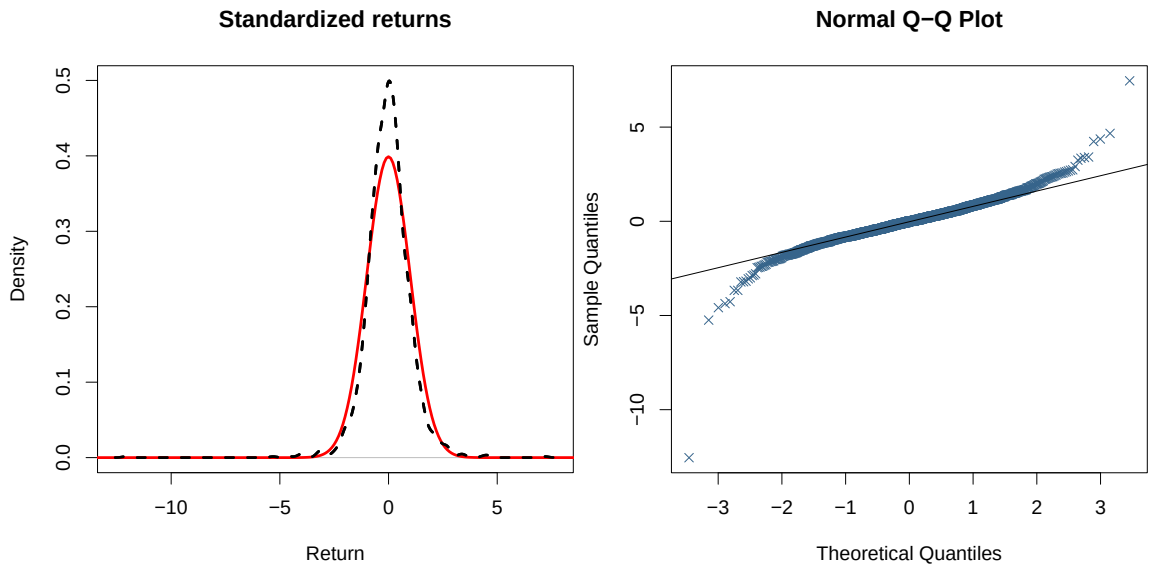

Figure 2.6: Left, kernel density estimate for the returns of Microsoft stocks standardized using the volatility estimated by a GARCH(1,1) model (in black with dots) and standard Gaussian density (in red). It can be appreciated the higher kurtosis of the kernel estimate. Right, normal quantile-quantile plot for the same standardized returns. The tails are still heavier than those of a normal distribution and a Kolmogorov-Smirnov test [Papoulis and Pillai, 2002] rejects the normality hypothesis with a p-value of $2.747 \cdot 10^{-05}$.

can be computed using the specification of the model (with its parameters fitted to some training data) and some empirical observations (test data different from the data used to fit the model). Such statistic should follow a fixed distribution (standard Gaussian or Bernoulli for example) under the null hypothesis and several of those statistics are usually obtained using different sets of training and test data. Finally, it is checked whether a function of the obtained statistics lays within some confidence interval of its distribution (with a fixed probability like 0.95, e.g. $[-1.959, 1.959]$ for the standard Gaussian) or not. In the latter case the model is rejected.

The backtesting process is generally performed throughout a sliding-window experiment. The time series of returns for a financial asset is split in overlapping windows of fixed size where each window is equal to the previous one moved forward one unit in time. The model is trained with the data from each window and then tested on the first return out of the window in order to generate one of the mentioned statistics. Below we describe several backtesting procedures with different ability to identify wrong models.

2.3.1 Test of Exceedances

The testing procedure suggested by [Kupiec, 1995] is probably the most widely used method to validate risk models. This technique is also at the core of the backtesting procedure used by the Bank for International Settlements for determining multiplication factors for capital requirements [Bas, 1996]. The main idea of Kupiec's method consists in checking whether the observed frequency of losses that exceed the α-VaR is consistent with the theoretical frequency $1 - \alpha$. In this way, if our model predicts a VaR at the 0.95 level of 1.000€ for

the next day, we would expect to lose more than 1.000€ with a probability 0.05 on such day.

In the sliding-window framework mentioned before, each statistic would be a variable that takes value 1 if the first return out of the window exceeds the α-VaR and 0 otherwise. Under the null hypothesis, the sum x of n of such statistics follows the binomial distribution

$$\mathcal{B}(x|n,\alpha) = \sum_{i=0}^{x} \left(\begin{array}{c} n \\ i \end{array} \right) (1-\alpha)^i \alpha^{n-i}. \tag{2.11}$$

The model is rejected if the value x lies outside a confidence interval of distribution (2.11) (e.g. $[\mathcal{B}^{-1}(0.025|n,\alpha), \mathcal{B}^{-1}(0.975|n,\alpha)]$ for a significance level of 0.05). The main advantage of the test of exedancees is its simplicity. However, the test lacks power or ability to reject wrong models [Dowd, 2005]. This is so because it throws away valuable information like the sizes of the exceedances or their temporal pattern (exeedances should be independent).

2.3.2 General Tests Based on The Berkowitz Transformation

A more powerful test would be to verify that observed returns follow the distribution that our model predicts. This way, in the sliding-window framework, we would like to check that the first return r_i out of the ith window follows the distribution $\mathcal{P}_i$ predicted by our model (fitted to the data within the ith window), $i = 1, \ldots, n$. The main trouble is that for each window i there is a different predictive distribution P_i and for each of such distributions there is only a single observation r_i, where $r_i \sim \mathcal{P}_i$ under the null hypothesis. The solution to the problem is described in [Berkowitz, 2001] and consists in performing the transformation $\Phi^{-1}(\mathcal{P}_i(r_i))$ of the returns, where Φ^{-1} is the inverse of the cumulative standard Gaussian distribution. As a result, under the null hypothesis we have that $\hat{r}_i = \Phi^{-1}(\mathcal{P}_i(r_i)) \sim \mathcal{N}(0,1)$ and we end up with n points $\{\hat{r}_i\}_{i=1}^n$ that should be standard normal distributed if the forecasts $\mathcal{P}_i$ of our model are right.

Applying the Berkowitz transformation we can now make use of any test for normality over $\{\hat{r}_i\}_{i=1}^n$ to validate the accuracy of our model. Some of the most well-known examples of such tests are the Kolmogorov-Smirnov [Papoulis and Pillai, 2002], the Anderson-Darling [Anderson and Darling, 1954], the Jarque-Bera [Jarque and Bera, 1987] or the Shapiro-Wilk [Shapiro and Wilk, 1965] tests.

2.3.3 Specialized Tests Based on The Functional Delta Method

The tests described in the previous section can be a wrong choice if we are only interested in verifying that the risk measures estimated by our model are right. For example, I would not mind if $\mathcal{P}_i$ is wrong as long as $\rho_{ES}(\mathcal{P}_i)$ is accurate enough. Such a situation can happen if the loss tail of $\mathcal{P}_i$ is correct but the rest of the distribution is wrongly modeled. The statistical tests described in [Kerkhof and Melenberg, 2004] allows us to determine the accuracy of a model for specifically estimating Value at Risk or Expected Shortfall. Those tests relay on the fact that the Berkowitz transformation is a monotonically increasing function and therefore it maps quantiles from one distribution to another[1]. This property creates a direct link between the deviations from normality of the empirical distribution $\mathcal{Q}_n$ for $\{\hat{r}_i\}_{i=1}^n$ and the deviations of the estimates $\{\mathcal{P}_i\}_{i=1}^n$ from the real unknown distributions

[1]Note that the Value at Risk and the Expected Shortfall for a distribution $\mathcal{P}$ can be interpreted as functions of the quantiles of $\mathcal{P}$ [Dowd, 2005].

for $\{r_i\}_{i=1}^n$. For example, if the left tail of $\mathcal{Q}_n$ is similar to the left tail of $\mathcal{N}(0,1)$ it would mean that the left tails of the estimate distributions $\{\mathcal{P}_i\}_{i=1}^n$ are right. The conclusion is that, for a risk measure ρ, the difference between $\rho(\mathcal{Q}_n)$ and $\rho(\mathcal{N}(0,1))$ is related to the accuracy of the model for estimating such risk measure.

The functional delta method [Vaart, 2000] is the mathematical tool that will allow us to determine whether the difference between $\rho(\mathcal{Q}_n)$ and $\rho(\mathcal{N}(0,1))$ is statistically significant or not, if it is too big we will reject the null hypothesis (that our model is an accurate tool for estimating risk by means of the measure ρ). If $\mathcal{Q}_n = n^{-1}\sum_{i=1}^n \delta_{\hat{r}_i}$ is the empirical distribution (here δ_x is the step function centered at x) of a sample $\{\hat{r}_i\}_{i=1}^{i=n}$ such that $\hat{r}_i \sim \mathcal{Q}$, $i = 1,\ldots,n$ and ρ is a functional which is Hadamard differentiable, then the functional delta method states that

$$\sqrt{n}(\rho(\mathcal{Q}_n) - \rho(\mathcal{Q})) \approx \rho'_{\mathcal{Q}}(\sqrt{n}(\mathcal{Q}_n - \mathcal{Q})) \approx \sqrt{n}\frac{1}{n}\sum_{i=1}^n \rho'_{\mathcal{Q}}(\delta_{\hat{r}_i} - \mathcal{Q})\,, \tag{2.12}$$

where the function $x \mapsto \rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q})$ is the influence function of the functional ρ. This influence function can be computed as

$$\rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q}) = \lim_{t\to 0}\frac{d}{dt}\rho((1-t)\mathcal{Q} + t\delta_x) \tag{2.13}$$

and measures the change in $\rho(\mathcal{Q})$ if an infinitesimally small part of $\mathcal{Q}$ is replaced by a point mass at x. In the last step of expression (2.12) we have made use of the linearity property of the influence function. The quantity $\rho(\mathcal{Q}_n) - \rho(\mathcal{Q})$ behaves as an average of independent random variables $\rho'_{\mathcal{Q}}(\delta_{\hat{r}_i} - \mathcal{Q})$ which are known to have zero mean and finite second moments. Therefore, the central limit theorem states that $\sqrt{n}(\rho(\mathcal{Q}_n - \mathcal{Q}))$ has a normal limit distribution with mean 0 and variance $\mathbb{E}_x[\rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q})^2]$ where

$$\mathbb{E}_x[\rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q})^2] = \int \rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q})^2\, d\mathcal{Q}(x) \tag{2.14}$$

We can then use the statistic

$$S_n = \frac{\sqrt{n}(\rho(\mathcal{Q}_n) - \rho(\mathcal{Q}))}{\sqrt{\mathbb{E}_x[\rho'_{\mathcal{Q}}(\delta_x - \mathcal{Q})^2]}} \xrightarrow{d} \mathcal{N}(0,1) \tag{2.15}$$

to determine if the difference $\rho(\mathcal{Q}_n) - \rho(\mathcal{Q})$ is statistically significant or not.

In [Kerkhof and Melenberg, 2004] it is proved that ρ_{VaR} and ρ_{ES} are Hadamard differentiable [Vaart, 2000] and if $\mathcal{Q}$ is standard Gaussian it is easy to show that

$$\mathbb{E}_x[\rho'_{VaR,\mathcal{Q}}(\delta_x - \mathcal{Q})^2] = \frac{\alpha(1-\alpha)}{\phi(\Phi(1-\alpha))^2} \tag{2.16}$$

$$\mathbb{E}_x[\rho'_{ES,\mathcal{Q}}(\delta_x - \mathcal{Q})^2] = \frac{-\Phi(1-\alpha)\phi(\Phi(1-\alpha)) + 1 - \alpha}{(1-\alpha)^2} - \frac{\phi(\Phi(1-\alpha))^2}{(1-\alpha)^2} + \Phi(1-\alpha)^2\frac{\alpha}{1-\alpha} + 2\Phi(1-\alpha)\frac{\phi(\Phi(1-\alpha))}{1-\alpha}\frac{\alpha}{1-\alpha}\,, \tag{2.17}$$

where Φ and ϕ are respectively the standard Gaussian distribution and density functions and the value α represents the fraction of best results used to compute the Value at Risk and the Expected Shortfall.

We point out that it is possible to implement the test of Exceedances described by [Kupiec, 1995] utilizing the functional delta method. We just have to make use of the functional

$$\rho_{Exc}(\mathcal{Q}) = \int_{-\infty}^{\infty} \left(\sum_{t=1}^{n} I_{(-\infty,V]}(y) \right) d\mathcal{Q}(y)\,, \tag{2.18}$$

which represents the average number of elements smaller than the constant V in a sample $\{x_i\}_{i=1}^n$ from distribution $\mathcal{Q}$ (this is, $x_i \sim \mathcal{Q}$, $1 \leq i \leq n$). If we let $V = \mathcal{Q}^{-1}(1-\alpha)$ this functional allows us to implement the binomial test for exceedances over the Value at Risk for the α fraction of best results. We just have to calculate the value $\mathbb{E}_x[\rho'_{Exc,\mathcal{Q}}(\delta_x - \mathcal{Q})^2]$ which turns out to be

$$\mathbb{E}_x[\rho'_{Exc,\mathcal{Q}}(\delta_x - \mathcal{Q})^2] = (1-\alpha)\alpha n^2\,. \tag{2.19}$$

Finally, in [Kerkhof and Melenberg, 2004] it is performed a complete study which compares the power of the tests for Value at Risk, Expected Shortfall and Exceedances. The results indicate that the most powerful test is the one for Expected Shortfall.

Chapter 3

Competitive & Collaborative Mixtures of Experts

This chapter is mainly based on [Hernández-Lobato and Suárez, 2006]. We compare the performance of competitive and collaborative strategies for mixtures of autoregressive experts with normal innovations for conditional risk measurement in financial time series. The prediction of the mixture of collaborating experts is an average of the outputs of the experts. In a competitive mixture the prediction is generated by a single expert. The expert that becomes activated is selected either deterministically (hard competition) or at random, with a certain probability (soft competition). The different strategies are compared in a sliding window experiment for the time series of log-returns of the Spanish stock index IBEX 35, which is preprocessed to account for its heteroskedasticity. Experiments indicate that the best performance for risk estimation is obtained by mixtures with soft competition.

3.1 Introduction

Machine learning and pattern recognition [Bishop, 2006] mainly deal with the problem of learning from examples. This is, given a sample of n paired observations $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ and a new example $\mathbf{x}_{n+1}$ whose paired component is unknown, we are requested to determine a probability distribution for $\mathbf{y}_{n+1}$. In order to solve the problem such distribution must *a priori* be restricted in some way. Otherwise, the set of plausible values for $\mathbf{y}_{n+1}$ would be too big and their distribution would be non-informative. We restrict that distribution by means of a probabilistic model $\mathcal{M}$ (e.g. a neural network or a decision tree) that captures the relationship between $\{\mathbf{x}_i, \mathbf{y}_i\}$ for any i. In this way, the frequentist solution to the learning problem is $\mathcal{P}(\mathbf{y}_{n+1}|\mathbf{x}_{n+1}, \mathcal{M})$ where the parameters of $\mathcal{M}$ have been fixed with $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ (by maximum likelihood for example). We previously saw in Chapter 2 that in the context of measuring market risk we are given a time series of returns $\{r_t\}_{t=1}^n$ from a financial asset and we have to determine the distribution $\mathcal{P}(r_{t+1}|\{r_t\}_{t=1}^n)$. This is a clear problem where we have to learn from examples and therefore machine learning techniques can be used here.

The mixture of experts paradigm [Jacobs et al., 1991] represents an application of the divide-and-conquer principle to the field of machine learning. The idea consists in building a complex model called *mixture* by means of several simpler models called *experts*. The

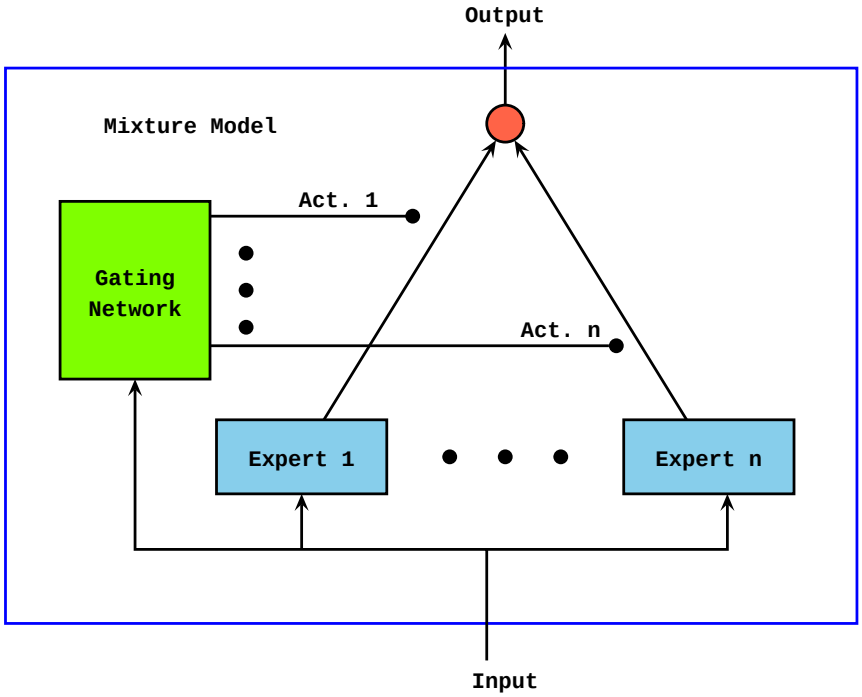

Figure 3.1: Architecture of a mixture of experts model. The input to the whole system is also the input to each expert and to the gating network. The gating network determines for each expert the probability of being activated. The output of the system is the output of the single expert which turns out to be activated.

input space to the mixture is divided into as many different regions as individual experts. A gating network performs this task in such a way that the boundaries between regions are *soft* (data points can belong to different regions with different probabilities). Each of the individual experts specially focuses on one of the regions and the output of the whole system is generated by the expert whose corresponding region is stochastically selected. We display the whole architecture of a mixture of experts on Figure 3.1 and we show the output of a gating network on Figure 3.2. The general procedure followed in order to train a mixture of experts is maximum likelihood. The Expectation Maximization algorithm with iterative reweighted least squares being employed in the M step [Bishop, 2006] can be used for such task if the experts and the gating network are generalized linear [Jordan and Jacobs, 1994]. Hierarchical mixtures of experts [Jordan and Jacobs, 1994] can be obtained if the individual experts are also mixture models. Finally, a fully Bayesian treatment of the hierarchical mixtures of experts architecture is described in [Bishop and Svensén, 2003].

Model complexity is an important factor to take into account while designing a mixture of experts. In machine learning, the frequentist view point to model complexity is known as the *bias-variance* trade-off [Bishop, 2006] which states that the error of a model can be decomposed into the sum of its bias and its variance. It turns out that simple and rigid models have high bias and low variance and flexible and complex models have lower bias but higher variance. This is the reason why very complex models have a worse generalization error and tend to overfit the training data. In a mixture of experts we can keep down

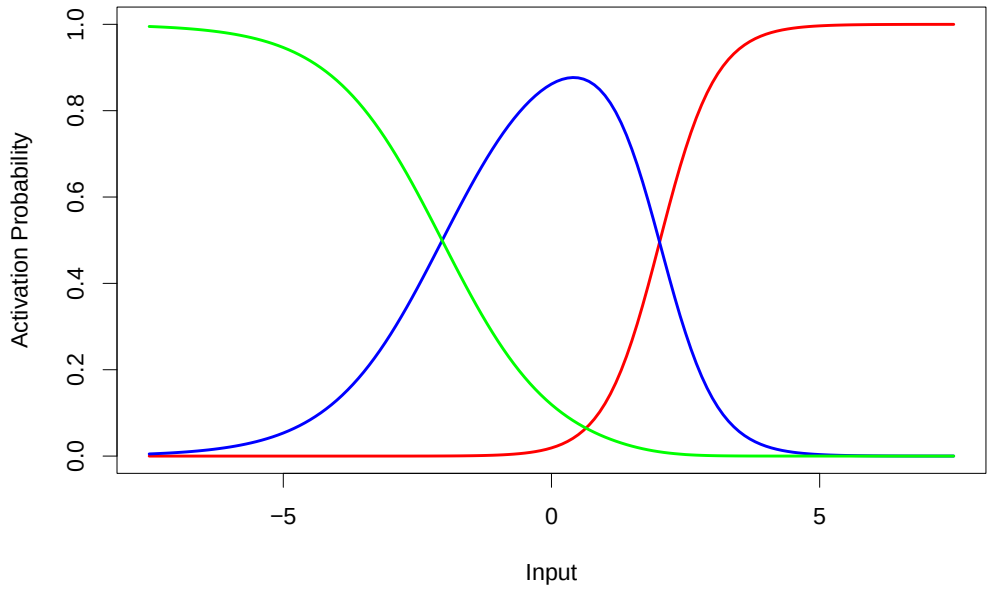

Figure 3.2: Activation probabilities generated by the gating network in a mixture of three experts. It can be appreciated how the input to the mixture is split into three soft regions. The *green* expert is more likely to be activated when the input is less than -2.5, the *blue* expert when the input is between -2.5 and 2.5 and the *red* expert when the input is greater than 2.5. The gating network was generalized linear, a single-layer neural network with a *softmax* [Bishop, 1995] activation function in the three output units.

complexity by reducing the number of experts in the mixture and by choosing simple models for the experts and the gating network (e.g. generalized linear models).

Mixtures of two GARCH processes with autoregressive components were successfully implemented in [Suarez, 2002] to account for correlations, extreme events (heavy tails) and heteroskedasticity (time-dependent variance) in financial time series. Hierarchical mixtures of up to three autoregressive processes were also applied to the analysis of financial time series in [Vidal and Suarez, 2003]. In both of the previous works the experts worked together following a soft competitive strategy as suggested in [Jacobs et al., 1991]. Soft competition implies that, for a fixed input, any expert can be stochastically selected for generating the output of the mixture. Nevertheless, there exist other possible strategies like hard competition (winner-take-all) [Jacobs et al., 1993] or collaboration. A hard competitive strategy requires that, for a fixed input, the output of the mixture is always generated by a single expert which is deterministically selected. On the other hand, collaboration implies that the output of the mixture is a weighted average of the output of each expert.

In this chapter we perform an exhaustive study among mixtures of autoregressive experts that employ collaborative, soft competitive and hard competitive strategies. The ability of the mixtures to accurately estimate market risk using different risk measures is evaluated. For this purpose we carry out a sliding-window experiment over the time series of returns for the Spanish stock index IBEX-35, where the series is specially preprocessed to account for its heteroskedasticity. Finally, we make use of the statistical tests described in Section 2.3 to discriminate among the different models.

3.2 Financial Time Series Models

The future prices of a financial asset that is freely traded in an ideal market are unpredictable. By arguments of market efficiency any expectations on the future evolution of the asset value should be immediately reflected in the current price. Hence, the time series of asset prices follows a stochastic process, where the variations correspond to new unexpected information being incorporated into the market price. We previously mentioned in Section 2.2.1 that, instead of modelling the time series of prices $\{S_t\}_{t=0}^T$, it is common to work with the quasi-stationary series of returns $\{X_t\}_{t=1}^T$. This latter series can be obtained by log-differencing the series of prices, namely

$$X_t = \log S_t - \log S_{t-1} = \log \frac{S_t}{S_{t-1}},\ 1 \leq t \leq T. \tag{3.1}$$

Before formulating a model based on mixtures of autoregressive experts for the series of returns, we perform a transformation to take into account its heteroskedastic structure. We previously described in Section 2.2.3 that GARCH(1,1) processes are among the most successful models for describing the time-dependent structure of volatility in financial time series. If the time series $\{X_t\}_{t=1}^T$ with mean μ follows a GARCH(1,1) model, then

$$\begin{aligned} X_t &= \mu + \sigma_t \varepsilon_t \\ \sigma_t^2 &= \gamma + \alpha(X_{t-1} - \mu)^2 + \beta\sigma_{t-1}^2\,, \end{aligned} \tag{3.2}$$

where $\{\varepsilon_t\}_{t=1}^T$ are iidrv's generated by a $\mathcal{N}(0,1)$ distribution and the parameters γ, α and β satisfy the constraints $\gamma > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta < 1$. Assuming that the time series of returns approximately follows a GARCH(1,1) process, it is then possible to obtain a homoskedastic time series $\{Z_t\}_{t=1}^T$ by performing the normalization

$$Z_t = \frac{X_t - \mu}{\sigma_t}\,, 1 \leq t \leq T, \tag{3.3}$$

where σ_t follows equation (3.2). The parameters μ, γ, α and β are estimated by maximizing the conditional likelihood of the GARCH(1,1) process to the series $\{X_t\}_{t=1}^T$.

Note that the GARCH process is not being trained in an optimal way. In particular the residuals $\{\varepsilon_t\}_{t=1}^T$ are not independent and their distribution is leptokurtic (peaked and heavy-tailed). Nonetheless, given that the deviations from independence and normality are small, the variance σ_t^2 estimated under the hypothesis of normal independent residuals should be a good approximation to the actual variance of the process.

On the left column, Fig. 3.3 displays the graph and autocorrelations of the time series of returns of the Spanish stock index IBEX 35 and, on the right column, the corresponding plots for the normalized time series (3.3). The features of this series are representative of typical time series of financial portfolio returns. The presence of medium-term correlations for the absolute values of the returns $\{X_t\}_{t=1}^T$ is a clear mark of heteroskedasticity. These autocorrelations are not present in the normalized series $\{Z_t\}_{t=1}^T$, which appears to be homoskedastic. We now focus on the sample autocorrelations of the normalized returns $\{Z_t\}_{t=1}^T$ which appear on the bottom-right of Fig. 3.3. It is quite apparent that there is a small but non-negligible correlation at the first lag which could be modeled by a first order autoregressive process [Hamilton, 1994] so that

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + \sigma\varepsilon_t \tag{3.4}$$

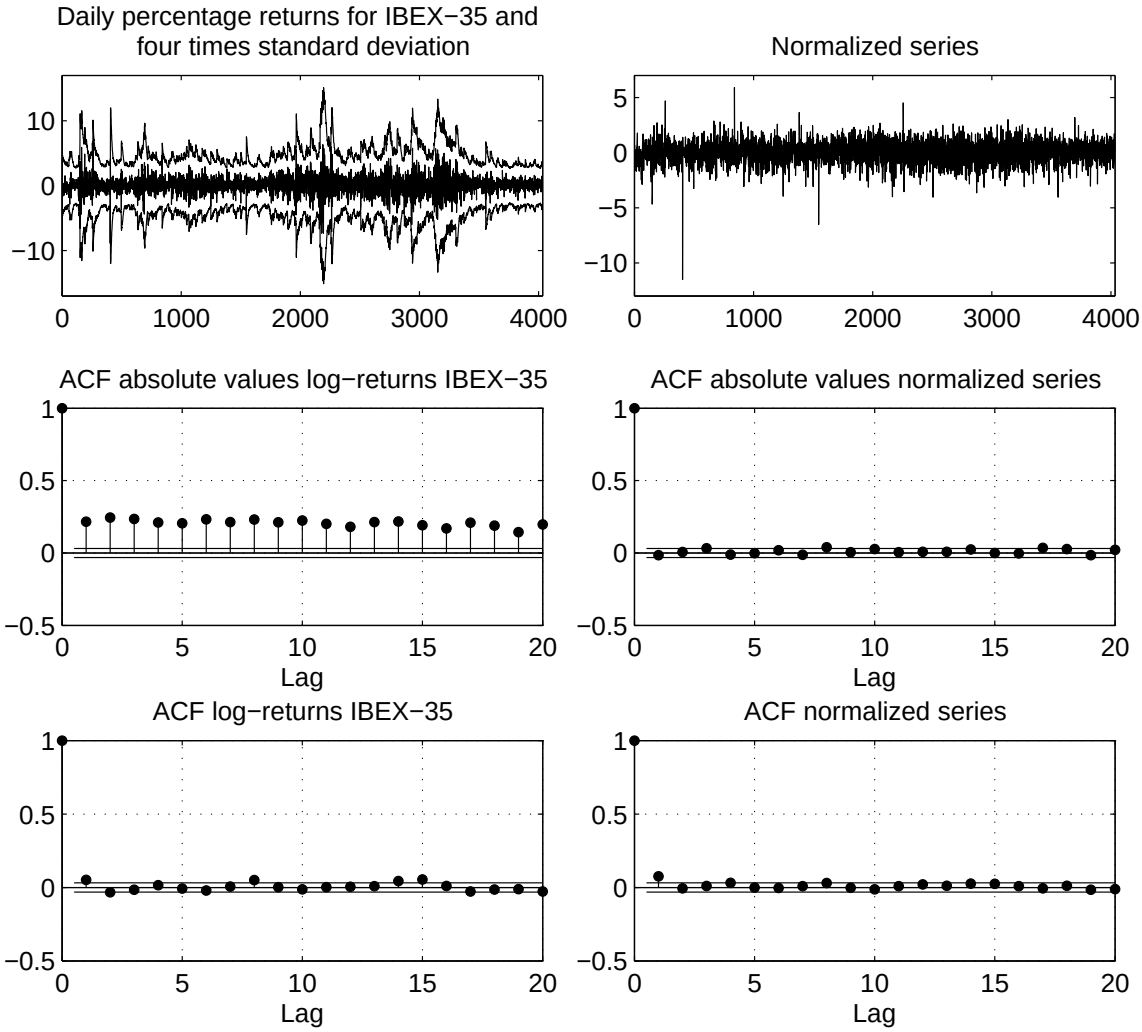

Figure 3.3: Daily returns (multiplied by 100) of the Spanish IBEX 35 stock index from 12/29/1989 to 1/31/2006 (4034 values) provided by [Sociedad de Bolsas, 2006]. Graphs on the left column correspond to the original series of returns (first row) and to the sample autocorrelation functions (ACF) of the absolute values of the returns (second row) and of the returns themselves (third row). The outer lines in the top left plot correspond to $\mu \pm 4\sigma_t$, where μ and σ_t are obtained from a fit to a GARCH(1,1) model (3.2). On the right column it is displayed the corresponding plots for the homoskedastic series obtained after normalization (3.3).

where $|\phi_1| < 1$, $\sigma > 0$ and $\varepsilon_t \sim \mathcal{N}(0,1)$. However, this simple $AR(1)$ process will not be able to account for the leptokurtosis in $\{Z_t\}_{t=1}^T$.

3.3 Mixtures of Autoregressive Experts

Our intention is to model the series of normalized returns $\{Z_t\}_{t=1}^T$ by a mixture of M first order autoregressive processes in a single level with no hierarchy [Jordan and Jacobs, 1994].

These models can be thought of as dynamical extensions of the mixture of Gaussian paradigm, which has been successfully applied to modeling the excess of kurtosis in the unconditional distribution of returns [Kon, 1984]. We intend to give an accurate and robust description of the conditional distribution of returns that can be used for market risk estimation. For this reason we evaluate the performance of the mixture models based not only on point predictions, but also on their capacity to model the whole distribution of returns, especially of extreme events, which are determinant for the calculation of risk measures like Value at Risk or Expected Shortfall.

The way in which the outputs of the $AR(1)$ experts are combined to generate a prediction is controlled by a gating network [Jordan and Jacobs, 1994] with a single layer, see Fig. 3.4. The input for this network is the same as the input for the experts (i.e., the delayed value of the normalized series, Z_{t-1}). The output layer contains as many nodes as the number of experts in the mixture. Their activation is modulated by a *softmax* [Bishop, 1995] function so that the outputs are within the interval $[0, 1]$ and add up to 1. Because of these properties they can be interpreted either as activation probabilities or as weights. In particular, if $\zeta_0^{(m)}$ and $\zeta_1^{(m)}$ are the parameters of the m-th node of the gating network, its outgoing signal is

$$g_m(Z_{t-1}) = \frac{exp(\zeta_0^{(m)} + \zeta_1^{(m)} Z_{t-1})}{\sum_j exp(\zeta_0^{(j)} + \zeta_1^{(j)} Z_{t-1})}, m = 1, 2, \ldots, M. \tag{3.5}$$

Let $\phi_0^{(m)}$, $\phi_1^{(m)}$ and σ_m be the parameters of the m-th $AR(1)$ expert. There are different strategies to determine how the outputs of the experts in the mixture are combined. We consider and compare three different paradigms: collaboration, hard competition and soft competition.

Collaboration. The output of the mixture is a weighted average of the outputs from each of the experts. These weights are determined by the output of the gating network. The output of the mixture is

$$Z_t = \sum_{m=1}^{M} g_m(Z_{t-1}) \left[\phi_0^{(m)} + \phi_1^{(m)} Z_{t-1} + \sigma_m \varepsilon_t \right], \tag{3.6}$$

Hard competition: Experts compete, so that only one expert is active at a given time. The output of the gating network is either 1 for m_t^*, the expert that generates the output, or 0 for the other experts. This strategy was proposed in [Jacobs et al., 1993]

$$Z_t = \phi_0^{(m_t^*)} + \phi_1^{(m_t^*)} Z_{t-1} + \sigma_{m_t^*} \varepsilon_t. \tag{3.7}$$

Soft competition: The output is generated by a single expert. However, in contrast to hard competition, every expert has a probability of being chosen to generate the output of the system. This probability is given by the output of the gating network so that the output of the mixture is

$$Z_t = \sum_{m=1}^{M} \xi_t^{(m)} \left[\phi_0^{(m)} + \phi_1^{(m)} Z_{t-1} + \sigma_m \varepsilon_t \right], \tag{3.8}$$

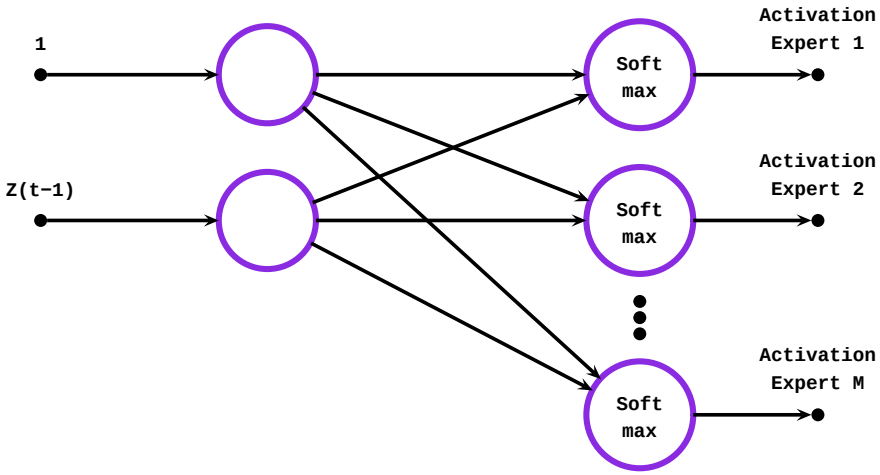

Figure 3.4: Gating network used in the mixtures of autoregressive processes. The gating network is a generalized linear model with a softmax activation function. The input to the gating network is the same as to the autoregressive experts: the value Z_{t-1}. Finally, there is an output unit for each expert in the mixture. Each output unit will determine the activation of its corresponding expert.

where the random variables $\{\xi_t^{(m)},\ m = 1, 2, \ldots, M\}$ take value one with probabilities $\{g_m(Z_{t-1}),\ m = 1, 2, \ldots, M\}$, respectively. At a given time t only one of them can have value 1, and the rest are zero. This strategy was proposed in [Jacobs et al., 1991] and used in [Suarez, 2002, Vidal and Suarez, 2003].

In all these models, we assume that $\{\varepsilon_t\} \sim$ IID $\mathcal{N}(0, 1)$ and require that $\{|\phi_1^{(m)}| < 1;\ m = 1, 2, \ldots, M\}$ to guarantee stationarity.

3.3.1 Training Procedure

In order to fit the parameters of the $AR(1)$ experts and of the gating networks to the time series $\{Z_t\}_{t=1}^T$ we condition the distribution of Z_t to Z_{t-1} and maximize the likelihood function. The expressions of the conditional likelihood for collaborative (CL), hard competitive (HC) and soft competitive (SC) mixtures of M experts, given a series of observations $\{Z_t\}_{t=2}^T$ and an initial value Z_1 are, respectively,

$$\begin{aligned}
\mathcal{L}_{CL}(\Theta; \{Z_t\} \mid Z_1) &= \prod_{t=2}^{T} \psi\left(Z_t, \sum_{m=1}^{M} g_m(Z_{t-1}) AR_m(Z_{t-1}), \sqrt{\sum_{m=1}^{M} g_m^2(Z_{t-1})\sigma_m^2}\right) \\
\mathcal{L}_{HC}(\Theta; \{Z_t\} \mid Z_1) &= \prod_{t=2}^{T} \prod_{m=1}^{M} \psi\left(Z_t,\ AR_m(Z_{t-1}),\ \sigma_m\right)^{g_m(Z_{t-1})} \\
\mathcal{L}_{SC}(\Theta; \{Z_t\} \mid Z_1) &= \prod_{t=2}^{T} \sum_{m=1}^{M} g_m(Z_{t-1}) \psi\left(Z_t,\ AR_m(Z_{t-1}),\ \sigma_m\right), \qquad (3.9)
\end{aligned}$$

where $\psi(x, \mu, \sigma)$ is the normal probability density function with mean μ and standard deviation σ evaluated at x, $\Theta = \{\phi_0^{(m)}, \phi_1^{(m)}, \sigma_m, \zeta_0^{(m)}, \zeta_1^{(m)},\ m = 1, 2, \ldots, M\}$ are the parameters that determine the model and $AR_m(Z_{t-1}) = \phi_0^{(m)} + \phi_1^{(m)} Z_{t-1}$. The previous expressions are maximized by a gradient-descent optimization algorithm applied to the logarithm of the likelihood, taking into account the restrictions of the AR parameters. We also restrict the parameters of the gating network to be in the interval [-50,50] in order to avoid floating point overflows in the calculation of the softmax function. The optimization routine *fmincon* from the Matlab Optimization Toolbox [Mathworks, 2002] is used.

One well-known problem of the maximum likelihood method is that there might be no global maximum of the likelihood function [MacKay, 2003] and such is precisely the case in our mixture models. For example, expert m can get anchored to a single data point in the sample if $\phi_0^{(m)} + \phi_1^{(m)} Z_{t-1} = Z_t$, $\sigma_m \to 0$ and $g_m(Z_{t-1}) > 0$, which causes a divergence in the likelihood function. This problems makes the optimization more and more difficult as the number of experts is increased or as the length of the series $\{Z_t\}$ is reduced. To circumvent this difficulty we adopt the solution proposed in [Hamilton, 1991] and modify the a priori probabilities of the variances of each expert in order to avoid that their values get too close to zero. The prior information is equivalent to a direct observation of T points known to have been generated by each expert and with sample variance $\hat{\sigma}^2$. Accordingly, the logarithmic conditional likelihood of each mixture of AR processes is modified and includes a term of the form

$$\sum_{m=1}^{M} -\frac{T}{2} log(\sigma_m^2) - \frac{T\hat{\sigma}^2}{2\sigma_m^2}\,. \tag{3.10}$$

In our experiments the values $T = 0.1$ and $\hat{\sigma}^2 = 1.5$ are used. The results are not very sensitive to reasonable choices of these parameters.

3.3.2 Validation Procedure

In order to test how accurately the different mixtures of experts fit a time series $\{Z_t\}_{t=1}^{T}$, we follow the approach described in Section 2.3 and apply the Berkowitz transformation. We transform each point from the series $\{Z_t\}_{t=2}^{T}$ to its percentile in terms of the conditional distribution specified by the mixture of experts (ME) and then apply the inverse of the standard normal cumulative distribution function. In this way, we obtain the series $\{Y_t\}_{t=2}^{T}$ so that

$$Y_t = \Psi^{-1}[cdf_{ME}(Z_t|Z_{t-1})], \tag{3.11}$$

where $\Psi^{-1}(u)$ is the inverse of the cumulative distribution function for the standard normal. The cumulative distribution functions of collaborative (CL) and competitive (CP) mixtures evaluated on Z_t and conditioned to Z_{t-1} are

$$\begin{aligned} cdf_{CL}(Z_t \mid Z_{t-1}) &= \Psi\left(Z_t,\ \sum_{m=1}^{M} g_m(Z_{t-1}) AR_m(Z_{t-1}),\ \sqrt{\sum_{m=1}^{M} g_m^2(Z_{t-1})\sigma_m^2}\right) \\ cdf_{CP}(Z_t \mid Z_{t-1}) &= \sum_{m=1}^{M} g_m(Z_{t-1}) \Psi\left(Z_t,\ AR_m(Z_{t-1}),\ \sigma_m\right)\,, \end{aligned} \tag{3.12}$$

respectively, where $\Psi(x, \mu, \sigma)$ is the normal cumulative distribution function with mean μ and standard deviation σ evaluated at x.

Table 3.1: p-values for the different statistical tests. The values highlighted in boldface correspond to the highest p-values for mixtures of 2 and 3 experts, respectively.

#experts	Strategy	VaR 99%	VaR 95%	Exc 99%	Exc 95%	ES 99%	ES 95%
1	-	0.01	0.93	0.07	0.95	$3 \cdot 10^{-8}$	$2 \cdot 10^{-3}$
2	CL	0.01	0.19	0.03	0.39	$2 \cdot 10^{-9}$	10^{-4}
	SC	**0.26**	**0.56**	**0.50**	**0.72**	**0.15**	**0.13**
	HC	0.05	0.48	0.07	0.60	$3 \cdot 10^{-6}$	$2 \cdot 10^{-3}$
3	CL	0.02	0.08	0.02	0.17	$8 \cdot 10^{-7}$	$2 \cdot 10^{-4}$
	SC	**0.44**	**0.31**	**0.39**	**0.54**	**0.16**	**0.10**
	HC	0.02	0.07	0.03	0.39	$4 \cdot 10^{-6}$	$5 \cdot 10^{-4}$

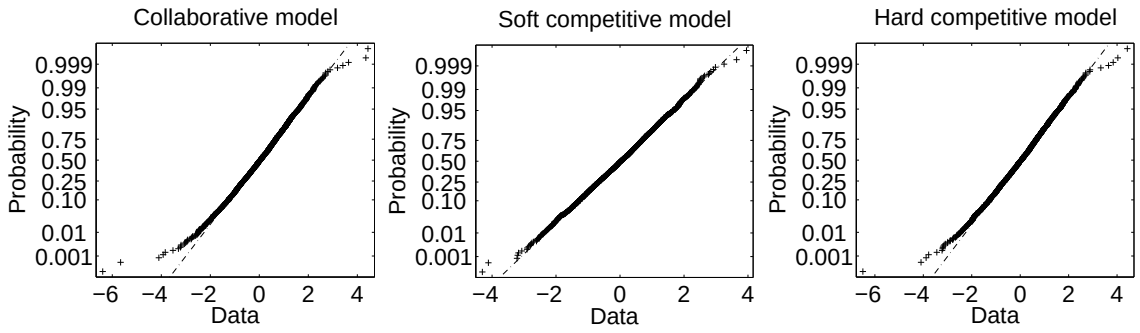

Figure 3.5: Normal probability plots of the transformed sample points for the models with 2 experts. Collaboration and Hard Competition strategies fail to correctly describe the loss tail of the distribution. Plots for the models with 3 experts are very similar.

The Berkowitz transformation is monotonic and preserves the rank order of the normalized returns (i.e. the tails of the distribution in Z_t are mapped into the tails of the distribution in Y_t). If the hypothesis that the values $\{Z_t\}_{t=2}^{T}$, given Z_1, have been generated by our model is correct, then the transformed values $\{Y_t\}_{t=2}^{T}$ should be distributed as a standard normal random variable. In consequence, it is possible to apply statistical tests for normality to these transformed values in order to determine whether the mentioned hypothesis should be rejected. However, instead of utilizing general tests for normality we make use of the tests for Value at Risk, Expected Shortfall and Exceedances described in Section 2.3.3 and based on the functional delta method.

3.4 Experiments and Results

We assess the accuracy of the models investigated by means of a sliding window analysis of the series of IBEX 35 returns (see Fig. 3.3). Each of the models is trained on a window containing 1000 values and then tested on the first out-of-sample point. The origin of the sliding window is then displaced by one point and the training and evaluation processes repeated. To avoid getting trapped in local maxima of the likelihood, we restart the optimization process several times at different initial points selected at random and

retain the best solution. Every 50 iterations in the sliding window analysis, we perform an exhaustive search by restarting the optimization process 2000 times for the HC models, 500 times for the CL and SC and 5 times for the GARCH(1,1) process. In the remaining 49 iterations we use the solution from the previous iteration as an initial value for a single optimization. Once an exhaustive optimization process is completed we restart the previous 50 optimizations (49 simple and 1 exhaustive) using the new solution found as the initial point and replace the older fits if the values of the likelihood are improved.

Table 3.1 displays the results of the statistical tests performed. All the models investigated perform well in the prediction of VaR and exceedances over VaR at a probability level of 95%. At a probability level of 99% the only models that cannot be rejected are mixtures of 2 and 3 experts with soft competition and mixtures of 2 experts with hard competition. The tests for Expected Shortfall are more conclusive and reject all models except mixtures of 2 and 3 experts with soft competition. Furthermore, the p-values obtained for the rejected models are very low, which indicates that they are clearly insufficient to capture the tails of the conditional distribution. This observation is confirmed by the normal probability plots displayed in Fig. 3.5. To summarize, soft competition between experts outperforms the other strategies considered. According to the experiments and statistical tests it is not possible to tell which of the mixtures (2 or 3 experts) performs best. More than 3 experts would probably lead to overfitting.

A Wilcoxon rank test [Wilcoxon, 1945] has been carried out to detect differences in mean square prediction error between models with the same number of experts. The only significant difference appears between soft competitive and collaborative models with 2 experts (i.e. it is possible to reject the hypothesis that those models have the same error. The p-value obtained is 0.03). A similar test indicates that there are no significant differences between the prediction error of a single $AR(1)$ compared with the mixtures of 2 and 3 experts with soft competition.

We analyze why collaboration and hard competition are less accurate than soft competition in capturing the tails of the conditional distribution for returns. The collaborative strategy models the conditional distribution as a single Gaussian whose mean and variance are a weighted average of the means and variances of the Gaussians that correspond to each of the experts. In hard competition the conditional distribution predicted is the Gaussian corresponding to the expert that is active at that particular time. Apparently a single Gaussian distribution, even with time-dependent mean and variance, can not account for the heavy tails of the distribution of returns (see Fig. 3.5). By contrast, the soft competition strategy predicts a time-dependent mixture of Gaussians, one Gaussian per expert. The resulting hypothesis space is more expressive and can account for the excess of kurtosis in the conditional distribution of returns. Hence, we conclude that the proper dynamical extension of the mixture of Gaussians paradigm to model the conditional probability of returns is a mixture of autoregressive experts with soft competition.

3.5 Summary

In this chapter we have investigated three models based on mixtures of $AR(1)$ processes with normal innovations for estimating conditional risk. The models differ only in the way the experts interact to generate a prediction: one model enforces collaboration between experts and the other two competition. In the hard competitive model only one expert (selected

deterministically) is active at each time to generate a prediction. In the soft competitive model, each expert has a probability, given by the output of the gating network, to generate the output of the system. The models are trained over a financial time series previously normalized by means of a GARCH(1,1) process to account for the time dependence in volatility. The accuracy of the models is tested by performing a sliding window experiment over the normalized daily returns of the Spanish stock index IBEX-35 from 12/29/1989 to 1/31/2006. Specialized statistical tests are carried out to measure the ability of each model to give accurate estimates of conditional risk measures, such as Value at Risk and Expected Shortfall. The results obtained indicate that the model with soft competition outperforms the other models. The relatively poor performance of the collaborative and hard competitive models can be ascribed to the fact that they remain within the normal paradigm, which is insufficient to capture the distribution of events in the tails of the distribution. The soft competitive strategy naturally extends the mixture of Gaussian paradigm and is able to model the distribution of extreme events.

Chapter 4

GARCH Processes with Non-parametric Innovations

This chapter is mainly based on [Hernández-Lobato et al., 2007]. Here, we introduce a procedure to estimate the parameters of GARCH processes with non-parametric innovations by maximum likelihood. We also design an improved technique to estimate the density of heavy-tailed distributions with real support from empirical data. The performance of GARCH processes with non-parametric innovations is evaluated in a series of experiments on the daily logarithmic returns of IBM stocks. These experiments demonstrate the capacity of the improved processes to yield a precise quantification of market risk. In particular, the model provides an accurate statistical description of extreme losses in the conditional distribution of daily logarithmic returns.

4.1 Introduction

We previously saw in Section 2.2.3 that standard GARCH processes can account for the time-dependent volatility in financial time series very accurately. However, after fitting a GARCH process to a series of returns, it is noticed that the empirical residuals are not Gaussian. In particular, the residuals are more peaked and show heavier tails than the Gaussian distribution, a fact which can be appreciated on Figure 2.6. For this reason, ad hoc GARCH processes with non-Gaussian heavy-tailed innovations were proposed, we refer to Section 2.2.3 for a review. All those models have one trait in common, they all constrain the innovations to belong to a *parametric* family of distributions. This is, the distribution for the innovations can be fully specified with only a few parameters, generally two or three. The advantage is that the estimation process is simpler. However, if the empirical innovations deviate significantly from the selected family the model will fail to provide an accurate measurement of risk. In this chapter we attempt to model the innovations of GARCH processes in a non-parametric way. In this way, we make very few assumptions about the form of the distribution for the innovations, expecting to obtain better risk estimation results. We point out that asset returns are unidimensional and consequently, non-parametric techniques applied to returns will not suffer the curse of dimensionality [Bishop, 2006].

Kernel density estimates [Parzen, 1962, Bishop, 2006] are one of the most successful

tools for performing non-parametric density estimation. Given a sample of data points $\{x_i\}_{i=1}^n$ where $x_i \sim f$, $i = 1, \ldots, n$, a kernel density estimate $\hat{f}_n$ for f would be

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{4.1}$$

where $h > 0$ is a *smoothing parameter* and K is a *kernel function* which satisfies the constraints $K(x) \geq 0$ and $\int K(x)dx = 1$ so that the resulting density estimate $\hat{f}_n$ is nonnegative everywhere and integrates to one. The properties of kernel density estimates have been thoroughly studied by several authors like [Silverman, 1986, Nadaraya, 1989, Devroye and Györfi, 1985, Devroye et al., 1997]. In the next paragraph we review some of the most important results obtained in [Silverman, 1986].

One way to measure the accuracy of a kernel estimate is through its mean integrated square error (MISE) which is defined as

$$\begin{aligned} \text{MISE}(\hat{f}_n) &= \int \mathbb{E}[(\hat{f}_n(x) - f(x))^2]\, dx & (4.2)\\ &= \int (\mathbb{E}[\hat{f}_n(x)] - f(x))^2\, dx + \int \mathbb{E}[(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)])^2]\, dx & (4.3)\\ &= \text{Bias}(\hat{f}_n) + \text{Variance}(\hat{f}_n), & (4.4) \end{aligned}$$

where the expectations are taken with respect to $\{x_i\}_{i=1}^n$, $x_i \sim f$, $i = 1, \ldots, n$. If the kernel K is a symmetric function such that $\int xK(x)\, dx = 0$ and $\int x^2 K(x)\, dx = k_2 \neq 0$ and the density f has continuous derivatives of all orders required, it can be shown that

$$\begin{aligned} \text{Bias}(\hat{f}_n) &\approx \frac{1}{4} h^4 k_2^2 \int f''(x)^2\, dx, & (4.5)\\ \text{Variance}(\hat{f}_n) &\approx \frac{1}{nh} \int K(x)^2\, dx. & (4.6) \end{aligned}$$

The first conclusion is that $\text{MISE}(\hat{f}_n) \to 0$ if and only if $h \to 0$ and $nh \to \infty$. The second conclusion is that the value for h that minimizes $\text{MISE}(\hat{f}_n)$ is

$$h_{opt} = n^{-\frac{1}{5}} k_2^{-\frac{2}{5}} \left\{ \int K(x)^2\, dx \right\}^{\frac{1}{5}} \left\{ \int f''(x)^2\, dx \right\}^{-\frac{1}{5}}. \tag{4.7}$$

The third conclusion is that if $h = h_{opt}$ then $\text{MISE}(\hat{f}_n)$ converges like $\mathcal{O}(n^{-\frac{4}{5}})$. However, we note that 4.7 depends on $\int f''(x)\, dx$, a value which we do not know. If we assume that f is Gaussian with standard deviation σ and we take K to be a standard Gaussian kernel, we have that

$$h_{opt} \approx 1.06\, n^{-\frac{1}{5}} \sigma \approx 1.06\, n^{-\frac{1}{5}} \hat{\sigma}, \tag{4.8}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the sample $\{x_i\}_{i=1}^n$. Instead of using (4.8), [Silverman, 1986] suggests using his *rule of thumb* which should behave quite well for a wide range of densities. The rule is

$$h_{opt} = 0.9\, min\left\{ \hat{\sigma}, \frac{\text{IRQ}}{1.34} \right\} n^{-\frac{1}{5}}, \tag{4.9}$$

where IRQ is the interquartile range of the sample $\{x_i\}_{i=1}^n$. There are better ways of choosing a value for the smoothing parameter than Silverman's rule, for example the *plug in* method described by [Sheather and Jones, 1991]. However, Silverman's rule is simpler and accurate enough.

Because of their good properties for non-parametric density estimation, we make use of kernel density estimates for modeling the distribution of the innovations in GARCH processes. However, kernel estimates are no *panacea* and it will be seen that they have trouble modelling heavy-tailed densities. Because of this, we suggest a new family of GARCH processes with non-parametric innovations whose parameters will be estimated in a transformed space to account for the leptokurtosis of the return distribution. It will be seen that the new models provide an accurate statistical description of extreme losses in the tail of the conditional distribution of daily returns. Furthermore, they can be used to compute measures of risk which are very precise.

4.2 Financial Time Series Models

Consider the time series of daily prices of a financial asset $\{S_t\}_{t=0}^T$. In general, it is common to model the time series of logarithmic returns

$$r_t = 100 \cdot log\left(\frac{S_t}{S_{t-1}}\right), t = 1, 2, \ldots \tag{4.10}$$

which, as described in Section 2.2.1, have more desirable properties than prices themselves. Typically, the autocorrelations between returns are small and short-lived. By contrast, the time series of financial asset returns have a time-dependent volatility: Large price variations (either positive or negative) are likely to be followed by price variations that are also large. We previously mentioned in Section 2.2.2 that such phenomenon can be observed in a plot of the autocorrelations in the powers of the absolute values of returns

$$C_\delta = corr(|r_{t+h}|^\delta, |r_t|^\delta), h = 1, 2, \ldots \tag{4.11}$$

These correlations turn out to be positive for large delays and generally obtain their highest value for $\delta = 1$, a fact which can be appreciated on Figure 4.1 for the returns of IBM stocks.

As already described, GARCH processes are time-series models that can successfully account for the time-dependent structure of the volatility in financial time series. In this work we consider power GARCH(1,1,δ) processes [Ding et al., 1993], which are a generalization of the standard GARCH(1,1) process. A time series $\{r_t\}_{t=1}^T$ follows a power GARCH(1,1,δ) process with normal innovations if

$$\begin{aligned} r_t &= \sigma_t \varepsilon_t \\ \sigma_t^\delta &= \gamma + \alpha |r_{t-1}|^\delta + \beta \sigma_{t-1}^\delta , \end{aligned} \tag{4.12}$$

where $0 < \delta \leq 2$, $\gamma > 0$, $\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta < 1$ and the innovations $\varepsilon_t \sim \mathcal{N}(0, 1)$ are distributed according to a standard Gaussian. The condition $\alpha + \beta < 1$ is sufficient to guarantee the existence of $\mathbb{E}[\sigma_t^\delta]$ and $\mathbb{E}[|r_t|^\delta]$ for any value of δ [Ding et al., 1993]. Power GARCH processes take into account the correlation between $|r_{t+1}|^\delta$ and $|r_t|^\delta$. The usual choice is $\delta = 2$, which corresponds to the standard GARCH process. The parameters of the model are then estimated by maximizing the model likelihood with standard optimization

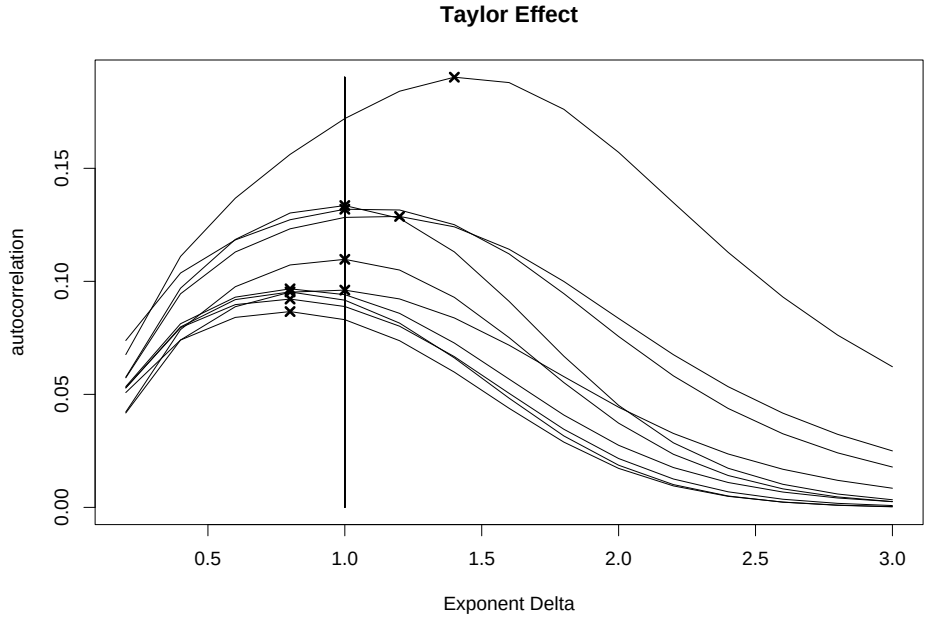

Figure 4.1: Plots of the autocorrelation function between $|r_t|^\delta$ and $|r_{t+h}|^\delta$ for $h \in \{1, ..., 10\}$. Each curve corresponds to a different value of $h = 1, 2, \ldots, 10$. The series $\{r_t\}_{t=1}^T$, with $T = 9190$ values, corresponds to the standardized returns of IBM stocks from 1962/07/03 to 1998/12/31. The maximum of each function is shown with a cross. All maxima are located close to the value $\delta = 1$.

algorithms that generally employ gradient descent. Nevertheless, an empirical analysis of the data (see Fig. 4.1) suggests using a value of δ closer to 1. In the experiments carried out in the present research, the value $\delta = 1$ is chosen and an optimization method that does not need to compute the gradient is used to maximize the likelihood. The value of $\delta = 1$ is preferred to $\delta = 2$ to allow for the possibility of infinite-variance models for the innovations (for instance, if the innovations are assumed to follow a stable distribution [Nolan, 2002], whose second and higher moments do not exist).

If r_t exhibits correlations with r_{t+h} for a range of values of h (usually h small) it is possible to add an autoregressive component to the GARCH process. The simplest one is an AR(1) process

$$\begin{aligned} r_t &= \phi_0 + \phi_1 r_{t-1} + \sigma_t \varepsilon_t \\ \sigma_t &= \gamma + \alpha |r_{t-1} - \phi_0 - \phi_1 r_{t-2}| + \beta \sigma_{t-1} \,, \end{aligned} \tag{4.13}$$

where $|\phi_1| < 1$ is a necessary condition for stationarity. Assuming Gaussian innovations, the likelihood for the parameters of this GARCH process given a time series $\{r_t\}_{t=1}^T$ is

$$\mathcal{L}(\boldsymbol{\theta} | \{r_t\}_{t=1}^T) = \prod_{t=1}^T \psi\left(\frac{u_t}{\sigma_t}\right) \frac{1}{\sigma_t} \,, \tag{4.14}$$

where $\boldsymbol{\theta} = \{\phi_0, \phi_1, \gamma, \alpha, \beta\}$, $u_t = r_t - \phi_0 - \phi_1 r_{t-1}$ is the empirical autoregressive residual and $\psi(x)$ is the standard Gaussian density function. To evaluate (4.14) we define $u_0 = 0$,

$u_1 = r_1 - \hat{\mu}$ and $\sigma_0 = \hat{\sigma}$ where $\hat{\sigma}$ and $\hat{\mu}$ are the sample standard deviation and sample mean of $\{r_t\}_{t=1}^T$.

If a GARCH process like the one described in (4.13) is fitted to the daily returns of a financial asset, the model captures the time-dependent volatility quite accurately. However, the empirical innovations of the model $u_t/\sigma_t = (r_t - \phi_0 - \phi_1 r_{t-1})/\sigma_t$ do not follow a standard Gaussian distribution. In particular, the empirical distribution of the residuals has larger kurtosis (i.e., heavier tails) than the Gaussian distribution. This mismatch in the loss tail of the distribution for the innovations reflects the fact that the model is unable to correctly describe extreme events, causing it to severely underestimate measures of market risk such as Value at Risk or Expected Shortfall. We propose to address this shortcoming by estimating the parameters of the GARCH process (4.13) assuming that the innovations ε_t follow an unknown distribution f which is estimated in a non-parametric way.

4.3 GARCH Processes with Non-parametric Innovations

The goal is to maximize (4.14) replacing ψ by f, the (unknown) density function for the innovations $\{\varepsilon_t = u_t/\sigma_t\}_{t=1}^T$. One possible solution to the problem is to use a non-parametric kernel density estimate $\hat{f}$:

$$\hat{f}(x\,;\,\{c_i\}_{i=1}^N) \;=\; \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-c_i}{h}\right), \tag{4.15}$$

where $\{c_i\}_{i=1}^N$ are the centers of N Gaussian kernels denoted by K and h is the smoothing parameter. If $c_1, ..., c_N \sim f$, $h \to 0$ and $Nh \to \infty$ as $N \to \infty$, then $\hat{f}$ can approximate f to any degree of precision. The parameter h can be automatically fixed (under some assumptions) as a function of $\{c_i\}_{i=1}^N$ following Silverman's *rule of thumb*. The function $\hat{f}$ can be seen as a radial basis function network [Bishop, 1995] that approximates the actual density. The network uses Gaussian densities as basis functions with standard deviation h and centers $\{c_i\}_{i=1}^N$ (the only free parameters); there is no bias term in the network and the output weights are $1/N$. Alternatively, $\hat{f}$ can be interpreted as a mixture of N Gaussians model.

If ψ is replaced by $\hat{f}$ in (4.14) then

$$\mathcal{L}(\boldsymbol{\theta}, \{c_i\}_{i=1}^N | \{r_t\}_{t=1}^T) = \prod_{t=1}^{T} \hat{f}\left(\frac{u_t}{\sigma_t}\,;\,\{c_i\}_{i=1}^N\right)\frac{1}{\sigma_t}, \tag{4.16}$$

where the parameters of the model are $\boldsymbol{\theta}$ and $\{c_i\}_{i=1}^N$. Assuming that there are as many Gaussians with centers $\{c_i\}_{i=1}^N$ as training data $\{r_t\}_{t=1}^T$ (N = T) and that h is held fixed, (4.16) can be maximized by iterating the following steps: First, (4.16) is maximized with respect to each of the $\{c_t\}_{t=1}^T$ holding $\boldsymbol{\theta}$ fixed. This is accomplished by setting $c_t = u_t/\sigma_t$, $t = 1, 2, \ldots, T$. Second, (4.16) is maximized with respect to $\boldsymbol{\theta}$ (holding $\{c_t\}_{t=1}^T$ fixed) using a non-linear optimizer. This last step is the same as the one used for calibrating a standard GARCH process. A maximum of (4.16) is obtained by iterating these steps until the likelihood shows no significant change in two consecutive iterations. Note that the parameter h in the model cannot be determined by maximum likelihood. The reason is that the value of the likelihood increases without bound as h tends to zero. A suitable value for h is computed on each iteration at the end of the first step using Silverman's *rule*

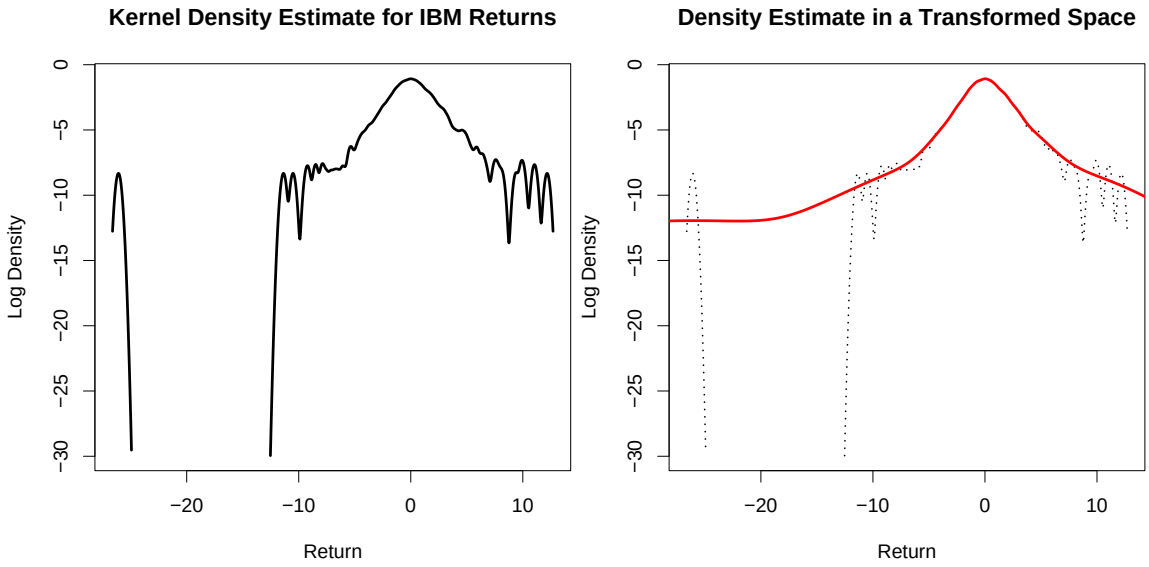

Figure 4.2: Left, logarithm of a kernel density estimate for the unconditional distribution of the returns of IBM stocks. The sample used to compute the estimate was the returns from 1962/07/03 to 1998/12/31. It can be noticed how the estimate fails to model the left and right tails of the distribution where several bumps appear. Each bump corresponds to a Gaussian kernel centered at an extreme observation. The smoothing parameter was chosen by means of Silverman's rule. Right, kernel density estimate performed over the transformed returns. A stable density was used in the transformation as described in the main text. It is noticeable how this time the tails are correctly modeled.

of thumb over the sample $\{c_t\}_{t=1}^T$. If T is sufficiently large, at convergence, $\hat{f}$ should be an accurate non-parametric estimate of the true distribution f. Note that after the first step in each iteration, $\hat{f}$ can be interpreted as a kernel density estimate of a density g where $u_1/\sigma_1, ..., u_T/\sigma_T \sim g$. This is the basis for the analysis carried out in the next section.

4.3.1 Density Estimation for Heavy-tailed Distributions

Density estimates based on Gaussian kernels like (4.15) do not work very well with heavy-tailed samples. The difficulties are particularly severe in the modeling of extreme events, which are crucial for the quantification of risk. The origin of this shortcoming is that samples from heavy-tailed distributions usually include very few points in the tails and kernel estimates tend to assign very low probability to regions with sparse samples. This fact can be observed on the left of Figure 4.2 for a kernel density estimate of the unconditional distribution for the returns of IBM stocks.

A solution to this problem consists in performing the density estimation in a transformed space where the kernel estimate is believed to perform better [Wand et al., 1991]. If we assume that such a transformation is known, the non-parametric estimate $\hat{f}$, which models the density $f \sim c_1, \ldots, c_N$ in the transformed space, is

$$\hat{f}(x\,;\,\{c_i\}_{i=1}^N\,;\,g_{\boldsymbol{\lambda}}) \;=\; |g'_{\boldsymbol{\lambda}}(x)|\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{g_{\boldsymbol{\lambda}}(x)-g_{\boldsymbol{\lambda}}(c_i)}{h}\right), \tag{4.17}$$

where $g_{\boldsymbol{\lambda}}$ is a function which maps the original space to the transformed space and $\boldsymbol{\lambda}$ is a set of parameters that specify $g_{\boldsymbol{\lambda}}$ within a family of monotonic increasing transformations. This process of estimating the density in a transformed space is similar to standard density estimation based on kernels with varying widths [Wand et al., 1991].

By arguments similar to those in [Silverman, 1986] and already described in Section 4.1, if the smoothing parameter h is optimally chosen, then the MISE of the density estimate in the transformed space is proportional to

$$\left\{ \int f''_{g_{\boldsymbol{\lambda}}}(x)^2 \, dx \right\}^{\frac{1}{5}} , \tag{4.18}$$

where $f_{g_{\boldsymbol{\lambda}}}$ is the density followed by the data in the transformed space, this is $f_{g_{\boldsymbol{\lambda}}} \sim g_{\boldsymbol{\lambda}}(c_1), \ldots, g_{\boldsymbol{\lambda}}(c_N)$. It turns out that the Gaussian density provides a more than reasonable low value for (4.18) [Wand et al., 1991]. As a result, the key aspect that should be considered in the choice of $g_{\boldsymbol{\lambda}}$ is that if $c_i \sim f$, then $g_{\boldsymbol{\lambda}}(c_i)$ should follow a distribution close to normality. A trivial transformation that achieves this is $\Phi^{-1}(F(c_i)) \sim \mathcal{N}(0,1)$, where F is the cumulative distribution for f and Φ^{-1} is the inverse of the standard Gaussian distribution. The difficulty is that if we knew F then we would also know f as $f = F'$ and the estimation process would not be necessary. Nevertheless, it is possible to approximate F by means of a parametric distribution $\hat{F}_{\boldsymbol{\lambda}}$ that can account for the heavy tails in F. The parameters $\boldsymbol{\lambda}$ of $\hat{F}_{\boldsymbol{\lambda}}$ can be estimated by maximum likelihood to $\{c_i\}_{i=1}^N$. The final transformation is $g_{\boldsymbol{\lambda}}(x) = \Phi^{-1}(\hat{F}_{\boldsymbol{\lambda}}(x))$. In this manner, a hybrid estimation is performed where the parametric model corrects the tails of the non-parametric estimate.

To incorporate this idea into the algorithm proposed in Section 4.3 the first step of such algorithm is divided into two parts. In the first one, the parameter $\boldsymbol{\lambda}$ of the transformation $g_{\boldsymbol{\lambda}}$ is found by maximizing the likelihood of $\hat{F}_{\boldsymbol{\lambda}}$ over the observed values $\{u_t/\sigma_t\}_{t=1}^T$. In the second part, the density of $\{u_t/\sigma_t\}_{t=1}^T$ is estimated by means of (4.17) where h is now determined using Silverman's *rule of thumb* in the *transformed space* as a function of $\{g_{\boldsymbol{\lambda}}(u_t/\sigma_t)\}_{t=1}^T$. The second step of the algorithm remains unchanged: $\hat{f}(x\,;\{c_i\}_{i=1}^N)$ is replaced in (4.16) by (4.17). One technical problem that remains is that the location and scale of $u_1, ..., u_T$ can be fixed in two different ways. One through the parameters $\boldsymbol{\theta}$ of the GARCH process and another one through $\hat{f}$ and the centers of the kernels $\{c_i\}_{i=1}^N$. This can lead to convergence problems and should be avoided. To address this problem $\hat{f}$ is forced to have always the same location and scale. This is achieved by standardizing $\{u_t/\sigma_t\}_{t=1}^T$ before the first step on each iteration of the algorithm. The pseudocode of the whole process followed to train a GARCH process with non-parametric innovations is displayed on Figure 4.3.

Family of Transformations

A parametric family of distributions $\hat{F}_{\boldsymbol{\lambda}}$ needs to be selected to fully specify the family of transformations $g_{\boldsymbol{\lambda}}(x) = \Phi^{-1}(\hat{F}_{\boldsymbol{\lambda}}(x))$. The choice used in our research is the family of stable distributions [Nolan, 2002] which should have enough flexibility to account for the empirical properties of the marginal distribution of returns [Cont, 2001]. This family of distributions is parameterized in terms of a location parameter, a scale parameter, a parameter describing the decay of the tails and, finally, a parameter allowing each tail to have a different behavior. Some examples of stable densities are displayed on Figure 4.4. We

Input: a series of returns $\{r_t\}_{t=1}^T$.

Output: The parameters $\boldsymbol{\theta}$ of a GARCH process and a density $\hat{f}$ for its innovations.

1. Initialize $\boldsymbol{\theta}$ randomly, $\mathcal{L}_{old} \leftarrow \infty$, $\mathcal{L}_{new} \leftarrow -\infty$.
2. while $|\mathcal{L}_{new} - \mathcal{L}_{old}| <$ tolerance.
   (a) Obtain the standardized scaled residuals $\{u_t/\sigma_t\}_{t=1}^T$ given $\boldsymbol{\theta}$ and $\{r_t\}_{t=1}^T$.
   (b) $\boldsymbol{\lambda} \leftarrow max_{\boldsymbol{\lambda}} \left\{ \prod_{t=1}^T \log \hat{F}'_{\boldsymbol{\lambda}}(u_t/\sigma_t) \right\}$.
   (c) $g_{\boldsymbol{\lambda}}(x) \leftarrow \Phi^{-1}(\hat{F}_{\boldsymbol{\lambda}}(x))$.
   (d) Obtain h by means of Silverman's rule over $\{g_{\boldsymbol{\lambda}}(u_t/\sigma_t)\}_{t=1}^T$.
   (e) $\hat{f}(x) \leftarrow |g'_{\boldsymbol{\lambda}}(x)| T^{-1} h^{-1} \sum_{t=1}^T K((g_{\boldsymbol{\lambda}}(x) - g_{\boldsymbol{\lambda}}(u_t/\sigma_t))/h)$.
   (f) Update $\boldsymbol{\theta}$ with the maximum likelihood estimate of a GARCH process with $\hat{f}$-innovations over the series $\{r_t\}_{t=1}^T$.
   (g) $\mathcal{L}_{old} \leftarrow \mathcal{L}_{new}$.
   (h) Store in $\mathcal{L}_{new}$ the log-likelihood of the model obtained at step (f).
3. Return $\boldsymbol{\theta}$ and $\hat{f}$.

Figure 4.3: Pseudocode with the steps followed in order to train a GARCH processes with non-parametric innovations.

also show on the right of Figure 4.2 a kernel estimate of the unconditional density for the transformed returns of IBM stocks, the stable family was used to perform the transformation as described above. Finally, other possible alternatives to the stable family would be normal inverse Gaussian distributions [Barndorff-Nielsen, 1997], hyperbolic distributions [Eberlein and Keller, 1995] or the superclass of these last two: the generalized hyperbolic distribution [Prause, 1999].

4.4 GARCH Processes with Stable Innovations

We point out that at each iteration of the proposed algorithm we are estimating the density g, where $u_1/\sigma_1, ..., u_T/\sigma_T \sim g$, twice. First, in a parametric way assuming g belongs to the family of stable distributions and, second, in a non-parametric way making use of the former estimate. It is possible that using only the first parametric estimate a better or equivalent solution than the one obtained by means of the non-parametric technique is reached. This would happen if the innovations of the GARCH process in (4.13) were actually stable. In this case ε_t would follow a stable distribution with 0 as location parameter and 1 as scale parameter: $\varepsilon_t \sim \mathbf{S}(a, b, 1, 0; 0)$, where a is the index of stability, or characteristic exponent, b is the skewness parameter, and the scale and location parameters are held fixed and equal to 1 and 0, respectively. The last parameter with value 0 indicates that the first parameterization proposed by Nolan [Nolan, 2002] is used. To account for this situation two alternative models are compared in our experiments. One where the distribution of the innovations is estimated in a non-parametric way as described in Section 4.3 and another one where the innovations are assumed to be stable. In this way, by comparing

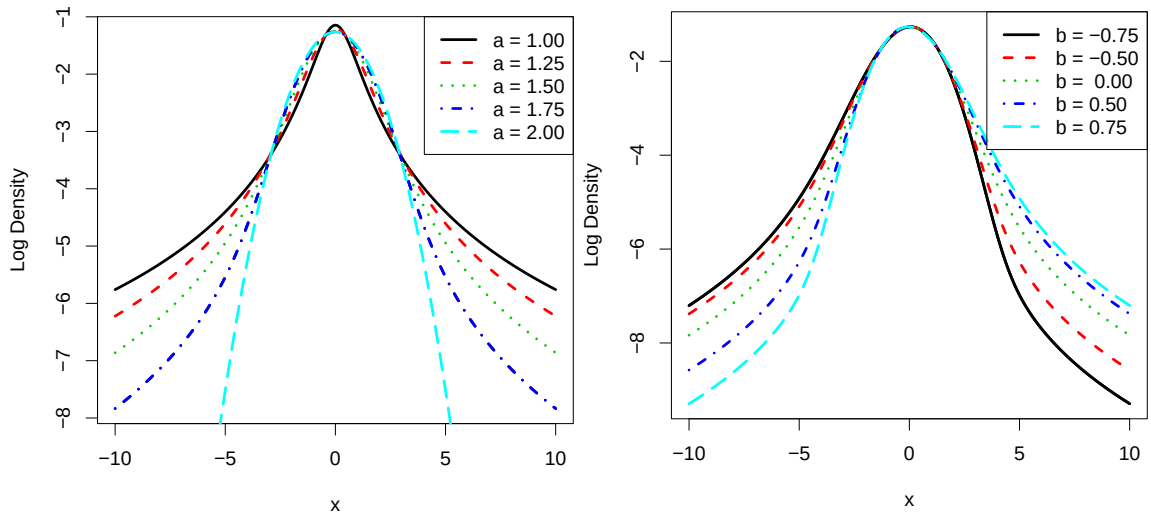

Figure 4.4: Plots of the logarithm of different stable densities with location 0 and scale 1 where the first parameterization proposed by [Nolan, 2002] is used. Left, plots of densities with parameter a (α in [Nolan, 2002] and with range from 0 to 2) for the heaviness of the tails equal to 1, 1.25, 1.5, 1.75 and 2 and parameter b (β in [Nolan, 2002] and with range from -1 to 1) for the asymmetry of the tails equal to 0 (symmetric tails). The stable density with parameter $a = 2$ has Gaussian tails. Right, plots of densities with parameter b equal to -0.75, -0.5, 0, 0.5, 0.75 and parameter a equal to 1.75. Negative values for b give more weight to the left tail and positive values give more weight to the right tail.

the performance of both models it is possible to determine whether implementing the non-parametric estimate represents an actual improvement. The parameters of the GARCH model with stable innovations, $\boldsymbol{\theta}$, a and b are also estimated by maximum likelihood.

4.5 Model Validation and Results

To assess the capacity of the proposed models to accurately quantify market risk the general backtesting procedure described in Section 2.3 is followed. We perform a sliding window experiment on the daily returns of IBM stocks from 1962/07/03 to 1998/12/31, a total of 9190 measurements (see Fig. 4.6). From the series of returns 8190 overlapping windows with T = 1000 elements each are generated. Each window is equal in length to the previous window but is displaced forward by one time unit (one day). The parameters of the models are estimated by maximum likelihood using the data included in each window. Finally, the performance of the models is tested on the first point to the right of the training window. The testing process consists in calculating the cumulative probability that a model assigns to the first point out of the window and then applying it the inverse of the standard Gaussian distribution (the Berkowitz transformation). This way, we obtain 8190 *test measurements* for each model that should follow a standard Gaussian distribution under the null hypothesis that the model is correct. The application of the statistical tests described in Section 2.3.3 to those *test measurements* allows us to validate the models for estimating risk one day ahead in the future.

The estimation techniques are implemented in R [R Development Core Team, 2005].

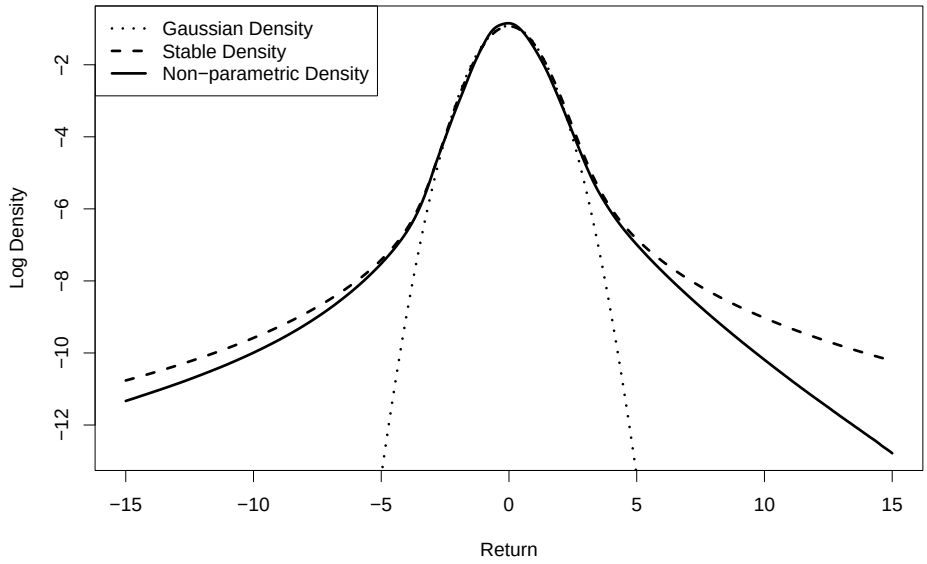

Figure 4.5: Plots in logarithmic scale of the different model density functions. The non-parametric density and the stable density have been obtained as averages of the densities estimated in the sliding window experiment. To compute these averages only the densities for one in every fifty windows are used (a total of 163 densities). The stable densities are scaled so that their standard deviation is 1.

The non-linear optimization method used to maximize the likelihood was the *downhill simplex method* included in the function *constrOptim*. To avoid that the optimization process gets trapped in a local maximum, the algorithm is rerun using different initial random values, and the best solution found is selected. Stable distributions have no general known closed form expressions for their density [Nolan, 2002]. To evaluate the stable density we employ the technique described in [Mittnik et al., 1997]. The density is computed exactly on a grid of 2^{13} equally spaced points. Outside this grid the density function is calculated by linear interpolation from the values in the grid. To evaluate the cumulative stable distribution we use the R package *fBasics* [Wuertz et al., 2004] that implements the method described by Nolan [Nolan, 2002]. Finally, the parameter a of stable densities is restricted to have values greater than 1.5, a bound which should be sufficiently low to account for the heavy tails in the distribution of returns.

Table 4.1 displays, for each model, the results of the statistical tests performed over the *test measurements* obtained as described in Section 4.5. The level of significance considered to reject the null hypothesis on each test is 5%. The GARCH model with stable innovations fails all the tests but the one for Exceedances at the 95% level. The *p-values* obtained by this model for the tests Exc 99% and ES 95% are rather low. In particular, the result for ES 95% reflects the failure of the model to accurately describe the loss tail of the conditional distribution of returns. The results of the GARCH model with non-parametric innovations are remarkable. The model obtains very high *p-values* in all tests except for one. In particular, the test for Expected Shortfall at the 99% level fails. This is the most demanding test and it requires the model to correctly account for extreme events. In our case the test fails because of one single extreme event, a loss of around 5% that took place on January 19, 1970 (see Fig. 4.6). Immediately prior to that observation, the value estimated

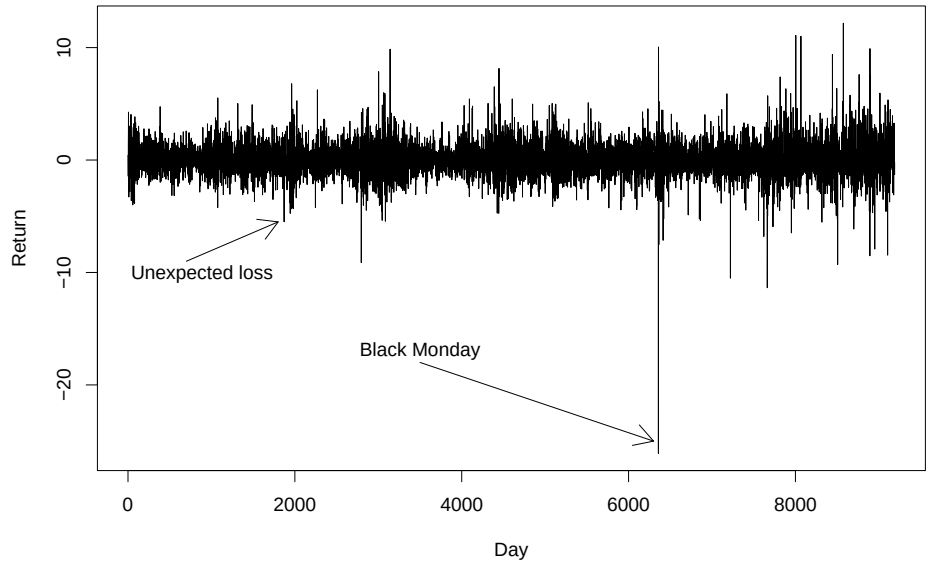

Figure 4.6: Returns of IBM stocks from 1962/07/03 to 1998/12/31. The loss on January 19, 1970 has been singled out. The model proposed has difficulties to account for such a loss. The loss on October 19, 1987 known as Black Monday, has also been marked. This loss is correctly modeled.

Table 4.1: *p-values* obtained from the statistical tests performed. From left to right tests for Value at Risk, Exceedances and Expected Shortfall at levels 99% and 95%. See Section 2.3.3 for a description of the tests.

Model	VaR 99%	VaR 95%	Exc 99%	Exc 95%	ES 99%	ES 95%
stable	1.5e-3	0.01	3.7e-5	0.08	0.026	4e-4
non-parametric	0.51	0.66	0.49	0.86	0.042	0.31
mix. of 2 exp.	5.7e-3	0.30	0.025	0.19	7e-7	7.2e-4

for a in the stable distribution that characterizes the transformation $g_{\boldsymbol{\lambda}}$ is fairly high 1.97, indicating that the tails are not very heavy and that the local fluctuations of the returns were expected to be small in size (smaller than the loss observed) with a high degree of confidence. If this point is removed, the *p-value* obtained in the test ES 99% is 0.31 and all the other tests give results above the level of significance 0.5. It is also interesting to see that the model can properly account for the 25% loss that took place on October 19, 1987 (Black Monday, see Fig. 4.6). The difference with the previous instance is that the value estimated for the parameter a right before Black Monday is fairly low (1.88), which signals that the distribution tails are very heavy and that large fluctuations have a non negligible chance of being observed. This low value of a originates in a sequence of large, but not extreme, fluctuations in the period immediately before Black Monday.

Figure 4.5 displays the average stable density and average non-parametric density obtained in the sliding window experiment. The tails of the stable and non-parametric densities are heavier than those of the standard Gaussian density. Moreover, those two distri-

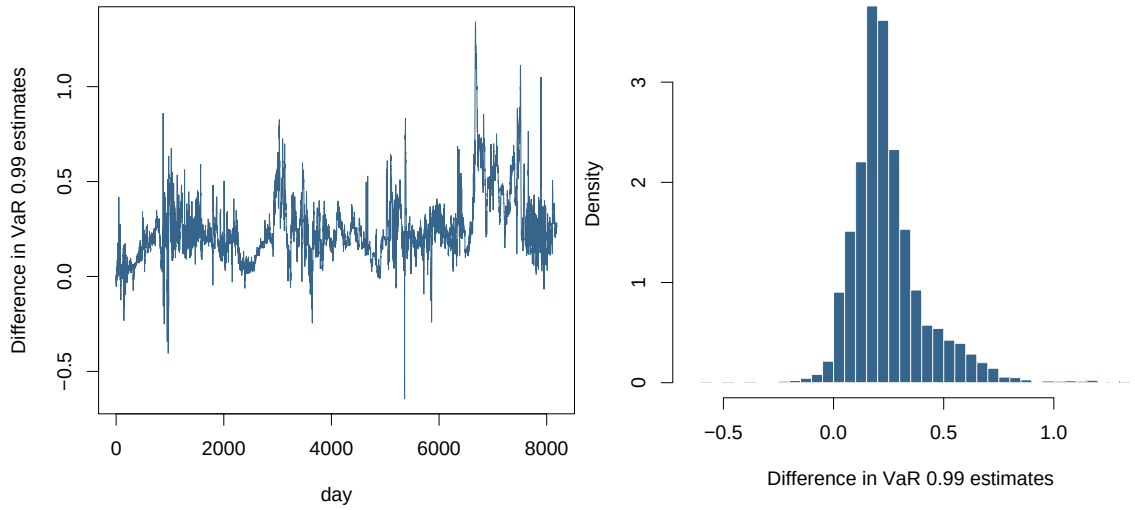

Figure 4.7: Left, time series with the difference between the estimates of the VaR at the 0.99 level from the non-parametric GARCH process and the stable GARCH process. It can be noticed that the series is most of the time positive, what indicates that the stable GARCH process provides estimates of risk which are generally lower than those of the non-parametric one. The units used to measure the VaR are percentage log-returns. Right, histogram of the series which appears on the left. The difference in VaR estimates are significantly positive.

butions are not symmetric, both densities assign a higher probability to returns close to 5 than to returns close to -5. However, the right tail of the non-parametric density displays a fast decay for returns greater than 5. This decay does not appear in the right tail of the stable density. This figure also shows how the tails of the non-parametric density are on average slightly less heavy than those of the stable distribution, specially in the right tail, corresponding to gains. This agrees with previous studies based on extreme value theory, which suggest that the tails of stable densities are probably too heavy to model asset returns [Longin, 1996, Jansen and de Vries, 1991].

Strikingly, even though the tails of the stable densities are on average heavier than those of the non-parametric densities, the GARCH process with stable innovations provides estimates of one-day-ahead Value at Risk at the 0.99 level which are consistently lower than those of the non-parametric GARCH process. This fact can be appreciated on Figure 4.7. Because the non-parametric model obtains much better results in the test for Value at Risk at the 0.99 level, we can conclude that the stable GARCH process infraestimates risk.

We also performed the sliding-window experiment with mixtures of 2 autoregressive experts. The experiment was designed in the same way as the one described in Chapter 3, the only difference was that we employed the returns of IBM stocks instead of those from the stock index IBEX 35. The last row of Table 4.1 shows the test results obtained by mixtures of 2 autoregressive experts. It can be noticed how the mixture model fails all the tests at the 0.99 level, specially the one for Expected Shortfall. The reason why mixtures of 2 experts have now obtained much worse results than in Chapter 3 is due to the fact that returns from individual stocks depart from normality more than those from stock indexes. This is so because the price of stock indexes is computed as a weighted average of the price

of individual stocks and the central limit theorem smooths extreme events away. Finally, we believe that mixtures of 3 autoregressive experts would obtain similar results.

4.6 Summary

A GARCH process with non-parametric innovations is proposed as a model for the series of daily stock returns. We have also described a procedure to estimate the parameters of the model and the density of the innovations by maximum likelihood in a transformed space. The density for the innovations is modeled using a non-parametric estimate based on kernels whose width is determined using Silverman's *rule of thumb*. In the transformed space the non-parametric estimate provides a more accurate model of heavy-tailed densities. A sliding window experiment using the returns of IBM stocks was carried out to evaluate the performance of the proposed model. The experiments indicate that GARCH processes with non-parametric innovations provide very accurate and sensitive estimates of risk measures, such as Value at Risk and Expected Shortfall. Furthermore, the new model outperforms soft competitive mixtures with 2 experts.

Chapter 5

Conclusions and Future Work

This chapter summarizes some of the main conclusions reached throughout the whole document. Furthermore, it provides some ideas that could be followed in order to extend the current research.

5.1 Conclusions

In the second chapter we described that any financial institution which holds an investment portfolio is exposed to market risk, a kind of risk which has recently increased because of globalization in financial markets. In order to be competitive and to protect from unexpected high losses, financial institutions must implement and validate risk measuring and managing systems. The process of market risk measuring requires us to choose a risk measure (if possible, a coherent one) and to know the distribution of future prices for a portfolio, a distribution which is generally unknown. The usual approach to tackle the latter problem consists in making use of a probabilistic model (e.g. GARCH process) for the price changes. The model is fitted with historical past data and then used to infer the distribution of future price changes. However, obtaining an accurate model is usually a great challenge because of the complex properties displayed by financial time series (heavy tails and heteroskedasticity). Once a probabilistic model is available, it must be checked that the risk measurements that it provides are accurate enough. This is performed by means of a backtesting procedure.

Throughout this document we have analyzed the performance of different models for estimating market risk and the accuracy of each of them was measured by means of advanced statistical tests based on the functional delta method. In the third chapter, mixtures of up to three autoregressive experts that work together using different strategies were compared. It turned out that mixtures which employ a soft competitive strategy outperformed the other models. This was due to the fact that soft competitive mixtures predict a future distribution for price changes which is a mixture of Gaussians (one Gaussian for each expert), a paradigm which can effectively account for the heavy tails of financial time series (a mixture of an unlimited number of Gaussians can approximate, up to any degree of precision, any density function). However, one drawback of soft competitive mixtures is that the number of experts (Gaussians) is limited due to overfitting and training cost. Finally, in the fourth chapter we improved GARCH processes by means of modelling their innovations in a

non-parametric way. The innovations were described by means of kernel density estimates which operate in a transformed space to better account for the leptokurtosis of financial time series. The mentioned kernel estimates can also be regarded as constrained mixtures of Gaussians. However, the difference with respect to soft competitive mixtures is that, in this case, it is feasible to make use of thousands of Gaussians without causing overfitting. The experiments demonstrate the superiority of GARCH processes with non-parametric innovations for performing market risk estimation.

5.2 Future Work

The research performed in this document could be extended in several ways. First, in all the experiments performed we have used daily returns for training and validating the models. With the computerization of financial markets it is easier to obtain intra-day returns and even tick-by-tick transaction data. The increase on available information allows to compute better estimates of volatility and to suggest new models [Andersen et al., 2003, Engle, 2000]. Second, in this document we have dealt mainly with models for the changes in price of single financial assets. However, financial institutions hold portfolios with many different assets in order to take advantage of the risk reduction caused by diversification. In such a situation it is necessary to model the dependence between individual assets if we want to obtain accurate measurements of risk for the whole portfolio. This can be achieved by means of correlations if we work with elliptical distributions, or copulas [Nelsen, 2006] if we deal with general non-elliptical distributions. Third, we have focused on measuring risk one day ahead on the future. Nevertheless, it is possible to study the ability of the described models to estimate risk with horizons which range from weeks to months. Monte Carlo methods can be used to compute the predicting distribution determined by the models when analytical solutions are not available. It would also be profitable to carry out experiments with returns from other financial assets. Fourth, it would be interesting to compare the performance of GARCH processes with non-parametric innovations against GARCH processes with normal inverse Gaussian or generalized hyperbolic innovations. Also, those parametric distributions could be used to carry out the transformation needed in the non-parametric density estimation step of the training process for the models described in Chapter 4. Finally, Expected Shortfall is not the best coherent risk measure [Dowd, 2005]. This is so because it is risk-*neutral* between tail-region outcomes and we usually assume that investors are risk-*averse*. The class of spectral risk measures do not have this problem [Acerbi, 2002].

Bibliography

[Bas, 1996] (1996). *Supervisory Framework for the Use of "Backtesting" in Conjunction with the Internal Models Approach to Market Risk Capital Requirements*. Bank for International Settlements.

[Acerbi, 2002] Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26:1505–1518.

[Andersen et al., 2003] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

[Anderson and Darling, 1954] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

[Artzner et al., 1999] Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

[Barndorff-Nielsen, 1997] Barndorff-Nielsen, O. E. (1997). Normal inverse gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24(1):1–13.

[Bengio et al., 2001] Bengio, Y., Lauzon, V.-P., and Ducharme, R. (2001). Experiments on the application of iohmms to model financial returns series. *IEEE Transactions on Networks Networks*, 12(1):113–123.

[Berkowitz, 2001] Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *The Journal of Business and Economic Statistics*, 19(4):465–474.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

[Bishop and Svensén, 2003] Bishop, C. and Svensén, M. (2003). Bayesian hierarchical mixtures of experts. *Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 57–64.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Bollerslev, 1986] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

[Bollerslev, 1987] Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3):542–547.

[Campbell et al., 1997] Campbell, J. Y., Lo, A. W., and Mackinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

[Cont, 2001] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Journal of Quantitative Finance*, 1:223–236.

[Davis and Mikosch, 1998] Davis, R. A. and Mikosch, T. (1998). The sample autocorrelations of heavy-tailed processes with applications to arch. *The Annals of Statistics*, 26(5):2049–2080.

[Devroye and Györfi, 1985] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 View*. John Wiley.

[Devroye et al., 1997] Devroye, L., Gyrfi, L., and Lugosi, G. (1997). *A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability)*. Springer.

[Ding et al., 1993] Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106.

[Dowd, 2005] Dowd, K. (2005). *Measuring Market Risk*. John Wiley & Sons.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience Publication.

[Eberlein and Keller, 1995] Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli*, 1(3):281–299.

[Engle, 2000] Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica*, 68(1):1–22.

[Fama, 1970] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2).

[Forsberg, 2002] Forsberg, L. (2002). On the normal inverse gaussian distribution in modelling volatility in the financial markets. *PhD Dissertation, Uppsala University, Sweden*.

[Forsberg and Bollerslev, 2002] Forsberg, L. and Bollerslev, T. (2002). Bridging the gap between the distribution of realized (ecu) volatility and arch modelling (of the euro): the garch-nig model. *Journal of Applied Econometrics*, 17:535–548.

[Franke and Diagne, 2006] Franke, J. and Diagne, M. (2006). Estimating market risk with neural networks. *Statistics & Decisions*, 24:233–253.

[Hamilton, 1991] Hamilton, J. D. (1991). A quasi-bayesian approach to estimating parameters for mixtures of normal distributions. *Journal of Business & Economic Statistics*, 9(1):27–39.

[Hamilton, 1994] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

[Hernández-Lobato et al., 2007] Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. (2007). GARCH processes with non-parametric innovations for measuring market risk. In de Sá, J. M., Alexandre, L., Duch, W., and Mandic, D., editors, *ICANN (2)*, volume 4669 of *Lecture Notes in Computer Science*, pages 718–727. Springer.

[Hernández-Lobato and Suárez, 2006] Hernández-Lobato, J. M. and Suárez, A. (2006). Competitive and collaborative mixtures of experts for financial risk analysis. In Kollias, S. D., Stafylopatis, A., Duch, W., and Oja, E., editors, *ICANN (2)*, volume 4132 of *Lecture Notes in Computer Science*, pages 691–700. Springer.

[Holton, 2003] Holton, G. A. (2003). *Value-at-Risk*. Elsevier.

[Holton, 2004] Holton, G. A. (2004). Defining risk. *Financial Analysts Journal*, 60(6).

[Jacobs et al., 1993] Jacobs, R. A., Jordan, M. I., and Barto, A. G. (1993). Task decompostiion through competition in a modular connectionist architecture: The what and where vision tasks. In *Machine Learning: From Theory to Applications*, pages 175–202.

[Jacobs et al., 1991] Jacobs, R. A., Jordan, M. I., Nowlan, S., , and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:1–12.

[Jansen and de Vries, 1991] Jansen, D. W. and de Vries, C. G. (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *The Review of Economics and Statistics*, 73(1):18–24.

[Jarque and Bera, 1987] Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172.

[Jordan and Jacobs, 1994] Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.

[Jorion, 1997] Jorion, P. (1997). *Value at Risk: The New Benchmark for Controlling Market Risk*. McGraw-Hill.

[Kerkhof and Melenberg, 2004] Kerkhof, J. and Melenberg, B. (2004). Backtesting for risk-based regulatory campital. *Journal of Banking and Finance*, 28(8):1845–1865.

[Kon, 1984] Kon, S. J. (1984). Models of stock returns–a comparison. *Journal of Finance*, 39(1):147–165.

[Kupiec, 1995] Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84.

[Longin, 1996] Longin, F. M. (1996). The asymptotic distribution of extreme stock market returns. *The Journal of Business*, 69(3):383–408.

[MacKay, 2003] MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

[Markowitz, 1991] Markowitz, H. (1991). *Portfolio Selection*. Blackwell Publishing.

[Mathworks, 2002] Mathworks (2002). *Matlab Optimization toolbox 2.2*. Mathworks, Inc., Natick, MA.

[Mikosch and Starica, 2000] Mikosch, T. and Starica, C. (2000). Limit theory for the sample autocorrelations and extremes of a garch(1,1) process. *The Annals of Statistics*, 28(5):1427–1451.

[Mittnik et al., 1997] Mittnik, S., Doganoglu, T., and Chenyao, D. (1997). Computing the probability density function of the stable paretian distribution. *Mathematical and Computer Modelling*, 29(10):235–240.

[Mittnik and Paolella, 2003] Mittnik, S. and Paolella, M. S. (2003). Prediction of financial downside-risk with heavy-tailed conditional distributions. *CFS Working Paper*, (2003/4).

[Morgan, 1996] Morgan, J. P. (1996). *RiskMetrics Technical Document*. Fourth Edition, New York.

[Nadaraya, 1989] Nadaraya, E. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic Publishers.

[Nelsen, 2006] Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer.

[Nolan, 2002] Nolan, J. P. (2002). *Stable Distributions*. Birkhauser.

[Panorska et al., 1995] Panorska, A. K., Mittnik, S., and Rachev, S. T. (1995). Stable garch models for financial time series. *Applied Mathematics Letters*, 8(5):33–37.

[Papoulis and Pillai, 2002] Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill.

[Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

[Prause, 1999] Prause, K. (1999). The generalized hyperbolic model: Estimation, financial derivatives and risk measures. *PhD Dissertation, University of Freiburg*.

[R Development Core Team, 2005] R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Shapiro and Wilk, 1965] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[Sheather and Jones, 1991] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.

[Silverman, 1986] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

[Sociedad de Bolsas, 2006] Sociedad de Bolsas (2006). *Histórico Cierres Índices Ibex*. http://www.sbolsas.es.

[Suarez, 2002] Suarez, A. (2002). volume 2415/2002 of *Lecture Notes in Computer Science*, chapter Mixtures of Autoregressive Models for Financial Risk Analysis. Springer Berlin / Heidelberg.

[Vaart, 2000] Vaart, V. D. (2000). *Asymptotic Statistics*. Cambridge University Press.

[Vidal and Suarez, 2003] Vidal, C. and Suarez, A. (2003). volume 2714/2003 of *Lecture Notes in Computer Science*, chapter Hierarchical Mixtures of Autoregressive Models for Time-Series Modeling. Springer Berlin / Heidelberg.

[Wand et al., 1991] Wand, M. P., Marron, J. S., and Ruppert, D. (1991). Transformations in density estimation. with discussion and a rejoinder by the authors. *Journal of the American Statistical Association*, 86(414):343–361.

[Weigend and Huberman, 1990] Weigend, A. S. and Huberman, B. A. (1990). Predicting sunspots and exchange rates with connectionist networks. Proc. of 1990 NATO Workshop on Nonlinear Modeling and Forecasting, Santa Fe, NM, 1990.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

[Wuertz et al., 2004] Wuertz, Diethelm, and Others, M. (2004). *fBasics: Financial Software Collection - fBasics*.