UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

DOTTORATO DI RICERCA IN
INGEGNERIA ED ECONOMIA DELL'INNOVAZIONE
XII CICLO

TESI DI DOTTORATO IN:

Evaluation of Volatility Forecasts

COORDINATORE:
Ch.ma Prof.ssa Alessandra Amendola

TUTOR:
Ch.ma Prof.ssa Alessandra Amendola

CANDIDATO:
Vincenzo Candila

ANNO ACCADEMICO 2012/2013

Contents

Keywords

- Volatility forecasts
- Volatility forecast evaluation
- Realized volatility
- Multivariate GARCH models
- Statistical loss function
- Economic loss function
- Consistent loss function
- Value at Risk
- Consistent Value at Risk estimator

Acknowledgments

This thesis represents the conclusion of my Ph.D., an incredibly wonderful experience, albeit very tough. It would have been impossible for me to complete these overwhelming three years without the support and encouragement of many people, who I wish to sincerely thank.

First of all, I would like to thank my supervisor Prof. A. Amendola. Without her guidance, this work would not have been even started. She has always been patient, encouraging during the most stressful and difficult times. And her advices have always proved to be so precious.

Then, I gratefully acknowledge Prof. C. Perna, Director of the Department of Economics and Statistics of the University of Salerno, for providing the necessary infrastructures and resources to accomplish my research work.

I would also like to say thank to Prof. G. Storti. His reviews along with our discussions have upgraded the quality of this work. I am also extremely thankful to Prof. P. Coretto. His lectures on *Limit Theorems*, his suggestions about the LaTeX formatting and the R environment have been widely appreciated. I am very grateful to Prof. M. Restaino for being a friend and for having always provided me with solutions to all the issues I have dealt with. I gratefully acknowledge Dr. M. Rizzo. I still remember how she helped me with reference to my first submission till the late evening. Even though we had only a few chances to speak about my work, Prof. G. Giordano has been a strong source of inspiration.

Last but not least, I am deeply grateful to Prof. C. Vitale. His lectures on the *Time Series Analysis* and *Stochastic Processes and Statistical Inference* have allowed me to grow interest in this matter.

In conclusion, I would like to thank E. Lazzazzera for her love, encouragement and patience. I also express a heartfelt acknowledgement to all my family: my late grandmother, my mother and father, my aunt Elena, my sister Simona and my brother in law Giacomo. They all have been by my side throughout these meaningful three years.

Introduction

The problem of the evaluation of volatility predictions is a challenging research field. A general formulation of this problem is:

$$VE = (h, \hat{h}_m, FI, EF(\cdot)), \tag{1}$$

where VE stands for volatility evaluation and:

- h is the (unobserved) volatility;
- $\hat{h}_m$ represents the volatility predictions of m competing models;
- FI stands for the Forecasting Issue;
- $EF(\cdot)$ is the evaluation function comparing each $\hat{h}_m$ to (a proxy of) h.

In this thesis the term volatility refers to the risk related to hold financial instruments like assets, bonds, and so forth. Unfortunately, the volatility is not directly observable. Despite the latent nature of the volatility, its estimate is fundamental in many empirical problems such as the asset allocation and risk management, for instance. During the last decades

many volatility measures have been proposed. A comprehensive survey on the volatility measures can be found in Andersen et al. (2002). Even though h is not directly observable, different proxies have been proposed in literature. A widely accepted volatility proxy is represented by the realized volatility. See Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2004) for the theoretical foundations of this approach, while a review is provided by McAleer and Medeiros (2008). If there are more than one asset, the term realized volatility is replaced by the term realized covariance.

With reference to $\hat{h}_m$ in (1), the estimate of volatility can be obtained through different models/approaches. In this work, the attention is focused on the GARCH models, both in the univariate and multivariate context, and on the models that derive the volatility as combination of the past proxies. Terasvirta (2006) provides an overview of the univariate GARCH models while Bauwens et al. (2006) survey the most important developments in multivariate ARCH-type modelling. Other methods can be used to estimate the volatility, such as the implied volatility method, the stochastic volatility approach, and so forth. A discussion on the implied volatility is in Dumas et al. (1998). A survey on the stochastic volatility is in the work of Ghysels et al. (1995) while Lehar et al. (2002) empirically test the differences between the GARCH and the stochastic volatility approach.

The discussion about the evaluation of the volatility predictions cannot disregard by the forecasting issue, denoted in (1) by FI. The FI concerns different aspects such as the width of the information set used to predict the volatility, how often the information set is updated, the combination of the forecasts, and so forth. A survey on the forecasting is found in Elliott and Timmermann (2007) while Timmermann (2006) presents an overview of the forecasts combination.

Last but not least, $EF(\cdot)$ in (1) represents the function/method used to evaluate the volatility predictions of a set of competing models to the

volatility proxy. These methods are generally divided in two approaches:

1. a statistical approach involving loss functions such as the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), etc.
2. an economic approach, based on utility functions or indirect methods, such as the Value at Risk (VaR) measures.

In a general forecasting framework, where the volatility predictions are only a possible application, the predictive ability of the models can be tested by the Diebold and Mariano (1995) and West (1996) tests, for the pairwise comparison and White (2000) and Giacomini and White (2006) tests, for the multiple comparison, for instance. In all these tests, the loss functions play a fundamental role, evaluating the distance between the actual (when observed) and the forecasted value of the variable of interest. When the volatility is the variable of interest, the situation becomes more difficult, due to its latent nature. Recently, an important innovation regarding the evaluation of volatility by means of the loss functions has been provided by Hansen and Lunde (2006) and Patton (2006) for the univariate framework and Laurent et al. (2013) for the multivariate one. The innovation regards the consistency or robustness of the loss function: a loss function is said to be consistent if the ranking of any two volatility forecasts is the same of the ranking that would be obtained if the true volatility had been observable. The tests cited above and the new contribution regarding the consistency of the loss functions belong to the statistical approach. Comparing the forecasts of a set of models to the volatility proxy, the statistical approach represents a direct method of evaluation.

Instead, the economic approach is said to be an indirect method of evaluation. In fact, it takes in consideration risk measures, whose inputs are the volatility predictions obtained by a set of models. Afterwards, these risk measures are evaluated. A popular risk measure is the Value at Risk (VaR). Duffie and Pan (1997) and Jorion (2007) present its overview.

If the mean-variance approach is used, the volatility forecast of a model is the key input to estimate the VaR. Then the performance of the model is evaluated looking at its VaR measures. Traditionally, the evaluation of the VaR measures is done examining the occurrences of VaR violation, where violation means the event that the observed negative daily return (representing a loss) is greater than the VaR. More specifically, the VaR evaluations in terms of violations is done through the Unconditional and Conditional Coverage tests, proposed by Kupiec (1995) and Christoffersen (1998), respectively. The former tests if the actual number of violations is statistically equal to the expected number of violations. The Conditional Coverage test verifies jointly the empirical rate of violations and their independence.

Research Questions

This work aims to investigate the performance of a set of competing models in terms of volatility forecast accuracy. The volatility predictions are compared cross-sectionally among the models and with respect to the volatility proxy, by means of statistical and economic approaches. In other words, the main interest is investigating the $EF(\cdot)$ in (1).

From the univariate point of view, the two aforementioned approaches are merged. The evaluation of the VaR measures is done through two loss functions, the first proposed by Lopez (1998) and the second is a new asymmetric loss function. The term asymmetric means that a model with an actual number of violations greater than the expected one is more penalized. The research questions are:

1. Is it possible to use the loss function in a VaR framework in order to evaluate the volatility predictions of a set of competing models? Does this approach bring an advantage when the statistical and the economic approaches fail to recognize the best model?

2. Is it possible to find a threshold that discriminates low from high

loss function values in order to evaluate the performances of the volatility models?

From the multivariate point of view, again the two approaches are considered but this time are explicitly compared. The research questions are:

3 Is the ranking of the competing models the same if a statistical and an economic loss functions are used?

4 Do the multivariate GARCH models have a worse forecast accuracy than that of the models that use the realized volatility to forecast $\hat{h}$?

Thesis Structure

The work is organized as follows. Chapter 1 concerns the first two sub-points of (1). First, the univariate and multivariate GARCH models are surveyed and then the nonparametric approach used to have the volatility proxy is discussed, both in the univariate and multivariate context. Chapter 2 is related to the FI in (1). The questions dealt with are (i) the forecasting schemes, that can be fixed, recursive or rolling; (ii) the evaluation of the forecast performances using the robust statistical loss function, the economic approach and the predictive ability tests; (iii) the forecasts combination in order to have a superior predictive ability. The answers to the research questions 1-2 are given in Chapter 3. Here the new asymmetric loss function is proposed and the block bootstrap of the high-frequency intraday return is used in order to have a consistent estimator of the distribution of the daily returns. This approach allows to estimate the VaR for any distribution of the daily returns. Moreover, the block bootstrap is re-used to estimate the threshold that discriminates low loss function values to high loss function values, once the VaR has been obtained. In this chapter, the analysis is done through a Monte Carlo experiment and an empirical analysis of a stock listed on the New York

Stock Exchange. Chapter 4 presents the answers to the research questions 4-5. Here an extensive Monte Carlo simulation is performed to compare the volatility predictions of fourteen models. The benchmark of the volatility is the realized covariance and the analysis is repeated as the proxy becomes more and more imperfect. In other words, the frequency at which the proxy is obtained varies from five minutes to one day. Finally, the conclusions with some suggestions for future research and the main references are reported.

Key Contributions

With reference to the volatility evaluation in the univariate framework, discussed in Chapter 3, this thesis uses the loss functions in a VaR framework. This approach is said to be mixed because it considers both the loss functions (statistical approach) and the VaR measures (economic approach). Moreover, the loss functions in a VaR framework works independently of the distribution of the daily returns, provided that their high-frequency increments are available. This approach provides an empirical, straightforward method to determine the threshold discriminating low from high loss function values, helping the forecaster to decide which model has a superior predictive ability. Moreover, a new asymmetric loss function is proposed that could be taking in consideration by the risk-averse agents. The indirect evaluation of the volatility models by means of the loss functions yields advantages in situations where the statistical approach, like the MSE, or the VaR tests fail to recognize the best model.

With reference to the volatility evaluation in the multivariate framework, presented in Chapter 4, the empirical analysis realized with different levels of proxy's precision shows that there is not a clear correspondence between the ranking produced by the robust loss function and that produced by the economic approach. For a risk manager, decisions only based on the VaR violations could lead to use models that are far from the

true (but unobserved) volatility. Moreover, models using a combination of the past realized covariances assure better forecast accuracy than that obtained from the multivariate GARCH models. Finally, the importance of the covariances in the evaluation of the forecast accuracy seems to be low. The ranking given by the loss function that uses only the variances is the same of the ranking given by the loss function that uses both variances and covariances.

1 The Volatility Models

1.1 Introduction

The volatility is a latent variable and it is not directly or uniquely calculable. With reference to the assets listed on exchange markets, the volatility is related to the variability: the greater is the variability of a stock, the greater is the "risk" of possible losses or possible gains. Thus, to measure the risk related to holding financial instruments it is usual to refer to some statistical variability indices, like the variance or the standard deviation. These statistical indices are used to quantify the risk of the considered instrument over that time period. Already in the 1960s many studies showed that financial time series exhibit time-varying variance (see the seminal paper of Mandelbrot (1963)). The quantification of the volatility is fundamental in order to price derivatives, hedge against portfolio risk, make decision about which instruments is suitable to buy, which to hold and which to sell, for instance. This explains why practitioners and researchers have such enormous interest on it.

Different volatility measurements have been proposed in literature. A classic distinction between these measures is the parametric/nonparametric nature of the procedure used to estimate the volatility. Andersen

et al. (2002) provide an excellent report on this issue. A specific class of statistical models that captures the "heteroskedasticity", that is the time-varying variance, is the GARCH (General AutoRegressive Conditional Heteroskedasticity) class of models, proposed by Engle (1982) and Bollerslev (1986). The GARCH models belong to the family of parametric approaches where the underlying data generating process is assumed to be known. In particular, the GARCH models explicitly consider a functional form such that the ex-ante expected volatility depends on directly observable variables. Also the Stochastic Volatility framework, reviewed by Taylor (1994), belongs to the family of parametric procedures. It aims to estimate the ex-ante expected volatility by means of directly observable variables but, in addition to the GARCH models, the Stochastic Volatility method also considers some additional, latent state variables. Instead, the nonparametric approach lacks of any functional form and estimates the ex-post volatility by means of the high-frequency data. For instance, the well known realized volatility (Andersen et al. (2003)) estimates the ex-post volatility summing the squared return realizations over a fixed time interval.

The aim of this chapter is to present the main GARCH models, from an univariate and multivariate point of view. These models represent different estimators of $\hat{h}_m$ in (1). In the last part of the chapter, first nonparametric approach for the volatility proxy and then its parametric modelization are illustrated. During last years, an very large amount of models have been presented. We focus only on the models that will be used in the empirical part of this work.

The remainder of this chapter is organized as follows. In Section 1.2 the notations that will be used throughout this work are showed. In Section 1.3 some empirical regularities of the asset returns are presented. Section 1.4 and 1.5 survey the univariate and multivariate GARCH models, respectively. Section 1.6 overviews the inference procedures to estimate the models presented and Section 1.7 reports the volatility proxy

framework

1.2 The notation

Throughout this work we use the following notations. Let r_t be the daily logarithmic returns at time t of a given stock:

$$r_t = p_t - p_{t-1},$$

where p_t is the logarithm of the observed price at time t. Let I^{t-1} be the Information Set at time $t-1$, used to estimate any moment of r_t, conditionally on it. I^{t-1} consists of all the information generally available concerning that stock. The conditional mean of r_t is:

$$\mathbb{E}(r_t|I^{t-1}) = \mu_t, \tag{1.1}$$

We focus on the second moment of r_t, that is commonly defined as conditional variance:

$$VAR(r_t|I^{t-1}) = h_t^2. \tag{1.2}$$

Within this framework, the volatility is generally defined as the square root of h_t^2.

With reference to the multivariate context, the bold character means a column vector. More specifically, let $\mathbf{r}_t$ a column vector of dimension $k \times 1$ such that each entry is given by the daily logarithmic difference of k assets. The transposition of a vector is denoted by the operator $'$, such that $\mathbf{r}_t'$ indicates a $1 \times k$ vector. All the matrices are indicated by a capital letter.

Let L be the lag or backshift operator such that $L^i r_t = r_{t-i}$ and $\alpha_p(L) = \alpha_1 L + \cdots + \alpha_p L^p$.

The expressions $\xrightarrow{a.s.}$, $\xrightarrow{p}$ and $\xrightarrow{d}$ denote the almost sure convergence, the convergence in probability and the convergence in distribution, respectively. The expression $\overset{d}{\approx}$ means that, when multiplied by a suitable factor, the convergence is in distribution.

1.3 Stylized Facts

In the asset returns there are some common features, defined as empirical regularities or "Stylized Facts". The volatility models aim to capture as many as possible of these stylized facts. In what follows, the main empirical regularities of the asset returns are presented. Some stylized facts will be displayed in figures illustrating the daily logarithmic returns of the Capital One Financial Corporation (COF) stock. The COF stock is listed on the New York Stock Exchange. Capital One Financial Corp. is a U.S.-based bank holding company specialized in credit cards, home loans, banking and savings products. The dataset, obtained from Tick Data, Inc.[1], covers the period from April 8, 1997 until July 18, 2008 (2839 trading days). It is characterized by severe changes in volatility dynamics. As done by Laurent et al. (2010), that worked more or less with the same time lapse, we divide the whole sample into three not overlapping sub-periods. In the first sub-sample (from April 3, 1998 to December 31, 2003, for a total of 1695 days), the data experience a considerable turbulence.

The second sub-sample consists of 607 trading days, from January 2, 2004 to May 31, 2006. In this sample the data experience a period of market stability.

The third sub-sample (537 days) starts on June 1, 2006 and ends on July 18, 2007. The last part of this sub-sample corresponds to the beginning of the recent financial crisis.

1.3.1 Volatility Clustering

The first stylized fact is the volatility clustering, discovered by Mandelbrot (Mandelbrot (1963)), that on 1963 wrote "$\cdots$ large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes$\cdots$".

[1] Tick Data is a provider of historical intraday market data.

In (1.1), if μ_t is assumed to be zero $\forall t$, the squared returns can be used as (a very noisy) estimator of the conditional variance. Looking at their correlogram there is a clear evidence of time-varying variance. Hence, the conditional variance at time t depends on the conditional variance at time $t-1$ and so forth. For an economic point of view, the explanation of volatility clustering is as follows. If a stock price falls down, it is expected that the investor sells. If many investors begin to sell, the stock value goes down more and more. Thus, changes of negative sign tend to be followed by other negative changes. In such periods the volatility is very high. Many papers highlight the evidence of volatility clustering, like Friedmann and Sanddorf-Köhle (2002) and Kirchler and Huber (2007).

The volatility clustering phenomenon is immediately observable when asset returns are plotted through time, as shown in Figure 1.1. In panel (A), corresponding to the first period, the scale of the daily returns is the greatest compared to the other two periods. Here, some considerable days of volatility clustering are visible in mid 2003. Without doubts, some evident episodes of volatility clustering are present in the third period (panel (C)). In this period, the first months do not experience great movements while the last months are characterized by huge changes of the daily returns.

The volatility clustering can be also visualized by means of the autocorrelations of the squared daily returns. Recalling that it is assumed $\mu_t = 0$. The autocorrelations of the squared daily returns show how much the variance is clustered in time. These autocorrelations are displayed in Figure 1.2. In panel (A) and (C) the autocorrelations till at least lag nine strengthen the idea of volatility clustering presence. In addition, panel (B) witnesses that in the second period there are not relevant volatility clustering episodes.

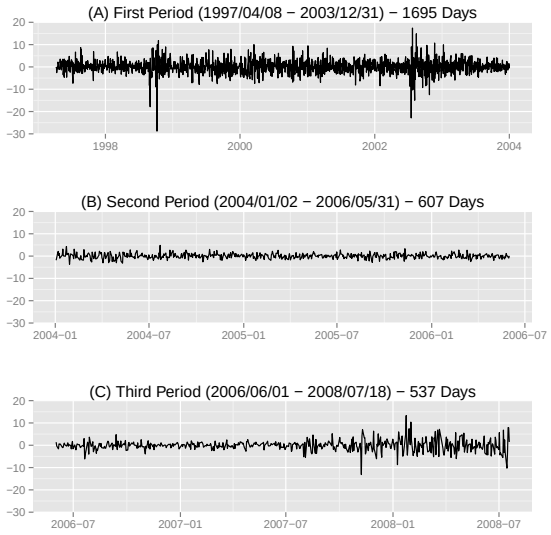


Figure 1.1 – COF daily returns

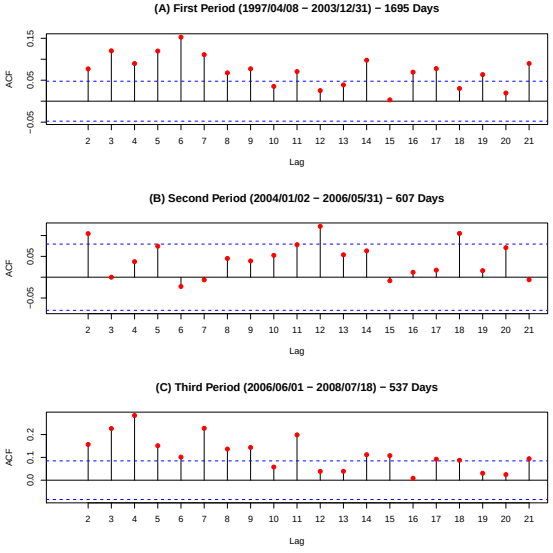


Figure 1.2 – COF squared returns autocorrelation

1.3.2 Thick Tails

It is generally supposed that the asset returns follow a Normal distribution. Many times, this assumption is rejected by empirical regularities. For instance, Rachev et al. (2005) examine 382 U.S. stocks, strongly rejecting this assumption. In fact, asset returns tend to be leptokurtic. This empirical regularity is called Fat or Thick tails. Assuming a Normal distribution, it is known that about 95% of the values are within a distance of two standard deviations from the mean. Instead, the unconditional distribution of the assets usually displays a different behaviour: in particular, there are (i) rare events with a small frequency (very far from the mean) but with a exceptionally low or high values (for example, a very bad performance caused by a failure's news spread or a very good performance caused by a merger between two companies appreciated by the market) (ii) many events with a high concentration around the mean, usually equal to zero. Despite all these drawbacks, the Normal distribution assumption is taken in order to make easier the inference procedure. However, in literature many distributions have been proposed to take into account the fat tails of the daily returns. For instance, the Student's t-distribution, the Generalized Error Distribution (GED) of Nelson (1991), the Generalized Hyperbolic Distribution (GHD) due to Eberlein and Keller (1995), the Noncentral t Distribution proposed by Harvey and Siddique (1999), and so forth, could be used. All these distributions allow to have thicker tails than those of the Normal distribution.

In Figure 1.3 the unconditional distributions of the COF daily logarithmic returns are plotted against a Normal distribution, for each subsample. In panel (B), where the period displayed is characterized by more stability, the Normal assumption could be not rejected. Instead, in panel (A) and (C), where the market experienced greater instability, the fat tails of the daily returns are immediately visible.

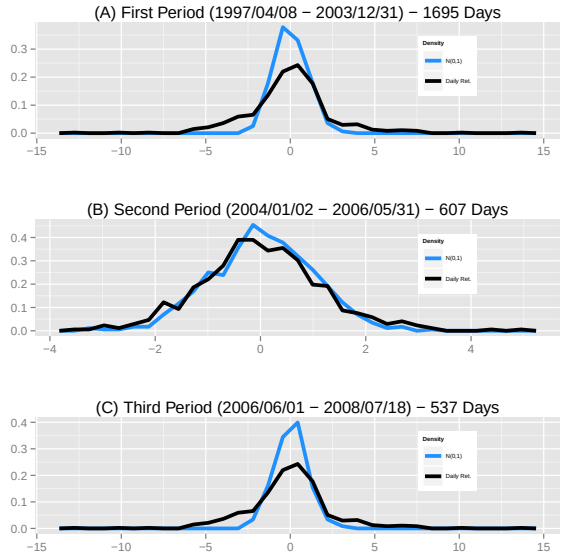


Figure 1.3 – COF fat tails

1.3.3 Leverage Effect

The "leverage effect" due to Black (1976) means that negative notices[2] have a different impact on the volatility. Typically, bad notices produce an increase of volatility, compared to good ones. In presence of past negative returns, the resulting volatility is often higher than the volatility that occurs in presence of same size (in absolute terms) positive returns. The reason is as follows: when the stock prices of a firm fall (and thus the returns become negative), the debts remain stable in the short period; hence, the debt-to-equity ratio (also called "leverage") increases. At this point, the market reacts to the fact that the future of

[2]For negative notices we refer to all the news concerning a "X" company, the economic sector in which the "X" company operates or the economic system as whole, such that the "X" stock value could be reduced.

that firm has an increased uncertainty and thus the stock price exhibits more variability. Evidences of the leverage effect are provided by Chen and Wang (2004), among others. In literature there have been proposed different models capable to capture the leverage effect, models that will be presented afterwards.

1.3.4 Calendar Anomalies

The so-called "calendar anomalies" have been largely studied in the literature (see French (1980) and Rogalski (1984), among others). The calendar anomalies can be distinguished in *January effect* and *Week-end* or *day of the week effect*.

The January effect means that the stock price tend to rise during the period starting on the last day of December and ending on the fifth trading day of January. This happens for several reasons, as indicated in Jacobs and Levy (1988). For instance it could be happen that at the end of the year the investors have to pay taxes and so there is less liquidity but afterwards in January the money availability becomes higher causing a re-buying of the stock.

With reference to the day of the week effect, the average return on Monday is sometimes significantly less than the average return over the other days of the week. The presence of day of the week effect can be tested inserting in the mean equation or in the variance equation a dummy variable which assumes a value equal to one if the day is Monday and zero otherwise. Then the unknown parameter linked to this dummy variable is estimated and its statistical significance is studied. This approach can be easily extended to verify which week day helps to explain the conditional mean and variance of the stock return by following the same methodology. Berument and Kiymaz (2001) find empirically that the week effect is present in both return and volatility equations.

We present some evidences of the calendar anomalies in Figures 1.4 and 1.5. These figures show the boxplots of the COF daily returns for

each month and each day of the week, respectively. In these boxplots, the top two largest values, both negative and positive, have been ruled out. With reference to the differences among the monthly averages and medians, we find some evidences mainly in the second and third sub-samples (panel (B) and (C)). In the first sub-sample, that is the largest, the differences among the months are attenuated by the length of the period. However, it is pretty clear that the means and the medians remain fairly stable during the summer periods while they tend to change more during the last months of the year. As regards to the week of the day effect, again the length of the period mitigates the differences among the daily averages and medians, such that the first period (panel (A)) does not highlight any relevant inequality. Instead, mainly in the second period, the daily averages and medians change over the weekdays.

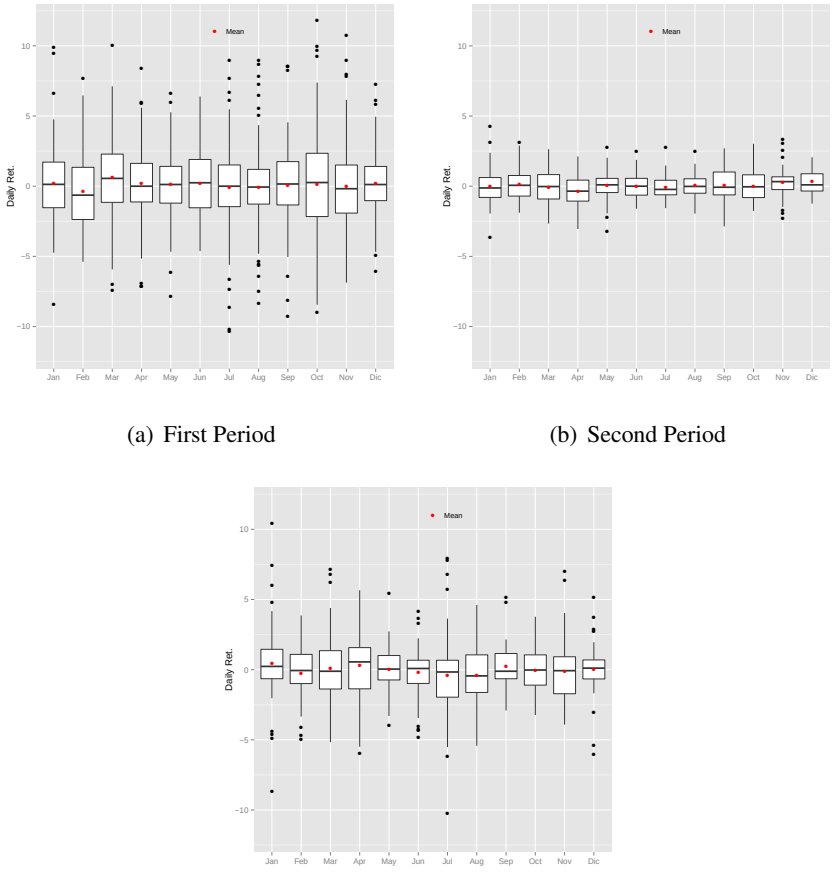
1.4 Univariate volatility models

In this section two approaches that aim to estimate $\hat{h}$ are overviewed: (i) an exploratory method, like the RiskMetrics approach; (ii) a stochastic method, that is GARCH and related models.

1.4.1 RiskMetrics

RiskMetrics estimator, proposed by the risk management group of JP-Morgan[3] in 1996, has the advantage of being easily calculated, given that it does not require any estimation procedure. It represents a good starting point for comparing the estimated conditional variance with conditional variances obtained with other methods. For the RiskMetrics estimator the variance at time t derives from a weighted mean of the variance at time

[3] JPMorgan Chase & Co. is an American multinational banking and financial services holding company.

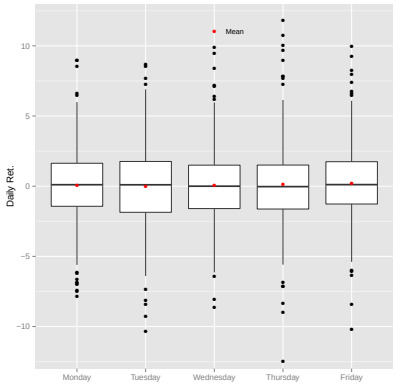


(a) First Period

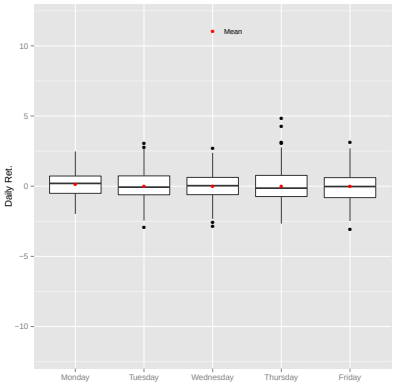
(b) Second Period

(c) Third Period

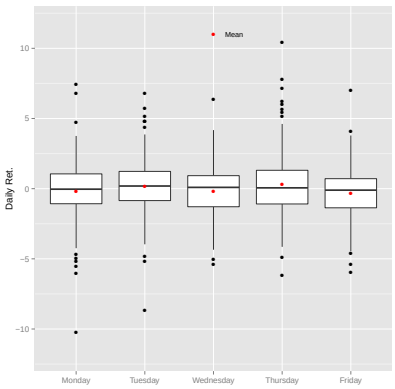
Figure 1.4 – COF monthly boxplot



(a) First Period



(b) Second Period



(c) Third Period

Figure 1.5 – COF weekly boxplot

$t-1$ and the squared returns at $t-1$:

$$\hat{\sigma}_t^2 = \lambda \hat{\sigma}_{t-1}^2 + (1-\lambda) r_{t-1}^2, \tag{1.3}$$

with $0 < \lambda < 1$, $\hat{\sigma}_0^2$ denoting the variance at time 0, usually equal to the unconditional sample variance. JPMorgan suggests a value of λ equal to 0.94 for daily returns and equal to 0.97 for weekly returns. Being a weighted moving average, the RiskMetrics approach gives more importance to the more recent observations. Moreover, the λ parameter provides the volatility persistence: if $\lambda \to 1$, then the memory of the estimator is very high, since the variance at time t is given mostly by the variance at time $t-1$. For other details, see the original technical document provided by Morgan (1996).

1.4.2 GARCH models

Let $\{r_t(\theta)\}$ denote a discrete time stochastic process with conditional mean and variance functions parametrized by the finite dimensional vector $\theta \in \Theta$, where Θ is the parametric space and θ_0 denotes the true value. As above mentioned, here r_t is assumed to be a scalar, with the obvious extensions to a multivariate framework treated in the next section. In the subsequent discussions we shall focus on $\{r_t\}$ process, but the same ideas may be extended to the situation in which we consider another stochastic process $\{\xi_t\}$ corresponding to the innovations from some more elaborate econometric model. In particular, let $\{y_t(\theta)\}$ denote the stochastic process of interest with conditional mean

$$\mu_t(\theta) = \mathbb{E}(y_t | I^{t-1}) \qquad t = 1, \cdots, T; \tag{1.4}$$

Hence, define the $\{\xi_t\}$ process by

$$\xi_t(\theta) = y_t - \mu_t(\theta) \qquad t = 1, \cdots, T; \tag{1.5}$$

The conditional variance for $\{\xi_t\}$ then equals the conditional variance of the $\{y_t\}$ process. The last two equations therefore provide the framework

on which we shall focus. In fact, the daily returns are directly observed. If in $\{r_t\}$ there is a conditional mean structure, $\mu_t(\theta)$ is estimated by means of the class of ARMA[4] model and then, on the fitted residuals, obtained through the difference expressed in (1.5), the volatility can be estimated. In order to avoid burdening the notation, we exclude this last case: in what follows, the focus will be on conditional variance of r_t, which is therefore assumed to be a weak *white noise* while μ_t is considered equal to zero.

The general conditional heteroskedastic model can be defined as

$$\begin{aligned} r_t &= h_t z_t \qquad t = 1, \cdots, T; \\ h_t^2 &= f(I^{t-1}, \theta), \end{aligned} \tag{1.6}$$

with z_t is a $i.i.d.$ *white noise* process with zero mean and unit variance, formally:

$$z_t \sim \mathcal{N}(0, 1). \tag{1.7}$$

As stated above, h_t^2 is the conditional variance of r_t at time t and the vector θ contains the unknown model's parameters. By construction, $r_t|I^{t-1} \sim \mathcal{N}(0, h_t^2)$, that is also r_t is conditionally normally distributed.

The general model in (1.6) includes several other models of conditional heteroskedasticity. Each model represents a different choice of the function $f(\cdot)$.

ARCH Model

The autoregressive conditional heteroskedasticity (ARCH) model proposed by Engle (1982) was the first parametric model able to capture the time-varying variance. While the first empirical applications of the ARCH models were concerned with modeling the uncertainty of the inflation rate, the methodology has subsequently found especially extensive application in capturing the temporal dependencies in the conditional

[4] See de Gooijer et al. (1985) for a review on AutoRegressive (AR), Moving Average (MA) and AutoRegressiveMovingAverage (ARMA) models.

variance of asset returns.

The ARCH(p) model, with p term denoting the number of lags, assumes $f(\cdot)$ as given by a linear combination of the p past values of r_t^2:

$$h_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j r_{t-j}^2 = \alpha_0 + \alpha_p(L) r_t^2. \tag{1.8}$$

Thus, the conditional variance of r_t in an ARCH(p) model is a linear function of the p squared lags. The unknown parameters in (1.8) are $\alpha_0, \alpha_1, \cdots, \alpha_p$.

Defining $v_t = r_t^2 - h_t^2$, the ARCH(p) model in (1.8) may be re-written as:

$$r_t^2 = \alpha_0 + \alpha(L) r_t^2 + v_t.$$

Since $\mathbb{E}(v_t | I^{t-1}) = 0$, the model corresponds directly to an autoregressive model of order p (noted as AR(p)) for the squared daily returns, r_t^2.

The parameter restrictions in (1.8) for positivity of the conditional variance are: $\alpha_0 > 0$ and $\alpha_j \geq 0$, for $j = 1, \cdots, p$.

The process is covariance stationary if and only if the sum of the positive autoregressive parameters is less than one. In this case the unconditional variance is equal to $Var(r_t) = \sigma^2 = \alpha_0/(1 - \alpha_1 - \cdots - \alpha_p)$.

It is easy to demonstrate that an ARCH model has a kurtosis larger than 3, using the law of iterated expectations[5]. Remember that the kurtosis is based on the forth moment of the data, in this case r_t. Let us consider the case that the conditional variance of r_t follows an ARCH(1)

[5]The law of iterated expectations states that if X is a random variable satisfying $\mathbb{E}(|X|) < \infty$) and Y is any random variable, on the same probability space, then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)),$$

i.e., the expected value of the conditional expected value of X given Y is the same as the expected value of X.

process. The (conditional) forth moment of r_t is:

$$\begin{aligned}\mathbb{E}(r_t^4|I^{t-1}) &= \mathbb{E}(h_t^4 z_t^4|I^{t-1}) \\ &= \mathbb{E}(z_t^4|I^{t-1})\mathbb{E}((h_t^2)^2|I^{t-1}) \\ &= 3(\alpha_0 + \alpha_1 r_{t-1}^2)^2, \end{aligned} \tag{1.9}$$

where (1.9) immediately comes from (1.7) and (1.8). Applying the law of iterated expectations, it follows:

$$\begin{aligned}\mathbb{E}(r_t^4) &= \mathbb{E}(\mathbb{E}(r_t^4|I^{t-1})) \\ &= 3\mathbb{E}(\alpha_0 + \alpha_1 r_{t-1}^2)^2 \\ &= 3(\alpha_0^2 + 2\alpha_0\alpha_1\mathbb{E}(r_{t-1}^2) + \alpha_1^2\mathbb{E}(r_{t-1}^4)) \\ &= 3(\alpha_0^2 + 2\alpha_1\frac{\alpha_0^2}{1-\alpha_1} + \alpha_1^2\mathbb{E}(r_{t-1}^4)).\end{aligned}$$

Assuming that the process is stationary both in variance and in the fourth moment, if $\mathbb{E}(r_t^4) = c$,

$$c = \frac{3\alpha_0^2(1-\alpha_1^2)}{(1-\alpha_1^2)(1-3\alpha_1^2)}.$$

Simple algebra then reveals that the kurtosis is

$$\kappa_{(r_t)} = \frac{\mathbb{E}(r_t^4)}{\mathbb{E}(r_t^2)^2} = \frac{\mathbb{E}(r_t^4)}{\sigma^4} = \frac{3(1-\alpha_1^2)}{(1-3\alpha_1^2)} \geq 3.$$

Hence, the unconditional distribution of r_t is leptokurtic. That is to say that the ARCH(1) process has tails heavier than the normal distribution, as desired. This property makes the ARCH model attractive because, as seen previously, the distribution of asset returns display tick tails.

The greatest limit of ARCH formulation is that it requires too many parameters (i.e., too many lagged daily returns) to model the persistence of the conditional variance. Some practitioners suggest a minimum order of $p = 14$ to model adequately the volatility. Obviously, the estimation of large dimensional parameter spaces becomes quite complicated when

the number of parameters increases. The estimation of ARCH models is normally done using the maximum likelihood (ML) method, whose application in the time series framework will be explained later. To overcome the over-parametrization of ARCH, Bollerslev (1986) proposed the generalized ARCH (GARCH) model which allows a more parsimonious parametrization of the conditional variance process.

GARCH model

A GARCH(p, q) model considers the conditional variance to be a linear function of the past p squared innovations and of the past q variances. Similar to the ARCH model, it also includes a constant. A GARCH(p, q) is formalized as:

$$h_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}^2 \tag{1.10}$$

$$= \alpha_0 + \alpha_p(L) r_{t-1}^2 + \beta_q(L) h_{t-1}^2. \tag{1.11}$$

If the conditional variance follows a GARCH(1,1) process, the coefficient α_1 measures the intensity of reaction of volatility to yesterday unexpected market return r_{t-1}^2 and the coefficient β_1 measures the effect of one lagged volatility. For $q = 0$, the process reduces to the ARCH(p) model, and for $p = q = 0$, r_t is reduced to a *white noise* process.

The conditional variance for a GARCH(p, q) model is well defined if all the coefficients are positive. For a GARCH($1, 1$) model, almost sure positivity of h_t^2 it assured if $\alpha_0 > 0$, $\alpha_1 \geq 0$ and $\beta_1 \geq 0$.

A GARCH(p, q) model with positive coefficients $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \cdots, p$, and $\beta_j \geq 0$, $j = 1, \cdots, q$, is covariance stationary if

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1,$$

as demonstrated by Lindner (2009), among others.

In empirical applications with financial data the estimate for $\alpha_p(L) +$

$\beta_q(L)$ turns out to be very close to unity. This provides an empirical motivation for the so-called integrated GARCH(p, q) model, commonly defined as IGARCH(p, q) model, introduced by Engle and Bollerslev (1986). In the IGARCH class of models a shock to the conditional variance is highly persistent.

As earlier mentioned, the GARCH formulation is more parsimonius in terms of parameters than the ARCH model. If an ARMA model often leads to a more parsimonious representation of the temporal dependencies in the conditional mean compared to an AR model, the GARCH(p, q) formulation in (1.10) provides a similar added flexibility over the linear ARCH(p) model. In fact, the GARCH model can be expressed as an ARCH(∞). Let us consider a GARCH(1,1) model as that of equation (1.12). Then h_{t-1}^2 is calculated in (1.13) and it is inserted in (1.12):

$$h_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 h_{t-1}^2 \tag{1.12}$$

$$h_{t-1}^2 = \alpha_0 + \alpha_1 r_{t-2}^2 + \beta_1 h_{t-2}^2 \tag{1.13}$$

$$\begin{aligned} h_t^2 &= \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1(\alpha_0 + \alpha_1 r_{t-2}^2 + \beta_1 h_{t-2}^2) \\ &= \alpha_0 + \beta_1\alpha_0 + \alpha_1 r_{t-1}^2 + \alpha_1\beta_1 r_{t-2}^2 + \beta_1^2 h_{t-2}^2 \end{aligned} \tag{1.14}$$

Going backwards with the same trick for k steps, with k large enough, the contribute of h_{t-k}^2 will be negligible, due to $\beta_1^k \to 0$, given that by construction $\beta_1 < 1$. Hence, a GARCH(1,1) can reproduce the same volatility pattern as an ARCH(∞).

In Figure 1.6 the conditional variances for some simulated data are illustrated. In particular, the data generating process is a GARCH(2,2) process with $z_t \sim \mathcal{N}(0, 1)$ and $T = 200$. The true parameters are $\alpha_0 = 0.01$, $\alpha_1 = 0.08$, $\alpha_2 = 0.02$, $\beta_1 = 0.75$ and $\beta_2 = 0.10$. In the figure the true conditional variance is reported with a thicker line in order to distinguish it to the forecasted variances produced by three GARCH models, that are: an ARCH(4), an ARCH(14) and a GARCH(1,1). All the models follow quite good the pattern of the true variance even though they exhibit large deviations from it. In particular, the ARCH(4) model experiences

the greatest distance from the true variance, while the GARCH(1,1) the smallest one.

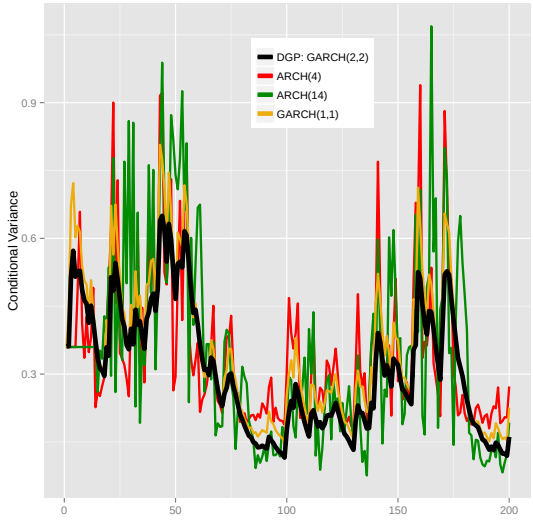


Figure 1.6 – GARCH(2,2) vs ARCH(4), ARCH(14) and GARCH(1,1) models

To conclude, the main advantage of the GARCH model is that it requires a lower number of parameters than the ARCH model, keeping all the good features of this last one: it is able to model the excess kurtosis of daily returns and capture the volatility clustering phenomenon. In particular, considering this latter and given that the GARCH models assume that conditional variances depend on realizations of past innovations and on lagged variances, the main appeal of this class of models is that “big” surprises, of either sign, increase uncertainty and therefore will be more likely to be followed by other “big" surprises. Another advantage of the model is its relative simplicity. However, the GARCH models present some drawbacks. A disadvantage of the standard GARCH models is that they cannot model asymmetries of the volatility with respect to the sign

of past shocks (leverage effect). This results from the squared form of the lagged shocks in (1.8) and (1.10). Therefore they have an effect on the level but no effect on the sign: bad news (identified by a negative sign of daily returns at time $t-1$) have the same influence on the volatility at time t as good news if the absolute values are the same. Moreover, the parameter restrictions that are required to ensure positivity of the conditional variance at every point of time represent another drawback.

An Asymmetric GARCH Model: the GJR-GARCH

The GJR-GARCH model of Glosten et al. (1993) gives a higher weight to negative returns, introducing a parameter that increases the conditional variance if $r_{t-1} < 0$. Formally, a GJR-GARCH(1,1) is

$$h_t^2 = \alpha_0 + (\alpha_1 + \gamma_1 I_{(r_{t-1}<0)})(r_{t-1}^2) + \beta_1 h_{t-1}^2, \qquad (1.15)$$

where $I_{(r_{t-1}<0)}$ is an indicator function assuming value one when the argument is true and zero otherwise. Thus, if there is a negative change of daily returns at time $t-1$, the volatility increases of $\alpha_1+\gamma_1$ at time t, capturing the leverage effect. This model has very similar behaviour to the classic GARCH, i.e. we have positive conditional variance, considering a GJR-GARCH(1,1), if:

$$\alpha_0 > 0 \qquad (\alpha_1 + \gamma_1)/2 \geq 0 \qquad \beta_1 \geq 0,$$

and covariance stationarity if:

$$(\alpha_1 + \gamma_1)/2 + \beta_1 < 1.$$

If the last condition is satisfied, the unconditional variance is

$$\sigma^2 = \alpha_0/(1 - (\alpha_1 + \gamma_1)/2 - \beta_1).$$

Another Asymmetric GARCH model: The E-GARCH

The Exponential GARCH (EGARCH) model, introduced by Nelson (1991), overcomes both the drawbacks mentioned above: the parameter restrictions and the lack of leverage effect in the standard GARCH models. The EGARCH models has the following form:

$$h_t^2 = \exp\left[a_0 + a_{1a}\frac{r_{t-1}}{h_{t-1}} + a_{1b}\left(\frac{|r_{t-1}|}{h_{t-1}} - \mathbb{E}\left[\frac{|r_{t-1}|}{h_{t-1}}\right]\right) + g_1 \log h_{t-1}^2\right]$$

For $r_t \sim \mathcal{N}(0, h_t^2)$ the standardised variable $\frac{r_t}{h_t}$ follows a standard normal distribution and consequently $\mathbb{E}\left[\frac{|r_t|}{h_t}\right] = \sqrt{\frac{2}{\pi}}$. The parameter a_{1a} captures the leverage effect. For "good news" $\left(\frac{r_{t-1}}{h_{t-1}} > 0\right)$ the impact of the innovation r_{t-1} is $(a_{1b} + a_{1a})\frac{r_{t-1}}{h_{t-1}}$ and for "bad news" $\left(\frac{r_{t-1}}{h_{t-1}} < 0\right)$ it is $(a_{1b} - a_{1a})\frac{r_{t-1}}{h_{t-1}}$. If $a_{1a} = 0$, h_t^2 responds symmetrically to $\frac{r_{t-1}}{h_{t-1}}$. To produce a leverage effect, a_{1a} must be negative.

Given that the EGARCH process is specified in terms of log-volatility implies that h_t^2 is always positive and, consequently, there are no restrictions on the parameters.

1.4.3 Final Considerations on the GARCH models

The previous models are only a small part of the wide literature on GARCH models that has been developed during last decades. For more details on the recent developments, see the work of Zivot (2009), where several univariate GARCH models are empirically estimated and compared. It is important to underline that we focused on these models because these will be used in the empirical part of the work.

Before the breakthrough work of Engle (1982), the time-varying variance were modelled by nonparametric procedures like a recursive estimation over time (Mandelbrot (1963)) or a moving variance estimation (Klein (1977)). After Engle's work, many of the stylized facts have been captured by the GARCH models. In fact, GARCH models are able to

capture the excess of kurtosis and the volatility clustering. Moreover, the GJR-GARCH can capture the leverage effect and the EGARCH model does not require any parameter restrictions.

With reference to $\hat{h}_m$ in (1), each univariate volatility model presented above will produce a volatility forecast. In the next chapter the attention will be focused on the methods used to compare the predictions, while in the next section the multivariate specification is presented.

1.5 Multivariate volatility models

When more than one stock is owned or analysed, other instruments have to be considered to estimate the volatility. In fact each conditional variance of a stock Y_1 could be explained by other current or lagged realizations of other stocks $Y_2, \cdots, Y_k$. The analysis becomes more complicated because the output of the volatility estimation is now a $k \times k$ matrix, generally called covariance matrix. In this framework, the principal diagonal of the covariance matrix collects the conditional variances of each stock while the conditional covariances lie on the extra-diagonal elements. By construction, the covariance matrix is symmetric but the restrictions needed to assure the positive definiteness are sometimes cumbersome as long as the number of stock increases. Then, the covariance matrix is normally used as a key input for the portfolio optimization, for instance. A portfolio optimization using the covariance matrix as key input is reviewed by Chan et al. (1999) while for the basis of the portfolio optimization, see the seminal papers of Markowitz (1952, 1959).

A general formulation for the multivariate GARCH (MGARCH) model is:

$$\mathbf{r}_t = H_t^{0.5}\mathbf{z}_t, \tag{1.16}$$

where $\mathbf{r}_t$ is a $k \times 1$ vector of k daily logarithmic returns at time t, $H_t^{0.5}$ is a $k \times k$ symmetric matrix, and $\mathbf{z}_t$ is a $k \times 1$ vector of $i.i.d.$ innovations at time t, such that $\mathbf{z}_t \sim \mathcal{MVN}(\mathbf{0}; I_k)$. $\mathcal{MVN}$ indicates the "Multivariate

Normal" distribution and I_k a $k \times k$ matrix of ones. It is assumed that $\mathbb{E}(\mathbf{r}_t) = \mathbf{0}$. Within this framework, $\mathbf{r}_t|I^{t-1} \sim \mathcal{MVN}(\mathbf{0}, H_t)$, given that:

$$\mathbb{E}(\mathbf{r}_t|I^{t-1}) = \mathbb{E}(H_t^{0.5}\mathbf{z}_t) = H_t^{0.5}\mathbb{E}(\mathbf{z}_t) = \mathbf{0} \tag{1.17}$$

and

$$\begin{aligned} VAR(\mathbf{r}_t|I^{t-1}) &= VAR(H_t^{0.5}\mathbf{z}_t) \\ &= (H_t^{0.5})VAR(\mathbf{z}_t)(H_t^{0.5})' \\ &= H_t I_k \\ &= H_t. \end{aligned} \tag{1.18}$$

Thus, in this framework, H_t is the conditional covariance matrix for the day t and it represents our object of interest. In the spirit of Bauwens et al. (2006), the MGARCH models will be presented following three non mutually exclusive approaches for determining H_t:

1. direct generalizations of the univariate GARCH models of Bollerslev (1986);
2. linear combinations of univariate GARCH models;
3. nonlinear combinations of univariate GARCH models.

Regardless of the approaches used to describe the models, a multivariate GARCH model should be flexible enough to be able to represent the dynamics of the conditional variances and covariances, from one hand. From the other hand, the specification should be parsimonious enough to allow for a relatively easy estimation of the model and an easy interpretation of the model parameters. Unfortunately, flexibility and parsimony go in opposite directions, making the MGARCH models a challenging research field. The positive definiteness of H_t represents another complication. One possibility is to derive conditions under which the conditional covariance matrices implied by the model are positive definite. However,

this method is often infeasible in practice. An alternative is to formulate the model in a way that positive definiteness is implied by the its structure.

1.5.1 Generalizations of the univ. GARCH model

This class of models derives H_t as a the direct generalization of the univariate GARCH model of Bollerslev (1986).

The VEC and DVEC models (Bollerslev et al. (1988))

The VEC(1,1) model is defined as:

$$h_t = C + A\eta_{t-1} + Gh_{t-1} \tag{1.19}$$

where:

$$h_t = vech(H_t);$$
$$\eta_{t-1} = vech(\mathbf{r}_{t-1}\mathbf{r}_{t-1}^{'}),$$

and $vech(\cdot)$ denotes the operator that stacks the lower triangular portion of a $k \times k$ matrix as a $k(k+1)/2 \times 1$ vector. C is a $k(k+1)/2 \times 1$ matrix; A and G are $k \times k$ matrix. The parameters in C, A and G have to be estimated, thus we have $(p+q)(k(k+1)/2)^2 + k(k+1)/2$ unknown parameters. The VEC model is covariance-stationary if the eigenvalues of $A + G$ are less than one in modulus. A representation of a VEC(1,1) for $k = 2$ is:

$$\begin{bmatrix} h_{1,t}^2 \\ h_{12,t} \\ h_{2,t}^2 \end{bmatrix} = \begin{bmatrix} c_{11} \\ c_{12} \\ c_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} r_{1,t-1}^2 \\ r_{12,t-1} \\ r_{2,t-1}^2 \end{bmatrix} +$$
$$= \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \begin{bmatrix} h_{1,t-1}^2 \\ h_{12,t-1} \\ h_{2,t-1}^2 \end{bmatrix},$$

where $r_{12,t-1}$ is the cross-product between the daily returns of the two stocks and $h_{12,t-1}$ is the covariance at time $t-1$. In the general VEC model, each element of H_t is a linear function of the lagged squared $\mathbf{r}_t$, of the cross products of $\mathbf{r}_t$ and lagged values of the elements of H_t. Even for low dimensions of k and small values of p and q the number of parameters is very large; for instance, for $k = 5$ and $p = q = 1$ the VEC model requires 465 parameters to estimate. This allows plenty of flexibility to capture most of the cited empirical regularities at a cost of an estimation very computationally demanding. Moreover, the conditions to ensure that the conditional covariance matrices are positive definite $\forall t$ are difficult to impose and verify.

A method reducing the number of parameters is the "variance targeting". Proposed by Engle and Mezrich (1996), the variance targeting estimates the matrix C in (1.19) as:

$$C = (I - A - G)S, \tag{1.20}$$

where S is the (unconditional) covariance matrix. This method can be used not only for the parameter's reduction in VEC model, but also in other models.

The same authors of the VEC suggested another simplifying assumption: the diagonal VEC (DVEC) model in which the A and G matrices are imposed to be diagonal. A representation of a DVEC for two stocks is:

$$\begin{aligned}\begin{bmatrix} h^2_{1,t} \\ h_{12,t} \\ h^2_{2,t} \end{bmatrix} &= \begin{bmatrix} c_{11} \\ c_{12} \\ c_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \begin{bmatrix} r^2_{1,t-1} \\ r_{12,t-1} \\ r^2_{2,t-1} \end{bmatrix} + \\ &= \begin{bmatrix} g_{11} & 0 & 0 \\ 0 & g_{22} & 0 \\ 0 & 0 & g_{33} \end{bmatrix} \begin{bmatrix} h^2_{1,t-1} \\ h_{12,t-1} \\ h^2_{2,t-1} \end{bmatrix}.\end{aligned}$$

The positive definitiveness of H_t is given expressing the equation (1.19) in terms of the Hadamard product[6] (denoted by $\odot$):

$$H_t = C + A \odot (r_{t-1} r'_{t-1}) + G \odot H_{t-1}. \tag{1.21}$$

As showed by Marcus and Minc (1992), H_t in (1.21) is positive definite $\forall t$ if C is positive definite and the A and G matrices are positive semi-definite.

The estimation of a DVEC is less difficult than that of the full VEC model because each equation can be estimated separately. But the diagonal VEC model, that contains $(p+q+1)k(k+1)/2$ parameters, seems to be too restrictive because no interaction is allowed between the different conditional variances and covariances.

The BEKK model (Engle and Kroner (1995))

In the BEKK model H_t is positive definite by construction. The model has the form

$$H_t = CC' + \sum_{i=1}^{p} \sum_{j=1}^{J} A_{ij}^{'} \mathbf{r}_{t-i} \mathbf{r}_{t-i}^{'} A_{ij} + \sum_{i=1}^{q} \sum_{j=1}^{J} G_{ij}^{'} H_{t-i} G_{ij} \tag{1.22}$$

where A_{ij} , G_{ij} , and C are $k \times k$ parameter matrices and C is lower triangular. The decomposition of the constant term into a product of two triangular matrices is needed to ensure positive definiteness of H_t. The BEKK model is covariance stationary if and only if the eigenvalues of $\sum_{i=1}^{p} \sum_{j=1}^{J} A_{ij} \otimes A_{ij} + \sum_{i=1}^{q} \sum_{j=1}^{J} G_{ij} \otimes G_{ij}$, where $\otimes$ denotes the Kronecker[7] product of two matrices, are less than one in modulus. Whenever $J > 1$

[6] Given two matrices of the same dimensions, the Hadamard product produces another matrix where each element ij is the product of elements ij of the original two matrices.

[7] The Kronecker product, denoted by $\otimes$, is an operation on two matrices of arbitrary size resulting in a block matrix. If A is an $m \times n$ matrix and G is a $p \times q$ matrix, then the Kronecker product A $\otimes$ G is the $mp \times nq$ block matrix.

an identification problem arises because there are several parametrizations that yield the same representation of the model.

The first order model is:

$$H_t = CC' + A(\mathbf{r}_{t-1}\mathbf{r}'_{t-1})A' + GH_{t-1}G' \tag{1.23}$$

The diagonal BEKK (DBEKK) model is obtained from the general BEKK model by constraining A and G to be diagonal. As for DVEC model, the main reason for introducing these constraints is the need for a more parsimonious model. The most restricted version of the diagonal BEKK model is the scalar BEKK with $A = aI$ and $B = bI$ where a and b are scalars and I is the identity matrix.

Engle and Kroner provided also sufficient conditions for the two models, BEKK and VEC, to be equivalent. They also gave a representation theorem that establishes the equivalence of diagonal VEC model and diagonal BEKK model.

The estimation of a BEKK model still involves somewhat heavy computations: the number of parameters, $(p+q)Jk^2 + k(k+1)/2$ in the full BEKK model, or $(p+q)Jk+k(k+1)/2$ in the diagonal one, is still quite large. Obtaining convergence may therefore be difficult because (1.22) is not linear in parameters. However, there is the advantage that the structure automatically ensures positive definiteness of H_t. In order to avoid these computational problems, it is typically to assume $p = q = J = 1$. In Table 1.1 the comparison between the parameters' number to estimate for each of the presented models is showed.

1.5.2 Linear combinations of univ. GARCH models

In the previous section, we have seen that VEC and BEKK are affected by the "curse of dimensionality". If the assets' number increases, then the number of parameters becomes very large such that not only it is difficult to interpret them but also a computational problem arises. One way to simplify the problem of the dimensionality is to use the so-called

Table 1.1 – Parameters' Number in VEC, DVEC, BEKK and DBEKK models

Models \ k	3	5	10	Formulation
VEC	78	465	6105	$(p+q)(k(k+1)/2)^2 + k(k+1)/2$
DVEC	18	45	65	$(p+q+1)k(k+1)/2$
BEKK[a]	24	65	255	$(p+q)Jk^2 + k(k+1)/2$
DBEKK[a]	12	25	75	$(p+q)Jk + k(k+1)/2$

[a] The number of parameters are obtained by setting $p = q = J = 1$.

"Factor Models", introduced by Engle et al. (1990). In this type of models it is assumed that the observations are generated by some underlying factors that are conditionally heteroskedastic and possess a GARCH-type structure. This approach reduces the dimensionality of the problem when the number of factors relative to the dimension of the return vector $\mathbf{r}_t$ is small. In this approach the intuition is that if the factors are uncorrelated, then they represent genuinely different common components driving the returns. Motivated by this consideration, several factor models with uncorrelated factors have been proposed in the literature (Diebold and Nerlove (1989), Forni et al. (2000), and many many others). In all of them, the observed daily returns $\mathbf{r}_t$ are assumed to be driven by some unobserved and uncorrelated variables noted as factors, through a linear invertible transformation W. In this work we present the model proposed by Alexander (2000), defined as the Orthogonal model because the factors are orthogonal between themselves. Afterwards, a generalization of the Orthogonal model is briefly illustrated.

Alexander's Orthogonal model (2001)

The principal idea of Alexander's Orthogonal (OGARCH) model is that the financial markets are influenced by many heterogeneous components such that it becomes fundamental to filter the relevant informations into a few factors influencing all the variables.

If we were interested in finding a set of observed risk factors influencing the volatility, we should use a matrix with interest rates, exchanges rates, commodity prices, and so forth. Instead, we can find risk factors using the $T \times k$ matrix of the original data to find a common factor structure, representing $m < k$ unobservable and uncorrelated (risk) factors. The common structure is estimated by a principal component analysis (*PCA*), that captures as much as possible of the information of the contained in the data by means of a reduced number of factors[8]. Practically, the *PCA* aims to exclude the noise from the data.

Let Y be the original matrix of daily returns. The *PCA* will give up to *k* uncorrelated stationary variables, called the principal components of Y. As aforementioned, we suppose that the observed data are given by a linear combination of the original data, that is the factors from the *PCA* noted as f_k, and by an error term, noted as e_i:

$$Y = \sum_{k=1}^{m} w_{ik} f_k + e_i.$$

We assume that, for $k = 1, \cdots, m$, $j \neq k$:

$$Cov(f_k, f_j) = 0;$$
$$Cov(f_k, e_i) = 0.$$

To calculate the conditional covariance of Y, the steps are as follows.

1. From the observed data matrix $\underset{T\times k}{Y}$, the standardized data matrix $\underset{T\times k}{X}$ is obtained. All the *k* columns of X have mean zero and variance one. For each column, the standardization is $(Y_{i,t} - \mu_i)/\sigma_i$, with $i = 1, \cdots, k$ and μ_i and σ_i indicating the unconditional mean and standard deviation of the $i - th$ column of Y.

[8] *PCA* is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

2. The *PCA* of $X'X/T$ is carried out, where $X'X/T$ represents the unconditional sample correlation matrix of the stock returns. Only the first *m* of *k* components are considered, on the basis of the highest eigenvalues associated to the matrix $X'X/T$. Let $\underset{k\times m}{W}$ be the matrix of the eigenvectors, also called the matrix of the "factor weights".

3. The principal components of Y are determined by the linear and invertible transformation W multiplied by X:

$$\underset{T\times m}{P} = \underset{T\times k}{X}\,\underset{k\times m}{W}\ .$$

Now, an univariate GARCH model is estimated for each $m-th$ column of P obtaining the diagonal matrix $D_t = diag(Var(P|I^{t-1}))$. Note that each column of the normalized matrix X could be rewritten as:

$$x_{i,t} = w_{i1}P_1 + \cdots + w_{ik}P_k, \text{ with } i = 1, \cdots, k, \tag{1.24}$$

if we used all the principal components instead of the first m.

4. Given that we are interested in the conditional covariance of the original matrix Y, equation (1.24) becomes:

$$y_i = \omega^*_{i1}P_1 + \cdots + w^*_{im}P_m + e_i,$$

where $\omega^*_{ij} = w_{ij}\sigma_i$, with $j = 1, \cdots, m$, and the error term e_i picks up the approximation from using only the first m of the k principal components. Hence, assuming that the unconditional error variance σ^2_e goes to zero as long as $m \to k$, the conditional covariance matrix of Y is:

$$\begin{aligned} Var(Y|I^{t-1}) &= W^*(Var(P|I^{t-1})W'^* + \sigma^2_e \\ Var(Y_t|I^{t-1}) &= W^*D_tW'^* + \sigma^2_e \\ Var(Y_t|I^{t-1}) &\approx W^*D_tW'^*, \end{aligned}$$

where W^* is the matrix form for $\omega^*_{ij} = w_{ij}\sigma_i$.

An advantage of the OGARCH is that there are no dimensional restrictions. Moreover, the factor model yields the desired reduction of parameters. Considering an orthogonal factor GARCH with m principal components written as OGARCH(m,1,1), the number of parameters to estimate is: $(k \times k)/2 + 3m$; if we have 10 asset returns and we consider the first two principal components, thus $m = 2$, we have to estimate 56 parameters, which are much less than 6105 or 255 parameters of VEC or BEKK, respectively. There is a clear advantage in transforming a complicated process in a more simple process, by splitting the estimation procedure in some sub-steps. The parameter constraints concerning m univariate GARCH, i.e applied on each column of P matrix, are:

$$\alpha_0 > 0; \quad \alpha_i, \beta_i \geq 0; \quad \alpha_i + \beta_i < 1; \quad i = 1, \cdots, m. \tag{1.25}$$

If (1.25) holds, the positivity of the conditional variance is assured as well as and covariance stationarity. Another advantage is that it is possible to estimate also different GARCH specifications for each column of P.

The greatest disadvantage is that the estimated conditional covariance matrix, that is the $k \times k$ matrix $Var(Y_t|I^{t-1})$, is positive semi-definite and not positive definite. If we use $m = k$ principal components and k univariate GARCH models have to be estimated, the positive definiteness of the conditional covariance matrix is assured by construction. However in this case there is no data reduction.

The generalized OGARCH model (van der Weide (2002))

The generalized OGARCH (GOGARCH) model is a natural generalization of the Alexander's orthogonal model. As seen before, The OGARCH model implicitly assumes that the observed data can be linearly transformed into a set of uncorrelated components. These unobserved components can be interpreted as a set of uncorrelated factors that drive the particular economy or market. However, the strongest assumption of the OGARCH is that the matrix linking the independent components to

the observed data is assumed to be orthogonal. The GOGARCH model relaxes this assumption such that when the link returns to be orthogonal, the OGARCH becomes a special case of the GOGARCH specification. In addition, the OGARCH model suffers from some identification problems: when the data exhibit weak correlation, the model has substantial difficulties to identify a matrix that is truly orthogonal. Instead, in the GOGARCH the data are not required to exhibit strong correlation for the method to work properly. Another difference between the OGARCH and GOGARCH model is that the latter model it is not possible to have fewer factors than assets.

The GOGARCH model formulation is as follows. As for the OGARCH, it is assumed that Y is driven by a linear combination of independent economic components P through the linear, constant over time and invertible map Z, such that $Y = PZ$. As for the OGARCH, the determination of the components P is carried out on the standardized data. Each of the k principal components is modelled as an univariate GARCH process. Relaxing the matrix notation:

$$\begin{aligned} p_{i,t}|I^{t-1} &\sim \mathcal{N}(0, h_{i,t}^2), \\ h_{i,t}^2 &= \alpha_{0,i} + \alpha_{1,i} p_{i,t-1}^2 + \beta_{1,i} h_{i,t-1}^2, \end{aligned} \tag{1.26}$$

with $i = 1, \cdots, k$. This implies that:

$$\begin{aligned} Y|I^{t-1} &\sim \mathcal{N}(0, V_t), \\ V_t &= ZH_tZ, \end{aligned} \tag{1.27}$$

with $H_t = diag(h_{i,t}^2, \cdots, h_{i,k}^2)$. As for the OGARCH, the previous framework can be extended allowing for a non-zero conditional mean of the original data in (1.27) or a different specification of the conditional variances in (1.26). For the practical identification issue of Z, see the original paper of var der Weide.

1.5.3 Nonlinear combinations of univ. GARCH models

In this section we present a set of models that may be viewed as non linear combinations of univariate GARCH models. These models specify separately, on the one hand, the individual conditional variances and on the other hand, the conditional correlation matrix or another measure of dipendence beetween the individual series. Thus, the most advantage of these models is that, by a multi-step procedure, they are more easily estimable.

Constant Conditional Correlation Models (1990) and Extensions

The conditional constant correlation (CCC) model, proposed by Bollerslev (1990), assumes that the correlations among k assets returns are are time-invariant. The CCC model is defined as:

$$H_t = D_t R D_t,$$

where R is the correlation matrix and $D_t = diag(h_{1,t}, \cdots, h_{k,t})$ is the conditional standard deviations at time t of the k asset returns. Being a correlation matrix, $\underset{k\times k}{R}$ has all ones on the principal diagonal. For this reason, the principal diagonal of H_t contains the conditional variances while the off-diagonal elements of the conditional covariance matrix are defined as follows:

$$[H_t]_{ij} = h_{i,t} h_{j,t} \rho_{ij}, \qquad i \neq j,$$

where ρ_{ij} denotes the correlation between the $i-th$ and $j-th$ asset. The standard deviations in D_t matrix are obtained taking the square root of the conditional variance as resulting by the already seen univariate GARCH(p, q) specification:

$$h_{i,t}^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_{ij}(r_{i,t-j}^2) + \sum_{j=1}^{q} \beta_{ij} h_{i,t-j}^2,$$

with $i = 1, \cdots, k$. The number of parameters of the CCC model, when $p = q = 1$, is $[k(k+5)]/2$.

The model with constant correlations is interesting as far as the interpretation is concerned. It facilitates the comparison between subperiods. One may independently estimate such submodels on separate subperiods and afterwards examine whether the correlation patters vary from one period to another. Nevertheless the approach presented above has the drawback of considering the assumption of time invariant correlations. For this reason, some models with time-varying correlation have been subsequently proposed.

An extension of the CCC-GARCH model was introduced by Jeantheau (1998). The Extended CCC- (ECCC-) GARCH model allows the past squared returns and variances of all series to enter the individual conditional variance equations. For instance, in the first order ECCC-GARCH model, the $i-th$ conditional variance does not depend only on its past squared returns and variance realizations, but also on the past squared returns and variance realizations of the other stocks:

$$\begin{aligned} h_{i,t}^2 =& \alpha_i + a_{11}(r_{i,t-1}^2) + \cdots + a_{1k}(r_{k,t-1}^2) + \\ & g_{11}h_{i,t-1}^2 + \cdots + g_{1k}h_{k,t-1}^2. \end{aligned}$$

Although the CCC-GARCH model is an attractive parametrization under many aspects, empirical studies have suggested that the assumption of constant conditional correlations may be too restrictive. This opens the doors to the dynamic conditional correlations (DCC) models.

Dynamic Conditional Correlation Models and Extensions

The dynamic conditional correlation (DCC) model is defined as:

$$H_t = D_t R_t D_t \tag{1.28}$$

In (1.28), the positive definiteness of H_t follows if all the conditional variances h_{it}^2 with $i = 1, \cdots, k$ are well-defined and the conditional cor-

relation matrix R_t is positive definite at each point in time. If D_t is the diagonal matrix of the conditional standard deviations as that of the CCC specification, during last years several specifications of R_t have been proposed.

Tse and Tsui (2002) impose a GARCH type of dynamics on the conditional correlations. In their specification, R_t is function of the past conditional correlations. More specifically, R_t is defined as

$$R_t = (1 - \alpha - \beta)S + aS_{t-1} + bR_{t-1}, \tag{1.29}$$

where S is the sample correlation matrix as defined as in CCC model, α and β are some non-negative scalar parameters such that $\alpha + \beta < 1$. S_t is the sample correlation matrix of the past M standardized residuals $\hat{\boldsymbol{\xi}}_{t-1}, \cdots, \hat{\boldsymbol{\xi}}_{t-M}$, where $\hat{\boldsymbol{\xi}}_{t-j} = \hat{D}_{t-j}^{-1}\mathbf{r}_{t-j}$, $j = 1, \cdots, M$. The positive definiteness of R_t is ensured by construction if S_{t-1} is positive definite. This happens when is $M \geq k$. The intercept in (1.29) corresponds to variance targeting specification.

The DCC model of Engle (2002), (DCC_E), the dynamic conditional correlation matrix is defined as follows:

$$R_t = diag(q_{11,t}^{-1/2}, \cdots, q_{kk,t}^{-1/2})\; Q_t \; diag(q_{11,t}^{-1/2}, \cdots, q_{kk,t}^{-1/2}); \tag{1.30}$$

$$Q_t = (1 - \alpha - \beta)Q + \alpha \mathbf{u}_{t-1}\mathbf{u}_{t-1}^{'} + \beta Q_{t-1}. \tag{1.31}$$

In (1.30) $q_{11,t}$ is the first element of the matrix Q_t as defined in (1.31). Moreover, α and β are non-negative scalar parameters such that $\alpha + \beta < 1$, Q is the $k \times k$ unconditional correlation matrix of the standardized errors $u_{i,t}$, obtained from $r_{i,t}/h_{i,t}$. Moreover, $\mathbf{u}_t$ is the vectorization of $u_{i,t}$, with $i = 1, \cdots, k$. Unfortunately, the procedure above ensures positive definiteness but does not generally produce valid correlation matrices. They are obtained by rescaling Q_t such that $R_t = (I \odot Q_t)^{-1/2} Q_t (I \odot Q_t)^{-1/2}$. Contrary to the DCC_T model, that calculates the conditional correlations as a weighted sum of past ones, the DCC_E model formulates R_t through Q_t that is written like a GARCH specification. In both

the DCC models presented above α and β in (1.29) and (1.31) are scalars, so that all the conditional correlations have the same dynamics. Unfortunately, this is a necessary condition to have R_t positive defined $\forall t$. To overcome this limitation, various generalizations of the DCC-GARCH model have been proposed. Billio et al. (2006) suggest a model imposing a BEKK structure on the conditional correlations. In their Quadratic Flexible DCC (GFDCC) GARCH model the matrix Q_t is defined as:

$$Q_t = C^{'} Q C + A^{'} \mathbf{u}_{t-1} \mathbf{u}_{t-1}^{'} A + B^{'} Q_{t-1} B,$$

where the matrices A, B, and C are symmetric, and Q is the unconditional covariance matrix of the standardized errors $\mathbf{u}_t$. To obtain stationarity, $C^{'} Q C$ has to be positive definite and the eigenvalues of $A + B$ has to be less than one in modulus. The number of parameters governing the correlations in the GFDCC-GARCH model in its fully general form is $3k(k+1)/2$ which is infeasible in large systems.

1.6 Inference Procedures

Maximum Likelihood methods

The estimation of the unknown parameters collected in the vector θ for the GARCH models is based on the maximization of a likelihood function constructed under two assumptions:

- the distribution of z_t in (1.6) or of $\mathbf{z}_t$ in (1.16) is assumed to be known;
- the maximum likelihood (ML) function must be at least three time differentiable, with continue partial derivatives in Θ. Such property is defined as a "regularity condition".

The primary appeal of the maximum likelihood technique stems from the well-known optimality conditions of the resulting estimators under ideal conditions (see Casella and Berger (1990) for details).

Prediction Error Decomposition

At it is pretty obvious, the time series exhibit serial dependence. The method of the maximum likelihood requires independent data such that without any modification, it is impossible to use the standard maximization of the likelihood. The so-called *prediction error decomposition* helps to overcome this problem.

Given two random variables x and y, the joint density function can be written as:

$$f(x,y) = f(y|x)f(x) \tag{1.32}$$

Thus, the joint density function can be easily expressed by a product of the conditional density function of y given x and the marginal density function of x. Using three random variables x, y and z, the joint density function can be factorized as:

$$f(x,y,z) = f(z|x,y)f(x,y) \tag{1.33}$$

Inserting (1.32) in (1.33) yields:

$$f(x,y,z) = f(z|x,y)f(x,y) = f(z|x,y)f(y|x)f(x) \tag{1.34}$$

Returning to k-dimensional daily returns $\mathbf{r}_t$ with $t = 1, \cdots, T$, the joint density function can be expressed as follows:

$$f(\mathbf{r}_1, \cdots, \mathbf{r}_T) = f(\mathbf{r}_T|\mathbf{r}_1, \cdots, \mathbf{r}_{T-1})f(\mathbf{r}_1, \cdots, \mathbf{r}_{T-1}); \tag{1.35}$$

$$= f(\mathbf{r}_T|I^{T-1})f(\mathbf{r}_1, \cdots, \mathbf{r}_{T-1}). \tag{1.36}$$

From (1.36), considering the marginal density function, we obtain:

$$f(\mathbf{r}_1, \cdots, \mathbf{r}_{T-1}) = f(\mathbf{r}_{T-1}|I^{T-2})f(\mathbf{r}_1, \cdots, \mathbf{r}_{T-2}). \tag{1.37}$$

Now we include (1.37) in (1.36) and so on such that the joint density function will be equal to:

$$f(\mathbf{r}_T|I^{T-1}), \cdots, f(\mathbf{r}_2|I^1)f(\mathbf{r}_1) \approx \prod_{t=2}^{T} f(\mathbf{r}_t|I^{t-1})$$

$$= \sum_{t=2}^{T} \log f(\mathbf{r}_t|I^{t-1}) \tag{1.38}$$

In (1.38) we assume that $f(\mathbf{r}_1)$ is not depending on the parameters vector to estimate. Moreover, the first observation $\mathbf{r}_1$ gives a negligible contribution to the ML function, if T is large. Hence, the maximum log-likelihood estimator is:

$$\ell(\mathbf{r}_t;\theta) = \sum_{t=2}^{T} \log f(\mathbf{r}_t;\theta), \tag{1.39}$$

where $f(\cdot)$ depends on the density that is chosen.

Univ. GARCH estimation

Let us start the presentation of the likelihood functions used to estimate the unknown parameters of the univariate GARCH models from the easiest case, the ARCH(p) model, like that of equation (1.8). T observed daily returns are available. The unknown parameters vector is $\theta = (\alpha_0, \alpha_1, \cdots, \alpha_p)$. Under the normality assumption for z_t, the likelihood function is obtained as:

$$\begin{aligned} f(r_T, r_{T-1}, \cdots, r_1) =& f(r_T|I^{T-1}) f(r_{T-1}|I^{T-2}) \cdots \\ & \cdots f(r_{p+1}|I^p) f(r_p, r_{p-1}, \cdots, r_1) \qquad (1.40) \\ =& \prod_{t=p+1}^{T} (2\pi h_t^2)^{-1/2} exp\left(-\frac{r_t^2}{h_t^2}\right) \times \\ & f(r_p, r_{p-1}, \cdots, r_1). \qquad (1.41) \end{aligned}$$

In (1.40) the prediction error decomposition has been used while in (1.41) h_t^2 stands for the conditional variance of r_t, estimated recursively. If T is large, $f(r_p, r_{p-1}, \cdots, r_1)$ can be dropped such that the likelihood to maximize is:

$$L(\theta|r_T, \cdots, r_{p+1}) = \prod_{t=p+1}^{T} (2\pi h_t^2)^{-1/2} exp\left(-\frac{r_t^2}{h_t^2}\right) \tag{1.42}$$

It is usual to maximize the log likelihood because it is simpler to handle and it does not modify the results of the estimation. Thus, the log

likelihood that is maximized in order to find the estimate of the $p+1$ parameters of an ARCH(p) is:

$$\begin{aligned}\ell_{\mathcal{N}}(\theta) &= \log L(\theta|r_T, \cdots, r_{p+1}) \\ &= \sum_{t=p+1}^{T} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(h_t^2) - \frac{1}{2}\frac{r_t^2}{h_t^2} \right]. \end{aligned} \tag{1.43}$$

Otherwise, one can start from $t = 2$ independently of the number of parameters to estimate. Hereafter this latter starting point will be used.

All the models belonging to the class of univariate GARCH models use the log likelihood of the equation (1.43) in order to estimate the unknown parameters collected in θ, when the distribution of z_t is Normal. The differences among the models are specified in the vector θ. For instance, for a GARCH(p,q) as that of equation (1.10), the unknown parameters are $\theta = (\alpha_0, \alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q)$.

As aforementioned, the Normal distribution of the daily returns is frequently questioned. Under Student's t distribution, the log likelihood is:

$$\begin{aligned}\ell_{ST}(\theta) = T \left\{ \log\Gamma\left(\frac{v+1}{2}\right) - \log\Gamma\left(\frac{v}{2}\right) - \frac{1}{2}[\pi(v-2)] \right\} + \\ - \frac{1}{2}\sum_{t=2}^{T} \left[\log(h_t^2) + (1+v)\log\left(1 + \frac{r_t^2}{v-2}\right) \right], \end{aligned} \tag{1.44}$$

where $\Gamma(\cdot)$ denotes the Gamma function and v the degrees of freedom. Note that the degrees of freedom are jointly estimated and it is imposed that $v > 2$.

Mult. GARCH estimation

When the estimation concerns the multivariate framework, the log likelihood for the model (1.16) under the Multivariate Normal assumption is (up to a constant):

$$\ell_{\mathcal{MVN}}(\theta) = -\frac{1}{2}\sum_{t=2}^{T}\log|H_t| - \frac{1}{2}\sum_{t=2}^{T}\mathbf{r}_t' H_t^{-1}\mathbf{r}_t, \tag{1.45}$$

where $|\cdot|$ indicates the determinant. Note that in (1.45) the $k \times k$ H_t matrix is inverted for each time period, making the estimation procedure quite complicated, even if iterative methods are used and even if T and k are small. If the distribution of $\mathbf{z}_t$ is not Multivariate Normal, as it frequently happens in the empirical world, a consistent estimator of the true parameters may still be achieved by using (1.45), as demonstrated by Bollerslev and Wooldridge (1992), provided that the conditional mean and variance are correctly specified. This approach is called quasi-maximum likelihood method and an its detailed overview is present in Gouriéroux (1997). Otherwise, the multivariate version of the likelihood for the Student's t-distribution can be used:

$$\begin{aligned}\ell_{M-ST}(\theta) = T\left\{\log\Gamma\left(\frac{v+1}{2}\right) - \log\Gamma\left(\frac{v}{2}\right) - \frac{1}{2}[\pi(v-2)]\right\} + \\ - \frac{1}{2}\sum_{t=2}^{T}\left[\log|H_t| + (1+v)\log\left(1 + \frac{(\mathbf{r}_t)'(\mathbf{r}_t)}{v-2}\right)\right].\end{aligned} \tag{1.46}$$

If the asymptotic normality of the maximum likelihood and quasi-maximum likelihood estimators in the univariate framework has been proven under quite mild assumptions (see Ling and McAleer (2003) for details), the asymptotic normality in the multivariate context is still a challenging research field. In fact, it has been proved for some specific MGARCH models, like the BEKK formulation by Comte and Lieberman (2003), the CCC specification in the already cited work of Ling and McAleer (2003) and the VEC model by Hafner and Preminger (2009) or assuming strong assumptions. Recent developments on this research field are provided by Francq and Zakoïan (2011).

Even though the log likelihood in (1.45) can be used to obtain the estimates of the DCC model unknown parameters, it can be show that a two step procedure for that class of models still yields a consistent (but

inefficient) estimator for θ. Due to the nature of the DCC model, the estimation can be split in two parts. For the first part, let θ_1 be the vector in which the unknown parameters for the k univariate GARCH models are collected. The log (or the quasi-log) likelihood is given by the sum of k univariate GARCH models:

$$\ell_1(\theta_1) = -\frac{1}{2}\sum_{t=2}^{T}\sum_{i=1}^{k}\left[\log(h_{i,t}^2) + \frac{r_{i,t}^2}{h_{i,t}^2}\right], \tag{1.47}$$

where $r_{i,t}$ represents the $i-th$ daily return at time t and $h_{i,t}^2$ its conditional variance. For the second part of the procedure, let θ_2 be the vector of the correlation parameters. A consistent but inefficient estimator of θ_2, given $\hat{\theta}_1$, is obtained by maximizing the following log likelihood:

$$\ell_2(\theta_2|\hat{\theta}_1) = -\frac{1}{2}\sum_{t=2}^{T}\left(\log|R_t| + \mathbf{u}_t^{'}R_t^{-1}\mathbf{u}_t\right). \tag{1.48}$$

As argued above, the split procedure works firstly estimating θ_1 by maximizing (1.47) and then, given $\hat{\theta}_1$, maximizing (1.48) in order to find θ_2. Interestingly, the sum of the two log likelihoods in (1.47) and (1.48) plus $\sum \mathbf{u}_t^{'}\mathbf{u}_t$ is equal to the total log likelihood in (1.45). Although the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are inefficient, they can be used as starting point for the total likelihood. This makes the procedure faster and the consequent estimator results to be asymptotically efficient, as showed by Engle and Sheppard (2001).

1.7 Models for the volatility proxy and extensions

In the previous section the univariate and multivariate GARCH models that will be used in the empirical part of this work have been presented. These models estimate ex-ante the expected volatility on the basis of some observed variables, implying the knowledge of the data generating process. Recall that the expression $VAR(r_t|I^{t-1})$ means the

expected conditional variance for the day t is obtained directly observing some variables at time $t-1$. Then, if one assumes that $VAR(r_t|I^{t-1})$ is driven by a linear combination of the past p squared daily returns, then the functional form that is used to model the conditional variance is: $VAR(r_t|I^{t-1}) = \alpha_o + \alpha_1 r_{t-1}^2 + \cdots + \alpha_p r_{t-p}^2$.

Instead, the nonparametric approach aiming to estimate the volatility uses the high-frequency data of day t in order to obtain an ex-post measure of the volatility for that day. In other words, this approach is completely data-driven. In the next section, the volatility estimator named realized volatility is illustrated. Then some models that parametrize the realized volatility will be presented.

1.7.1 The realized volatility

Let the continuous log price process of a generic liquid asset be determined by the following stochastic differential equation

$$dp(t+\tau) = \mu(t+\tau)d\tau + \sigma(t+\tau)dW(t+\tau), \tag{1.49}$$

with $0 \leq \tau \leq 1$ and $t = 1, \cdots, T$. In (1.49), $p(t+\tau)$ is the log price at time $t+\tau$, $\mu(t+\tau)$ is a drift component, $\sigma(t+\tau)$ is the instantaneous or spot volatility and $W(t+\tau)$ is a standard Brownian motion. Moreover, $\sigma(t+\tau)$ and $W(t+\tau)$ are assumed to be orthogonal. As pointed out by Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002), the daily returns r_t, given in this framework by $r_t = p(t) - p(t-1)$, are Normal distributed conditionally to the Information Set generated by the paths of $\mu(t+\tau-1)$ and $\sigma(t+\tau-1)$. Formally, let $\Im_t = \{\mu(t+\tau-1), \sigma(t+\tau-1)\}_{\tau=0}^{\tau=1}$ be the Information Set. Hence:

$$r_t|\Im_t \sim \mathcal{N}\left(\int_0^1 \mu(t+\tau-1)d\tau, \int_0^1 \sigma(t+\tau-1)d\tau\right).$$

We are interested in the *integrated* variance, also defined as the true volatility:

$$IV_t = \int_0^1 \sigma^2(t+\tau-1)d\tau. \tag{1.50}$$

Unfortunately, IV_t is not directly observable, because it is measured in continuous time and prices are only observed at discrete time. Hence, let us introduce the formulation for sampling the prices at discrete time. Suppose that the interval [0,1] of equation (1.50) is partitioned in D intervals. It normally happens that $D_t \neq D_{t+1}$, such that each observation day has a different number of intervals. But for ease of notation, we consider only the case of $D_t = D, \forall t$. The grid of observation times is defined as $\Psi = \{\tau_0, \cdots, \tau_D\}$, with $0 = \tau_0 < \cdots < \tau_D = 1$. The $d-th$ interval has a length of $\zeta_d = \tau_d - \tau_{d-1}$. If $D \to \infty$, than $\zeta_d \to 0$. Different sampling schemes have been proposed in literature in order to find the optimal interval length. In the most common, all the intervals are equally spaced such that $\zeta_{d,D} = 1/D$. For details on other different schemes, see Hautsch (2012), among others.

Once that the prices have filled the grid Ψ, the $d-th$ intraday return for the day t is:

$$r_{t,d} = p_{t,d} - p_{t,d-1}.$$

Taking advantage of the concept of quadratic variation (QV) and of the availability of high-frequency data, the integrated variance can be estimated. With our notation, the interval [0,1] has been divided in D adjacent, small intervals. The QV is defined as the limit of the sums of the squared differences between the prices registered for each interval, when the length of these intervals goes to zero (or, equally, the number of intervals goes to infinity):

$$QV_t = \lim_{D\to\infty} \sum_{d=1}^{D} (p_{t,d} - p_{t,d-1})^2. \tag{1.51}$$

In this framework[9], the quadratic variation is equal to the integrated variance. Equation (1.51) gives the input to compute the so-called realized volatility (RV_t), obtained by summing the squared intraday returns at

[9]There are no micro-structure noise and jumps.

higher frequencies. The realized volatility is a consistent estimator of IV_t, as showed by Andersen et al. (2003). Formally:

$$RV_t = \sum_{d=1}^{D} (r_{t,d})^2, \tag{1.52}$$

The asymptotic distribution of the realized volatility, as evidenced by Barndorff-Nielsen and Shephard (2002), is:

$$D^{1/2}(2IQ_t)^{-1/2}(RV_t - IV_t) \xrightarrow{d} \mathcal{N}(0,1),$$

where IQ_t is the integrated quarticity, defined as:

$$IQ_t = \int_0^1 \sigma^4(t+\tau-1)d\tau.$$

Micro-structure noise

It is well known that there are important implications in choosing a small or a high frequency at which the prices are sampled. From one hand, if the frequency D represents a day (so we have a very small frequency) and if the price at the end is the same of the price at the beginning of the trading day, then the corresponding realized volatility would be zero, even though the prices had experienced huge variations during the time interval. More importantly, equation (1.51) is no longer valid. Recall that the realized volatility is a consistent estimator of IV_t if the number of intervals D increases. From the other hand, if the frequency is very high, some problems may occur. In fact, it has been documented that increasing to infinity the sampling frequency does not take to the real volatility but to a noise estimation, due to presence of micro-structure noise (Bandi and Russell (2004)). What are these market micro-structure frictions? These frictions are some complications that make the observed price noisy, i.e. with an error term. One example is the discreteness of prices: we are only able to observe the prices at discrete time and thus the

prices we observe are not efficient, in the sense that they do not immediately incorporate the notices as these are diffused. Still, another example is the bid-ask spread, that is the inevitable gap between buying and selling prices. For these reasons, the observed price does not represent the actual price. Instead, the observed (log) price p_t is given by the actual (and unobserved) price p_t^* plus an error term. For the $d-th$ interval:

$$p_{t,d} = p_{t,d}^* + e_{t,d}, \tag{1.53}$$

where e_t is the so-called micro-structure noise. Then, the observed intraday return is:

$$\begin{aligned} r_{t,d} &= p_{t,d} - p_{t,d-1} \\ &= p_{t,d}^* + e_{t,d} - p_{t,d-1}^* + e_{t,d-1} \\ &= r_{t,d}^* + n_{t,d}, \end{aligned} \tag{1.54}$$

where $r_{t,d}^* = p_{t,d}^* - p_{t,d-1}^*$ and $n_{t,d} = e_{t,d} - e_{t,d-1}$. Equation (1.54) implies that RV_t is a biased estimator of IV_t:

$$\begin{aligned} RV_t &= \sum_{d=1}^{D} (r_{t,d}^* + n_{t,d})^2 \\ &= \sum_{d=1}^{D} (r_{t,d}^*)^2 + 2\sum_{d=1}^{D} (r_{t,d}^*)(n_{t,d}) + \sum_{d=1}^{D} (n_{t,d})^2. \end{aligned} \tag{1.55}$$

The expectation of RV_t conditionally on the true returns r^* is:

$$\mathbb{E}(RV_t | r^*) = RV_t^* + 2D\mathbb{E}(e_{t,d}^2) \tag{1.56}$$

Following Bandi and Russell (2004), we assume that e_t is a zero mean, $i.i.d.$ random variable, independent of the actual price process p_t^*. Hence, if the sampling frequency D is very high, then the observed realized volatility based on observed daily returns of equation (1.54) is inconsistent and it diverges from IV_t as much as $D \to \infty$. In fact, as showed

by Zhang et al. (2005), increasing the sampling frequency leads to a bias of discretization, such that:

$$RV_t \overset{d}{\approx} IV_t + \underbrace{2D\mathbb{E}(e_{t,d}^2)}_{\text{bias due to noise}} + \underbrace{\left[\underbrace{4D\mathbb{E}(e_{t,d}^4)}_{\text{due to noise}} + \underbrace{\frac{2}{D}\int_0^1 \sigma_t^4 dt}_{\text{due to discretization}}\right]^{1/2}}_{\text{total variance}} N(0,1) \tag{1.57}$$

In (1.57), if the number of intervals D increases, the bias due to noise increases and from the other hand the bias due to the dicretization decreases. In this case, the total variance in (1.57) will decrease, given that the intraday returns, based on more and more infinitesimal intervals, show smaller variations. Instead, if the number of intervals decreases, the bias due to the noise decreases but the bias due to the discretization increases. Now, the total variance increases. What has been cited is generally called the bias-variance trade-off such that the choosing of the sampling frequency is crucial when the high-frequency data are used.

Instead of using a grid Ψ based on one observation for each transaction, one could sample prices at lower frequencies, based on a new grid Ψ^{sm} such that $\Psi^{(sm)} \subset \Psi$. The maximum number of intervals associated to $\Psi^{(sm)}$ is $D^{(sm)} < D$. The realized volatility estimator is now:

$$RV_t^{(sm)} = \sum_{d=1}^{D^{(sm)}} r_{t,d}^2.$$

Unfortunately, $RV_t^{(sm)}$ is still a biased estimator of IV_t, as evidenced by Aït-Sahalia et al. (2005), among others. Equation (1.57) changes now in:

$$RV_t^{(sm)} \overset{d}{\approx} IV_t + \underbrace{2D^{(sm)}\mathbb{E}(e_{t,d}^2)}_{\text{bias due to noise}} + \underbrace{\left[\underbrace{4D^{(sm)}\mathbb{E}(e_{t,d}^4)}_{\text{due to noise}} + \underbrace{\frac{1}{D^{(sm)}}\int_0^1 \sigma_t^4 dt}_{\text{due to discretization}}\right]^{1/2} N(0,1)}_{\text{total variance}} \quad (1.58)$$

It can be easily seen that if the bias is reduced, the total variance is increased. In order to take into account the problem of the micro-structure noise, several methods have been proposed (for instance, Zhang et al. (2005), Bandi and Russell (2006)), many of which are based on the optimal sampling. The presentation of these different methods is beyond the goal of this work, such that we focus on a very concise discussion about the most used frequencies.

In the applied works, the most popular frequency is the 5 minutes sample, chosen by Liu (2009) for an evaluation of a portfolio composed of the 30 Dow Jones index constituents. Other works instead provide evidence of optimal samplings ranging between 30 and 65 minutes as in the paper of de Pooter et al. (2008).

Realized Covariance

The extension to the multivariate is the following. For the day t, the log prices of k assets, denoted by $\mathbf{p}_t$, are assumed to be driven by the following continuous time diffusion process:

$$d\mathbf{p}(t+\tau) = \boldsymbol{\mu}(t+\tau)d\tau + \Sigma(t+\tau)d\mathbf{W}(t+\tau), \quad (1.59)$$

with again $0 \leq \tau \leq 1$ and $t = 1, \cdots, T$. In (1.59), $\boldsymbol{\mu}(t+\tau)$ is the multivariate drift component at time $t+\tau$, $\Sigma(t+\tau)$ is the spot co-volatility and $\boldsymbol{W}_t$ is the $k-$dimensional Brownian motion. As in (1.49), also here

$\Sigma(t+\tau)$ is assumed to be orthogonal to $\mathbf{W}(t+\tau)$. If in the univariate case, the quantity of interest was the *integrated* variance (equation (1.50)), in the multivariate case the quantity of interest is the *integrated* covariance, denoted by ICV_t:

$$ICV_t = \int_0^1 \Sigma(t+\tau)\Sigma(t+\tau)^{'} d\tau \tag{1.60}$$

As seen in the univariate case, the quadratic variation of $\mathbf{p}_t$ over the same interval equals the *integrated* covariance such that a consistent estimator of (1.60) is given by:

$$\begin{aligned} RCV_t &= \sum_{d=1}^{D} \{\mathbf{p}_{t,d} - \mathbf{p}_{t,d-1}\}\{\mathbf{p}_{t,d} - \mathbf{p}_{t,d-1}\}^{'} \\ &= \sum_{d=1}^{D} \boldsymbol{r}_{t,d}\boldsymbol{r}_{t,d}^{'}. \end{aligned} \tag{1.61}$$

Thus, the more the intervals are shrunk, such that $D \to \infty$, the more precise the realized covariance is. Formally:

$$\text{if } D \to \infty \text{ then } RCV_t \xrightarrow{p} ICV_t. \tag{1.62}$$

Two kinds of problems affect the realized covariance estimator of equation (1.61). First, if the intervals are more and more small, it becomes practically impossible to have the data synchronized. Asynchronous data cause distortion such that what stated in (1.62) is no longer valid. Second, also in the multivariate case the micro-structure noise is a pressing problem that leads the researcher to pay attention to the sampling frequency. The solution to these problems represents a challenging and new research field. To the best of our knowledge, there are not so many estimators robust to micro-structure noises and asynchronous data. One is the estimator recently proposed by Aït-Sahalia et al. (2010).

The realized covariance as expressed by equation (1.61) represents the benchmark employed to compare different estimates produced by a

set of volatility models. Moreover, the realized covariance can be used in some parametric models that explicitly use a functional form to forecast the future values of the conditional covariance matrix, within the general framework of equation (1.16). These models will be presented in the next two paragraphs.

1.7.2 The Rolling Covariance model

The first parametric model that employs the past realized covariance realizations to forecast the conditional covariance matrix has been proposed by Fleming et al. (2003). In particular, the estimation of H_t is carried out through a weighting scheme of the past realized covariances and past conditional covariances. Formally:

$$H_t = exp(-\alpha)H_{t-1} + \alpha exp(-\alpha)RCV_{t-1}, \tag{1.63}$$

where α, the optimal decay rate, is the only unknown parameter to estimate by maximizing the log likelihood in (1.45). The rolling scheme of (1.63) has some advantages. First of all, it is a very parsimonious model. Second, the exponentially weighted estimators, like that of the rolling covariance, generally have the smallest asymptotic mean squared error, as pointed out by Foster and Nelson (1994). Third, the rolling scheme assures the positive definiteness of H_t provided that $\alpha > 0$. The estimator of H_t from this model should be, theoretically, more efficient than that of an estimator resulting from models based on the cross products of the daily returns. In fact, as evidenced by Andersen and Bollerslev (1998), the variance of the cross-products[10] of the daily returns is larger than the variance of the cumulative squared intraday returns, as long as D increases.

[10]The cross-products are largely present in the MGARCH models. For instance, they are present in equation (1.22) for the estimation of the BEKK model.

1.7.3 The CAW model

The Conditional Autoregressive Wishart (CAW) model, recently proposed by Golosnoy et al. (2012), assumes that the realized covariance RCV_t follows a central Wishart distribution, given the past recorded in the information set I^{t-1}:

$$RCV_t|I^{t-1} \sim \mathcal{W}_k(\nu, S_t/\nu), \tag{1.64}$$

where $\nu > k-1$ is the scalar degree of freedom and S_t/ν is the $k \times k$ symmetric positive definite matrix. The idea to model the realized covariance following a Wishart process is due to Gouriéroux et al. (2009). They proposed a Wishart autoregressive process of order p (WAR(p)) as the updating structure of the realized covariance, but unfortunately the number of parameter to estimate is too high even if the number of stocks is low: for $p = 1$, the unknown parameters are $3k^2/2 + k/2 + 1$.

Instead, the CAW model directly relates the conditional covariance matrix to the realized covariance, letting the number of parameters considerably decrease. In fact, following Anderson (1984), it can be shown that $\mathbb{E}(RCV_t|I^{t-1}) = H_t$, the conditional covariance matrix. Hence, different specifications of the CAW model are present in literature. For instance, Golosnoy et al. (2012) propose some specifications based on the BEKK updating structure. The first is the scalar CAW with covariance targeting, that is:

$$H_t = (1 - a^2 - b^2) * \overline{RCV} + a^2 * RCV_{t-1} + b^2 H_{t-1}, \tag{1.65}$$

where a and b are scalars to estimate and $\overline{RCV}$ is the sample realized covariance.

A second specification of the CAW model is the diagonal CAW with covariance targeting, that is:

$$H_t = \overline{RCV} - A\overline{RCV}A^{'} - B\overline{RCV}B^{'} + ARCV_tA^{'} + BH_{t-1}B^{'}, \tag{1.66}$$

where A and B are two diagonal matrices to estimate. The estimation of (1.65) and (1.66) is done by maximizing the following log-likelihood, provided by Golosnoy et al. (2012):

$$\ell(\theta_w) = \sum_{t=2}^{T} \left\{ -\frac{\nu n}{2}\log(2) - \frac{k(k-1)}{4}\log(\pi) - \sum_{i=1}^{k} \log\Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2}\log|\frac{RCV_t}{\nu}| + \frac{\nu-k-1}{2}\log|RCV_t| - \frac{1}{2}tr(\nu H_t^{-1} RCV_t) \right\}, \tag{1.67}$$

where $w = 1, 2$ is the suffix indicating the CAW models presented above. Hence, the unknown parameters are: $\theta_1 = (\nu, a, b)$ for the scalar CAW and $\theta_2 = (\nu, A_{11}, \cdots, A_{kk}, B_{11}, \cdots, B_{kk})$ for the diagonal CAW. The scalar and diagonal CAW can be used to estimate $\hat{h}_m$ in (1). They have the advantages to directly incorporate the realized covariance in their updating structure. Both the models will be used in the empirical part of this work, when different volatility models will be compared from an economic and statistical point of view. The scalar and diagonal CAW are only two of possible specifications available in literature. For instance, another formulation that can be used within the Wishart distribution framework is DCC-type specification proposed by Bauwens et al. (2012).

2 The Forecasting Problem

2.1 Introduction

At the beginning of this work we presented the volatility evaluation problem as a problem involving many aspects. Among all these aspects the forecasting issue assumes an important rule. One question regards how the volatility predictions are obtained. Another question is the evaluation of the forecasts produced by a set of models. Related to this latter question there is the problem of the forecasts evaluation when the actual value of the variable of interest is never observed. For instance, this happens when we aim at evaluating the volatility forecasts. All these issues will be discussed in this chapter, organized as follows. In Section 2.2 the formalization of the forecasting problem is given. Due to the usually very large amount of data available to the forecaster, how can we deal with it? In order to predict the variable of interest, we could use all the data from the beginning to the last in-sample observation or we could use only the last recent observations. The problem concerns the forecasting schemes, illustrated in Section 2.3. In Section 2.4 different methods for the evaluation of the forecasts are presented: the Mincer-Zarnowitz regression, the robust statistical loss functions, the economic loss functions, the pair-

wise and multiple predictive ability tests. The combination of forecasts is finally discussed in Section 2.5.

2.2 A formal statement of the problem

The *forecasting* can be defined as the process that aims to provide informations on future values of one (or more) variable of interest. Let x be the value to forecast and $\hat{x}$ the forecast, depending on all the relevant informations I available at the time of forecast and on the function f used to manipulate these informations for obtaining $\hat{x} = f(I)$. The process of forecasting relies on a set of elements, that are: $\{x, \hat{x}, f(I)\}$. It often happens that there are too much informations collected in I such that it is difficult to understand which are relevant and which not. A second important element of the forecasting process is played by the loss function. Because it is almost impossible that $\hat{x} = x$, the loss function is needed to check how much each forecast error costs. The loss functions will be described in the next paragraph. The function $f(I)$ has been illustrated in the previous chapter with reference to the volatility forecasting. The just illustrated framework considers the point estimate, meaning that x is a single number. Another possible forecasting types are the interval forecasts or forecasts of the conditional distribution of the variable of interests. We focused only on the point estimate, while the latter two are discussed in Chatfield (1993) and Tay and Wallis (2002), respectively. In what follows, only the evaluation of the point forecasts will be discussed.

Loss Function

A loss function $\mathcal{L}$ maps the distance between the actual and forecasted value. Formally:

$$\mathcal{L}(x, \hat{x}). \tag{2.1}$$

Let $\hat{x}_i$ be the forecast of the variable of interest produced by the model i through the function f_i such that $\hat{x}_i = f_i(I)$. We assume that there are

m models, such that $i = 1, \cdots, m$. Let $\dot{X}$ be the set of all the estimates, with $\hat{x}_i \in \dot{X}$. The set of estimates are ranked by the loss function: the smaller the loss function is, the better that estimate is. If a loss function has the following three features, it is said to be *well defined*:

1. $\mathcal{L}(\cdot, \cdot)$ is continuous in $\dot{X}$ and it is minimized at $\hat{x}^*$ which represents the optimal forecast.

2. $\mathcal{L}(\cdot, \cdot)$ is such that the optimal forecast $\hat{x}^*$ equals the true value, formally:
$$\hat{x}^* = \underset{\hat{x} \in \dot{X}}{\operatorname{argmin}} \mathcal{L}(x, \hat{x}) \Longleftrightarrow \hat{x}^* = x. \tag{2.2}$$

3. The loss function gives zero loss then $\hat{x}^* = x$.

In addition to the previous properties, a loss function yields an increasing penalty if the distance between x and $\hat{x}$ increases. A loss function is said to be symmetric if:

$$\mathcal{L}(x - \hat{x}) = \mathcal{L}(x + \hat{x}) \tag{2.3}$$

It is common to use symmetric loss functions, even though sometimes the forecaster needs to consider the asymmetric loss (see Patton and Timmermann (2007), for instance). The most common symmetric loss function is the Mean Squared Error (MSE), based on the squared of the forecast error:

$$\mathcal{L}(x, \hat{x}) = \left[(x - \hat{x})^2\right]. \tag{2.4}$$

Another common loss function is the Mean Absolute Error (MAE):

$$\mathcal{L}(x, \hat{x}) = |x - \hat{x}|. \tag{2.5}$$

MSE and MAE belong to the family of loss functions considered by Elliott et al. (2003):

$$\mathcal{L}(x, \hat{x}; p, \alpha) = \left[\alpha + (1 - 2\alpha) I_{(x - \hat{x} < 0)}\right] \cdot |x - \hat{x}|^p, \tag{2.6}$$

where $I_{(x-\hat{x}<0)} = 1$ if its argument is true. The class of loss functions in Equation (2.6) gives the lin-lin (piece-wise linear) loss function when $p = 1$ and the asymmetric quadratic loss function when $p = 2$. The MAE belongs to the lin-lin class when $\alpha = 2$ while the MSE can be obtained setting $p = 2$ and $\alpha = 0.5$.

Another set of loss function is studied by Varian (1975). He proposed the lines loss:

$$\mathcal{L}(x, \hat{x}; b) = exp\left[b(x - \hat{x})\right] - b(x - \hat{x}) - 1, \tag{2.7}$$

where b is the parameter that controls the asymmetry. If $b < 0$, the overpredictions ($\hat{x} > x$) costs more than equally large underpredictions and vice versa.

2.3 Forecasting schemes

As mentioned above, all the relevant informations needed to forecast x are collected in the information set I. The information set could be enormously large such that it is difficult to recognize the variables that are really able to make a forecast for x. For instance, it could happen that most of the variables are highly correlated between themselves such that some variables bring no additional information for forecasting x. Moreover, if the number of variables increases, then the bias of the forecast decreases but at a cost of an increased forecast variance. When the variable of interest is the volatility, the problem of variable selection is reduced because generally in this framework the forecasts are only based on the observed daily returns. Instead, the variable selection for other variables of interest still represents a challenging research field. Given that this work focuses on the volatility evaluation, we will not consider more the variable selection problem, assuming that only the relevant variables are collected in the information set. Because we act in a time series framework, the information set changes with the time. As seen in the previous

chapter, let I^t be the information set observed at time t to predict the variable of interest at time $t+1$, with $t = \{1, \cdots, T\}$. The variable of interest now is indexed by x_{t+1} and its forecast by $\hat{x}_{t+1}$. There are three forecasting schemes normally used that update the information set: the recursive, rolling and fixed schemes. In order to explain these schemes, let us consider the case of the h periods ahead forecast, till the period $T+h$.

In the recursive scheme, the forecaster uses all the information available at time T to forecast x_{T+1}. Once the forecast $\hat{x}_{T+1}$ has been carried out, the forecast for the period $T+2$ uses the increased information set from 1 to $T+1$. Practically, if x_{T+1} becomes observed, then the information set includes it. The procedure is repeated til $T+h$, such that the information set expands as the number of forecasting periods increase.

The rolling scheme uses a fixed length information set that considers only the last w observations. Practically, $\hat{x}_{T+1}$ uses an information set starting on $T-w+1$ and ending on T. Once $\hat{x}_{T+1}$ is obtained, then the rolling window considers the period from $T-w+2$ to $T+1$. At $t = T+1$, x_{T+1} is observed and it enters the Information Set. The procedures ends at time $T+h-1$, when the last forecast for the period $T+h$ is computed. The rolling scheme has the advantage of using only the last w informations, that are more likely to influence the forecast than the oldest informations. As drawback, this method requires to decide the parameter w.

The fixed scheme considers exclusively the information from 1 to T, neither updating the Information Set nor using a rolling window. Thus, even the forecast for h periods ahead is based on the information available at time T.

Whenever the forecasting scheme is, when the forecast is made for a period $T+h$, this forecast is said *out-of-sample* or *pure out-of-sample*. Suppose that the variable of interest is observed till T. Before forecasting out of the sample, it is very common to split the entire dataset in two

parts: $t_1 = \{1, \cdots, L\}$ and $t_2 = \{L+1, \cdots, T\}$, with $L < T$. Let T_1 and T_2 be the length of the in-sample and out-of-sample periods, respectively, such that $T_1 + T_2 = T$. The forecasts made on the basis of the information set t_1 for the period t_2 are called *pseudo out-of sample*, because the variable of interest is indeed observed. This procedure allows to compare different forecasting models before going completely out of the sample. However, the problem of choosing the sample split arises. As argued by Hansen and Timmermann (2012), some predictive ability tests (that will be discussed afterwards) critically depend on when the split is made. And more importantly, there are not universally accepted guidelines in order to decide how and when split the sample. Ideally, a superior predictive model as well as the predictive test should be robust to the different split points. A robust predictive test is provided in the cited work of Hansen and Timmermann (2012). A discussion about when split the sample can be found in Clark and McCracken (2011). With reference to the split choice, there is a clear trade-off between large T_1 and small T_2 and vice versa. In the former case, the forecasts will be more accurate. Instead, if T_2 is bigger than T_1, the accuracy of the forecasts evaluation will be more precise.

Returning to the three forecasting schemes, we graphically show how each works by means of a simulation of a GARCH(1,2) process. In particular, the data generating process is such that:

$$\begin{aligned} r_t \sim & \mathcal{N}(0, h_t^2), \\ h_t^2 = & 0.01 + 0.05 r_{t-1}^2 + 0.70 h_{t-1}^2 + 0.20 h_{t-2}^2. \end{aligned}$$

We set $T = 1000$. Then, four different GARCH models are used in order to estimate the variable of interest: the conditional variance. The models are: a GARCH(1,1), an ARCH(4), a GJR(1,1) and a EGARCH(1,1). The models employ the in-sample period of length $T_1 = 900$ to estimate the unkwown parameters obtaining the estimate of $\hat{h}_{t+h}^2$. The pseudo out-of-sample period consists of 100 days. There is a daily re-fitting estimation:

once $\hat{h}^2_{t+h}$ is obtained, r_{t+h} becomes observed, with $h = 1, \cdots, 100$. We use all the previously cited forecasting schemes. The window's width for the rolling scheme is 200 and 400. In Figure 2.1 the different conditional variances estimated by the GARCH(1,1) model and the different schemes are plotted together with the true conditional variance. All the predictions, independently of the forecasting schemes, follow the pattern of the true variance. However, the rolling schemes, mainly when the width of the window is small, gives the less accurate estimates. The volatility estimates of the ARCH(4) models are illustrated in Figure 2.2. Being this model largely misspecified, the pattern of the estimated variances is quite far from the true variance. Moreover, with reference to the rolling scheme, using a smaller or bigger width does not lead a considerable difference. In Figures 2.3 and 2.4 are displayed the variance predictions of the GJR and EGARCH models. Both these specifications are misspecified but surprisingly, the latter model works quite well, showing an small distance from the true variance even when the rolling scheme with the smallest width is used.

2.4 Forecasts evaluation

This work focuses on the evaluation of the volatility predictions obtained by a set of models. Given that the first input of the loss function (2.1) is never observed, this framework requires specific instruments to evaluate the volatility predictions. In other words, if we aim to evaluate the volatility predictions, we have to take care of the latent nature of x in equation (2.1). If in Economics the forecasted value, after a certain amount of time, becomes directly observed, this does not happen for the volatility. In this case, the volatility evaluation would be impossible by means of the classical loss function approach as presented above. But, thanks to some recent developments, the loss functions can be used. The first development is the using of a consistent estimator of the true volatil-

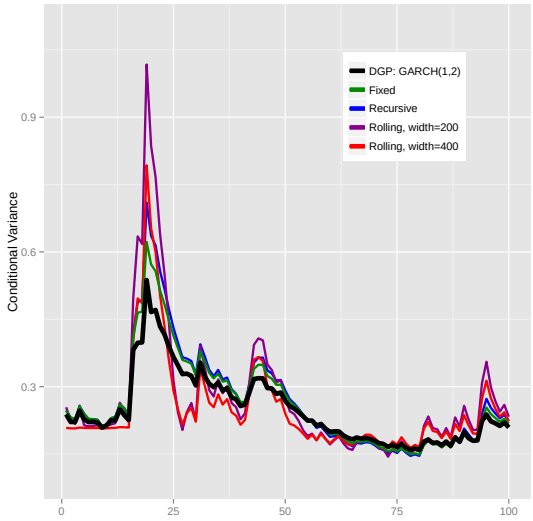


Figure 2.1 – Forecasting schemes comparison. GARCH(1,2) vs GARCH(1,1)

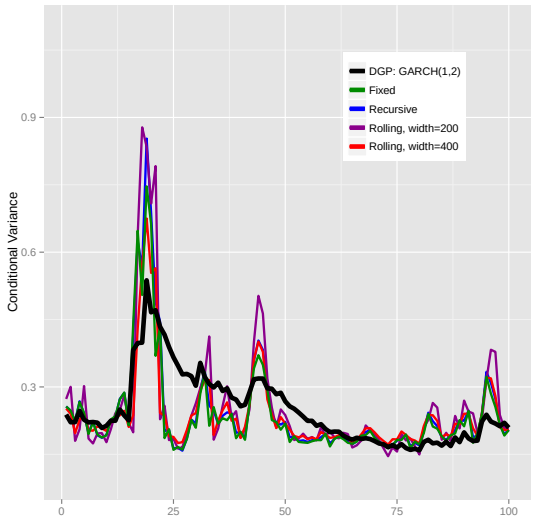


Figure 2.2 – Forecasting schemes comparison. GARCH(1,2) vs ARCH(4)

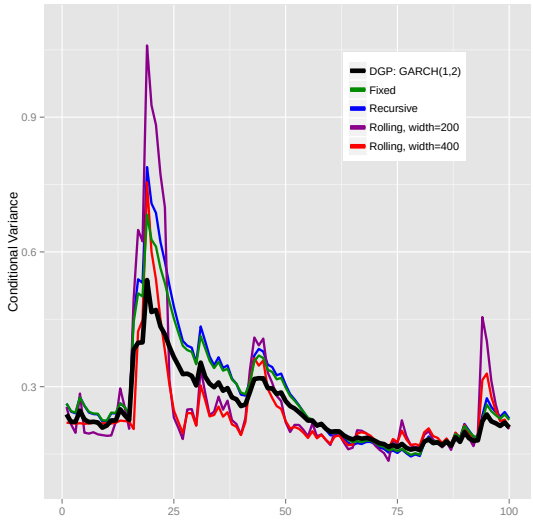


Figure 2.3 – Forecasting schemes comparison. GARCH(1,2) vs GJR(1,1)

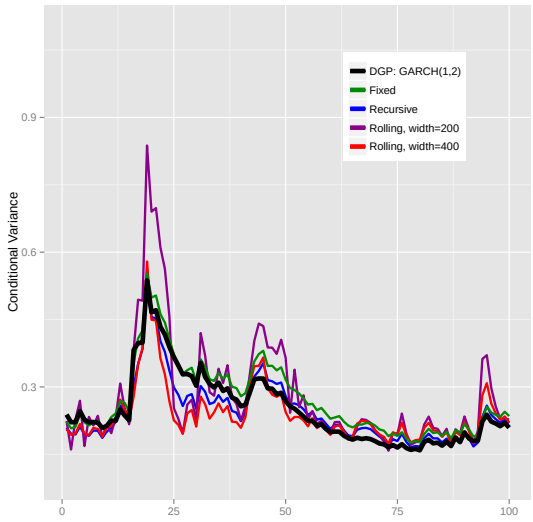


Figure 2.4 – Forecasting schemes comparison. GARCH(1,2) vs EGARCH(1,1)

ity. And, as pointed out in the previous chapter, a consistent estimator of the unobserved volatility is represented by the (square root of the) realized volatility. A second development has been provided by Hansen and Lunde (2006) and Patton (2006): the consistent or robust loss functions. Naturally, one could be interested in evaluating the forecasts regarding some observable economic variables, but this is beyond the aim of this work. In what follows, we illustrate some methods generally used for the evaluation of volatility forecasts and some methods generally used for the standard forecasts evaluation. The Mincer-Zarnowitz regression, the robust (statistical) loss functions and the economic loss functions belong to the former methods, the pairwise and multiple predictive ability tests to the latter ones.

2.4.1 The Mincer-Zarnowitz Regression

Suppose the entire sample has been split in two parts, as described above. The regression proposed by Mincer and Zarnowitz (1969) considers the pseudo out-of-sample forecasts for verifying if the forecast errors for h-ahead predictions are mean zero and are uncorrelated with I^t. The regression is:

$$x_{t+h} = \beta_0 + \beta_1 f(I^t) + u_{t+h}, \quad \text{with } t = L+1, \cdots, T-h, \tag{2.8}$$

where u_{t+h} is the error term. For simplicity, let us suppose that $h = 1$. The regression (2.8) means that the predictions obtained using the information available at time t and a forecasting model $f(\cdot)$ should be as much as possible close to the true value x_{t+1}. If this happens, the forecast error $\hat{u}_{t+1} = x_{t+1} - \hat{\beta}_0 - \hat{\beta}_1 f(I^t)$ is such that $\mathbb{E}(\hat{u}_{t+1}|I^t) = 0$. This is controlled by testing jointly $\beta_0 = 0$ and $\beta_1 = 1$. If the test is rejected, it means that the information I^t has not been adequately exploited: other better forecasts are available given the information set.

With reference to the volatility evaluation, (2.8) changes in:

$$RV_t = \beta_0 + \beta_1 \hat{h}^2_{t,m} + u_{t+1}, \quad \text{with } t = L+1, \cdots, T-1 \tag{2.9}$$

where RV_t is the realized volatility as defined in the previous chapter and $\hat{h}^2_{t,m}$ if the conditional variance forecasted by a model m. Andersen and Bollerslev (1998) used (2.9) to demonstrate the goodness of the forecasting ability of the standard GARCH models. To be precise, they used (2.9) considering the left hand side with the squared daily returns r_t^2. The resulting R^2s of such regression ranged between 0.026 and 0.047, with $\hat{h}^2_{t,m}$ obtained from a GARCH(1,1) specification. These disappointing results were coherent with the literature[1]. However, as argued by the same authors, the low R^2s derive from the noises of r_t^2. More specifically, the population R^2 is defined as $R_p^2 = VAR(\hat{h}_t^2)/VAR(r_t^2)$. It can be show that this quantity exists if $\kappa\alpha_1^2 + \beta_2^2 + 2\alpha_1\beta_1 < 1$, where α_1 and β_1 are the coefficient of the GARCH(1,1) and κ is the kurtosis of the innovations. Using the standard framework of equation (1.6), $\kappa = \mathbb{E}(z_t^4)$. If the population R^2 exists, Andersen and Bollerslev (1998) found the following inequality:

$$R_p^2 = \frac{\alpha_1^2}{1 - \beta_1^2 - 2\alpha_1\beta_1} < \frac{1}{\kappa}. \tag{2.10}$$

Recalling that κ identifies the tickness of the distribution tails, if the innovations are normally distributed, then $\kappa = 3$ such that the population R^2 is at most equal to 1/3. If the data exhibit fat tails, R_p^2 is even smaller. From this two important consequences follow. Firstly, the low R^2s of the regression (2.9) when on the left hand side there are the squared daily returns do not say that the forecast performance of the GARCH models is low. Secondly, and more importantly, if one aims to use the Mincer-Zarnowitz regression, then it would be appropriate to use a good proxy of the latent volatility. As argued by Andersen and Bollerslev (1998), r_t^2 are a noisy estimator of the true volatility. This comes from the already seen equation (1.6): $r_t = h_t z_t$. If the interest is on h_t^2 and one uses r_t^2 as its proxy, then $r_t^2 = h_t^2 z_t^2$ such that r_t^2 is far from being a good approximation of h_t^2, due to z_t^2.

[1] Similar results are provided in West and Cho (1995) and Jorion (1995), among others.

To conclude, the Mincer-Zarnowitz regression is a valid method to evaluate the pseudo out-of-sample forecasts of a set of volatility models when on the left hand side a good proxy of the true volatility is provided. Otherwise, all the models will show low forecasting abilities.

2.4.2 Robust statistical loss functions

As mentioned above, the concept of robustness of the loss functions has been provided by Hansen and Lunde (2006) and Patton (2006). These authors gave the conditions under which an univariate loss function is said to be consistent (or alternative robust, in the sense that the loss function is robust to the noise present in the volatility proxy). The consistency of the loss function allows to rank any volatility forecasts even though the true volatility is not observed and the volatility proxy replacing the true volatility is observed with some noise. The definition of consistency is as follows.

Definition 2.1. A well defined loss function, $\mathcal{L}$, is said to be "robust" if the ranking of any two (possibly imperfect) volatility forecasts, $h_{m,t}$ and $h_{l,t}$, with $m \neq l$, by expected loss is the same if the ranking is done using the true conditional variance, σ_t^2 , or its proxy, $\hat{\sigma}_t^2$. That is:

$$\begin{aligned} &\mathbb{E}(\mathcal{L}(\sigma_t^2, h_{l,t})) \geq \mathbb{E}(\mathcal{L}(\sigma_t^2, h_{m,t})) \Longleftrightarrow \\ &\mathbb{E}(\mathcal{L}(\hat{\sigma}_t^2, h_{l,t})) \geq \mathbb{E}(\mathcal{L}(\hat{\sigma}_t^2, h_{m,t})), \end{aligned} \tag{2.11}$$

for any $\hat{\sigma}_t^2$ such that its conditional mean is equal to the unobserved volatility: $\mathbb{E}(\hat{\sigma}_t^2 | I^{t-1}) = \sigma_t^2$. This means that $\hat{\sigma}_t^2$ is a conditionally unbiased volatility proxy of the true but observed volatility σ_t^2.

Definition 2.1 states that if a volatility forecast given by any model l is expected to be worse compared to the true volatility, then the same volatility forecast given by any model l is expected by be worse compared to the volatility estimate $\hat{\sigma}_t^2$. In the aforementioned work of 2006, Patton gives conditions under which a loss function is consistent. Moreover, he

provides some examples of robust loss functions, for which Definition 2.1 holds. A set of robust and nonrobust loss functions is summarized in Table 2.1.

Table 2.1 – Loss function examples

	Formulation	Robustess
MSE	$(\hat{\sigma}^2 - h)^2$	Yes
QLIKE	$\log h + \frac{\hat{\sigma}^2}{h}$	Yes
MSE-LOG	$(\log \hat{\sigma}^2 - \log h)^2$	No
MSE-SD	$(\hat{\sigma} - h^{0.5})^2$	No
MAE	$\lvert\hat{\sigma}^2 - h\rvert$	No

In an multivariate context, the (univariate) loss function presented above becomes:

$$\mathcal{L}(\Sigma_t, H_t), \tag{2.12}$$

where Σ_t is the true but unobservable covariance matrix and H_t is its estimate. The consistency definition for the multivariate context changes as follows.

Definition 2.2. A well defined loss function of type (2.12) is said to be "robust" if

$$\begin{aligned} &\mathbb{E}(\mathcal{L}(\Sigma_t, H_{l,t})) \geq \mathbb{E}(\mathcal{L}(\Sigma_t, H_{m,t})) \iff \\ &\mathbb{E}(\mathcal{L}(\hat{\Sigma}_t, H_{l,t})) \geq \mathbb{E}(\mathcal{L}(\hat{\Sigma}_t, H_{m,t})). \end{aligned} \tag{2.13}$$

The assumptions, provided by Laurent et al. (2013), under which Definition 2.2 holds are:

A1: Σ_t replaced by any conditionally unbiased proxy;

A2: the well defined $\mathcal{L}(\cdot,\cdot)$ is twice continuously differentiable with respect to $\hat{\sigma}_t$ and h_t, indicating the element of matrices $\widehat{\Sigma}_t$ and H_t, respectively;

A3: the second derivative $\frac{\partial^2 \mathcal{L}(\Sigma_t, H_t)}{\partial \sigma_{l,t} \partial \sigma_{m,t}}$ is finite and independent of H_t, $\forall l, m$.

However, not all the most common loss functions are consistent. Laurent et al. (2013) define a family of consistent loss functions based on the (observed) forecast error $\widehat{\Sigma}_t - H_t$, that assumes the following quadratic form:

$$\mathcal{L}(\widehat{\Sigma}_t, H_t) = vech(\widehat{\Sigma}_t - H_t)' \Lambda \, vech(\widehat{\Sigma}_t - H_t), \tag{2.14}$$

where $vech(\cdot)$ is the operator that stacks the lower triangular portion of a matrix into a vector and Λ is a matrix that assigns the weights to each element of forecast error matrix $\widehat{\Sigma}_t - H_t$. Changing opportunely Λ in (2.14), infinite loss functions may be created. However, only if the previous assumptions hold, then the loss function is consistent. In the empirical part of this work, we will consider four specifications of Λ, in order to have two symmetric and two asymmetric loss functions. The specifications are summarized in Table 2.2.

- insert Table 2.2 about here -

The Euclidean distance is the matrix version of the Mean Squared Error distance. The matrix Λ is a diagonal matrix of ones, such that each forecast error term is first squared and then summed, $\forall t$. The squared weighted Euclidean distance considers only the variances: the matrix Λ is a diagonal matrix of zeros and ones, such that only the forecast error for the variances is computed. The *over prediction* version of the Mahalanobis distance penalizes the negative forecast errors, that are present when the forecasted value is larger than the correspondent value of the volatility proxy. If this happens, the diagonal matrix Λ is such that the negative terms of the forecast error $\hat{\Sigma}_t - H_t$ count twice. In conclusion,

the inverse of the Mahalanobis distance just presented is considered. The *under prediction* of the Mahalanobis distance penalizes the cases in which the forecast error is positive, meaning that the forecasted value has been under predicted. If this happens, the diagonal Λ matrix is such that each positive term of the forecast error matrix counts twice, $\forall t$.

2.4.3 Economic loss function

The economic loss function provides an indirect evaluation of the risk, because we do not directly assess the distance between a volatility proxy and the volatility as obtained by a forecasting model. In this work, we present two approaches belonging to this evaluation method: the utility function and the VaR measure methods. Let us start with the first.

Utility function approach

As mentioned above, the conditional covariance matrix does not have an intrinsic value but it represents the key input for pricing options, derivatives, and so forth. The utility function approach considers this idea in order to indirectly compare the volatility predictions of different models. Suppose an estimated conditional covariance $\widehat{H}_{m,t}$ has been obtained (we operate in the multivariate world such that the univariate framework could be considered an its generalization) by a model m, $\forall t$. Suppose also that a realized covariance estimates has been obtained. Let $\widehat{\Sigma}_t$ be the realized covariance estimate. $\widehat{\Sigma}_t$ and $\widehat{H}_{m,t}$ can be separately used to find the time-varying weights maximizing the portfolio returns (or equivalently minimizing the portfolio risk), subject to some constraints (Markowitz (1952)). A typical portfolio optimization problem is:

$$\begin{aligned} \min_{\mathbf{w}_t} \quad & \mathbf{w}_t^{'} \widehat{H}_{m,t} \mathbf{w}_t \\ s.t. \quad & \mathbf{w}_t \boldsymbol{\iota} = 1, \end{aligned} \tag{2.15}$$

where $\mathbf{w}_t$ is a $k \times 1$ vector of weights and $\boldsymbol{\iota}$ is a $k \times 1$ vector of ones. As usual, k is the number of stocks. The constraint $\mathbf{w}_t\boldsymbol{\iota} = 1$ means that no short sellings are allowed. The optimization problem of (2.15) is repeated $\forall t$, giving a set of time-varying weights, denoted by $\hat{\mathbf{w}}_t$. Then, let $r^p_{t,m} = \hat{\mathbf{w}}_t^{'}\mathbf{r}_t$ be the portfolio daily return when the model m is used for calculating the conditional covariance matrix. The utility function approach evaluates the performance of a model looking at the utility produced by that model. For instance, Fleming et al. (2001) propose the following quadratic utility function:

$$U(m,t) = W_0\left((1 + r_f + r^p_{t,m}) - \frac{\gamma}{2(1+\gamma)}(1 + r_f + r^p_{t,m})^2\right), \tag{2.16}$$

where W_0 represents the amount of wealth, r_f the risk-free interest rate and γ the coefficient of risk aversion. One can evaluate any model by simply summing $U(m,t)$ $\forall t$ or determining the quantity Δ that makes indifferent a model m by a model l, with $l \neq m$:

$$\sum_{t=1}^{T} U(m,t) = \sum_{t=1}^{T} U(l,t) - \Delta. \tag{2.17}$$

Obviously, If $\Delta > 0$, then model l is preferred. This method can be used also for the comparison of the utility produced when $\widehat{\Sigma}_t$ and $\widehat{H}_{m,t}$ are plugged into the minimization problem of equation (2.15).

However, the utility function approach exhibits some drawbacks. It depends on many variables, such as the risk-free interest rate and the risk-aversion coefficient. Even though these variables are exogenous, mainly the former is difficult to hypothesize as completely unchanging, during all the sample period. Another problem is the nature of the chosen utility function. How do the results change when the utility function changes?

Value at Risk approach

A popular ex ante risk measure is the Value at Risk (VaR), that is defined as the potential losses that a portfolio exhibits over a defined period

for a given confidence interval. Firstly developed by JPMorgan Chase & Co. in the early 1990s, the VaR has been extensively used during the last two decades. For a comprehensive overview of the Value at Risk and its measures, look at Duffie and Pan (1997) and Jorion (2007). Even though the VaR has been criticized for its statistical properties (see Artzner et al. (1999)), it still has an important role in Basel III (for a definition of the 'stressed' VaR, see the EBA Guidelines on Stressed Value at Risk, 2012). If a violation is defined as the occurrence of the portfolio's loss greater than the VaR, a model that exhibits good risk performances is a model for which the number of these violations is closer to the prefixed number of expected violations, that depend on confidence interval and period. For instance, if 100 trading days and a VaR with a 95% confidence interval are considered, a risk manager expects that the violations occur 5 times. An interesting discussion about the violations of the VaR can be found in Gencay and Selcuk (2004). If the VaR is violated many times, more than the expected ones, there is an underestimation of risk. Instead, if the VaR is violated few times, there is an overestimation of risk. In this latter case, the portfolio holder owns a large amount of capital, supporting a high opportunity cost: he could invest capital in other business, for instance. In this framework, only the case of a risk-averse portfolio holder is considered, such that, for him, the less the VaR is violated, the better it is. Practically, the VaR can be used to indirectly evaluate different models: each model gives a volatility prediction that represents the input for the VaR measures, which are subsequently evaluated in terms of violations. The problem is when and how much the VaR measures are good and hence, consequently, a model will be considered better than another.

The VaR measures can be estimated either parametrically or nonparametrically. The mean-variance approach, that will be presented in a while, belongs to the first approach while the Historical Simulation and the Monte Carlo methods belong to the latter. In the chapter 3 a third method, belonging to the family of nonparametric approach, will be re-

ported. Lima and Neri (2007) review all the VaR calculation methods while in Markovich (2008) the nonparametric methods are discussed.

Let us now focus on the mean-variance approach for the calculation of the VaR measures. This parametric methods requires the knowledge of the standard deviations of the portfolio (or of a generic asset) and of the daily returns distribution. Continuing with the notation used in the previous paragraph, let $h^{2,p} = \mathbf{w}_t^i H_{m,t} \mathbf{w}_t$ be the portfolio variance obtained after some optimization routines. If the univariate approach is considered, then $r_{t,m}^p$ and $h_{t,m}^{2,p}$ collapse to $r_{t,m}$ and $h_{t,m}^2$, respectively. For ease of notation, we suppress the index m indicating the $m-th$ model. Moreover, let $\mathbb{E}(r_t^p) = 0$. Formally, the Value at Risk for the day t, denoted by VaR_t, for long trading position is obtained as:

$$VaR_t = r_t^p + f_\alpha h_t^p, \tag{2.18}$$

where f_α is the left quantile at $\alpha\%$ of the distribution f, representing the portfolio daily returns distribution. For instance, if the daily returns distribution is Normal, then $r_t^p \sim \mathcal{N}(0, h_t^{2,p})$. The sense of α is that

$$Pr(r_t^p < VaR_t) = \alpha. \tag{2.19}$$

In other words, α represents the probability that the portfolio loss on day t exceeds VaR_t. To be more concrete, let us suppose that the presumed distribution of the portfolio returns is Normal. Then suppose that $r_t^p = 0$ and $h_t^p = 0.075$. The daily VaR at 5% confidence level is given by $-1.64 \times 0.075 = -0.1233$, because of the left quantile at 5% for the Normal distribution is about -1.64. Thus, we expect at a 95% confidence interval that the portfolio loss will not exceed the value of 12.3% daily. The VaR computed for a longer time horizon is as follows:

$$VaR_t(T) = (r_t^p + f_\alpha h_t^p) \times \sqrt{T}, \tag{2.20}$$

Let us continue with the previous example but the time horizon is now a month. Given that there 20 trading days in a month, the monthly VaR

becomes $-0.1233 \times \sqrt{20} = -55.14\%$. This means that with a 95% confidence the portfolio losses are not expected to overcome the value of 55.14% in a given month.

Observing the VaR measures and portfolio daily returns, we can define the sequence of hit function $\{I_t\}_{t=1}^{T}$ as the number of VaR violations occurring in a given time period, where the hit function at time t is obtained as follows:

$$I_t = \begin{cases} 1 & \text{if } r_t^p < VaR_t \\ 0 & \text{if } r_t^p \geq VaR_t \end{cases} \tag{2.21}$$

The assessment of a model through the VaR and the hit function can be done in many methods, among which there are the Time Until First Failure (TUFF) test, the Unconditional coverage test and the Independence test. The TUFF test, proposed by Kupiec (1995), reports the first day in which a VaR violation occurs. The Unconditional Coverage (UC) test, always proposed by Kupiec, tests if the empirical frequency of violations is statistically equal to the prefixed α. The Independence test, due to Christoffersen (1998), checks if the VaR violations are clustered in time or are independently distributed over time. In this work we only consider the last two.

The null hypothesis of the UC test is $H_0 : \mathbb{E}[I_t] \equiv \pi = \alpha$, where π stands for the unconditional probability of a violation of a model. Assuming the independence of I_t for each t, the likelihood of the hit sequence will be given by the productory of a Bernoulli random variable, that is:

$$\ell(\pi) = \prod_{t=1}^{T} (1-\pi)^{1-I_{t+1}} \pi^{I_{t+1}} = (1-\pi)^{T_0} \pi^{T_1},$$

where T_0 and T_1 the number of zeros and ones, that is the number of non-violations and violations of the VaR in the sample, respectively. Let $\hat{\pi} = T_1/T$ the observed number of violations in the sample. If we insert $\hat{\pi}$ into the likelihood, we have:

$$\ell(\hat{\pi}) = (1 - T_1/T)^{T_0} (T_1/T)^{T_1}.$$

Under the null hypothesis that $\pi = \alpha$, we have the following likelihood:

$$\ell(\alpha) = \prod_{t=1}^{T}(1-\alpha)^{1-I_{t+1}}\alpha^{I_{t+1}} = (1-\alpha)^{T_0}\alpha^{T_1}.$$

Finally, we can check the null by using the likelihood test

$$LR_{UC} = -2\log\left[\ell(\alpha)/\ell(\hat{\pi})\right] \sim \chi^2_1. \tag{2.22}$$

Also the independence test is based on a likelihood ratio test. If the hit sequence is dependent over time such that it can be defined as a first-order Markov sequence with the following transition probability matrix:

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix},$$

where π_{01} is the probability that, given today being a non-violation, tomorrow a violation occurs, meaning that $I_t = 0$ for today and $I_{t+1} = 1$ for tomorrow. Moreover, π_{11} is the probability of a violation tomorrow given today being a violation ($I_t = I_{t+1} = 1$). Conversely, the probability of a non-violation following a non-violation is denoted as $(1-\pi_{01})$ and the probability of a non-violation following a violation as $(1-\pi_{11})$. Let the likelihood function of the first-order Markov process be

$$\ell(\Pi_1) = (1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}},$$

where T_{ij} is the number of observations with a j following an i. Now, π_{01} and π_{11} can be estimated by taking the first derivatives of $\ell(\Pi_1)$. It is easy to demonstrate that:

$$\begin{aligned}\frac{\partial\ell(\Pi_1)}{\partial\pi_{01}} =&[(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}][-T_{00}(1-\pi_{01})^{T_{00}-1}\pi_{01}^{T_{01}} + \\ &(1-\pi_{01})^{T_{00}}T_{01}\pi_{01}^{T_{01}-1}] = 0.\end{aligned}$$

This yields to:

$$\hat{\pi}_{01} = \frac{T_{01}}{T_{00}+T_{01}} \quad \text{and} \quad \hat{\pi}_{11} = \frac{T_{11}}{T_{10}+T_{11}}.$$

Because the probability has to sum to one, we have $\hat{\pi}_{00} = 1 - \hat{\pi}_{01}$ and $\hat{\pi}_{10} = 1 - \hat{\pi}_{11}$. At this point, we can formulate the null hypothesis of the independence test: $\pi_{01} = \pi_{11} = \pi$, based on the observation that under independence a violation tomorrow should not depend on today value of the hit function. The null hypothesis is tested using a likelihood ratio that assumes the form:

$$LR_{ind} = -2\log\left[\ell(\hat{\pi})/\ell(\hat{\Pi}_1)\right] \sim \chi^2_1,$$

where $\ell(\hat{\pi})$ has already been defined for the Unconditional Coverage test.

In the empirical part, we will use both the Unconditional and the Independence test in a jointly test as proposed by Christoffersen. In fact, this test indicated as Conditional Coverage (CC) test, verifies the empirical rate of failures and the independence of the violations jointly. The CC test is given by following likelihood ratio test:

$$LR_{CC} = -2\log[\ell(p)/\ell(\hat{\Pi})] \sim \chi^2_2,$$

that equals to test the hypothesis $\pi_{01} = \pi_{11} = \alpha$. Moreover, it is such that $LR_{CC} = LR_{UC} + LR_{ind}$.

The main drawbacks of the VaR approach in order to evaluate the volatility predictions are: (i) it is not a coherent risk measure; (ii) it ignores what happens in the tail of the distribution.

The coherency of a risk measure is defined in terms of a set of properties that the risk measures should have. The properties are: monotonicity, subadditivity, homogeneity, translational invariance. Artzner et al. (1999) detail all these properties. With reference to the Value at Risk, it does not have the subadditivity property. The subadditivity property states that evaluating the function for the sum of two elements of the domain always returns something less than or equal to the sum of the function's values at each element. Formally, in the VaR framework, the subadditivity property means that:

$$VaR_t(r_t^p) \leq VaR_t(X_1) + VaR_t(X_2),$$

where X_1 and X_2 are two stocks forming the portfolio whose daily returns are as usually expressed by r_t^p.

The other drawback of the VaR is that it ignores what happens in the tail of the distribution. In particular, if the VaR is a threshold of the daily return distribution for the day t, then it does not take any information once that threshold has been exceeded. For this reason, the Conditional VaR $cVaR_t$, also named Expected Shortfall has been proposed. It is equal to:

$$cVaR_t = \mathbb{E}\left[r_t^p | r_t^p < VaR_t\right].$$

The $cVaR_t$ is a coherent measure of risk, given that the property of subadditivity here holds.

2.4.4 Pairwise Predictive Ability Testing

The previous methods allow the comparison of volatility forecasts produced by m models. In some circumstances, it may be useful to perform a pairwise comparison, meaning that the models analysed are two (such that $m = 2$). In the next two paragraphs the Diebold-Mariano, West and Giacomini tests will be briefly illustrated. These tests have been extensively used, not only for the volatility forecasts evaluation but also for the evaluation of heterogeneous economic variables.

Diebold-Mariano Test

Diebold and Mariano (1995) propose to test the equivalence of the forecasting performance by means of a t-test, based on the difference between the estimated losses. In particular, let $\varepsilon^i_{t+1|t}$ be the forecast error made at time $t+1$ for the forecast computed at time t for the model i, with $i = 1, 2$. Outside the volatility framework, the forecast error is:

$$\varepsilon^i_{t+1|t} = x_{t+1} - f_i(I^t),$$

where f_i denotes the model i producing the forecast. Within the volatility framework, x_{t+1} can be replaced by a volatility proxy and $f_i(I^t)$ one of

the model seen in the previous chapter. The h-step ahead forecasts, with $h \geq 1$, are assumed to be computed for the period $t_2 = L+1, \cdots, T$ for a total of T_2 forecasts. Thus, the series of the forecast errors is denoted by $\{\varepsilon^i_{t+h|t}\}^T_{t=L+1}$.

The accuracy of each model can be naturally exploited by the loss functions, such as the MSE, MAE, and so forth. The Diebold-Mariano test checks if on average the two losses are equal. The null hypothesis is:

$$H_0: \quad \mathbb{E}\left[\mathcal{L}(\{\varepsilon^1_{t+h|t}\}^T_{t=L+1})\right] = \mathbb{E}\left[\mathcal{L}(\{\varepsilon^2_{t+h|t}\}^T_{t=L+1})\right]. \tag{2.23}$$

Let $d_t = \mathcal{L}(\varepsilon^1_{t+h|t}) - \mathcal{L}(\varepsilon^2_{t+h|t})$ be the loss differential between the model 1 and model 2 at time t by means of the loss function $\mathcal{L}$. The null hypothesis in (2.23) can be now written as

$$H_0: \quad \mathbb{E}[d_t] = 0. \tag{2.24}$$

Let $\overline{d}$ be the sample mean of d_t, that is:

$$\overline{d} = \frac{1}{T_2} \sum_{t=L+1}^{T} d_t. \tag{2.25}$$

The Diebold-Mariano test statistic is:

$$S = \frac{\overline{d}}{\sqrt{V(\overline{d})}}. \tag{2.26}$$

If $\overline{d}$ is an estimated but observed quantity, its variance $V(\overline{d})$ requires some attention because of the autocorrelation of d_t. In fact, each forecast error series will be serially correlated given that overlapping data are used. In particular, it can be show that $\{\varepsilon^i_{t+h|t}\}^T_{t=L+1}$ follows a $MA(h-1)$ process. If it is assumed that all the autocorrelations of order h or higher of d_t are zero, then it can be proven that the denominator of equation (2.26) is consistently estimated by:

$$V(\overline{d}) \approx T_2^{-1} \left[\gamma_0 + 2 \sum_{k}^{h-1} \gamma_k\right], \tag{2.27}$$

where γ_k is the $k-th$ autocovariance of d_t:

$$\gamma_k = T_2^{-1} \sum_{k+1}^{T_2} (d_t - \overline{d})(d_{t-k} - \overline{d}). \tag{2.28}$$

Hence, the test statistic (2.26) becomes:

$$S_c = \frac{\overline{d}}{\sqrt{\widehat{V}(\overline{d})}}, \tag{2.29}$$

where $\widehat{V}(\overline{d})$ is obtained through (2.27). Under the null hypothesis, $S_c \xrightarrow{d} \mathcal{N}(0,1)$. Finally, if $|S_c| > 1.96$, then one can argue that the model 1 and model 2 do not have equal predictive accuracy at 5% significance level. In this case, the sign of S_c indicates which model has a better forecast accuracy. If $S_c > 0$, then model 1 has greater forecast errors such that model 2 should be preferred and vice versa.

The Diebold-Mariano test has some unquestionable advantages. It can be applicable for any loss functions, for multi-period forecasts, and the forecast errors can be non-Normal distributed, nonzero-mean and serially and contemporaneously correlated. However, such test has at least three drawbacks. Firstly, it can be applied when the competing models are non-nested. When the models are nested, other tests have to be considered, like that proposed by Clark and McCracken (2001), among others. Secondly, the Diebold-Mariano test does not take into account the parameter estimation errors that derive from the models used to forecast. It only considers the forecast errors, that indeed could depend on the uncertainty inherent to $f_i(I^t)$. In fact, the asymptotic variance of $\overline{d}$ could be influenced by the in-sample parameter estimation. A brief overview of the solutions to this problem are discussed in the next paragraph. Thirdly, the asymptotic distribution of S_1 is seriously affected by the small size of the sample. It often happen that the T_2, the length of the pseudo out-of-sample period, is small due to several reasons. For example, the period used for the forecasting is small by itself. Possible solution/modification

of the Diebold-Mariano test for small sample size are provided in Harvey et al. (1997), among others.

West and Giacomini-White Tests

In this paragraph the West (1996) and Giacomini and White (2006) tests are briefly presented. Both tests concern the equal predictive ability like the Diebold-Mariano test but differently fro this latter, they take into account the estimation uncertainty. Recall that the whole sample period can be split in two not overlapping subperiods $t_1 = \{1, \cdots, L\}$ and $t_2 = \{L+1, \cdots, T\}$, with T_1 and T_2 representing the length of the two periods, respectively. Let us suppose that $f_i(I^t)$ uses the period t_1, the in-sample period, to estimate the parameters generating the forecasts for the pseudo out-of-sample period t_2. West proposes an asymptotic test for $T_1, T_2 \to \infty$, working under a squared loss function, whose null hypothesis is:

$$H_0 : \mathbb{E}\left[(x_{t+1} - f_1(\theta^*))^2 - (x_{t+1} - f_2(\theta^*))^2\right] = 0. \tag{2.30}$$

In (2.30), the forecasting models $f_1(\cdot)$ and $f_2(\cdot)$ depend explicitly on the parameter θ estimated in the in-sample period. This means that these forecasting models can be exclusively parametric. θ^* indicates that the parameters are considered at their population values, i.e. interpretable as probability limits of the parameter estimates when L_1 goes to infinity. The statistical test proposed by West is:

$$W_c = \frac{1}{T_2} \sum_{t=L+1}^{T-h} \frac{d_t(\hat{\theta}_t)}{\hat{\sigma}}, \tag{2.31}$$

where $d_t(\hat{\theta}_t)$ and $\hat{\sigma}$ are the loss differential and the standard deviation depending on the possible uncertainty of the in-sample parameter estimation, respectively. West (1996) shows how to construct $\hat{\sigma}$ identifying many cases in which the parameter uncertainty does not count asymptotically. But, if the denominator of the Diebold-Mariano test is quite simple to calculate, $\hat{\sigma}$'s computation may be difficult, given that it depends on

the forecasting scheme (recursive, rolling or fixed) and the two models. Another drawback is that the West statistic works adequately only for non-nested models, as pointed out by Clark and McCracken (2001).

The Giacomini-White (GW) test differs from the West test under many points. As pointed out by their authors, the (GW) test has a null hypothesis of equal conditional predictive ability, while (2.30) has a null of equal unconditional predictive ability. Why conditional? Because the forecasting models are now function of the estimated parameters (while in the West test they were function of the population parameters). The null hypothesis is:

$$H_0 : \mathbb{E}\left[\left(x_{t+1} - f_1(\hat{\theta}_t)\right)^2 - \left(x_{t+1} - f_2(\hat{\theta}_t)\right)^2 |I^t\right] = 0. \qquad (2.32)$$

The main difference between the West and the GW test is that the former focuses only on the forecasting model (whose parameters are evaluated at their population values), while the latter takes into account the whole forecast method, consisting of the forecasting model, the estimation procedure, the window size of the data to use, the possible non stationarity of the data generating process, and so forth. In fact, the GW test allows for the parameters changing over time, differently from the West test. Moreover, other important differences occur between the two approaches. First, the GW test can be applied on nested models. This is an important aspect when the forecaster considers as benchmark a simple model and he needs to compare it to a more sophisticated version (with a larger number of parameters) of this benchmark model. Second, as evidenced by Giacomini and White, the GW test is capable to prefer simpler models to larger correctly specified models if these latter have a large amount of uncertainty with reference to the parameters estimation.

As drawback, the GW test excludes the possibility to use the recursive forecasting scheme, while it allows for the rolling and fixed schemes. This is because the asymptotic environment of the GW test considers a fixed in-sample size T_1 while T_2 increases to infinity, even though some

recent works allow to use the GW test with a recursive scheme in special cases (see Clark and McCracken (2009) for details).

2.4.5 Multiple Predictive Ability Testing

In the previous section the pairwise testing has been discussed. However, this is only a special case of the multiple predictive ability testing, where a set of competing models is compared. More specifically, the forecaster could be interested in testing the predictive ability of a set of models against a benchmark, in order to find the best model or the subset of models with equal forecast accuracy. This framework will be illustrated in the next two paragraphs.

The Reality Check

Let $d_{t+h}^{(i)} = \mathcal{L}_{t+h}^{(0)} - \mathcal{L}_{t+h}^{(i)}$ be the loss differential between the benchmark, identified by 0, and the $i-th$ model, with $i = 1, \cdots, m$. If $d_{t+h}^{(i)} > 0$, then at time $t+h$ the model i outperforms the benchmark, because its loss is smaller than that of the benchmark. Instead, if $d_{t+h}^{(i)} < 0$, the model i has worse forecast accuracy with respect to that of the benchmark. Hence, the idea is to test that none of the m models is better than the benchmark. The Reality Check, proposed by White (2000), formalizes this idea as

$$H_0 : \max_{i=1,\cdots,m} \mathbb{E}\left[\{d_{t+h}^{(i)}\}_{t=L+1}^{T}\right] \leq 0$$
$$H_1 : \max_{i=1,\cdots,m} \mathbb{E}\left[\{d_{t+h}^{(i)}\}_{t=L+1}^{T}\right] > 0$$

If the null hypothesis is not rejected, then the benchmark is never outperformed. Otherwise, if the null is rejected, the model i outperforms the benchmark. Note that if $m = 2$, then the reality check is nothing else than the Diebold-Mariano test. Let $\mathbf{d}_{t+h} = (d_{t+h}^{(1)}, \cdots, d_{t+h}^{(m)})'$ and

$\mathbb{E}\left(\{\mathbf{d}_{t+h}\}_{t=L+1}^{T}\right) = \boldsymbol{\mu}$. Now, the null hypothesis can be written as:

$$H_0: \quad \boldsymbol{\mu} \leq \mathbf{0}.$$

White makes the following assumption on $\mathbf{d}_{t+h}$:

A1: The vector $\{\mathbf{d}_{t+h}\}$ is strictly stationary on a strong mixing base of size $-(2+\delta)(r+\delta)/(r-2)$, with $r > 2$, $\delta > 0$, where $\mathbb{E}\left(|\{\mathbf{d}_{t+h}\}|^{r+\delta}\right) < \infty$ and $Var(d_{t+h}^{(i)}) > 0, \forall i = 1, \cdots, m$.

For simplicity, let us suppose that there is only one step ahead forecast, such that $h = 1$. Recall that we are in the pseudo out-of-sample period, whose length is T_2. Given A1, the central limit theorem can be applied such that

$$T_2^{-1}(\overline{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{MVN}(\mathbf{0}, \Omega), \tag{2.33}$$

where $\overline{\mathbf{d}} = T_2^{-1}\sum_{t=L+1}^{T}\mathbf{d}_t$ and $\Omega = aVar(T_2^{-1}(\overline{\mathbf{d}} - \boldsymbol{\mu}))$, with $aVar(\cdot)$ indicating the asymptotic variance of the argument. As argued by Clark and McCracken (2001), the asymptotic normality does not hold when the benchmark is nested in all alternative models (under the null hypothesis) and the parameters are estimated recursively. White uses the following test statistic,

$$T^{RC} = \max(T_2^{1/2}\overline{d}_1, \cdots, T_2^{1/2}\overline{d}_m), \tag{2.34}$$

where $\overline{d}_i = T_2^{-1}\sum_{t=L+1}^{T} d_t^{(i)}$. The asymptotic null distribution of (2.34) is based on $T_2^{1/2}\overline{\mathbf{d}} \sim \mathcal{MVN}(\mathbf{0}, \widehat{\Omega})$. Unfortunately, the estimation of the $\widehat{\Omega}$ is not simple, mainly if m is large. To handle this problem, White suggests a bootstrap procedure in order to derive the p-values of the test. The bootstrap procedure relies on the stationary bootstrap of Politis and Romano (1994), provided that A1 holds. Other assumptions needed to use the Reality check are (i) at least one model has to be non nested with the benchmark; (ii) the uncertainty deriving from the parameters estimation does not count asymptotically. Some recent developments in the Reality Check literature are discussed in Corradi and Swanson (2011).

The Superior Predictive Ability test

The Superior Predictive Ability (SPA) test, due to Hansen (2005), modifies the White's procedure obtaining a test more powerful and less sensitive to the inclusion of poor and irrelevant forecasts than the Reality Check. This result is achieved by two modifications of the Reality Check procedure. First, a studentized test statistic is employed. Second, a data dependent null distribution is used to incorporate additional sample information such that the irrelevant forecasts are identified. This allows for a reduction of their influence on SPA test. More specifically, the studentized SPA test statistic modifies (2.34) in:

$$T^{SPA} = \max \left[\max_{i=1,\cdots,m} \frac{T_2^{1/2} \overline{d}_i}{\widehat{\omega}_i}, 0 \right], \tag{2.35}$$

where $\widehat{\omega}_i$ is a consistent estimator of $\omega_i = \left[Var(T_2^{1/2} \overline{d}_i) \right]^{1/2}$.

The data dependent null distribution is based on $\mathcal{MVN}(\hat{\boldsymbol{\mu}}^c, \widehat{\Omega})$, where $\hat{\boldsymbol{\mu}}^c$ is an estimator for $\boldsymbol{\mu}$ that conforms for the null hypothesis. Hansen suggests the estimator:

$$\hat{\boldsymbol{\mu}}^c = \overline{d}_i I_{\{T_2^{1/2} \overline{d}_i / \widehat{\omega}_i \leq -\sqrt{2 \log \log T_2}\}}, \tag{2.36}$$

where $I_{\{\cdot\}}$ denotes the Indicator function. As for the Reality Check, this framework does not allow for the comparison of nested models when the recursive scheme is used.

The Model Confidence Set

The idea of Model Confidence Set (MCS), due to Hansen et al. (2011), is to explore the circumstances in which the Reality Check and the Superior Ability Test null hypotheses are not rejected. In this case, we only know that none of the m models outperforms the benchmark or at least some (or all) have equal predictive ability. Hence, the aim is to repeat the analysis step-by-step ruling out the models with the worse performances

in order to retain all the models with equal forecasting performances. These models form the so called Model Confidence Set, that is a set of models with equal predictive ability. The models belonging to the MCS are determined by the following procedure:

1. $\forall i, j$, with $i \neq j$ and $i, j = 1, \cdots, m$, the null hypothesis of equal predictive ability as in the Diebold-Mariano test (equation 2.23) is evaluated by the statistic

$$MCS_c = \max_{i,j} t_{i,j}, \tag{2.37}$$

   where again $t_{i,j}$ is the Diebold-Mariano test statistic in (2.29).

2. If the null is not rejected, all the m models entry in the MCS, given that they all have the same predictive ability. Instead, if the null is rejected, then the model with the highest average loss are eliminated and step 1 is repeated till the null is not rejected.

As for the Reality Check and the Superior Predictive Ability, the p-value for the test in step 1 is obtained by a bootstrap procedure. For details about the bootstrap procedure, see the appendix of the cited paper of Hansen et al. (2011).

2.5 Combining forecasts

In the previous sections some methods to find the best model or the set of models with equal predictive ability have been presented. This is a possible solution when the forecaster has many forecasting models. Another possible solution is to combine all or some models in order to create a kind of "super" forecasting model. An excellent review of the forecast combinations is provided by Timmermann (2006). There are many reasons advocating the utility of the forecast combinations. For instance, given that the data could be affected by structural breaks, a model i could adapt very slowly while another model j very quickly. Because of the

difficulty to recognize on time the structural breaks, the combinations of forecasts is expected to outperform the individual models. Evidences of this intuition are illustrated in Pesaran and Timmermann (2007). Another possible justification for the forecast combinations is the presence of misspecification biases and measurement errors in the information sets used by each model to forecast. Combining the forecasts makes the resulting "super" model more robust to these misspecification and measurement errors.

However, there are some drawbacks in using the forecast combinations. First of all, the forecaster has to deal with the problem of finding the optimal weights. And, with reference to the last reason advocating the utility of the forecast combinations, the research of optimal weights may be severely affected by the presence of misspecification and measurement errors.

The problem of forecast combinations can be formalized as follows. As done previously, suppose that the realizations of the variable of interest x_t are observed for h periods ahead and these realizations are indicated by x_{t+h}. Moreover, suppose that the m individual forecasting models have forecasted the variable of interest for the same h periods ahead. These forecasts are denoted by $\hat{\mathbf{x}}_{t+h} = (x_{t+h}^{(1)}, \cdots, x_{t+h}^{(m)})'$. We aim to find a vector of optimal weights $\mathbf{w}_{t+h}^*$ reducing the multivariate vector $\hat{\mathbf{x}}_{t+h}$ to a scalar through the function $g(\cdot)$. Formally, the combined forecast is denoted by:

$$x_{t+h}^{(c)} = g(\hat{\mathbf{x}}_{t+h}; \mathbf{w}_{t+h}). \tag{2.38}$$

Equation (2.38) means that the combined forecast is obtained through an opportune function linking $\mathbf{w}_{t+h}$ to $\hat{\mathbf{x}}_{t+h}$. In (2.38) the weights are assumed to be time-varying but they can always be considered as time-invariant. A classical distinction concerning the function $g(\cdot)$ is its linear or non-linear nature.

The optimal weights $\mathbf{w}_{t+h}^*$ are determined solving the problem of

finding the smallest forecast error deriving from the unavoidable difference between x_{t+h} and $x_{t+h}^{(c)}$. The resulting loss function $\mathcal{L}(e_{t+h}^{(c)})$, for simplicity, is assumed to depend only on the forecast error $e_{t+h}^{(c)}$ of the combination, with $e_{t+h}^{(c)} = x_{t+h} - g(\hat{\mathbf{x}}_{t+h}; \mathbf{w}_{t+h})$. For instance, the uncertainty of the parameters estimation for each model does not play any rule. Hence, the optimal weights $\mathbf{w}_{t+h}^{*}$ solve the problem:

$$\mathbf{w}_{t+h}^{*} = \min_{\mathbf{w}_{t+h}} \mathbb{E}\left[\mathcal{L}(e_{t+h}^{(c)}(\mathbf{w}_{t+h}))|\hat{\mathbf{x}}_{t+h}\right], \tag{2.39}$$

where in $e_{t+h}^{(c)}(\mathbf{w}_{t+h})$ it has been evidenced that the forecast error depends on the weights vector.

The minimization problem in (2.39) generally does not admit a closed-form solution. Some solutions can be achieved by imposing restrictions on the loss function, on the function $g(\cdot)$, and so forth. One common solution is to consider the linear combination of equal weights. In this case, we have that

$$x_{t+h}^{(c)} = \frac{1}{m}\boldsymbol{\iota}'\hat{\mathbf{x}}_{t+h}, \tag{2.40}$$

where $\boldsymbol{\iota}$ is a $m \times 1$ vector of ones. This specification of forecasts combination usually represents the benchmark used to compare other more complicated combinations. For instance, one may consider to rule out the worse models by giving them a weight equal to zero, use a nonlinear combination methods, and so forth.

Appendix

Table 2.2 – Specifications of loss functions used in the work

	Name	form of Λ	symmetry
$\mathcal{L}_E$	Euclidean distance	$\Lambda = I_k$	symmetric
$\mathcal{L}_{SE}$	Squared weighted Euclidean distance	$\lambda_{i,i} > 0$ and $\lambda_{i,j} = 0$	symmetric
$\mathcal{L}_{M-O}$	Mahalanobis distance over prediction vers.	$\lambda_{i,i} = \lambda_{i,j} = 1 + I_{\{ov\}}$, where $I_{\{ov\}} = \begin{cases} 0 & \text{if } \hat{\sigma}_t - h_t \geq 0 \\ 1 & \text{if } \hat{\sigma}_t - h_t < 0 \end{cases}$	asymmetric
$\mathcal{L}_{M-U}$	Mahalanobis distance under prediction vers.	$\lambda_{i,i} = \lambda_{i,j} = 1 + I_{\{ov\}}$, where $I_{\{ov\}} = \begin{cases} 0 & \text{if } \hat{\sigma}_t - h_t \leq 0 \\ 1 & \text{if } \hat{\sigma}_t - h_t > 0 \end{cases}$	asymmetric

Notes: λ_{ii} indicates the element of Λ referring to the variance element ii of the forecast error matrix; λ_{ij} to the covariance element ij, with $i, j = 1, \cdots, k$. $\mathcal{L}_E$ equally weights the variance and covariance elements. $\mathcal{L}_{SE}$ weights only the variance elements. $\mathcal{L}_{M-O}$ penalizes the over predictions, such that if there is an over prediction at time t, the loss function counts twice that forecast error. $\mathcal{L}_{M-U}$ penalizes the under predictions, such that if there is an under prediction at time t, the loss function counts twice that forecast error.

3

Evaluation of Volatility Forecasts in a VaR framework

3.1 Introduction

In the previous chapter two methods for the evaluation of volatility forecasts have been discussed: the statistical and economic approaches. In this chapter we present a new method, that mixes the previous two. In particular we focus on the evaluation of volatility forecasts by means of the VaR measures and then we evaluate these measures through the loss functions. In this context, the loss function compares the losses of a portfolio (or of an asset) to the VaR quantities. More violations occur, greater the value of the loss function is, worse the correspondent model is. As far as we know, in literature there are not so many loss functions specifically used for evaluating the VaR measures. Lopez (1998) introduces the Magnitude loss function that takes into account the distance between the loss and the VaR measures when a violation occurs. The Firm loss function of Sarma et al. (2003) adds to the former the opportunity cost of capital faced by the firm when there is not a violation. Both loss functions are symmetric: they equally penalize models whose number of violations is larger or smaller than the expected one. As stated above, in this scenario

the portfolio holder does not prefer models with a large number of violations. Moreover, the Firm loss function experiences the problem of choosing the opportunity cost of capital. In this chapter, a new asymmetric loss function is proposed. The term asymmetric means that the models with a number of violations larger than the expected one are penalized more than the models with a smaller number of violations.

The aim of this chapter is to investigate the opportunity to use the loss functions in a VaR framework in order to evaluate the volatility predictions of a set of competing models. More specifically, we are interested in verifying if this method can be helpful to find the best model, independently of the distributional assumptions of the daily returns.

Following Bowers and Heaton (2013), (block) bootstrapping the intraday increments of a generic asset allows to have consistent estimators of any characteristics of that asset's daily return, independently of its distribution, as desired. Hence the VaR measures are obtained as a quantile of the estimated distribution of the daily return for that day.

As it is known, the loss function gives a value - named numerical score - representing the average of the distance between the two inputs for all the sample period considered. In the previous chapter, the two inputs were the actual and forecasted values of the variable of interest, generally and the volatility proxy and the volatility prediction, specifically. In this framework, the two inputs are the observed daily returns and the estimated VaR measures. Hence, a question arises: can we find a method to discriminate high numerical scores from low numerical scores? Our answer is positive and it is based again on the block bootstrap method. The proposed simple method helps the researcher, the portfolio holder and the risk manager to empirically find the volatility model with a superior predictive ability in situations where the statistical and economic approaches are not in accordance with the choice of the best model.

The aim of the chapter is pursued by means of a Monte Carlo experiment and an analysis carried out on a stock listed on the New York Stock

Exchange, the same used for the graphical description of the stylized facts discussed in Chapter 1. In the simulation, a data generating process (DGP) following a GARCH(1,1) with Normal innovations is considered. Then the intraday increments are simulated such that the variance of each increment is equal to the daily variance divided by the number of intraday observations. The set of competing models are seven misspecified models: a Riskmetrics, three GARCH(1,1) and three GJR-GARCH(1,1) models, each with different specifications. The evaluation of their performances relies on three accuracy measures. Let a *rejection* be the event that a numerical score lies above the threshold. The first accuracy measure is the frequency of rejections of the DGP with those of each model. The second accuracy measure is the frequency at which the model's numerical score lies below the chosen threshold. The third accuracy measure is the frequency at which the model ranks first (meaning that the relative numerical score is the smallest among all the models). The linear combination of these measures gives a synthetic value, varying between zero and one. This value allows not only to compare cross-sectionally the models, but also to compare the goodness of accuracy among the loss functions.

In the empirical analysis the volatility of the stock of the Capital One Financial Corporation is evaluated. If in the Monte Carlo experiment the evaluation of forecasts is made in a pure-in-sample perspective, here the analysis concerns the out-of-sample perspective, by means of a rolling window of length 250 days. Again, the VaR measures are obtained by block bootstrapping the intraday increments. Finally the thresholds, one for each loss function, are calculated in order to evaluate the performances of a set of competing models.

We start in Section 3.2 by presenting the method used to estimate the daily VaR measures. In Section 3.3 the two loss functions in a VaR framework are illustrated. In the same section the three accuracy measures that will used in the Monte Carlo simulation in Section 3.4 are presented. The

results of the volatility evaluation, when the real data are used, are illustrated in Section 3.5. Section 3.6 concludes.

3.2 Nonparametric estimation of the VaR measures

Suppose that a sequence of N intraday increments for a generic day t and a generic asset is observed. Formally, the sequence[1] is denoted by: $\chi_{t,N} = \{q_{t,1}, \cdots, q_{t,N}\}$, such that

$$q_{t,n} = p_{t,n} - p_{t,n-1}, \quad \forall n = 2, \cdots, N. \tag{3.1}$$

In (3.1), $p_{t,n}$ denotes the observed intraday log price at day t and n^{th} intraday frequency, with $n = 1, \cdots, N$. By definition, $q_{t,1} = 0$, $\forall t$. Thus, the open-to-close daily return $r_{t,N}$ is given by $r_{t,N} = \sum_{n=2}^{N} q_{t,n}$. We aim to find the Value at Risk measure for the daily return $r_{t,N}$ on the basis of the observation of the intraday increments. Recall that the Value at Risk is nothing else than a quantile of the presumed distribution of the daily returns. If the data were $i.i.d.$, the distribution of any statistic of interest (mean, median, etc.) could be obtained simply re-sampling with replacement B times the original sample and calculating B statistics of interest. The B statistics of interest represent the bootstrap distribution for that statistic. Unfortunately, the previous method proposed by Efron (1979) cannot be applied on the sequence $\chi_{t,N}$ because of the lack of independence of the time series. But a solution exists. It is the Stationary Bootstrap (SB) of Politis and Romano (1994), that allows for data dependency. In the SB, for each day t, B re-sampled intraday sequences are calculated, each of length N. A re-sampled intraday sequence is formed by N sampled blocks, whose average block length depends on the dependence exhibited within $\chi_{t,N}$. Like Bowers and Heaton (2013), we use

[1] We use a slightly different notation here indicating the intraday increments or returns. In Chapter 1 the intraday increments were denoted by $r_{t,d}$. Here they are denoted by $q_{t,n}$. This is because we need to make clearer their difference with the daily returns depending on N daily increments, indicated with $r_{t,N}$.

the procedure described in Patton et al. (2009) in order to estimate the average block length, procedure that entirely depends on the observed data. Once the bootstrapped sequence is obtained, the resulting summation represents a re-sampled daily return, independent of the original one, but generated by the same distribution, as $\{N, B\} \to \infty$. Any moment or quantile of the original return can be now estimated by means of the B *i.i.d.* sequence of the re-sample daily returns. Formally, let $r^*_{t,N,b}$ be a re-sampled daily return given the summation of N blocks independently drawn from the sequence $\chi_{t,N}$ and average block length depending only on $\chi_{t,N}$. Let $r^*_{t,N} = B^{-1} \sum_{b=1}^{B} r^*_{t,N,b}$. By construction, $\mathbb{E}^* r^*_{t,N} \equiv r_{t,N}$. Let

$$q_{t,n} = N^{-1} X_{t,n}. \tag{3.2}$$

As argued by Bowers and Heaton (2013), (3.2) allows for increments of different length, as normally happens. The assumptions needed to assure that $r^*_{t,N}$ and $r_{t,N}$ converge in distribution, given in Gonçalves and White (2002) and Gonçalves and de Jong (2003), are[2]:

A1: X_n is $L_{2+\delta}$ NED (Near Epoch Dependent) of size -1 on a strong mixing base of size $-(2+\delta)(r+\delta)/(r-2)$, with $r > 2$, $\delta > 0$.

A2: $\mathbb{E}|X_n|^{r+\delta} < \infty, \forall n$.

A3: The homogeneity condition described in Gonçalves and White (2002) holds for $\mathbb{E}X_n$.

A4: $\mathbb{E}(r_{t,N}) = 0$.

Many kinds of different weak dependence specifications are allowed by assumptions 1-3. In addition, the assumption 1 allows for micro-structure effects. More delicate is the assumption 4, that may be violated. If one is interested in the estimation of naturally centred moments of the daily returns like the variance, skewness and so forth, assumption 4 does not

[2]For sake of simplicity, the suffix t is omitted.

count. But, if one is interested in the estimation of quantiles like the VaR, assumption 4 is needed to assure the convergence in distribution, as defined in the following proposition:

Proposition 1: Given (3.2), if Assumptions 1-4 hold, then:

$$\sup_{x \in R} \left| P\left(r_{t,N} \leq x\right) - P^*\left(r^*_{t,N} - \mathbb{E}^* r^*_{t,N} \leq x\right) \right| \xrightarrow{P} 0, \text{ as } N \to \infty, \tag{3.3}$$

where P^* denotes that the probability is conditional on $\chi_{t,N}$. Proposition 1 means that the *i.i.d.* sequence $r^*_{t,N,b}$, with $b = 1, \cdots, B$, after having subtracted $r_{t,N}$, converges in distribution to that of $r_{t,N}$, whatever it is, as $\{N, B\} \to \infty$. Now, the sequence $r^*_{t,N,b}$ can be used to consistently estimate any moments or quantile of the original daily return. Because we are interested in finding a consistent estimator of the VaR in order to use it in a loss function, let $r^*_{t,N,[b]}$ indicate the b^{th} quantile of $r^*_{t,N,b}$. A consistent estimator of the 5% VaR for the day t is given by:

$$\widehat{VaR}_t = r^*_{t,N,[0.05B]} - \mathbb{E}^* r^*_{t,N}. \tag{3.4}$$

Equation (3.4) allows to estimate the true VaR independently of the daily return distribution. The drawback of this pretty simple method is the computational timing it requires. For each trading day t, first the average block length has to be calculated, then the B independent bootstrap samples are obtained. If B and T are quite large, the timing could become a problem.

3.3 VaR and loss function approach

Instead of using conditional coverage test as described in the previous chapter, the VaR measures can be evaluated by the loss function (LF). Precisely, in this context, the LF compares the profits and losses of an hypothetical portfolio to the VaR measures. Generally speaking, a good

model should have not only a number of VaR violations in line with the expected one, but also a distance between the losses and the VaR as much as possible small. Then, the comparison between the competing models is done looking at their numerical scores: the smaller the numerical score for model i is, the better the model i is. In this work we consider two loss functions.

The first LF due to Lopez (1998) is called Magnitude loss function, because it relies on the magnitude, the distance between the losses and the VaR measures only when the former are greater than the latter. Denoted by $Loss_M$ at time t it is equal to:

$$Loss_{M,t} = \begin{cases} 1 + (r_t - VaR_t)^2 & \text{if } r_t < VaR_t \\ 0 & \text{if } r_t \geq VaR_t \end{cases} . \tag{3.5}$$

Then, the mean of equation (3.5) is: $NS_M = T^{-1} \sum_{t=1}^{T} Loss_{M,t}$, where the acronym NS stands for *Numerical Score*.

The second LS is a new asymmetric function, denoted by NS_A. It penalizes more the model with a number of violations greater than the expected one, denoted by $\alpha_0 = \alpha \cdot T$. Let $\hat{\alpha}$ be the number of times that the losses are larger than the VaR and R be the quantity that expresses the difference between α_0 and $\hat{\alpha}$, such that: $R = (\hat{\alpha} - \alpha_0)/\alpha_0$. If $R > 0$, the model has experienced more violations than expected. Thus, to penalize it, we set up a quantity denoted by P, by using the exponential function:

$$P = \begin{cases} exp(R) & \text{if } R > 0 \\ 1 & \text{if } R \leq 0 \end{cases} . \tag{3.6}$$

The following step is similar to the construction of $Loss_M$ at time t:

$$Loss_{A,t} = \begin{cases} 1 + P \cdot (r_t - VaR_t)^2 & \text{if } r_t < VaR_t \\ 0 & \text{if } r_t \geq VaR_t \end{cases} . \tag{3.7}$$

In the last step, we obtain the final value of the loss function as: $NS_A = T^{-1} \sum_{t=1}^{T} Loss_{A,t}$. The trick of using the exponential function in (3.6) makes a model overpenalized if it has a lot of violations.

As aforementioned, we aim to investigate the opportunity to use the loss functions in a VaR framework. To reach this goal, we propose the following method for the construction of the threshold discriminating low from high numerical scores. The procedure is again based on the block bootstrap. The threshold is an empirical quantile of the distribution of the numerical score, when $\widehat{VaR}$ is used. For brevity, we omit the suffix of the Magnitude or Asymmetric loss functions. Let $LS = \{Loss_1, \cdots, Loss_T\}$ be the sequence of zeros and nonzeros for the Magnitude or Asymmetric LF when $\widehat{VaR}$ is used. Before proceeding with the usual block bootstrap for depending data, we note that the construction of the threshold should be exactly based on a number of nonzeros equal to a_0. If this does not happen, the resulting threshold is upward or downward biased, when the number of nonzeros are greater or smaller than α_0, respectively. Let us start with the first case. To take into account this problem, we split the procedure in three steps.

1. Let LS_{ps} be the a new sequence drawn from LS such that the number of nonzeros are equal to α_0. The computation of LS_{ps} is based on a simple random replacement of the nonzeros with the zeros. Obliviously, the length of LS_{ps} is T.

2. LS_{ps} contains dependent data. The usual procedure of block bootstrap is used, after having calculated the average block length as discussed above. Let $NS^*_{b,s}$ be the b^{th} bootstrap re-sample of LS_{ps}. Let $NS^*_s = B^{-1} \sum_{b=1}^{B} NS^*_{b,s}$.

3. Step 1-2 are repeated S times such that a sequence of NS^*_s is observed. Every time step 1 is repeated, the random replacement of nonzeros changes such that LS_{ps} will be always different.

The threshold is calculated as the 80% quantile of the NS^*_s distribution, with $s = 1, \cdots, S$, that is $TR = NS^*_{[0.8S]}$. We use the 80% quantile because the same quantile was chosen by Lopez (1998), even though he

opted for another (parametric) procedure for the threshold calculation. Naturally, the Magnitude and Asymmetric loss functions will have different thresholds if the number of violations is greater than the expected one. If a numerical score of a volatility model lies above the threshold, that model is considered rejected.

If the number of violations when $\widehat{VaR}_t$ is used is less than α_0, the threshold as obtained by the bootstrap procedure is downward biased. As before, we use a three steps procedure. Steps 2-3 are identical to those presented above while the first step changes. The data for the block bootstrap are collected in LS_{ns}, which comes from the random replacing of the zeros in LS with the mean of nonzeros. How many zeros are random replaced? Till the point in which the nonzeros are exactly equal to α_0.

When the sequence LS has a number of nonzeros exactly equal to α_0, then the stationary bootstrap can be directly applied.

The accuracy measures and their linear combination

In order to indirectly compare the volatility predictions of a set of competing models, a linear combination of three accuracy indexes is proposed. This allows to synthesize in one value the goodness of the volatility predictions for each model. Even though the setting of Monte Carlo experiment will be exposed in the next section, we have to underline that the accuracy measures are based on how many times each competing model reaches the best results. These measures check the number of times that an event happens. The number of times here is indexed by R, where R is the acronym for Replicate. The illustration of the three events, one for each accuracy measure, is explained as follows.

Let v be the number of times for which the DGP is above the threshold. The DGP in this framework is identified with the process using the $\widehat{VaR}_t$ estimates. We impose the following restriction:

$$0 < v \leq R, \tag{3.8}$$

meaning that the DGP lies above the threshold at least one time till a maximum of R times. Let J_i the number of times for which the model i is above the threshold, with $i = 1, \cdots, m$. Differently from the DGP, we impose that:

$$0 \leq J_i \leq R, \qquad \forall i \tag{3.9}$$

The first accuracy measure, indicated as Acc_1, matches the occurrences of replicates of the model i exceeding the threshold with those of the true model, in percentage terms. Let $\{v\}$ and $\{J_i\}$ be the replicates that lie above the threshold for the DGP and the model i, respectively. For example, there could happen that the true model experiences 3 replicates above the thresholds, say the $14-th$, $19-th$ and $27-th$. In this case, $v = 3$ and $\{v\} = \{14, 19, 27\}$. With reference to the generic volatility model, there could happen it experiences 5 replicates above the thresholds, that are $\{J_i\} = \{2, 4, 6, 19, 25\}$. In this extremely simplified case, the generic model has a low accuracy measure, because it presents a joint occurrence of exceedings only for the replicate $R = 19$. Formally, the first accuracy measure, for the model i, is obtained as:

$$Acc_{1,i} = length(\{v\} \cap \{J_i\})/v \tag{3.10}$$

Given (3.8), $Acc_{1,i} = 0$ if and only if there is any coincidence between a exceeding replicate of the true model and a exceeding replicate of the model i. This happens when $\{v\} \cap \{J_i\} = \emptyset$. Vice versa, $Acc_{1,i} = 1$ if and only if all the J exceedings of the model i coincide with the v exceedings of the true model. Thus, $Acc_{1,i}$ varies in the interval [0,1].

The second accuracy measure, indicated with $Acc_{2,i}$, represents the percentage of replicates for the model i that lie below the thresholds. Formally, it is obtained as:

$$Acc_{2,i} = 1 - \frac{J_i}{R}. \tag{3.11}$$

If $J_i = 0$, then $Acc_{2,i} = 1$, meaning that the model i has all the numerical scores below the thresholds. If $J_i = 1$, then $Acc_{2,i} = 0$. In this

unlucky circumstance, the model i has all the numerical scores above the thresholds, meaning that it experiences very bad volatility predictions to make always violated the VaR measures. Hence, $0 \leq Acc_{2,i} \leq 1$.

The third accuracy, RF_i, represents the frequency at which the model i ranks first. Formally, it is:

$$RF_i = R^{-1} \sum_{r=1}^{R} CF_{i,r}, \tag{3.12}$$

where

$$CF_{i,r} = \begin{cases} 1 & \text{if } NS^*_{i,r} : \ \nexists\, NS_{l,r} < NS^*_{i,r}, \forall l \neq i \\ 0 & \text{otherwise} \end{cases}.$$

$RF_i = 0$ if and only if the model i always experiences the highest numerical scores, compared to the numerical scores of the other $l \neq i$ models. Instead, $RF_i = 1$ if and only if the model i presents, for all the replicates, the smallest numerical scores. Thus, also the third accuracy measure lies in the interval [0,1].

The linear combination of the accuracy measures is denoted by AM_i, such that $AM_i = 3^{-1}[Acc_{1,i} + Acc_{2,i} + RF_i]$. Hence, AM_i varies in the interval [0,1]. The higher AM_i is, the better the performance of the model i is, because it has all the replicates below the threshold, it has the same performance of the DGP and its numerical scores are low compared to those of the other models. The construction of AM_i makes comparable all the competing models and all the loss functions.

3.4 Simulation experiment

In this section the setting of the Monte Carlo simulation is illustrated. The data generating process has the usual univariate GARCH representation, that is: $r_t = h_t z_t$, with $z_t \sim \mathcal{N}(0,1)$ and $h_t^2 \sim GARCH(1,1)$ process. In particular:

$$h_t^2 = 0.01 + 0.10 r_{t-1}^2 + 0.85 h_{t-1}^2, \quad t = 1, \cdots, 1099. \tag{3.13}$$

The parameters have been chosen following Christoffersen (1998). The first 99 observations are used as a warm-up period. The rest of the sample is used to simulate the intraday returns in order to have an estimate of the VaR for each period by means of the stationary bootstrap. The simulation of the intraday increments for $t = 100$ is as follows. In $t = 100$, $h^2_{t=100}$ becomes observed. For sake of simplicity, the sequence of increments are assumed to be drawn independently from a Normal distribution, whose variance is constant and equal to $(1/N)h^2_{t=100}$. The increments for the day $t = 100$ are thus $q_{t=100,1}, \cdots, q_{t=100,N}$. The daily return at time $t = 100$ is obtained as: $r_{t=100,N} = \sum_{n=1}^{N} q_{t=100,n}$. Again, $h^2_{t=101}$ is estimated through (3.13), given that $r_{t=100,N}$ is observed now. Then, the constant variance of the increments is calculated in order to obtain the increments. The procedure is repeated recursively till $t = 1099$, such that we obtain a sequence of 1000 daily returns, whose conditional variance is described in (3.13) and whose intraday increments are observed. More specifically, we set a number of intraday increments N equal to 390 and a number of replicates R equal to 200. Because the standard daily trading period is 6 hours and 30 minutes, we have setted a number of increments exactly equal to 390 in order to have one increment for each minute. Once obtained the sequence of increments, the stationary bootstrap, as illustrated above, is used to have $\widehat{VaR_t}$, for $t = 100, \cdots, 1099$ and $R = 1, \cdots, 200$. Finally, the thresholds for each loss function and each replicate are obtained.

The set of competing models are showed in Table 3.1. We consider a RiskMetrics, three GARCH(1,1) and three GJR-GARCH(1,1) models, all misspecified. These three models have been described in Chapter 1. Note that the true α_1 and β_1 parameters in (3.13) sum to 0.95, like all the parameters of the models M2-M7. But if M2 and M3 exhibit small deviation from the true DGP, M4 presents a higher distance from it. Instead, the three GJR models will be biased at time t whenever $r_{t-1} > 0$. If M5 has a slight bias, M6 and M7 have a larger bias.

Table 3.1 – The configuration of the models used in the simulation

	M1	M2	M3	M4	M5	M6	M7
Model	RM	G(1,1)	G(1,1)	G(1,1)	GJR(1,1)	GJR(1,1)	GJR(1,1)
α_1	-	0.07	0.05	0.25	0.05	0.01	0.15
β_1	-	0.88	0.90	0.70	0.88	0.90	0.70
γ	-	-	-	-	0.02	0.04	0.10

Notes: RM stands for RiskMetrics, G(1,1) stands for GARCH(1,1) and GJR(1,1) stands for GJR-GARCH(1,1). For the RiskMetrics model, $\lambda = 0.94$. Models M2-M7 have all the constant $\alpha_0 = 0.01$.

The evaluation of volatility performances is made in a pure-in-sample perspective. For all the sample period, the conditional variances and then the VaR measures are obtained and the performances of the models are evaluated by using a statistical and an economic approaches and the loss function approach in a VaR framework. With reference to the statistical approach, the MSE is used. Note that the MSE is a robust loss function, as discussed in Chapter 2. For model i, the MSE is equal to $\mathbb{E}[h_t^2 - \hat{h}_{t,i}^2]^2$, where h_t^2 comes from (3.13) and $\hat{h}_{t,i}^2$ from the formulation of the conditional variance as expressed by the model i. Because we aim to compare the MSE function within the Monte Carlo framework, the frequency at which each model has the smallest MSE is calculated. As regards to the economic approach, the Conditional Coverage (CC) test, illustrated in the Chapter 2 at page 81, is reported. Recall that the null hypothesis of the CC test jointly checks if the violations are independent distributed over time and the actual number of violations are coherent with the expected one. We calculate the frequency at which the CC test is not rejected. With reference to the loss function approach in a VaR, we calculate the global accuracy measure for each loss function. The results are showed in Table 3.2.

Table 3.2 – The results of the volatility evaluation

	M1	M2	M3	M4	M5	M6	M7
MSE	0.1050	0.2700	0.6250	0.0000	0.0000	0.0000	0.0000
CC test	0.8950	0.9550	0.8950	0.6050	0.8600	0.6550	0.2750
AM_M	0.3883	0.4567	0.5767	0.3667	0.3900	0.3800	0.3400
AM_A	0.3917	0.4533	0.5833	0.3700	0.3867	0.3800	0.3367

Notes: The first row shows the frequency at which the MSE is the smallest for each model. The second row shows the frequency at which the Conditional Coverage test is not rejected. The third and fourth rows show the global accuracy measures for the Magnitude and Asymmetric loss functions, respectively.

In the table it is clear that the statistical approach does not prefer any GJR models. Instead, it awards about 63% of times the model 3. The CC test does not bring any advantage to the analysis, given that four models are awarded: the models 1-3 and the model 5 reach the highest percentages of not rejections of the Conditional Coverage test. Interestingly, the loss function approach in a VaR framework awards model 3. This means that M3 has a same behaviour compared to the that of the DGP (first accuracy measure), its numerical scores are below the thresholds (second accuracy measure) and its numerical scores are the smallest among all the models (third accuracy measure). Moreover, the global accuracy measure when the asymmetric loss function is used reaches a slightly greater value than that of Magnitude loss function (0.5833 against 0.5767). In situations where more than one model have the same performance, this could be an advantage.

To conclude this section, we argue that the using of loss function in a VaR framework helps the research of the best model, when the economic approach does not lead to a clear decision. The results of this approach are consistent with those obtained from a robust statistical loss function. This opens the door to apply this method in an empirical analysis, as done in the next section.

3.5 Empirical Analysis

In this section we aim at investigating the use of the loss function in a VaR framework for the situations in which the traditional tests do not clearly identify the best model. Differently from the previous section, we use empirical data and the forecasting scheme is now updating by means of a rolling window, in a pure out-of-sample perspective. Because here there are not a set of replicates, the construction of the threshold is carried out only twice, once for each loss function. Once the threshold has been obtained when the $\widehat{VaR_t}$ is used, each model is evaluated simply looking at its exceeding of the threshold. Moreover, we study the ratio between the numerical score deriving from the usage of $\widehat{VaR_t}$ and the numerical score as produced by the volatility model. As in the Monte Carlo simulation, the MSE and the CC test are provided.

The data we use here have been already presented in Chapter 1. The dataset consists of the one-minute trade prices of the Capital One Financial Corporation (COF) stock. In this chapter, we only use the first sub-sample. In particular, from the whole sample the first 250 days have been excluded to initialize the algorithms such that the sample used for the analysis consists of 1445 days. This period starts on April 3, 1998 and ends on December 31, 2003. Figure 3.1 shows the realized volatility, expressed in percentage terms, obtained sampling the prices at 5 minutes. It can be seen that the considered periods grasps the peak of the Dot-com boom (March 2000), the burst and the its consequences. The highest peak of volatility is registered on October 7, 1998. Other peaks of volatility are registered on March, 2000 (burst of the speculative bubble), September, 2001 (attack to the twin towers) and July, 2002 (bankruptcy of WorldCom).

The set of competing models are illustrated in Table 3.3. The models are the same of those used in the Monte Carlo simulation but this time the parameters are estimated recursively by a rolling window of width 250. In order to make quicker the estimation, it has been considered a re-fitting

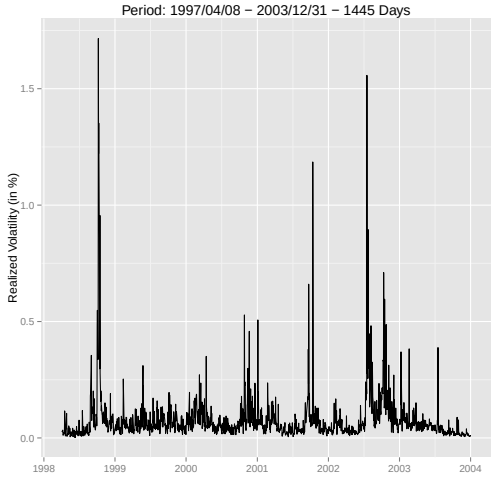


Figure 3.1 – COF realized volatility

period of 5 days. In total, there have been produced 1445 forecasts and 289 re-fitting periods. The different distributions specified for models 2-7 allow to take into account the stylized facts cited in Chapter 1. For instance, even though it has not been reported, the Normality test[3] on the daily returns strongly rejects the null hypothesis.

Table 3.3 – The configuration of the models used in the empirical part

	M1	M2	M3	M4	M5	M6	M7
Model	RM	G(1,1)	G(1,1)	G(1,1)	GJR(1,1)	GJR(1,1)	GJR(1,1)
$z_t \sim$	-	$\mathcal{N}(0,1)$	$t(v)$	$sk-t(v,\xi)$	$\mathcal{N}(0,1)$	$t(v)$	$sk-t(v,\xi)$

Notes: $t(v)$ and $sk-t(v,\xi)$ represent t and skewed-t distributions, respectively. v stands for the degrees of freedom and ξ for the skewness parameter.

The results of the volatility evaluation are showed in Table 3.4. If we only look at the MSE (statistical approach), the best model is M6, even though it exhibits more violations than expected, that is almost 6%

[3]The Jarque and Bera (1980) test.

against the 5%. The economic approach does not help at all, because the CC test is always rejected: no model clearly emerges. And here the $\widehat{VaR}$ and threshold come into the game. Rows 4-5 show the performances of the models when the Magnitude loss function is used. Model 7 is rejected because its numerical score lies above the threshold. Row 5 represents the ratio between the numerical score of each model and the numerical score when $\widehat{VaR}$ is used. The closer the ratio is to one, the better that model is, given that it has on average the same behaviour of the data generating process[4], in terms of number of violations and distance between the daily returns and the Value at Risk. And the Magnitude loss function awards M4. Unfortunately, there are too many models below the threshold. The situation becomes clearer if the Asymmetric loss function is used, as showed in rows 6-7. For this loss function, all the GJR models have numerical scores above the threshold. The choice shrinks to models 1-4 and looking at the ratio, also the Asymmetric loss function awards M4. Finally, we argue that, for the period considered and our mixed approach, M4 is the best model. This result does not represent a big surprise, because M4 is a GARCH(1,1) with $z_t \sim sk - t(v, \xi)$, distribution that is able to take into account the possible negative skewness and fat tails of the data. The innovation here is that in situations in which the statistical and economic approaches do not bring any significant help to find the best model, the loss function in a VaR framework does.

3.6 Conclusion

The evaluation of volatility forecasts by means of statistical or economic approaches may lead to non-unique conclusions. For this reason, the investigation of the opportunity to use a mixed approach considering the VaR measures evaluated by means of the loss functions has been the

[4]Whatever the DGP is, bootstrapping the increments allows to have a consistent estimator of the true Value at Risk.

Table 3.4 – Forecasts comparison

	M1	M2	M3	M4	M5	M6	M7
MSE	0.1411	0.1523	0.1464	0.1484	0.1346	0.1299	0.1351
$\hat{\alpha}$	0.0484	0.0457	0.0533	0.0526	0.0567	0.0595	0.0630
CC test	RH_0	RH_0	RH_0	RH_0	RH_0	RH_0	RH_0
$NS_M < TR_M$	Yes	Yes	Yes	Yes	Yes	Yes	No
$NS_M/NS_{T,M}$	0.9224	0.8697	1.0145	1.0013	1.0802	1.1329	1.1986
$NS_A < TR_A$	Yes	Yes	Yes	Yes	No	No	No
$NS_A/NS_{T,A}$	0.9258	0.8724	1.0197	1.0061	1.0865	1.1409	1.2083

Notes: The MSE is multiplied by 1000. $\hat{\alpha}$ is the frequency of violations. CC test is the conditional coverage test. $NS_M < TR$ is Yes if the numerical score of a model lies below the threshold. $NS_M/NS_{T,j}$ is the ratio between the numerical score of a model and the numerical score obtained with $\widehat{VaR}$, for $j = A, M$ if the loss function is the Magnitude or the Asymmetric loss function.

aim of this chapter. In this framework, the loss function evaluates the distance between the daily return and the VaR measure. A new asymmetric loss function, penalizing more the model with an actual number of violations larger than the expected ones, has been proposed. With reference to the term violation, we intend the occurrence that the daily return is greater than the VaR. Differently from the traditional methods normally used for calculating the VaR measures (the mean-variance, the Monte Carlo simulation and the Historical Simulation), the stationary bootstrap of the intraday returns has been chosen to have a consistent estimator of the VaR, in the spirit of Bowers and Heaton (2013). Then the VaR measures have been used to find the threshold discriminating low from high loss function values. A Monte Carlo experiment and an analysis carried out on a stock listed on the New York Stock Exchange have been the instruments in order to evaluate the possibility to use the loss function in a VaR framework.

In the Monte Carlo experiment, seven misspecified models belonging to the family of GARCH models have been evaluated from an in-sample

perspective. The data generating process (DGP) has been a GARCH(1,1) with Normal innovation. If the economic approach has not been able to recognize the model closer to the DGP, the statistical approach as well as the mixed approach have been capable of doing it. Based on the construction of a global accuracy measure, the loss function in a VaR framework has correctly signalled the model closer to the DGP.

As regards to the analysis carried on a stock listed on the New York Stock Exchange, the observed intraday returns have been used to estimate the daily VaR measures. Then the thresholds have been obtained and again the results of the statistical, economic and mixed approach have been explored. For the period considered, the GARCH(1,1) model, with innovations distributed as a skew-t distribution, has been awarded as the best model by the statistical and mixed approach. To conclude, using the loss functions in a VaR framework has been helpful in situations in which the traditional approaches do not clearly determine the best model. Moreover, the new Asymmetric loss function has appeared to have discriminated better the good from the bad models.

4

A Comparison of the Forecasting Performances of Multivariate Volatility Models

4.1 Introduction

In the previous chapter the evaluation of the forecasting performances concerned the univariate framework and the mixed approach. Here we focus on the multivariate context, evaluating empirically a set of competing models from a statistical and economic point of view. We aim to answer to the following questions: are there significant differences between the ranking of models using a statistical and economic loss function? If we compare models belonging to the MGARCH family to models that re-parametrize the realized covariance (like the CAW or the Rolling Covariance), do we always find that the latter forecast better?

The answers to these questions are given by a Monte Carlo experiment. In particular, the data generating process we set is a trivariate continuous-time stochastic process, where we assume that each instantaneous variance is the GARCH(1,1) diffusion as proposed by Andersen and Bollerslev (1998).

The chapter is organised as follows. Section 4.2 illustrates the methods practically used to take into account the non-normality of the daily

returns for the VaR estimation. Section 4.3 briefly overviews the volatility models used in this work. The setting of the Monte Carlo experiment is in Section 4.4 and the answers to our questions are in Section 4.5. Section 4.6 concludes.

4.2 Non-normality VaR estimations

To take into account the distribution of the portfolio returns, possibly different from the Gaussianity, the VaR forecasts used in this chapter are based on two different assumptions: the assumption of Normal distribution and the assumption of skewed and leptokurtic distribution of the returns. If the distribution of the returns is Normal, then f_α in (2.20) is replaced by the left α quantile of the standard Normal distribution. But, if the returns exhibit skewness and severe kurtosis, as it has well documented in literature (for details, see Cont (2001)), then the VaR forecasts based on the Normal distribution assumption could be misleading. In order to explicitly consider the skewness and kurtosis of the returns, two approaches giving different values to f_α are used: the modified VaR approach and the skewed Student's t distribution. The first is due to Favre and Galeano (2002) that proposed an alternative version of the VaR quantile, through the use of a Cornish Fisher expansion. This modified version quantifies the α quantile as follows:

$$z_{cf} = z_c + [(z_c^2 - 1)S]/6 + [(z_c^3 - 3z_c)K]/24 - [(2z_c^3 - 5z_c)S^2]/36, \quad (4.1)$$

where z_c is the α quantile of the Normal distribution, S and K are the skewness and the excess kurtosis of the daily returns, respectively. Note that if $S = K = 0$, as in the case of Normal distribution, then $z_{cf} = z_c$, and the modified VaR collapses to the standard VaR.

The other approach considers the standardized skewed Student's t distribution (in short, $sk - t$). Following Bauwens and Laurent (2002), assuming that the daily returns r_t are as usual formalized with $r_t = h_t z_t$,

where z_t is an $i.i.d.$ process with $E(z_t) = 0$ and $Var(z_t) = 1$ and h_t^2 representing the conditional variance, the excess of kurtosis and the skewness can be accommodated by using the (standardized) skewed-t distribution for z_t, formally $z_t \sim sk - t(0, 1, \xi, v)$. The log transformation of ξ measures the skewness: if $log(\xi) > 0$ the distribution is skew to the right and vice versa. The parameter v represents the degrees of freedom. This formalization is a generalization of the Student's t distribution: if $\xi = 1$, then $sk - t$ collapses to a standard Student's t distribution. Assuming that the conditional variance h_t^2 could be modelized by an univariate GARCH, such that $h_t^2 = \alpha_0 + \alpha_1 r_t^2 + \beta_1 h_{t-1}^2$, in the empirical part of this chapter we derive the quantile at $\alpha\%$ of the $sk - t$ distribution, after having obtained the unknown parameters ξ and v, by maximizing the following log likelihood:

$$\begin{aligned} \ell_t(\theta) \;=\; & \log\left(\frac{2}{\xi + \frac{1}{\xi}}\right) + \log\Gamma\left(\frac{v+1}{2}\right) - 0.5\pi(v-2) - \log\Gamma\left(\frac{v}{2}\right) + \\ & \log\frac{s}{h_t} - 0.5(1+v)\log\left[1 + \frac{(sz_t + m)^2 \xi^{-2J_t}}{v-2}\right], \end{aligned} \tag{4.2}$$

where $\theta = (\omega, a, b, \xi, v)$, $z_t = r_t/h_t$,

$$m = \frac{\Gamma\left(\frac{v-1}{2}\right)(v-2)^{0.5}}{\pi^{0.5}\Gamma\left(\frac{v}{2}\right)}\left(\xi - \frac{1}{\xi}\right),$$

$$s = \left[\left(\xi^2 + \frac{1}{\xi^2} - 1\right) - m^2\right]^{0.5},$$

and

$$J_t = \begin{cases} 1 & \text{if } z_t \geq -\frac{m}{s} \\ -1 & \text{if } z_t < -\frac{m}{s} \end{cases}.$$

4.3 The models for the volatility

In this section we briefly mention the set of competing models used in the Monte Carlo experiment, referring to the Chapter 1 for their discussion. We consider 9 specifications for the conditional covariance matrix that are frequently used in practice, that are: scalar, diagonal and full BEKK (the scalar BEKK model used in this chapter is the version with covariance targeting (Engle and Mezrich (1996)). Moreover, the DCC (Engle (2002)) and GOGARCH (van der Weide (2002)) models are considered. The univariate GARCH specifications for the conditional variance used in DCC and GOGARCH models are the GARCH, the GJR and the IGARCH formulations. These models have been chosen because most of them had been used in the work of Laurent et al. (2013). Table 4.1 provides the functional form for H_t for each model.

- insert Table 4.1 about here -

There have been considered also two models that parametrize the realized volatility. These models are the Rolling Covariance and the scalar and diagonal CAW. The Rolling Covariance, discussed in Chapter 1 at page 57, requires the estimation of only one parameter. This is done by maximizing the log likelihood in (1.45). Recall that its formulation is $H_t = exp(-\alpha)H_{t-1} + \alpha exp(-\alpha)RCV_{t-1}$, where α is the parameter to estimate. In this chapter three specifications of the Rolling Covariance model are assessed. The first uses the realized covariance sampling the prices at 5 minutes, the second sampling at 15 minutes and the third at 30 minutes. The scalar and diagonal CAW have been discussed in Chapter 1 at page 58. Here we use the formulation presented in that circumstance.

4.4 Monte Carlo experiment

In this section we illustrate the setting of the Monte Carlo experiment through which we investigate the performance of the competing models

with reference to the forecast accuracy, from a statistical and economic point of view. For sake of simplicity, we consider a portfolio only composed by three assets. Let $\mathbf{p}_t$ denote the trivariate vector of the log prices at time t, where t represents a day. As done in Chapter 1, we assume that $\mathbf{p}_t$ is driven by the following stochastic differential equation, where for brevity the drift term is omitted:

$$d\mathbf{p}(t) = \Sigma(t)d\mathbf{W}(t). \tag{4.3}$$

In (4.3), $\Sigma(t)$ is the spot co-volatility and $\boldsymbol{W}_t$ is the $k-$dimensional Brownian motion, assumed to be orthogonal to $\Sigma(t)$. The instantaneous volatility is collected in $\Theta(t) = \Sigma(t)\Sigma(t)'$, that in our case is:

$$\Theta(t) = \begin{pmatrix} \Theta_{11}(t) & \Theta_{12}(t) & \Theta_{13}(t) \\ \Theta_{21}(t) & \Theta_{22}(t) & \Theta_{23}(t) \\ \Theta_{31}(t) & \Theta_{32}(t) & \Theta_{33}(t) \end{pmatrix} = \begin{pmatrix} \sigma_1^2(t) & \sigma_{12}(t) & \sigma_{13}(t) \\ \sigma_{21}(t) & \sigma_2^2(t) & \sigma_{23}(t) \\ \sigma_{31}(t) & \sigma_{32}(t) & \sigma_3^2(t) \end{pmatrix},$$

where $\sigma_{ij}(t) = \sigma_i(t)\sigma_j(t)\rho_{ij}(t)$, with $i, j = 1, 2, 3$ and $i \neq j$. The model for $\sigma_i^2(t)$ is the GARCH(1,1) diffusion studied in Andersen and Bollerslev (1998): $d\sigma_i^2(t) = (\omega - \theta\sigma_i^2(t))dt + \lambda\sigma_i^2(t)db_i(t)$, where $b_i(t)$ is a standard Brownian motion independent of $\mathbf{W}(t)$. Following Dovonon et al. (2010), we set $\omega = 0.636$, $\theta = 0.035$ and $\lambda = 0.236$ for each i, that differs from each other for the initial point. Then, we set the instantaneous correlation $\rho_{ij} = (e^{2x(t)} - 1)/(e^{2x(t)} + 1)$, where x follows the GARCH diffusion: $dx(t) = (0.0192 - 0.03x(t))dt + 0.018x(t)db_1(t)$, where b_1 is a standard Brownian motion. To make different the three instantaneous correlations, we divide each of them by a random number sampled from an Uniform distribution. The solution for all the stochastic differential equations has been obtained by an Euler discretization method, based on a equally spaced time increments. If these time increments are small, then we can approximate $t_i - t_{i-1}$ with dt. We have chosen a time increments of 0.0001. For details on the stochastic differential equation solutions and Euler discretization scheme, see Iacus (2008). In this framework, the

trivariate vector $\mathbf{p}_t$ has been simulated 500 times for a period of two years (730 days). From the continuous time process (4.3) we drawn 288 observations per day, i.e. one observation each 5 minutes, for a total of 210,528 simulated observations (for each replicate). The forecasting sample, that is the period used to evaluate the forecast accuracy of the models, is of one year, in a pure in-sample perspective.

In this context, we have simulated three high-frequency prices following a data generating process with GARCH diffusion as variances and with time-varying instantaneous correlations. In our idea each simulated price should approximate that of an actual stock. For this reason, we do not consider all the simulated prices (288 per day) but only a subset, consisting of 78 observations at most per day. This is because we aim to approximate the prices' behaviour only during the standard trading day[1]. Then, we use these high-frequency data to obtain the volatility proxy (RCV_t) for different level of price aggregation. In other words, the volatility models are evaluated with respect to the RCV when its quality deteriorates: from 5 minutes (highest quality) to daily sampling frequency. All the simulations have been carried out using R 2.15.3.

4.5 Comparison of the multivariate forecasts

In the first part of this section, the results of the evaluation of the volatility forecasts by means of the statistical approach are showed. The loss functions used here have been illustrated in Table 2.2. We refer to page 74 for their description. Recall that the research questions are: (i) Is the ranking of the models the same if we use a statistical or an economic loss functions? (ii) Do the MGARCH models have a worse forecast accuracy than that of the rolling Covariance and CAW models? The answers to the questions are given by using the simulated data presented

[1]The trading day is the time span that a particular Stock Exchange is open. The time span is usually from 9:30 a.m. to 4:00 p.m., for a total of 390 minutes.

above, needed to have 500 replicates of the conditional covariance matrix, denoted as $H_{t,i}$ for the $i-th$ model (summarized in Table 4.1, with $i = 1, \cdots, 14$) and the 500 replicates of the volatility proxy, that is the realized covariance, for $t = 1, \cdots, 730$. The forecasting sample is of one year, in a pure in-sample perspective: all conditional covariance matrices are computed for the whole period and the last year has been used for the statistical and economic loss evaluation.

Let us start with the rankings produced by the statistical loss functions. The evaluation is performed with respect to the deterioration of the volatility proxy, obtained sampling the intraday returns at lower frequencies. It is well known that, in absence of micro-structure frictions, the higher the frequency is, the better the proxy is (Bandi and Russell (2004)). For instance, using the data at 30 minutes will produce a realized covariance less noisy than using the data at 300 minutes. Having simulated the data by ourselves, we do not take in consideration the problem of the micro-structure frictions.

Symmetric loss functions

The results of the symmetric loss functions are summarized in Tables 4.2 and 4.3. Here the loss functions are the Euclidean distance and the squared weighted Euclidean distance. The former considers the distance between each element of the volatility proxy matrix and $H_{t,i}$ while the latter considers only the diagonal elements, excluding the covariance entries. We do not find any significant difference between the two loss functions: the ranking of the models is almost the same. This is coherent with the literature that states the larger importance of variances with respect to the covariances. For the Euclidean distance loss function, it results that when the volatility proxy is computed with the 5 minutes sampling frequency, the scalar CAW ranks first about 54% of the times and the Rolling Covariance sampled at 5 minutes 41%. When the quality of the proxy deteriorates, first the Rolling Covariance at 5 minutes and then at

30 minutes emerge.

- insert Table 4.2 about here -

The same ranking (with different frequencies) can be observed for the squared weighted Euclidean distance. Recall that this loss function does not consider the covariance terms. We argue that, as stated in literature, the covariances do not bring a significant advantage with respect to the ranking produced by the loss function considering all the entries of the covariance matrix. For instance, the scalar CAW ranks first about 46% of the times, a sensible lower percentage with respect to that of Euclidean distance loss function.

- insert Table 4.3 about here -

Asymmetric loss functions

The frequencies at which each model is ranked first when the asymmetric loss functions are used are presented in Tables 4.4 and 4.5. In the first table the results for the penalizing over predictions Mahalanobis distance are reported, while in second there are those of the penalizing under predictions loss function. For the former loss function, the Rolling Covariance at 30 minutes ranks first when the sampling frequency drops from 10 minutes to 390 minutes (i.e. the daily frequency). Only when the volatility proxy is computed with 5 minutes frequency, the Rolling Covariance at 15 minutes ranks first about 50% of the times. We can state that the Rolling Covariance model seldom produces forecasts larger than the volatility proxy. This is an important result in a portfolio management optical. Not surprisingly, when the under prediction version of the Mahalanobis distance is used, the ranking of the model has different patterns. For a good quality of the realized covariance, the best model is the scalar CAW: when the realized covariance is obtained by using 5 minutes frequency, this model ranks first about 32% of the cases. Then,

a multivariate GARCH model, for the first time, emerges: the iGOGARCH, that is the closest to the volatility proxy 25% of the times, for the realized covariance at 20 minutes. Finally, as seen also for the other loss functions, the Rolling Covariance at 30 minutes, when the quality of the proxy is low, always ranks first.

- insert Tables 4.4 and 4.5 about here -

To sum up, when the statistical loss functions are used, the results award the models that directly use the realized covariance, as expected, excluding the too-parametrized diagonal CAW, which rarely ranks first. We can state that the statistical loss functions reward the model with less parameters to estimate, given that the Rolling Covariance and the scalar CAW have only one unknown parameter. Moreover, the frequently used multivariate GARCH models never rank first, except for the iGOGARCH model when used in some circumstances. After having reported the results of the statistical loss function, the next step is to check if the same ranking is obtained when the economic loss functions are used.

Economic approach

The Christoffersen test is the method used to indirectly rank the set of forecasting models within the economic approach. The sample period used for such evaluation is of one year, as done for the statistical evaluation. The solution for applying the CC test in the multivariate framework is to assign equal weights to the stocks in order to transform the multivariate problem in an univariate one. Once obtained the conditional portfolio returns and variance, the VaR measures are estimated. These are calculated by means of the mean-variance approach, even though they could be obtained by the bootstrap procedure discussed in the previous chapter. In the spirit of Bauwens and Laurent (2002), we assign 3 different vectors of constant weights to the daily returns and daily conditional covariance matrix $H_{m,t}$, in order to obtain the portfolio mean and variance:

$w_1 = (1/3, 1/3, 1/3)$, $w_2 = (0.5, 0.2, 0.3)$ and $w_3 = (1.4, -0.2, -0.2)$.

The Christoffersen test for all the models, with the one day ahead VaR for $\alpha = 0.05$ and the quantile for the long position obtained by means of the normal distribution, the Cornish Fisher expansion and the skewed Student's t distribution is reported in Table 4.6.

- insert Table 4.6 about here -

Looking at the table, the smaller the frequency is, the better that model is, because each frequency indicates how many times the test has been rejected over the 500 replicates. First of all, we note the differences between the frequencies when the VaR is computed with the normal distribution or the Cornish Fisher expansion and the frequencies when the VaR is computed with the skewed Student's t distribution. These latter are smaller even though we do not work with real financial data that suffer from the kurtosis excess and skewness. Secondly, the rejections of the Christoffersen test are not only few but are also similar in the number among the models: when using the scalar CAW, the vector weights w_1 and the normal distribution, these rejections are about 6% (smallest value) against about 10% of the worse model, that is the DCC. Using the weights w_1, the best model is the GJR-GOGARCH, for all the specifications of the quantile. Instead, using the weights w_3, the best models are the diagonal CAW and Rolling Covariance. We can state that there is not too much correspondence between the ranking of the statistical loss functions and that of the economic loss function. In fact, the GJR-GOGARCH model and the diagonal CAW never rank first when the statistical loss functions are used. Instead now these models yield to best economic performances. The issue is worth further consideration. Our idea is to look at the average of the VaR violations among the models, where for *average* we refer to the average number of VaR violations for all the replicates. Intuitively, for each replicate, we should have a value close to 0.05. The aim of taking the averages among all the replications is to approximate the whole

behaviour of the model, independently of the each single replicate. The averages of the VaR violations for all the replicates are reported in Table 4.7.

- insert Table 4.7 about here -

The highest average of VaR violations is of the GOGARCH model, for all the univariate specifications. Excluding the GOGARCH model, all the averages are closer to the expected values 0.05. Combining this information with those furnished in Table 4.6, we can state that when the economic loss function is used alone there is no model that clearly emerges as the "best" model. This is because neither changing the weights nor the distributional function for the returns, we find the same ranking of the statistical function. In this sense, the economic function used here yields results that diverge from those of the statistical loss function. As already seen in the previous chapter, the analysis of the volatility models exclusively based on the economic approach could lead to misleading conclusions.

4.6 Conclusion

The research questions of this chapter have been: (i) what about the differences between the ranking of a set of competing multivariate models when a statistical and economic approaches are used? (ii) Is the forecast accuracy of the MGARCH models similar to that of the models that directly use the realized covariance? The answers have been obtained by using a Monte Carlo experiment replicating 500 times a trivariate continuous-time stochastic process. Hence, fourteen models have been compared from a statistical and economic point of view, using a forecasting sample of one year. The forecasts have been only made from an in-sample perspective. The statistical loss functions taken in consideration in this chapter have been: (i) the matrix version of the Mean Squared

Error function, named Euclidean distance; (ii) the squared weighted Euclidean distance, considering as input only the variances; (iii) the Mahalanobis distance penalizing the over predictions; (iv) the Mahalanobis distance penalizing the under predictions. The economic loss function has used the Value at risk methodology. In particular the results of the Christoffersen test have been studied. The Christoffersen jointly tests if the number of violations are coherent with the expected number of violations and if the violations are not clustered in time (i.e. independence hypothesis). Given the VaR is sensible to the underlying assumption on the distribution of returns, three methods to calculate the VaR have been employed: the standard method considering the Normal distribution of the returns, the Corner Fisher expansion and the skewed Student's t distribution.

The answers to the research questions have been: first, a clear correspondence between the rankings resulting from the statistical and economic loss functions has not appeared. A portfolio manager that had only used economic criteria for his decision would have preferred models not exactly closer to the volatility proxy. The statistical based ranking, when the volatility proxy is good, in the sense that it is based on high frequency, has awarded the scalar CAW and the rolling Covariance models, for the symmetric and asymmetric loss functions, respectively. When the quality of the proxy deteriorates, meaning that the realized covariance is computed using low frequencies, up to the use of daily returns, the statistical approach has always ranked first the rolling Covariance model. Instead the economic loss function methodology used here has resulted to be very sensible to the choice of the distribution and the weights such that no model clearly has emerged. Second, looking at the statistical loss function, the MGARCH models have yielded worse forecast accuracy than that of the realized covariance based models, if these latter are parsimonious. Moreover, the impact of the covariances for the rankings has seemed to be irrelevant, given the ranking based only on variances has

been the same of the ranking based on both variances and covariances. Finally, if a portfolio manager is interested in studying the models such that the over predictions are rare, the Rolling Covariance at 30 minutes has resulted to be the best model when the negative forecast errors are more penalized. A negative forecast error is present when the forecasted value is larger than the correspondent value of the volatility proxy.

Appendix

Table 4.1 – Forecasting models functional forms

Model	Multivariate GARCH models	Parameters
sBEKK(1,1)	$H_t = (1-A-B)^*\overline{H} + A\mathbf{r}_{t-1}\mathbf{r}_{t-1}^{'}A + GH_{t-1}G$	2
dBEKK(1,1)	$H_t = CC^{'} + A\mathbf{r}_{t-1}\mathbf{r}_{t-1}^{'}A + GH_{t-1}G$	$k(k-1)+$ $2k$
BEKK(1,1)	$H_t = CC^{'} + A\mathbf{r}_{t-1}\mathbf{r}_{t-1}^{'}A^{'} + GH_{t-1}G^{'}$	$5(5k+1)/2$
DCC(1,1)	$H_t = D_tR_tD_t$ $R_t = diag(q_{11,t}^{-1/2},\cdots,q_{kk,t}^{-1/2})Q_tdiag(q_{11,t}^{-1/2},\cdots,q_{kk,t}^{-1/2})$ $D_t = diag(h_{11,t}^{1/2},\cdots,h_{kk,t}^{1/2})$ $u_t = D_t^{-1}\mathbf{r}_t$ $Q_t = (1-\alpha-\beta)\overline{Q} + \alpha u_t u_t^{'} + \beta Q_{t-1}$	$2+k+$ $univ$
GOG(1,1)	$V^{-1/2}\mathbf{r}_t = Lf_t$ $H_t = V^{-1/2}LZ_tLV^{-1/2}$ $Z_t = diag(\sigma_{f_{1t}}^2,\cdots,\sigma_{f_{kt}}^2)$ $L = P\Lambda^{1/2}U, U = \prod_{i<j}R_{i,j}(\delta_{i,j}), -\pi \leq \delta_{i,j} \leq \pi$	$k(k-1)/2+$ $univ$
Univariate GARCH models in D_t and Z_t ($l = 1,\cdots,k$)		
GARCH(1,1)	$h_{l,t}^2 = \alpha_{l,0} + \alpha_{l,1}r_{l,t-1}^2 + \beta_{l,1}h_{l,t-1}^2$	$3k$
GJR(1,1)	$h_{l,t}^2 = \alpha_{l,0} + (\alpha_{l,1} + \gamma_l I_{(r_{l,t-1<0})})(r_{l,t-1}^2) + \beta_{l,1}h_{l,t-1}^2$ $I_{(r_{l,t-1<0})} = 1$ if $r_{l,t-1} < 0$	$4k$
IGARCH(1,1)	$h_{l,t}^2 = \alpha_{l,0} + \alpha_{l,1}r_{l,t-1}^2 + \beta_{l,1}h_{l,t-1}^2$ $\alpha_{l,1} + \beta_{l,1} = 1, \forall l$	$3k$
Model	Realized Covariance based models	Parameters
Rolling Cov.	$H_t = exp(-\alpha)H_{t-1} + \alpha exp(-\alpha)RCV_{t-1}$	1
sCAW	$H_t = (1-a^2-b^2) * \overline{RCV} + a^2RCV_t + b^2H_{t-1}$	2
dCAW	$H_t = \overline{RCV} - A\overline{RCV}A^{'} - B\overline{RCV}B^{'} +$ $ARCV_tA^{'} + BH_{t-1}B^{'}$	$2k$

Notes: sBEKK: scalar BEKK; dBEKK: diagonal BEKK; GOG: GOGARCH; Rolling Cov.: Rolling Covariance; sCAW: scalar CAW; dCAW: diagonal CAW.

Table 4.2 – Frequencies at which each model exhibits the smallest loss. *Symmetric loss function: Euclidean distance*

	sBEKK	dBEKK	BEKK	DCC	gjrDCC	iDCC	GOG	gjrGOG	iGOG	RC_{5m}	RC_{15m}	RC_{30m}	sCAW	dCAW
5 min	0.000	0.000	0.000	0.002	0.000	0.004	0.002	0.000	0.004	0.410	0.000	0.000	0.540	0.038
10 min	0.000	0.000	0.000	0.004	0.000	0.000	0.006	0.000	0.000	0.492	0.048	0.000	0.426	0.024
15 min	0.000	0.002	0.000	0.012	0.002	0.000	0.006	0.002	0.006	0.422	0.236	0.016	0.284	0.012
20 min	0.000	0.002	0.002	0.008	0.002	0.002	0.008	0.002	0.010	0.436	0.066	0.000	0.422	0.040
30 min	0.002	0.000	0.000	0.036	0.012	0.000	0.010	0.002	0.000	0.114	0.352	0.378	0.096	0.004
40 min	0.006	0.006	0.002	0.038	0.018	0.002	0.006	0.008	0.000	0.118	0.326	0.356	0.112	0.002
50 min	0.012	0.002	0.002	0.044	0.018	0.000	0.002	0.006	0.000	0.048	0.160	0.638	0.078	0.000
60 min	0.004	0.006	0.002	0.050	0.024	0.000	0.014	0.026	0.000	0.122	0.300	0.340	0.114	0.004
120 min	0.010	0.016	0.020	0.054	0.052	0.004	0.012	0.060	0.002	0.110	0.226	0.306	0.118	0.014
180 min	0.084	0.018	0.000	0.034	0.052	0.000	0.018	0.038	0.020	0.016	0.068	0.606	0.052	0.002
240 min	0.086	0.018	0.000	0.038	0.056	0.000	0.018	0.040	0.020	0.014	0.064	0.600	0.048	0.002
300 min	0.092	0.018	0.000	0.042	0.056	0.000	0.022	0.042	0.024	0.014	0.066	0.578	0.048	0.002
360 min	0.086	0.018	0.000	0.038	0.056	0.000	0.018	0.040	0.020	0.014	0.064	0.600	0.048	0.002
Daily	0.076	0.014	0.000	0.036	0.038	0.000	0.070	0.088	0.062	0.014	0.058	0.502	0.046	0.000

Table 4.3 – Frequencies at which each model exhibits the smallest loss. *Symmetric loss function. Squared weighted Euclidean distance*

	sBEKK	dBEKK	BEKK	DCC	gjrDCC	iDCC	GOG	gjrGOG	iGOG	RC_{5m}	RC_{15m}	RC_{30m}	sCAW	dCAW
5 min	0.000	0.000	0.000	0.016	0.002	0.002	0.010	0.000	0.012	0.448	0.008	0.004	0.460	0.046
10 min	0.000	0.004	0.002	0.016	0.000	0.000	0.012	0.000	0.006	0.552	0.054	0.004	0.322	0.032
15 min	0.000	0.006	0.000	0.014	0.006	0.004	0.012	0.002	0.014	0.444	0.242	0.028	0.218	0.014
20 min	0.000	0.006	0.002	0.016	0.006	0.008	0.016	0.000	0.018	0.468	0.070	0.008	0.346	0.040
30 min	0.010	0.002	0.000	0.056	0.012	0.002	0.012	0.002	0.000	0.088	0.312	0.430	0.078	0.002
40 min	0.014	0.004	0.000	0.044	0.016	0.004	0.014	0.010	0.000	0.102	0.310	0.394	0.092	0.002
50 min	0.024	0.006	0.002	0.046	0.026	0.002	0.004	0.008	0.000	0.030	0.134	0.654	0.068	0.000
60 min	0.008	0.008	0.008	0.048	0.028	0.002	0.008	0.018	0.004	0.106	0.308	0.376	0.084	0.000
120 min	0.016	0.022	0.020	0.050	0.048	0.006	0.006	0.050	0.006	0.106	0.236	0.330	0.100	0.012
180 min	0.070	0.012	0.000	0.034	0.038	0.002	0.028	0.072	0.032	0.016	0.060	0.596	0.042	0.002
240 min	0.066	0.014	0.002	0.034	0.052	0.002	0.034	0.060	0.030	0.016	0.066	0.586	0.040	0.002
300 min	0.078	0.014	0.000	0.034	0.044	0.002	0.034	0.076	0.034	0.016	0.062	0.570	0.038	0.002
360 min	0.066	0.014	0.002	0.034	0.052	0.002	0.034	0.060	0.030	0.016	0.066	0.586	0.040	0.002
Daily	0.066	0.012	0.002	0.036	0.036	0.002	0.070	0.100	0.076	0.016	0.050	0.506	0.032	0.000

Table 4.4 – Frequencies at which each model exhibits the smallest loss. *Asymmetric loss function. Mahalanobis distance, over prediction version*

	sBEKK	dBEKK	BEKK	DCC	gjrDCC	iDCC	GOG	gjrGOG	iGOG	RC_{5m}	RC_{15m}	RC_{30m}	sCAW	dCAW
5 min	0.002	0.002	0.002	0.016	0.002	0.002	0.018	0.000	0.000	0.160	0.498	0.186	0.112	0.004
10 min	0.006	0.006	0.000	0.036	0.008	0.000	0.010	0.000	0.002	0.044	0.138	0.668	0.082	0.004
15 min	0.022	0.006	0.000	0.044	0.012	0.000	0.004	0.006	0.000	0.034	0.094	0.708	0.074	0.002
20 min	0.026	0.010	0.000	0.044	0.010	0.000	0.010	0.002	0.000	0.030	0.104	0.694	0.074	0.002
30 min	0.056	0.018	0.000	0.054	0.026	0.000	0.000	0.006	0.000	0.022	0.086	0.678	0.060	0.002
40 min	0.062	0.020	0.000	0.052	0.030	0.000	0.006	0.008	0.000	0.018	0.082	0.664	0.064	0.002
50 min	0.078	0.018	0.000	0.048	0.036	0.000	0.006	0.008	0.002	0.018	0.082	0.652	0.060	0.000
60 min	0.074	0.018	0.000	0.052	0.040	0.000	0.006	0.008	0.000	0.018	0.080	0.658	0.056	0.002
120 min	0.080	0.020	0.000	0.046	0.060	0.000	0.008	0.008	0.006	0.018	0.074	0.638	0.054	0.000
180 min	0.066	0.016	0.000	0.036	0.036	0.000	0.076	0.094	0.078	0.014	0.060	0.482	0.046	0.000
240 min	0.070	0.016	0.000	0.034	0.038	0.000	0.080	0.094	0.078	0.014	0.060	0.478	0.042	0.000
300 min	0.072	0.016	0.000	0.038	0.036	0.000	0.068	0.102	0.080	0.014	0.060	0.474	0.044	0.000
360 min	0.070	0.016	0.000	0.034	0.038	0.000	0.080	0.094	0.078	0.014	0.060	0.478	0.042	0.000
Daily	0.074	0.014	0.000	0.038	0.034	0.000	0.064	0.108	0.080	0.014	0.062	0.474	0.042	0.000

Table 4.5 – Frequencies at which each model exhibits the smallest loss. *Asymmetric loss function. Mahalanobis distance, under prediction version*

	sBEKK	dBEKK	BEKK	DCC	gjrDCC	iDCC	GOG	gjrGOG	iGOG	RC_{5m}	RC_{15m}	RC_{30m}	sCAW	dCAW
5 min	0.012	0.038	0.008	0.064	0.010	0.058	0.066	0.004	0.162	0.102	0.004	0.012	0.326	0.142
10 min	0.012	0.052	0.014	0.052	0.014	0.074	0.060	0.010	0.204	0.076	0.006	0.008	0.270	0.156
15 min	0.018	0.072	0.014	0.058	0.022	0.072	0.072	0.004	0.188	0.088	0.006	0.010	0.246	0.138
20 min	0.038	0.090	0.026	0.038	0.022	0.096	0.044	0.016	0.250	0.062	0.008	0.008	0.164	0.146
30 min	0.014	0.082	0.016	0.058	0.022	0.056	0.066	0.018	0.170	0.114	0.006	0.008	0.264	0.114
40 min	0.028	0.104	0.038	0.040	0.032	0.072	0.038	0.024	0.220	0.072	0.008	0.010	0.208	0.118
50 min	0.010	0.072	0.032	0.036	0.042	0.052	0.062	0.024	0.166	0.130	0.006	0.014	0.262	0.096
60 min	0.050	0.132	0.068	0.016	0.026	0.068	0.040	0.046	0.220	0.048	0.004	0.012	0.156	0.122
120 min	0.090	0.144	0.154	0.008	0.026	0.062	0.012	0.086	0.198	0.020	0.002	0.002	0.082	0.118
180 min	0.026	0.014	0.000	0.048	0.042	0.000	0.002	0.018	0.000	0.020	0.110	0.656	0.060	0.004
240 min	0.022	0.026	0.068	0.050	0.108	0.002	0.008	0.052	0.000	0.048	0.120	0.434	0.058	0.008
300 min	0.032	0.024	0.034	0.040	0.094	0.000	0.002	0.048	0.000	0.038	0.106	0.534	0.042	0.010
360 min	0.022	0.026	0.068	0.050	0.108	0.002	0.008	0.052	0.000	0.048	0.120	0.434	0.058	0.008
Daily	0.064	0.018	0.002	0.044	0.048	0.000	0.010	0.024	0.006	0.016	0.076	0.642	0.054	0.000

Table 4.6 – Frequencies at which the Christoffersen Test is rejected

Weights	$(1/3, 1/3, 1/3)$			$(.5, .2, .3)$			$(.1.4, -.2, -.2)$		
	VaR	Mod. VaR	Skew VaR	VaR	Mod. VaR	Skew VaR	VaR	Mod.VaR	Skew VaR
sBEKK	0.082	0.090	0.058	0.082	0.092	0.048	0.072	0.090	0.050
dBEKK	0.096	0.086	0.048	0.094	0.088	0.048	0.094	0.104	0.058
BEKK	0.084	0.096	0.056	0.070	0.100	0.052	0.104	0.098	0.066
DCC	0.108	0.102	0.066	0.088	0.100	0.064	0.072	0.086	0.070
gjrDCC	0.100	0.112	0.066	0.090	0.100	0.046	0.070	0.082	0.068
iDCC	0.090	0.096	0.058	0.088	0.096	0.060	0.078	0.086	0.086
GOG	0.072	0.076	0.042	0.070	0.068	0.042	0.056	0.078	0.078
gjrGOG	0.074	0.084	0.042	0.062	0.062	0.032	0.070	0.092	0.056
iGOG	0.068	0.070	0.040	0.074	0.070	0.042	0.060	0.072	0.037
RC_{5m}	0.082	0.096	0.064	0.090	0.068	0.050	0.062	0.082	0.036
RC_{15m}	0.086	0.090	0.048	0.088	0.090	0.048	0.074	0.088	0.046
RC_{30m}	0.076	0.078	0.044	0.092	0.080	0.046	0.062	0.074	0.034
sCAW	0.084	0.092	0.062	0.082	0.076	0.046	0.072	0.084	0.044
dCAW	0.062	0.080	0.062	0.088	0.070	0.058	0.050	0.056	0.042

Table 4.7 – Averages of the VaR violations for $\alpha = 0.05$

Weights	$(1/3, 1/3, 1/3)$			$(.5, .2, .3)$			$(.1.4, -.2, -.2)$		
	VaR	Mod. VaR	Skew VaR	VaR	Mod. VaR	Skew VaR	VaR	Mod.VaR	Skew VaR
sBEKK	0.050	0.051	0.047	0.051	0.051	0.047	0.050	0.050	0.046
dBEKK	0.049	0.049	0.046	0.049	0.049	0.046	0.049	0.049	0.046
BEKK	0.048	0.048	0.048	0.049	0.049	0.046	0.048	0.049	0.045
DCC	0.051	0.051	0.048	0.051	0.051	0.047	0.051	0.051	0.047
GJR_DCC	0.050	0.050	0.047	0.050	0.050	0.047	0.051	0.051	0.047
iDCC	0.049	0.049	0.046	0.049	0.049	0.046	0.049	0.049	0.045
GOG	0.079	0.079	0.076	0.077	0.077	0.074	0.072	0.072	0.068
gjrGOG	0.079	0.079	0.076	0.076	0.076	0.073	0.070	0.070	0.067
iGOG	0.079	0.079	0.076	0.075	0.076	0.073	0.068	0.068	0.064
RC_{5m}	0.052	0.052	0.048	0.052	0.052	0.048	0.052	0.052	0.048
RC_{15m}	0.053	0.053	0.050	0.053	0.053	0.050	0.053	0.053	0.049
RC_{30m}	0.055	0.055	0.052	0.055	0.055	0.052	0.055	0.055	0.051
sCAW	0.051	0.051	0.048	0.051	0.051	0.048	0.051	0.051	0.047
dCAW	0.048	0.048	0.048	0.048	0.048	0.045	0.043	0.044	0.040

Conclusion

In the Introduction four research questions have been proposed. In Chapter 3 the answers to the first two questions have been given while the last two questions have been answered in Chapter 4. In this chapter we summarize each answer. Afterwards, some suggestions for future researches are proposed.

The first research question has been: "Is it possible to use the loss function in a VaR framework in order to evaluate the volatility predictions of a set of competing models? Does this approach bring an advantage when the statistical and the economic approaches fail to recognize the best model?". The answers have been positive to both the two sub-questions. It frequently happens that the economic approach used to evaluate the volatility prediction fails to recognize the best model. In such circumstance, the merged approach considering the loss function in a VaR framework, has represented a valid help. Moreover, this approach is based on a VaR estimator, as proposed by Bowers and Heaton (2013), that is completely data-driven: it does not depend on any pre-assumed distributional form of the daily returns. In addition, we have proposed an asymmetric loss function evaluating the daily returns to the VaR mea-

sures. The term asymmetric means that models with a greater number of violations than the expected one are more penalized. It has emerged that the asymmetric loss function is able to slightly discriminate better the volatility models than the Magnitude loss function, proposed by Lopez (1998).

The second research question was: "Is it possible to find a threshold that discriminates low from high loss function values in order to evaluate the performances of the volatility models?". Again, the answer has been positive: the knowledge of a threshold, based on the VaR estimator discussed previously, has helped the discrimination of the models when these are evaluated by the loss functions in a VaR framework. The construction of the threshold as well as the VaR estimator have been based on the stationary bootstrap of Politis and Romano (1994).

The third and the fourth research questions have concerned the multivariate context. The third has been: "Is the ranking of the competing models the same if a statistical and an economic loss functions are used?". The answer has been negative, as showed in Chapter 4. If the economic loss function methodology is highly sensible to the distribution of the returns such that no model has clearly emerged, the statistical loss function has awarded the models re-parametrizing the realized volatility, as expected. More specifically, for both the symmetric and the asymmetric statistical loss functions the scalar CAW and the Rolling Covariance models have emerged, when the volatility proxy quality is high. When the quality of the volatility proxy decreases, the statistical loss function has always rewarded the Rolling Covariance. This means that in the Monte Carlo simulation used to compare fourteen models, the models based on the multivariate GARCH models have never ranked first. And this answers also the fourth research question: "Do the multivariate GARCH models have a worse forecast accuracy than that of the models that use the realized volatility to forecast $\hat{h}$?". Interestingly, the covariances in the evaluation of the forecast accuracy has seemed to have low importance.

The ranking given by the loss function using only the variances has been the same of the ranking given by the loss function using both variances and covariances.

For future research, many questions remain open. First of all, it would be interesting to check if the set of models that are below the threshold coincides with the set of models within the Model Confidence Set of Hansen et al. (2011). Then, the set of economic loss functions could be broadened. A natural choice could be using the Dynamic Quantile (DQ) test, proposed by Engle and Manganelli (2004), that verifies if the probability of getting a VaR violation at time $t + 1$ is independent of any variable observed at time t. An observed variable could be the contemporaneous or lagged VaR estimate of the model that uses the realized covariance at different sampling frequencies, instead of the conditional covariance matrix resulting from a multivariate GARCH model, in order to have a direct link between the statistical loss function and the economic loss function. Moreover, what could be further explored are the reasons of the different ranking between statistical and loss functions.

Bibliography

Aït-Sahalia, Y., J. Fan, and D. Xiu (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association 105*(492), 1504–1517.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies 18*(2), 351–416.

Alexander, C. (2000). A primer on the orthogonal garch model. *manuscript ISMA Centre, University of Reading, UK*.

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review 39*(4), 885–905.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2002). Parametric and nonparametric volatility measurement. NBER Technical Working Papers 0279, National Bureau of Economic Research, Inc.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (second ed.). New York, NY: Wiley.

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

Bandi, F. M. and J. R. Russell (2004). Microstructure noise, realized volatility, and optimal sampling. Econometric Society 2004 Latin American Meetings 220, Econometric Society.

Bandi, F. M. and J. R. Russell (2006). Separating microstructure noise from volatility. *Journal of Financial Economics 79*(3), 655–692.

Barndorff-Nielsen, O. E. and Shephard (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B 64*(2), 253–280.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925.

Bauwens, L. and S. Laurent (2002). A new class of multivariate skew densities, with application to garch models. CORE Discussion Papers 2002020, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Bauwens, L., S. Laurent, and J. V. Rombouts (2006). Multivariate garch models: a survey. *Journal of applied econometrics 21*(1), 79–109.

Bauwens, L., G. Storti, and F. Violante (2012). Dynamic conditional correlation models for realized covariance matrices. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Berument, H. and H. Kiymaz (2001). The day of the week effect on stock market volatility. *Journal of economics and finance 25*(2), 181–193.

Billio, M., M. Caporin, and M. Gobbo (2006). Flexible dynamic conditional correlation multivariate garch models for asset allocation. *Applied Financial Economics Letters 2*(02), 123–130.

Black, F. (1976). The pricing of commodity contracts. *Journal of financial economics 3*(1), 167–179.

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

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The Review of Economics and Statistics*, 498–505.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy 96*(1), 116–31.

Bollerslev, T. and J. M. Wooldridge (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews 11*(2), 143–172.

Bowers, C. and C. Heaton (2013). Bootstrapping daily returns. *Available at SSRN 2361393*.

Casella, G. and R. L. Berger (1990). *Statistical inference*, Volume 70. Duxbury Press Belmont, CA.

Chan, L. K., J. Karceski, and J. Lakonishok (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. *Review of Financial Studies 12*(5), 937–974.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics 11*(2), 121–135.

Chen, Y.-S. and S.-T. Wang (2004). The empirical evidence of the leverage effect on volatility in international bulk shipping market. *Maritime Policy & Management 31*(2), 109–124.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–62.

Clark, T. and M. McCracken (2009). Nested forecast model comparisons: a new approach to testing equal accuracy. *FRB of St. Louis Working Paper No*.

Clark, T. E. and M. W. McCracken (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics 105*(1), 85–110.

Clark, T. E. and M. W. McCracken (2011). Advances in forecast evaluation. *Federal Reserve Bank of St. Louis Working Paper Series*.

Comte, F. and O. Lieberman (2003). Asymptotic theory for multivariate garch processes. *Journal of Multivariate Analysis 84*(1), 61–84.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*(2), 223–236.

Corradi, V. and N. R. Swanson (2011). The white reality check and some of its recent extensions.

de Gooijer, J. G., B. Abraham, A. Gould, and L. Robinson (1985). Methods for determining the order of an autoregressive-moving average process: A survey. *International Statistical Review/Revue Internationale de Statistique*, 301–329.

de Pooter, M., M. Martens, and D. van Dijk (2008). Predicting the daily covariance matrix for s&p 100 stocks using intraday data - but which frequency to use? *Econometric Reviews 27*(1-3), 199–229.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3).

Diebold, F. X. and M. Nerlove (1989). The dynamics of exchange rate volatility: a multivariate latent factor arch model. *Journal of Applied Econometrics 4*(1), 1–21.

Dovonon, P., S. Goncalves, and N. Meddahi (2010). Bootstrapping realized multivariate volatility measures. MPRA Paper 40123, University Library of Munich, Germany.

Duffie, D. and J. Pan (1997). An overview of value at risk. Technical report, Stanford University.

Dumas, B., J. Fleming, and R. E. Whaley (1998). Implied volatility functions: Empirical tests. *The Journal of Finance 53*(6), 2059–2106.

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

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1–26.

Elliott, G., I. Komunjer, and A. G. Timmermann (2003). *Estimating loss function parameters*. Number 3821. Centre for Economic Policy Research.

Elliott, G. and A. G. Timmermann (2007). Economic forecasting. CEPR Discussion Papers 6158, C.E.P.R. Discussion Papers.

Engle, R. and J. Mezrich (1996). Garch for groups. *Risk 9*, 36–40.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*(4), 987–1007.

Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics 20*(3), 339–50.

Engle, R. F. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric Reviews 5*(1), 1–50.

Engle, R. F. and K. F. Kroner (1995). Multivariate simultaneous generalized arch. *Econometric Theory 11*(01), 122–150.

Engle, R. F. and S. Manganelli (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*, 367–381.

Engle, R. F., V. K. Ng, and M. Rothschild (1990). Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics 45*(1), 213–237.

Engle, R. F. and K. Sheppard (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate garch. NBER Working Papers 8554, National Bureau of Economic Research, Inc.

Favre, L. and J. Galeano (2002). Mean-modified value-at-risk optimization with hedge funds. *Journal of Alternative Investmen 5*, 21–25.

Fleming, J., C. Kirby, and B. Ostdiek (2001). The economic value of volatility timing. *The Journal of Finance 56*(1), 329–352.

Fleming, J., C. Kirby, and B. Ostdiek (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics 67*(3), 473–509.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

Foster, D. P. and D. B. Nelson (1994). Continuous record asymptotics for rolling sample variance estimators.

Francq, C. and J.-M. Zakoïan (2011). Qml estimation of a class of multivariate asymmetric garch models. *Econometric Theory 28*(1), 179.

French, K. R. (1980). Stock returns and the weekend effect. *Journal of financial economics 8*(1), 55–69.

Friedmann, R. and W. G. Sanddorf-Köhle (2002). Volatility clustering and nontrading days in chinese stock markets. *Journal of economics and business 54*(2), 193–217.

Gencay, R. and F. Selcuk (2004). Extreme value theory and value-at-risk: relative performance in emerging markets. *International Journal of Forecasting 20*(2), 287–303.

Ghysels, E., A. Harvey, and E. Renault (1995). *Stochastic volatility*. CIRANO.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. Staff Report 157, Federal Reserve Bank of Minneapolis.

Golosnoy, V., B. Gribisch, and R. Liesenfeld (2012). The conditional autoregressive wishart model for multivariate stock market volatility. *Journal of Econometrics 167*(1), 211–223.

Gonçalves, S. and R. de Jong (2003). Consistency of the stationary bootstrap under weak moment conditions. *Economics Letters 81*(2), 273–278.

Gonçalves, S. and H. White (2002). The bootstrap of the mean for dependent heterogeneous arrays. *Econometric Theory 18*(6), 1367–1384.

Gouriéroux, C. (1997). *ARCH models and financial applications*. Springer.

Gouriéroux, C., J. Jasiak, and R. Sufana (2009). The wishart autoregressive process of multivariate stochastic volatility. *Journal of Econometrics 150*(2), 167–181.

Hafner, C. M. and A. Preminger (2009). Asymptotic theory for a factor garch model. *Econometric Theory 25*(2), 336.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics 23*(4).

Hansen, P. R. and A. Lunde (2006). Consistent ranking of volatility models. *Journal of Econometrics 131*(1-2), 97–121.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hansen, P. R. and A. Timmermann (2012). Choice of sample split in out-of-sample forecast evaluation.

Harvey, C. R. and A. Siddique (1999). Autoregressive conditional skewness. *Journal of financial and quantitative analysis 34*(4).

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting 13*(2), 281–291.

Hautsch, N. (2012). *Econometrics of financial high-frequency data*. Springer.

Iacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations: With R Examples*. New York: Springer. ISBN 978-0-387-75838-1.

Jacobs, B. I. and K. N. Levy (1988). Calendar anomalies: Abnormal returns at calendar turning points. *Financial Analysts Journal*, 28–39.

Jarque, C. M. and A. K. Bera (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters 6*(3), 255–259.

Jeantheau, T. (1998). Strong consistency of estimators for multivariate arch models. *Econometric theory 14*(1), 70–86.

Jorion, P. (1995). Predicting volatility in the foreign exchange market. *The Journal of Finance 50*(2), 507–528.

Jorion, P. (2007). *Value at risk* (3. ed. ed.). New York, NY: McGraw-Hill.

Kirchler, M. and J. Huber (2007). Fat tails and volatility clustering in experimental asset markets. *Journal of Economic Dynamics and Control 31*(6), 1844–1874.

Klein, B. (1977). The demand for quality-adjusted cash balances: Price uncertainty in the us demand for money function. *The Journal of Political Economy*, 691–715.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. Finance and Economics Discussion Series 95-24, Board of Governors of the Federal Reserve System (U.S.).

Laurent, S., J. V. Rombouts, and F. Violante (2010). On the forecasting accuracy of multivariate garch models. *Journal of Applied Econometrics, forthcoming*.

Laurent, S., J. V. Rombouts, and F. Violante (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics 173*(1), 1–10.

Lehar, A., M. Scheicher, and C. Schittenkopf (2002). Garch vs. stochastic volatility: Option pricing and risk management. *Journal of Banking & Finance 26*(2), 323–345.

Lima, L. and B. Neri (2007). Comparing value-at-risk methodologies. *Brazilian Review of Econometrics 27*(1).

Lindner, A. M. (2009). Stationarity, mixing, distributional properties and moments of garch (p, q)–processes. In *Handbook of financial time series*, pp. 43–69. Springer.

Ling, S. and M. McAleer (2003). Asymptotic theory for a vector arma-garch model. *Econometric Theory 19*(2), 280–310.

Liu, Q. (2009). On portfolio optimization: How and when do we benefit from high-frequency data? *Journal of Applied Econometrics 24*(4), 560–582.

Lopez, J. A. (1998). Methods for evaluating value-at-risk estimates. *Economic Policy Review* (Oct), 119–124.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business 36*, 394.

Marcus, M. and H. Minc (1992). *A survey of matrix theory and matrix inequalities*, Volume 14. Courier Dover Publications.

Markovich, N. (2008). *Nonparametric analysis of univariate heavy-tailed data: research and practice*, Volume 753. John Wiley & Sons.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York : Wiley.

McAleer, M. and M. Medeiros (2008). Realized volatility: A review. *Econometric Reviews 27*(1-3), 10–45.

Mincer, J. A. and V. Zarnowitz (1969). The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, NBER Chapters, pp. 1–46. National Bureau of Economic Research, Inc.

Morgan, J. (1996). *Riskmetrics: technical document*. Morgan Guaranty Trust Company of New York.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.

Patton, A. (2006). Volatility forecast comparison using imperfect volatility proxies. Research Paper Series 175, Quantitative Finance Research Centre, University of Technology, Sydney.

Patton, A., D. N. Politis, and H. White (2009). Correction to "automatic block-length selection for the dependent bootstrap". *Econometric Reviews 28*(4), 372–375.

Patton, A. J. and A. Timmermann (2007). Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics 140*(2), 884–918.

Pesaran, M. H. and A. Timmermann (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics 137*(1), 134–161.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Rachev, S. T., S. V. Stoyanov, A. Biglova, and F. J. Fabozzi (2005). An empirical examination of daily stock return distributions for us stocks. In *Data Analysis and Decision Support*, pp. 269–281. Springer.

Rogalski, R. J. (1984). New findings regarding day-of-the-week returns over trading and non-trading periods: A note. *The Journal of Finance 39*(5), 1603–1614.

Sarma, M., S. Thomas, and A. Shah (2003). Selection of value-at-risk models. *Journal of Forecasting 22*(4), 337–358.

Tay, A. S. and K. F. Wallis (2002). Density forecasting: a survey. *Companion to Economic Forecasting*, 45–68.

Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical finance 4*(2), 183–204.

Terasvirta, T. (2006). An introduction to univariate garch models. Working Paper Series in Economics and Finance 646, Stockholm School of Economics.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting 1*, 135–196.

Tse, Y. K. and A. K. C. Tsui (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics 20*(3), 351–362.

van der Weide, R. (2002). Go-garch: a multivariate generalized orthogonal garch model. *Journal of Applied Econometrics 17*(5), 549–564.

Varian, H. R. (1975). A bayesian approach to real estate assessment. *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, 195–208.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 1067–1084.

West, K. D. and D. Cho (1995). The predictive ability of several models of exchange rate volatility. *Journal of Econometrics 69*(2), 367–391.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Zhang, L., P. A. Mykland, and Y. Ait-Sahalia (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association 100*, 1394–1411.

Zivot, E. (2009). Practical issues in the analysis of univariate garch models. In *Handbook of financial time series*, pp. 113–155. Springer.