BestMasters

Simona Roccioletti

Backtesting Value at Risk and Expected Shortfall

Springer Gabler

BestMasters

Mit „BestMasters“ zeichnet Springer die besten Masterarbeiten aus, die an renommierten Hochschulen in Deutschland, Österreich und der Schweiz entstanden sind. Die mit Höchstnote ausgezeichneten Arbeiten wurden durch Gutachter zur Veröffentlichung empfohlen und behandeln aktuelle Themen aus unterschiedlichen Fachgebieten der Naturwissenschaften, Psychologie, Technik und Wirtschaftswissenschaften.
Die Reihe wendet sich an Praktiker und Wissenschaftler gleichermaßen und soll insbesondere auch Nachwuchswissenschaftlern Orientierung geben.

Springer awards „BestMasters“ to the best master's theses which have been completed at renowned universities in Germany, Austria, and Switzerland. The studies received highest marks and were recommended for publication by supervisors. They address current issues from various fields of research in natural sciences, psychology, technology, and economics. The series addresses practitioners as well as scientists and, in particular, offers guidance for early stage researchers.

Simona Roccioletti

Backtesting Value at Risk and Expected Shortfall

Springer Gabler

Simona Roccioletti
Guilianova, Italy

Master Thesis, University of Applied Sciences (bfi) Vienna, Austria, 2015

BestMasters
ISBN 978-3-658-11907-2 ISBN 978-3-658-11908-9 (eBook)
DOI 10.1007/978-3-658-11908-9

Library of Congress Control Number: 2015958558

Springer Gabler

Printed on acid-free paper

Springer Gabler is a brand of Springer Fachmedien Wiesbaden
Springer Fachmedien Wiesbaden is part of Springer Science+Business Media
(www.springer.com)

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Christian Cech for his excellent guidance, his advices and corrections that greatly improved the work.

I am also grateful to Prof. Umberto Cherubini, who prompted me to study this subject and guided me right from the start.

I wish to express my sincere thanks to Dr. Carlo Acerbi, who provided insight and expertise that greatly assisted the research, although he may not agree with all of the interpretations/conclusions of this thesis.

I take this opportunity to express gratitude to all of the Department faculty members for their help and kindness.

I would like to thank my QF and ARIMA collegues, who shared with me the best moments of this course of study.

Still more I am grateful to my family, who gave me the opportunity to pursue a college career and who has always supported me.

Then I would like to express my gratitude to my lifelong friends, who never had any doubts about my " final success" and who always encouraged me to do the best.

Finally, I would like to thank the person I love, for always being with me ...
wherever I go ...

Simona Roccioletti

Contents

List of Figures

List of Tables

Abbreviations

VAR	**V**alue at **R**isk
ES	**E**xpected **S**hortfall
GARCH	**G**eneralized **A**utoregressive **C**onditional **H**eteroskedasticity
TCE	**T**ail **C**onditional **E**xpectation
SE	**S**quared **E**rror
AE	**A**bsolute **E**rror
APE	**A**bsolute **P**ercentage **E**rror
RE	**R**elative **E**rror
LR	**L**ikelihood **R**atio
POF	**P**ercentage **O**f **F**ailure
CR	**C**apital **R**equirements
EVT	**E**xtreme **V**alue **T**heory
HS	**H**istorical **S**imulation
MISE	**M**ean **I**ntegrated **S**quared **E**rror
MC	**M**onte **C**arlo
i.i.d.	**i**ndipendent and **i**dentically **d**istributed
mf	**m**ultiplication **f**actor

Symbols

$\mathcal{A}$	acceptance set
$\mathbb{R}$	set of real numbers
$\mathcal{Q}$	set of probability measures
Q	risk neutral probability measure
$\mathcal{P}(\cdot)$	penalty function
$\mathcal{S}(\cdot)$	scoring function
$\mathcal{O}$	observation domain
A	action domain
$L(\cdot)$	loss function
$T(\cdot)$	statistical functional

Ω	probability space
χ	set of real-valued functions
ρ	risk measure
α	(VaR - ES) confidence level
β	(Expectiles) confidence level
μ	mean
σ	standard deviation
ν	degree of freedom
ϵ	innovation
Δ	time horizon

Abstract

Value at Risk (VaR) has always been the most popular risk measure in finance. Nonetheless, it has been constantly and heavily criticized for not being a subadditive (and hence a coherent) risk metric. Recently, Expected Shortfall (ES) has been accepted as a conceptually superior risk measure: it is coherent and, being an average, it is sensible to the size of potential losses beyond VaR itself. However, it seems that there are drawbacks in using this metric too. A recent study by Gneiting [25] has disclosed that ES is not *elicitable* and hence not *backtestable*. At the same time, Acerbi and Szekely [2] claim that *elicitability* has nothing to do with backtesting and present three model-free non parametric backtesting methodologies. In this thesis we review several valuable studies about the properties we should require from a risk measure and more important we investigate the issue related to the backtesting of ES. The main contribution of this work is the application of “Test 1” and “Test 2” developed by Acerbi and Szekely [2]. Indeed, in the empirical analysis we perform both the VaR and ES backtesting of different models and for five global market indexes: S&P 500, DAX, FTSE 100, NIKKEI and EURO STOXX 50 . . .

Chapter 1

Introduction

Risk Management is one of the most crucial skills of financial practitioners like banks, corporate treasuries, portfolio management firms and insurance companies; managing risks means identify and evaluate uncertainties in risk factors through pertinent and meaningful techniques.

A central issue in modern risk management is the measurement of risk: we need to quantify risk in a way that is easily understandable and interpretable.

Although there exist different approaches to measure the risk of a financial position, we measure it in terms of probability distributions.

Tools that map random variables to real numbers are called *measures*, in our specific case *risk measures*. Accordingly, in this work a risk measure is understood as a means providing a risk evaluation in form of capital amount, which is needed as a buffer against unexpected future losses.

The most well known and currently used risk measures are Value at Risk (VaR) and Expected Shortfall (ES).

In order to obtain an appropriate and prudential judgment, we should require some virtues from a risk measure. Artzner et al. [5] formalized some desirable properties in a set of axioms and introduced the label "Coherent" for any risk metric satisfying them all.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9_1, © Springer Fachmedien Wiesbaden 2016

It turned out that, differently from ES, VaR is not a coherent risk measure[1].
Consequently, by 1998, academic research began to criticize VaR for a number of different reasons. First of all it does not observe the sub-additivity property, meaning that it does not always take into account diversification benefits; then it says nothing concerning the "what if" questions (e.g. what could happen in case we experience a loss higher than the designed quantile). In other words, VaR completely ignores the magnitude of losses in the tail of the distribution.
In order to remedy the deficiencies of VaR, ES has been proposed as a better risk measure. Also the Basel Committee of Banking Supervision have suggested to move from VaR to ES in the interest of mitigating the weaknesses encountered using VaR for determining economic capital requirements in the context of market risk (BCBS [7]).
Expected Shortfall is sub-additive and hence it is a coherent risk measure.
Moreover, being an average, it is sensitive to the sizes of the potential losses beyond the given threshold.[2]
Despite these strengths, it seems there are drawbacks in using this kind of metric too. Cont et al. [16] show that there is a fundamental theoretical conflict between sub-additivity and robustness of risk measurement procedures for spectral measures of risk (which are popular coherent risk measure generalizing ES).
It comes forth that ES lacks robustness with respect to small changes in the data set used to estimate the loss distribution, and also that the estimation method has a great impact on ES sensitivity properties.
Furthermore, Gneiting [25] has recently argued that backtesting ES is burdensome (or even impossible) because ES is not elicitable. According to Gneiting [25], functionals for which meaningful point forecasts and forecast performance comparisons are possible are called elicitable. One important example of elecitable functionals are quantiles; thus VaR is elicitable.
"The fact that ES is not elicitable could be a partial explanation for the difficulties with robust estimation and backtesting."

[1] We should specify that this statement is not always true. Later, we will examine some cases in which VaR is a coherent risk measure.

[2] Here, we refer to the significance level.

Actually, at the time Gneiting [25] was writing, there were few works concerning the evaluation of ES forecasts and the specification of appropriate statistical tests. Besides, (from the author's point of view) the approaches already existing did not allow for a direct comparison and ranking of the predictive performance of competing forecasting methods.

The discussion conducted so far brings forth a legitimate question.

Is there any interesting coherent risk measure that is also an elicitable functional? Apparently, the answer is positive and translate into the concept of Expectiles. Expectiles are introduced by Newey and Powell [39] as the minimizers of an asymmetric quadratic loss function. They have been suggested as substantially better alternatives to both VaR and ES by virtue of their compliance with the properties mentioned above.

Despite their attractive features, the underlying idea of Expectiles is less intuitive than that of VaR and ES. Further, Expectiles do not observe the comonotonic additive principle which implies that in applications they may fail to detect risk concentration due to non-linear dependencies.

At this point there is not a strong evidence to justify an overall replacement of actual risk measures by this recent competitor[3].

Recently Acerbi and Szekely [2] have published a paper in which they join the current debate over ES versus VaR.

Above all, they argue that the belief under which the ES can not be back-testable is erroneous: they claim that the mathematical property of elicitability concerns model selection rather than model testing.

These statements imply that the concept of elicitability is almost irrelevant for the aim of choosing a regulatory risk standard and hence that ES could qualify as a better risk measure than VaR.

Acerbi and Szekely [2] also introduce three model-free, nonparametric ES backtesting methodologies, which seems to be more powerful that the Basel VaR test.

In conclusion, there exists a confusing debate about risk measures, their theoretical properties and their practical backtesting methodologies. Each risk measure

[3] This is the reason why we do not concentrate on Expectiles: for the sake of completeness we describe them in the theoretical part of the work, but we do not study them empirically.

has its strengths and drawbacks and none of them seems perfect.
Consequently, trying to understand the whole picture could help (for example) financial intermediaries in deciding which risk measure to resort to and for which reason.

The purpose of this work is to investigate the issue related to the backtesting of ES. In particular, we attempt to understand which is the true link (if any) between elicitability and backtesting and try to clarify why elicitability means (or does not) backtestable. Subsequently, we investigate a way for backtesting ES.
We describe three new methodologies to perform the backtesting of ES (i.e. the three tests by Acerbi and Szekely [2]) and try to compare them to VaR tests.
The goal of the thesis is then to empirically perform the Expected Shortfall backtesting for some global market indexes: S&P 500, DAX, FTSE 100, NIKKEI and EURO STOXX 50.

The thesis is divided in two parts.
In the first one a qualitative analysis is developed. After the introductory section (Chapter 1) we recall the main definitions and properties of a risk measure, emphasizing the differences between Expected Shortfall, Value at Risk and Expectiles (Chapter 2). We devote an entire chapter (Chapter 3) to discuss the recently introduced property of elicitability, its implications and relevance and finally we focus on backtesting: on VaR backtesting which is theoretically simple and on ES backtesting which instead, according to many authors, seems to be very complicated (Chapter 4).
In the second part a quantitative analysis is carried out. For the above mentioned time series we estimate the Value at Risk and the Expected Shortfall according to five different models: Normal, Student's t, Kernel and two GARCH models (with normal and Student's t innovations respectively). Eventually, we perform the VaR and ES backtesting for all the models considered (Chapter 5).
To carry out the empirical analysis we use MATLAB computing environment and programming language.

Chapter 2

Risk Measures and their Properties

We have seen that a risk measure can be thought of as a map from spaces of probability distributions to real numbers.

In this chapter we will provide a formal definition of risk measure and describe all the crucial properties it should satisfy.

2.1 Definition of risk measure

Given some "reference instrument", there is a natural way to define a measure of risk by describing how close or far a position is from acceptance by the regulator (Artzner et al. [5]).

Let Ω be the set of states of nature and assume it is finite.

Let $\mathcal{X}$ be the set of all risks, i.e. the set of all real-valued functions X $\in \mathcal{X}$, which represent the final net worth of an instrument, or of a portfolio of instruments, for each element of Ω.

Definition 2.1. A risk measure $\rho(X)$ is a mapping from $\mathcal{X}$ into $\mathbb{R}$.

A measure of risk allows us to express the riskiness of a position with just one number. Obviously, the riskier a position is, the higher its measure of risk will be. When positive, the number $\rho(X)$ assigned by the measure ρ to the risk X will be

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9_2, © Springer Fachmedien Wiesbaden 2016

interpreted as the amount of capital an agent has to add to the risky position X to make it an acceptable position. On the contrary, if $\rho(X) < 0$, the cash amount $-\rho(X)$ can be pulled out from the already being acceptable position and invested in a more profitable way.

Thus, it seems that the concept of measure of risk is strictly related to that of acceptability.

Indeed, Artzner et al. [5] state that sets of acceptable future net worths are of primary importance and need to be considered when describing allowance or rejection of a risky position.

An acceptance set $\mathcal{A}$ is a class of final net worths accepted by a regulator.

Depending on the degree of tolerance of a specific regulator, we can have different families of acceptable positions. Hence, each acceptance set identifies a risk measure returning acceptable positions, and each risk measure identifies an acceptance set of admissible positions.

Definition 2.2. Given the total rate of return r on a reference instrument, the risk measure associated with the acceptance set $\mathcal{A}$ is the mapping from $\mathcal{X}$ to $\mathbb{R}$, denoted by $\rho_{\mathcal{A},r}$, and defined by

$$\rho_{\mathcal{A},r}(X) = \inf\{m | m \cdot r + X \in \mathcal{A}\}$$

Definition 2.3. The acceptance set associated with a risk measure ρ is the set denoted by $\mathcal{A}_\rho$ and defined by

$$\mathcal{A}_\rho = \{X \in \mathcal{X} | \rho(X) \leq \ 0\}$$

From def.2.2, we see that a measure of risk of an unacceptable position is interpreted as the minimum extra capital (m) we need to invest in the reference instrument, for letting the future value of the modified position to be acceptable.

From def.2.3, we observe that the risk measure of an acceptable position is less than or equal to zero.

Consequently, we can affirm that the risk measure of an (initially) unacceptable position is exactly the amount of cash m through which the position $X + m$ has a null risk measure.

In the following sections, we will present and describe the three most popular risk measures in the financial environment.

2.2 Value at Risk

Value at Risk is probably the most widely used risk measure in finance. It has become the classic measure that financial executives use to quantify market risk. Indeed, when RiskMetrics announced Value at Risk as its measure of risk in 1996, the Basel Committee on Banking Supervision enforced financial institutions to meet capital requirements based on VaR estimates.

We will now provide a formal definition.

Consider a portfolio of risky assets and a fixed time horizon Δ.

Suppose we have estimated the loss distribution associated to this portfolio and denote by $F_L(l) = P(L \leq \ l)$ its distribution function. [1]

We would like to define a statistic based on F_L which evaluates the level of risk associated to the holding of our portfolio over the time period Δ.

An noticeable applicant is the maximum possible loss, i.e. $\inf\{l \in \mathbb{R} : F_L(l) = 1\}$. Nonetheless, it is not difficult to believe that some models entail an unbounded support for F_L , meaning that the maximum possible loss would be infinity.

Value at Risk is a smart and straightforward extension of maximum loss: the idea is simply to replace "maximum loss" by "maximum loss which is not exceeded with a given high probability", the so called confidence level (McNeil et al. [37]).

Definition 2.4 (Value-at-Risk)**.** Given some confidence level [2] $\alpha \in (0, 1)$, the VaR of our portfolio at the confidence level α is given by the smallest number l such that the probability that the loss L exceeds l is no larger than $(1 - \alpha)$.

[1] Practitioners are often involved with the so called profit-and-loss (P&L) distribution. However in risk management we are mainly concerned with the probability of large losses and hence we often drop the P from P&L. Moreover, since actuarial risk theory is a theory of positive random variables, we will use the convention that losses are positive numbers and we will focus on the upper (right) tail of the loss distribution L.

[2] α denotes the confidence level, not to be confused with the significance level, which is equal to $1 - confidence\ level$. Hence, here α could assume values like 0.95 or 0.99, not 0.05 nor 0.01.

Formally,

$$VaR_\alpha(L) = \inf\{l \in \mathbb{R} : P(L > \ l) \leq \ 1-\alpha\} = \inf\{l \in \mathbb{R} : F_L(l) > \ \alpha\} \qquad (2.1)$$

In probabilistic terms, VaR is thus simply a quantile of the loss distribution.
We recall that, given some df L, the generalized inverse $F^{\leftarrow}$ is called the *quantile function* of L.

$$q_\alpha(L) := F^{\leftarrow}\ (L) = \inf\{l \in \mathbb{R} : F_L(l) \geq \ \alpha\}$$

Thusly, we get:

$$VaR_\alpha(L) = q_\alpha(L)$$

Typical values for α are α=0.95, α=0.99 or higher.
In market risk management the time horizon Δ is usually 1 or 10 days, while in credit risk management Δ is usually one year.
VaR has the advantage of being very intuitive and the great popularity this instrument has reached, is essentially due to its conceptual clarity.
However, VaR has two important drawbacks:

i) it does not fulfill (in general) the property of sub-additivity, meaning that it does not award diversification benefits;

ii) it is tail insensitive. It tells us that in the $\alpha \cdot 100\%$ of the cases the loss will not be greater than a certain level, but it does not give us any clue about the size of the loss in the remaining $(1-\alpha)\cdot 100\%$ of the cases.

In Section 2.5 we will deepen the discussion about the properties a good risk measure should have.

2.3 Expected Shortfall

Risk professionals have been looking for a coherent alternative to Value at Risk for many years.

Expected Shortfall turns out to be a natural choice to appeal to, when VaR is unable to distinguish between portfolios embodying different levels of risk.

In fact, little imagination is necessary to construct portfolios with identical VaR and dangerously divergent levels of tail risk, namely the risk in the $(1-\alpha)\%$ worst cases.

It seems that instead of asking what could be the minimum loss incurred in a presupposed percentage of worst cases, we should be concerned with the expected loss sustained in that portion of unfortunate possibilities.

It is effortless to figure out that, if the loss distribution function is continuous, then the statistical quantity which answers the questioning above, is simply given by the conditional expected value above the quantile, that is the $Tail\ Conditional\ Expectation$:

$$TCE_\alpha(L) = E\left\{L|L \geq \ VaR_\alpha\right\} \tag{2.2}$$

For more general distribution however, this statistic does not fit really well our purposes, since the event $\{L \geq \ VaR_\alpha\}$ may happen to have a probability larger than our set of selected worst cases.

Indeed TCE may violate the sub-additivity property on a general distribution.

In order to account for more general distribution we need to expand the definition of TCE. This leads to the following formulation.

Definition 2.5 (Expected Shortfall)**.** For a loss L with $E(|L|) < \infty$ and distribution function F_L, the Expected Shortfall at confidence level $\alpha \in (0,1)$ is defined as

$$ES_\alpha(L) = \frac{1}{1-\alpha} \cdot \left(E(L; L \geq \ q_\alpha) + q_\alpha \cdot (1 - \alpha - P(L \geq \ q_\alpha))\right) \tag{2.3}$$

where q_α is the α-quantile of F_L.

Equation 2.3 may look complicated. Nevertheless, the concept it expresses is that cited above. Hence equation 2.3 is simply the mathematical translation of what we are looking for.

To have a more transparent view, we should clarify some notions: the term $q_\alpha \cdot (1 - \alpha - P(L \geq \ q_\alpha))$ has to be interpreted as the exceeding part which need to be added to the expected value $E(L; L \geq \ q_\alpha)$ when the event $\{L \geq \ q_\alpha\}$ has probability larger than $(1 - \alpha)$.

When, on the contrary, $P(L \geq \ q_\alpha) = 1 - \alpha$, as is always the case if the probability distribution is continuous, the term vanishes and equation 2.3 reduces to equation 2.2 .

Said differently, for an integrable loss L with continuous distribution function F_L and any $\alpha \in (0, 1)$, we have that

$$ES_\alpha(L) = \frac{E(L; L \geq \ q_\alpha(L))}{1 - \alpha} = E(L|L \geq \ VaR_\alpha) \tag{2.4}$$

and hence

$$ES_\alpha(L) = TCE_\alpha(L) \tag{2.5}$$

The comprehensibility of the ES_α can be perceived making use of an equivalent definition, which handles the ES_α as a combination of expected values.

There exists an analogous representation to equation 2.3 which reveals the close link with the parameter α and the distribution function F_L.

Literally, employing the left generalized inverse function of F_L,

$$F_L^{\leftarrow}\ (u) = \inf\{u \in \mathbb{R} : F_L(u) \geq \ \alpha\}$$

one can handily illustrate that ES_α can be expressed as the mean of $F^{\leftarrow}\ (L)$ on the significance level interval:

$$ES_\alpha(L) = \frac{1}{1 - \alpha} \cdot \int_\alpha^1 q_u(F_L)\, du. \tag{2.6}$$

Expected Shortfall is thus related to VaR by

$$ES_\alpha(L) = \frac{1}{1-\alpha} \cdot \int_\alpha^1 VaR_u(L)\, du. \tag{2.7}$$

This is the principal and most used formulation of ES_α.

Its mathematical manageability makes it especially suitable for studying the analytical properties of ES_α. For instance, one of the property that differentiate the ES_α from its competitors, specifically the continuity in α, is manifest in equation 2.7, while it is not in equation 2.3.

To sum up, Expected Shortfall comes out to be a coherent risk measure; however it has some weaknesses as well.

In particular we will see that ES is not an *elicitable* risk measure.

2.4 Expectiles

Given that Value at Risk is not coherent and Expected Shortfall is not elicitable, many authors have been searching recently, for a risk measure sharing both the properties of coherence and elicitability.

Possible candidates are Expectiles. In this section we are going to see what Expectiles are and why they are so interesting as a potential risk measure.

We will define Expectiles starting from the work of Newey and Powell [39].

We know that a quantile can be defined as the minimizer of an asymmetric loss function:

$$q_\alpha(L) := \arg\min_{l\in\mathbb{R}} E[\alpha \cdot (L-l)^- + (1-\alpha)\cdot(L-l)^+]$$

Similarly Newey and Powell [39] have defined the Expectile as the minimizer of an asymmetrically weighted squared loss function.

Definition 2.6 (Expectiles)**.** For $0 < \beta < 1$ and square-integrable L, the $\beta -$ *Expectile* $e_\beta(L)$ is defined as

$$e_\beta(L) = \arg\min_{l \in \mathbb{R}} E[(1-\beta)\ max\ (L-l,0)^2 + \beta\ max\ (l-L,0)^2] \tag{2.8}$$

Note that, as for the variance, the notion of Expectile requires finite second moment.

Expectiles can also be defined in terms of their first order condition (f.o.c henceforth). Indeed the β-Expectile, $e_\beta(L)$, is the unique solution l of the following equation:

$$(1-\beta)\ E[max\ (L-l,0)] = \beta\ E[max\ (l-L,0)] \tag{2.9}$$

Consequently, $e_\beta(L)$ satisfies

$$e_\beta(L) = \frac{\beta\ E[L\ \mathbf{1}_{\{L<e_\beta(L)\}}] + (1-\beta)\ E[L\ \mathbf{1}_{\{L\geq\ e_\beta(L)\}}]}{\beta\ P[L<e_\beta(L)] + (1-\beta)\ P[L \geq\ \ e_\beta(L)]}$$

From all the above definitions we can understand that the Expectiles are specified by the weighted conditional expectations of the random variable L.

Strictly speaking Expectiles are a measure of the right tail and left tail expected values of the loss L. Thus, while VaR does not consider any of the two tails of the loss distribution and ES only examines one of them, Expectiles assess both the tails with different weights.

By way of explanation, they balance the left and right tail of the distribution, so that the ratio between the expected positive and negative deviations from e_β will be equal to a predetermined constant:

$$\frac{E[(L-e_\beta(L))^+]}{E[(L-e_\beta(L))^-]} = \frac{\beta}{(1-\beta)}$$

Supposing that $\beta = 1-\beta = \frac{1}{2}$, the ratio will be equal to 1, implying that the means of the left and the right deviations from e_β are equal.

We should mention that in Rossi [42] the risk measure associated to expectiles is indicated as EVaR.

Specifically, for a loss L with $E(L^2) < \infty$ and distribution function F_L, we have

$$EVaR_\beta(L) = e_\beta(L)$$

Before going on and describing the very important properties of a risk metric, we would like to make a parallelism with Section (2.1) and define a risk measure using the notion of acceptance set (see def. 2.3).

This is done, mainly, to let the reader better understand the concept of EVaR. Hence, we can define our three risk measures through the concept of acceptance set.

For a generic random variable X, the VaR acceptance set is:

$$\mathcal{A}_{VaR_\alpha} = \{X \in \mathcal{X} : P(X < 0) \leq \; 1 - \alpha\}$$

For a generic random variable X, the ES acceptance set is:

$$\mathcal{A}_{ES_\alpha} = \left\{X \in \mathcal{X} : \frac{1}{1-\alpha} \cdot \int_0^{1-\alpha} q_u(F_X)\, du \geq \; 0\right\}$$

For a generic random variable X, the EVaR acceptance set is:

$$\mathcal{A}_{EVaR_\beta} = \left\{X \in \mathcal{X} : \frac{E[X^+]}{E[X^-]} \geq \; \frac{1-\beta}{\beta}\right\}$$

Thus, when we refer to VaR, a position is included in the acceptance set if the probability of a loss does not exceed the fixed level $(1 - \alpha)$.

In the case we are dealing with ES, a position is said acceptable if the average loss in the worst $(1 - \alpha) \cdot 100\%$ cases is not greater than zero and, for this reason, we should say that ES is more conservative than VaR.

Finally, looking at EVaR, a position can be considered acceptable if its gain-loss ratio is greater than the prearranged value $\frac{1-\beta}{\beta}$.

2.5 Coherent risk measures

We are now going to outline the subset of coherent risk measures.

The notion of coherence was introduced by Artzner et al. [5] and currently, it is a fundamental concept related to the acceptability of a risk measure.

In reality, Acerbi and Tasche [3] state that if a measure turns out to be not coherent, then it simply cannot be named as *risk measure*:

> "To avoid confusion, if a measure is not coherent we just choose not to call it a risk measure at all."

So, let us describe what *coherence* means.

Definition 2.7. A risk measure is *coherent* if it satisfies the following four axioms:

Axiom 1. Translation Invariance

For all $X \in \mathcal{X}$ and for all $m \in \mathbb{R}$, we have

$$\rho(X+m) = \rho(X) - m \tag{2.10}$$

Axiom 2. Sub-additivity

For all $X_1 \in \mathcal{X}$ and $X_2 \in \mathcal{X}$, we have

$$\rho(X_1 + X_2) \leq \ \rho(X_1) + \rho(X_2) \tag{2.11}$$

Axiom 3. Positive Homogeneity

For all $X \in \mathcal{X}$ and for all $\tau > 0$, we have

$$\rho(\tau \cdot X) = \tau \cdot \rho(X) \tag{2.12}$$

Axiom 4. Monotonicity

For all $X_1 \in \mathcal{X}$ and $X_2 \in \mathcal{X}$ with $X_1 \leq \ X_2$, we have

$$\rho(X_1) \geq \ \rho(X_2) \tag{2.13}$$

The axiom of *Translation Invariance* illustrates that adding (or deducting) a risk-free amount to (from) a portfolio and investing it in the reference instrument, results in a decrease (increase) of the risk of the position by exactly the same amount.

From this property we can derive a well-known fact in finance:

if $m = \rho(X)$, then

$$\rho(X + m) = \rho(X + \rho(X)) = \rho(X) - \rho(X) = 0$$

Hence, it is possible to hedge an underwritten risky position, by simply adding a certain amount of risk-free instruments in the portfolio.

The axiom of *Sub-additivity* reminds us of the diversification theory.

In fact, thanks to diversification benefits, a portfolio composed by several assets will be strictly less risky than a portfolio made up by a single instrument, provided that the correlation among the assets is different from 1.

In this sense, we can say that the sub-additivity property sets an upper bound to the risk of a portfolio and thus to the amount of regulatory capital we need to allocate. Only when there is a well-founded possibility that the sources of these risks may act altogether, the global risk of a portfolio will be the sum of the risks of its components.

According to Artzner et al. [5], the *Sub-additivity* axiom captures the essence of how a risk measure should behave, especially in the aggregation (disaggregation) of portfolios, and for this reason it is probably the key feature of a risk metric.

The property of *Positive Homogeneity* underlines that, if (for instance) the exposure to a specific position doubles, then the risk measure related to that position doubles as well. However, in the case that the position size directly influences risk

(think for example of liquidity risk), we should account for any possible repercussion (e.g. difficulty in liquidate the position), and we might expect the risk will increase more than twice.
Hence, we could have $\rho(\tau \cdot X) \geq \ \tau \cdot \rho(X)$.
Nonetheless, from the *Sub-additivity* axiom we know that $\rho(\tau \cdot X) \leq \ \tau \cdot \rho(X)$. Indeed:

$$\rho(\tau \cdot X) = \underbrace{\rho(X + X + \cdots + X)}_{\tau-times} \leq \ \underbrace{\rho(X) + \rho(X) + \cdots + \rho(X)}_{\tau-times} = \tau \cdot \rho(X)$$

.

Accordingly, the only possible solution is the equality as expressed in Axiom 3.
Finally, the *Monotonicity* axiom explains that if, in each state of the world, the position X_2 performs always better than position X_1, then the risk associated to X_1 should be higher than that related to X_2.

2.5.1 Coherence of VaR

As previously mentioned, the sub-additivity property fails to be valid for Value at Risk (*in general*), meaning that it is not a coherent risk measure.
Thus we have:

$$VaR_{\alpha}(L_1 + \cdots + L_d) > VaR_{\alpha}(L_1) + \cdots + VaR_{\alpha}(L_d) \tag{2.14}$$

We have seen that for a sub-additive measure, portfolio diversification always lead to risk reduction, while for metrics like VaR this fact may not hold anymore. Whenever we try to get the whole picture, we immediately realize that sub-additivity is a necessary property for capital adequacy requirements.
Consider a financial institution made of several businesses: if the regulatory capital is computed using a 'bottom-up' approach (and using a non sub-additive risk measure), how could we be sure that the final number we get, consistently estimates the whole risk of the enterprise?
Hence, a decentralization of risk management using VaR is troublesome, since we

cannot be certain that, by aggregating VaR numbers for different branches, we will end up with a bound for the overall risk of the company.

When looking at aggregated risk $\sum_{i=1}^{n} L_i$, the equation 2.14 becomes:

$$VaR_\alpha(\sum_{i=1}^{n} L_i) > \sum_{i=1}^{n} VaR_\alpha(L_i)$$

Previously, we have stated that VaR is not sub-additive *in general*.

Indeed, whether or not it is the case depends on the properties of the joint loss distribution and the standard situations in which our measure comes out to be sub-additive, are the following:

i) The random variables are independent and identically distributed, as well as positively regularly varying.

ii) The random variables have an elliptical distribution.

iii) The random variables have an Archimedean survival dependence structure.

What the first context directly implies is that, when the loss distribution is normal, VaR does not have any problems.

Of course, it will not be a big issue to assess the tail risk since, with normality assumption, VaR is a scalar multiple of standard deviation and consequently also a scalar multiple of ES (which we know is good in providing information about tail losses).

Moreover, being a multiple of standard deviation, "Normal" VaR satisfies the sub-additivity axiom.

Howsoever, it should be noted that the real world has nothing to do with the Gaussian one, and also that in Gaussian world everything is proportional to standard deviation, so everything is sub-additive, and there is nothing special with VaR.

As regard the other two more complicated situations, the interested reader should refer to Danìelson et al. [17], Embrechts et al. [19] and Embrechts et al. [20].

Furthermore Embrechts et al. [21] has introduced a numerical algorithm which provides boundaries for the VaR of high-dimensional (inhomogeneous) portfolios, and Kratz [33] has studied the topic of the evaluation of VaR of aggregated heavy tailed risks.

2.5.2 Coherence of ES

Expected Shortfall fulfills all the four axioms above and so it is a coherent risk measure.

Thence, many authors believe that ES is an excellent substitute for VaR in risk management applications. Undoubtedly, some of the warnings that apply for VaR, such as the daintiness of the estimation procedure, the reliability of approximations and the consistency of the assumptions, should also be considered in this case. The only difference is that, if correctly estimated, the ES will give less misleading answers.

From the practical point of view, the goodness of Expected Shortfall is contingent on the stability of the estimation approach and on the finding of effective backtesting methodologies.

We will address this topic later on, placing particular emphasis on the robustness of risk measurement procedures and on the relationship between backtesting and a property called *elicitability*.

2.5.3 Coherence of Expectiles

Expectiles can be a coherent risk measure under some restrictions. In fact, they satisfy all the axioms, except the one of sub-additivity.

Actually, expectiles are sub-additive if $\beta \geq \frac{1}{2}$, although super-additive if $\beta \leq \frac{1}{2}$:

Sub-additivity. For two random variables $X_1, X_2 \in \mathcal{X}$, and for $\beta \geq \frac{1}{2}$, we have

$$e_\beta(X_1 + X_2) \leq e_\beta(X_1) + e_\beta(X_2)$$

Super-additivity. For two random variables $X_1, X_2 \in \mathcal{X}$, and for $\beta \leq \frac{1}{2}$, we have

$$e_\beta(X_1 + X_2) \geq \ e_\beta(X_1) + e_\beta(X_2)$$

Hence, in the moment we look at expectiles for risk measurement purposes , we should keep in mind that it is a coherent risk measure only in the case the confidence level is above 0.5.

What is worth mention is that expectiles, not only fulfill the other three axioms, but also comply with a stricter property:

Strong Monotonicity. For two random variables $X_1, X_2 \in \mathcal{X}$, and for $\beta \in (0, 1)$, if $X_1 \leq \ X_2$ and $P(X_1 < X_2) > 0$, we have

$$e_\beta(X_1) < e_\beta(X_2)$$

This is a more rigorous axiom than the monotonicity one.

Indeed strong monotonicity guarantees that, given $X_1 \leq \ X_2$, if the probability of the event "X_1 is strictly less than X_2" is non-null, then the above inequality involving $e_\beta(X_1)$ and $e_\beta(X_2)$, always holds.

On the contrary, simple monotonicity could end up with $e_\beta(X_1) = e_\beta(X_2)$, even when there is a strictly positive probability for the event "$X_1 < X_2$."

2.6 Risk Measures: a deeper view

In this brief section we will provide some other significant properties a risk measure should comply with. In truth, we believe they reserve some care.

2.6.1 Convexity

The concept of convex measure of risk is an extension of that of coherent risk measure discussed formerly.

Following the reasoning of Föllmer and Schied [24], we start with a comprehensible characterization of a risk metric:

A coherent measure of risk ρ originates from some family $\mathcal{Q}$ of probability measures, by gauging the expected loss under $Q \in \mathcal{Q}$ and then considering the worst result as Q varies over $\mathcal{Q}$:

$$\rho(X) = \sup_{Q\in\mathcal{Q}} E_Q[-X]$$

In real world, however, there exist a large number of situations in which, the risk borne in an investment, evolves in a nonlinear way with the size of a position. Recall the example (about liquidity risk) we made for explaining the property of positive homogeneity.

In this context, we should relax the assumption of sub-additivity and positive homogeneity and require the weaker property of *Convexity*:

$$\rho(\lambda X_1 + (1 - \lambda X_2)) \leq \ \lambda\rho(X_1) + (1 - \lambda)\rho(X_2)$$

for any $\lambda \in (0, 1)$ and any (risky positions) $X_1, X_2 \in \mathcal{X}$.

Also here, we can notice the direct link with the notion of diversification: the risk of a diversified portfolio, which in this case is $\lambda X_1 + (1 - \lambda X_2)$, is less or equal to the weighted average of individual risks.

Assume now that $0 \in \mathcal{X}$ and that $\mathcal{X}$ is closed under the addition of constants.

Definition 2.8 (Convex Measure of Risk). A map $\rho : \mathcal{X} \to \mathbb{R}$ will be called a *convex measure of risk* if it satisfies the condition of *convexity*, *monotonicity* and *translation invariance*.

When we deal with *normalized convex* measure of risk, i.e. $\rho(0) = 0$, the quantity $\rho(X)$ can be understood as a "margin requirement".

Remind that a margin is the minimum amount of capital we have to add to a risky position and invest in a risk-free asset, in order to make that position "acceptable".

We conclude this subsection by enunciating the fundamental representation theorem for convex measures of risk.

Theorem 2.9. *Suppose $\mathcal{X}$ is the space of all real-valued functions on a finite set Ω. Then, $\rho : \mathcal{X} \to \mathbb{R}$ is a convex measure of risk if and only if there exists a "penalty function" $\alpha : \mathcal{P} \to (-\infty, \infty]$ such that*

$$\rho(X) = \sup_{Q \in \mathcal{P}} \Big(E_Q[-X] - \alpha(Q) \Big)$$

The function α satisfies $\alpha(Q) \geq \; -\rho(0)$ for any $Q \in \mathcal{P}$, and it can be taken to be convex and lower semicontinuous on $\mathcal{P}$.

Observe that the structure theorem of coherent risk measure is a special case of the above. Accordingly, it is not surprising that any positive homogeneous and subadditive risk measure is also convex. Expected shortfall and expectiles are both convex risk measures. Obviously, Value at risk is not.

2.6.2 Comonotonic Additivity

Earlier, we have given considerable importance to the sub-additivity axiom.
Now we would like to present a similar result, which can be very useful for financial purposes. First of all we should define the concept of *comonotonicity*.

Definition 2.10. Two random variables X_1 and X_2 are said to be comonotonic if, given a third random variable Y, there exist two monotonic increasing function f_1 and f_2, such that

$$X_1 = f_1(Y) \textit{ and } X_2 = f_2(Y)$$

This simply means that, for instance, two risky positions X_1, X_2 are perfectly and also positively dependent on the same source of risk Y, i.e.

$$X_1 \uparrow \iff X_2 \uparrow$$

Obviously, the perfect positive dependence implies the maximum degree of correlation.
Given these clarifications, we can now turn to the concerned definition.

Definition 2.11 (Comonotonic Additivity). A risk measure $\rho(\cdot)$ is *comonotonically additive*, if for any comonotonic random variables X_1, X_2, it holds that

$$\rho(X_1 + X_2) = \rho(X_1) + \rho(X_2)$$

The reason why this property extremely matters, is intuitive and again related to the notion of diversification.
As a matter of fact, if two different risky positions perfectly depend on the same risk factor, they should not benefit from diversification effects.
Thus, in risk management, we should always use comonotonic additive risk metrics.
VaR and ES both comply with this property, while Expectiles do not.

2.6.3 Law Invariance

A risk measure is defined as *law invariant* if it depends entirely on the distribution of the random variable associated to it. More precisely:

Definition 2.12 (Law Invariance). Consider two random variables X_1, X_2 and their corresponding distribution functions F_{X_1}, F_{X_2}. A risk measure $\rho(\cdot)$ is a *law invariant* risk measure if

$$F_{X_1} = F_{X_2} \implies \rho(X_1) = \rho(X_2)$$

The importance of this property lies on the fact that, for assessing the risk level of a position, we apply the risk measure on the loss distribution, which is estimated from empirical data.
Hence, in the interest of making this approach accurate we need that, every time the variables follow identical distributions, the measure returns the same level of risk.
A direct consequence of this fact is that, whenever we deal with risk measures that are not law invariant, we could not evaluate the riskiness of a position through the loss distribution.

As underlined in Acerbi [1], the majority of risk measures used in finance (including VaR, ES and Expectiles) comply with this property.

2.6.4 Robustness

Another important issue when coping with risk measures is robustness.
A risk measure is said to be *robust* if it is quite insensible to measurement errors. Cont et al. [16] observe that measuring the risk of a financial portfolio involves two steps: estimating the loss distribution from empirical data and computing the risk metric ρ which, as usual, is a map assigning a number to each random payoff. They also point out that, even if these two operations have always been studied separately, they are strictly related (at least in applications) and their connection is of great importance when opting for a risk measure, rather than another.
What could be really problematic is the sensitivity of risk measures to mis - specification errors in the portfolio loss distribution. We care so much about the portfolio loss distribution for the simple reason that VaR, ES and also Expectiles are directly estimated from it. In light of this, it could be useful to evaluate the risk estimators' sensitivity, for instance computing their relative change when a new observation in the data set is included.
The figure below, taken from Cont et al. [16], exhibits this measure of sensitivity of VaR and ES as a function of the size of a point added.

We can understand that the ES is much more sensitive than VaR to a change in the data set, especially when large observations are added.
Thus, ignoring robustness may lead to meaningless results, since small measurement errors in the estimation procedure can have a huge impact on our outcomes. Moreover, looking at figure 2 (Cont et al. [16]), we can observe that different estimation procedures for the same risk measure and the same portfolio display divergent sensitivities to a new outlier.
Hence, depending on whether the loss distribution is estimated directly from historical data or approximated through a parametric model, the sensitivity of risk measures varies.

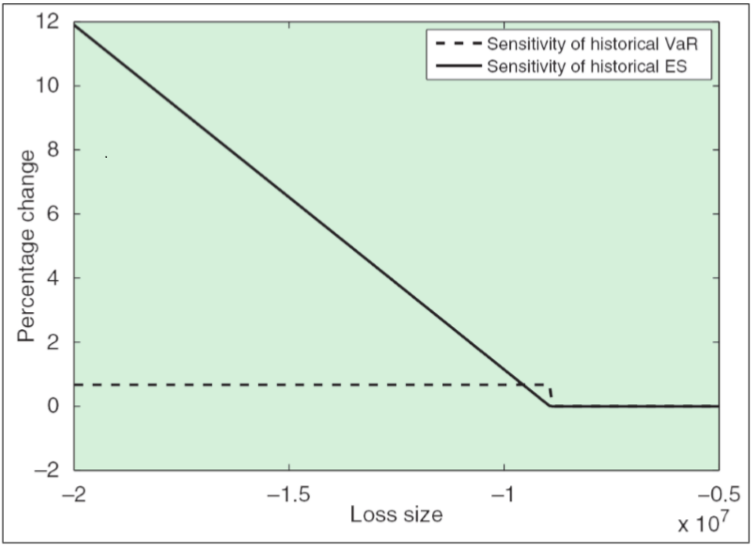


FIGURE 2.1: Empirical sensitivity (in percentage) of the historical VaR and historical ES, as in Cont et al. [16]

Concluding, Cont et al. [16] show us two important results:

i) While Expected Shortfall has the advantage of being a coherent risk measure, it seems to lack robustness. On the contrary, VaR appears to be quite insensible to data modifications.

ii) The choice of the estimation method has a considerable impact on the sensitivity of the risk measure.

These fundings motivate the claim according to which there exists a conflict between coherence (or better sub-additivity) of a risk measure and the robustness of its statistical estimators.

Without going any further on the subject, we shall mention that robustness is usually investigated in terms of continuity (either with respect to the weak topology or considering the Wasserstein distance).

For the sake of conformity, we should report that Expectiles are not robust with respect to the weak topology but, as proved in Bellini et al. [9], they are robust

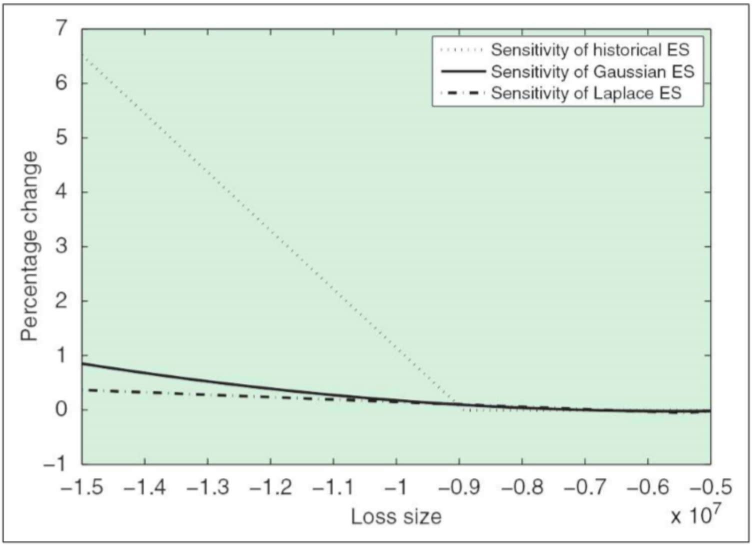


FIGURE 2.2: Empirical sensitivity (in percentage) of the $ES_0.01$ estimated with different methods, as in Cont et al. [16]

with respect to the Wasserstein distance.

The interested reader should refer to Cont et al. [16] and Stahl et al. [46] for further insights.

Chapter 3

Elicitability

As we have seen before, a risk measure has to be estimated from historical data. In order to reach the best possible point estimate, we have to make several choices concerning models, methods and parameters.

For this reason, it is crucial to be able to validate and compare competing estimation procedures. In a decision-theoretic framework, statistical functionals (e.g. risk measures) for which such validation and performance comparison are possible, are called *elicitable*.

In this chapter we will understand what *elicitability* means and what are its main implications. We will follow the approach of Gneiting [25], who has recently investigated the issue from a broad perspective.

3.1 Evaluate Point Forecasts

Typically, point forecasting procedures are evaluated using an error measure or scoring functions; classic examples are the absolute error and the squared error.

A simple representation of a possible performance criterion is:

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(x_i, y_i) \tag{3.1}$$

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9_3, © Springer Fachmedien Wiesbaden 2016

where n is the number of forecast cases, $x_1, \cdots, x_n$ denote the point forecasts and $y_1, \cdots, y_n$ the realizations. The function S takes as arguments both our predictions and the verifying observations, and we refer to it as a *scoring function*.
We generally want the scoring function to be as small as possible, since this would mean that our forecasts do not deviate much from reality. More appropriately, we say that scoring functions are *negatively oriented*.
Table 1 lists the four most famous scoring functions.

$S(x,y) = (x-y)^2$	squared error	(SE)
$S(x,y) = \|x-y\|$	absolute error	(AE)
$S(x,y) = \|(x-y)/y\|$	absolute percentage error	(APE)
$S(x,y) = \|(x-y)/x\|$	relative error	(RE)

TABLE 3.1: Some commonly used scoring functions.

The absolute error and the squared error are of the *prediction error* form, as they depend on the forecast error only; they are also *symmetric*. The absolute percentage error and the relative error do not share any of these two features. They are adopted for strictly positive quantities only.
In his article, Gneiting [25] argues that evaluating point forecasts by using 'some'[1] scoring function(s), which could not be consistent for a specified functional, may lead to meaningless conclusions.
He carries out a simulation study whose results are counterintuitive and disconcerting: he constructs a realistic example in which the performance of skillful statistical forecasts is ranked worse than a single mindless prediction, when in fact assessed by means of 'some' scoring functions.
To make this concept more perceptible, we report two extracts respectively from an article by J.Engelberg et al. [29], and a paper by Murphy and Daan [38]:

> "Our concern is prediction of real-valued outcomes [...]. In this case the users of point predictions sometimes presumes that forecasters report the means of their subjective probability distributions; that is their

[1] What is meant for 'some' is the aimless and unguided choice of the scoring functions we need to use.

> best point predictions under square loss. However, forecasters are not specifically asked to report subjective means. Nor are they asked to report subjective median or modes, which are best predictors under other loss functions. Instead, they are simply asked to 'predict' the outcome or to provide their 'best prediction', without definition of the word 'best'. In the absence of explicit guidance, forecasters may report different distributional features as their point predictions."

> "It will be assumed here that the forecasters receive a 'directive' concerning the procedure to be followed [...] and that it is desirable to choose an evaluation measure that is consistent with this concept. An example may help to illustrate this concept. Consider a continuous [...] predicant, and suppose that the directive states 'forecast the expected (or mean) value of the variable'. In this situation, the mean square error measure would be an appropriate scoring rule, since it is minimized by forecasting the mean of the (judgmental) probability distribution. Measures that correspond with a directive in this sense will be referred to as consistent scoring rules (for that directive)."

In agreement with these standpoints, Gneiting [25] states that "guidance" or "directives" for effective point forecasting can be given in two complementary ways:

i) by disclosing the scoring function ex ante to the forecaster, allowing him/her to tailor the point predictor to the given function and to issue the optimal forecast, i.e. the Bayes rule:

$$\hat{x} = \arg\min_{x} E_F[S(x, Y)] \tag{3.2}$$

where the random variable Y is distributed according to the distribution F.

ii) by requesting a specific functional of the forecaster's predictive distribution (e.g. the mean or a quantile) and applying any scoring function that is consistent with that functional.

We now develop a theoretical framework for the assessment of point predictions and explain the hinted notions from a mathematical point of view.
The basic ingredients we need are:

- an *observation domain* O,
- a class $\mathcal{F}$ *of probability measures* on the observation domain,
- an *action domain* A and
- a *loss function* $L : A \times O \to [0, \infty)$, where $L(a, o)$ denotes the loss incurred when we take the action a and the observation o materializes.

Following Granger and Machina [26], we assume that the observation and the action domain coincide and that this common domain is a subset of the Euclidean space equipped with the Borel σ-algebra:

$$D = O = A \subseteq \mathbb{R}$$

Moreover, we designate the loss function as a scoring function
$S : \mathcal{D} = D \times D \to [0, \infty)$, where $S(\mathbf{x}, \mathbf{y})$ represents the penalty due to the occurrence of $\mathbf{y} \in D$ when $\mathbf{x} \in D$ is predicted.
Notice that we actually work in dimension $d = 1$, in which any connected domain is simply an interval I. Primarily, we are concerned with the real line, $I = \mathbb{R}$ or with the nonnegative half-axis, $I = [0, \infty)$ or $I = (0, \infty)$.
Then, a *scoring function is any mapping* $S : I \times I \to [0, \infty)$.
We impose our scoring functions on intervals, to satisfy the following assumptions:

Assumption 1. $S(x, y) \geq 0$, *with equality if* $x = y$.

Assumption 2. $S(x, y)$ *is continuous in* x.

Assumption 3. *The partial derivative $\partial_x S(x, y)$ exists and is continuous in x whenever $x \neq y$*

The first assumption is merely a nonnegative condition (technically non restrictive); the second and the third correspondingly regard continuity and differentiability with respect to the point forecast x, and they are justified by the fact that a loss function multiplied by a strictly positive constant should not change the nature of the optimal point forecast. Furthermore, the optimization problem in equation 3.2 is posed in terms of the first argument x.
We also require our scoring function to possess two suitable interconnected properties, *homogeneity* and *equivariance*.

Definition 3.1 (**Homogeneity or Scale Invariance**). A scoring function S on a general prediction-observation (PO) domain $\mathcal{D} = D \times D$ is *homogeneous* of *order* b if

$$S(c\mathbf{x}, c\mathbf{y}) = |c|^b S(\mathbf{x},\ \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in D$ *and* $c \in \mathbb{R}$ such that $c\mathbf{x} \in D$ and $c\mathbf{y} \in D$.

Definition 3.2 (**Equivariance**). A scoring function S on a general PO domain $\mathcal{D} = D \times D$ is *equivariant* with respect to some class $\mathcal{H}$ of injections $h : D \to D$ if

$$\arg\min_x E_F[S(\mathbf{x}, h(\mathbf{Y}))] = h\Big(\arg\min_x E_F[S(\mathbf{x}, \mathbf{Y})]\Big)$$

for all $h \in \mathcal{H}$ and all probability distributions F that are concentrated on D.

3.1.1 Consistency

As we have clarified what a scoring function is, we now turn to define the notion of *statistical functional*.

Definition 3.3 (**Statistical Functional**)**.** A Statistical Functional is a potentially set-valued mapping from a class of probability distributions $\mathcal{F}$, to a Euclidean space.

In the current context, we require that the functional

$$T : \mathcal{F} \to \mathcal{P}(D), \quad F \mapsto T(F) \subseteq D,$$

maps each distribution $F \in \mathcal{F}$ to a subset $T(F)$of the domain $D \subseteq \mathbb{R}$.

Often $T(F)$ is single valued, with the set-valued quantile functionals being a leading exemption.

We began this section quoting, among others, Murphy and Daan (1985, p. 391), who stressed the importance of identify what *consistency* for a given functional, means.

It is time to give an accurate explanation of this concept.

Definition 3.4 (**Consistency of a Scoring Function**)**.** The scoring function S is *consistent* for the functional T relative to the class $\mathcal{F}$ if

$$E_F S(\mathbf{t}, \mathbf{Y}) \leq \ E_F S(\mathbf{x}, \mathbf{Y}) \tag{3.3}$$

for all probability distributions $F \in \mathcal{F}$, all $\mathbf{t} \in T(F)$ and all $\mathbf{x} \in D$.

It is *strictly consistent* if it is consistent and equality in equation 3.3 implies that $\mathbf{x} \in T(F)$.

To give an example, we can think of the squared error scoring function $S(x, y) = (x - y)^2$, the most used 'penalizing' function in academia:

-it is *consistent* for the mean functional relative to the class of probability measures on the real line with finite first moment;

-it is *strictly consistent* for the mean functional relative to the class of probability measures with finite second moment.

It turns out that consistency is the dual of the optimal point forecast property;

this indicates that there is a close conjunction between the task of finding optimal point forecasts and that of evaluating point predictions.

Theorem 3.5. *The scoring function S is consistent for the functional T, relative to the class $\mathcal{F}$, if and only if, given any $F \in \mathcal{F}$, any $\mathbf{x} \in T(F)$ is an optimal point forecast under S.*

By way of explanation, this theorem declares that the set of scoring functions that are consistent for a specific functional is identical to the set of the loss functions under which that functional is an optimal point forecast.

3.1.2 Back to Elicitability

Although the term *elicitability* was forged in recent times by Lambert et al. [35], the general idea dates back to the exceptional work of Osband [40].

Definition 3.6 (Elicitability). The functional T is *elicitable* relative to the class $\mathcal{F}$ if there exists a scoring function S that is strictly consistent for T relative to $\mathcal{F}$.

Undoubtedly, if T is elicitable relative to the class $\mathcal{F}$, then it is elicitable relative to any subset $\mathcal{F}\prime \subseteq \mathcal{F}$.
In order to fix ideas, we now provide an example of $elicitable functional$.
Take into account the squared error scoring function: $S(x, y) = (x - y)^2$.
We can observe that $S(x, y)$ is minimized by the mean functional.
Expressly, set $f(y) = E[S(x, y)]$. Then, we have:

$$f(y) = E[(x - y)^2] = E[x^2 - 2xy + y^2] = E[x^2] - 2yE[x] + y^2$$

To minimize the expression above, we compute the first derivative with respect to y, and set it equal to zero.

$$f_0(y) = -2E[x] + 2y = 0, \implies y* = E[x].$$

Furthermore, it emerges that this function is strictly consistent for the mean functional relative to the set of probability distributions on $\mathbb{R}$, whose second moment is finite.

Hence, means or expectations are elicitable.

Savage [44] was the first who recognized the class of scoring function consistent for the mean functional, and here we report his disclosure.

Theorem 3.7 (Savage)**.** *Let $\mathcal{F}$ be the class of probability measure on the interval $I \subseteq \mathbb{R}$ with finite first moment. Then the following holds:*

(a) The mean functional is elicitable relative to the class $\mathcal{F}$

(b) Suppose that the scoring function S satisfies assumptions (1), (2) and (3) on the Prediction-Observation domain. Then S is consistent for the mean functional, relative to the class of the compactly supported probability measures on I, if and only if it is of the form

$$S(x, y) = \phi(y) - \phi(x) - \phi^{'}(x)(y - x) \tag{3.4}$$

where ϕ is a convex function with subgradient $\phi^{'}$ on I.

(c) If ϕ is strictly convex, the scoring function 3.4 is strictly consistent for the mean functional, relative to the class of the probability measure F on I for which both $E_F[Y]$ and $E_F[\phi(Y)]$ exist and are finite.

Banerjee et al. [6] call a function of the type 3.4 a *Bregman function.*

Carrying forward, we introduce a a remarkable theorem presented by Gneiting [25], who stresses its variety of applications.

Theorem 3.8. *Let the functional T be defined on a class $\mathcal{F}$ of probability distributions which admit a density f, with respect to some dominating measure on the domain D. Consider a measurable weight function $\omega : D \to [0, \infty)$.*

Let $\mathcal{F}^{(\omega)} \subseteq \mathcal{F}$ denote the subclass of the probability distributions in $\mathcal{F}$ which are such that $\omega(\mathbf{y})f(\mathbf{y})$ has finite integral over D, and the probability measure $\mathcal{F}^{(\omega)}$

with density proportional to $\omega(\mathbf{y})f(\mathbf{y})$ belongs to $\mathcal{F}$. Define the functional

$$T^{\omega} : \mathcal{F}^{(\omega)} \to \mathcal{P}(I), \quad F \mapsto T^{\omega}(F) = T(\mathcal{F}^{(\omega)}), \tag{3.5}$$

on this subclass $\mathcal{F}^{(\omega)}$. Then the following holds:

(a) If T is elicitable, then T^{ω} is elicitable.

(b) If S is consistent for T relative to $\mathcal{F}$, then

$$S^{(\omega)}(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{y})S(\mathbf{x}, \mathbf{y}) \tag{3.6}$$

is consistent for $T^{(\omega)}$ relative to $\mathcal{F}^{(\omega)}$.

(c) If S is strictly consistent for T relative to $\mathcal{F}$, then $S^{(\omega)}$ is strictly consistent for $T^{(\omega)}$ relative to $\mathcal{F}^{(\omega)}$.

State differently, a weighted scoring function is consistent for the functional $T^{(\omega)}$, which takes the initial functional T and applies it to the probability measure whose density is proportional to the product of the weight function and the starting density.

For instance, if we consider the following scoring function

$$S_{\beta}(x, y) = \left|1 - \left(\frac{y}{x}\right)^{\beta}\right|$$

we can see it is of the form of equation 3.5, with initial scoring function $S(x, y) = |x^{-\beta} - y^{-\beta}|$ and weight function $\omega(y) = y^{\beta}$, defined on $D = (0, \infty)$.

Given that the scoring function S is consistent for the median functional, the scoring function S_{β} will be consistent for the β-median functional, $med^{(\beta)}(F)$.

Specifically, the class $S_{\beta}(x, y)$ of scoring functions defined above, comprehends both the absolute percentage error (when $\beta = -1$), and the relative error (if $\beta = 1$) scoring functions.

Now, we are going to present another theorem due to Osband [40]. It provides a further essential characterization of an elicitable functional.

Theorem 3.9 (Osband)**.** *If a functional is elicitable, then its level sets are convex in the following sense : if $F_0 \in \mathcal{F}, F_1 \in \mathcal{F}$ and $\rho \in (0,1)$ are such that* *$F_\rho = (1-\rho)F_0 + \rho F_1 \in \mathcal{F}$, then $\mathbf{t} \in T(F_0)$ and $\mathbf{t} \in T(F_1)$ imply $\mathbf{t} \in T(F_\rho)$.*

This result allow us to easily understand that, for example, the sum of two quantiles, not having convex level sets, is not an elicitable functional.

3.2 Elicitability of VaR

We have pointed out that a functional is said to be elicitable, if there exists a scoring function strictly consistent for it. Hence, in order to see whether the quantile of a distribution is an elicitable functional, we should verify the presence of this type of scoring functions. Thomson [47] and Saerens [43] provide a first characterization of the scoring functions that are consistent for a quantile and then for the Value at Risk. Here, we record a theorem that compiles their findings.

Theorem 3.10 (Thomson, Saerens)**.** *Let $\mathcal{F}$ be the class of the probability measures on the interval $I \subseteq \mathbb{R}$, and let $\alpha \in (0,1)$.* [2]
Then the following holds:

(a) The α-quantile functional is elicitable relative to the class $\mathcal{F}$.

(b) Suppose that the scoring function S satisfies assumptions (1), (2) and (3) on the Prediction-Observation domain $\mathcal{D} = I \times I$. Then S is consistent for the α-quantile relative to the class of the compactly supported probability measures on I if, and only if, it is of the form

$$S(x,y) = (\mathbb{1}(x \geq \; y) - \alpha)(g(x) - g(y)) \tag{3.7}$$

where g is a nondecreasing function on I and $\mathbb{1}$ denotes the indicator function .

[2] Note that in this theorem α represents the probability level, i.e. $1-$ the confidence level.

(c) If g is strictly increasing, the scoring function above is strictly consistent for the α-quantile relative to the class of the probability measure F on I for which $E_F[g(Y)]$ exists and is finite.

Gneiting [25] calls a function of the type of the equation 3.7 *generalized piecewise linear* (GPL) of order $\alpha \in (0, 1)$. The motivation behind this name is a mathematical one. Indeed (3.7) is piecewise linear only after applying a nondecreasing transformation. Notice that, any GPL function is equivariant with respect to the class of the nondecreasing transformations, just as the quantile functional is equivariant under monotone mappings (Koenker [32]). We conclude that quantiles are elicitable.

3.3 Elicitability of ES

We have shown that Expected Shortfall is a coherent risk measure.
However, it is not elicitable.
Weber [48] and Gneiting [25] have shown that ES does not have convex level sets and hence, by mean of the Osband theorem 3.9, it does not comply with the property of elicitability.
The following theorem essentially shows that, consistent scoring functions for the Expected Shortfall are not available.

Theorem 3.11. *The ES functional is not elicitable relative to any class of $\mathcal{F}$ of probability distributions on the interval $I \subseteq \mathbb{R}$, that contains the measures with finite support, or the finite mixtures of the absolutely continuous distributions with compact support.*

According to many authors, this fact makes it difficult to use the ES as a predictive measure of risk and, in some ways, it demonstrates why there exist almost no literature about the evaluation of ES forecasts.
We will analyze this issue in more deeply in the next chapter.
For what we are concerned now, we should emphasize the work by Emmer et al.

[22]. They state that, even if it lacks the elicitability property, Expected Shortfall is conditionally elicitable.

Definition 3.12 (**Conditional Elicitability**)**.** The functional T on $\mathcal{F}$ is called *conditionally elicitable* if there exist two functionals $\tilde{\gamma}$ and γ such that

$$T(F) = \gamma(F, \tilde{\gamma}(F))$$

where $\tilde{\gamma}$ is elicitable with respect to $\mathcal{F}$ and γ is such that γ_c defined by

$$\gamma_c : \mathcal{F} \to 2^{\mathbb{R}}, \ F \mapsto \gamma(F, c) \subset \mathbb{R}$$

is elicitable relative to $\mathcal{F}$ for all $c \in \mathbb{R}$.

If we think of Expected Shortfall as a combination of $E[L|L \geq \ c]$ and $c = q(L)$, where L denotes the loss, we immediately realize that ES functional is elicitable conditionally on VaR.
Stated differently, ES is elicitable under the condition of evaluating only one tail of the distribution. In fact, when we consider the values of the random variable L that are beyond the quantile at confidence level α, the ES represents their expected value. Hence, since we know that the mean functional satisfies the elicitability property, we can affirm the following:

i) ES is not elicitable if we deal with the entire loss distribution;

ii) ES is elicitable if we consider only the tail marked by VaR; moreover it is elicitable solely with respect to that "portion" of distribution.

3.4 Elicitability of Expectiles

In section 2.4 we have introduced the expectiles as the minimizer of an asymmetrically weighted squared loss function.[3]

We will now see that they have properties similar to those of quantiles.

Indeed, as well as the α-quantile matches the Bayes rule (3.2) under an asymmetric piecewise linear scoring function, the same way the β-expectile (whenever the second moment of F is finite,) corresponds to the Bayes rule under the asymmetric piecewise quadratic scoring function:

$$S_\beta(x, y) = |\mathbb{1}(x \geq \; y) - \beta|(x - y)^2 \tag{3.8}$$

The following theorem identifies the class of scoring function consistent for expectiles. The attentive reader should notice similarities with both the Bregman and the GPL families.

Theorem 3.13. *Let $\mathcal{F}$ be the class of the probability measures on the interval $I \subseteq \mathbb{R}$ with finite first moment, and let $\beta \in (0, 1)$. Then the following holds:*

(a) The β-expectile functional is elicitable relative to the class $\mathcal{F}$.

(b) Suppose that the scoring function S satisfies assumptions (1), (2) and (3) on the PO domain $\mathcal{D} = I \times I$. Then S is consistent for the β-expectile relative to the class of the compactly supported probability measures on I if, and only if, it is of the form

$$S(x, y) = |\mathbb{1}(x \geq \; y) - \beta|(\phi(y) - \phi(x) - \phi^{'}(x)(y - x)) \tag{3.9}$$

where ϕ is a convex function with subgradient $\phi^{'}$ on I.

(c) If ϕ is strictly convex, the scoring function above is strictly consistent for the β-expectile relative to the class of probability measure F on I for which both $E_F[Y]$ and $E_F[\phi(Y)]$ exist and are finite.

[3] Note that here we use the letter F (instead of L) to denote the loss function.

From all these considerations, it figures out that elicitability of expectiles is a simple corollary of their definition. Therefore, expectiles occur to be the only example of an *elicitable coherent risk measure*.
This result does not seem to be totally unexpected; in effect, expectiles can be seen as a generalization of both the mean (after having added asymmetry in the loss function) and the quantiles.
It seems as if Expectiles were ideal to fill up the deficiencies of both VaR and ES. However, the reader should remember what we have said about comonotonic additivity: for $\frac{1}{2} < \beta < 1$ Expectiles are not *comonotonically additive*.
This means that they could underestimate the risk exposure of a financial position, since they do not detect the dependence of different instruments on the same risk factors.
Nevertheless, unable to deny their engaging statistical properties, we hope that the aptitude of Expectiles as risk measures will be further researched.

Find below a table that summarize what we have discussed so far.
In the next chapter we are going to examine whether (and how) the good properties a risk measure should have, influence forecast verification and backtesting.

Property	VaR_α	ES_α	Expectiles($\beta \geq \frac{1}{2}$)
Coherence		X	X
Convexity		X	X
Comonotonic additivity	X	X	
Law Invariance	X	X	X
Robustness *w.r.t. the weak topology*	X		
Robustness *w.r.t. the Wasserstein distance*	X	X	X
Elicitability	X		X
Conditional Eliticitability	X	X	X

Table 3.2: Properties of standard risk measures as in Emmer et al. [22], extended.

Chapter 4

Backtesting

Regulators experience the crucial but onerous task of establishing capital requirements for financial intermediaries.
Such capital requirements have two important, conflicting goals:

1. prevent institutions from taking exceptional risks (in this thesis we concentrate on market risk).
2. not preclude them from working at one of their fundamental business, namely trading risk.

Hence, there exists a sort of trade off regulators need to assess; the basic ingredients they necessitate are: a risk measurement method, a backtesting procedure and multiplications factors (based on the outcome of the backtesting procedure).
We have already evaluated the pros and cons of VaR and ES as two potential risk measurement methods, and we know that VaR is still preferred by many experts, due to its perceived superior performance in case of backtesting.
In this chapter we are going to describe what backtesting is, and how it can be implemented for both Value at Risk and Expected Shortfall risk metrics.[1]

[1] The reader should be informed that the backtesting of Expectiles goes beyond the scope of the thesis.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9_4, © Springer Fachmedien Wiesbaden 2016

4.1 The Backtesting Idea

What does backtesting mean? According to Kerkhof and Melenberg [31] it is a final diagnostic check on a risk model carried out by the risk management team, i.e. a set of statistical procedures aimed at inspecting whether real losses, observed ex post, are in compliance with what has been predicted.

Actually, the term backtesting is used in several distinct ways in finance. Generally, backtesting denotes either

1. an evaluation of the theoretical, presupposed performance of a planned trading strategy;
2. the assessment of financial risk models, by means of historical data on risk forecasts and profit and loss realizations.

It should be clear that our discussion will focus on the second issue, that is on considering the ex ante risk measure predictions, and testing them against the ex post verified losses.

Typically, the choice of the backtesting methodology is contingent on the kind of forecasts available. As stated in Emmer et al. [22] there exist backtesting methods for:

i) *Point forecasts*, e.g the value of a random variable (Y); they are usually defined in terms of conditional expectation, with respect to the accessible information at the time the forecast is made.

$$E[Y_{t+k}|\mathcal{F}(Y_s.s \leq \ t)]$$

ii) *Probability range forecasts or interval forecasts*; they outline an interval in which the forecast value is prophecy to be found, with a certain, predetermined probability p.

iii) *Forecasts of the complete probability distribution*

$$P[Y_{t+k} \leq \ . \, |\mathcal{F}(Y_s.s \leq \ t)]$$

.

VaR and ES forecasts belong to the second group.

Recently, Gneiting [25] has questioned the possibility of directly backtesting ES, as a result of the fact it is not *elicitable*.

Undoubtedly, we can affirm that elicitability provides a natural methodology to perform backtesting. However, this does not necessarily mean that a non-elicitable functional cannot be backtested.

Bellini and Bignozzi [8] state that, if we want to provide an incentive to an expected score minimizer forecaster to give an accurate assessment of a statistical functional, and if the functional at issue is elicitable, then the forecaster is induced to report a correct forecast by means of the expected score. Hence, if the functional (say T) is elicitable, a natural statistic to perform backtesting is given by the average expected score

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} S(T_i, Y_i), \quad n \in \mathbb{N}$$

where $T_1, \cdots, T_n$ are point forecasts of the functional T and $Y_1, \cdots, Y_n$ are outcomes of the random variable Y.

This well argue example suggests that implementing the backtesting procedure might be easier for a risk measure which satisfies the elicitability property, than for one that does not.

Despite all, our question has not been answered yet. We will try to find out whether there exist any real binding links between elicitability and backtesting.

In order to reach our goal, in section 4.3 we will refer to a very recent paper by Acerbi and Szekely [2], who argue that one of the most popular risk measures, although not elicitable, can be backtested.

4.2 Backtesting VaR

Since the VaR revolution in 1994, the supervisory authorities immediately recongnized the need for VaR backtesting methodologies. Indeed, the first researches on backtesting were published few years later, see for instance Kupiec [34] (1995) and Hendricks [28] (1996).

Here we will follow the approach taken by Christoffersen [14], who has been studying and reviewing this subject from 1998 until recent times.

A popular backtesting procedure is based on the so called *violation process* or *hit sequence* (as it is named in Christoffersen [14]).

We now provide a concise explanation of the reference topic.

Consider a continuous loss distribution. By definition of VaR at confidence level α, we have that the violation probability of the VaR number equals $1 - \alpha$:

$$P(L > VaR_\alpha(L)) = 1 - \alpha$$

.

Hence, we can delineate the violation process as

$$I_{t+1}(\alpha) = \mathbb{1}_{\{L(t+1) > VaR_\alpha(L(t+1))\}} \tag{4.1}$$

It follows that the hit sequence returns 1 if the loss in day t+1 is larger than the predicted VaR number, and 0 otherwise. To backtest a model, we need to build up a sequence $\{I_{t+1}\}_{t=1}^{T}$ (where T indicates the number of days in the testing period), revealing when the past exceedances materialized.

Certainly, we should expect a 1 with probability $1-\alpha$ and a 0 with probability α. Following this line of reasoning, we will say that a risk model for VaR estimations

has *correct unconditional coverage* if

$$P(I_{t+1} = 1) = E[I_{t+1}] = 1 - \alpha$$

and *correct conditional coverage* if

$$P_t(I_{t+1} = 1) = E_t[I_{t+1}] = 1 - \alpha$$

Note that correct conditional coverage implies correct unconditional coverage but not vice versa. Moreover, this model satisfies the *independence condition* if

$$I_{t+1}(\alpha) \ and \ I_{s+1}(\alpha) \ are \ independent \ for \ s \neq t$$

Roughly speaking, a correct unconditional coverage indicates that the proportion of VaR violations is not significantly different from $1 - \alpha$ across the days.
Instead, a correct conditional coverage suggests that the model gives a VaR hit with the right probability on every day, provided all the information available the day before.
Finally, independence means that exceedances are not clustered over time.
Under these terms, VaR hits are independent and identically distributed Bernoulli random variables, with success probability $1 - \alpha$;
Thus, we can think of VaR backtesting as questioning the hypothesis:

$$H_0 = I_{t+1} \sim \ i.i.d. \ Bernoulli(1 - \alpha)$$

More precisely, we need to investigate two different hypotheses:

1. the unconditional coverage hypothesis; [2]

$$H_0 = E[I_{t+1}] = \pi = 1 - \alpha$$

2. the conditional coverage hypothesis.

$$H_0 = E_t[I_{t+1}] = \pi_{t+1|t} = 1 - \alpha$$

[2] π is the sample average.

A crucial issue is the choice of the significance level of a test. It should be selected taking into account the costs of making two types of errors:

i) Type I error: probability of rejecting a correct model.

ii) Type II error: probability of not rejecting a wrong model.

Increasing the significance level implies larger Type I errors but smaller Type II errors and vice versa.[3]
According to Jorion [30], a statistically powerful test would efficiently minimize both of these probabilities.

4.2.1 Unconditional Coverage Tests

We first want to test if the unconditional probability of a violation in the risk model, π, significantly differs from the conjectured probability, p.
In this case it is sufficient to verify whether the number of violations, i.e. the 1s in the sequence, follows a binomial distribution: [4]

$$f(x) = \binom{T}{x} p^x (1-p)^{T-x}$$

As the number of observations increase, the binomial distribution can be approximated with a normal distribution; therefore, for discussing the null hypothesis, we can use a simple mean test:

$$z = \frac{x - pT}{\sqrt{p(1-p)T}} \approx N(0,1)$$

where pT is the expected number of exceptions and $p(1-p)T$ their variance.
We can also carry out unconditional coverage tests as likelihood ratio tests.

[3] Notice that a smaller Type II error implies a greater power of the test; indeed Power = 1− Type II error.

[4] To simplify notations, we call the number of exceptions x and set $1 - \alpha = p$.

The most widely popular likelihood ratio tests have been suggested by Kupiec [34], and here we will briefly explain one of them.

POF-Test[5]

The null hypothesis of the $Proportion\ Of\ Failure$ test is

$$H_0 : p = \hat{p} = \frac{x}{T}$$

and the idea is to find out whether there is a large discrepancy between the observed failure rate, $\hat{p}$ and the theoretical failure rate p.
The test is easily constructed as

$$LR_{uc} = -2ln\Big(\frac{(1-p)^{T-x}p^x}{[1-(\frac{x}{T})]^{T-x}(\frac{x}{T})^x}\Big) \tag{4.2}$$

Under the null hypothesis, that is under the assumption of a correct model specification, LR_{uc} is asymptotically χ_1^2 (chi-squared distributed) with one degree of freedom.
In case the value of the statistic is higher than the critical value of the χ_1^2 distribution, the null hypothesis will be rejected and the model recognized as incorrect. We should underline that the power of the test increases when the sample size gets larger. Thus, when a considerable amount of data is at our disposal, we are able to reject an inaccurate model without much difficulty.
Another shortcoming of the POF-Test is that it ignores the time when losses occur. As a consequence, it may fail to reject a model that produces clustered VaR violations and, coincidentally, clustered VaR violation are just what risk managers want to avoid, since large losses in rapid progression are more likely to lead to tragic events.
This is the fundamental reason why there exists one more type of coverage test, namely the conditional coverage tests.

[5] Also known as Kupiec Test.

4.2.2 Conditional Coverage Tests

There is strong evidence of time-varying volatility in daily asset returns as pointed out in Andersen et al. [4]. If the risk model fails to recognize such behavior the VaR will reply late to changing market conditions and VaR breaches will appear clustered over time.

Pritsker [41] highlights the gravity of this situation when VaR numbers are computed through Historical Simulation.

Therefore, prior to formulate a conditional coverage test, we need to address the independence problem: in an accurate model the exception today should not depend on whether or not an exception occurred the previous day.

The simplest way to test for dynamics in time series analysis, is examining the autocorrelation function and then conducting the well known Portmanteau or Ljung-Box type tests.

We will use this approach for backtesting as well.

Let γ_k be the autocorrelation at lag k for the violation process.

Plotting the autocorrelation function (for $k = 1, \cdots, m$) will show us the degree of interconnection between an exceedance in one of the last m trading days and a hit today.

Then the null hypothesis will be

$$H_0 : \gamma_k = 0 \ \ fork = 1, \cdots, m$$

The test statistic takes the form

$$LB(m) = T(T+2)\sum_{k=1}^{m} \frac{\gamma_k^2}{T-k} \sim \chi_m^2$$

where χ_m^2 denotes the chi-squared distribution with m degrees of freedom.

According to Berkowitz et al. [11], setting $m = 5$ provides good testing power in a realistic daily VaR backtesting setting.

Of course, we can also adopt a likelihood approach for carrying out independence tests.

Christoffersen's Interval Forecast Test

Assume that the dependence structure of the hit sequence can be depicted as a first order Markov chain with the subsequent transition probability matrix

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

We can interpret the numbers in the matrix as follow:

(a) π_{01} is the probability of having a violation tomorrow, conditional on today having no violations;

(b) π_{11} is the probability of tomorrow being a violation given today is also a violation;

(c) $1-\pi_{01}$ is the probability of a non-violation following a non-violation;

(d) $1-\pi_{11}$ is the probability of a non-violation succeeding a violation.

If we have available a sample of T observations, then we can write the likelihood function of the first-order Markov process as

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

where $T_{ij},\ i,j = 0,1$ is the number of observations with a j succeeding an i. Finding the Maximum Likelihood (ML) estimates means taking the first derivatives with respect to π_{01} and π_{11}, and setting them equal to zero.
We will end up with

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

And given the definition of probability, we have

$$\hat{\pi}_{00} = 1 - \hat{\pi}_{01}, \quad \hat{\pi}_{10} = 1 - \hat{\pi}_{11}$$

We are interested in discovering whether $\hat{\pi}_{01}$ statistically differs from $\hat{\pi}_{11}$, and more precisely whether $\hat{\pi}_{11}$ is larger than $\hat{\pi}_{01}$.
Indeed this would imply that it will be more likely to have two consecutive exceedances, than to have a violation following a non-violation.
We can test the independence hypothesis

$$H_0 : \hat{\pi}_{01} = \hat{\pi}_{11}$$

with the subsequent likelihood ratio test

$$LR_{ind} = -2ln\left(\frac{(1-\pi)^{T_{00}+T_{10}}\ \pi^{T_{01}+T_{11}}}{(1-\pi_{01})^{T_{00}}\ \pi_{01}^{T_{01}}\ (1-\pi_{11})^{T_{10}}\ \pi_{11}^{T_{11}}}\right) \tag{4.3}$$

where $\pi = \frac{T_{01}+T_{11}}{T_{00}+T_{01}+T_{10}+T_{11}}$
Asymptotically, the test follows a Chi-squared distribution with one degree of freedom, $LR_{ind} \sim \chi^2_1$.
Practically speaking, when carrying out the LR_{ind} test we may encounter in samples where $T_{11} = 0$. In that case, we substitute the aforementioned likelihood function with

$$L(\mathbf{\Pi_1}) = (1 - \hat{\pi}_{01})^{T_{00}}\ \hat{\pi}_{01}^{T_{01}}$$

Ultimately, we care about simultaneously testing both the properties of a good VaR model: the correct failure rate and the independence of exceptions.
We can test jointly for independence and correct coverage using the conditional coverage test

$$LR_{cc} = LR_{uc} + LR_{ind} \tag{4.4}$$

and $LR_{cc} \sim \chi^2_2$.
Christoffersen's framework allows inspecting whether the reason for not passing the test is caused by inaccurate coverage, clustered violations, or even both.

Nevertheless, Campbell [13] warns us that there are cases in which the model passes the joint test, but not the individual ones.

Therefore it is recommended to run the independence and the coverage test separately, even when the null of correct conditional coverage is not rejected.

Mixed Kupiec Test

Christoffersen's interval forecast test is unable to capture all forms of dependencies, since it considers only the association between outcomes of two successive days.

It is possible that likelihood of VaR violation today does not depend on whether a violation occurred yesterday, but whether the violation occurred, for instance, a week ago (Campbell [13]).

Accordingly, Haas [27] introduces an improved test for both independence and coverage, mixing the ideas of Kupiec and Christoffersen.

This Mixed Kupiec test also measures the time between exceptions, being able (at least potentially) to capture various form of dependence.[6]

In order to construct this *mixed* test, we need to build up the Likelihood Ratio statistic for each exception i.

$$LR_i = -2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)$$

where v_i is the time between exceptions i and $i-1$.

Once calculated the LR- statistics, we conduct an independence test assuming that there are n violations.

$$LR_{ind^n} = \sum_{i=2}^{n}\left[-2ln\left(\frac{p(1-p)^{v_i-1}}{\left(\frac{1}{v_i}\right)\left(1-\frac{1}{v_i}\right)^{v_i-1}}\right)\right] - 2ln\left(\frac{p(1-p)^{v-1}}{\left(\frac{1}{v}\right)\left(1-\frac{1}{v}\right)^{v-1}}\right)$$

[6] Actually, a similar test has been proposed also by Christofferssen and Pelletier [15]. This test is more powerful, and at the same time it does not require any additional information, compared to the Christoffersen's interval forecast test. The underlying idea is that, in case of independence, the VaR violations should be unconditioned of the time that has passed since the last exception. Hence we need a measure for the "waiting time" between two events.
Christofferssen and Pelletier [15] define the $no-hit\ duration$ as $D_i = t_i - t_{i-1}$. A correct model with coverage rate p should have an expected conditional duration of $\frac{1}{p}$ days and the no-hit duration should have no memory (Christofferssen and Pelletier [15]).

LR_{ind^n} is distributed as a χ^2 with n degrees of freedom.
Equivalently to what we have done above, we can combine the two test to obtain the *Mixed Kupiec Test*:

$$LR_{mix} = LR_{un} + LR_{ind^n}$$

The LR_{mix}-statistic is χ^2 with $n+1$ degrees of freedom.

4.2.3 Backtesting with Information Variables

Despite their simplicity, the tests analyzed so far might not have much power to disclose incorrect risk models, as they only handle information on past VaR hits. To overcome this limit, Christoffersen [14] propose to consider also the information in past relevant market variables.
The basic idea is to test the model trying to explain when and why a violation occur and we manage to do this, increasing the information set and consequently the power of our testing methodologies.
If we define the q-dimensional vector of information variables available at time t as X_t, then the null hypothesis of a well-specified risk model can be written as

$$H_0 : P(I_{t+1} = 1|X_t) = p \Leftrightarrow E[I_{t+1} - p|X_t] = 0$$

The hypothesis claims that the conditional probability of having a VaR exceedance on day t +1 should be independent of the information variables observed at time t and it should naturally coincide to the promised VaR coverage rate, p.
This corresponds to say that the conditional expectation of the hit sequence minus p should be equal to 0.
Engle and Manganelli [23] adopt a regression quantile approach and propose a dynamic quantile test to investigate the null hypothesis:

$$DQ = \frac{(I-p)'X(X'X)^{-1}X'(I-p)}{Tp(1-p)} \quad \sim \chi^2_q$$

where X is the $T \cdot q$ matrix of information variables and $(I - p)$ is the $T \cdot 1$ vector of hits, where each component has been subtracted by the expected coverage rate p.

Berkowitz et al. [10] observe that carrying out the DQ test using lagged VaR from a GARCH model and lagged violation, gives good results (also in term of power) in a sensible daily VaR examination.

We have defined X_t as the vector of information variables, so it is reasonable to ask what kind of variables we should include in our analysis.

Answering this question is a difficult task; indeed the choice of the information variables to enter depends on the particular portfolio at hand.

It should not be surprising that variables (reasonably) highly correlated with the future volatility of the portfolio should be included in the vector (e.g. in equity portfolios option implied volatility measures would be an obvious candidate, while in bond portfolios we would prefer variables such as term spreads and credit spreads).

4.2.4 Regulatory Framework

The backtesting framework established by the Basel Committee in 2013 is based on a regular comparison of the bank's daily risk measure with the realized profit or loss ("trading outcome").

The risk measures are intended to be larger than all but a certain fraction of the trading outcomes, namely the confidence level of the risk measure.

Comparing the risk measures with the trading outcomes simply means that the bank counts the number of times that the risk measures were larger than the trading outcome. The fraction actually covered can then be compared with the intended level of coverage to gauge the performance of the bank's risk model(BCBS [7]).

The model backtesting will be based on a VaR measure calibrated at the 99th percentile confidence level. Banks need to produce $T = 250$ risk measures forecasts.

According to the Committee, an exception or an outlier occurs when either the loss of a trading desk registered in a day of the backtesting period is higher than the corresponding daily risk measure given by the model.

In the case when either the P&L or the risk measure is not available or impossible to compute, it will count as an outlier (BCBS [7]).

In this framework the risk capital requirement depends not only on the portfolio risk, but also on the outcome of the the backtesting procedure:

$$CR_t = mf_t \cdot \rho(L_t)$$

where CR states for Capital Requirement, mf for multiplication factor and ρ is the risk measure computed on the loss distribution L. The subscript t indicates that all is calculated at time t, using the information up to time $t-1$.

The term mf_t is determined in consonance with the backesting results.

The Basel Committee ranks backtesting outcomes according to three categories: green, yellow and red zones.

Properly, the approach is called "Traffic Light Approach".

These three zones have been delineated and their boundaries chosen in order to balance type I and type II errors: under the assumption that the model covers 99% of outcomes, the probabilities of making the type I error (i.e. an erroneous rejection of an accurate model) after having selected a given number of exceptions as a threshold for rejecting a model, have been calculated.

In accordance to these probabilities, the threshold number for each of the three categories has been set.

Precisely, the Basel Committee has decided not to designate a single boundary:

> "[...] there is no threshold number of exceptions that yields both a low probability of erroneously rejecting an accurate model and a low probability of erroneously accepting all of the relevant inaccurate models. It is for this reason that the Committee has rejected an approach that contains only a single threshold."

The table above illustrates the Committee's fixed thresholds for the three zones and the increase in the scaling factor, which reflects the backtesting performance

Zone	Number of exception	Plus factor	Cumulative Probability
Green Zone	0	0.00	8.11
	1	0.00	28.58
	2	0.00	54.32
	3	0.00	75.81
	4	0.00	89.22
Yellow Zone	5	0.40	95.88
	6	0.50	98.63
	7	0.65	99.60
	8	0.75	99.89
	9	0.85	99.97
Red Zone	$\geq$ 10	1.00	99.99

Table 4.1: Traffic light approach (Basel Committee, 2013).

based on a sample of 250 observations.[7]

The table shows the plus factors which are needed to compute the multiplication factor. Indeed, $mf_t = 3 +$ plus factor.

As we can see, the Green Zone admits up to four violations in a sample of 250 observations.

Since an accurate model providing 99% coverage would be quite likely to produce as many as four exceptions, there is a little reason for concern raised by backtesting results that fall in this range (BCBS [7]).

With a number of exceptions going from five to nine, we are in the Yellow Zone. Outcomes in this range may suggest both accurate and inaccurate models. Clearly, suspicious about the validity of the model should grow as the number of violations approaches nine. This is reflected also by the increase in the scaling factor.

Finally, the tumble in the red zone generally leads to the conclusion that there exists a problem within the bank's model. As the reader can realize, it is extremely

[7] For other sample sizes, the boundaries should be deduced by calculating the binomial probabilities associated with true coverage of 99%. The yellow zone begins at the point such that the probability of obtaining that number or fewer exceptions equals or exceeds 95%. Similarly, the beginning of the red zone is defined as the point such that the probability of obtaining that number or fewer exceptions equals or exceeds 99.99%. (BCBS [7])

implausible that a well-specified model would autonomously generate ten or more exceptions from a sample of 250 trading outcomes.

Therefore, the regulator generally increase the multiplication factor from 3 to 4 and start exploring the deficiencies of the model at hand. The supervisor may require the bank to enhance its model right away.

According to Kerkhof and Melenberg [31], the backtesting procedure implemented by the Basel Committee described above has some serious shortcomings.

It assumes that, under the null hypothesis, the exceedances are *i.i.d*, while empirical evidence shows that this might not be the case.

Moreover, the above procedure does not take estimation risk into account which reveals itself in the fact that $\hat{\rho}(L)$ is not necessarily equal to $\rho(L)$.

Estimation errors in exceedances may be caused by inaccurate predictions of the risk measures, which in turn could be due to the limited amount of data at our disposal. A final drawback regards the treatment of the information; in fact summing up all our knowledge in one number (the risk measure) and then into one characteristic (outstrip or not a threshold), we lose significant information about the loss distribution.

Thus, we should entrust, in any case, also the backtesting methodologies we have previously described.

4.3 Backtesting ES

The disclosure in 2011 that the Expected Shortfall is not elicitable, spread out some skepticism about the possibility of its backtesting.

In October 2013, the Basel Committee chose to replace the VaR with the ES for establishing the capital burden of internal based models, but preserved VaR as the measure to backtest in the common way. This decision was criticized and the fact that there was no intention to deepen the discussion about Expected Shortfall backtesting methodologies widened the belief that there should be something wrong with ES as a risk measure.

Not everyone, however, was convinced.

Indeed (as Acerbi and Szekely [2] point out), if elicitability is really a necessary condition for backtesting a risk metric, why has VaR never been backtested by exploiting this property?

And what about the few but estimable works on ES backtesting such that the one by Kerkhof and Melenberg [31] where the authors even state that:

> "contrary to common belief, ES is not harder to backtest than VaR [...] Furthermore, the power of the test for ES is considerably higher."
> ?

At some point, some discordant opinions began to arise.

We will now present three model-independent, non-parametric methodologies for backtesting Expected Shortfall, which were introduced in the literature just some months ago by Acerbi and Szekely [2].

First of all we need to characterize our testing framework: we take into account only unconditional coverage tests for ES and we assume that independence of tail event occurrences is tested separately through visual inspection of VaR violation clusters.

We let L_t represent a bank's loss in day t for each t belonging to the testing period $(t = 1, \cdots, T)$. These losses are distributed according to a *real* but *unknowable* distribution F_L and forecasted by a *predictive* distribution P_L, the one also used to compute the VaR and ES risk metrics.

The random variables $\vec{L} = \{L_t\}$ are assumed to be independent, but not identically distributed. Moreover, there is no restriction about the variability of F_L and P_L over time.

We only assume that the distributions are continuous and strictly increasing; the reason behind this is that, under these assumptions, the ES can be seen as a Tail Conditional Expectation (equation 2.2 and equation 2.4).

Finally, to be in line with Basel VaR tests, our null hypothesis assumes that the forecast is accurate, while the alternative hypotheses are chosen to be only in the direction of risk underestimation (i.e. they presume that our risk measure predictions are too low).

4.3.1 Test 1.

The first test presented in Acerbi and Szekely [2] follows exactly from the idea of Tail Conditional Expectation. Indeed, given that

$$ES_{\alpha,t}(L) = E[L_t|L_t > VaR_{\alpha,t}(L)]$$

by simple algebra, we obtain

$$E\Big[\ \frac{L_t}{ES_{\alpha,t}(L)} - 1\Big|L_t - VaR_{\alpha,t}(L) > 0\ \Big] = 0$$

Consider now that the $VaR_{\alpha,t}(L)$ model has been already tested; we would like to investigate the magnitude of the verified violations against model predictions.
Let $I_t = (L_t - VaR_{\alpha,t}(L) > 0)$, be the indicator function of a VaR violation.
If $N_t = \sum_{t=1}^{T} I_t\ > 0$, we can easily construct the following test statistics:

$$Z_1(\vec{L}) = \frac{\sum_{t=1}^{T} \frac{L_t \cdot I_t}{ES_{\alpha,t}}}{N_T} - 1 \tag{4.5}$$

What we have done here is just constructing an average of normalized VaR violations ($\frac{L_t \cdot I_t}{ES_{\alpha,t}}$).
The underlying null hypothesis is

$$H_0 : P_t^{[1-\alpha]} = F_t^{[1-\alpha]},\ \forall t$$

where $P_t^{[1-\alpha]}(l) = min\big(1, \frac{1-P_t(l)}{1-\alpha}\big)$, is the tail distribution for $l > VaR_{\alpha,t}$
The alternatives are [8]

$$\begin{aligned} H_1 : ES^F_{\alpha,t} \geq\ & ES_{\alpha,t},\ \forall t\ and\ >\ for\ some\ t \\ VaR^F_{\alpha,t} =\ & VaR_{\alpha,t},\ \forall t \end{aligned}$$

Notice that the predicted VaR_α model is still correct under H_1; this actually means that this test is subordinated to a preliminary (non rejected) VaR test.
In fact, being an average over VaR exceedances, this test is absolutely indifferent to an exaggerated number of exceptions and hence, for it to make sense, we must validate the VaR model first.
Under these conditions, we can claim that

1. $E_{H_0}[Z_1|N_T > 0] = 0$
2. $E_{H_1}[Z_1|N_T > 0] > 0$

Proof. 1. Under the null hypothesis, we have that $ES_{\alpha,t} = ES^P_{\alpha,t} = ES^F_{\alpha,t}$ and (consequently) $I_t = I_t^P = I_t^F$.
Hence, conditioning first on the I_t's and then using the independence of the L_t [9], we get

$$\begin{aligned} E_{H_0}[Z_1|N_T > 0] &= E_{H_0}\Big[\frac{1}{N_T}\sum_{t=1}^{T}\frac{L_t I_t}{ES_{\alpha,t}}\ -1\Big|N_T > 0\Big] \\ &= E_{H_0}\Big[E_{H_0}\Big[\frac{1}{N_T}\sum_{t=1}^{T}\frac{L_t I_t}{ES_{\alpha,t}}\Big|I_1, \cdots, I_T\Big]\ -1\Big|N_T > 0\Big] \\ &= E_{H_0}\Big[\frac{1}{N_T}\sum_{t=1}^{T} I_t\frac{E_{H_0}[L_t|I_t^F]}{ES_{\alpha,t}}\ -1\Big|N_T > 0\Big] \\ &= E_{H_0}\Big[\frac{1}{N_T}\sum_{t=1}^{T} I_t\frac{ES^F_{\alpha,t}}{ES_{\alpha,t}}\ -1\Big|N_T > 0\Big] \\ &= 0 \end{aligned}$$

[8] $VaR^F_{\alpha,t}$ and $ES^F_{\alpha,t}$ denote the value of the risk measures when $L \sim F_L$.
[9] The independence of the L_t random variables allows us to condition on a single I_t.

The last step follows from H_0 and from the definition of N_T, which is the overall number of exceptions.

2. Assume H_1. Given that $VaR^F_{\alpha,t} = VaR_{\alpha,t}$, we get $I^F_t = I^P_t = I_t$ also in this case. Moreover, remembering that $ES^F_{\alpha,t} \geq ES_{\alpha,t} > 0$, along the same reasoning as before, we obtain

$$\begin{aligned} E_{H_1}[Z_1|N_T > 0] &= E_{H_1}\left[\frac{1}{N_T}\sum_{t=1}^{T} I_t \frac{E_{H_1}[L_t|I_t]}{ES_{\alpha,t}} - 1\middle|N_T > 0\right] \\ &= E_{H_1}\left[\frac{1}{N_T}\sum_{t=1}^{T} I_t \frac{E_{H_1}[L_t|I^F_t]}{ES_{\alpha,t}} - 1\middle|N_T > 0\right] \\ &= E_{H_1}\left[\frac{1}{N_T}\sum_{t=1}^{T} I_t \frac{ES^F_{\alpha,t}}{ES_{\alpha,t}} - 1\middle|N_T > 0\right] \\ &> 0 \end{aligned}$$

□

Thus, the attained value $Z_1(\vec{l})$ is expected to be zero, and indicates a problem when it turns out to be positive.

The idea behind this test is the same as in the test developed by McNeil and Frey [36], who perform ES backtesting in a GARCH-EVT context. They construct a test based on model residuals, which under the null hypothesis of correct ES predictions, should behave like an i.i.d. sample with mean zero. To investigate the null hypothesis, they use a bootstrap test that makes no assumption about the underlying distribution of the residuals.

4.3.2 Test 2.

The second test we are going to describe, derives from the representation of the Expected Shortfall as an unconditional expectation:

$$ES_{\alpha,t}(L) = E\left[\frac{L_t \cdot I_t}{1-\alpha}\right]$$

From the above equation we can delineate the following test statistic

$$Z_2(\vec{L}) = \frac{\sum_{t=1}^{T} \frac{L_t \cdot I_t}{ES_{\alpha,t}}}{T(1-\alpha)} - 1 \tag{4.6}$$

The null and the alternative hypothesis will be

$$\begin{aligned} &H_0 : P_t^{[1-\alpha]} = F_t^{[1-\alpha]},\ \forall t \\ &H_1 : ES_{\alpha,t}^F \geq\ ES_{\alpha,t},\ \forall t\ and\ >\ for\ some\ t \\ &\quad VaR_{\alpha,t}^F \geq\ VaR_{\alpha,t},\ \forall t \end{aligned}$$

As before, we can state that

1. $E_{H_0}[Z_2] = 0$

2. $E_{H_1}[Z_2] > 0$

Proof. 1. From the identity $ES_{\alpha,t} = E\Big[\frac{L_t I_t}{1-\alpha}\Big]$, we easily obtain

$$E_{H_0}\Big[\frac{L_t I_t}{1-\alpha} \cdot \frac{1}{ES_{\alpha,t}} \ - 1\Big] = 0,\ \forall t$$

Then, by definition of Z_2 we have that

$$\begin{aligned} E_{H_0}[Z_2] =&\ E_{H_0}\Big[\sum_{t=1}^{T} \frac{L_t \cdot I_t}{T \cdot (1-\alpha) \cdot ES_{\alpha,t}} - 1\Big] \\ =&\ \frac{1}{T}\sum_{t=1}^{T} E_{H_0}\Big[\frac{L_t I_t}{1-\alpha} \cdot \frac{1}{ES_{\alpha,t}} \ - 1\Big] \\ =&\ 0 \end{aligned}$$

2. Under H_1 we have that $VaR_{\alpha,t}^F \geq\ VaR_{\alpha,t}$; this implies that $I_t^F \leq\ I_t$ and thus that $ES_{\alpha,t}^F \geq\ ES_{\alpha,t}$ for all t and $ES_{\alpha,t}^F >\ ES_{\alpha,t}$ for some t. Then, we

get

$$
\begin{aligned}
E_{H_1}[Z_2] = &\ E_{H_1}\Big[\sum_{t=1}^{T}\frac{L_t \cdot I_t}{T\cdot(1-\alpha)\cdot ES_{\alpha,t}} - 1\Big] \\
= &\ \frac{1}{T}\sum_{t=1}^{T} E_{H_1}\Big[\frac{L_t I_t}{1-\alpha}\Big]\cdot\frac{1}{ES_{\alpha,t}} \ -1 \\
= &\ \frac{1}{T}\sum_{t=1}^{T}\frac{ES^F_{\alpha,t}}{ES_{\alpha,t}} \ -1 \\
> &\ 0
\end{aligned}
$$

□

The last inequality follows from the fact that, given H_1, the mean ratio will be strictly higher than 1.

Remarkably, we do not exploit the independence property of the L_t for proving the above proposition.

We can also observe that this second test evaluates both frequency and magnitude of the $(1-\alpha)$-tail events. Indeed, from equation 4.5 and equation 4.6 we have

$$Z_2(\vec{L}) + 1 = (1 + Z_1(\vec{L}))\cdot\frac{N_T}{T(1-\alpha)}$$

and then the following relationship holds[10]

$$Z_2(\vec{L}) = (1 + Z_1(\vec{L}))\cdot\frac{N_T}{T(1-\alpha)} - 1 \tag{4.7}$$

We should point out that both tests 1 and 2 could have been defined under the weaker null hypothesis

$$
\begin{aligned}
H_0^{'} : ES^F_{\alpha,t} &= ES_{\alpha,t},\ \ for\ all\ t \\
VaR^F_{\alpha,t} &= VaR_{\alpha,t},\ \ for\ all\ t
\end{aligned}
$$

[10] Remember that $E_{H_0}[N_T] = T\cdot(1-\alpha)$.

and all the results presented above would have held true.

However, this choice would have complicated the assessment of the significance of the tests. Indeed, in order to calculate the p-value, $p = P_Z(Z(\vec{l}))$, of a realization $Z(\vec{l})$, we need to conduct a Monte Carlo simulation (i.e. simulate the distribution of the test statistic P_Z under H_0). Specifically, we have to

i) simulate independent $L_t^i \sim P_t, \quad \forall t, \forall i = 1, \cdots, M$

ii) compute $Z^i = Z(\vec{L}^i)$

iii) estimate $p = \sum_{i=1}^{M} \frac{(Z^i > Z(\vec{l}))}{M}$

where M is a suitably large number of scenarios.

Then, given a significance level[11] ϕ, the test is finally not-rejected if $p > \phi$.

As we can see, contrary to what is needed for VaR, for backtesting ES it may be necessary to keep memory of the entire predictive distribution P_t.

In truth, for the tests examined so far, it is sufficient to record the $(1-\alpha)$-tail $P_t^{[1-\alpha]}$ of the foreseen distribution, since $L_t \cdot I_t$ can be simulated considering $I_t \sim \ Bernoulli(1-\alpha)$.

Moreover, Acerbi and Szekely [2] have demonstrated that, thanks to the remarkable stability of Z_2 critical levels across different distribution types, there is no need to do a Monte Carlo test (and therefore no need to store predictive distributions), when this second test is developed. [12]

Testing Z_2 demands to record only two numbers per day, the magnitude $L_t \cdot I_t$ of a $VaR_{\alpha,t}$ violation and the forecasted $ES_{\alpha,t}$.

[11] Here we mean the significance level of the test.

[12] Table 4 in Acerbi and Szekely [2] exhibits the 5% and the 0.01% significance thresholds for Z_2 (like the ones in the Basel traffic-light approach), for Student-t distributions with different degrees of freedom and mean. It turns out that these thresholds diverge considerably only for greatly heavy tailed distributions and, anyway, the most serious error would be to overestimate the risk. As the reader can imagine this is not a prime issue, as extremely heavy tails represent a danger by themselves.

4.3.3 Test 3.

The last test proposed by Acerbi and Szekely [2], reflects the idea (already existent in Berkowitz et al. [11]) that it is possible to backtest the tail(s) of a model by checking if the observed ranks $U_t = P_t(L_t)$ are i.i.d. $U(0,1)$, as they should be if the predictive distribution is the right one.[13] The rank U_t is simply the cumulated probability associated to the observed loss L_t.

Indeed, consider a model that at end of each day generates a cumulative distribution estimate for the next day's return (P_t). This means that, given the actual performance, we are able to compute the probability (implied by the risk model) of experiencing a loss below the current. We denote this transform probability U_t. Now, if we are using the correct risk model to approximate the real unknowable loss distribution, then we should not be able to predict such a thing.

For this reason, the time series of observed ranks should be distributed independently over time as a Uniform (0,1) variable.

In other words, to backtest the tail of our loss distribution, we have to check if the random variables $\vec{U} = \{U_t\}$ are uniformly distributed in the interval (0,1).

Since here we are interested in backtesting ES, it is necessary to adapt this idea to create a specific test for our risk metric.

First of all, we must assign to each quantile its weight in terms of money and hence, we need to construct an ES estimator based on i.i.d. draws, $\vec{Y} = \{Y_t\}$ from a general (but continuous and strictly increasing) distribution:

$$\hat{ES}_{\alpha}^{(N)} = \frac{1}{[N(1-\alpha)]} \sum_{i}^{[N(1-\alpha)]} Y_{i:N} \tag{4.8}$$

[13] We should underline that it is also possible to backtest the full distribution.
However, in risk management, we are not concerned in backtesting the entire distribution, since this may lead us to reject a model that explain well the tails, but not the interior of the distribution. Given that we are dealing with the Loss distribution, we are interested only in backtesting the upper tail.

where [x] is the integer part of x, and $Y_{i:N}$ denotes ordered statistics.

Keeping this in mind, we can define the following test statistics

$$Z_3(\vec{X}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{ES}_\alpha^{(T)}(P_t^{-1}(\vec{U}))}{E_V\left[\hat{ES}_\alpha^{(T)}(P_t^{-1}(\vec{V}))\right]} - 1 \tag{4.9}$$

where $\vec{V}$ are i.i.d $U(0,1)$.

The idea here is to recalculate the ES as a mean above the quantile $P_t^{-1}(\vec{U})$, for every past day $t = 1, \cdots, T$, and then take an average of the result.

It can be proved that the estimator in equation 4.8, is biased. This is the reason why, in the above test statistic we do not normalize by $ES_{\alpha,t}$ as in the other two tests, but in order to compensate for the bias, we divide by a finite sample estimate.[14]

As we can expect, the underlying hypothesis involve now the entire distributions:[15]

$$H_0 : P_t = F_t, \forall t$$
$$H_1 : P_t \succeq F_t, \forall t \; and \succ for \; some \; t$$

Also in this case we can conclude that:

1. $E_{H_0}[Z_3] = 0$
2. $E_{H_1}[Z_3] > 0$

Proof.
1. Under H_0 we have that $U_t \sim U(0,1)$; moreover, by definition, also $\vec{V}$ are i.i.d. $U(0,1)$ and hence, our ratio reduces to 1 for all t.

   Then, the claim easily follows.

2. Given H_1, we have that the variables $P_t^{-1}(U) \sim F_t$ are stochastically dominated by $P_t^{-1}(V) \sim P_t$.

[14] The denominator can be computed analytically as:

$$E_V\left[\hat{ES}_\alpha^{(T)}(P_t^{-1}(\vec{V}))\right] = \frac{T}{[T(1-\alpha)]}\int_0^1 I_{1-p}(T - [T(1-\alpha)], [T(1-\alpha)])P_t^{-1}(p)dp$$

where $I_x(a,b)$ is a regularized incomplete beta function.

[15] $(\succeq) \succ$ denotes (weak) first order stochastic dominance.

Consequently, $E_{H_1}[\hat{ES}_\alpha^{(T)}(P_t^{-1}(\vec{U}))] \geq E_V[\hat{ES}_\alpha^{(T)}(P_t^{-1}(\vec{V}))]$, $\forall t$ and $>$ for some t. Loosely speaking, the considered ratio will be higher or equal than one for all t, and strictly higher for some t.

Then, the conclusion is straightforward.

□

Compared to the other two tests, this one is definitely less intuitive. Nevertheless, as the reader can imagine, it is extremely general.

4.3.4 Power of the tests

To evaluate the power of their tests, Acerbi and Szekely [2] conduct some experiments based on Student-t distributions.

They choose Student-t as a reference distribution because by modifying the degree-of-freedom parameter v, it is possible to consider a huge number of distributions differing a lot in terms of heavy tails.

It turns out that, whenever we assume alternative hypothesis exploiting distributions with different volatilities, Z_2 will be the most powerful test, while Z_3 and Z_1 will perform better in case of H_1 distributions differing only w.r.t. tail indexes.

Moreover the authors compare the results obtained with those of VaR Basel test and discover that their Expected Shortfall backtesting methodologies exhibit more power than the standard Value at Risk backtest.

In view of this, we recognize that contrary to common belief backtesting Expected Shortfall is possible and (in general) not really troublesome.

There are no theoretical limitations in carrying out the three tests presented above: they seems to be a viable alternative to the currently used VaR tests.

We conclude this chapter and consequently our discussion about *elicitability and backtesting* quoting again Acerbi and Szekely [2] :

> " [...] Elicitability has in fact nothing to do with backtesting."
>
> "Elicitability allows to compare in a natural way (yet not the only way)

> different models that forecast a statistics in the exact same sequences of events [...]. But this is *model selection* not *model testing*. It's a relative ranking not an absolute validation."

Thus, even if financial intermediaries may call for elicitable functionals for selecting the best model among different competing ones, they do not need elicitability for validating individual models on an absolute scale.

Chapter 5

Empirical Analysis

In this chapter we present some concrete backtesting results for both Value at Risk and Expected Shortfall.

The major scope of this work is to apply the new ES backtesting methodologies (described in Chapter 4) to real financial time series.

5.1 Data

Our empirical analysis is based on univariate data.

In particular, we have chosen five global market indexes[1]: S&P 500, DAX, FTSE 100, NIKKEI and EURO STOXX 50.

In order to compute our risk measures and perform their backtesting, we use daily observations (specifically daily adjusted close prices) from 1/01/2000 to 15/03/2015: we have 3843 daily prices and consequently 3842 log-returns[2].

We estimate the loss function by means of five different models (which we present later), using a rolling window of 250 days. In this way we end up with 3592 VaR and ES estimates.

[1] Data has been downloaded from Bloomberg platform.

[2] Here we refer to the case of the S&P 500 Index. Indeed given some calendar divergences, we get few more prices for the Eurozone Indexes and few less for the Nikkei Index.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9_5, © Springer Fachmedien Wiesbaden 2016

All time series exhibit two well known facts in finance: leptokurtosis and heteroskedasticity.

Figure 5.1 shows the S&P 500's log returns for the considered time period and its histogram[3]; we can observe that daily log returns have zero mean but exhibit volatility clusters. In addition, the histogram reveals that the distribution is skewed to the right and presents heavy tails.

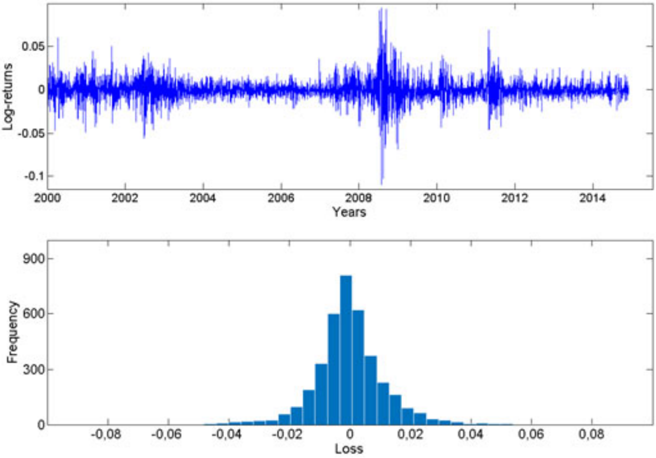


FIGURE 5.1: S&P 500 log-returns & Histogram.
(Losses are positive numbers)

Figure 5.2 shows the Quantile-Quantile plot of the S&P 500 versus the standard normal distribution. It highlights the leptokurtosis of the time series: when the left end pattern is below the line corresponding to normal quantiles and the right end is above it, the sample distribution is said to be heavy tailed.

5.2 Models

There exist different approaches for estimating the loss distribution of a portfolio (in our case a single time series). We can classify them into three categories:

[3] In this section we use the S&P 500 as the reference time series. Similar figures for the other indexes are provided in AppendixB.

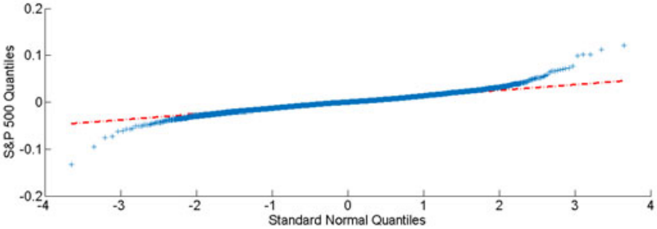


FIGURE 5.2: QQ plot - S&P 500 vs Standard Normal.

non parametric historical simulation methods; parametric models based on known distributions; and methods exploiting econometric models for capturing volatility dynamics. We are going to use one or two models for each of these classes.

5.2.1 Normal Distribution

The normal distribution is one of the standard methods used in finance to measure market risk. Its popularity is basically due to its simplicity: it provides a straightforward solution for estimating the loss distribution of a financial portfolio.
In particular the VaR and the ES for the Gaussian distribution are given by:

$$VaR_\alpha = \mu + \sigma\ \Phi^{-1}(\alpha) \ \ and \ \ ES_\alpha = \mu + \sigma\ \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}$$

where Φ denotes the standard normal distribution function and ϕ its density (for a formal proof see McNeil et al. [37], pag. 45).
However, this plainness is achieved at the cost of some unrealistic assumptions. As we have already seen, the loss distribution of a financial series' log-returns is heavier-tailed than the Gaussian one. This implies that our risk measures, which are based on the right hand tail will tend to underestimate the risk.
This effect amplify, the more we move towards higher confidence levels.
Although the hypothesis of log-returns being normally distributed is often rejected in both theoretical and empirical studies, in real world the normal distribution still plays a lead role.
This is the reason why it has been included in our study as the *base case*.

5.2.2 Student's t-distribution

Since log-returns of our time series are leptokurtic, to better estimate our risk measures we need a distribution with more probability mass in the tails.

The natural first choice is the Student's t-distribution; the standard Student's t-distribution has zero mean, and variance determined through its degrees of freedom parameter. In view of fitting our data in a proper way, we perform a location-scale transformation through which we can easily shift the center of the distribution and affect its dispersion by increasing (decreasing) the scale parameter.

Assume that the loss variable L is such that $(L-\mu)/\sigma$ has a standard t distribution with ν degrees of freedom. We write $L \sim t(\nu, \mu, \sigma)$ and we call t a t-locationscale distribution with mean μ and variance $\nu\sigma^2/(\nu-2)$, with $\nu > 2$. Notice that σ is not the standard deviation of the distribution.

Also in this case we are able to compute VaR and ES with no much effort:

$$VaR_\alpha = \mu + \sigma \; t_\nu^{-1}(\alpha) \;\; and \;\; ES_\alpha = \mu + \sigma \; \frac{g_\nu(t_\nu^{-1}(\alpha))}{1-\alpha} \left(\frac{\nu + (t_\nu^{-1}(\alpha))^2}{v-1}\right)$$

where t_ν denotes the distribution function and g_ν the density of the standard t (McNeil et al. [37], pag. 46).

Figure 5.3 shows the Normal and Student's t distributions fitted to the S&P 500. We can easily understand how much more the normal distribution underestimates the the tail risk w.r.t the Student's t.

Nevertheless, the Student't is far from being completely faithful to real extreme losses.

5.2.3 Kernel Density Estimation

Kernel smoothed densities belong to the class of non parametric historical simulation (HS) methods.

In the HS-approach the loss distribution of a portfolio is simply given by the empirical distribution of its past gains and losses. Basically, we avoid making

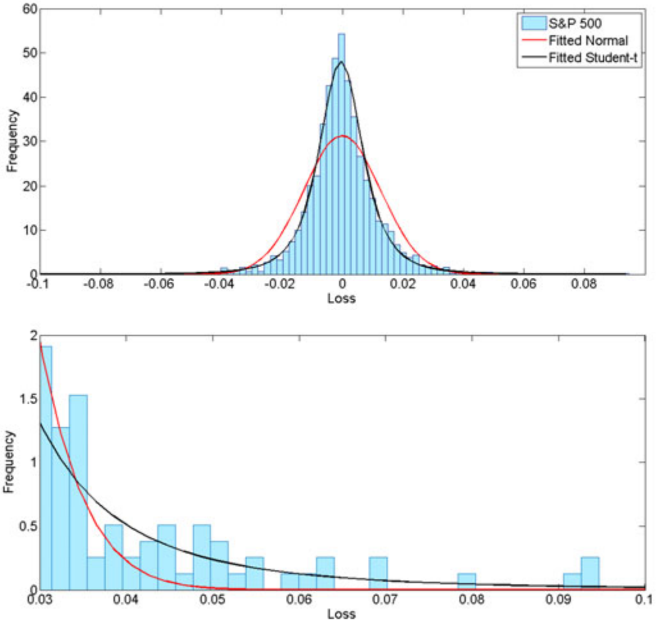


FIGURE 5.3: S&P 500 vs Fitted Normal & Student-t

assumptions about the distribution from which observations are drawn and ”let the data speak”. Although this method is easy to implement, it suffers from some serious drawbacks. First, the success of the approach is highly dependent on our ability to collect sufficient quantities of relevant data; second, this is an unconditional model and so we need a number of extreme scenarios in the historical record to provide more informative estimates of the tail of the loss distribution (McNeil et al. [37]); and finally estimates of extreme quantiles are notoriously difficult and tend to be inefficient.

To partially offset these shortcomings, we have smoothed the empirical distribution through a Kernel estimator. The kernel estimator is defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - Xi}{h}\Big)$$

where n is the sample size and h is the smoothing parameter (also called *window width* or *bandwidth*). The kernel function $K(\cdot)$ is everywhere non-negative and satisfies the following condition

$$\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$$

Thus, $\hat{f}$ will be a density function itself and will share all the continuity and differentiability properties of the kernel function $K(\cdot)$.

There exist different types of kernel functions. In our empirical study we have chosen the Gaussian one.

The most difficult decision you have to take when using a kernel density estimator is the choice of the bandwidth: small values of h reduce the bias of the estimator but increase its variance; large values of h have exactly the opposite effect.

When the Gaussian kernel is used the optimal window width (i.e. the one which minimize the Mean Integrated Squared Error) is

$$h_{opt} = 1.06 \; s \; n^{-1/5}$$

where s is the sample standard deviation.

However, for heavy tailed distributions it is preferable to use a robust measure of variability, namely the R inter-quantile range of the underlying normal distribution (Silverman [45], pages 45-47):

$$h_{opt} = 0.79 \; R \; n^{-1/5}$$

Given the above discussion about leptokurtosis of financial time series we have selected this second option.

Figure 5.4 illustrates how the Gaussian kernel smooths the data.

In order to fit the distributions described so far, we use the MATLAB function `fitdist (distname, x)`. Then, for computing the VaR we invert the obtained

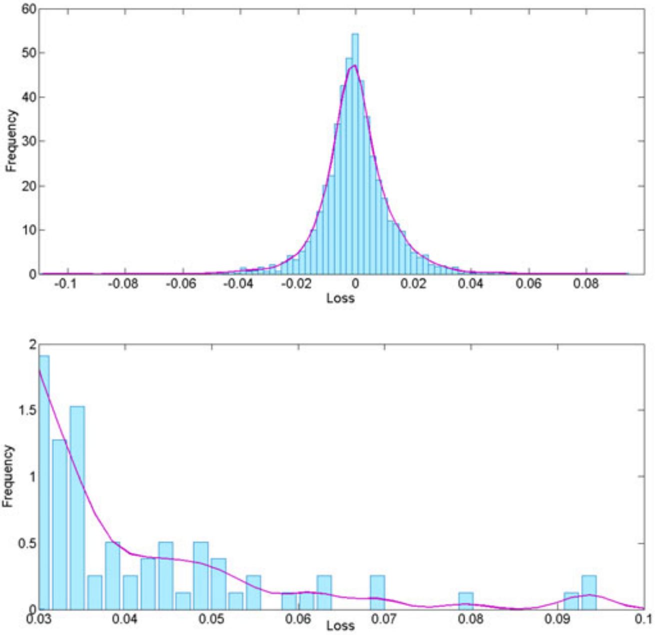


Figure 5.4: S&P 500 vs Fitted Gaussian Kernel

distribution using the MATLAB function `icdf(pd,x)`. Finally, the ES is calculated using a discretized version of equation 2.7 (see MATLAB code in Appendix A).

5.2.4 GARCH Models

Econometric models of volatility dynamics such as Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models provide VaR and ES estimates which reflect the current volatility background (Bollerslev [12]).

In this context we assume that the behavior of log-returns can be expressed by[4]:

$$r_t = \sigma_t \cdot \epsilon_t$$

[4] We are assuming that the mean of daily log-returns is zero.

where ϵ_t is the innovation term and σ_t the volatility.
The fact that σ depends on time indicates that the volatility is not supposed to be constant and hence we need to model it with a conditional econometric model; more precisely we use the parsimonious but efficient GARCH(1,1) model:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

Regarding the innovation distribution we consider two cases:

1. ϵ_t are normal random variables independently and identically distributed;
2. ϵ_t are Student's t random variables independently and identically distributed;

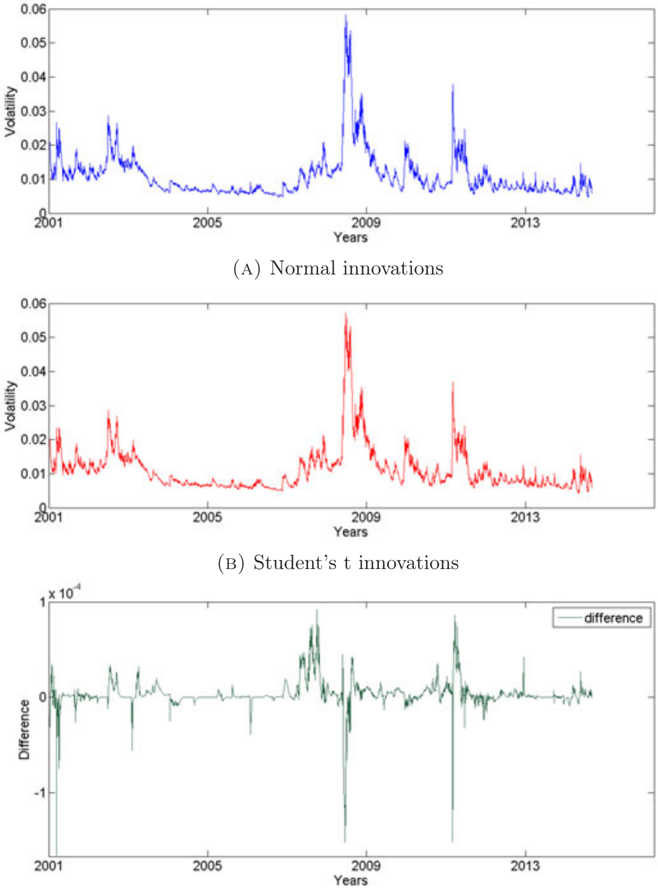
A GARCH type model can be fitted by Maximum Likelihood (ML). For each time $t \in [251, 3844]$, the coefficients of the model are estimated on the 250 previous observations. Then, σ_{t+1} is estimated using standard 1-step forecast; the first value for σ_t^2 that has to be used for the forecast is the average daily variance evaluated on the first 250 observations of log-returns (see the MATLAB code in Appendix A).
The behavior of the estimated volatility is displayed in Figure 5.6.
Given the conditional standard deviation, the VaR and the ES can be easily estimated as follow

$$VaR = \sigma_{t+1}\ \epsilon_q \ \ and\ \ ES = \sigma_{t+1}\ E[\ \epsilon \mid \epsilon > \epsilon_q\ ]$$

where ϵ_q is the upper quantile of the marginal distribution of ϵ_t, which by assumption does not depend on t.

5.3 Backtesting results

Given that a single backtesting procedure can never be enough to evaluate the goodness of a model, we apply two tests for each risk measure.



(A) Normal innovations

(B) Student's t innovations

(C) Difference (numbers of the order of 10^{-4})

FIGURE 5.5: Conditional standard deviations estimated by the GARCH(1,1) model.

Precisely the Kupiec test (or POF test) and the Christoffersen's Interval Forecast test for the VaR, and Test 1 and Test 2 by Acerbi and Szekely [2] for the ES. Without loss of generality, we consider the 99% confidence level for the VaR and the 97.5% for the ES. This shift in level was proposed in 2003 by Kerkhof and Melenberg [31] and has been recently adopted by the Basel Committee (BCBS [7]). $ES_{97.5\%}$ is correctly chosen to equal $VaR_{99\%}$ for Gaussian tails and penalize heavier tails (for a more detailed explanation refer to Kerkhof and Melenberg [31]).

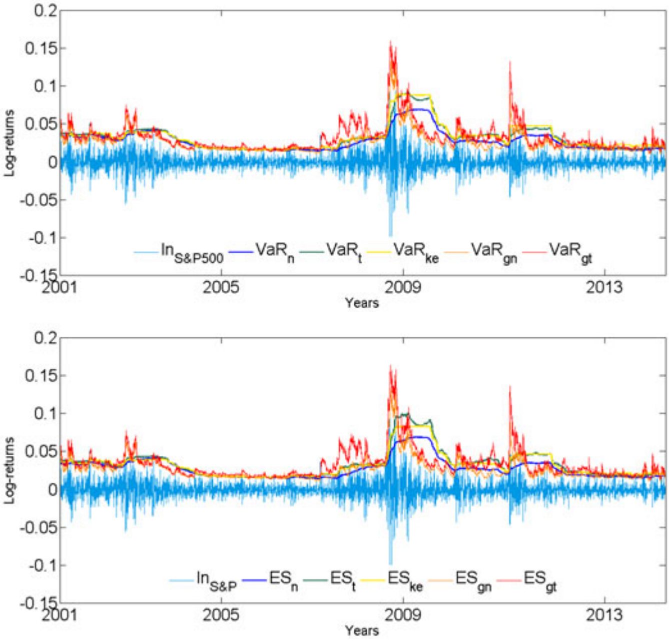


FIGURE 5.6: S&P 500: VaR and ES estimates

Figure 5.6 shows the estimated $VaR_{99\%}$ and $ES_{97.5\%}$ for each of the five models described above. We can easily recognize that the two measures are of equal magnitude.

In truth, since Test 1 is subordinated to a preliminary (non rejected) VaR test we also backtest the $VaR_{97.5\%}$.

5.3.1 VaR results

The five models we have fitted to our time series produce significantly different VaR estimates. To make this fact visible we have included a zoomed version of figure 5.6, figure 5.8.

As expected, the GARCH models promptly adapt to changes in volatility and hence display a very different behavior with respect to other models. At first look, their VaR estimates seem quite accurate.

On the opposite side we have the Normal model whose estimates are always below that of Student's t and Kernel ones; it seems to systematically estimate a lower level of the risk.

Notice however that after a period of high volatility (for example year 2008) our three unconditional models need a lot of time to re-adjust to sensible levels and VaR predictions turn out to be too conservative.

We can also observe that, except for the GARCH models, Value at Risk violations appear in clusters.

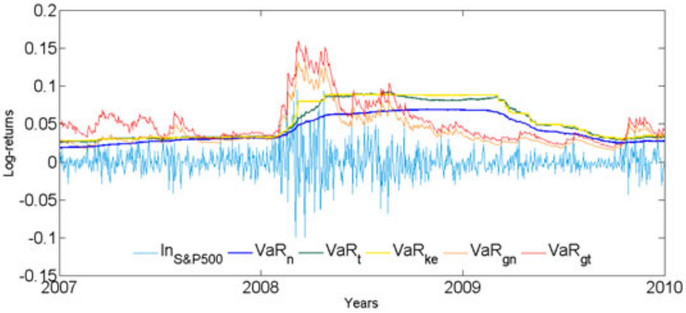


FIGURE 5.7: S&P 500: VaR estimates for different models (2007-2010).

Below we provide two tables: the first one shows the number of observations and the expected number of exceptions for each time series and each VaR confidence level; the second one displays the number of observed violations for each time series and each model.

The only two models that do not considerably exceed the thresholds are the Kernel density estimator and the GARCH with Student't innovations.

Given that it is quite interesting to find out when and why these violations occur we have grouped them by years (see figure 5.8).

	Confidence Level	**Number of observations**	**Expected number of exceptions**
S&P 500	97.5%	3592	90
	99.0%	3592	36
FTSE 100	97.5%	3728	93
	99.0%	3728	37
DAX	97.5%	3645	91
	99.0%	3645	36
NIKKEI	97.5%	3516	88
	99.0%	3516	35
EURO STOXX 50	97.5%	3666	92
	99.0%	3666	37

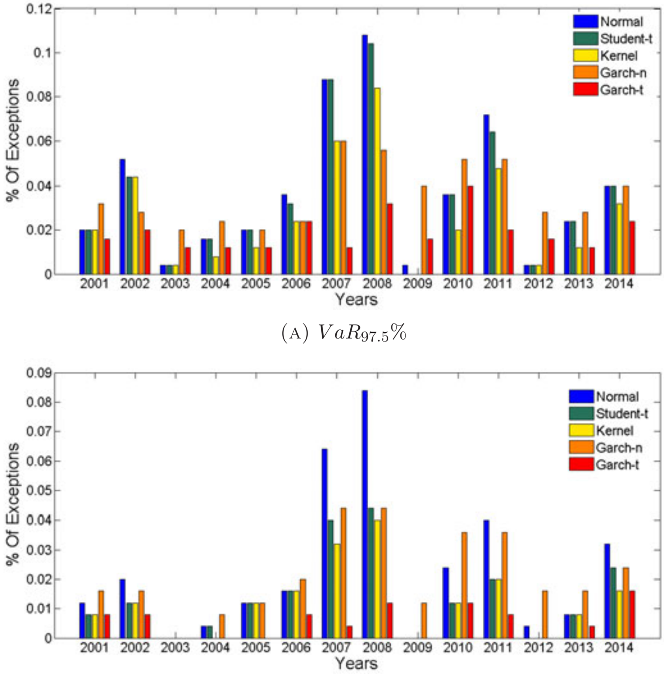
Table 5.1: Expected number of exceptions

	Normal	**Student't**	**Kernel**	**Garch- n**	**Garch- t**
S&P	137	130	95	132	71
500	84	50	44	79	22
FTSE	138	134	94	132	87
100	91	70	48	72	40
DAX	136	130	95	131	75
	77	54	37	58	24
NIKKEI	116	115	90	116	69
	75	58	41	54	33
EURO	119	115	92	119	73
STOXX 50	80	63	42	66	38

Table 5.2: Observed number of exceptions per model

Not surprisingly, in 2008 we observe the highest frequency of VaR exceptions and notably, the second worst year is not 2009 but 2007. This reflects the fact that our models calibrated in the biennium 2007-2008 finally succeed in reacting to the experienced volatility.

This is not to be considered a good thing: we would like to have models that quickly adapt to market conditions. The only model that shows the desired consistency is the GARCH-t.



(A) $VaR_{97.5}\%$

(B) $VaR_{99}\%$

FIGURE 5.8: S&P 500 - Percentage of VaR exceptions

In order to make our results legible also from the point of view of the Basel backtesting framework, we provide a table with the equivalent threshold for the Traffic Light Approach. They are calculated using the binomial probabilities associated to the considered true levels of coverage (97.5% and 99%) for a sample size of 3630 days.

The chosen sample size is the mean of the total number of observations in each time series. Reasonably, the identified thresholds work for all indexes.[5]

[5] The yellow zone begins at the point such that the probability of obtaining at maximum that number of exceptions equals or exceeds 95%. The starting point of the red zone is the one for which the same probability equals or exceeds 99.99%.

Zone	**97.5 %**	**99 %**
Green Zone	≤ 105	≤ 45
Yellow Zone	106-120	46-59
Red Zone	≥ 121	≥ 60

Table 5.3:

Traffic Light Approach based on a sample of 3630 observations.

If we look at tables 5.2 and 5.3 simultaneously, we can realize what would be the judgement of the Basel Committee for our five models: only the Kernel (with one exception) and the GARCH-t models fall inside the Green Zone.

Here we should add a comment about the Green Zone we have delineated: it is clear that the number of $VaR_{97.5\%}$ ($VaR_{99\%}$) exceedances must not be greater than 105 (45); however, a number considerably lower than the one expected (see table 5.1), may indicate that the model regularly overestimates the risk.

As a rule of thumb, a number of Value at Risk violations lower than 75 (26)[6], should signal too conservative models.

In order to see if the amount of exceptions is significantly high from a statistical point of view, we perform the Kupiec Test[7]. If statistically too many or too few exceptions are observed, the model is rejected.

The results for the S&P 500[8] index are shown in table 5.4. The entry "Test Outcome" is marked according to a significance level $\phi = 5\%$. Anyhow, given the p-value it is easy to verify which would be the outcome with another choice of ϕ.

We should point out that the GARCH-t model does not pass the test because the number of exceptions is overall too low, i.e. the risk level is overestimated.

While this is not an issue from regulators' point of view, it reasonably is from a bank's perspective whose main business is trading risk.

Looking at the tables for the other indexes we find out everywhere the same

[6] The binomial probability of observing a number of exceptions less or equal than that values is indeed about 5%.

[7] See MATLAB code "VaR_tests" in Appendix A.

[8] Similar tables for the other indexes are provided in Appendix B.

situation: the Normal, the Student's t and the GARCH-n models do not work at all. The kernel model and the GARCH-t model perform quite well in all the occasions.

	Kupiec (POF) Test				
Title : S&P 500	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic	21.97	16.25	0.30	17.81	4.34
$\mathbf{LR_{un}}$	47.21	4.96	1.71	38.89	6.32
P-value	0.0000	0.0001	0.5819	0.0000	0.0371
	0.0000	0.0258	0.1906	0.0000	0.0036
Test	X	X	√	X	X
Outcome	X	X	√	X	X

TABLE 5.4: S&P 500 - Kupiec Test results

The second test we conduct is the Christoffersen's Interval Forecast test (results are shown in table 5.5). We recall that this is a joint test of conditional coverage that combines the Kupiec test and the Christoffersen's Independence test (see equations 4.3 and 4.4).

Hence, the three steps we have to follow for obtaining the final results are[9]:

1. calculate the probabilities of having a violation tomorrow, conditional on today having a (no) violation(s);

2. counting the number of days T_{00} T_{01} T_{10} T_{11};

3. compute LR_{ind} (equation 4.3) and finally LR_{cc} (equation 4.4).

Also in this case the entry "Test Outcome" is marked according to a significance level $\phi = 5\%$. The results are mostly the same as the ones of Kupiec test.

[9] See also MATLAB code "VaR_tests" in Appendix A.

Notice that the great performance of the Kernel model in Kupiec test for the 97.5% confidence level is undermined by the failure of the model in this second test. Evidently, the few violations experienced occur in clusters.

Likewise the previous case, for the other time series we recognize the same circumstance: there is not even a case in which the Normal, the Student's and the GARCH-n models work. The Kernel model alternates successes to failures and the GARCH-t model behaves properly.

We can say for sure that the assumption of normality does not hold in financial time series, and even that the one of conditional normality does not reveal true. Regarding this latter case we should specify that the GARCH-n model usually works accurately for a confidence level $\alpha \leq 95\%$; for higher confidence levels (e.g. the ones discussed in our analysis) this approach often underestimate the conditional quantile (see the backtesting results in McNeil and Frey [36]).

Also the Student's t model proves to be inaccurate: even if it accounts for extreme events, he does not adapt to changes in volatility. This leads to inconsistent VaR estimates in periods of stress and high volatility.

Accordingly, the model that embodies both these features, the GARCH with Student's t innovation, shows the greatest performance.

	Christoffersen's Interval Forecast Test				
Title : S&P 500	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{LR_{cc}}$	27.33 50.71	21.20 6.38	6.12 2.80	17.97 39.27	5.78 6.59
P-value	0.0000 0.0000	0.0001 0.0412	0.0496 0.2460	0.0001 0.0000	0.0555 0.0370
Test Outcome	X X	X X	X √	X X	√ X

TABLE 5.5: S&P 500 - Christoffersen's Interval Forecast Test results

In commenting the results we should take into account the fact that the backtesting is performed over a period that includes the crisis of 2008. Precisely, in autumn 2008 stock prices were affected by macroeconomics events more than usual and experienced abnormally high volatility.

Given this context, we are not surprised by the many failures of our models[10] and instead we would like to emphasize the achievements of the Kernel model and above all of the GARCH-t one.

For further insights about the topic -"Value at Risk models and the financial crises of 2008" - we remind the interested reader to the work of Degiannakis et al. [18].

5.3.2 ES results

The five models we have fitted on our data behave differently and produce quite divergent VaR estimates. This is true also for ES estimates.

Since the tests we apply for backtesting Expected Shortfall are interested not only in the presence or absence of a violation but also in its magnitude, we provide a figure which enables us to observe both the things (figure 5.9).

Figure 5.9 shows the scatter plot of the S&P 500 log-returns[11]. The orange circles denote losses bigger than the $VaR_{97.5\%}$, while the red ones correspond to losses bigger than the $ES_{97.5\%}$ (and consequently also bigger than the $VaR_{97.5\%}$).

To justify this distinction we recall the test statistics of Test 1 and Test 2:

$$Z_1(\vec{L}) = \frac{\sum_{t=1}^{T} \frac{L_t \cdot I_t}{ES_{\alpha,t}}}{N_T} - 1 \quad and \quad Z_2(\vec{L}) = \frac{\sum_{t=1}^{T} \frac{L_t \cdot I_t}{ES_{\alpha,t}}}{T(1-\alpha)} - 1$$

In both Z_1 and Z_2 we have the summation of losses L_t greater than the VaR_α (α in our case is 97.5%) normalized by the ES_α at the same confidence level.

The orange circles tell us that there is a violation of Value at Risk but it is not so

[10] This is not to say that the considered models would perform well in other circumstances: even if there is a justifiable reason to believe that in a more stable market environment they would work better, we do not have studied that possibility (i.e. we do not have evidence to prove the claim).

[11] We remind that losses are positive numbers.

massive to exceed also the Expected Shortfall. This type of exception contributes to increase the value of the test statistic, but not in a “problematic” way[12].

[12] We recall that the value of Z_1 and Z_2 is zero under the null hypothesis and greater than zero under the alternative one; this means that a positive value of the test statistic indicates a problem.

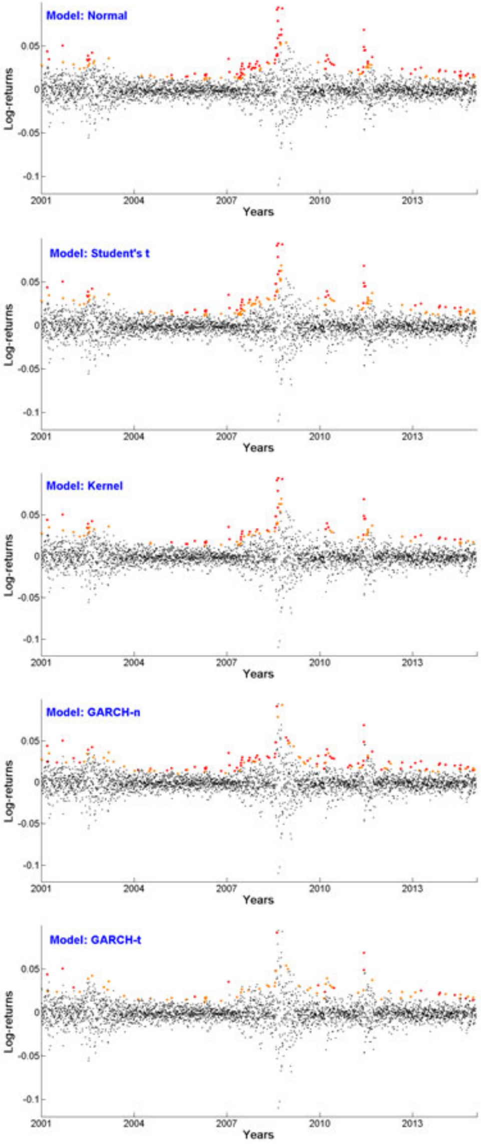


FIGURE 5.9: S&P 500: Log-returns and exceptions of $VaR_{97.5\%}$ (orange circles) and $ES_{97.5\%}$ (red circles).

Indeed, in the (extremely positive) case that all the violations belong to this "orange" type, the value of Z_1 (or Z_2) will be for sure lower than zero; this will lead to a non-rejection of the null hypothesis and hence of the model[13]. The red circles reveal that something is going wrong: the losses are not only bigger than the Value at Risk, but also bigger than the Expected Shortfall. They contribute to the increase in value of the test statistic in a "dangerous" way.

In the (extremely negative) case in which all the violations belong to this "red" type, the value of Z_1 (or Z_2) will be for sure greater than zero. Then, the rejection of the model will depend on the p-value associated to the resulting value of the test statistic.

Given these considerations, we expect that a model whose scatter plot exhibits "a lot" of red points will gain an higher value of the test statistic (and hence an higher probability of failing the tests) than a model presenting fewer "red type" violations[14].

Table 5.6 below displays the number of "orange type" and "red type" exceptions for the five models considered.

Notice that the sum of the two type of exceptions per model gives exactly the observed number of exceptions of $VaR_{97.5\%}$ in table 5.2.

	Normal	**Student's t**	**Kernel**	**Garch-n**	**Garch-t**
Orange Type violations	50	81	50	53	49
Red Type violations	87	49	45	79	22

TABLE 5.6: S&P 500 - Number of "orange type" and "red type" exceptions.

Table 5.7 shows the results of Test 1 for each model. With a significance level $\phi = 5\%$, no model except the GARCH-t passes the test.

We highlight that the failure of the Student's t model is due to the fact that it

[13] Since Test 1 is indifferent to an exaggerated number of exceptions, in this case we must verify whether the model passes a coverage VaR test before non-rejecting it.

[14] We remark that this is not statistical evidence but rather visual inspection.

passes neither the Kupiec nor the Christoffersen's test.

Since this model has not a good coverage, Test 1 outcome is simply meaningless. The result obtained for the GARCH-t is coherent with its scatter plot and with the total number of violations observed.

It is worth remarking that the p-values are computed using Monte Carlo simulations (see MATLAB code in Appendix A); each of the test statistic has its own distribution and we cannot compare the values of Z_1 and the associated probability between models.

For reasons of clarity we have included figure 5.10, which displays the Monte Carlo distribution of Z_1 for the Normal model and the GARCH-t one: the x-axis range is not the same for both the histograms.

	TEST 1				
Title : **S&P 500**	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{Z_1}$	0.1875	0.0217	0.0527	0.1128	-0.0088
P-value	0.0000	0.1359	0.0014	0.0000	0.6794
Test **Outcome**	X	X	X	X	√

Table 5.7: S&P 500 - Test 1 results

Table 5.8 shows Test 2 outcomes. The only difference with the previous results is the success of the Kernel model.

Contrary to the previous case, here we can also compare the values of the test statistic to that of other models.

Indeed, Acerbi and Szekely [2] demonstrate that the critical levels for Z_2 display remarkable stability across different distribution types.

This basically means that, given the same backtesting period, Z_2 takes values (more or less) in the same range, no matter the model considered.

Figure 5.11 highlights this fact. It shows the Z_2 Monte Carlo distribution for the

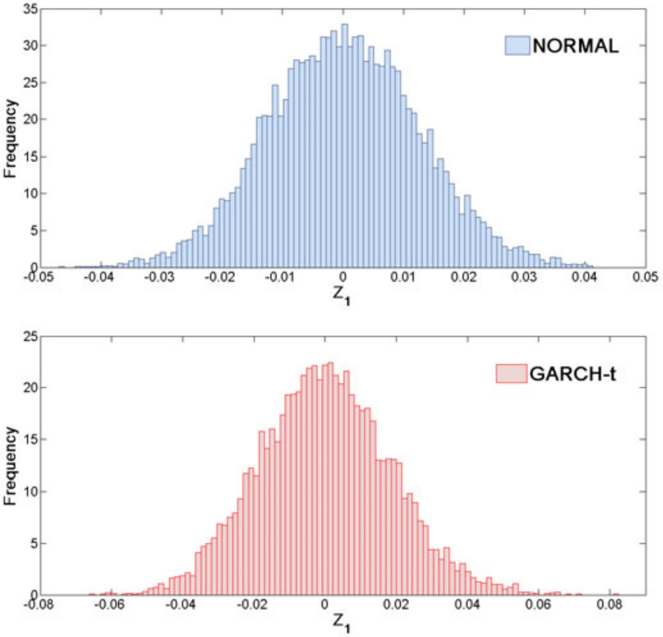


Figure 5.10: MC distributions for Z_1

Normal model and the GARCH-t one: even if these two models make very different assumptions about the S&P 500 log-returns distribution, the support of the two histograms is essentially the same.

We also notice that the test statistic for the GARCH-t model is negative and the corresponding p-value is closed to one. This may indicate that the model overestimates the shortfall risk.

Looking at the figure 5.11 (bottom panel) we realize that the value of Z_2 for the GARCH-t model lies in the left hand tail of the Monte Carlo distribution.

Although the studied tests are born to detect models that systematically underestimate the shortfall risk, it could be important also understand whether a model constantly overestimates that risk.

	TEST 2				
Title : **S&P 500**	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{Z_2}$	0.8116	0.4791	0.1137	0.6358	-0.2163
P-value	0.0000	0.0000	0.1415	0.0000	0.9833
Test **Outcome**	X	X	✓	X	✓

Table 5.8: S&P 500 - Test 2 results

Given a realized value of Z and a simulated distribution, drawing such conclusion is straightforward[15].

The shortfall risk is overestimated if

$$z^* \;<\; q_\phi^{MC} \tag{5.1}$$

or equivalently if

$$P(Z \leq z^*) \;<\; \phi \tag{5.2}$$

where z^* is the realized value of Z and q_ϕ^{MC} is the lower ϕ-quantile of the distribution, e.g. the quantile at 5% confidence level (see figure 5.12).

In our specific case the cumulative probability evaluated at $z^* = -0.2163$ equals:

$$P(Z \leq z^*) = P(Z_2 \leq -0.2163) = 1 - P(Z_2 > -0.2163) = 1 - \text{p-value} = 0.0167$$

This means that with a significance level $\phi = 5\%$ we have to reject the null hypothesis of consistent risk estimation (see section 4.3) and we conclude that the GARCH-t model overestimate the shortfall risk.

We get the same results for all the indexes (see tables in Appendix B) except for the FTSE 100. In this case the test statistic is negative but very close to zero and the associated cumulative probability is strictly greater than the chosen ϕ (i.e.

[15] With Z we denote a general test statistic. It could be either Z_1 or Z_2.

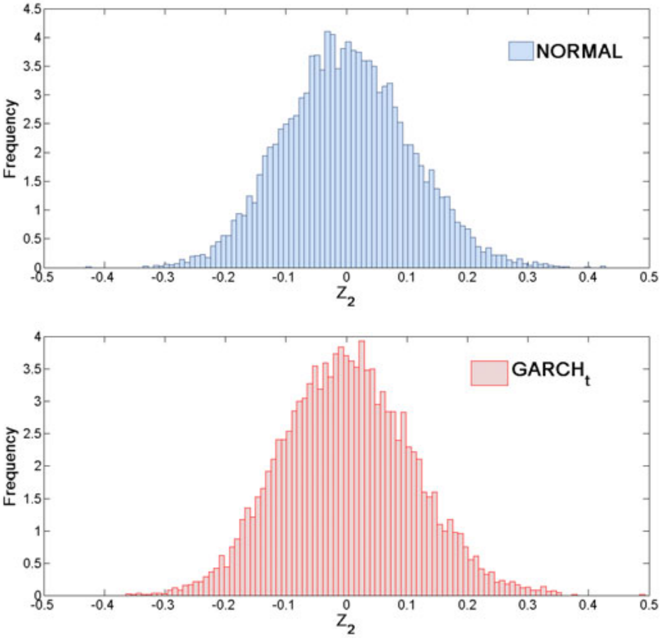


FIGURE 5.11: S&P 500 - MC distribution for Z_2

0.4647).

Summing up, according to Test 1 no model works except for the GARCH-t, while Test 2 claims that the Kernel model works quite well and the GARCH-t seems to overestimate the risk.

Since almost no model passes the VaR Kupiec test for the 97.5% confidence level[16], in the majority of cases we have to regard as invalid the Test 1 outcomes and rely on Test 2 ones.

Also in this case we have to be cautious in commenting the results: the backtesting period is quite long and includes different market situations (e.g. the 2008 crisis). Although this may help to understand how much a model is consistent, it may also give us a misleading picture of the models' performances.

In case a model is rejected we are not able to figure out the reason: it could be

[16] The Kupiec test results for the other four indexes can be found in Appendix B.

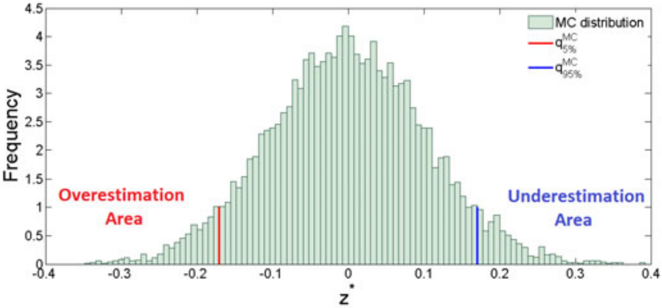


FIGURE 5.12: Overestimation an Underestimation Areas

that the model does not work for the entire period, or that it does not work now but in the past did, or finally that it did not work in the past but probably now does. In order to make everything a little bit clearer we provide two more figures: figure 5.13 and figure 5.14 show the value of the test statistics for each calendar year and for "cumulative" years respectively.
They have to be read as follows:

1. figure 5.13 tells us which would have been the values of Z_1 or Z_2 if we had performed the backtesting year by year (i.e. considering only the previous 250 days);

2. figure 5.14 tells us which would have been the values of Z_1 or Z_2 if we had performed the backtesting considering cumulative years: the first year and then the first two, three, etc.

Thanks to these charts we can see that the 2008 crisis has actually played a very important role in determining models' performances.
We can argue something more if we consider the case of Z_2 calculated in each calendar year: for a backtesting period of 250 days we can compare the Z_2 values with the critical levels provided by Acerbi and Szekely [2], namely 0.70 ($\phi = 5\%$) and 1.8 ($\phi = 0.01\%$).
Precisely, if the value of the statistical test is higher than 0.70 we reject the model at 95% confidence level (think of the Yellow Zone in Traffic Light Approach), if it

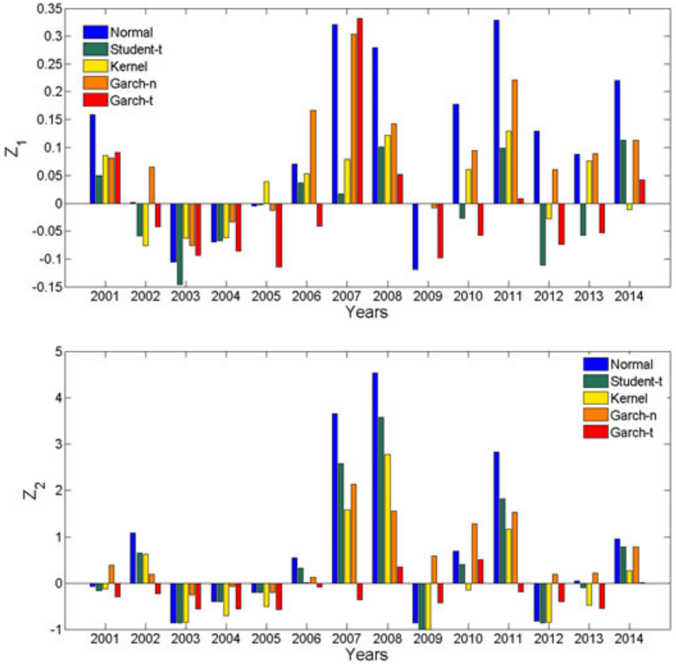


FIGURE 5.13: S&P 500 - Z_1 and Z_2 in each calendar year

is higher than 1.8 we reject the model with a 99.99% confidence level (it would be like falling in the Red Zone).

As expected, in the biennium 2007-2008 almost all the bars are above the thresholds (the GARCH-t model is a resounding exception).

It is worth recalling again that this comparison (with the mentioned critical levels) can be done only for Z_2 and for a backtesting period T=250 days.

The last consideration we have to make regards figure 5.14. It is likely that in the industry "cumulating years" does not make sense at all. In order to properly judge a model and find out the causes of its failure, we should instead look at what happens day by day. We should identify the days when the model has produced more (or larger) violations than expected, and then verify the reason.

Nevertheless our figure just wants to show what has broadly happened over years: once again it seems that the turning point was at the back of the crisis.

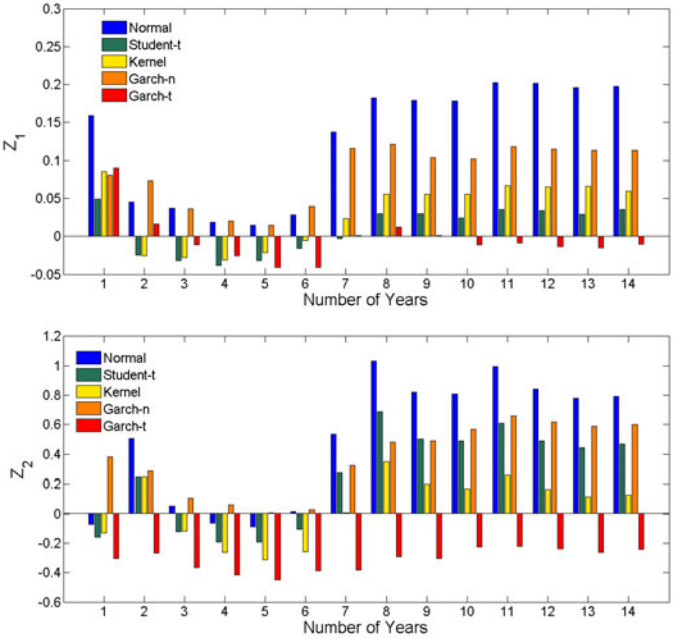


FIGURE 5.14: S&P 500 - Z_1 and Z_2 cumulative in years

Chapter 6

Conclusions

The popularity of Value at Risk in finance is unquestionable. It is the most widely used measure of risk.

However in recent years an increasing number of risk professionals starts thinking about replacing it with another risk measure: the Expected Shortfall.

In this work we have seen that, due to the lack of subadditivity, VaR is not a coherent risk measure. Even worse, it is not sensible to the tail risk of the loss distribution. Conversely, the Expected Shortfall is coherent and by definition sensitive to the sizes of potential losses beyond the threshold α. Nonetheless it carries a pitfall too, not being elicitable. This drawback is considered a weakness by many authors, who claim that a non elicitable functional cannot be backtestable.

To make the whole picture clearer we have analyzed numerous theoretical studies and we have reached the conclusion that elicitability does not mean backtestable.

Therefore we have focused our attention on the backtesting issue and we have investigated three new Expected Shortfall backtesting methodologies.

The most important contribution of this work is the application of “Test 1” and “Test 2” by Acerbi and Szekely [2] on real data, precisely on five global market indexes.

Differently to the well known VaR tests (e.g. Kupiec Test), “Test 1” and “Test 2” require the storage of the predictive cumulative distribution function for each day of the backtesting procedure. Thus, they may be relatively difficult to implement.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9, © Springer Fachmedien Wiesbaden 2016

In case of VaR backtesting we only need to count the number of exceptions or at most to record time between exceptions; the job consists of applying a test (i.e. computing the test statistic), observing the outcome and in case something turns out to be strange plotting a graph with the realized violations over time.

Poorly speaking, we just have to figure out if and when an exception occurs.

In case of Expected Shortfall backtesting we need something more: what matters is not only whether there is a violation or not, but also how large this violation is.

Another point is the interpretation of results. We have seen that one out of five models fitted on our data shows up to overestimate the risk.

Given that banks' main job is trading risk, overestimating it is actually a problem.

When performing Value at Risk backtesting (through Kupiec Test for example), a model that overestimates the risk is automatically rejected.

In case of Expected Shortfall backtesting this fact does not hold in general.

"Test 2" (for example) would return a really high p-value for the "wrong" model and one can be tempted to not reject it.

This is to say that a superficial interpretation of the results could lead to erroneous conclusions more easily in case of Expected Shortfall backtesting than in that of Value at Risk one.

Everything considered, we think that the Expected Shortfall will gain more and more visibility both in academia and industry and that its role as a risk measure will be sooner widely accepted. We also expect that it will never completely substitute the Value at Risk, but we believe that the correct and coordinated usage of both risk metrics can only have positive effects and bring improvements everywhere.

Appendix A

MATLAB Code

A.1 MATLAB variables

```
close all;
clear all;
clc
title= xlsread('datathesis.xlsx','foglio1','B2:B3844');
r_title=-price2ret(title);
dummy=zeros(t,l-t);
x_values = -0.2:0.005:0.3;
alpha = [.975,.99];
la=length(alpha);
t = 250;
l = length(r_title);
qe = zeros(la,l-t);
qn = zeros(la,l-t);
qt = zeros(la,l-t);
qke = zeros(la,l-t);
qga = zeros(la,l-t);
qgat = zeros(la,l-t);
ESe = zeros(1,l-t);
ESn = zeros(1,l-t);
ESt = zeros(1,l-t);
```

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9, © Springer Fachmedien Wiesbaden 2016

```
ESke = zeros(1,l-t);
ESga = zeros(1,l-t);
ESgat = zeros(1,l-t);
mu = zeros(1,l-t);
sigma = zeros(1,l-t);
r=zeros(1,l-t);
v = zeros(1,l-t);
s = zeros(1,l-t);
st = zeros(1,l-t);
const=zeros(1,l-t);
const2=zeros(1,l-t);
coeff= zeros(2,l-t);
coeff2= zeros(2,l-t);
total=0;
tot=0;
nsim = 10000;
M=10000;
Y= zeros(l-t,M);
Data= zeros(l-t,M);
Data2= zeros(l-t,M);
GAN= zeros(l-t,M);
GAT= zeros(l-t,M);
mdl = garch(1,1);
mdl2 = garch(1,1);
mdl2.Distribution = 't';
st1=zeros(2,5);
st2=zeros(2,5);
test1=zeros(2,5);
test2=zeros(2,5);
test12=zeros(2,5);
vex1=zeros(1,5);
vex2=zeros(1,5);
Z=zeros(1,5);
Z2=zeros(1,5);
test3=zeros(1,5);
test4=zeros(1,5);
p=0.025;
phi=0.05;
```

```
BtrialsN=rand(l-t,M)<p;
BtrialsT=rand(l-t,M)<p;
BtrialsKE=rand(l-t,M)<p;
BtrialsGN=rand(l-t,M)<p;
BtrialsGT=rand(l-t,M)<p;
```

A.2 ESTIMATION OF RISK MEASURES

A.2.1 Normal model

```
for j=1:l-t
    dummy(:,j)= r_title(j:t-1+j);
end

for k= 1:(l-t)
    disp(l-t-k)
    fitnormal = fitdist(dummy(:,k),'Normal');
    mu(k) = fitnormal.mu;
    sigma(k) = fitnormal.sigma;
    r(k) = iqr(fitnormal);
    for i=1:M
        if  BtrialsN(k,i)==1
            uniform= unifrnd(0.975,.999);
            Data(k,i) = icdf(fitnormal,uniform);
        end
    end
    qn(:,k) = icdf(fitnormal,[.975,.99]);
    cfl=.975:.000005:0.999;
    values = icdf(fitnormal,cfl);
    ESn(1,k) = mean(values)
end
```

A.2.2 Student's t model

```
for k=1:l-t
    disp(l-t-k)
    fitt = fitdist(dummy(:,k),'tLocationScale');
    v(k) = fitt.nu;
    for i=1:M
        if  BtrialsT(k,i)==1
            uniform= unifrnd(0.975,.999);
            Data2(k,i) = icdf(fitt,uniform);
        end
    end
    cfl=.975:.000005:0.999;
    qt(:,k) = icdf(fitt,[.975,.99]);
    values = icdf(fitt,cfl);
    ESt(1,k) = mean(values);
end
```

A.2.3 Kernel model

```
for k=1:l-t
    disp(l-t-k)
    bw=.79*r(k)*t^(-1/5);
    fitke = fitdist(dummy(:,k),'Kernel','Kernel','normal','Width',bw);
    for i=1:M
        if  BtrialsKE(k,i)==1
            uniform= unifrnd(0.975,.999);
            Y(k,i) = icdf(fitke,uniform);
        end
    end
    cfl=.975:.000005:0.999;
    qke(:,k) = icdf(fitke,[.975,.99]);
    values = icdf(fitke,cfl);
    ESke(1,k) = mean(values);
end
```

A.2.4 Garch with normal innovations

```
cfl=.975:.000005:0.999;
values1=norminv(cfl);

for k= 1:(l-t)
    disp(l-t-k)
    count=-1;
    while true
        try
            count=count+1;
            mdlp = estimate(mdl,r_title(k+count:t-1+k+count));
            break;
        catch
             total=total+1;
        end
    end
    const(1,k)=mdlp.Constant;
    coeff(1,k)=mdlp.ARCH{1};
    coeff(2,k)=mdlp.GARCH{1};
    if k == 1
        s0 = var(r_title(1:t));
        s(k) = forecast (mdlp, 1, 'Y0', r_title(t-1+k), 'V0', s0);
    else
        s(k) = forecast (mdlp, 1, 'Y0', r_title(t-1+k), 'V0', s(k-1));
    end

    epsilon1=norminv(.975);
    epsilon12=norminv(.99);
    qga(1,k)=sqrt(s(k))*epsilon1;
    qga(2,k)=sqrt(s(k))*epsilon12;
    ESga(k)=mean(values1)*sqrt(s(k));
    for i=1:M
        if  BtrialsGN(k,i)==1
            value=unifrnd(0.975,.999);
            eps=norminv(value);
            GAN(k,i)=sqrt(s(k))*eps;
        end
    end
```

A.2.5 Garch with Student's t innovations

```
    counter=-1;
    while true
        try
            counter=counter+1;
            mdlp2 = estimate(mdl2,r_title(k+counter:t-1+k+counter));
            break;
        catch
            tot=tot+1;
        end
    end
    const2(1,k)=mdlp2.Constant;
    coeff2(1,k)=mdlp2.ARCH{1};
    coeff2(2,k)=mdlp2.GARCH{1};
    dof= mdlp2.Distribution.DoF;
    if k == 1
        st0 = var(r_title(1:t));
        st(k) = forecast (mdlp2, 1, 'Y0', r_title(t-1+k), 'V0', st0);
    else
        st(k) = forecast (mdlp2, 1, 'Y0', r_title(t-1+k), 'V0', st(k-1));
    end
    epsilon2=tinv(.975,dof);
    epsilon22=tinv(.99,dof);
    qgat(1,k)=sqrt(st(k))*epsilon2;
    qgat(2,k)=sqrt(st(k))*epsilon22;
    values2=tinv(cfl,dof);
    ESgat(k)=mean(values2)*sqrt(st(k));
    for i=1:M
        if  BtrialsGT(k,i)==1
            value=unifrnd(0.975,.999);
            eps2=tinv(value,dof);
            GAT(k,i)=sqrt(st(k))*eps2;
        end
    end
end
```

A.3 Value at Risk Tests

```
function [t1,t2,test1,test2,test12,vex,vex0,vex1,vex2] = VaR_tests(q,r)

    TP=length(q);
    t=250;
    exc = zeros(2,TP);
    vex=zeros(14,1);
    vex0=zeros(14,1);
    p=[.025,.01];
    T00=zeros(TP-1,2);
    T01=zeros(TP-1,2);
    T10=zeros(TP-1,2);
    T11=zeros(TP-1,2);
    t1=zeros(2,1);
    t2=zeros(2,1);
    test1=zeros(2,1);
    test2=zeros(2,1);
    test12=zeros(2,1);
    slv=.95;
    for i=1:TP
        if q(1,i) > r(t+i)
            exc(1,i) = 0;
            exc(2,i)= 0;
        else
            exc(1,i) = 1;
            if q(2,i)> r(t+i)
                exc(2,i) =0;
            else
                exc(2,i) =1;
            end
        end
    end
    add=-t;
    for j=1:14
        add=add+t;
        vex0(j)=sum(exc(1,1+add:t+add+5))/t;
        vex(j)=sum(exc(2,1+add:t+add+5))/t;
    end
    vex1=sum(exc(1,:));
    vex2=sum(exc(2,:));

    for j=1:2
        vio= sum(exc(j,:));
        nv= 1-p(j);
        nx=TP-vio;
        fra=vio/TP;

        t1(j)=-2*log((nv^nx*p(j)^vio)/((1-fra)^nx*fra^vio))

        if t1(j)> chi2inv(slv,1)
            test1(j)=0;
        else
            test1(j)=1;
        end
    end

    for j=1:2
        for i=1:TP-1
            if (q(j,i)>r(t+i)) && (q(j,i+1)>r(t+i+1))
                T00(i,j)=1;
            else
                T00(i,j)=0;
            end
            if (q(j,i)>r(t+i)) && (q(j,i+1)<r(t+i+1))
                T01(i,j)=1;
            else
                T01(i,j)=0;
            end
            if (q(j,i)<r(t+i)) && (q(j,i+1)>r(t+i+1))
                T10(i,j)=1;
            else
                T10(i,j)=0;
            end
            if (q(j,i)<r(t+i)) && (q(j,i+1)<r(t+i+1))
                T11(i,j)=1;
            else
                T11(i,j)=0;
            end
        end

        noy= sum(T01(:,j));
        nono=sum(T00(:,j));
        yno=sum(T10(:,j));
        yy=sum(T11(:,j));
        pi01= noy/(nono+noy);
        pi11= yy/(yno+yy);
        pi= (noy+ yy)/(TP-1);

        if yy==0
            t2(j)= -2*log(((1-pi)^(nono+yno)*pi^(noy+yy))/...
                    ((1-pi01)^nono*pi01^noy));
        else
            t2(j)= -2*log(((1-pi)^(nono+yno)*pi^(noy+yy))/...
                    ((1-pi01)^nono*pi01^noy*(1-pi11)^yno*pi11^yy));
        end

        if  t2(j)> chi2inv(slv,1)
            test2(j)=0;
        else
            test2(j)=1;
        end

        if t1(j)+t2(j)> chi2inv(slv,2)
            test12(j)=0;
        else
            test12(j)=1;
        end
    end

end
```

```
[st1(:,1),st2(:,1),test1(:,1),test2(:,1),test12(:,1),
    vex1(:,1),vex2(:,1)] = VaR_tests(qn,r_title);

[st1(:,2),st2(:,2),test1(:,2),test2(:,2),test12(:,2),
    vex1(:,2),vex2(:,2)] = VaR_tests(qt,r_title);

[st1(:,3),st2(:,3),test1(:,3),test2(:,3),test12(:,3),
    vex1(:,3),vex2(:,3)] = VaR_tests(qke,r_title);

[st1(:,4),st2(:,4),test1(:,4),test2(:,4),test12(:,4),
    vex1(:,4),vex2(:,4)] = VaR_tests(qga,r_title);

[st1(:,5),st2(:,5),test1(:,5),test2(:,5),test12(:,5),
    vex1(:,5),vex2(:,5)] = VaR_tests(qgat,r_title);
```

A.4 Expected Shortfall Tests

```
function [ Z, S2, S3 ] = Z_1( q,es,r )


    t=250;
    TP=length(q);
    I=0;
    ratio=zeros(1,TP);
    s2=zeros(1,TP);
    S2=zeros(14,1);
    S3=zeros(14,1);
    for i=1:TP
        if r(t+i)> q(1,i)
            I=I+1;
            ratio(i)=r(t+i)/es(1,i);
            s2(i)=1;
        end
    end
    count=-t;
    for j=1:14
        count=count+t;
        S2(j)=sum(ratio(1+count:t+count))/sum(s2(1+count:t+count))-1;
        if j==1
            S3(j)= S2(j);
        else
            S3(j)=sum(ratio(1:t+count))/sum(s2(1:t+count))-1;
        end
    end

    Z(1)=sum(ratio(:))/I-1;

end

[Z(1),S(:,1),S3(:,1)]= Z_1(qn,ESn,r_title);
[Z(2),S(:,2),S3(:,2)]= Z_1(qt,ESt,r_title);
```

```
[Z(3),S(:,3),S3(:,3)]= Z_1(qke,ESke,r_title);
[Z(4),S(:,4),S3(:,4)]= Z_1(qga,ESga,r_title);
[Z(5),S(:,5),S3(:,5)]= Z_1(qgat,ESgat,r_title);

function [Z2,S] = Z_2( q,es,r)


    TP=length(q);
    t=250;
    ratio=zeros(1,TP);
    p=[.025,.01];
    S=zeros(14,1);
    for i=1:TP
        if r(t+i)> q(1,i)
            ratio(i)=r(t+i)/es(1,i);
        end
    end

    count=-t;
    for j=1:14
        count=count+t;
        S(j)=sum(ratio(1+count:t+count))/(t*p(1))-1;
    end
    Z2=sum(ratio(:))/(TP*p(1))-1;

end

[Z2(1),S2(:,1)]= Z_2(qn,ESn,r_title);
[Z2(2),S2(:,2)]= Z_2(qt,ESt,r_title);
[Z2(3),S2(:,3)]= Z_2(qke,ESke,r_title);
[Z2(4),S2(:,4)]= Z_2(qga,ESga,r_title);
[Z2(5),S2(:,5)]= Z_2(qgat,ESgat,r_title);
```

A.5 Monte Carlo p-values

```
for model=1:5

    switch model

        case 1
            [SN,SN2,ZsimN,ZsimN2]=test_statistic(Data,ESn,Z,Z2,1);
            pvN=sum(SN)/M;
            pvN2=sum(SN2)/M;

            if pvN> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

            if pvN2> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

        case 2

            [ST,ST2,ZsimT,ZsimT2]=test_statistic(Data2,ESt,Z,Z2,2);
            pvT=sum(ST)/M;
            pvT2=sum(ST2)/M;

            if pvT> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end
            if pvT2> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

        case 3

            [SKE,SKE2,ZsimKE,ZsimKE2]=test_statistic(Y,ESke,Z,Z2,3);
            pvKE=sum(SKE)/M;
            pvKE2=sum(SKE2)/M;

            if pvKE> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

            if pvKE2> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

        case 4
            [SGA,SGA2,ZsimGN,ZsimGN2]=test_statistic(GAN,ESga,Z,Z2,4);
            pvGN=sum(SGA)/M;
            pvGN2=sum(SGA2)/M;

           if pvGN> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

            if pvGN2> phi
                 disp('PASS')
            else
```

```
                 disp('FAIL')
            end

        case 5
            [SGAT,SGAT2,ZsimGT,ZsimGT2]=test_statistic(GAT,ESgat,Z,Z2,5);
            pvGT=sum(SGAT)/M;
            pvGT2=sum(SGAT2)/M;

            if pvGT> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

            if pvGT2> phi
                 disp('PASS')
            else
                 disp('FAIL')
            end

    end

end
```

Appendix B

Figures

B.1 DAX

Here we illustrate all the figure related to the DAX index.

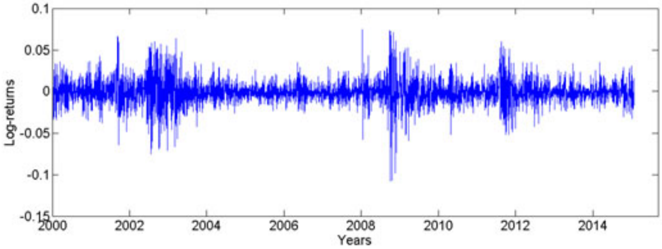


FIGURE B.1: DAX log-returns (Losses are positive numbers).

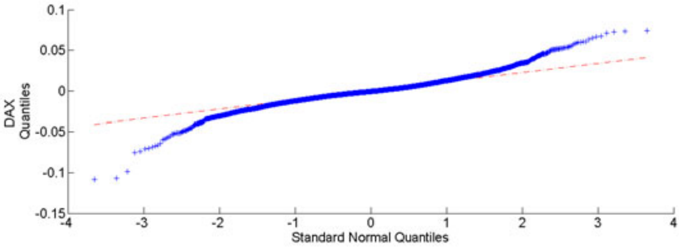


FIGURE B.2: DAX - QQ plot.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9, © Springer Fachmedien Wiesbaden 2016

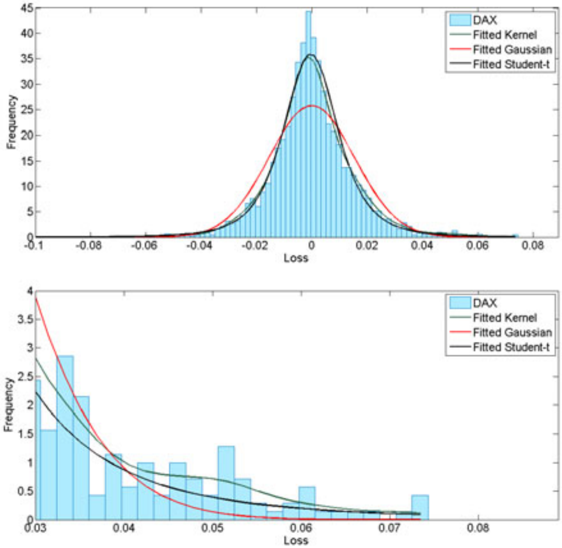


FIGURE B.3: DAX vs Fitted Distributions.

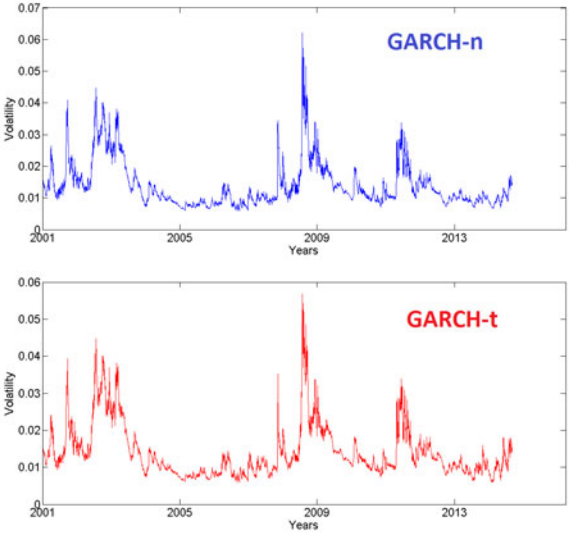


FIGURE B.4: DAX - σ GARCH models

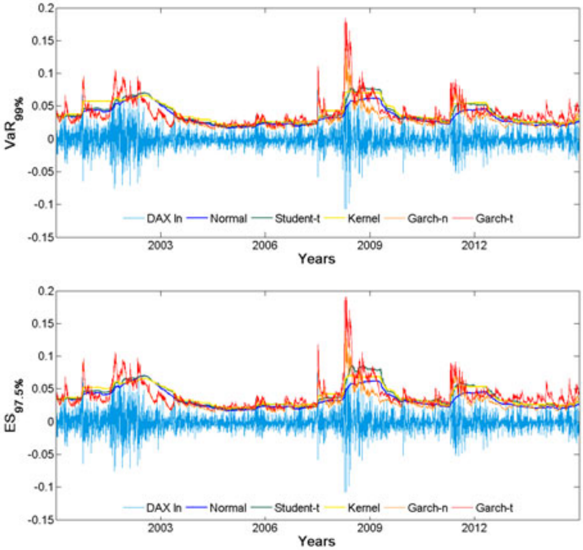


FIGURE B.5: DAX - VaR and ES estimates.

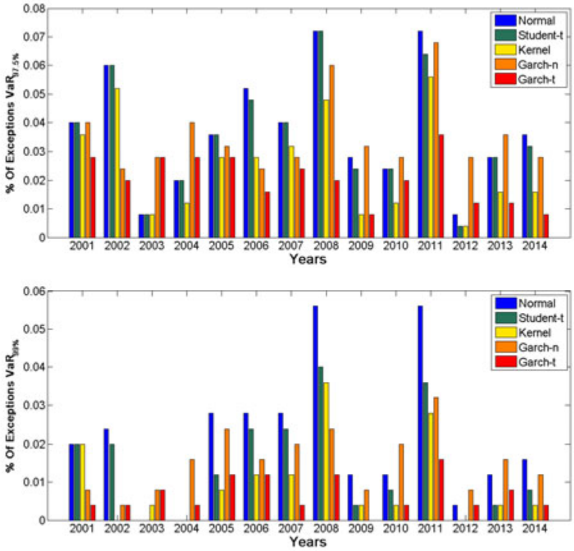


FIGURE B.6: DAX - VaR exceptions

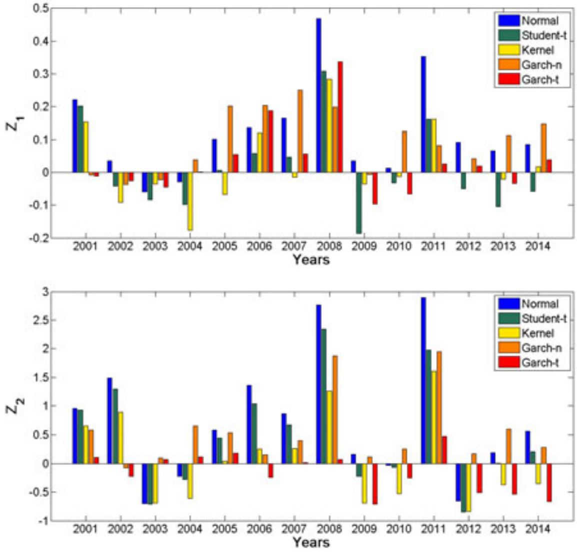


FIGURE B.7: DAX - Z_1 Z_2 per calendar year.

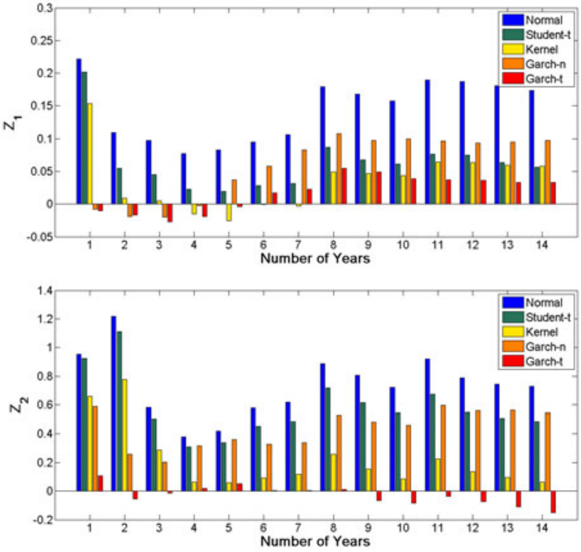


FIGURE B.8: DAX - Z_1 Z_2 cumulative in years.

	Kupiec (POF) Test				
Title : DAX	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{LR_{un}}$	19.76 34.55	15.08 7.44	0.17 0.09	15.82 10.92	3.10 4.88
P-value	0.0000 0.0000	0.0001 0.0064	0.6811 0.9259	0.0001 0.0009	0.0782 0.0272
Test Outcome	X X	X X	√ √	X X	√ X

Table B.1: DAX - Kupiec Test results

	Christoffersen's Interval Forecast Test				
Title : DAX	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{LR_{cc}}$	36.55 64.78	31.35 14.31	28.65 0.77	15.84 10.93	3.33 6.94
P-value	0.0000 0.0000	0.0000 0.0008	0.0000 0.6811	0.0004 0.0042	0.1892 0.0311
Test Outcome	X X	X X	X √	X X	√ X

Table B.2: DAX - Christoffersen's Interval Forecast Test results

	TEST 1				
Title : DAX	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic Z_1	0.1737	0.0562	0.0532	0.0944	0.0244
P-value	0.0000	0.0103	0.0007	0.0000	0.0845
Test Outcome	X	X	X	X	√

TABLE B.3: DAX - Test 1 results

	TEST 2				
Title : DAX	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic Z_2	0.7521	0.5072	0.0983	0.5737	-0.1566
P-value	0.0000	0.0000	0.1720	0.0000	0.9357
Test Outcome	X	X	√	X	√

TABLE B.4: DAX - Test 2 results

B.2 FTSE 100

Here we illustrate all the figure related to the FTSE index.

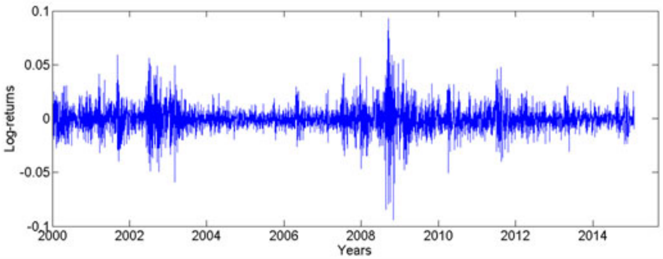


FIGURE B.9: FTSE 100 log-returns (Losses are positive numbers).

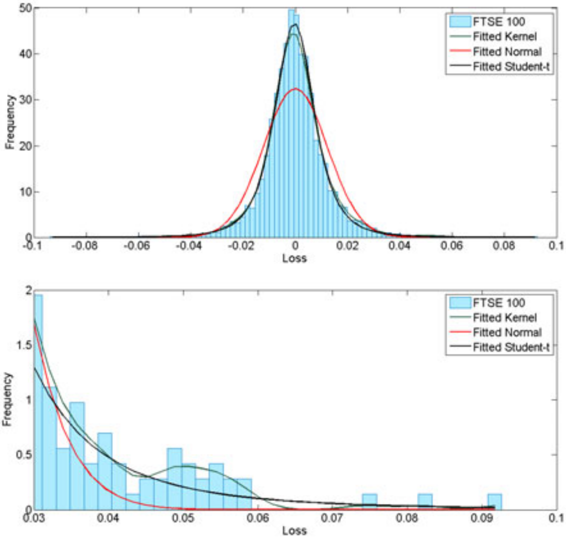


FIGURE B.10: FTSE vs Fitted Distributions.

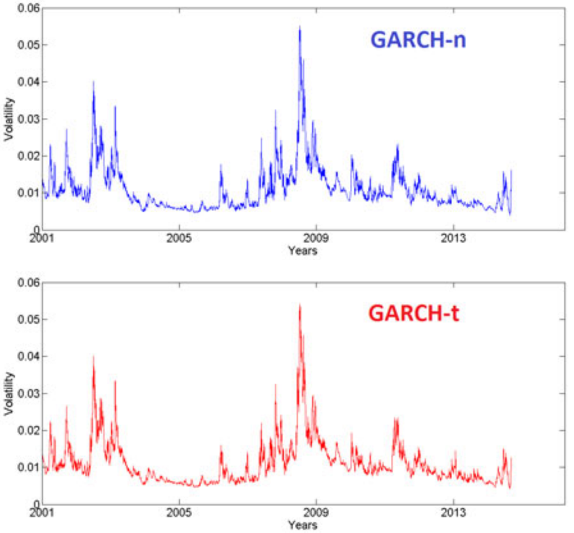


FIGURE B.11: FTSE 100 - σ GARCH models.

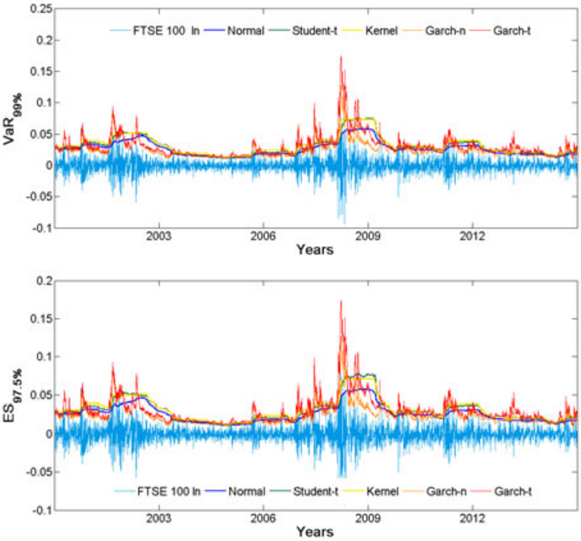


FIGURE B.12: FTSE 100 - VaR and ES estimates.

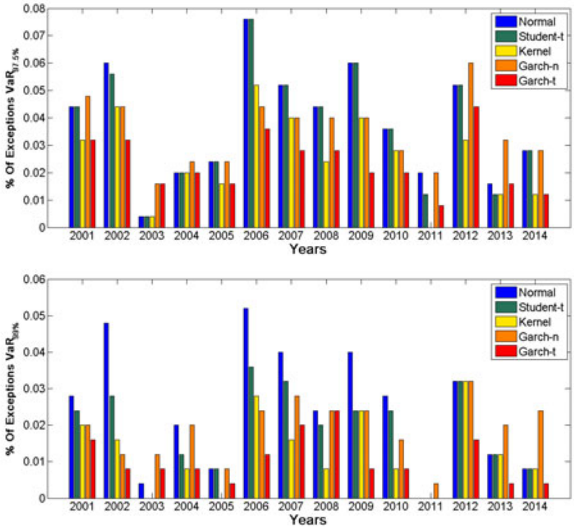


FIGURE B.13: FTSE 100 - VaR exceptions

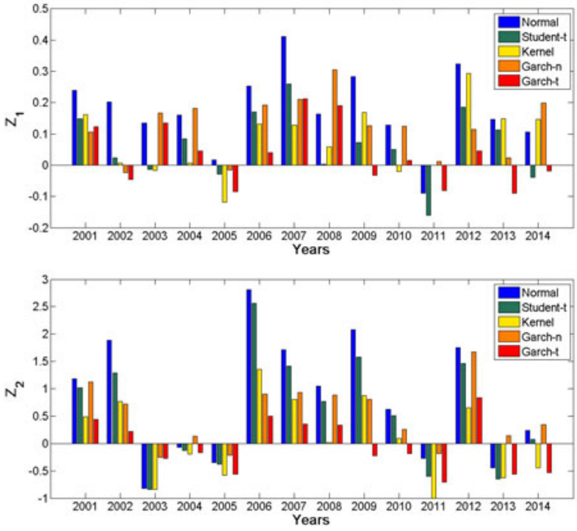


FIGURE B.14: FTSE 100 - Z_1 Z_2 per calendar year.

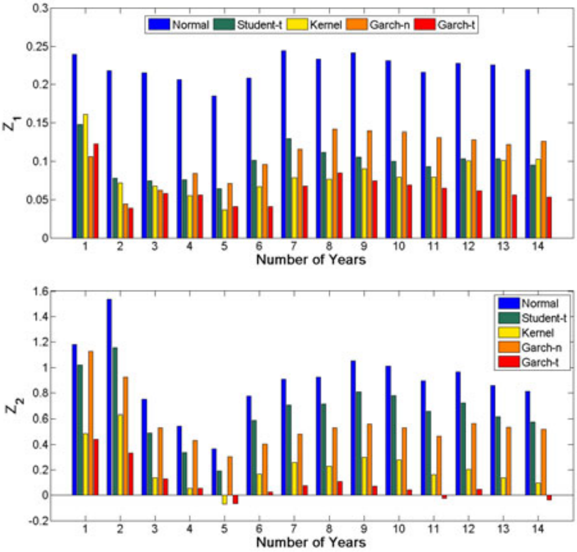


FIGURE B.15: FTSE 100 - Z_1 Z_2 cumulative in years.

	Kupiec (POF) Test				
Title : FTSE 100	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{LR_{un}}$	22.67 61.34	19.31 25.68	0.16 3.97	14.72 25.69	0.42 0.87
P-value	0.0000 0.0000	0.0000 0.0000	0.6901 0.0464	0.0001 0.0000	0.5124 0.6569
Test Outcome	X X	X X	√ X	X X	√ √

Table B.5: FTSE 100 - Kupiec Test results

	Christoffersen's Interval Forecast Test				
Title : FTSE 100	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic $\mathbf{LR_{cc}}$	40.48 76.18	38.63 34.63	11.00 12.05	15.81 25.93	0.85 1.06
P-value	0.0000 0.0000	0.0000 0.0000	0.0041 0.0024	0.0004 0.0000	0.6526 0.5870
Test Outcome	X X	X X	X X	X X	√ √

Table B.6: FTSE 100 - Christoffersen's Interval Forecast Test results

	TEST 1				
Title : FTSE 100	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic Z_1	0.2284	0.1029	0.1066	0.1294	0.0611
P-value	0.0000	0.0004	0.0000	0.0000	0.0004
Test Outcome	X	X	X	X	X

Table B.7: FTSE 100 - Test 1 results

	TEST 2				
Title : FTSE 100	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic Z_2	0.8721	0.6335	0.1520	0.6000	-0.0092
P-value	0.0000	0.0000	0.0739	0.0000	0.5353
Test Outcome	X	X	√	X	√

Table B.8: FTSE 100 - Test 2 results

B.3 NIKKEI

Here we illustrate all the figure related to the NIKKEI index.

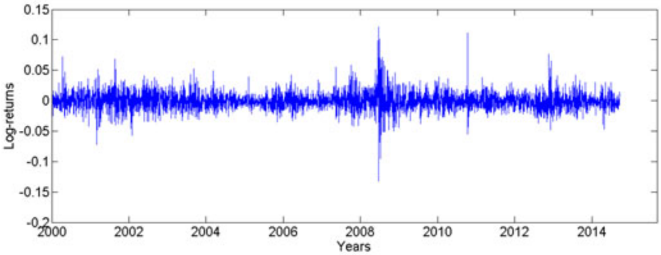


Figure B.16: NIKKEI log-returns (Losses are positive numbers).

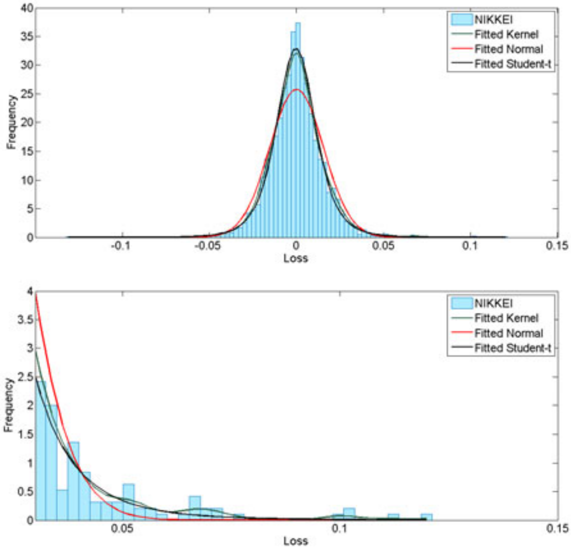


FIGURE B.17: NIKKEI vs Fitted Distributions.

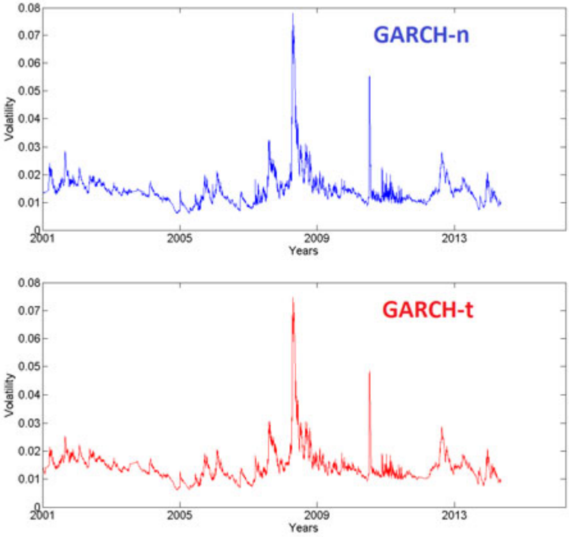


FIGURE B.18: NIKKEI - σ GARCH models.

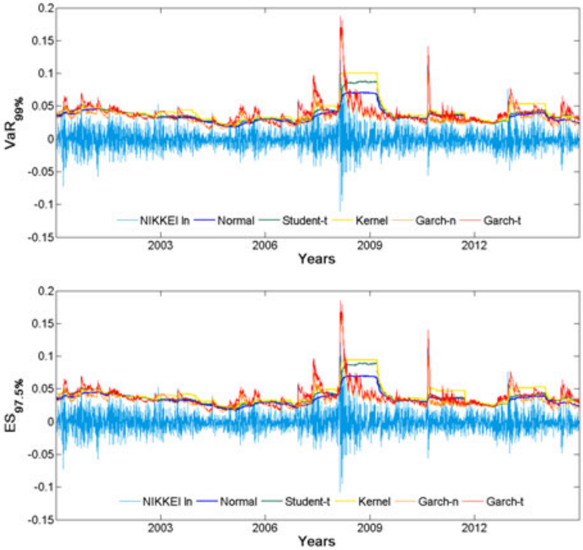


FIGURE B.19: NIKKEI - VaR and ES estimates.

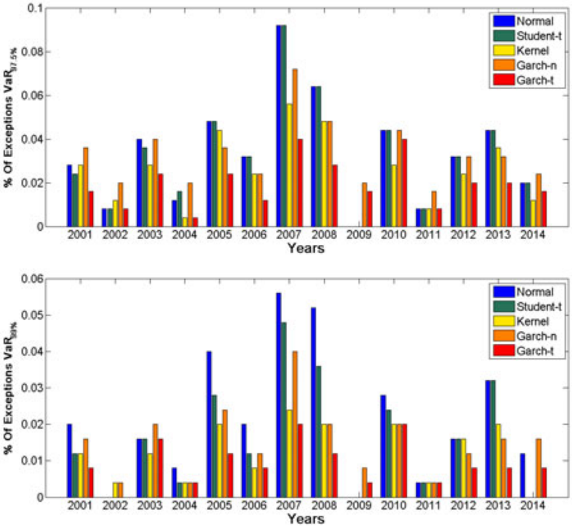


FIGURE B.20: NIKKEI - VaR exceptions

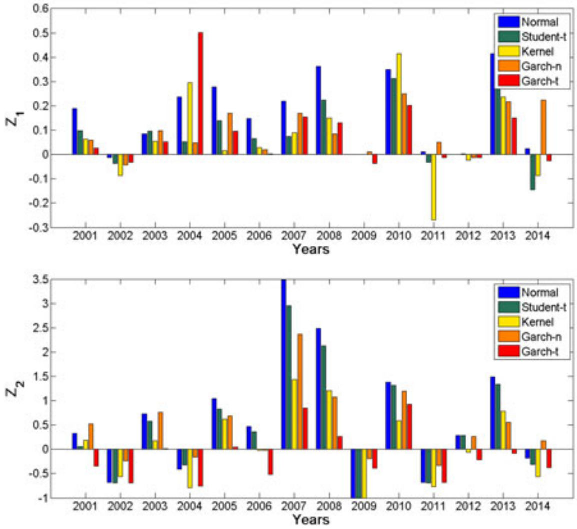


Figure B.21: NIKKEI - Z_1 Z_2 per calendar year.

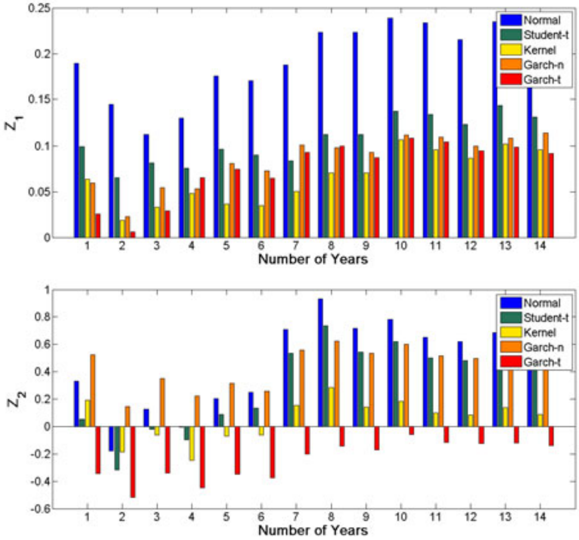


Figure B.22: NIKKEI - Z_1 Z_2 cumulative in years.

	Kupiec (POF) Test				
Title : NIKKEI	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic	8.40	7.84	0.01	8.40	4.48
$\mathbf{LR}_{\mathbf{un}}$	34.44	12.54	0.93	8.77	0.13
P-value	0.0037	0.0051	0.9246	0.0037	0.0342
	0.0000	0.0004	0.3339	0.0031	0.7127
Test	X	X	√	X	X
Outcome	X	X	√	X	√

Table B.9: NIKKEI - Kupiec Test results

	Christoffersen's Interval Forecast Test				
Title : NIKKEI	**Normal**	**Student' t**	**Kernel**	**Garch-n**	**Garch-t**
Test Statistic	12.17	11.76	2.93	9.53	4.76
$\mathbf{LR}_{\mathbf{cc}}$	39.35	18.23	7.22	10.00	1.13
P-value	0.0023	0.0028	0.2306	0.0085	0.0933
	0.0000	0.0001	0.0269	0.0067	0.5690
Test	X	X	√	X	√
Outcome	X	X	X	X	√

Table B.10: NIKKEI - Christoffersen's Interval Forecast Test results

	TEST 1				
Title : NIKKEI	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic Z_1	0.2257	0.1309	0.0958	0.1142	0.0916
P-value	0.0000	0.0000	0.0002	0.0000	0.0004
Test Outcome	X	X	X	X	X

Table B.11: NIKKEI - Test 1 results

	TEST 2				
Title : NIKKEI	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic Z_2	0.6180	0.4800	0.0849	0.4708	-0.1429
P-value	0.0000	0.0000	0.2134	0.0000	0.9111
Test Outcome	X	X	$\surd$	X	$\surd$

Table B.12: NIKKEI - Test 2 results

B.4 EURO STOXX 50

Here we illustrate all the figure related to the EURO STOXX 50 index.

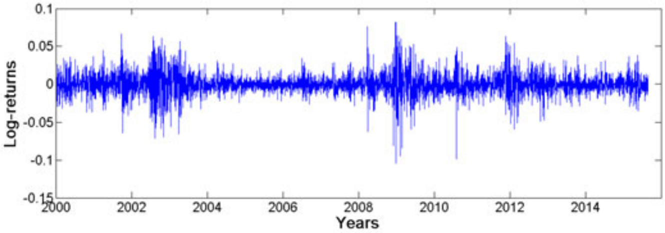


FIGURE B.23: EURO STOXX 50 log-returns (Losses are positive numbers).

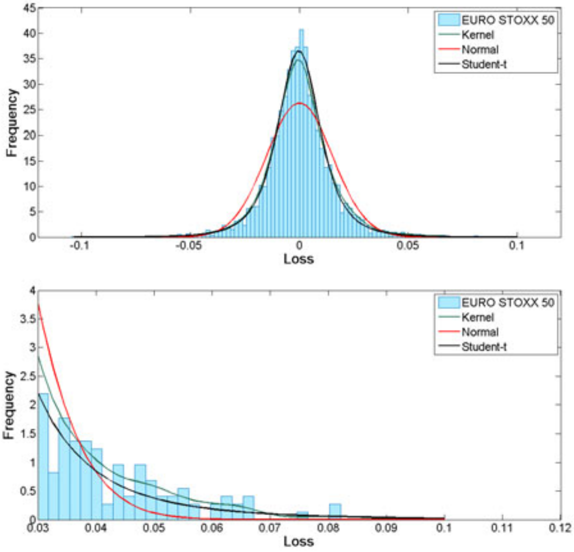


FIGURE B.24: EURO STOXX 50 vs Fitted Distributions.

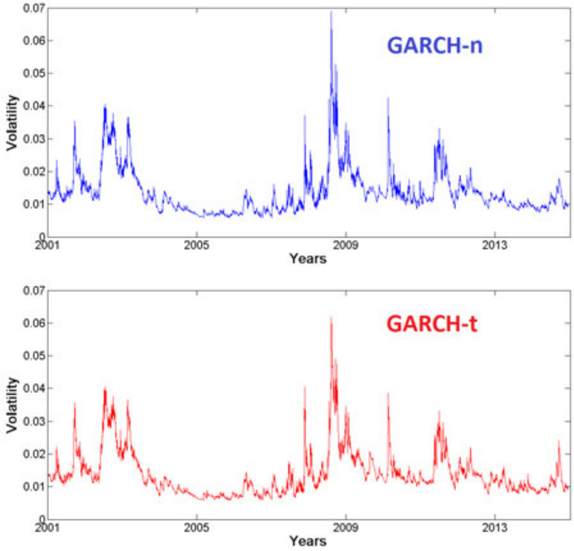


FIGURE B.25: EURO STOXX 50 - σ GARCH models.

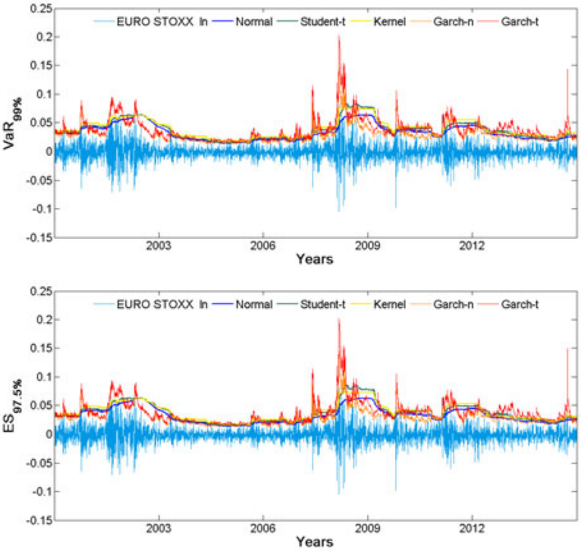


FIGURE B.26: EURO STOXX 50 - VaR and ES estimates.

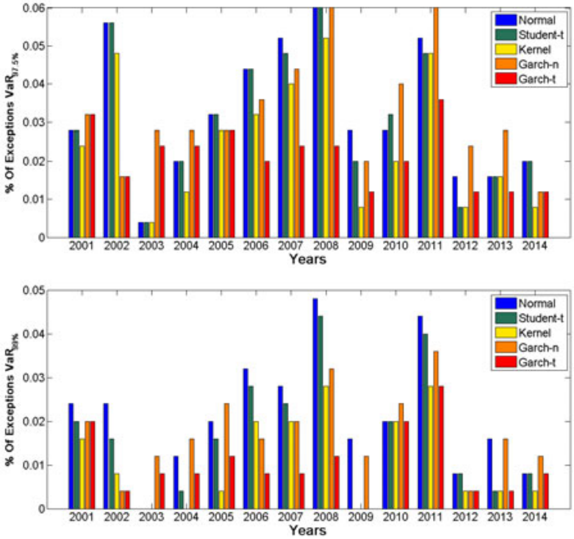


FIGURE B.27: EURO STOXX 50 - VaR exceptions

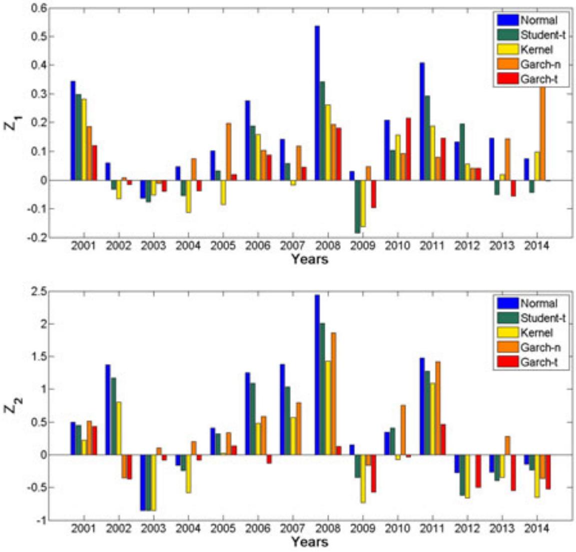


FIGURE B.28: EURO STOXX 50 - Z_1 Z_2 per calendar year.

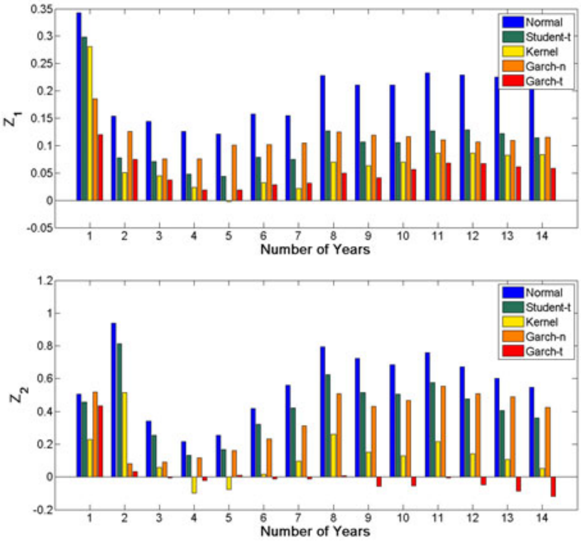


FIGURE B.29: EURO STOXX 50 - Z_1 Z_2 cumulative in years.

	Kupiec (POF) Test				
Title : EURO STOXX 50	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic **LR_{un}**	7.67 38.71	5.67 15.75	0.01 0.75	7.68 19.18	4.16 0.05
P-value	0.0056 0.0000	0.0173 0.0001	0.9684 0.3854	0.0056 0.0000	0.0414 0.8237
Test **Outcome**	X X	X X	√ √	X X	X √

TABLE B.13: EURO STOXX 50 - Kupiec Test results

	Christoffersen's Interval Forecast Test				
Title : EURO STOXX 50	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic LR_{cc}	9.90 48.37	8.37 27.26	6.65 1.73	7.90 12.22	4.24 0.72
P-value	0.0071 0.0000	0.0152 0.0000	0.0360 0.4216	0.0333 0.0000	0.1222 0.6977
Test Outcome	X X	X X	X ✓	X X	✓ ✓

Table B.14: EURO STOXX 50 - Christoffersen's Interval Forecast Test results

	TEST 1				
Title : EURO STOXX 50	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic Z_1	0.2242	0.1185	0.0861	0.1176	0.0558
P-value	0.0000	0.0000	0.0000	0.0000	0.0012
Test Outcome	X	X	X	X	X

Table B.15: EURO STOXX 50 - Test 1 results

	TEST 2				
Title : EURO STOXX 50	Normal	Student' t	Kernel	Garch-n	Garch-t
Test Statistic Z_2	0.5900	0.4039	0.0906	0.4515	-0.1012
P-value	0.0000	0.0001	0.1988	0.0001	0.8331
Test Outcome	X	X	✓	X	✓

Table B.16: EURO STOXX 50 - Test 2 results

Bibliography

[1] C. Acerbi. Spectral measures of risk: a coherent representation of subjective risk aversion. *Journal of Banking Finance*, 26(17), March 2002.

[2] C. Acerbi and B. Szekely. Backtesting expected shortfall. (27), October 2014.

[3] C. Acerbi and D. Tasche. Expected shortfall: a natural coherent alternative to value at risk. (9), May 2001.

[4] T. Andersen, T. Bollerslev, P. Christoffersen, and F. Diebold. Practical volatility and correlation modeling for financial market risk management. *M. Carey and R. Stulz (eds.), Risks of Financial Institutions, University of Chicago Press for NBER*, 2005.

[5] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9, July 1999.

[6] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51, July 2005.

[7] BCBS. Fundamental review of the trading book: A revised market risk framework. *Second Consultative Paper*, 2013.

[8] F. Bellini and V. Bignozzi. Elicitable risk measures. (4), December 2013.

[9] F. Bellini, B. Klar, A. Müller, and E. R. Gianin. Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54, 2014.

S. Roccioletti, *Backtesting Value at Risk and Expected Shortfall,* BestMasters,
DOI 10.1007/978-3-658-11908-9, © Springer Fachmedien Wiesbaden 2016

[10] J. Berkowitz, P. Christoffersen, and D. Pelletier. Evaluating value-at-risk models with desk-level data. *Manuscript North Carolina State University*, 2007.

[11] J. Berkowitz, P. Christoffersen, and D. Pelletier. Evaluating value-at-risk models with desk-level data. 2007.

[12] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 1986.

[13] S. Campbell. A review of backtesting and backtesting procedure. *Finance and Economics Discussion Series, Division of Research & Statistics and Monetary Affairs, Federal Reserve Board*, 2005.

[14] P. Christoffersen. Backtesting. November 2008.

[15] P. Christofferssen and P. Pelletier. Backtesting value at risk: A duration based approach. *Journal of Empirical Finance*, 2, 2004.

[16] R. Cont, R. Deguest, and G. Scandolo. Robustness and sensitivity analysis of risk measurament procedures. *Quantitative Finance*, 10(1), February 2010.

[17] J. Danìelson, B. Jorgenson, G. Samorodnitsky, M. Sarma, and C. De Vries. Fat tails, var and subadditivity. *Journal of Econometrics*, 172, 2013.

[18] S. Degiannakis, C. Floros, and A. Livada. Evaluating value at risk models before and after the financial crisis of 2008: International evidence. *Managerial Finance*, 38, 2012.

[19] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. *Cambridge University Press*, 2002.

[20] P. Embrechts, J. Neslehovà, and M. Wüthrich. Additivity properties for value at risk under archimedean dependence and heavy-tailedness. *Insurance: Mathematics and Economics*, 44, 2009.

[21] P: Embrechts, G. Puccetti, and L. Rüschendorf. Model uncertainty and var aggregation. *Journal of Banking & Finance*, 37, 2013.

[22] S. Emmer, M. Kratz, and D. Tasche. What is the best risk measure in practice? a comparison of standard measures. *ESSEC working paper*, 2013.

[23] R. F. Engle and S. Manganelli. Caviar: Conditional autoregressive value-at-risk by regression quantiles. *Journal of Business and Economic Statistics*, 22, 2004.

[24] H. Föllmer and A. Schied. Convex and coherent risk measures. (8), October 2008.

[25] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, (24), January 2012. URL `http://dx.doi.org/10.1198/jasa.2011.r10138`.

[26] C. W. J. Granger and M. J. Machina. Forecasting and decision theory. 1, 2006.

[27] M. Haas. New methods in backtesting. 2001.

[28] D. Hendricks. Evaluation of value-at-risk models using historical data. *Economic Policy Review, Federal Reserve Bank of New York*, April 1996.

[29] J.Engelberg, C. F. Manski, and J. Williams. Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, 27, 2009.

[30] P. Jorion. Value at risk, the new benchmark for managing financial risk. 2001.

[31] J. Kerkhof and B. Melenberg. Backtesting for risk-based regulatory capital. *Journal of Banking & Finance*, 28, 2004.

[32] R. Koenker. Quantile regression. *Cambridge: Cambridge University Press*, 2005.

[33] M. Kratz. There is a var beyond usual approximations. *ESSEC working paper*, 2013.

[34] P. Kupiec. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3, 1995.

[35] N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions: the highlights. 7, November 2008.

[36] A. J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. June 1999.

[37] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management*. Princeton University Press, 2005.

[38] A. H. Murphy and H. Daan. Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences*, 1985.

[39] W. K. Newey and J. L. Powell. Asymmetric least square estimation and testing. *Econometrica*, 55, July 1987.

[40] K. H. Osband. Providing incentives for better cost forecasting. 1985.

[41] M. Pritsker. The hidden dangers of historical simulation. *Finance and Economics Discussion Series 2001-27*, 2001.

[42] G. De Rossi. *Staying ahead on downside risk. Optimizing Optimization: The Next Generation of Optimization Applications and Theory*. Academic Press, 2009.

[43] M. Saerens. Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks1*, 11, 2000.

[44] L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, December 1971.

[45] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1998.

[46] G. Stahl, J. Zheng, R. Kiesel, and R. Rühlicke. Conceptualizing robustness in risk management. (24), May 2012. URL SSRN:http://ssrn.com/abstract=2065723.

[47] W. Thomson. Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, 20, 1979.

[48] S. Weber. Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16, 2006.