Backtesting Expectiles with Moment Conditions

Jesús Armando de Ita Solis

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science (Mathematics) at

Concordia University

Montréal, Québec, Canada

August 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Jesús Armando de Ita Solis

Entitled: Backtesting Expectiles with Moment Conditions

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality. Signed by the final Examining Committee:

________________________________ Thesis co-supervisor

Dr. Y. Lu

________________________________ Thesis co-supervisor

Dr. M. Mailhot

________________________________ Examiner

Dr. X. Meng

________________________________ Examiner

Dr. K. Kim

Approved by ____

Dr. L. Popovic (Graduate Program Director)

____

Dr. P. Sicotte (Dean of Faculty)

____

Date

Abstract for MSc

Backtesting Expectiles with Moment Conditions

Jesús Armando de Ita Solis

Concordia University, 2023

Under the current regulations, banks and insurance companies have the option to use their own internal models to monitor their risk. To this end, Value-at-Risk (VaR) and the Expected Shortfall (ES) are typically used as the risk measures to compute their capital requirements. Nevertheless, both present flaws, such as the lack of coherence for VaR and lack of elicitability for ES. Recently, expectile has attracted much attention as a potential alternative to VaR and ES. However, the literature on expectile is mainly focused on its statistical inference, and just few traditional backtesting procedures have been proposed. This thesis proposes a traditional backtesting procedure for the expectile and considers its application on financial data.

Acknowledgements

I am truthfully grateful to my supervisors, Professor Mélina Mailhot and Professor Yang Lu, for providing essential guidance through this incredible journey. On top of their valuable support, they also had a vital role in the realization of this work by supplying the necessary tools through their courses.

Moreover, I would like to thank all the members, friends, and colleagues in the department for all their help and company during this process. Everything summed up made my time at Concordia a place to remember.

I would also like to thank my family and partner for always being there for me. This work is dedicated to you.

Finally, I would like to express how fortunate and deeply grateful I feel for the place life has brought me to right now, and for all that is to come.

Contents

List of Tables

Chapter 1

Introduction

Financial institutions are required to compute risk measures for their portfolios, in order to determine their capital reserves. The Basel II Accord[1] and the Solvency II Directive[2] state that financial institutions have the option to use their own internal models for this purpose. The historical standard risk measure is the Value-at-Risk (VaR), but it was later replaced by the Expected Shortfall (ES) for banks and Swiss insurers, among other institutions. However, VaR and ES both suffer from several theoretical downsides. For instance, the VaR does not satisfy the subadditivity property, which is essential in the context of risk management since it provides incentives to diversify risk exposure (Artzner et al., 1999). On the other hand, while ES is subadditive, it has the disadvantage of not being elicitable on its own (Gneiting, 2011). Elicitability allows to rank different candidate internal models based on the accuracy of their risk measure estimates. However, the ES is only jointly elicitable along with the VaR (Fissler et al., 2015). This means that many of the tasks involving ES, such as backtesting, should be conducted jointly with VaR. For example, suppose that we have two models,

[1] Basel Committee on Banking Supervision (2004)

[2] Directive 2009/138/EC of the European Parliament and of the Council (2009)

with A dominating B and where the scores of both models are in terms of the joint score function of the couple (VaR, ES). In this case, we cannot conclude that model A provides more accurate ES forecasts than model B just in terms of the ES. In other words, it is difficult to find the best model for the current ES-based regulatory purpose.

Expectile, on the other hand, is both subadditive and elicitable. First introduced in a regression context by Newey and Powell (1987), it has gained much interest in finance (Bellini and Di Bernardino, 2017; Girard et al., 2021). It is considered easier to estimate than the VaR (Daouia et al., 2018), and its statistical inference has been considered extensively, in either a parametric (Nolde and Ziegel, 2017), semi-parametric (Daouia et al., 2018), or non-parametric framework (Holzmann and Klar, 2016).

Despite its theoretical advantages and the well documented literature on their estimation, to date, expectile has yet to be applied by banks and insurers as a risk measure. One explanation for this lack of success is that currently there are very few (traditional) backtesting methods. The objective of a backtesting procedure is to test whether a given model provides acceptable risk measure estimates (at a given level) by comparing them with the realized sequence of Profit and Losses (P&L). While the backtesting literature for VaR and ES is extensive[3], the literature on backtesting for expectile is still in its infancy. To our knowledge, only Bellini et al. (2019) propose a (traditional) backtesting procedure focused on expectile. Their approach is based on the Probability Integral Transformation (PIT), inspired from the VaR and ES backtesting literature (Costanzino and Curran, 2015; Du and Escanciano, 2017; Löser et al., 2018; Gordy and McNeil, 2020).

Indeed, the definition of VaR, and to a lesser extent the definition of ES, are cumulative distribution function (CDF) based, in the sense that a quantile is involved in their

[3] For backtesting literature on VaR and ES see: Emmer et al. (2013); Acerbi and Szekely (2014); Nolde and Ziegel (2017).

definition. This explains why many of the existing backtesting procedures for VaR and ES involve the Probability Integral Transformation (PIT). Expectile, on the other hand, is defined through the second moment[4] which depends on the associated CDF in an indirect way. This suggests that expectile backtesting procedures based on the PIT might not be the most suitable approach because the expectile of the PIT of the P&L and the expectile of the P&L are not equal. Hence, it is not guaranteed that a model that passes a backtest based on the PIT will pass a backtest based on the P&L sequence.

The purpose of this thesis is to propose a traditional backtesting procedure for expectiles. The proposed backtest is simple to implement, since it only requires institutions to report the daily P&L as well as the associated expectile. In particular, it does not require quantities such as the PIT.

The subsequent chapters of the thesis are organized as follows: Chapter 2 reviews the expectile's definition, compares its properties with those of the VaR and ES, and shows some methods to estimate expectile. Chapter 3 surveys the literature on backtesting VaR, ES, and expectile. Chapter 4 develops the backtesting procedure for expectile, reports results of some Monte Carlo simulations, an empirical application with S&P 500 data, and concludes.

[4] Indeed, expectiles are defined through the optimization of the second order moment of a function.

Chapter 2

Risk Measure Properties: Expectile vs. ES and VaR

Both VaR and ES are popular risk measures that have been used in the past to calculate capital requirements as dictated by the Basel Committee on Banking Supervision (BCBS). Initially, VaR was used as the industry standard, but it was later replaced by the ES in the aftermath of the 2007 GFC. Nonetheless, ES still presents flaws when used for risk management purposes as we will see in this chapter. In the following sections, we review the definition of expectile as an alternative to VaR and ES. Additionally, the sections cover the notion of coherence, elicitability and review some methods to estimate expectiles.

2.1 Definition of the Expectile

The expectile (alternatively L_2-quantile) is introduced by Newey and Powell (1987) for regression purposes, but it has since gained popularity in the quantitative risk management literature (Bellini and Di Bernardino, 2017; Daouia et al., 2018).

The expectile of a random variable Y with finite variance at a given level $\alpha \in [0,1]$ is defined as the minimizer of the following optimization problem:

$$e_\alpha(Y) = \underset{x \in \mathbb{R}}{\arg\min}\, E[S^{(e)}(x,Y)], \tag{2.1}$$

where the scoring function $S^{(e)}(x,y)$ is defined as:

$$S^{(e)}(x,y) = \alpha(y-x)_+^2 + (1-\alpha)(y-x)_-^2. \tag{2.2}$$

This function is asymmetric for all $\alpha \in (0,1)$, except when $\alpha = 0.5$.

Alternatively, the expectile can be characterized by the first order condition[1] (Bellini and Di Bernardino, 2017):

$$\frac{\mathbb{E}\Big[(Y-e(\alpha))^+\Big]}{\mathbb{E}\Big[(Y-e(\alpha))^-\Big]} = \frac{1-\alpha}{\alpha} \tag{2.3}$$

This provides a financial interpretation of the expectile, as the amount of capital to be added to the position in order to have a sufficiently high (expected) gain-loss ratio (Bellini and Di Bernardino, 2017).

This alternative definition is valid as long as Y has a finite mean (without necessarily having a finite variance).

Expectile is a particular case of the M-quantiles proposed by Breckling and Chambers (1988). Additionally, they can also be interpreted[2] as L_p-quantiles, which are a parametric subset of the m-quantiles. L_p-quantiles are introduced by Chen (1996) following Breckling and Chambers (1988). Let a random variable Y, then the L_p-quantile is defined as:

[1] In this case, $e_{0.5}(Y) = E[Y]$

[2] See Philipps (2022) for more interpretations of expectiles.

$$L_{p,\alpha}(Y) = \underset{x\in\mathbb{R}}{\arg\min}\, E[|\alpha - \mathbb{1}\{Y < x\}| \cdot |Y - x|^p], \tag{2.4}$$

where $\alpha \in (0,1)$ and $p \geq 0$.

In particular, the VaR and expectile are obtained when $p = 1$ and $p = 2$, respectively. Both risk measures can be expressed through an optimization problem, unlike ES. While VaR at level $\alpha = 0.5$ represents the median of a distribution, the expectile at that same level accounts for the mean of the distribution.

2.2 Coherence

Artzner et al. (1999) introduced the concept of coherence for risk measures in a financial context.

Definition 2.2.1. *Consider Ω to be the finite set of outcomes of an experiment. Let G be the set of all risks, which is the set of all real-valued functions on Ω. Since Ω is finite, G can be alternatively denoted by $\mathbb{R}^n$, where n = card(Ω). Let $\rho(\cdot)$ be a risk measure which is a mapping from G into $\mathbb{R}$. Then $\rho(\cdot)$ is coherent if it satisfies the following four following axioms:*

i) Translation invariance. We say that $\rho(\cdot)$ is translation invariant if, for all random variables $X \in G$ and all real numbers α, we have $\rho(X + \alpha) = \rho(X) - \alpha$.

In particular, if $\alpha = \rho(X)$ then $\rho(X + \alpha) = 0$

The intuition behind the translation invariance property is that if we add an extra fixed amount of cash to a portfolio (increase in equity of a certain position), then the level of risk of that portfolio will be reduced by the amount added in cash.

Note that some research works, such as in McNeil et al. (2015), give a different interpretation by increasing the risk with positive sign, and some decrease the risk with negative sign. Originally, negative sign is used by Artzner et al. (1999).

ii) Subadditivity. We say that $\rho(\cdot)$ is subadditive if, for all random variables X_1 and $X_2 \in G$, $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$.

This property provides incentives for financial institutions to diversify their portfolios, which will reduce the overall risk of the position.

iii) Positive Homogeneity. We say that $\rho(\cdot)$ is positive homogeneous if, for all $\lambda \geq 0$ and all $X \in G$, $\rho(\lambda X) = \lambda \rho(X)$.

In other words, if we increase the size of the position linearly, then its risk will increase proportionally.

iv) Monotonicity. We say that $\rho(\cdot)$ is monotone if, for all random variables X and $Y \in G$ with $X \leq Y$, we have $\rho(Y) \leq \rho(X)$.

In other words, the risk measure should preserve the ordering between risks.

Artzner et al. (1999) show that VaR is not coherent since it is not subadditive, while the ES is coherent. Bellini et al. (2014) show that expectile is coherent.

2.3 Elicitability

Definition 2.3.1 (see Nolde and Ziegel (2017))**.** *Let $\rho(\cdot)$ be a risk measure and $\Phi = (\rho_1, \rho_2, ..., \rho_n)$ a vector of n risk measures where $n \geq$ 1, and let a scoring function S: $\mathbb{R}^n$ × $\mathbb{R} \rightarrow \mathbb{R}$ be strictly consistent for Φ if:*

$$\mathbb{E}(S(\rho_1(X), \rho_2(X), ..., \rho_n(X)),X) < \mathbb{E}(S(r,X)), \tag{2.5}$$

for all predictions $r = (r_1,...,r_k) \neq \Phi = (\rho_1,...,\rho_n)$ *and all risk variables* X*.*

Then, the vector Φ *of risk measures is elicitable if it has a strictly consistent scoring function.*

In the above definition, if $n = 1$, then we say that ρ_1 is elicitable on its own. Indeed, if a risk measure is defined through an optimization program, it is automatically elicitable. This is the case of the VaR and expectile with scoring functions:

$$S^{(VaR)}(x,y) = \alpha(y-x)_+ + (1-\alpha)(y-x)_-, \tag{2.6}$$

$$S^{(Expectile)}(x,y) = \alpha(y-x)_+^2 + (1-\alpha)(y-x)_-^2, \tag{2.7}$$

respectively.

If we can only find a vector of size $n \geq 2$ containing ρ_1, then ρ_1 is only jointly elicitable with $\rho_2, \rho_3, ..., \rho_n$.

For instance, this is the case of ES, which is only jointly elicitable along with VaR Fissler et al. (2015). In fact, the potential of expectile in risk management is partially explained by their status as the only elicitable and coherent risk measure (Nolde and Ziegel, 2017).

The importance of elicitability depends on the type of backtesting procedure that is used. There exist two types of backtests: traditional backtests and comparative backtests.

Comparative backtests allow us to compare different models in terms of their resulting scores (Nolde and Ziegel, 2017). However, to run those comparative tests, the risk measure selected to perform backtesting must be elicitable, which is the case of the expectile. Comparative backtesting is useful when financial institutions want to select the best candidate model as their internal model. Nevertheless, as pointed out by Bellini et al. (2019), "comparative backtesting does not give any information about the validity of the considered forecasting models: one can compare extremely poor models without noticing" (p. 3).

The main objective of the traditional approach is to verify that the risk measure forecasts are correct at a given confidence level. In other words, it is used to determine whether a given model provides acceptable risk measure forecasts. Traditional backtests need to be implemented if an internal model is used. In case both the standard model and the internal model fail the traditional test, we should run a comparative test to see which model is best.

The focus of this thesis is to propose a traditional backtesting procedure for expectile.

2.4 Computing and estimating expectiles

The statistical properties of the expectile have been studied extensively in the literature. Bellini and Di Bernardino (2017) compare the properties of the expectiles with the ones of the VaR and the ES, and analyze the expectile's asymptotic behavior. Daouia et al. (2018) argue that "inference on expectiles is much easier than inference on quantiles, and their estimation makes more efficient use of the available data since weighted least squares rely on the distance to data points, whereas *empirical* quantiles utilize only the information on whether an observation is below or above the predictor" (p. 264). Besides, Daouia et al. (2018) also mention that it is easier to compute the expectile than the VaR, since the expectile provides a series of smooth curves which makes its loss function continuously differentiable, hence, it can lead to higher predictive accuracy.

There are three methods to compute expectiles:

Non-parametric: The first is the non-parametric method, also known as the filtered Historical Simulation (FHS).

FSH is based on the Historical Simulation method (HS), which makes use of historical returns in order to estimate a risk measure such as VaR or ES. Pérignon and Smith (2010) report that as of 2009, three out of four banks were using HS in order to compute their VaR. According to Christoffersen (2011), HS is widely used in practice since it is easy to implement as we do not have to estimate any parameters using MLE or any other method. However, HS considers the same weight to all assets in the trading period, which might present a problem since old information might have less impact on future returns than what past immediate prices would contribute. For that reason, the Weighted Historical Simulation method (WHS) works as an improved version of HS, because it assigns heavier weights to most recent returns. Finally, while WHS is a good alternative to account for the effect of past immediate information, it might not address outliers effectively. The best approach to control the presence of extreme events is the FHS, since the method applies smoothing techniques to account for the behaviour of financial markets (Barone-Adesi et al., 2002).

The FHS method will be used to compute the expectile in the Monte Carlo section.

In order to obtain the *empirical* expectile we use the iterative minimization of the least asymmetric weighted squares (LAWS) expression (Sobotka and Kneib, 2012):

$$\sum_{i=1}^{N} \omega_i(\kappa)(x_i - e_\kappa)^2, \tag{2.8}$$

where e_κ is the expectile at level κ, x_i is a sample element, and

$$\omega_i(\kappa) = \kappa \mathbb{1}\{x_i > e_\kappa\} + (1-\kappa)\mathbb{1}\{x_i < e_\kappa\}. \tag{2.9}$$

Parametric: The second is the fully parametric estimation method where we assume the distribution of the residuals typically as Normal, t-Student or skewed-t distributed. This method is used to compute the *theoretical* expectile involved in the backtest proposed in Section 4.2 and Section 4.3.

The LAWS formula in equation (2.8) is used to compute the *empirical* expectile. Alternatively, if the CDF is known in closed form, the *theoretical* expectile of a distribution at level κ is computed by solving the following equation [refer to the supplementary material document for Nolde and Ziegel (2017)]:

$$\kappa = \frac{zF(z) - G(z)}{2(zF(z) - G(z)) + m - z}, \tag{2.10}$$

where $z = e(\kappa)$ is the *theoretical* expectile, m is the mean, F the CDF and G the partial moment function of z defined as:

$$G(z) = \int_{-\infty}^{z} uf(u)du. \tag{2.11}$$

For instance, if $X \sim N(\mu,\sigma^2)$, then we obtain:

$$\kappa = \frac{\sigma\varphi(\frac{z-\mu}{\sigma}) + (z-\mu)\Phi(\frac{z-\mu}{\sigma})}{2\sigma\phi(\frac{z-\mu}{\sigma}) + (z-\mu)(2\Phi(\frac{z-\mu}{\sigma}) - 1)}, \tag{2.12}$$

where φ and Φ are the density and distribution functions of the standard normal distribution.

If, instead, X is t-Student distributed with $v > 1$ degrees of freedom, then:

$$\kappa = \frac{\frac{v+z^2}{v-1}t_v(z) + zT_v(z)}{2\frac{v+z^2}{v-1}t_v(z) + z(2T_v(z)-1)}, \tag{2.13}$$

where t_v and T_v are the density and distribution functions, respectively, of the t-Student distribution.

Semi-parametric: Finally, the Extreme Value Theory estimation (EVT) is a semi-parametric estimation method (Daouia et al., 2018). One of the functions of extreme value distributions is to estimate the tails of conditional distributions. For instance, McNeil and Frey (2000) show that conditional EVT outperforms the unconditional EVT procedure, and also improves the results of the GARCH-modelling with normally distributed error terms (GARCH type models are used to estimate volatility). Particularly, for the ES, they verify with some datasets that the generalized pareto distribution (GPD) based method gives better estimates of the ES than other methods.

Girard et al. (2021) propose a theory to estimate extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise. The method suggested is also applicable in the presence of high-dimensional covariates.

In general, all methods naturally have some disadvantages. The drawback of the parametric estimation is misspecification error, whilst the non-parametric estimation methods suffers from a slower convergence rate. Finally, the semi-parametric EVT method only uses "extreme" observations, and as a consequence, they also have a slower convergence rate.

Chapter 3

The Backtesting Literature

Let us now review the literature of backtesting VaR, ES and expectile.

3.1 Backtesting VaR

This section covers two major types of backtesting VaR: tests based on the sequence of violations and duration-based tests.

3.1.1 Tests based on the sequence of violations

The first backtesting on VaR is proposed by Christoffersen (1998), where the concept of violation process is introduced as:

$$I_t = \begin{cases} 1, & \text{if } R_t < -VaR_t(\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

In traditional backtesting, violations or exceptions for VaR and ES occur when the return in a portfolio is below the risk measure. In another context, it could be seen when losses exceed risk measure forecasts.

Christoffersen (1998) proposes interval model free forecasts applied to VaR, since model misspecification is the main reason of computing poor interval forecasts. The tests of interval forecasts, which do not rely on a distribution assumption, are created by using the combination of an indicator variable and a general conditioning set. They count the sequence of intervals that are efficient, and then the proportion is compared against the true coverage p. A sequence of $\text{VaR}_\alpha(t)$ is efficient with respect to $\mathcal{F}_{t-1}$ if:

$$\mathbb{E}_t\Big[\mathbb{1}\{X_t > R_t\}|\mathcal{F}_{t-1}\Big] = 1-\alpha \tag{3.2}$$

almost surely, where $R_t = \text{VaR}_t(\alpha)$.

The methodology proposed by Christoffersen (1998) works when higher-order moment[1] dynamics exist. According to De Clerk and Savel'ev (2022), higher order moments are used to "study the applicability of certain Generalised AutoRegressive Conditional Heteroskedasticity (GARCH) models for mimicking price dynamics" (p. 1). In other words "we can get an insight to the distribution of price change and how it varies over time" (p. 1). In the presence of higher-order dynamics, only testing for Unconditional Coverage (UC)[2] is insufficient since it does not take into account dependence, which is a possible scenario. For this reason, it is stressed the importance of detecting clustering[3] in violations when analyzing a portfolio of returns. Therefore, Christoffersen (1998) proposes two tests: the first verifies the independence assumption, to check if violations

[1] Higher-order moment dynamics is related to the statistical concept of High Order Statistics (HOS), which accounts for functions of third power degree or more. HOS statistics are used to estimate shape parameters such as the skewness (3rd moment) or kurtosis (4th moment) (Kendall et al., 1946).

[2] UC means that the unconditional probability of the VaR should not be significantly higher than the α level, otherwise we will have an overly conservative VaR. The probability should also not be too small because that will increase the risk of loss (Christoffersen, 1998).

[3] According to Mandelbrot (1963), volatility clustering in asset prices occur when "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes."

are distributed independently (no clustering in violations), and the second is a joint test, commonly known as test of conditional coverage (CC), checks for both independence and correct unconditional coverage (UC).

The test proposed in this thesis could be both UC or CC, depending on which conditions we are considering when building the statistical test. In particular, depending on which instrumental function we choose to run the backtest, it can be either UC or CC based. If the moment condition considered are based on lagged values of the *identification function* (refer to section 4.1) then we would be building a Conditional CC test, otherwise it would be an UC test.

3.1.2 Duration-based tests

Another stand of literature tests VaR using durations, which are the time elapsed between successive violations (or hits). According to Christoffersen and Pelletier (2004), Christoffersen (1998) has "relatively small power in realistic small sample settings" (p. 84). Also, while Christoffersen (1998) tests the independence hypothesis through a Markov Chain alternative, which is not the most efficient way to test for independence since it has small power when time dependence in violations is present, Christoffersen and Pelletier (2004) test based on the duration of days between the violations of VaR, can be splitted into the UC test and the independence test but under different assumptions. For instance, we can consider different distributions for the time elapsed between violations/hits, such as the exponential or the Weibull distribution. In other words, we can assume that durations follow a specific distribution.

Not only continuous distributions can be considered for the durations: as an example, Berkowitz et al. (2011) propose the "geometric test" assuming that the durations follow a geometric distribution, and where the null hypothesis states that such durations have the memory-less property. The procedure is based on a discrete duration test with the basis of the Likelihood Ratio test. Finally, they provide evidence of volatility dynamics and non-normality in the data used to test the method. Additionally, they also found that volatility dynamics are not captured by the Historical Simulation method, consequently, clustering in violations might go undetected.

Candelon et al. (2011) extend previous methods by including orthogonal moment conditions (GMM) into duration-based tests. While Bontemps and Meddahi (2012) used the J-statistic, which relies on the moments defined by the orthonormal polynomials based on the geometric distribution, the GMM Duration-based approach uses discrete lifetime distributions compared to the continuous approach used by Christoffersen and Pelletier (2004). The advantages are that the UC hypothesis, as well as the independence assumption and the CC hyphotesis can be analyzed separately (three components). Moreover, the optimal weight matrix of the test is known beforehand, and the approach does not require to make specific assumptions on the distribution of the alternative hypothesis as compared to previous duration-based methods. It is found that under Monte Carlo simulations, the GMM test proposed by Candelon et al. (2011) outperforms other backtests; particularly, methods based on the Likelihood Ratio (LR) test.

This duration based test is further extended by Pelletier and Wei (2016), where a geometric-VaR test is introduced. The test relies on a hazard function, which is a product of the combination of the hazard rate defined in the geometric test and the VaR forecast provided by the VaR test. The geometric-VaR can also be splitted into three components, which are the test for UC, the test for the dependence structure in durations, and finally, the last test "examines whether the probability of getting a violation depends on the VaR forecasts" (p. 727). They conclude that the Geometric approach has more power than previously proposed Duration-based tests.

3.2 Backtesting ES

This section reviews the literature on backtesting ES, which is divided into two types of tests: the first is based on multiple VaR tests, and the second is based on cumulative violations.

3.2.1 Tests based on multiple VaR tests

The literature on ES backtesting is more recent than the one of the VaR. One difficulty of backtesting ES is that it is not elicitable as a risk measure. Therefore, we cannot express it as an optimization problem (Weber, 2006; Gneiting, 2011). Rather, it should be defined through VaR, as the conditional expectation of the loss, if the loss is higher than the VaR. This explains why most backtests of ES involve, to a certain extent, backtests of VaR.

Colletaz et al. (2013) define the notions of exception and super exception, where the latter happens at the extreme tail of the distribution of returns (or P&L distribution). They define a VaR exception as $r_t < VaR_t(\alpha)$, and the VaR super exception as $r_t < VaR_t(\alpha')$, where $\alpha' < \alpha$, and r_t is the return at time t. The null hypothesis of their backtest is a joint hypothesis that checks whether the probability of both an exception and a super exception is α and α', respectively. Colletaz et al. (2013) test the hypothesis through a likelihood ratio test or using a hit regression test.

Another research where the use of VaR is present is introduced by Kratz et al. (2018), where they propose a backtesting procedure for the ES based on tests of VaR violations at various significance levels. Thus, it is an implicit way to backtest the ES. This is motivated by the following approximation first proposed by Emmer et al. (2013):

$$ES_\alpha \approx \frac{1}{4}[q(\alpha) + q(0.75\alpha + 0.25) + q(0.5\alpha + 0.5) + q(0.25\alpha + 0.75)] \tag{3.3}$$

where $q(\alpha) = VaR\alpha$.

The downside of this approach is the approximation error in (3.3) since we are approximating ES using quantiles.

3.2.2 Backtests based on cumulative violations

Du and Escanciano (2017) propose an unconditional and a conditional backtest. In order to build the tests, they introduce the concept of cumulative violations as the integral of the violations over the coverage level in the left tail:

$$H_t(\alpha) = \frac{1}{\alpha} \int_0^\alpha h_t(u)\, du, \tag{3.4}$$

where $h_t(\alpha) = \mathbb{1}(\ X_t \leq -VaR_t(\alpha)\ )$ is the hit at time t and X_t is the bank's revenue at time t. Then, they deduce by Fubini's theorem that the mean of $H_t(\alpha)$ is $\alpha/2$. Hence, their unconditional backtest uses a t-test to evaluate the null hypothesis stated as $E[H_t(\alpha,\theta_0)] = \alpha/2$, where the parameter θ_0 used in the generalized error distribution (GED) $u_t(\theta_0)$, can be estimated through a conditional maximum likelikhood estimator (CMLE), for instance. On the other hand, their conditional backtest is a Portmanteau Box-test which, according to them, is an analogue backtest for ES based on VaR-backtests proposed by (Christoffersen, 1998; Berkowitz et al., 2011).

In conclusion, the backtest by Du and Escanciano (2017) indirectly involves the VaR.

3.3 Backtesting Expectiles

One fundamental difference between backtesting expectile and VaR is that there is no concept of "violation" or "exception" for expectiles since the former is a moment based risk measure rather than a quantile based one. At the same time, even though expectile does not have the concept of violation since they are moment-based risk measures, this would be another point that explains why it has not been adopted by regulators and financial institutions. VaR and ES concept of violation simplifies and explains an undesired scenario which is the exceedance of a loss over the risk measure. Explaining an undesirable scenario through expectile is less intuitive since now the focus is not whether returns (losses) have fallen (surpassed) the quantile, but on the distance between the sample point and the expectile.

Moreover, despite there is some literature on traditional and comparative backtesting procedures for expectiles (Nolde and Ziegel, 2017), to our present knowledge, the only research mainly focused on traditional backtests for expectiles was proposed by Bellini et al. (2019), where the asymptotic distributions of empirical scores (realized scores) and realized identification functions are studied for normal and uniform i.i.d. samples.

First, following Newey and Powell (1987), they state that expectile is the minimizer of the expected value of the quadratic scoring function in (2.7) and identification function:

$$I^{(e)}(x,y) = \alpha(y-x)_{+} - (1-\alpha)(y-x)_{-}. \tag{3.5}$$

In the case of the identification function, the expectile is the only solution to equality $E[I^{(e)}(x,Y)] = 0$, $Y \in L^1$, as showed by Newey and Powell (1987).

Next, they define the realized score and the realized identification functions, which are the empirical versions of the expected values of the quadratic scoring function [equation 2.7)] and the identification function [equation (3.5)], respectively:

Let Y_k be an i.i.d. sample, for $k \in \{1, 2, ..., n\}$, then their definition of the realized score and realized identification functions, which are the empirical versions of the expected scores and identification functions are:

$$\hat{S}_n^{(e)}(x) = \frac{1}{n}\sum_{k=1}^{n} S^{(e)}(x,Y_k), \tag{3.6}$$

where $S^{(e)}(x,Y_k)$ is the scoring function of the expectile [see equation (2.7)], which is evaluated at x: the argument that minimizes the expected score [see equation (2.1)]. Since expectile minimizes the expected score and $\hat{S}_n^{(e)}(x)$ is close to the expected score by law of large numbers, we expect the realized score to be minimized in the expectile as well.

Secondly, the realized identification function is defined as:

$$\hat{I}_n^{(e)}(x) = \frac{1}{n}\sum_{k=1}^{n} I^{(e)}(x,Y_k), \tag{3.7}$$

where $I^{(e)}(x,Y_k)$ is the identification function also evaluated at x, which is the expectile: the value that solves the expression $E[I^{(e)}(x,Y)] = 0$.

After defining the realized functions, they describe two methods: one based on the simulation of realized scores and the second based on the PIT. In both methods, the null hypothesis is stated as: the forecasting model is correct.

The first backtest is based on the simulation of the realized score defined in (3.6). In this test, the comparison of the realized score is made against a simulated score. If $\hat{S}_n^{(e)}(x)$ is way higher than its mean under the H_0, then the null hypothesis is rejected. In case it is not rejected, they conclude that the forecasting model is adequate. The downside of this approach is that the critical value of the backtest does not have a closed form and must be computed by simulation.

The second backtest procedure involves the PIT, which transforms a continuously random variable to a standard uniform random variable by applying the (conditional) CDF (i.e. the daily return). First, they compute the PIT of the estimated models, and then the realized score $\hat{S}_n^{(e)}(x)$ and the realized identification function $\hat{I}_n^{(e)}(x)$ as if the model were uniform i.i.d.

Further, when using the realized score $\hat{S}_n^{(e)}(x)$ in their backtest, as stated in the first method, the null hypothesis should be rejected if it is sufficiently far from the asymptotical mean of $\hat{S}_n^{(e)}(x)$ under H_0. On the other hand, when using the realized identification function $\hat{I}_n^{(e)}(x)$, the null hypothesis should be rejected if $\hat{I}_n^{(e)}(x)$ is sufficiently far from zero. In other words, they analyze if there is a change (increase or decrease) in the realized identification function, since it should stay sufficiently close to zero because the expectile is the solution to equality $E[I^{(e)}(x,Y)] = 0$.

Finally, through some simulated examples, they also conclude their backtests that use scoring functions such as (2.7), have more power detecting conditional mean misspecifications than their backtests that use identification functions [see equation (3.5)].

The downsides of the PIT approach are: First, if the regulator needs to compute the PIT, then the financial institution would have to disclose its internal model, which can raise confidentiality concerns. Second, and more importantly, the procedure is focused on the expectile of the PIT; nevertheless, the expectile of the PIT of the P&L is not the same as the expectile of the P&L. In other words, there is no simple one-to-one correspondence between the expectiles of the PIT and the expectiles of the P&L. Thus, a model that passes a PIT-based backtest does not necessarily pass a P&L-based backtest. Namely, it is not possible to check whether the capital reserve for that financial institution is adequate by using their procedure[4].

[4] A similar issue exists also for many ES backtesting procedures, such as Du and Escanciano (2017), which also backtest the ES of the PIT instead of backtesting the ES of P&L directly.

Chapter 4

Methodology

In this chapter, we propose a traditional backtest for expectile, run some Monte Carlo simulations to compute its power and size, and show an empirical application with real data.

4.1 The test

In order to test the performance of a model in terms of the accuracy of its expectile estimate[1], we run a backtesting procedure that consists of comparing the daily observed return denoted as X_t for each date t, with the conditional expectile $e_t(\kappa)$ computed at time t.

For such a test, following Nolde and Ziegel (2017) we first state the next definitions:

Let $\Psi = (\rho_1, ..., \rho_n)$. The sequence of forecasts $\{e_t(\kappa)\}_{t \in \mathbb{N}}$ is calibrated for Ψ on average if

[1] We test the expectile in terms of the expectile estimate. We do not test whether expectile is an adequate risk measure since, as shown in Chapter 2, it is acceptable as it is elicitable and coherent.

$$\mathbb{E}[\Psi(e_t,X_t)] = 0, \forall\, t \in \mathbb{N}. \tag{4.1}$$

Furthermore, if we condition on past information, then we say that the sequence of predictions of forecasts $\{e_t(\kappa)\}_{t\in\mathbb{N}}$ is conditionally calibrated for Ψ if

$$\mathbb{E}[\Psi(e_t,X_t)|\mathbb{F}_{t-1}] = 0, \text{almost surely}, \forall\, t \in \mathbb{N}, \tag{4.2}$$

where $\Psi(e_t,X_t)$ is the identification function defined as:

$$\Psi(e_t,X_t) = \kappa(X_t - e_t(\kappa))^+ - (1-\kappa)(X_t - e_t(\kappa))^-. \tag{4.3}$$

Finally, we can state the null hypothesis of the backtest as:

$\mathcal{H}_0$: The sequence of forecasts $\{e_t(\kappa)\}_{t\in\mathbb{N}}$ is conditionally calibrated for Ψ.

Just for explanatory purposes, the null hypothesis could be interpreted as: the bank's internal model produces sufficiently accurate expectile forecasts.

If we do not reject the null hyphotesis, we conclude that the model passes the test. If $\mathcal{H}_0$ is rejected, then we conclude that the model fails the test (i.e. is not acceptable).

Starting off with the context, suppose that a bank has an internal model, which is used to compute the daily predictive distribution of next day's return, as well as the associated conditional expectile $e_t(\kappa)$ at level $\kappa \in (0,1)$ and time t characterized by:

$$\kappa\mathbb{E}_t[(X_t - e_t(\kappa))^+] = (1-\kappa)\mathbb{E}_t[(X_t - e_t(\kappa))^-] \tag{4.4}$$

where the symbol $\mathbb{E}_t$ means conditional expectation with respect to all the available information up to the start of the trading date t.

We compute the *conditional* expectile at time t as

$$e_t(\kappa) = \sigma_t \cdot e(\kappa), \tag{4.5}$$

where σ_t should be estimated, and $e(\kappa)$ is the *theoretical* expectile of the residuals ϵ_t, which is computed using one of the methods explained in Section 2.4. This method of estimating the *conditional* expectile is similar to the one used by McNeil and Frey (2000).

The objective is to test whether equation (4.4) is satisfied by the data, that is to say, if the internal calculation of $e_t(\kappa)$ over the period $t \in \{1, ..., n\}$ is acceptable, at some confidence level 1 - α, for instance, 95%.

The thesis is based on the use of moment conditions suggested by Nolde and Ziegel (2017). Indeed, (4.4) implies that

$$\mathbb{E}\Big[F_t[\kappa(X_t - e_t(\kappa))^+ - (1-\kappa)(X_t - e_t(\kappa))^-]\Big] = 0, \tag{4.6}$$

where F_t is called instrumental function and can be any deterministic function of previous market variables such as X_{t-1}, X_{t-2}, ... and $e_t(\kappa)$, $e_{t-1}(\kappa)$, ...

Thus, we have at our disposal an infinity of moment conditions that the sequences of expectiles $e_t(\kappa)$ should satisfy, which allow us to freely choose the instrumental function. See some natural candidates highlighted below:

Example 1: $F_t = \frac{1}{e_t(\kappa)}$. This is motivated by the fact that the term $\alpha(X_t - e_t(\kappa))^+ - (1-\alpha)(X_t - e_t(\kappa))^-$ likely features heteroscedasticity when t varies. Hence, dividing it by the expectile, which acts as a volatility measure at time t, could help removing such heteroscedasticity.

Example 2: $F_t = 1$, to test whether $\alpha(X_t - e_t(\kappa))^+ - (1-\alpha)(X_t - e_t(\kappa))^-$ is correctly centered. In the backtesting literature, this test is often called as simple conditional calibration test (Nolde and Ziegel, 2017).

Example 3: $F_t = \frac{X_{t-1}}{e_t(\kappa)^3}$, in which the numerator is chosen to test whether the impact of the leverage is correctly taken into account by the bank's internal model. The denominator, in turn, is chosen to control for heteroscedasticity.

Example 4: We can also use some exponential moving average filter to construct a volatility measure from the observed daily return only. For instance, we could use, alternatively:

$$F_t = \frac{1}{\sigma_t^2},$$

where $\sigma_t^2 = \alpha\sigma_{t-1}^2 + (1-\alpha)X_t^2$, and α is some pre-determined constant, say $\alpha = 0.95$.

Example 5: We can construct some empirical skewness measures, such as $F_t = \frac{X_{t-1}^3}{e_t(\kappa)^5}$.

All the above-mentioned examples share one common property: they only depend on past observations of (X_t) and $e_t(\kappa)$. Consequently, the bank does not need to disclose completely its own internal model. It is only required to report its daily risk measure $e_t(\kappa)$.

The condition that the *empirical* mean of the identification function is equal to zero is not sufficient when dealing with time series to characterize a Conditional Coverage test, since that would only characterize an Unconditional Converge Test. Therefore, when working with time series, extra conditions should be considered in order to deal with the dependence of past returns because the identification function would no longer be i.i.d. For this reason, we condition on past information and use the conditionally calibrated sequence of forecasts.

Moreover, depending on how we build the instrumental functions, the test can be UC-based or CC-based. If we consider instrumental functions based on lagged identification functions such as $F_t = (1, \Psi(e_{t-1}, X_{t-1}))^T$, as suggested by Giacomini and White (2006), then we would be building a Conditional Coverage test. In general, we can consider $F_t = (1, \Psi(e_{t-1}, X_{t-1}), ..., \Psi(e_{t-k}, X_{t-k}))^T$ where $k \geq 1$ (see, Engle and Manganelli, 2004; Kuester et al., 2006).

The moment conditions in equation (4.6) can now be explicitly tested using observations of return data X_t and the sequence of expectiles $e_t(\kappa)$ provided by the bank. Let's assume that we use m moment conditions, where $m \geq 1$. We denote the vector of instrumental functions $F_t = (F_{1,t}, F_{2,t}, ..., F_{m,t})$. Then, under appropriate stationarity assumptions[2] detailed in Theorem 4 named *Unconditional Predictive Ability Test* by Giacomini and White (2006), we get approximately:

$$\frac{1}{n}\sum_{t=1}^{n} F_t \Psi(e_t, X_t) \sim \mathcal{N}(0, \Omega), \tag{4.7}$$

where Ω is the covariance matrix of the vector $F_t\Psi(e_t, X_t)$, and where $\Psi(e_t, X_t)$ is the identification function.

Thus, we can conduct a Wald-test (Nolde and Ziegel, 2017) which is defined as:

$$\hat{W} = n \; (\frac{1}{n}\Sigma_{t=1}^{n} F_t \; \Psi(e_t, X_t) \;)^T \; \hat{\Omega}_n^{-1} \; (\frac{1}{n}\Sigma_{t=1}^{n} F_t \; \Psi(e_t, X_t) \;), \tag{4.8}$$

[2] One minor difficulty is that because we have time series data, the distribution of $\frac{1}{n}\sum_{t=1}^{n} F_t\Psi(e_t, X_t)$ is asymptotically Gaussian only under some conditions [see e.g. McLeish (1975); Hannan (1979); Wooldridge and White (1988)]. The conditions about the Gaussian asymptotics are frequently assumed in the economics/finance literature since Hansen (1982).

where the empirical variance-covariance matrix can be obtained using general result on weakly stationary time series (Giacomini and White, 2006, Theorem 1):

$$\hat{\Omega}_n = \frac{1}{n}\Sigma_{t=1}^{n}(F_t\ \Psi(e_t{,}X_t))\ (F_t\ \Psi(e_t{,}X_t))^T. \tag{4.9}$$

In particular, when $m = 1$ (we only use one instrumental function), the test statistic $\hat{W}$ is simplified as in the following inequality:

$$\hat{W} = \frac{n(\frac{1}{n}\Sigma_{t=1}^{n}F_t\ \Psi(e_t{,}X_t)\ )^2}{\Sigma} > z, \tag{4.10}$$

where Σ is simply the variance of $F_t\Psi(e_t{,}X_t)$, and where z is the critical value of the chi-squared distribution with $v = 1$ degrees of freedom. In fact, $v = m$ for all m.

If the inequality (4.10) is true, then we reject the null hypothesis, and reject the bank's internal model. Conversely, if we do not reject the null hypothesis, then there is not enough evidence to conclude that the expectile forecasts estimated by the internal model are poor.

4.2 Monte Carlo Simulation of size and power of the test

In this section, we compute the size and the power of the test, and analyze how those metrics behave when the sample size n varies, when we set a different level for the expectile κ, or when we change the distributional assumption of the innovations ϵ_t.

To initialize the simulation study, we first simulate daily returns X_t using a GARCH(1,1) time series model which we will refer to as the correct model, or as the true data generator. We start by specifying the parameters of the model as:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \beta_1 X_{t-1}^2 \tag{4.11}$$

$$X_t = \sigma_t \epsilon_t, \ t \in \{1, 2, ..., n\}, \tag{4.12}$$

where $\alpha_0 = 1 \times 10^{-6}$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and e_t are the innovations standard Normal, or t-Student[3] distributed with v degrees of freedom. Moreover, in order to make the GARCH(1,1) simulation stable, we verify the stationarity conditions: $\alpha_1 > 0$, $\beta_1 > 0$, and $\alpha_1 + \beta_1 < 1$. In the first set of tables, we will run a GARCH model with Normal innovations.

We set the variance at time $t = 1$ to be the unconditional variance of the model:

$$\sigma_1^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}. \tag{4.13}$$

Once we have the volatility vector σ_t, we use equation (4.5) to compute the vector of daily conditional expectiles $e_t(\kappa)$ associated with the returns X_t.

[3] The variance of the residual in the t-Student case needs to be standardized, in order to make it unitary.

Before running the Wald-test, we define six instrumental functions that we will use in the computation of the Wald-statistic:

Instrumental Function 1: $F_t = \frac{1}{e_t(\alpha)}$

Instrumental Function 2: $F_t = 1$

Instrumental Function 3: $F_t = \frac{X_{t-1}}{e_t(\alpha)^3}$

Instrumental Function 4: $F_t = \frac{1}{\sigma_t^2}$

Instrumental Function 5: $F_t = \frac{X_{t-1}^3}{e_t(\alpha)^5}$

Instrumental Function 6: $F_t = \frac{1}{e_{t-1}(\alpha)}$

Next, using X_t, $e_t(\kappa)$ and F_t, we run a Wald-test as previously described in Section 4.1. This procedure is replicated NSim times (number of simulations), in order to calculate the proportion of times the test is rejected. These proportions are two major metrics commonly known as the size and power of the test, and are used to measure the performance of a statistical test. Depending on which metric we are calculating, we will make a different assumption on the model to estimate σ_t, which is needed to calculate the conditional expectile as in equation (4.5).

4.2.1 Computing the Size of the test

The size of the test, alternatively called Error Type I, is defined as:

- Error Type I = P(Reject $\mathcal{H}_0$ | $\mathcal{H}_0$ is True)

Error Type I (Size) is interpreted as the probability of falsely rejecting the correct model.

In order to calculate Error Type I, we compute σ_t under the hypothesis that the returns X_t come from the previously stated GARCH(1,1), which is the true data generator. Then, we use equation (4.5) to compute the conditional expectile.

When we compute the *theoretical* expectile $e(\kappa)$ of the residuals ϵ_t, used to estimate the *conditional* expectile, we utilize the R functions "enorm" and "et" within the "expectreg" library, depending on the assumption of the residuals being Normal or t-Student distributed, and then we multiply $e(\kappa)$ by the daily volatility of the model σ_t as in equation (4.5).

Then we use equation (4.8) to compute the Wald-statistic $\hat{W}$. Finally, we verify inequality (4.10), where z is a chi-square quantile $\chi^2_{\alpha,v}$ with arbitrary $\alpha = 95\%$ and v degrees of freedom, which depend on the number of instrumental functions that we are using. If inequality (4.10) is satisfied, then we reject the null hypothesis $\mathcal{H}_0$ and conclude that the model fails the test. This whole procedure counts as the first iteration of a simulation. We run this procedure $NSim$ times (number of simulations) in total, and calculate the proportion of successful tests.

Since $\alpha = 95\%$ is the arbitrary level chosen for the critical value, we expect Error Type I results to be close to 1 - α if n is sufficiently large, given that we expect the null hypothesis to be rejected $(1 - \alpha)\%$ of the times.

4.2.2 Computing the Power of the test

The second metric is called the power of the test, and it can be expressed as 1 - Error Type II:

- Error Type II = P(Accept $\mathcal{H}_0$ | $\mathcal{H}_0$ is False) = 1 - P(Reject $\mathcal{H}_0$ | $\mathcal{H}_0$ is False)

Error Type II (1 - Power) is the probability of incorrectly accepting the wrong model. Alternatively, the power represents the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true. A higher power value indicates a better performance of the test.

Let us now explain how to compute Error Type II, the complement of the power. As in the computation of Error Type I, we also need to compute the conditional expectile using (4.5). In this case, instead of assuming that the daily volatility σ_t comes from the previously stated GARCH(1,1) which is the true data generator, we calculate σ_t "incorrectly" by a wrong model. For instance, we can compute σ_t using the HS method [see Section 2.4]. Under this method, σ_t is simply estimated as the sample variance:

$$\hat{\sigma^2} = \Sigma_{t=1}^{n} \frac{(X_t - E[X_t])^2}{n-1}, \ \forall \, t \tag{4.14}$$

Another way to incorrectly compute the expectile is by assuming that the returns follow a different model than the GARCH(1,1) used to generate the true data. For example, we can assume a GARCH(1,2) instead.

A priori, we expect that Error Type II will decrease as n increases if the periods of observations overlap. Two periods of time overlap if the observations of the first period are contained inside another one larger in number of observations. For instance, if days 1 to 250 of the first period are chronologically contained in the second period of time that has 500 observations, then we can say that the former overlaps with the latter. Hence, the larger is the length of the period, the more information the backtest has available to run the procedure. Consequently, the test will detect the wrong model more efficiently.

It is important to highlight that in the following tables, we show results where the *size-adjusted-power* has not been applied. We show raw results of the size and the power of the test. On a second note, we have excluded instrumental functions 3 and 5 in the results of the following sections, since we ran some preliminary simulations using the HS method, and both Error Type I and II results were not conclusive. Error Type I did not always approached to $1 - \alpha$, and Error Type II was quite large for both instrumental functions, and did not converged to zero as we increased the sample size n.

4.2.3 Impact of the sample size on Error Type I and Error Type II

Table 4.1 shows what happens to Error Type I and II when we use an extremely large number of observations n. We expect the Error Type I to converge to 1 - α, and Error Type II to plunge to zero. Note that Error Type II is not exactly zero for $n =$ 20,000. In this specific case, running these procedures takes quite a long time because n is large, and the number of simulations also affect the computational time when we run the process for more than 2,000 simulations.

It was previously mentioned that instrumental functions 3 and 5 alone show bad performance since Error Type I and II are very large. For instrumental function 6, we did not see any improvement compared to instrumental function 1 which is expected since they are of the same nature. As for Error Type I, we see that it is close to 1 - α across all instrumental functions (except 3 and 5). For this reason, we discard instrumental functions 3, 5, and 6, and keep 1, 2 and 4 for the rest of the simulations.

Another finding to highlight is to observe the results of Error Type I and Error Type II when we add more instrumental functions. One must question whether Error Type II decreases by adding more instrumental functions. From the final results we observe that in some cases Error Type I slightly increases to 8.5% for $\kappa = 90\%$. Here we are computing all results with 5,000 simulations. In some cases, the size will get closer to 1 - α if the number of observations n and number of simulations (NSim) are large enough, and if κ is not close to 100%. The choice of instrumental functions will affect the results as well.

(a) Error Type I

IF/n	250	750	5,000	20,000
1	0.0608	0.054	0.0524	0.0556
2	0.0626	0.0522	0.0494	0.051
4	0.0572	0.0562	0.0536	0.0534
1-2	0.0918	0.0712	0.0514	0.0526
1-4	0.088	0.0596	0.0486	0.0542
2-4	0.0888	0.0644	0.052	0.0546
1-2-4	0.1294	0.0892	0.0586	0.0538

(b) Error Type II

IF/n	250	750	5,000	20,000
1	0.8764	0.8066	0.2792	0.0016
2	0.863	0.8042	0.3522	0.01
4	0.8844	0.817	0.304	0.0016
1-2	0.8394	0.8126	0.373	0.0036
1-4	0.848	0.8218	0.369	0.0034
2-4	0.8422	0.812	0.367	0.0034
1-2-4	0.7992	0.8002	0.4192	0.0058

Table 4.1: Impact of the sample size on Error Type I and II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e - 6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. The returns of the wrong model come from a GARCH(1,2) model with $\alpha_0 = 1e - 6$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and $\beta_2 = 0.01$. The number of simulations is 5,000, $\kappa = 90\%$, and $\alpha = 95\%$.

To conclude, we observe that, in terms of convergence to (1 - α), the best instrumental function is number 2, followed by 4 and finally 1 for Error Type I when n is very large. For Error Type II, we see that instrumental function 1 showed better results when n = 5,000. The worst case for n = 20,000 is when using instrumental function 2, since Error Type II is the highest in that column.

When using two or more instrumental function in the construction of the Wald-statistic, we see a decrease in Error Type II when n = 250. However, when analyzing Error Type I, it is slightly higher for n = 250, being the combination of instrumental functions 1-2-4 the worst of all scenarios. Taking into account the aforementioned, we observe that for n = 250, and when adding more instrumental functions, there is a trade-off between both errors since Error Type II decreases as Error Type I increases.

As we analyze the extreme case where n = 20,000, we conclude that there is not significant difference in Error Type I since all combinations of instrumental functions roughly converge to 1 - α = 5%.

4.2.4 Impact of the choice of wrong model on Error Type II

In this section, we compute the conditional expectile assuming that volatility σ_t used to compute the conditional expectile $e_t(\kappa)$ comes from a different model, which we will call the "wrong model". This wrong model is different than the original GARCH(1,1) used to simulate the true data. This way, we will be able to compute Error Type II.

There are various choices of wrong models that we can use to compute the volatility used to compute the conditional expectile. One option is to compute the volatility using the Historical Simulation method, where in this case, we simply use the sample variance as in equation 4.14, instead of computing the volatility using the recursive formula of the GARCH(1,1) model. A second option is to compute the volatility using the recursive formula of a GARCH(1,2) model which is different from the true model:

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2, \tag{4.15}$$

$$X_t = \sigma_t \epsilon_t,\ t \in \{1, 2, ..., n\}, \tag{4.16}$$

with parameters:

$$\alpha_0 = 1 \times 10^{-6},\ \alpha_1 = 0.1,\ \beta_1 = 0.888,\ \beta_2 = 0.01.$$

We must check that the stationarity condition $0 < \alpha_1 + \Sigma\beta_i < 1$ is satisfied for this GARCH(1,2) model.

Now, since we already have the daily returns X_t previously simulated by the GARCH(1,1) model (the true data generator), we insert them into equation (4.15) to obtain the volatility of the GARCH(1,2) model (the wrong model). Note that the GARCH(1,2) model is almost the same as the GARCH(1,1) computed previously, because they have the same parameter values α_0, α_1, and β_1. The only difference is that the GARCH(1,2) has an extra parameter β_2, which is quite small in this specific case. Since β_2 is small, we do not expect to have a significant difference in the daily volatility values of the GARCH(1,2) and the volatility values of the GARCH(1,1).

As stated before, the GARCH(1,2) model used to compute the "wrong" expectile is very similar to the GARCH(1,1). On the contrary, if we choose β_2 to be larger value, for example 0.15, the results for Error Type II will change drastically since the daily volatilities of both models will be far from each other. We would expect Error Type II to be much lower as the wrong model will be easier to detect. In the following tables, we present the results from different backtesting procedures using a GARCH(1,1) model as the true data generator, against a GARCH(1,2) as the wrong model whose volatility is computed with the recursive formula stated before.

In summary, we will set the GARCH(1,2) as the wrong model. As mentioned before, if its parameters are close to the true data generator GARCH(1,1), we will expect a large Error Type II. Conversely, if both model's parameters are not similar (the models are quite different from each other), then Error Type II should be very small since it will be easier to detect that the wrong model is not close to the true model.

Table 4.2 shows the comparison between the results of a GARCH(1,2) whose parameters are quite far from the true model (Panel A), and a GARCH(1,2) with parameters close to the true model (Panel B). For the model that we consider is "far" from the true model, we set $\alpha_0 = 1e - 6, \alpha_1 = 0.1, \beta_1 = 0.6$ and $\beta_2 = 0.01$. The only differences with the original GARCH(1,1) defined before are the extra parameter β_2, and that while β_1 of the GARCH(1,1) is set as $\beta_1 = 0.888$, the β_1 parameter of the GARCH(1,2) model is set to 0.6. As expected, the backtest shows that Error Type II results in Panel A are close to zero since it is easy detect that one model is quite different from the other.

Moving on to the results in Panel B, we find the backtest results obtained by using a GARCH(1,2) as the wrong model, with parameters very similar to the ones of the true model. Recall that the parameters of the GARCH(1,1), which is the true data generator are: $\alpha_0 = 1e - 6, \alpha_1 = 0.1, \beta_1 = 0.888$. In this case, the GARCH(1,2) has the same values for all its parameters. The extra parameter is set as $\beta_2 = 0.01$. Because β_2 is small, we see a larger Error Type II in comparison to the results in Panel A, since the parameters of both models are quite close. As expected, this makes the backtest to have less power in order to detect the wrong model.

Furthermore, in Panel C, we present the results obtained using the HS method. Since the computation of the daily volatility σ_t with the HS method is constant for all values

(a) GARCH with parameters far from the true model

IF/n	750	5000
1	0	0
2	0	0
4	0	0
1-2	0	0
1-4	0	0
2-4	0	0
1-2-4	0	0

(b) GARCH with parameters close to the true model

IF/n	750	5000
1	0.8066	0.2792
2	0.8042	0.3522
4	0.8170	0.3040
1-2	0.8126	0.3730
1-4	0.8218	0.3690
2-4	0.8120	0.3670
1-2-4	0.8002	0.4192

(c) Historical Simulation method

IF/n	750	5000
2	0.3572	0.2980

Table 4.2: Impact of the choice of the wrong model on Error Type II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. In Panel A we observe the results of the wrong model that come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.6$, and $\beta_2 = 0.01$. In Panel B we observe the results of the wrong model that come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and $\beta_2 = 0.01$. The number of simulations is 5,000, $\kappa = 90\%$, and $\alpha = 95\%$.

of time t, the results are the same for instrumental functions 1, 2, 4 and 6. Thus, we only use instrumental function 2. Given that the conditional expectile depends on σ_t it will be constant too. Indeed, all instrumental functions that involve σ_t or $e_t(\kappa)$ will be constant over t. For this reason, we only use instrumental function 2 for the results in Panel C in Table 4.2.

Next, in Table 4.3 we show the results of a backtesting procedure with a different true data generator GARCH(1,1) model with different parameters as the previous one. In other words, we use a different "true model" to generate the returns. We set $\alpha_1 = 0.1, \beta_1 = 0.8$ (before β_1 was 0.888). After we run the simulations, we conclude that there is no improvement in Error I when setting different parameters for the GARCH(1,1) model (the correct model). However, Error Type II is lower in Table 4.1 since the sum of the parameters of the model used to compute the results of that table is closer to 1. In general, Error Type II tends to decrease as the sum of the parameters of the GARCH model $\alpha_1 + \beta_1$ approach to 1. If the sum of the parameters of the model is close to one, its volatility will be highly persistent, and as a consequence, the past volatility values have bigger influence over future volatility values.

Finally, Table A.1 in the Appendix is useful to run a simulation check. The objective of the table is simply to check that if the wrong model is in fact the correct model, then the results for Error Type I should be the complement of results for Error Type II.

4.2.5 Impact of the number of simulations on both errors

In Table 4.4, in the first column we show the results using 5,000 simulations. Then we reproduce the same procedure using 10,000 simulations to analyze if there are important variations in the results.

We conclude that we can keep the number of simulations to 5,000, because the results

(a) Error Type I

IF/n	250	750	1500
1	0.0578	0.0536	0.0534
2	0.0564	0.0536	0.0544
4	0.057	0.0536	0.0532
1-2	0.0868	0.0636	0.0576
1-4	0.0788	0.0624	0.0568
2-4	0.0824	0.0618	0.058
1-2-4	0.1356	0.0874	0.0752

(b) Error Type II

IF/n	250	750	1500
1	0.9148	0.8966	0.8612
2	0.9136	0.8938	0.8584
4	0.9156	0.896	0.8628
1-2	0.8864	0.8946	0.8754
1-4	0.8904	0.8986	0.8752
2-4	0.8884	0.8972	0.8766
1-2-4	0.8356	0.869	0.8592

Table 4.3: Impact of the choice of the wrong model on Error Type I and II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.80$. The returns of the wrong model come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.80$, and $\beta_2 = 0.01$. The number of simulations is 5,000, $\kappa = 90\%$, and $\alpha = 95\%$.

(a) Error Type I

IF/NSim	5,000	10,000
1	0.0498	0.0507
2	0.0554	0.0550
4	0.0502	0.0505
1-2	0.0654	0.0663
1-4	0.0616	0.0615
2-4	0.0644	0.0634
1-2-4	0.0912	0.0883

(b) Error Type II

IF/NSim	5,000	10,000
1	0.8044	0.8046
2	0.7958	0.7979
4	0.8168	0.8142
1-2	0.8102	0.8054
1-4	0.8126	0.8114
2-4	0.8098	0.8057
1-2-4	0.7942	0.7941

Table 4.4: Impact of the number of simulations on Error Type I and II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. The returns of the wrong model come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and $\beta_2 = 0.01$. The number of simulations is 5,000 and 10,000, $n = 750$, $\kappa = 90\%$, and $\alpha = 95\%$.

do not change substantially. This saves time and computational costs when running the backtesting procedures.

4.2.6 Impact of the level of the conditional expectile

Table 4.5 shows how the results for Error Type I and II change when we set a different κ level used to compute the *theoretical* expectile $e(\kappa)$ of the residual ϵ_t which consequently affects the computation of the *conditional* expectile $e_t(\kappa)$. We replicate the same procedure for different instrumental functions, and across various combinations.

As displayed in Panel A, Error Type I results are intuitive since they stay closer to 1 - α for $\kappa = 0.90$ and 0.95. The worst scenario occurs when we set κ at 99% and use three instrumental functions because Error Type I is above 10%, which is too far from 1 - α. In Panel B, we clearly see that Error Type II is higher when $\kappa = 0.90$. Although the results for $\kappa = 0.95$ are slightly better than the ones when $\kappa = 0.99$ when using one instrumental function, the improvement is negligible. When adding more instrumental functions, Error Type II clearly increases in most cases.

4.2.7 Impact of the distribution of the residual

As stated before, the computation of the conditional expectile at time t is as follows:

$$e_t(\kappa) = \sigma_t \cdot e(\kappa), \tag{4.17}$$

where $e(\kappa)$ is the *theoretical* expectile of standard Normal or t-Student residuals ϵ_t. In the previous simulations, we have computed the conditional expectile assuming that the residuals follow a Normal distribution. In this section, we now assume that the distribution of the residuals is t-Student with v degrees of freedom.

(a) Error Type I

IF/κ	0.9	0.95	0.99
1	0.0524	0.0494	0.0546
2	0.0494	0.0518	0.0516
4	0.0536	0.0518	0.0562
1-2	0.0514	0.0538	0.075
1-4	0.0486	0.0492	0.06
2-4	0.052	0.0508	0.0664
1-2-4	0.0586	0.07	0.1126

(b) Error Type II

IF/κ	0.9	0.95	0.99
1	0.2792	0.179	0.1918
2	0.3522	0.245	0.2494
4	0.304	0.1982	0.2128
1-2	0.373	0.2534	0.2488
1-4	0.369	0.2464	0.2442
2-4	0.367	0.2468	0.2428
1-2-4	0.4192	0.288	0.2564

Table 4.5: Impact of the level of the conditional expectile on Error Type I and II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. The returns of the wrong model come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and $\beta_2 = 0.01$. The number of simulations is 5,000, n = 5,000, and α = 95%.

The objective of Table 4.6, is to analyze how Error Type I and II develop as we increase the number of degrees of freedom used to compute the *theoretical* expectiles, and also to compare these t-Student results with the ones of the Normal distribution.

As we know, a t-Student distribution converges to a Normal distribution as the degrees of freedom go to infinity. Empirically, we would expect that 30 degrees of freedom are sufficient for the t-Student values to be very close to the ones of a Normal distribution. However, the convergence of the expectiles of a t-Student distribution might be slower, and require more degrees of freedom to converge to the expectiles of a standard Normal distribution. In other words, given the same number of degrees of freedom, the quantiles of a t-Student distribution will be closer to the quantiles of the Normal distribution, than what the expectiles of a t-Student will be from the expectiles of a standard Normal distribution.

Refer to Table 4.6 to see the Error results where the distribution of the *theoretical* expectile is t-Student across various degrees of freedom. We show that for 5,000 simulations and a sample size n of 5,000, Error Type I increases when adding more instrumental functions to the computation of the Wald-statistic. Moreover, note that for instrumental function 2, Error Type I seems to slightly go away from $1 - \alpha$ as the number of degrees of freedom increase. Conversely, it approaches approaches to $1 - \alpha$ when using the combination of instrumental functions 1 and 2 at the same time.

Furthermore, when analyzing Error Type II, we note that it increases as we add more instrumental functions to compute the Wald test statistic, however it decreases when increasing the number of degrees of freedom used to compute the conditional expectile.

We conclude that for this case, the Normal distribution shows better Error Type II results than the ones of the t-Student distribution. This could be explained by the fact that a GARCH model with t-Student innovations and a small value of degrees of freedom has higher volatility, and as a consequence, it is more difficult to backtest. In the case of Error Type I, the Normal model seems to perform better t-Student when we use more instrumental function to compute the Wald-statistic, however, there is not substantial difference in the results when using one or two instrumental functions.

(a) Error Type I

IF/v	4	6	8	30	N(0,1)
1	0.0524	0.0518	0.0484	0.0528	0.0472
2	0.0506	0.0526	0.0522	0.0530	0.0490
4	0.0530	0.0512	0.0494	0.0532	0.0462
1-2	0.0638	0.0626	0.0568	0.0588	0.0528
1-4	0.0584	0.0590	0.0510	0.0582	0.0478
2-4	0.0554	0.0576	0.0530	0.0566	0.0508
1-2-4	0.0830	0.0820	0.0694	0.0674	0.0610

(b) Error Type II

IF/v	4	6	8	30	N(0,1)
1	0.5616	0.4560	0.3852	0.2984	0.2856
2	0.5956	0.5206	0.4596	0.3760	0.3542
4	0.5854	0.4840	0.4194	0.3314	0.3102
1-2	0.6180	0.5324	0.4726	0.3940	0.3766
1-4	0.6152	0.5294	0.4670	0.3876	0.3714
2-4	0.6056	0.5240	0.4646	0.3884	0.3708
1-2-4	0.6172	0.5586	0.5110	0.4314	0.4162

Table 4.6: Impact of the distribution of the residual on Error Type I and II. Results obtained from a backtesting procedure utilizing instrumental functions 1, 2, 4 and their respective combinations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. The returns of the wrong model come from a GARCH(1,2) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, $\beta_1 = 0.888$, and $\beta_2 = 0.01$. The number of simulations is 5,000, $n =$ 5,000, $\kappa = 0.9$, and $\alpha = 95\%$.

4.3 Empirical application

Summary of the methodology In this last section, we apply the backtesting method proposed to log-returns defined as $X_t = ln(P_t) - ln(P_{t-1})$ where P_t is the closing price of the S&P 500 at time t. The data recovered contains 2,541 elements and goes from May 28th, 2013 to May 26th of 2023. The datasource is the financial platform Investing: https://ca.investing.com/indices/us-spx-500.

After converting 2,541 closing prices to 2,540 log-returns, we arbitrarily split the data into in-sample data and out-of-sample data. The first half has 2,040 log-returns and will be used to estimate the parameters of a GARCH(1,1) model. The second half with 500 elements will be used to run the backtesting procedure suggested in the thesis. Note that both the in and out-of-sample data sum up to 2,540; one element less than the original data given that we computed the log difference of the closing prices. Once we have the log-returns, we estimate and substract the mean to the vector of log-returns, since a zero-mean vector is necessary to estimate the GARCH model parameters.

First, we use the R package 'rugarch' to estimate the parameters ω, α_1 and β_1 of the GARCH(1,1) model, and then we verify the condition of stationarity: $0 < \omega + \alpha_1 + \beta_1 < 1$. Next, we compute the daily volatility for all times t up to 2,540, using the first real return and the unconditional variance $\sigma_1^2 = \frac{\omega}{1-\alpha_1-\beta_1}$ as the first volatility. With the volatility, we compute the last 500 daily conditional expectiles. Finally, we run the Wald-test to either accept or reject the null hypothesis for a GARCH(1,1) with Normal or t-Student innovations.

We recall the Null Hypothesis as:

$\mathcal{H}_0$: The expectile forecasts are correct.

If we reject $\mathcal{H}_0$, we say that the procedure "Fails" the test, otherwise, if the Null

Hypothesis is not rejected, we say that the procedure "Passes" the test.

4.3.1 Impact of the level of the conditional expectile

In this section we indicate whether or not a model passes the backtest. We fit the data to GARCH(1,1) models with either Normal or t-Student innovations. We also compare the normal-GARCH(1,1) vs. normal-GARCH(1,2).

In Table 4.7, in Panel A we observe that for the normal-GARCH(1,1) model, the procedure passes the test in all cases of instrumental functions when $\kappa = 99\%$, and also in two cases when $\kappa = 95\%$ for the first and second instrumental functions. The model only passes the test at $\kappa = 90\%$ when using the second instrumental function. On the other hand, in Panel B where the model has t-Student innovations, the model passes the test in the exact same cases that where the normal-GARCH(1,1) model passed the test. However, to make a distinction to conclude which model might be better, we can observe to the Wald-statistic of each model. We can conclude that the normal-GARCH model is a better fit to the data since the t-Student model has smaller Wald-statistics when $\kappa = 90\%$, but the Normal GARCH model shows smaller Wald-statistics when κ = 95 or 99%. This might be expected because we also found that the GARCH model with Normal innovations had more power than the t-Student model [refer to Section 4.2.7]. Although Monte Carlo simulations are a good indicator to determine how well a model might perform in practice, the performance of the backtest also depends on the dataset used. In this specific scenario, the normal-GARCH seem to perform better than the t-Student model, nonetheless, there will be cases where a model with t-Student innovations will better fit the data.

In Table 4.7 we observe the same Pass/Fail results for Panel A and C, which means that both models might be similar. When taking a closer look to both the critical values

(a) normal GARCH(1,1)

IF	CV_{95}	W_{90}	W_{95}	W_{99}	$H_{0,90}$	$H_{0,95}$	$H_{0,99}$
1	3.84	5.18	2.18	0.1	Fail	Pass	Pass
2	3.84	2.05	1.02	0.05	Pass	Pass	Pass
4	3.84	9.95	4.24	0.27	Fail	Fail	Pass
1-2	5.99	15.52	5.99	0.28	Fail	Fail	Pass

(b) *t*-Student GARCH(1,1)

IF	CV_{95}	W_{90}	W_{95}	W_{99}	$H_{0,90}$	$H_{0,95}$	$H_{0,99}$
1	3.84	4.45	2.97	2.25	Fail	Pass	Pass
2	3.84	1.58	1.56	2.22	Pass	Pass	Pass
4	3.84	9.03	5.32	2.6	Fail	Fail	Pass
1-2	5.99	14.88	6.62	2.26	Fail	Fail	Pass

(c) normal GARCH(1,2)

IF	CV_{95}	W_{90}	W_{95}	W_{99}	$H_{0,90}$	$H_{0,95}$	$H_{0,99}$
1	3.84	5.72	2.57	0.19	Fail	Pass	Pass
2	3.84	2.38	1.27	0.11	Pass	Pass	Pass
4	3.84	10.72	4.8	0.4	Fail	Fail	Pass
1-2	5.99	15.86	6.31	0.35	Fail	Fail	Pass

Table 4.7: Impact of the level of the conditional expectile in real data. Results obtained from a backtesting procedure utilizing instrumental functions 1 and 2. From left to right, we have the Chi-square critical value at 95%, then the Wald-statistic at κ levels 90, 95 and 99%. The last three columns indicate "Fail" when we reject the Null Hypothesis at $\alpha\% = 95\%$. We estimate the parameters using 2,540 daily log-returns from the S&P500 where the in-sample data size is 2,040 and the out-of-sample size is 500.

and the test statistics of the two GARCH models, we observe that the statistical tests for the GARCH(1,1) model are slightly closer to the critical value than the ones of the GARCH(1,2). This might suggest that the GARCH(1,1) is closer to not reject the null hypothesis, and as a consequence, indicates that the model better fits the data compared to the GARCH(1,2).

4.3.2 Impact of the split of the data

In Table 4.8, we find that for $\kappa = 90\%$, the 70-30 split passes the test in 2 occasions whereas the 80-20 split in only one. In the last column, we observe better results with the 80-20 split when $\kappa = 99\%$ since the backtesting procedure passes the test in all cases. This could be explained by the fact that we are using more data to estimate the parameters of the GARCH model compared to the 70-30 split, and we run a backtest with less number of log-returns forecasts into the future (20% as in-sample data compared to 30%).

(a) t-Student GARCH(1,1) with 80-20 split

IF	CV_{95}	W_{90}	W_{95}	W_{99}	$H_{0,90}$	$H_{0,95}$	$H_{0,99}$
1	3.84	4.45	2.97	2.25	Fail	Pass	Pass
2	3.84	1.58	1.56	2.22	Pass	Pass	Pass
4	3.84	9.03	5.32	2.6	Fail	Fail	Pass
1-2	5.99	14.88	6.62	2.26	Fail	Fail	Pass

(b) t-Student GARCH(1,1) with 70-30 split

IF	CV_{95}	W_{90}	W_{95}	W_{99}	$H_{0,90}$	$H_{0,95}$	$H_{0,99}$
1	3.84	3.29	2.6	4.23	Pass	Pass	Fail
2	3.84	0.39	0.83	5.8	Pass	Pass	Fail
4	3.84	8.87	5.49	3.77	Fail	Fail	Pass
1-2	5.99	13.69	6.26	7.27	Fail	Fail	Fail

Table 4.8: Impact of the split of the data (empirical application). Results obtained from a backtesting procedure utilizing instrumental functions 1 and 2. From left to right, we have the Chi-square critical value at 95%, then the Wald-statistics at κ levels 90, 95 and 99%. The last three columns indicate "Fail" when we reject the Null Hypothesis at $\alpha\% = 95\%$. In the first case, we use 2,540 daily log-returns from the S&P500 where the in-sample data size is 2,040 and the out-of-sample size is 500. In panel B we use 1,790 log-returns for the in-sample data and 750 for the out-of-sample data.

4.4 Extensions

4.4.1 Realized volatility approach and Bayesian Paradigm

In the Monte Carlo Simulations section, we used GARCH models to simulate daily returns based on the assumption that the residuals follow a parametric distribution. Using time series model qualifies as a robustness analysis. However, it would also be interesting to analyze what happens if we abandon the approach of time series models, and use a model-free approach instead, say, the realized volatility approach which is quite prevalent in the financial literature nowadays (Barndorff-Nielsen and Shephard, 2002).

For that purpose, we can fix a time window. See an instance of the realized variance formula based on the daily centered log-returns below:

$$\sigma^2_{realized} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} X_t^2}, \tag{4.18}$$

where X_t are the centered log-returns at day t, and n the total number of trading days in a period of interest..

In contrast to the HS method, we replicate this for a set of sub-samples (rolling window-based estimation) to finally combine them together to find an estimate of the realized volatility for each day. So, the realized volatility at time t is not constant over time compared to the HS method.

In this case, the question of interest is: Would the results still hold if we use a model-free approach, for instance, the *realized volatility* approach?

Finally, another alternative approach is to do it through a Bayesian Paradigm. In this case, a prior distribution (arbitrary) should be used, in order to backtest the data from a Bayesian focus. As an example, Zelvyte and Arnsdorf (2023) discuss the easy implementation of the procedure, and explain how the posterior distribution allows an easy showcase of the results. In addition, they criticize the frequentist approach with arguments such as the small power that backtests with a frequentist method have, for instance, by discussing how one can arrive to the wrong conclusion when interpreting the p-value.

4.5 Conclusion

This thesis proposes a traditional backtesting procedure for expectiles that is easy to implement, since we only require to reveal public information such as the daily returns and the daily conditional expectile. The simplicity of this backtest preserves companies' confidentiality because it does not require to reveal the company's internal model. Instead, it only requires to compute the daily volatility of the model of choice, which can be estimated from, for instance, GARCH models as shown in Section 4.3. Furthermore, the implementation of the test is straightforward as observed in Section 4.1 where the construction of the Wald test statistic is easily performed using daily log-returns, daily conditional expectiles, and instrumental functions. Further, the comparison of the Wald test statistic is simply done against a chi-square critical value in order to conclude if the model passes or fails the test.

Moreover, Error Type I and Error Type II were analyzed based on the effect of the sample size, model selection, level of the expectile, distribution of the model's innovations. Indeed, the versatility of the backtest allows us to the procedure not only with simple time series models, but with more complex models with autoregressive components. In addition, new instrumental functions were proposed and tested both individually and in combinations of two or three at the same time.

For further research, we suggest proposing more instrumental functions to run the test as they will provide different metrics to compute the size and power of the test. Particularly, in the case of instrumental function 6, we suggest to find a method to find which is optimal lagged value to be used in such instrumental function depending on the data used.

Appendix A

Simulations

(a) Error Type I

IF/n	250	750	1500
1	0.0616	0.0498	0.0544
2	0.0676	0.0554	0.0538
4	0.0608	0.0494	0.053
1-2	0.0986	0.0656	0.0646
1-4	0.0914	0.0612	0.0574
2-4	0.0962	0.0668	0.0616
1-2-4	0.1334	0.0908	0.0774

(b) Error Type II

IF/n	250	750	1500
1	0.9384	0.9502	0.9456
2	0.9324	0.9446	0.9462
4	0.9392	0.9506	0.947
1-2	0.9014	0.9344	0.9354
1-4	0.9086	0.9388	0.9426
2-4	0.9038	0.9332	0.9384
1-2-4	0.8666	0.9092	0.9226

Table A.1: Simulation check. Error Type I and II results obtained from a backtesting proceedure utilizing instrumental functions 1, 2, 4, and running 5,000 simulations. The true data comes from a GARCH(1,1) model with $\alpha_0 = 1e-6$, $\alpha_1 = 0.1$, and $\beta_1 = 0.888$. The level is $\kappa = 90\%$. The returns of the wrong model are computed assuming that the wrong model is in fact the correct model; hence, Error Type I = 1 - Error Type II.

Bibliography

Acerbi, C. and Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11):76–81.

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

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

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (2002). Backtesting derivative portfolios with filtered historical simulation (fhs). *European Financial Management*, 8(1):31–58.

Basel Committee on Banking Supervision (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework. `https://www.bis.org/publ/bcbs107.htm`.

Bellini, F. and Di Bernardino, E. (2017). Risk management with expectiles. *European Journal of Finance*, 23(6):487–506.

Bellini, F., Klar, B., Müller, A., and Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48.

Bellini, F., Negri, I., and Pyatkova, M. (2019). Backtesting VaR and expectiles with realized scores. *Statistical Methods & Applications*, 28(1):119–142.

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating Value-at-Risk models with desk-level data. *Management Science*, 57(12):2213–2227.

Bontemps, C. and Meddahi, N. (2012). Testing distributional assumptions: A GMM aproach. *Journal of Applied Econometrics*, 27(6):978–1012.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S. (2011). Backtesting Value-at-Risk: a GMM duration-based test. *Journal of Financial Econometrics*, 9(2):314–343.

Chen, Z. (1996). Conditional lp-quantiles and their application to the testing of symmetry in non-parametric regression. *Statistics & probability letters*, 29(2):107–115.

Christoffersen, P. (2011). *Elements of financial risk management*. Academic press.

Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2(1):84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

Colletaz, G., Hurlin, C., and Pérignon, C. (2013). The risk map: A new tool for validating risk models. *Journal of Banking & Finance*, 37(10):3843–3854.

Costanzino, N. and Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation*, 9(1):21–31.

Daouia, A., Girard, S., and Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):263–292.

De Clerk, L. and Savel'ev, S. (2022). An investigation of higher order moments of empirical financial data and their implications to risk. *Heliyon*, 8(2):e08833.

Directive 2009/138/EC of the European Parliament and of the Council (2009). Solvency ii directive 2009. `https://eur-lex.europa.eu/eli/dir/2009/138/2021-06-30`.

Du, Z. and Escanciano, J. C. (2017). Backtesting expected shortfall: accounting for tail risk. *Management Science*, 63(4):940–958.

Emmer, S., Kratz, M., and Tasche, D. (2013). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, 18(2):31–60.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of business & economic statistics*, 22(4):367–381.

Fissler, T., Ziegel, J. F., and Gneiting, T. (2015). Expected shortfall is jointly elicitable with Value at Risk-implications for backtesting. *Risk Magazine*.

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

Girard, S., Stupfler, G., and Usseglio-Carleve, A. (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *Annals of Statistics*, 49(6):3358–3382.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gordy, M. B. and McNeil, A. J. (2020). Spectral backtests of forecast distributions with application to risk management. *Journal of Banking & Finance*, 116:105817.

Hannan, E. J. (1979). The central limit theorem for time series regression. *Stochastic Processes and their Applications*, 9(3):281–289.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054.

Holzmann, H. and Klar, B. (2016). Expectile asymptotics. *Electronic Journal of Statistics*, 10(2):2355–2371.

Kendall, M. G. et al. (1946). The advanced theory of statistics. *The Advanced Theory of Statistics.*, (2nd Ed).

Kratz, M., Lok, Y. H., and McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, 88:393–407.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89.

Löser, R., Wied, D., and Ziggel, D. (2018). New backtests for unconditional coverage of expected shortfall. *Journal of Risk*, 21(4):1–21.

Mandelbrot, B. (1963). New methods in statistical economics. *Journal of Political Economy*, 71(5):421–440.

McLeish, D. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(3):165–178.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3-4):271–300.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition.* Princeton University Press.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, pages 819–847.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4):1833–1874.

Pelletier, D. and Wei, W. (2016). The geometric-VaR backtesting method. *Journal of Financial Econometrics*, 14(4):725–745.

Pérignon, C. and Smith, D. R. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, 34(2):362–377.

Philipps, C. (2022). Interpreting expectiles. *Available at SSRN 3881402.*

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4):755–767.

Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16(2):419–441.

Wooldridge, J. M. and White, H. (1988). Some invariance principles and central limit theorems for dependent heterogeneous processes. *Econometric Theory*, 4(2):210–230.

Zelvyte, M. and Arnsdorf, M. (2023). Bayesian backtesting for counterparty risk models. *Journal of Risk Model Validation*, 17(2).