Faculté des sciences

Backtesting of risk measures in dynamic models.

Auteur·es : Catarina Monteiro dos Ramos
Promoteur·rices : Christian Hafner
Lecteur·rices : Pierre Ars
Année académique 2024-2025

Preface.

This thesis, entitled Backtesting of Risk Measures in Dynamic Models, was prepared in partial fulfillment of the requirements for the Master's degree in Actuarial Sciences at the Université Catholique de Louvain. Its primary aim is to provide a comprehensive overview of three tail risk measures, namely Value-at-Risk, Expected Shortfall, and Expectiles, along with their theoretical properties and practical backtesting within dynamic volatility frameworks, such as GARCH and GJR-GARCH.

Chapters 1 and 2 introduce the motivation to move beyond traditional variance-based risk metrics and define the three risk measures. Chapter 3 develops the stochastic volatility models used to generate and estimate the risk measures, while Chapter 4 presents the backtests performed to validate them. Chapter 5 applies these concepts to historical S&P 500 data, comparing estimation and backtesting results under normal and Student-t innovations. The final chapter summarizes key insights and suggests directions for future research.

I hope that this work not only deepens our understanding of backtesting methodologies but also offers practical guidance for risk managers seeking robust, tail-focused tools in volatile markets.

Acknowledgment.

I am deeply grateful to my supervisor, Prof. Christian Hafner, for his expert guidance, constant availability, and patient support throughout the process of preparing this thesis.

I would also like to thank the professors at the Université Catholique de Louvain for sharing their extensive knowledge and delivering such high-quality instruction throughout my degree.

Finally, I owe my sincere thanks to my parents for their encouragement, understanding, and patience throughout my studies.

Contents

Chapter 1

Introduction.

Financial risk management is essential to ensure the stability and solvency of financial institutions. For this purpose, risk measurement is used through various risk measures. Risk measures can be defined as statistical amounts characterizing the portfolio's loss distribution over a given time period, focusing primarily on the tails of the distribution to capture extreme events. Traditional measures such as variance proved inadequate for quantifying tail risks, leading to the development and adoption of more sophisticated tail risk measures in both industry and regulation.

The introduction of J.P. Morgan's RiskMetrics methodology in 1994 popularized Value-at-Risk (VaR) as the standard metric for market risk. Since then, the most widely used risk measure by financial institutions has become the value at risk (VaR). The Basel I Accord of 1988 laid the foundation for global bank capital requirements, but it was the later amendments and Basel II (2004) that formally incorporated VaR into regulatory practice. Under Basel II guidelines, banks could use internal VaR models to determine capital charges for market risk. VaR was then included in the Basel Committee's market risk framework and in Europe's Solvency II framework for insurers. However, the global financial crisis of 2007-2008 exposed weaknesses in solely relying on VaR, a non-coherent risk measure that does not reflect the severity of losses beyond the VaR threshold. In response, the Basel III reforms, specifically the Fundamental Review of the Trading Book (FRTB) finalized in 2016, announced a shift from VaR to Expected Shortfall (ES) as the official market risk metric. In addition, the Swiss Solvency Test for insurers also replaced VaR with ES for risk management purposes. In fact, the ES better captures tail risk and the magnitude of extreme losses that VaR might overlook. Namely, the VaR at level α represents the threshold loss that will not be exceeded with probability α over a given time period, while the ES is the expected loss given that the loss exceeds the VaR at level α. Although the ES is a coherent risk measure, the ES also has a limitation, namely that it is not elicitable, which makes comparing models more challenging. The lack of elicitability of ES in comparison with the elicitable risk measure VaR initially raised concerns that ES might be difficult or even impossible to backtest, fueling debate over the Basel Committee's choice to adopt ES in place of VaR. However, Acerbi and Szekely (2014) demonstrated that ES can be backtested with appropriate methods and introduced several model-independent backtesting approaches for the ES. Given the respective limitations of VaR and ES, researchers have explored alternative risk measures that could overcome both shortcomings. One such alternative is the expectile, a less conventional risk measure that has recently gained attention because it is the only risk measure that is coherent and elicitable. An expectile can be thought of as a tail-focused percentile-like measure defined via an asymmetric least squares criterion. However, expectiles are not yet part of regulatory regimes or common industry practice, partly because they are relatively new to financial risk management and partly because expectile backtesting procedures have not been well-established until recently, namely in the paper by Armando et al. (2024).

Backtesting methods are used to analyze the performance of estimated risk measures. The purpose of this thesis is to get an overview of the backtesting methods for the three main risk measures, namely Value-at-Risk, Expected Shortfall, and Expectile, in the context of dynamic models such as stochastic volatility, and to investigate the performance of these backtesting methods. To do so, this thesis first estimates each risk measure, including the VaR, the ES, and the expectile, and then backtests the risk measures. The estimation of the risk measures and the backtesting methods for the value at risk and the expected shortfall of this thesis are mainly retrieved from the book by McNeil et al. (2015). The backtesting methods used for the expectile come from the paper by Armando et al. (2024).

The remainder of this thesis is organized as follows. Chapter 2 defines the three main risk measures, VaR, ES, and expectile. It also discusses the theoretical properties of these risk measures, namely coherence and elicitability, and examines how VaR, ES, and expectile each satisfy or fall short of these properties. Chapter 3 describes the GARCH family of models used in detail, namely GARCH(1,1) and GJR-GARCH(1,1). Simulation procedures are described to generate dynamic market scenarios, where returns are simulated using a GARCH(1,1) model with normally distributed innovations, and the risk measures are estimated for the simulated returns. A comparison between risk measures is also made. Chapter 4 introduces various backtesting methods for the risk measures, applies the backtesting methods to the simulated GARCH(1,1) returns, and looks at their performance by doing a bootstrap. Chapter 5 presents an empirical application of these concepts. Risk measures are estimated on real financial data (daily S&P 500 returns) and their performance is evaluated by backtesting using GARCH and GJR-GARCH models with normally distributed innovations. This practical study illustrates how VaR, ES, and expectile forecasts behave in a real-world context and compares their effectiveness in capturing market risk. Chapter 5 then looks at the improvement of the risk measures estimates and of the backtesting results when we apply the GARCH and GJR-GARCH models with Student-t distributed innovations to the S&P500 returns, especially GJR-GARCH models with Student-t distributed innovations. Chapter 6 concludes the thesis by summarizing the key findings, discussing their implications for financial risk management, and suggesting avenues for future research. Through this structure, the thesis aims to provide a comprehensive understanding of risk measures and their backtesting methods.

Chapter 2

Risk Measures

2.1 Overview

2.1.1 Value-at-risk

Most financial institutions use the value at risk, denoted by VaR, as a market risk measure. That is, the value at risk affected two major regulatory frameworks, namely the Basel Committee regulatory guidelines, called the Basel Framework, and the Solvency II framework.

The Value-at-Risk can be seen as the solution to an optimization problem where the loss function is given by equation (2.1).

$$S^q_\alpha(y, L) = |1_{\{L \leq y\}} - \alpha||L - y|, \tag{2.1}$$

where the probability level α is between 0 and 1, L is the loss, and y, namely the VaR, is the variable we are looking for: we seek y such that the scoring function S^q_α, where the superscript q indicates that it is the quantile score used for the Value-at-Risk at level α, is minimized (McNeil et al., 2015). The scoring function is the absolute value of the difference between an indicator function where the loss is less than or equal to y and α multiplied by the absolute value of the difference between the loss and y. Note that the indicator is 1 if the condition is true and 0 otherwise. The VaR, i.e. the quantile at the α level, is the one that minimizes this loss function. In other words, $y = F^{\leftarrow}_L(\alpha)$ is the only variable that minimizes the expected score $S^q_\alpha(y, L)$, where F_L is the distribution function of the loss.

The value at risk of a loss L at the confidence level α is defined as the lowest l to ensure that the probability that the loss L is greater than that number l is lower than or equal to $1 - \alpha$, as defined by McNeil et al. (2015):

$$VaR_\alpha = VaR_\alpha(L) = F^{\leftarrow}_L(\alpha) = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\}, \tag{2.2}$$

and for continuous strictly increasing distributions F, $VaR_\alpha(L) = F^{-1}(q)$.

2.1.2 Expected Shortfall

The Expected Shortfall (ES) is defined by McNeil et al. (2015) as the average of VaR_q at all levels $q \geq \alpha$, that is:

$$ES_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 VaR_q(L)\, dq. \tag{2.3}$$

Alternatively, the expected shortfall is the average of all possible losses conditional on the fact that we are at the tail of the distribution, formally only for continuous losses:

$$ES_\alpha = E(L|L \geq VaR_\alpha), \tag{2.4}$$

i.e., the expected loss knowing that the loss is larger than or equal to the value at risk. It follows from the definition of the expected shortfall that the ES is always higher than (or equal to) the VaR, and by how much depends on the distribution, i.e., the fatter the tail, the farther away the ES is from the VaR.

2.1.3 Expectile

Similarly to the value at risk, the expectile is also a risk measure that minimizes a loss function, i.e. the expectile is the y that minimizes $E(S_\alpha^e(y, L))$. The scoring function is the same as the loss function for the VaR, except that instead of the absolute value of the difference between L and y, we have squares, so we replace L1 with L2. The corresponding scoring function S_α^e, where the superscript e indicates the expectile score, is given by the following equation:

$$S_\alpha^e(y, L) = |1_{\{L \leq y\}} - \alpha|(L-y)^2, \tag{2.5}$$

(McNeil et al., 2015). The solution to this minimization problem, that is, the y obtained by minimizing this scoring function, is no longer a VaR but an expectile. When we replace the absolute value with a square, we get a loss function that appears smoother but still asymmetric with respect to α. Note that the scoring function only depends on α and we are looking for the value of the expectile, denoted y, which minimizes the scoring function for a given loss L and a given α.

The y that minimizes the scoring function (2.1) is the value at risk, namely the quantile, and the y that minimizes the scoring function (2.5) is the expectile.

The α-expectile denoted by $e_\alpha(L)$ can be defined by the only solution y of the following equation:

$$\alpha E[(L-y)^+] - (1-\alpha)E[(L-y)^-] = 0, \tag{2.6}$$

where $v^+ = \max(x, 0)$, $v^- = \max(-x, 0)$, and we assume that $\mathcal{M} := L^1(\Omega, \mathcal{F}, P)$ is the set of all integrable random variables L with $E|L| < \infty$ and $L \in \mathcal{M}$ (McNeil et al., 2015).

To help the reader understand the difference between the risk measures, it is useful to compare the loss functions used to derive the VaR and the expectile. The VaR loss function (often called the pinball loss) is given by

$$\rho_\alpha(u) = \begin{cases} (1-\alpha)|u|, & u < 0, \\ \alpha|u|, & u \geq 0, \end{cases}$$

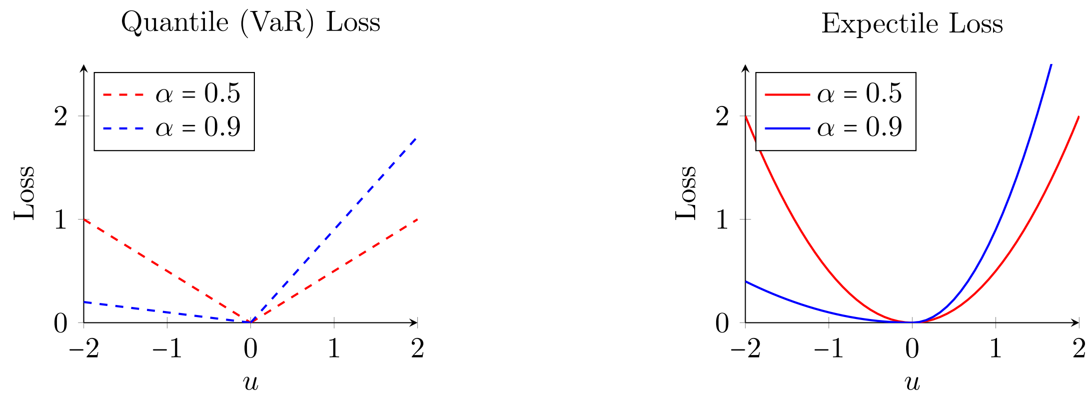
as introduced by Koenker and Bassett (1978) (see also Embrechts et al. (2015) for its application in risk management). This loss is piecewise linear and features a distinct kink at $u = 0$.

In contrast, the expectile loss function replaces the absolute error with a squared error, resulting in

$$\rho_\alpha(u) = \begin{cases} (1-\alpha)u^2, & u < 0, \\ \alpha\, u^2, & u \geq 0, \end{cases}$$

This formulation, which was proposed by Newey and Powell (1987), results in a smooth and differentiable loss function that facilitates gradient-based optimization and is more sensitive to larger deviations, a useful property in tail-risk assessment.

Figure 2.1.1 illustrates these differences by comparing the two loss functions for two values of α: 0.5 and 0.9. The left panel shows the quantile (VaR) loss function with its piecewise-linear structure, while the right panel displays the expectile loss with its quadratic, smooth shape. Note that for a α of 0.5, the value at risk becomes the median and the expectile becomes the mean.



(a) Quantile (VaR) Loss *(b) Expectile Loss*

Figure 2.1.1: Comparison of the loss functions for the quantile (VaR) and expectile approaches. The quantile loss is piecewise linear, while the expectile loss is quadratic and smooth.

We look at a numerical example for a standard normal $\mathcal{N}(0,1)$, for which the formulas for the VaR and the ES become, for the right tail of the distribution:

$$VaR_\alpha = \Phi^{-1}(\alpha)$$

$$ES_\alpha = \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}$$

Finally, the α-expectile e_α of $L \sim \mathcal{N}(0,1)$ is the unique solution y of the FOC given by equation (2.6):

$$\alpha\,\mathbb{E}\big[(L-y)^+\big] \;=\; (1-\alpha)\,\mathbb{E}\big[(y-L)^+\big].$$

Since for the standard normal we obtain

$$\mathbb{E}\big[(L-y)^+\big] = \phi(y) - y\big[1-\Phi(y)\big] \quad \text{and} \quad \mathbb{E}\big[(y-L)^+\big] = \phi(y) + y\,\Phi(y),$$

it follows Daouia et al. (2020) that e_α solves

$$(2\alpha-1)\,\phi(e_\alpha) \;=\; e_\alpha\Big[\alpha\big(1-\Phi(e_\alpha)\big) + (1-\alpha)\Phi(e_\alpha)\Big].$$

Table 2.1.1 presents the Value-at-Risk, Expected Shortfall, and Expectile for a standard normal loss distribution at confidence levels $\alpha = 0.90, 0.99, 0.999$. For $L \sim \mathcal{N}(0,1)$, we observe that

$$|ES_\alpha| > |VaR_\alpha| \quad \text{and} \quad |e_\alpha| < |VaR_\alpha|,$$

for all $\alpha \in \{0.90, 0.99, 0.999\}$, where we note that, in absolute value, the expected shortfall is in all cases greater than the value at risk, which is in turn in all cases greater than the expectile.

α	VaR_α	ES_α	e_α
0.90	1.2816	1.755	0.8616
0.99	2.3263	2.6652	1.7174
0.999	3.0902	3.3671	2.4358

Table 2.1.1: VaR, ES and expectile for $L \sim \mathcal{N}(0,1)$

2.2 Properties of Risk Measures

2.2.1 Coherence

In risk management, the concept of coherence is crucial when deciding which risk measure a financial institution selects. For a risk measure to be considered a coherent risk measure, it has to satisfy four axioms, namely monotonicity, subadditivity, positive homogeneity, and translation invariance. In this section, we briefly define a coherent risk measure and discuss how risk measures such as Value-at-Risk (VaR), Expected Shortfall (ES), and Expectile compare with respect to coherence.

Definition 1. *As defined by Artzner et al. (1999), a risk measure ρ is a coherent risk measure if the function $\rho : \mathbb{R} \to \mathbb{R}$ satisfies the following four axioms for any two financial positions X and Y and any scalar $\lambda > 0$:*

- *Axiom 1: Monotonicity: If $X \geq Y$ almost surely, then*

$$\rho(X) \geq \rho(Y).$$

  This property ensures that a position with uniformly higher losses is assigned a higher risk.

- *Axiom 2: Subadditivity:*

$$\rho(X+Y) \leq \rho(X) + \rho(Y).$$

  Subadditivity illustrates diversification, it guarantees that combining positions does not increase risk.

- *Axiom 3: Positive Homogeneity:*

$$\rho(\lambda X) = \lambda \rho(X).$$

  This property indicates that scaling a position by a positive factor scales the risk measure by the same factor.

- *Axiom 4: Translation Invariance:*

$$\rho(X+c) = \rho(X) + c,$$

  for any constant c. This means that adding a risk-free asset (or a fixed amount of cash) increases the risk measure by that amount.

When choosing which risk measure to use, financial institutions might look at which risk measures satisfy these axioms and are coherent risk measures. In particular, an important property is subadditivity because it encourages diversification, as it implies that aggregating risks does not result in an overestimation of the overall risk.

However, the choice of the best risk measure to apply is not solely based on coherence, since the most used risk measure, namely the VaR, is not a coherent risk measure because in some cases it can violate the subadditivity property. Although VaR satisfies monotonicity, positive homogeneity, and translation invariance, its lack of subadditivity can lead to counterintuitive situations where the VaR of a combined portfolio exceeds the sum of the individual VaRs.

On the other hand, ES is a coherent risk measure because it satisfies all four axioms. In particular, its subadditivity property ensures that risk is appropriately diversified across different positions. Although ES is theoretically appealing, one practical limitation is that it is not elicitable on its own, which dissuaded some financial institutions to use it because they believed that its lack of elicitability would complicate the backtesting of the ES.

Although the expectile is the less commonly used risk measure compared to VaR and ES, it is gaining popularity due to its coherence and elicitability.

2.2.2 Elicitability

An important concept in risk management and forecasting is elicitability. A risk measure is said to be elicitable if there exists a scoring function in which the true value of the risk measure uniquely minimizes the expected scoring function. The value at risk and the expectile are both elicitable risk measures, but the expected shortfall is not elicitable. This section investigates what makes a risk measure elicitable.

Definition 2. *For a random variable of interest Y and its realisation y, a risk measure denoted by ρ of Y is considered elicitable, if there is a function S, i.e. a (consistent) scoring function for the risk measure ρ, for which an estimate $\hat{\rho}$ minimizes the expected value of the scoring function $E[S(\hat{\rho}, Y)]$, (Armando et al., 2024).*

An alternative definition of an elicitable risk measure is based on the first-order (moment) condition of the minimization framework. In particular, Armando et al. (2024) let the partial derivative of S w.r.t. $\hat{\rho}$, i.e. $I(\hat{\rho}, Y)$, denote an identification function of the risk measure ρ, then suppose that $\hat{\rho} = \rho$ is the value that minimizes $E[S(x, Y)]$, we obtain:

$$E[I(\hat{\rho}, Y)] = 0 \tag{2.7}$$

The scoring function for the expectile is defined as follows:

$$S(\hat{e}, y) = \alpha[(y - \hat{e})^+]^2 + (1 - \alpha)[(y - \hat{e})^-]^2, \tag{2.8}$$

where $(y - \hat{e})^+ = \max(y - \hat{e}, 0)$ and $(y - \hat{e})^- = \min(y - \hat{e}, 0)$ and $\hat{e}$ represents the estimated expectile. This scoring function is referred to as the asymmetric least square loss, as for any chosen $\hat{e}$, the function $S(\hat{e}, y)$ is asymmetric about $\hat{e}$ for all $\alpha \in (0, 1)$, except when $\alpha = 0.5$, (Armando et al., 2024). In particular, the identification function for the expectile is the partial derivative of the scoring function, given by equation (2.10), with respect to $\hat{e}$:

$$I(\hat{e}, y) = \alpha(y - \hat{e})^+ - (1 - \alpha)(y - \hat{e})^-, \tag{2.9}$$

as defined by Armando et al. (2024).

The canonical identification function for the value at risk is defined as

$$I(\hat{q}, y) = 1 - \frac{1}{\alpha} 1_{\{y < \hat{q}\}}. \tag{2.10}$$

2.3 Conclusion

In this chapter, we introduce and compare three important risk measures: Value-at-Risk (VaR), Expected Shortfall (ES), and Expectile. VaR is widely used by financial institutions due to its simplicity and regulatory acceptance, but it has limitations since it does not capture the severity of extreme losses beyond the given threshold. To overcome this, the expected shortfall offers a clearer picture by averaging all potential extreme losses beyond VaR, thus better capturing tail risks. The expectile presents a balanced approach, using a smoother quadratic loss function that facilitates numerical estimation. Ultimately, choosing between these risk measures depends on regulatory standards, practical ease of use, and the specific risk scenario at hand.

Then, we examine two key properties of risk measures: coherence and elicitability. Coherent risk measures, characterized by the four axioms, i.e. monotonicity, subadditivity, positive homogeneity, and translation invariance, are particularly appealing in risk management due to their logical consistency and encouragement of diversification. Although VaR is widely used, its limitation is its lack of subadditivity, which can lead to unintuitive outcomes. In contrast, both the expected shortfall (ES) and the expectile satisfy coherence criteria, making them theoretically preferable from a risk-diversification point of view.

We also define the concept of elicitability, which concerns whether a risk measure can be reliably estimated by minimizing a suitable scoring function. We found that VaR and expectile are elicitable, facilitating straightforward backtesting and estimation procedures. In contrast, ES is not elicitable, posing practical challenges in its implementation. Nevertheless, given its coherence, ES remains highly valuable and used in practice. The insights from this chapter provide a solid theoretical foundation, which we will build upon through practical estimation and backtesting techniques in the following chapters.

Chapter 3

Dynamic Models for Risk Measures

Until now we were in a deterministic setting, and now we move to a dynamic setting, where we look at the conditional distribution of returns. In this chapter, we simulate return series based on dynamic models, namely the symmetric GARCH(1,1) and the asymmetric GJR-GARCH(1,1) model, and we estimate the corresponding value at risk, expected shortfall, and expectile conditional on historical data at each point in time.

3.1 GARCH(1,1) Model

To investigate whether the backtesting methods for the different risk measures are adequate and to evaluate their performance, we first simulate the GARCH(1,1) return series. The GARCH(1,1) model is commonly used to simulate returns series because it captures volatility clustering. Volatility clustering occurs when periods of high volatility are followed by periods of high volatility and similarly for low volatility.

This phenomenon of time-varying volatility was first formally modeled by Engle (1982) in his Autoregressive Conditional Heteroskedasticity (ARCH) framework. Engle (1982) addressed the problem that traditional time series models assumed constant variance, an assumption at odds with the observed clustering of large and small fluctuations in financial data. Bollerslev (1986) built on this idea by proposing the Generalized ARCH (GARCH) model, which allowed a more parsimonious representation of volatility dynamics by including lagged conditional variance terms. The GARCH$(1,1)$ specification thus captures the persistence in volatility shocks with just a few parameters, effectively modeling the fact that large returns (positive or negative) tend to be followed by large returns, and small returns by small returns, over time.

GARCH models have proven extremely useful for modeling financial return series because they address two major empirical features: volatility clustering and fat tails. By allowing volatility to change over time, these models generate periods of high variance and low variance, leading to an unconditional return distribution with much heavier tails (leptokurtosis) than would be predicted under a constant-volatility assumption. In other words, when volatility clusters, extreme returns (both large gains and losses) are more probable than under a homoskedastic model. This ability to capture both the persistence in volatility and the tendency for extreme returns has made GARCH$(1,1)$ a well-known model in financial econometrics. It often provides a good first approximation to the volatility dynamics of many assets (across equities, bonds, exchange rates, etc.), and it serves as a foundation for more complex risk modeling techniques in finance.

The purpose of first simulating return series instead of looking directly at real data of financial time series is that we already know the model that generates the return data. So, first, we need to fully understand the backtesting methods and check if they work with simulated data, where we know the true model that generated the data. Here, we know everything, and we expect the backtesting methods to work. If it does not work, then something is probably wrong with the programming. So, we will start by controlling the scenario in which we know the underlying reality. Once that is clear, we will apply the methods to real data. If they do not perform as expected, we will know we might need to adjust the distribution or perhaps fine-tune the GARCH model, for example.

Suppose that the model for the returns r_t is given by the following equation:

$$r_t = \mu_t + \sigma_t z_t, \ t \in \mathbb{Z} \tag{3.1}$$

where $\mu_t = E(r_t|\mathcal{F}_{t-1})$ is the conditional mean, $\sigma_t^2 = \mathbb{V}\text{ar}(r_t|\mathcal{F}_{t-1})$ is the conditional time-varying volatility, $\mathcal{F}_{t-1}$ is the information set at $t-1$ and $z_t \sim N(0,1)$. The assumption of standard normally distributed innovations does not change the fact that returns are fat tailed, it is indeed the distribution conditional on the past that is normally distributed. That is why from now on we will look at the conditional distribution and we add the subscript t to the value at risk, for example, denoted VaR_t. Note that μ_t can represent a non-zero risk premium, reflecting investors' compensation for bearing risk, which is typically positive over longer investment horizons. However, for simplicity and due to empirical observations that daily financial returns generally have means very close to zero and negligible linear autocorrelation, we assume $\mu_t = 0$ for the rest of the analysis. Indeed, financial return series typically do not exhibit significant linear autocorrelation, implying that past returns do not provide predictive power regarding future returns. However, empirical evidence indicates the presence of volatility clustering, which means that large price movements tend to be followed by similarly large movements, independent of sign. Furthermore, negative returns typically impact future volatility more strongly than positive returns of the same magnitude, a phenomenon known as the leverage effect. Although this motivates the use of asymmetric volatility models (such as GJR-GARCH), we initially focus on the simpler symmetric GARCH(1,1) model.

The standard GARCH(1,1) model, i.e. a Generalized AutoRegressive Conditional Heteroskedasticity model, introduced by Engle (1982) and Bollerslev (1986), is used to simulate the returns and to fit

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{3.2}$$

where $\epsilon_t = \sigma_t z_t$. We have three parameters, ω, α and β. To choose the parameters of the GARCH(1,1) model, we look at different possible values for α and β such that their sum is smaller than one to ensure stationarity. To select the parameters that we keep for the rest of the analysis, we look at the graph of the estimated conditional time-varying volatility and choose the parameters that better reflect the true volatilities of real return series. We come to the conclusion that a GARCH(1,1) model with $\mu = 0$, $\omega = 0.001$, $\alpha = 0.083$ and $\beta = 0.909$ reflects quite well what we would expect from volatilities of real return series. Figure 3.1.1 depicts the simulated volatility of GARCH (1,1) and the corresponding estimated GARCH(1,1) volatility with 1,000 observations.

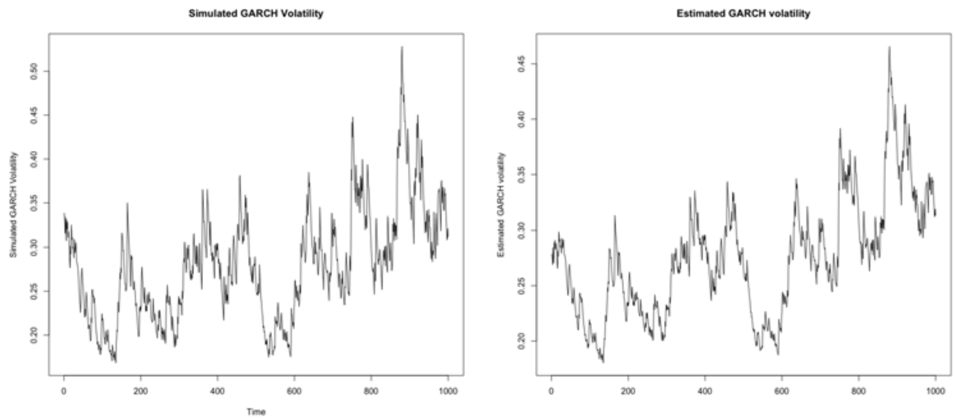


Figure 3.1.1: Simulated and estimated GARCH(1,1) volatility for n=1,000

The simulated GARCH(1,1) returns, the estimated VaR and the estimated ES for n = 1,000 are shown in Figure 3.1.2. The 1% VaR is the threshold such that the probability of returns falling below it is 1%. The VaR represents a limit of extreme negative returns that is unlikely to be exceeded in a normal market. The VaR line is mostly below the simulated return series, periods when the VaR becomes more negative demonstrate that the model predicts an increased likelihood of extreme losses represented by extreme negative returns, which often happens with higher market volatility. There is a VaR-exceedance whenever the returns fall below the VaR red line. The ES represents the average loss if the VaR threshold is actually breached, so the average loss whenever the returns fall below the VaR. The ES is more sensitive to the tail of the distribution and thus gives more information about the worst-case scenarios. From the plot and numerically, we confirm that the ES always falls below the VaR.

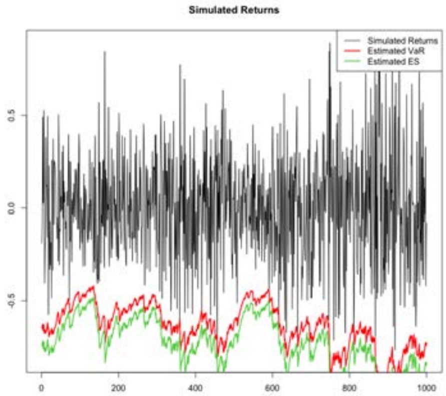


Figure 3.1.2: The simulated GARCH(1,1) returns, the estimated VaR and the estimated ES for n=1,000.

Table 3.1.1 presents the estimated parameters for the simulated GARCH(1,1) model with normally distributed innovations, estimated by maximum likelihood via the ugarchfit function in the R package rugarch. Although the parameter estimates are close to the true values used in the simulation ($\omega = 0.001$, $\alpha = 0.083$, $\beta = 0.909$), small differences emerge due to sample variability and estimation uncertainty. Nonetheless, the sum of parameters α and β remains high, confirming strong volatility persistence consistent with the specified simulation parameters.

Parameter	Estimate	Std. Error
ω	0.0009	0.0005
α	0.0497	0.0126
β	0.9392	0.0154

Table 3.1.1: Parameters estimated for the simulated GARCH(1,1) Normal model

3.1.1 GJR-GARCH(1,1) Model

While the standard GARCH(1,1) model captures volatility clustering, it assumes that positive and negative return shocks have a symmetric impact on future volatility. However, empirical evidence suggests that negative shocks (bad news) often increase volatility more than positive shocks of the same magnitude. This asymmetrical response is commonly referred to as the leverage effect. To account for the leverage effect, Glosten, Jagannathan, and Runkle (1993) introduced the GJR-GARCH model (named for its authors), which modifies the GARCH specification by adding an indicator term that differentiates between negative and positive shocks.

The GJR-GARCH(1,1) conditional variance equation can be written as:

$$\sigma_t^2 = \omega + \alpha\,\epsilon_{t-1}^2 + \gamma\,I_{t-1}\,\epsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2\,, \tag{3.3}$$

where I_{t-1} is an indicator variable that equals 1 if $\epsilon_{t-1} < 0$ (a negative shock) and 0 otherwise. In this formulation, ϵ_{t-1}^2 represents the last period's squared innovation (return shock). The key new term is $\gamma\,I_{t-1}\,\epsilon_{t-1}^2$, which allows the model to respond differently to bad news. If $\gamma > 0$, a negative shock at time $t-1$ contributes $\alpha+\gamma$ (in effect) times the squared shock to the variance at time t, while a positive shock of the same size contributes only α times its squared value. In this way, negative returns have a larger impact on future volatility than positive returns of equal magnitude, directly incorporating the leverage effect into the volatility dynamics (the usual parameter restrictions for non-negativity and stationarity are adjusted for this asymmetric model, for example, one sufficient set of conditions is $\omega > 0$, $\alpha \geq 0$, $\gamma \geq 0$, $\beta \geq 0$, and $\alpha+\beta+\frac{1}{2}\gamma < 1$ under a symmetric innovation distribution) (Glosten et al., 1993).

Fitting the GJR-GARCH(1,1) model to real data presents a few practical challenges. The inclusion of an asymmetry term means that there is an extra parameter to estimate (and additional constraints to ensure that the conditional variance remains positive for all observations), which often necessitates a reasonably large sample size to obtain reliable estimates. If the true leverage effect in the data is weak, introducing this extra term may yield little improvement in model fit and can even risk overfitting. It is also important to specify an appropriate conditional error distribution (e.g., Gaussian vs. Student-t), assuming normally distributed errors in the presence of heavy tails can distort volatility estimates and hypothesis tests. However, when a pronounced leverage effect is present (as is commonly the case for equity returns), the GJR-GARCH(1,1) model can deliver more accurate volatility forecasts and improved risk measure estimates by capturing the typically higher spike in volatility following negative returns.

3.2 Estimation of the VaR

Assuming that the innovations are normal with mean 0 and variance 1, the conditional loss distribution is normal with mean μ_t and variance σ_t^2 and the one-step-ahead value-at-risk calculated at date t-1 with time-varying volatility can then be computed by

$$VaR_t(\alpha) = \mu_t + \Phi^{-1}(\alpha)\sigma_t, \tag{3.4}$$

where Φ is the standard normal distribution function and $\Phi^{-1}(\alpha)$ denotes the α-quantile (namely the left-quantile at α%) of $\mathcal{N}(0,1)$, i.e. the inverse cumulative distribution function (quantile function) of the standard normal distribution at level α (McNeil et al., 2015).

The equation (3.4) of the value at risk for $z_t \sim N(0,1)$ can be proved as follows, given that $r_t = \mu_t + \sigma_t z_t$, $z_t \sim N(0,1)$. By definition, we have :

$$P(r_t > VaR_t(\alpha)|\mathcal{F}_{t-1}) = 1 - \alpha$$

$$P(\mu_t + \sigma_t z_t > VaR_t(\alpha)|\mathcal{F}_{t-1}) = 1 - \alpha$$

$$P(z_t > \frac{VaR_t(\alpha) - \mu_t}{\sigma_t}|\mathcal{F}_{t-1}) = 1 - \alpha$$

As z_t is i.i.d., the conditional probability equals the unconditional probability, namely:

$$P(z_t > \frac{VaR_t(\alpha) - \mu_t}{\sigma_t}) = 1 - \alpha$$

By symmetry of the Gaussian distribution,

$$P(z_t \leq \frac{\mu_t - VaR_t(\alpha)}{\sigma_t}) = 1 - \alpha$$

Given that z_t is standard normally distributed with cumulative distribution function $\Phi(x)$,

$$\Phi(\frac{\mu_t - VaR_t(\alpha)}{\sigma_t}) = 1 - \alpha$$

$$\frac{\mu_t - VaR_t(\alpha)}{\sigma_t} = \Phi^{-1}(1 - \alpha)$$

$$VaR_t(\alpha) = \mu_t - \Phi^{-1}(1 - \alpha)\sigma_t$$

By the symmetry of the Gaussian distribution, $\Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha)$, we obtain the final equation for the value at risk:

$$VaR_t(\alpha) = \mu_t + \Phi^{-1}(\alpha)\sigma_t$$

3.3 Estimation of the ES

To estimate the expected shortfall, we use two different formulas, then backtest the expected shortfall estimators, and select the formula that performs best for the rest of the analysis. The first formula we used for the expected shortfall estimation is the following.

$$ES_t(\alpha) = \mu_t + \sigma_t ES_\alpha(Z), \tag{3.5}$$

where $ES_\alpha(Z)$ is the expected shortfall of the innovations Z that are assumed to be standard normal, and here we take as innovations the standardized residuals of the fitted GARCH(1,1) model (McNeil et al., 2015).

Assuming that the innovations are standard normal, the expected shortfall can be computed as follows if we are looking at the right tail of the distribution for an $\alpha = 99\%$:

$$ES_t(\alpha) = \mu_t + \sigma_t \frac{\phi(\Phi^{-1}(\alpha))}{1 - \alpha}, \tag{3.6}$$

which is the expected shortfall for the Gaussian loss distribution defined by McNeil et al. (2015).

The formula (3.6) can be proved as follows. We first take formula (3.5):

$$ES_t(\alpha) = \mu_t + \sigma_t ES_\alpha(Z)$$

$$ES_t(\alpha) = \mu_t + \sigma_t E[Z|Z > \Phi^{-1}(\alpha)]$$

Note that $E[Z|Z > \Phi^{-1}(\alpha)] = \dfrac{1}{1-\alpha}\int_{\Phi^{-1}(\alpha)}^{\infty} z\phi(z)\,dz$,

where $\phi(z) = \dfrac{1}{\sqrt{2\pi}}e^{-\dfrac{1}{2}z^2}$ and $\phi(z)' = -z\dfrac{1}{\sqrt{2\pi}}e^{-\dfrac{1}{2}z^2} = -z\phi(z)$

From that we get:

$$E[Z|Z > \Phi^{-1}(\alpha)] = \frac{1}{1-\alpha}[-\phi(z)]_{\Phi^{-1}(\alpha)}^{\infty}$$

$$E[Z|Z > \Phi^{-1}(\alpha)] = \frac{1}{1-\alpha}\phi(\Phi^{-1}(\alpha))$$

since at ∞ the density is 0.

Therefore, we finally get the formula for losses (right tail):

$$ES_t^L(\alpha) = \mu_t^L + \sigma_t^L \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}$$

However, here we are looking at the returns, and so we are interested in the formula for $\alpha = 1\%$, that is, the left tail of the returns distribution. The losses are defined as $L_t := -R_t$, which means that the conditional mean for the losses $\mu_t^L = -\mu_t^R$ and for an $\alpha = 99\%$ the ES for the losses $ES_t^L(\alpha) = -ES_t^R(1-\alpha)$. Note that, by the symmetry of the normal distribution, $\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$, so we get the following formula for estimating the expected shortfall for returns (left tail):

$$ES_t^R(\alpha) = \mu_t^R - \sigma_t^R \frac{\phi(\Phi^{-1}(1-\alpha))}{\alpha} \tag{3.7}$$

$$ES_t^R(\alpha) = \mu_t^R - \sigma_t^R \frac{\phi(\Phi^{-1}(\alpha))}{\alpha}$$

(McNeil et al., 2015).

The second formula we use to estimate the expected shortfall is given by equation (3.7) to which we compare equation (3.5). The ES formula that gives the best results in terms of the one that provides the lowest proportion of p-values of the violation residuals is the formula (3.7), so it is the one we keep for the rest of the analysis.

3.4 Estimation of the expectile

The expectile forecast at time t is conditional on the dynamics of the simulated GARCH(1,1) data. The conditional expectile is defined by :

$$e_t = \mu_t + \sigma_t e_\epsilon, \tag{3.8}$$

where μ_t represents the conditional mean and σ_t represents the conditional volatility, which are both estimated using the fitted GARCH(1,1) model, and e_ϵ is the standard expectile at level α for a normally distributed innovation sequence ϵ_t, approximated by a Monte Carlo simulation of 10 000 $\mathcal{N}(0,1)$ draws (Armando et al., 2024).

The numerical estimation of the expectile e_t is done by a numerical optimization. The asymmetric least-squares loss function given by equation (3.9) is minimized using the optimize function in R. The function searches for the value of the expectile e_t that balances the weighted residuals above and below the expectile. The conditional expectile forecast e_t is dynamically updated for each time step, based on the estimated GARCH parameters. The asymmetric least-squares loss function from equation (3.9) can be alternatively expressed as :

$$S(\hat{e}, y) = \alpha \sum_{y \geq \hat{e}} (y - \hat{e})^2 + (1 - \alpha) \sum_{y < e} (y - \hat{e})^2 \tag{3.9}$$

The simulated GARCH(1,1) returns and the estimated expectile for n=1,000, 10,000, and 100,000 are illustrated in Figure 3.4.1.

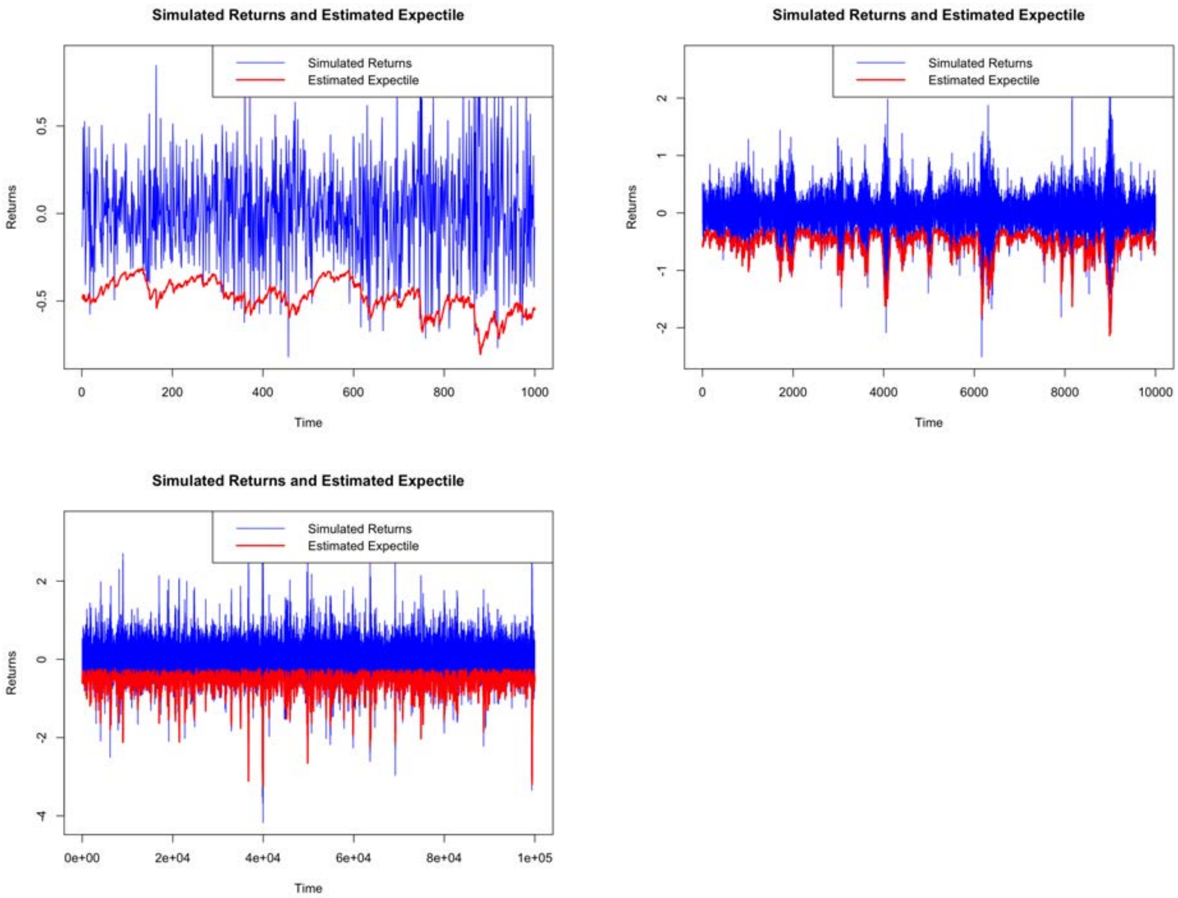


Figure 3.4.1: The simulated GARCH(1,1) returns and the estimated expectile for n=1,000, 10,000 and 100,000.

3.5 Comparison of Risk Measures

In this section, we compare risk forecasts obtained from the dynamic model using three different risk measures, Value-at-Risk (VaR), Expected Shortfall (ES), and Expectiles, all evaluated at the 1% risk level. The comparison is based on a simulated GARCH(1,1) dataset, which allows us to illustrate how these measures behave under identical conditions. We begin by describing the theoretical relationships between VaR, ES, and expectiles at the same confidence level. We then verify these relationships using our simulation results.

VaR and ES are standard quantile-based risk measures. $\mathrm{VaR}_t(1\%)$ is the 1^{st} percentile of the loss distribution (the loss threshold that is exceeded only 1% of the time), while $\mathrm{ES}_t(1\%)$ is the average loss in the worst 1% of cases (i.e. the expected loss conditional on being in the 1% tail beyond the VaR level). The expectile at the 1% level is an alternative risk measure defined via an asymmetric least-squares criterion. Intuitively, a 1% expectile places a high weight on lower-tail outcomes, producing a risk threshold that targets extreme losses but in a different way than the quantile-based VaR.

Theoretical ordering results establish that, for a given tail probability such as 1%, ES is the most extreme (most conservative) of these measures. In particular, one can show that $|ES_t(1\%)| \geq |\text{VaR}_t(1\%)|$ by definition, reflecting the fact that ES averages over the worst losses beyond the VaR quantile, and that $|ES_t(1\%)| \geq |\text{Expectile}_t(1\%)|$ as well as explained by Tadese and Drapeau (2020). On the other hand, there is no general guarantee that the 1% expectile will exceed (or even equal) the 1% VaR, the ordering between expectiles and quantiles can vary with the shape of the distribution (Koenker, 1993). In many practical situations (especially with fat-tailed data), the expectile at a given nominal level tends to be less extreme than the corresponding VaR, meaning that it implies a lower loss level than the VaR at the same probability (Daouia et al., 2018). However, in very heavy-tailed regimes, expectiles may even exceed the corresponding VaR as shown by Bellini et al. (2014).

The simulation evidence aligns with these theoretical expectations. Using 1,000 observations generated from a GARCH(1,1) process, we computed one-step-ahead forecasts of the 1% VaR, 1% ES and 1% expectile at each time step. The results show that ES is consistently the largest (most conservative) risk estimate among the three. In our simulation, there were zero instances (0 out of 1,000) in which the ES forecast was less conservative (i.e. indicated a smaller loss) than the corresponding VaR or expectile forecast, confirming that $|ES_t(1\%)| \geq |\text{VaR}_t(1\%)|$ and $|ES_t(1\%)| \geq |\text{Expectile}_t(1\%)|$ in every observation (no violations).

Moreover, the expectile forecasts were uniformly less extreme than the VaR forecasts. In all 1,000 observations, the 1% expectile yielded a higher loss threshold (closer to zero) than the 1% VaR. Thus, we did not observe cases where the expectile fell below (was more extreme than) the VaR at the same 1% level. These empirical findings underscore that, in our GARCH simulation, the expectile was consistently the least conservative measure, whereas the ES was the most conservative.

Figure 3.5.1 provides a visual illustration of the differences between the three risk measures for a subset of the data. The figure plots the 1% VaR, ES, and expectile forecasts for the first 100 simulated observations (for clarity, a zoomed-in portion of the series). As expected, the ES series (green line) is consistently the lowest line on the graph (most negative values), indicating the largest anticipated losses. The VaR series (red line) lies in the middle, and the expectile series (blue line) remains closest to zero (least negative). There are no intersections between these lines during the shown period, reflecting a stable ordering at each time point. Specifically, $ES_t(1\%)$ provides the furthest-tail (most conservative) estimate, $VaR_t(1\%)$ is slightly less extreme, and the expectile is the most moderate. This visualization reinforces our earlier statements: ES remains the most conservative risk measure throughout, VaR is intermediate, and the expectile offers the mildest risk assessment among the three. Note that we obtain similar plots and results for real data when we compare the risk measures.

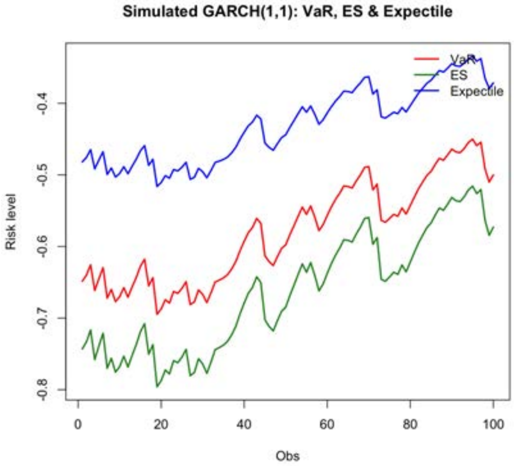


Figure 3.5.1: Comparison of risk measure forecasts at the 1% level for the first 100 simulated GARCH(1,1) observations. The ES (green) is the lowest (most conservative) at each time point, the VaR (red) is in the middle, and the expectile (blue) is closest to zero (least conservative). This ordering holds consistently, illustrating that $|ES| \geq |VaR|$ and $|ES| \geq |Expectile|$ at all times in the simulation, while the expectile never falls below the VaR.

3.6 Conclusion

This chapter presents an analysis of dynamic risk modeling and highlights the comparative behavior of different risk measures. We employ a GARCH(1,1) volatility model to generate time-varying forecasts for VaR, ES, and expectiles at a 1% risk level and examine their performance. We also define the GJR-GARCH(1,1) volatility model, which will be useful for our analysis of real data because it captures the leverage effect. A key insight from our study is that the choice of risk measure significantly influences the estimated level of risk. Specifically, ES consistently produced more conservative (higher in absolute value) loss estimates than VaR, and VaR, in turn, was generally more conservative than the expectile. This confirms the theoretical ranking of the risk measures and illustrates it in a controlled simulation setting, and the results of the order of the risk measures do not change when we look at real data. The results also emphasize the importance of properly accounting for tail risk. A measure like ES, which captures the severity of extreme losses, can provide a very different risk perspective than the expectile. In our study, the expectile clearly is lower in absolute value than the VaR and ES. Overall, the chapter's findings underscore that using dynamic models together with a tail risk measure offers a robust framework for evaluating extreme financial risks. They also highlight the need for careful selection of risk measures in practice to ensure adequate risk assessment.

Chapter 4

Backtesting Methods

4.1 Backtesting VaR

The quality of the value at risk estimation is evaluated by backtesting methods. The VaR-violations or VaR-exceedances must satisfy two properties for the model to be considered accurate, namely the VaR-exceedances divided by the observations must equal the VaR-level and the VaR-exceedances must be independent from one another, so they should not depend on the past.

Several classical backtesting methods exist to evaluate the accuracy of VaR models, among which are the Kupiec (1995) Proportion of Failures (POF) test and the Christoffersen (1998) independence test. The Kupiec test specifically evaluates whether the frequency of observed VaR violations aligns with the expected frequency implied by the chosen VaR confidence level. However, it does not test whether these violations occur independently over time. On the other hand, the Christoffersen test explicitly checks for the independence of exceedances, ensuring that violations are not clustering in time. However, it does not directly assess whether the total number of exceedances matches the expected number.

We chose the Dynamic Quantile (DQ) test by Engle and Manganelli (2004) precisely because it addresses both aspects simultaneously. It jointly evaluates whether the frequency of exceedances corresponds to the chosen VaR level and whether these exceedances are independently distributed over time. Thus, the DQ test provides a more comprehensive assessment in a single unified framework, making it particularly suitable for validating dynamic VaR models. In this section, we present the Dynamic Quantile (DQ) test as introduced by Engle and Manganelli (2004). All the formulas and derivations below are adapted from their work.

The backtesting method for the value at risk is the Dynamic Quantile (DQ) test, which tests both the property of the value at risk level and the property of independence of the value at risk exceedances. We define the variable Hit for α close to 0 (long positions).

$$Hit_t(\alpha) = I\{r_t < VaR_t(\alpha)\} - \alpha, \tag{4.1}$$

and given the linear regression

$$Hit_t(\alpha) = \beta_0 + \sum_{i=1}^{p} \beta_i Hit_{t-i}(\alpha) + \epsilon_t, \tag{4.2}$$

and the Dynamic Quantile test tests the null hypothesis $H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0$ where p is set equal to 5.[1] Assuming that the null hypothesis holds, the statistic of the DQ test, denoted by DQ_α, converges in distribution to a chi-square variable with $p+1$ degrees of freedom. In particular, we have

$$DQ_\alpha = \frac{\text{Hit}(\alpha)' X (X'X)^{-1} X' \text{Hit}(\alpha)}{\alpha(1-\alpha)} \xrightarrow{d} \chi^2_{p+1} \quad \text{as} \quad T \to \infty.$$

In this expression, the vector

$$\text{Hit}(\alpha) = \Big(\text{Hit}_1(\alpha), \ldots, \text{Hit}_T(\alpha)\Big)'$$

contains the hit indicators, and X is the $T \times (p+1)$ matrix whose t-th row is composed of a constant term along with the p lagged variables related to $\text{Hit}_t(\alpha)$.

Table 4.1.1 reports the DQ test statistics and corresponding p-values for three sample sizes (n = 1,000, 10,000 and 100,000). Although the test statistics grow larger as n increases (2.6352 → 4.4302 → 17.2081), only at n = 100,000 does the p-value (0.028) fall below the conventional significance level of 5%, indicating that for small to moderate samples (n = 1,000 or 10,000), the null hypothesis is not rejected, but for n = 100,000 the DQ test suggests a significant misspecification.

Number of observations (n)	Statistic	P-value
1,000	2.6352	0.9551
10,000	4.4302	0.8164
100,000	17.2081	0.028

Table 4.1.1: The results of the DQ test.

4.2 Backtesting ES

The first backtesting method we investigate for the expected shortfall is a violation-based test by checking if the violation residuals different than zero have a mean equal to 0. Later, we also have to test whether those violation residuals are independent.

The violation residual is defined by

$$K_t = \left(\frac{L_t - ES_\alpha^{L,t}}{ES_\alpha^{L,t} - \mu_{L,t}} \right) I_{\{L_t > VaR_\alpha^t\}}, \tag{4.3}$$

[1]We choose a lag value of five because it is the usual choice in empirical literature.

where the violation residual is computed by the product of an indicator function that equals one once there is a VaR-violation, i.e. when the loss is greater than the value at risk or equivalently when the return is lower than the VaR at time t, and the difference between the losses and the respective expected shortfall, and there is further standardization for theoretical reasons, where we divide by the difference between the expected shortfall and the conditional mean (set equal to 0 here) (McNeil et al., 2015).

Since in this analysis we look at the returns' distribution, we transform equation (4.3) for the returns R_t by replacing $L_t := -R_t$. If the sample mean for the returns R is:

$$\hat{\mu}_R = \frac{1}{N}\sum_{t=1}^{N} R_t,$$

then the sample mean for the loss L is literally the negative of the sample mean of R:

$$\hat{\mu}_L = \frac{1}{N}\sum_{t=1}^{N} L_t = \frac{1}{N}\sum_{t=1}^{N}(-R_t) = -\hat{\mu}_R.$$

Using a similar line of reasoning, the expected shortfall for the loss L at level α is literally the negative of the expected shortfall for the returns R at level $(1-\alpha)$, i.e. $ES^L(\alpha) = -ES^R(1-\alpha)$. And the value at risk for the loss L is literally the negative of the value at risk for the returns R, i.e. $VaR^L = -VaR^R$.

Therefore, the violation residual that we estimate can be obtained from equation (4.3) as follows:

$$K_t = \left(\frac{L_t - ES_\alpha^{L,t}}{ES_\alpha^{L,t} - \mu_{L,t}}\right) I_{\{L_t > VaR_\alpha^t\}}$$

$$K_t = \left(\frac{-R_t - (-ES_\alpha^{R,t})}{-ES_\alpha^{R,t} - (-\mu_{R,t})}\right) I_{\{-R_t > -VaR_\alpha^{R,t}\}}$$

$$K_t = \left(\frac{-R_t + ES_\alpha^{R,t}}{-ES_\alpha^{R,t} + \mu_{R,t}}\right) I_{\{R_t < VaR_\alpha^{R,t}\}}$$

We finally get the formula used to estimate the violation residuals for the expected shortfall:

$$K_t = \left(\frac{R_t - ES_\alpha^{R,t}}{ES_\alpha^{R,t} - \mu_{R,t}}\right) I_{\{R_t < VaR_\alpha^{R,t}\}} \tag{4.4}$$

Once we have estimated the model, we obtain the estimated mean $\hat{\mu_t}$, which was set equal to 0, and the estimated volatility $\hat{\sigma_t}$. We can calculate the expected shortfall of the estimated standardized residuals, which represents the innovations Z, and thus obtain the expected shortfall (ES) of the return series r_t, resulting in a series ES_α^t, a measure of the expected shortfall for each day, computed by Equation (3.5). With this, we can construct the violation residuals, or the residuals of violations K_t. The violation residual is computed by the product of an indicator function that equals one once there is a VaR-violation, and the difference between the losses and the respective expected shortfall divided by the difference between the expected shortfall and the conditional mean (set equal to 0 here). The key point is that this residual should, on average, be equal to zero.

Suppose that we have a loss distribution in plot 4.2.1 and we look at the right tail of the loss distribution, we have the value at risk (denoted VaR). We are considering the part of the distribution greater than the value at risk, so to the right of VaR, and by definition the expected shortfall (ES) is to the right of VaR, which must be greater than or equal to the VaR. This is our estimation, we have data points between the VaR and the ES, and other points larger than the ES. We simply want to construct the difference between the losses and the expected shortfall, standardize it, and see if this series K_t has an average of zero. Otherwise, if it is different, if it is positive, for example, it means that we are underestimating the true expected shortfall. If it is negative, it means that we are overestimating the true expected shortfall, and thus it would result in a backtest that does not pass.

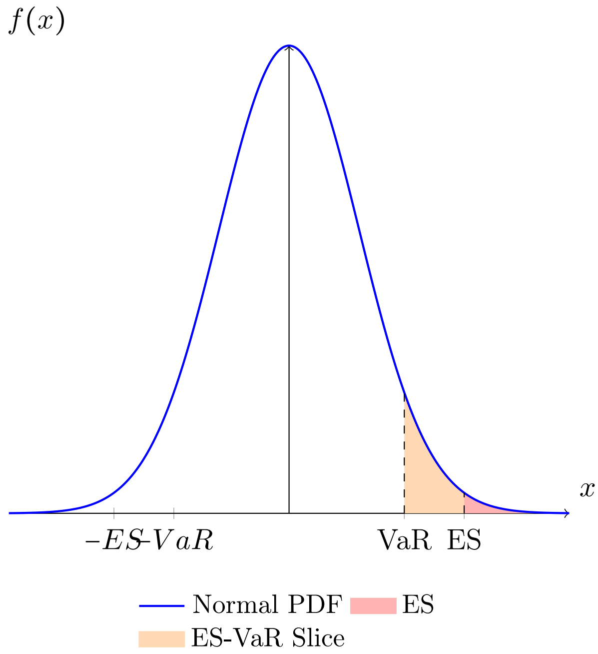


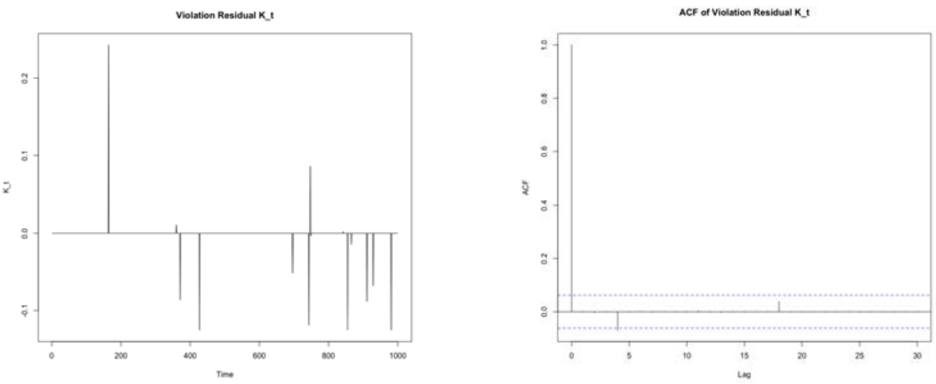
Figure 4.2.1: Example of a normal loss density function, with the VaR and the ES on the right tail.

To test whether the mean of the violation residuals equals zero, we first compute the series $K_t : K_1, ..., K_T$, that is our sample. Then we can simply calculate the mean $\bar{K} = \frac{1}{T}\sum_t K_t$ and perform a t-test, which means: simply calculate the ratio between the mean and its standard deviation, the whole multiplied by $\sqrt{T}$, i.e. the t-test statistic:

$$\frac{\bar{K}}{\sqrt{\hat{Var}(K_t)/T}} = \sqrt{T}\frac{\bar{K}}{\sqrt{\hat{Var}(K_t)}} \sim N(0,1), \tag{4.5}$$

where the variance is the empirical variance, that is, we just calculate the variance of the series K_t. This is simply the t-test statistic, which should be under the null hypothesis H_0 standard normally distributed. The formal null hypothesis of the t-test is $H_0 : E(K_t) = 0$, namely under the null hypothesis the average of the violation residuals K_t is zero.

Note that there is an indicator function, so many of these K_t values are zero, which means that this applies only when K_t is nonzero, positive or negative, but not zero. We do not look at cases where we are below the value at risk, where we do not have an extreme situation, we only consider the cases where the loss exceeds the value at risk. So, for these cases, we get the series of K_t values that are non-zero, positive or negative, and then we take the mean, standardize it with the estimated standard deviation divided by $\sqrt{T}$.



(a) The violation residual K_t *(b) The ACF of the violation residual K_t*

Figure 4.2.2: (a) The violation residual K_t. (b) The ACF of the violation residual K_t.

The results for the 1,000 simulated returns give a mean of the violation residuals of -0.0385, a variance of the violation residuals of 0.0043, 6 violations, a t-statistic of -1.4359 and a p-value of 0.2105. Since the p-value is larger than the α of 5%, we fail to reject the null hypothesis, stating that the average of the violation residuals K_t is zero. Table 4.2.1 shows the results of the t-test on the violation residuals K_t for sample sizes 1,000, 10,000 and 100,000, and for each of the sample sizes we fail to reject the null hypothesis stating that the average of the violation residuals K_t is zero.

Number of observations (n)	T-Statistic	P-value
1,000	-1.4359	0.2105
10,000	-0.5456	0.5866
100,000	-0.7469	0.4553

Table 4.2.1: The results of the t-test on the violation residuals K_t.

To test whether the violation residuals are independent, there are no tests done in the literature, so we first look at the autocorrelation function (ACF) of K_t depicted in Figure 4.2.2b. In this plot, all autocorrelation values in lags from 1 onward seem to lie within the 95% confidence bands depicted by the dashed blue lines. In other words, none of the autocorrelations at positive lags is statistically significant, namely, there is no evidence of significant autocorrelation in the violation residuals at any lag beyond zero.

We then do a Portmanteau test (Box-Pierce) on the violation residuals. The test statistic of the Portmanteau test is

$$Q(K) = T \sum_{i=1}^{K} \hat{\rho}_i^2, \tag{4.6}$$

with asymptotic distribution χ^2_{K-2}, the null hypothesis is rejected at level $\alpha = 5\%$ if $Q(K) > \chi^2_{K-2;1-\alpha}$ so if $Q(K) > \chi^2_{K-2;95\%}$, (McNeil et al., 2015). The test statistic is Q(5)=6.3986 for 5 lags[2] and the p-value is equal to 0.2693. The results show that the Portmanteau test did not find any significant autocorrelation in the violation residuals up to lag 5. That is, the p-value (0.2693) is much higher than the significance level of 5%, we fail to reject the null hypothesis, so there is no evidence of significant autocorrelation in our residuals at these lags.

4.2.1 One-sided test for risk underestimation of the ES by Acerbi & Szekely and Spring

In this section, we reproduce a backtest similar to the backtest 2 by Acerbi and Szekely (2014) that was also explained in detail by Spring (2021). The purpose of this one-sided backtest is to check whether the estimated expected shortfall underestimates the real risk represented by the real expected shortfall. However, we perform the two-sided backtest to verify whether the estimated expected shortfall underestimates or overestimates the real risk.

The test statistic Z is first computed as follows:

$$Z(R) = \sum_{t=1}^{T} \frac{R_t I_t}{T \alpha ES_{\alpha,t}} - 1, \tag{4.7}$$

where the indicator function $I_t = I_{\{R_t < VaR_t\}}$ is 1 for returns that are lower than the value at risk, and for the left tail of the returns distributions, for $\alpha = 1\%$, the ES is calculated as $ES_t(\alpha) = \mu_t - \sigma_t \dfrac{\phi(\Phi^{-1}(1-\alpha))}{\alpha}$ (Acerbi and Szekely, 2014).

Spring (2021) uses a bootstrap procedure to estimate the p-value. To reproduce it, we simulate the distribution of the statistic Z under the null hypothesis $H_0 : Z = 0$ to get then the p-value of the backtest. We start by simulating m independent bootstrap trials by first resampling the estimated returns with replacement. Then, we fit a GARCH(1,1) model to the bootstrap sample, extract the fitted μ and σ for each bootstrap sample, and estimate the VaR and ES for each bootstrap sample. We then compute the Z statistic for each bootstrap sample, denoted Z^i and finally estimate the p-value as follows:

[2] Five lags are used instead of 10 lags because there are only 6 violations and so 6 observations in the violations vector K_t. But for the other Portmanteau tests in this thesis a lag of 10 is used. We should not choose a lag that is too small because we could risk missing important autocorrelations, but if we choose a lag too large, the test loses power.

$$p = \frac{1}{m} \sum_{j=1}^{m} I_{Z^i > Z(R)}, \tag{4.8}$$

(Spring, 2021).

We initially simulate 10,000 returns observations from a GARCH(1,1) process, the test statistic Z is -0.2354 and for m=1,000 bootstrap trials the bootstrap p-value equals 0.981. So we fail to reject the null hypothesis, i.e. the ES model adequately captures tail risk.

4.3 Backtesting Expectile

Backtesting of the expectile is performed to evaluate whether the estimated expectile satisfies certain statistical properties.

The two main properties of a well-defined expectile forecast are:

- Unconditional Coverage (UC): The expectile captures the correct proportion of extreme observations.
- Independence (IND): The identification process of the expectile does not exhibit autocorrelation.

We separately test the Unconditional Coverage and Independence conditions.

The Unconditional Coverage (UC) Test evaluates the null hypothesis H_0 :

$$E[I(\hat{e}_t, R_t)] = 0 \tag{4.9}$$

and the Wald test is used to assess the Unconditional Coverage condition for the expectile with the following test statistic:

$$\frac{(\sum_{t=1}^{T} I(\hat{e}_t, R_t))^2}{T\hat{\sigma}^2} \sim \chi_1^2 \text{ as } T \to \infty, \tag{4.10}$$

where T is the sample size, $\hat{\sigma}^2$ is the sample variance of the sequence $(I(\hat{e}_t, R_t))$ and the p-value is calculated under a chi-squared distribution with 1 degree of freedom, i.e. meaning that under H_0, the Wald test statistic converges asymptotically to the chi-squared distribution with 1 degree of freedom when the sample size T increases to infinity (Armando et al., 2024).

For the Independence (IND) Tests, the IND property is tested using three transformed processes derived from the identification function:

- $C_1(t)$: the expectile exceedance process. This transformed process is similar to the value at risk exceedances used for the DQ test defined in equations (4.1) and (4.2). We define the binary indicator sequence which equals 1 whenever the returns is lower than the estimated expectile denoted by $C_1(t)$ and defined as follows:

$$C_1(t) = 1_{R_t < \hat{e}_t}, \tag{4.11}$$

$(C_1(t))$ is called the expectile exceedance process. Under H_0, the expectile exceedance process equals the binary indicator sequence $(1_{(\epsilon_t < \hat{e}_\epsilon)})$, which equals 1 whenever the innovation sequence ϵ_t is lower than its expectile e_ϵ, and since (ϵ_t) is i.i.d. so is $(C_1(t))$ under H_0 (Armando et al., 2024).

- $C_2(t)$: Normalized identification process:

$$C_2(t) = \frac{I(\hat{e}_t, R_t)}{\hat{\sigma}_t}, \tag{4.12}$$

where we normalize the identification process by $\hat{\sigma}_t$.

- $C_3(t)$: Normalized squared excess return:

$$C_3(t) = \frac{(R_t - \hat{e}_t)^2}{\hat{\sigma}_t^2}, \tag{4.13}$$

where we take the squared gap between the actual return R_t and its expectile $\hat{e}_t$, and then scale by the variance estimate.

The Box-Pierce/ Portmanteau test defined in (4.6) is used to check for autocorrelation in each process.

The implementation of the expectile backtest is performed by first computing the identification function of the expectile $I(\hat{e}_t, R_t)$ for the UC test and by calculating the Wald statistic (4.10). Then, we compute the three transformed processes derived from the identification function, i.e. $C_1(t)$, $C_2(t)$ and $C_3(t)$, for the IND tests. And finally, we perform Box-Pierce tests on $C_1(t)$, $C_2(t)$ and $C_3(t)$ to evaluate independence.

The results of the Unconditional Coverage Test using the Wald statistic are summarized in Table 4.3.1. For a sample size of 1,000 observations, the test statistic is 2.74 with a corresponding p-value of approximately 0.098, which exceeds the conventional 5% significance level. Hence, we do not reject the null hypothesis at the 5% level. For larger samples, i.e. 10,000 and 100,000 observations, the Wald statistics are 0.017 and 0.509 with p-values of 0.896 and 0.476, respectively. These high p-values indicate that as the sample size increases, the test provides strong evidence that the model's risk forecasts satisfy the unconditional coverage condition. Figure 4.3.1 illustrates the ACF of the processes C_1, C_2 and C_3 for n=1,000.

Number of observations (n)	Wald Statistic	P-value
1,000	2.7397	0.0979
10,000	0.0171	0.8961
100,000	0.5088	0.4757

Table 4.3.1: The results of the Unconditional Coverage Test (Wald Test) for n=1,000, 10,000 and 100,000 simulated returns.

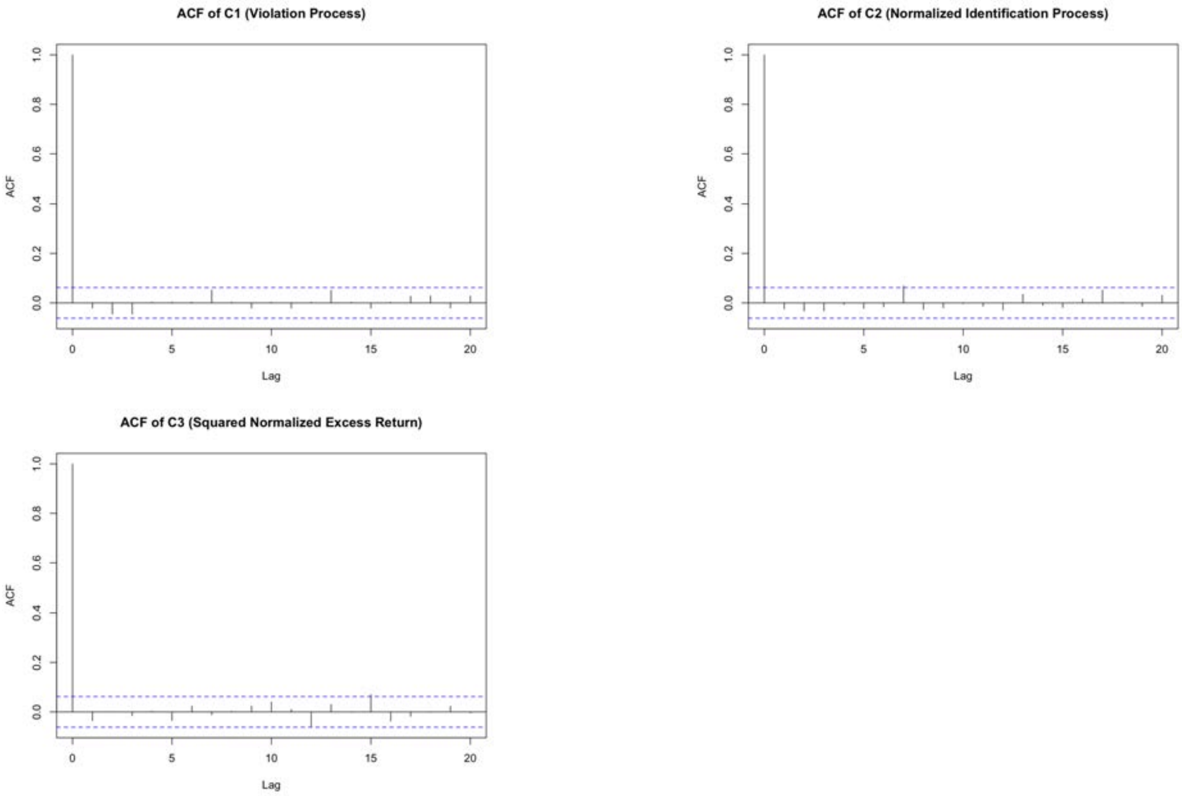


Figure 4.3.1: The ACF of the processes C_1, C_2 and C_3 for n=1,000.

Table 4.3.2 reports the results of the Box-Pierce independence tests for three candidate processes used in the expectile backtesting framework for $n = 1,000$, 10,000, and 100,000. For each process, the test evaluates the null hypothesis that the corresponding time series is serially uncorrelated, namely that the process is i.i.d.. A high p-value indicates that we cannot reject the null hypothesis, whereas a low p-value, i.e. below 5% would show evidence against independence. For the process $C_1(t)$ (the expectile exceedance process), the p-values are relatively high across all sample sizes (0.66, 0.12, and 0.25 for $n = 1,000$, 10 000, and 100 000, respectively), which indicates that there is no statistically significant autocorrelation. Similarly, $C_2(t)$ (the normalized identification process) has p-values of 0.49, 0.06, and 0.37, respectively, again providing no evidence to reject the null hypothesis of independence. In contrast, for the process $C_3(t)$ (the normalized squared excess return), while the p-values are high for $n = 1,000$ and $n = 10,000$ (0.85 and 0.82), the p-value drops to 0.001 for $n = 100,000$. This significant result at the 5% level suggests that when a very large sample is considered, the normalized squared excess return process exhibits statistically significant autocorrelation. This finding may indicate that the assumption of independence for $C_3(t)$ is questionable in large samples, which could be a result of the underlying dynamics of the risk measure or the sensitivity of the squared transformation to small deviations. Overall, the results suggest that while the expectile exceedance process $C_1(t)$ and the normalized identification process $C_2(t)$ satisfy the independence assumption in a range of sample sizes, caution should be exercised when using the normalized squared excess return $C_3(t)$, particularly with very large datasets.

Number of observations (n)	Independence Test Statistic	P-value
$C_1(t)$: the expectile exceedance process		
1,000	7.6844	0.6596
10,000	15.3492	0.1198
100,000	12.4888	0.2537
$C_2(t)$: Normalized identification process		
1,000	9.4057	0.4941
10,000	17.6376	0.0614
100,000	10.7998	0.3733
$C_3(t)$: Normalized squared excess return		
1,000	5.6329	0.8451
10,000	5.9387	0.8204
100,000	29.4613	0.001

Table 4.3.2: The results of the Independence Box-Pierce Tests for the processes C_1, C_2 and C_3 for n=1,000, 10,000 and 100,000 simulated returns.

4.4 Bootstrap of GARCH(1,1) Simulations

In this section, we investigate further whether the value at risk, the expected shortfall and the expectile are estimated correctly by verifying if the backtests for the VaR, ES and expectile are performing well using a bootstrap of 1,000 GARCH(1,1) simulated returns. In particular, we assess the finite-sample performance of the Value at Risk (VaR), the Expected Shortfall (ES) and the expectile at level $\alpha = 1\%$ by performing a Monte Carlo bootstrap using GARCH(1,1) simulations. We generate $1,000$-length return series under a fixed GARCH(1,1) specification and repeat this procedure for B bootstrap iterations, where $B \in \{100, 1,000, 10,000\}$. For each simulated path, we compute the corresponding backtest p-values:

- VaR backtest: Dynamic Quantile (DQ) test.
- ES backtest: P-value of the t-test on the violation residuals K_t.
- Expectile backtests:
  - Unconditional Coverage (UC) Wald test on the identification function.
  - Independence tests (Box–Pierce) on the transformed processes C_1, C_2, and C_3.

We then record the proportion of p-values below the 5% significance level for each test across B iterations and the proportion should be smaller than 5%. Table 4.4.1 displays these proportions (in percent) for the three choices of B.

Backtest	$B = 100$	$B = 1,000$	$B = 10,000$
VaR: DQ test	6%	8.7%	8.87%
ES: K_t statistic	15%	8.9%	8.58%
Expectile: UC test	5%	3.7%	4.01%
Expectile: BP(C_1)	9%	6.1%	5.63%
Expectile: BP(C_2)	13%	9.9%	10.03%
Expectile: BP(C_3)	5%	4.3%	4.68%

Table 4.4.1: Proportion of rejections (%) at 5% nominal level for different bootstrap sizes B for 1,000 simulated returns.

For $B = 100$, the VaR test rejects 6% of the time and the ES test rejects 15% (both above the nominal 5%, especially ES). As B increases to 1,000 and 10,000, the ES rejection rate becomes 8.9% and 8.58%, respectively, while the VaR rejection remains elevated at 8.7% and 8.87%. The expectile UC test is exactly 5% for $B = 100$ but becomes conservative for $B = 1,000$ (3.7%) and $B = 10,000$ (4.01%). Among the independence tests, C_1 drops from 9% for $B = 100$ to about 5.63% for $B = 10,000$, C_2 remains over-rejective (13% $\rightarrow$ 10.03%), and C_3 is 5% for $B = 100$ then slightly conservative (4.3%, 4.68%). These results appear in Table 4.4.1.

We then look at a bootstrap of 10,000 GARCH(1,1) simulated returns, and we simulate 10,000 GARCH(1,1) returns each having 10,000 observations, and the results are given in Table 4.4.2. The proportion of p-values of the DQ test that are smaller than 5% is 5.39% and the proportion of p-values of the violation residuals K_t t-test which are smaller than 5% is 5.22%. The value at risk and the expected shortfall are better estimated once we simulate more observations (10,000 instead of 1,000) and do a bootstrap with more iterations (10,000 instead of 1,000) since the proportion of p-values smaller than 5% is close to 5% for both the DQ test and the violation residuals t-test. A similar phenomenon is observed for the expecile backtests, since the proportion of p-values of the expectile UC test which are smaller than 5% is 2.61% and the independence backtests also improve in general.

Backtest	$B = 10,000$
VaR: DQ test	5.39%
ES: K_t statistic	5.22 %
Expectile: UC test	2.61%
Expectile: BP(C_1)	4.69%
Expectile: BP(C_2)	6.56 %
Expectile: BP(C_3)	4.73%

Table 4.4.2: Proportion of rejections (%) at 5% nominal level for different bootstrap size B=10,000 for 10,000 simulated returns.

4.5 Conclusion

In this chapter, we outline and implement a range of backtesting procedures, namely the Dynamic Quantile test for VaR, the t-test on ES violation residuals, and the Unconditional Coverage and independence tests for expectiles, before presenting our empirical findings. In general, the Dynamic Quantile test validates both coverage and independence for VaR except for very large samples (n=100,000), while the ES backtests confirm that violation residuals have mean zero and are uncorrelated for practical sample sizes. The UC tests for the expectiles show that the expectile captures the correct proportion of extreme observations. For the independence tests on the expectiles, independence holds consistently for C_1 and C_2, but C_3 shows autocorrelation only in very large samples (n=100,000). Bootstrap experiments demonstrate that larger series lengths and more resamples bring rejection rates closer to the nominal 5% level for all methods.

Chapter 5

Empirical Application to S&P 500 Returns

5.1 Introduction

In this chapter, we estimate the different risk measures and apply the backtesting methods discussed in the previous chapters to real-world data, specifically focusing on the S&P 500 index returns. The S&P 500 represents a comprehensive and widely followed benchmark for the U.S. equity market, making it a pertinent choice for our application of the Value-at-Risk (VaR), Expected Shortfall (ES), and Expectile risk measures. By analyzing these measures on actual financial data, we aim to evaluate their effectiveness, compare their performance, and gain insight into their practical implications in risk management.

5.2 Data Definition and Preprocessing

Daily adjusted closing prices of the S&P 500 index are retrieved from Yahoo Finance (2025) for the period from February 22, 2005 to February 20, 2025. This long sample period includes various market regimes, providing a robust basis for analyzing extreme losses using a GARCH(1,1) framework. The focus here is on the empirical estimation of Value at Risk (VaR), Expected Shortfall (ES), and Expectile, along with comprehensive backtests.

The plot of the S&P 500 index from February 22, 2005 to February 20, 2025 and the respective real returns are shown in Figure 5.2.1. We notice a boom before the financial crisis and a significant decline in the S&P 500 index during the 2008 financial crisis. After the 2008 financial crisis, the index recovers and continues on a generally upward trend. A significant decline is observed in 2020 due to the COVID-19 pandemic, the S&P500 experienced one of its fastest market declines in history, dropping more than 30% in a few weeks. The S&P 500 index recovers and shows an upward trend. In 2022, the S&P 500 fell sharply as persistently high inflation forced the Federal Reserve to tighten monetary policy, and increased economic uncertainty made investors more cautious. Finally, the S&P 500 index recovers and increases.

Next, we transform these index levels into log returns using the formula:

$$r_t = \log(S_t) - \log(S_{t-1}) \tag{5.1}$$

where S_t is the index level on day t. Figure 5.2.1 displays these daily log returns. Although most of the returns are close to zero, there are clear spikes in volatility during periods of market stress. In particular, the 2008 financial crisis and early 2020 exhibit higher and more persistent volatility, whereas in more stable times, the fluctuations are relatively low. These large volatilities during crises underline the sensitivity of the market to external shocks, although overall the index tends to recover its upward trajectory over time.

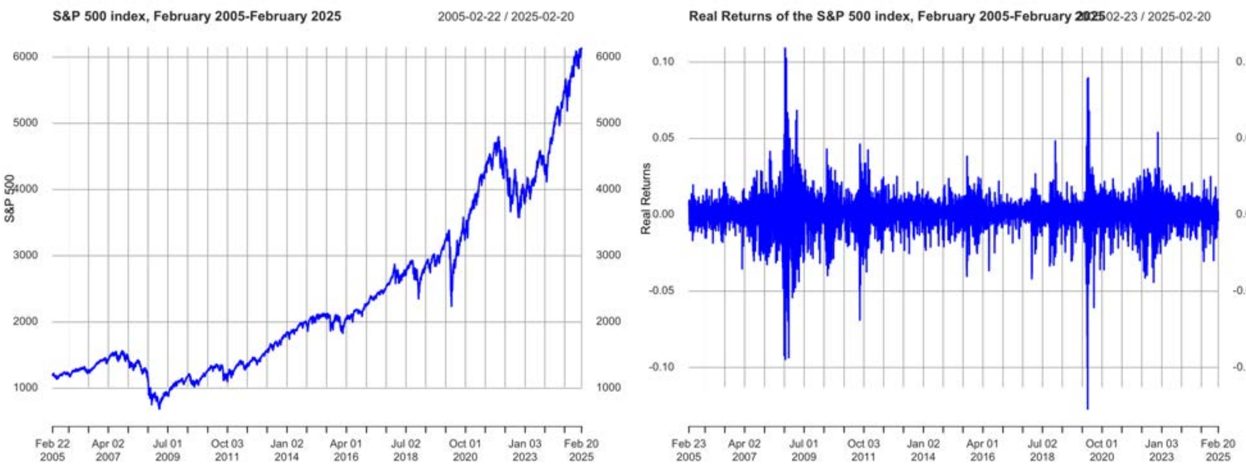


Figure 5.2.1: The S&P 500 index from February 22, 2005 to February 20, 2025 and the respective real returns.

An Augmented Dickey-Fuller test is performed on the first difference of the log of the S&P 500 and the statistic is -17.344 and the p-value equals 0.01 for a lag order of 17. Therefore, we reject the null hypothesis that the S&P 500 returns has a unit root, i.e. the return series is stationary.

5.3 Model Estimation and Risk Measure Computation

The estimated parameters of the GARCH(1,1) model with normally distributed innovations applied to the real S&P 500 returns are presented in Table 5.3.1. The estimates confirm the persistence of significant volatility, indicated by the sum of the parameters α and β approaching unity, while the small value of ω highlights a relatively low long-term variance level.

Parameter	Estimate	Std. Error
ω	2.6179e-06	8.027e-07
α	1.2481e-01	9.6854e-03
β	8.5305e-01	1.0814e-02

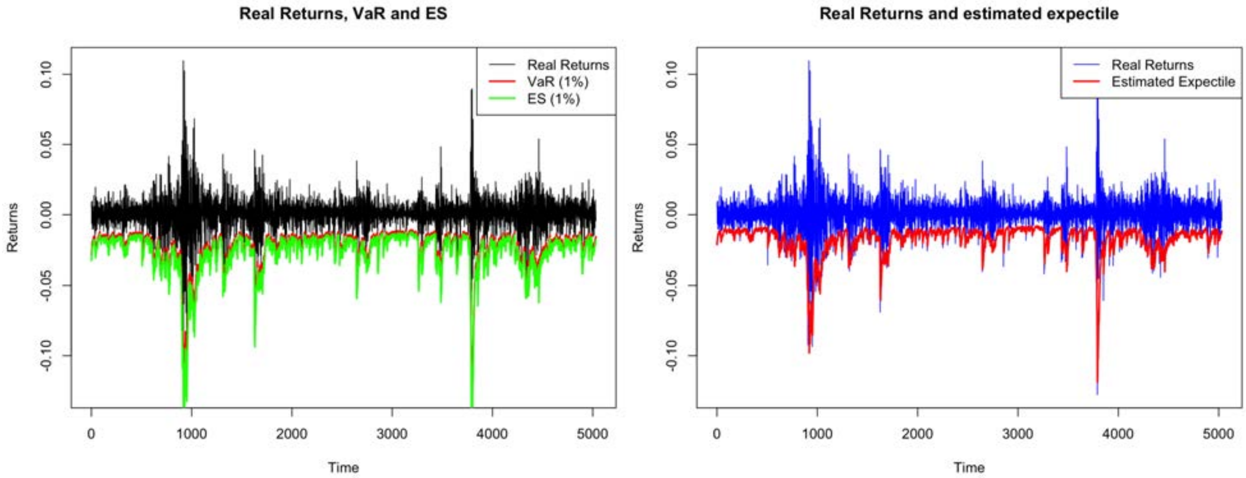
Table 5.3.1: Parameters estimated for the real data GARCH(1,1) Normal model

A GARCH(1,1) model was estimated on the log returns of the S&P 500 index, following the same specification and estimation procedures detailed in previous sections. The conditional volatilities are obtained from the GARCH(1,1) model, the VaR is estimated by equation (3.4), the ES by equation (3.7) and the expectile through the minimization of the asymmetric loss function given by equation (3.9). The estimation procedures remain consistent with those used for the simulated data, ensuring a coherent comparison between simulation and real data applications.

5.4 Empirical Risk Measure Results

5.4.1 Risk Measure Estimates

The fitted GARCH(1,1) model generated a series of conditional volatility estimates that were used to estimate the VaR, ES, and expectile forecasts at the 1% confidence level. Figure 5.4.1a illustrates the daily real returns of the S&P 500 (black) alongside the 1% Value at Risk (red) and the 1% Expected Shortfall (green). The VaR line represents the threshold below which the daily return is not expected to fall more than 1% of the time, assuming that the model is correctly specified. We observe that most returns lie above this VaR boundary, but occasional violations occur, indicating periods of increased market stress. The ES line goes further than the VaR by estimating the average loss on days when the return breaches that VaR threshold, thereby capturing deeper tail-risk events. Comparing the real returns to these risk measures shows how VaR provides a boundary for expected losses, while ES offers additional insight into the severity of extreme market movements. As expected, the ES always falls below the VaR. Figure 5.4.1b illustrates the daily real returns of the S&P 500 (blue) compared with the estimated expectile (red). This alternative risk measure lies below most of the returns, capturing the behavior in the lower tail of the distribution. In particular, during episodes of increased volatility or negative market shocks, the expectile adjusts downward to reflect a higher downside risk, thus highlighting the magnitude of extreme losses.



(a) The Real Returns, the Estimated VaR and ES of the S&P 500.

(b) Real Returns and Estimated Expectile of the S&P 500.

Figure 5.4.1: (a) VaR&ES vs. Real Returns; (b) Expectile vs. Real Returns for the S&P 500.

5.4.2 Backtesting of VaR, ES and Expectile

Backtesting was performed on the S&P 500 returns to evaluate the accuracy of the estimated risk measures.

Concerning the VaR backtesting, a hit sequence was constructed by identifying the days when the observed return fell below the estimated VaR. The empirical violation rate was found to be approximately 2.29%, with 115 exceedances observed. This rate exceeds the nominal 1% level, suggesting that the model might underestimate the tail risk during extreme market conditions. Moreover, the Dynamic Quantile (DQ) test strongly rejects the null hypothesis of the correct model specification (with a p-value effectively equal to zero and a statistic equal to 108.5728). These results suggest that the model underestimates tail risk during periods of increased market volatility.

Concerning the ES backtesting, the violation residuals were calculated by comparing the realized losses with the ES estimates. For the real returns, the mean of the violation residuals K_t (denoted as $\bar{K}$) is 0.1179 with a variance of 0.0852, based on 115 observed violations. The t-test statistic for these residuals is 4.3294, with a corresponding p-value of 0.0032%. Furthermore, the Portmanteau test on the violation residuals yielded the test statistic $Q(10) = 14.962$ for 10 lags with a p-value of 0.1334, indicating no significant serial correlation. Together, these results do not support the robustness of the ES estimates at significance levels of 1% or of 5%.

For the expectile backtest, an unconditional coverage test was performed. The test produced a Wald statistic of 25.74 with a p-value on the order of 3.9×10^{-7}, indicating statistically significant deviations from the expected coverage. Moreover, additional diagnostic tests using the Box-Pierce test were conducted on several candidate processes derived from the expectile framework. Specifically, the expectile exceedance process (C_1) produced a statistic of 22.87 (p-value=0.0112), the normalized identification process (C_2) returned a statistic of 16.44 (p-value=0.0877) and the normalized squared excess return process (C_3) produced a statistic of 81.28 (p-value=2.82×10^{-13}). These mixed results, with a test indicating significant serial dependence and others not, suggest that the expectile approach may be particularly sensitive to model specification issues when applied to real data.

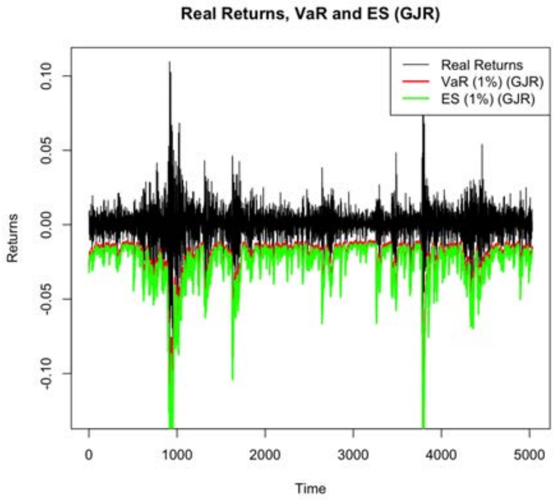
	GARCH(1,1) with $\mathcal{N}$ innovations on real S&P500 returns	
Backtest	**Statistic**	**P-value**
VaR: DQ test	108.5728	0
ES: K_t t-test	4.3294	0.0032%
ES: BP(K_t violations)	14.962	13.34 %
Expectile: UC test	25.7446	$3.8972 \cdot 10^{-5}$%
Expectile: BP(C_1)	22.8703	1.1236%
Expectile: BP(C_2)	16.4423	8.7652 %
Expectile: BP(C_3)	81.2804	$2.8155 \cdot 10^{-11}$%

Table 5.4.1: The results of the backtesting methods for the different risk measures for the GARCH(1,1) model with normally distributed innovations.

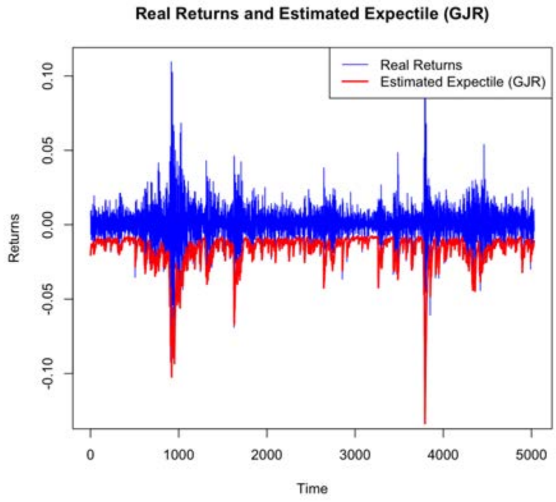
5.5 GJR-GARCH(1,1) Model: An Alternative Specification and Its Backtesting Results

Due to the empirical challenges observed when backtesting risk measures with the GARCH(1,1) model on real data, we explore an alternative volatility specification that allows for asymmetry in the conditional variance. To this end, we estimate a GJR-GARCH(1,1) model on the S&P 500 index returns. The asymmetric structure of the GJR model permits negative shocks to affect future volatility stronger than positive shocks, and thereby yields a potentially improved representation of tail risk.

Figure 5.5.1 illustrates the risk measures (VaR, ES and expectile) and the real returns under the GJR-GARCH(1,1) model with normally distributed innovations.



(a) Real Returns, VaR, and ES of the S&P 500 under the GJR-GARCH model.



(b) Real Returns and Estimated Expectile of the S&P 500 under the GJR-GARCH model.

Figure 5.5.1: (a) VaR&ES vs. Real Returns under the GJR-GARCH model; (b) Expectile vs. Real Returns for the S&P 500 under the GJR-GARCH(1,1) model with normally distributed innovations.

Table 5.5.1 reports the parameter estimates of the GJR-GARCH(1,1) model with normally distributed innovations, applied to the real S&P 500 returns. The significant and positive estimate of the leverage parameter γ indicates clear evidence of the leverage effect, which means that negative shocks tend to have a stronger impact on volatility than positive shocks. Furthermore, the high persistence in volatility is confirmed by the sum of α, $\gamma/2$, and β being close to one.

Parameter	Estimate	Std. Error
μ	3.2879e-04	9.6624e-05
ω	2.6532e-06	2.6179e-07
α	9.7726e-03	3.218e-03
γ	1.8871e-01	1.2971e-02
β	8.6877e-01	7.1515e-03

Table 5.5.1: Parameters estimated for the real data GJR-GARCH(1,1) Normal model

Backtesting Results

The dynamic Value-at-Risk (VaR) forecast is computed as

$$VaR_t(\alpha) = \mu_t + \sigma_t \, \Phi^{-1}(\alpha), \text{ for } \alpha = 1\%$$

where the conditional mean μ_t and volatility σ_t come from the GJR-GARCH model. In our application, the proportion of VaR violations (i.e. the hit rate) is approximately 2.2% (111 violations out of 5031 observations), a rate that is slightly higher than the nominal 1% level. The Dynamic Quantile (DQ) test for the VaR yields a high test statistic (92.32) with an extremely small p-value ($\sim 1.1 \times 10^{-16}$), suggesting that despite the asymmetry of the model, there remains evidence of underestimation of tail risk during extreme market periods.

To evaluate the performance of the ES forecasts, we perform backtesting methods by calculating the violation residuals K_t. The mean of the violation residuals K_t, denoted as $\bar{K}_{\mathrm{GJR}}$, is 0.1472 and the sample variance of K_t is 0.1127. The corresponding t-test statistic is 4.6191 with a p-value of 0.0011%. These results indicate that, under the GJR-GARCH(1,1) model, the ES forecasts differ significantly from the expected coverage level, as the t-test rejects the null hypothesis. In addition, diagnostic tests (including the Portmanteau test and the analysis of autocorrelation functions) were performed on the violation residuals, and the violation residuals show no significant autocorrelation with a Portmanteau p-value of 0.7256 for 10 lags.

In the case of expectiles, the risk measure is determined by minimizing the asymmetric loss function given by equation (3.9). The expectile forecast computed under the GJR-GARCH framework provides an alternative view on tail risk. To backtest the expectile, an unconditional coverage test is performed, yielding a Wald statistic of 32.27 and a p-value on the order of 1.34×10^{-8}, a strong indication that the expectile forecast deviates from the nominal level.

Additional diagnostics are performed using the three candidate processes derived from the expectile framework, i.e. the expectile exceedance process (C_1), the normalized identification process (C_2), and the normalized squared excess return (C_3). For the GJR specification, the Portmanteau tests on these processes revealed mixed outcomes. Although the test on C_2 does not indicate a significant serial correlation (p-value ≈ 0.9), both C_1 (p-value ≈ 0.035) and C_3 (p-value $\approx 4.6 \times 10^{-5}$) indicate a significant serial correlation. This result suggests that the expectile backtest is quite sensitive to the model specification in the tail.

	GJR-GARCH(1,1) with $\mathcal{N}$ innovations on real S&P500 returns	
Backtest	**Statistic**	**P-value**
VaR: DQ test	92.3151	$1.1102 \cdot 10^{-14}$%
ES: K_t t-test	4.6191	0.0011%
ES: BP(K_t violations)	6.9987	72.56 %
Expectile: UC test	32.2671	$1.3437 \cdot 10^{-6}$%
Expectile: BP(C_1)	19.477	3.4606%
Expectile: BP(C_2)	5.026	88.9438 %
Expectile: BP(C_3)	37.5414	$4.56 \cdot 10^{-3}$%

Table 5.5.2: The results of the backtesting methods for the different risk measures for the GJR-GARCH(1,1) model with normally distributed innovations.

5.6 Student-t GARCH Models on Real S&P 500 Returns

In this section, we deepen our analysis of volatility dynamics by fitting heavy-tailed GARCH models to daily log-returns of the S&P 500 index. Focusing first on the symmetric GARCH(1,1) specification with Student-t innovations, we derive analytic expressions for both Value-at-Risk (VaR) and Expected Shortfall (ES), discuss the necessary standardization to unit variance, and present formal backtests. We then extend to the asymmetric GJR-GARCH(1,1) variant to capture leverage effects, comparing its tail-risk predictions to the symmetric case.

5.6.1 Model Specification and Motivation

Let r_t denote the daily log-return of the S&P 500. A classical GARCH(1,1) framework assumes

$$r_t = \mu_t + \epsilon_t, \qquad \epsilon_t = \sigma_t z_t,$$

with the conditional variance following

$$\sigma_t^2 = \omega + \alpha\,\epsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2.$$

Under normal errors, this model often underestimates the probability of extreme returns. To accommodate heavier tails, we assume

$$z_t \sim t_\nu(0,1),$$

i.e. innovations follow a Student-t distribution with $\nu > 2$ degrees of freedom, mean zero, and variance $\mathbb{V}\mathrm{ar}(z_t) = \nu/(\nu-2)$. To work with unit-variance innovations, we define the standardized variable

$$\tilde{z}_t = \sqrt{\frac{\nu-2}{\nu}}\,z_t,$$

which satisfies $\mathbb{E}[\tilde{z}_t] = 0$ and $\mathbb{V}\mathrm{ar}(\tilde{z}_t) = 1$. Thus, in the risk calculations below, the quantiles and densities of the standard Student-t distribution are scaled by $\sqrt{(\nu-2)/\nu}$.

5.6.2 Risk Measure Formulas for a Standardized Student-t

Suppose a random variable R representing returns satisfies

$$\frac{R-\mu^*}{\sigma^*} \sim t_\nu,$$

with density $g_\nu(\cdot)$ and cumulative distribution function $t_\nu(\cdot)$ where we assume $\nu > 2$ to have a finite variance $\mathbb{V}\mathrm{ar}(t_\nu) = \dfrac{\nu}{\nu-2}$. For a given probability level $\alpha = 1\%$:

$$VaR_\alpha(R) = \mu^* + \sigma^*\,t_\nu^{-1}(\alpha), \tag{5.2}$$

$$ES_\alpha(R) = \mu^* - \sigma^*\,\frac{g_\nu\big(t_\nu^{-1}(\alpha)\big)}{\alpha}\,\frac{\nu + \big(t_\nu^{-1}(\alpha)\big)^2}{\nu-1}. \tag{5.3}$$

Here $t_\nu^{-1}(\alpha)$ denotes the α-quantile of a standard t_ν distribution, and

$$g_\nu(x) = \frac{\Gamma\big((\nu+1)/2\big)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\Big(1+\frac{x^2}{\nu}\Big)^{-\frac{\nu+1}{2}}.$$

Because the raw Student-t has variance $\nu/(\nu-2)$, we multiply the quantile and density arguments by $\sqrt{(\nu-2)/\nu}$ to ensure unit variance before re-scaling by σ_t. Therefore, for our given probability level $\alpha = 1\%$, we use the following formulas to estimate the value at risk and the ES by using the scale parameter $\sqrt{(\nu-2)/\nu}$, if $\mathbb{E}(R_t) = \mu_t$ and $\mathbb{V}\mathrm{ar}(R_t) = \sigma_t$, then:

$$VaR_\alpha(R) = \mu_t + \sigma_t \sqrt{\frac{(\nu-2)}{\nu}}\, t_\nu^{-1}(\alpha), \tag{5.4}$$

$$ES_\alpha(R) = \mu_t - \sigma_t \sqrt{\frac{(\nu-2)}{\nu}}\, \frac{g_\nu\big(t_\nu^{-1}(\alpha)\big)}{\alpha} \frac{\nu + \big(t_\nu^{-1}(\alpha)\big)^2}{\nu - 1}. \tag{5.5}$$

5.6.3 Empirical Implementation

Applying the above in practice involves three steps:

1. **Parameter Estimation.** Fit the symmetric GARCH(1,1) with t-errors to historical returns, obtaining $\hat{\omega}, \hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\nu}$.

2. **One-step Risk Forecasts.** Compute conditional volatility for the returns $\hat{\sigma}_t$ and then, for a confidence level $\alpha = 1\%$:

$$\widehat{VaR}_t(\alpha) = \hat{\mu}_t + \hat{\sigma}_t \sqrt{\frac{\hat{\nu}-2}{\hat{\nu}}}\, t_{\hat{\nu}}^{-1}(\alpha),$$

$$\widehat{ES}_t(\alpha) = \hat{\mu}_t - \hat{\sigma}_t \sqrt{\frac{\hat{\nu}-2}{\hat{\nu}}}\, \frac{g_{\hat{\nu}}\big(t_{\hat{\nu}}^{-1}(\alpha)\big)}{\alpha} \frac{\hat{\nu} + \big(t_{\hat{\nu}}^{-1}(\alpha)\big)^2}{\hat{\nu} - 1}.$$

   Expectile Estimation under Student-t Innovations. Instead of assuming normal errors, we exploit the fact that an expectile at level α minimizes the asymmetric least-squares loss

$$S(\hat{e}, r_t) \;=\; \alpha \sum_{r_t \geq \hat{e}} (r_t - \hat{e})^2 \;+\; (1-\alpha) \sum_{r_t < \hat{e}} (r_t - \hat{e})^2.$$

   Under Student-t residuals with ν degrees of freedom, we first draw a large Monte Carlo sample $z_i \sim t_\nu(0,1)$, standardize via $\tilde{z}_i = \sqrt{\frac{\nu-2}{\nu}}\, z_i$, and numerically solve for the standardized expectile $\hat{e}_{\text{std}}$. The conditional expectile forecast is then

$$\hat{e}_t \;=\; \hat{\mu}_t + \hat{\sigma}_t\, \hat{e}_{\text{std}}.$$

3. **Backtesting.** For the VaR backtest, we test the coverage and independence of the VaR by performing the Dynamic Quantile (DQ) test. For the ES backtest, we apply a t-test on the violation residuals K_t for zero mean and a Portmanteau test for serial correlation with 10 lags. For the expectile backtests, an Unconditional Coverage (UC) test is performed, and a Portmanteau test is performed on each of the transformed processes C_1, C_2 and C_3 with 10 lags.

Table 5.6.1 displays the estimated parameters for the GARCH(1,1) model with Student-t distributed innovations fitted to real S&P 500 returns. The shape parameter ν, estimated around 5.58, suggests that the return distribution exhibits heavy tails, supporting the use of the Student-t distribution. The persistence of volatility remains substantial, as indicated by the sum of α and β, which approaches unity.

Parameter	Estimate	Std. Error
μ	8.7671e-04	9.8434e-05
ω	1.6856e-06	1.3902e-06
α	0.1352	2.5085e-02
β	0.8604	2.2088e-02
ν (shape)	5.5835	0.5707

Table 5.6.1: Parameters estimated for the real data GARCH(1,1) Student-t model

Table 5.6.2 shows the results of the backtesting methods (the statistics and p-values) for the different risk measures for the Student-t GARCH(1,1) on real S&P500 returns. We use a 5 % significance level throughout. For the symmetric t-GARCH, the DQ statistic (36.67) far exceeds its 5 % critical value and yields a very low p-value (0.0013%), indicating that the VaR forecasts under-cover true tail risk. In contrast, the ES backtest K_t statistic returns a p-value of 0.9973, showing no evidence of miscoverage, i.e. the ES forecasts do not significantly differ from the expected coverage level. The Portmanteau test performed on the violation residuals fails to reject the null hypothesis of linear independence at the 5% level.

For the expectiles, the unconditional coverage test produces p-value of 0.008, implying that the 1 % expectile forecasts for the symmetric t-GARCH deviate significantly from the nominal coverage. For the Box–Pierce tests on the transformed series C_1, C_2 and C_3, only C_2 fails to reject independence at the 5 % level, suggesting remaining serial dependence.

	Student-t GARCH(1,1) on real S&P500 returns	
Backtest	**The Statistic**	**P-value**
VaR: DQ test	36.6672	0.0013%
ES: K_t statistic	-0.0034	99.73%
ES: BP(K_t violations)	3.2777	97.41 %
Expectile: UC test	6.962	0.8%
Expectile: BP(C_1)	19.731	3.2%
Expectile: BP(C_2)	6.348	78.5%
Expectile: BP(C_3)	67.895	0%

Table 5.6.2: The results of the backtesting methods for the different risk measures for the Student-t GARCH(1,1) on real S&P500 returns.

5.6.4 Asymmetric Extension: GJR-GARCH

To capture the empirically observed leverage effect, where negative shocks raise volatility more than positive ones, we fit a GJR-GARCH(1,1) model with Student-t distributed innovations to the real data and recompute all risk measures and backtests as above. Comparing symmetric and asymmetric fits highlights how skewed volatility responses alter tail risk estimates.

Table 5.6.3 presents the parameter estimates for the GJR-GARCH(1,1) model with Student-t distributed innovations fitted to real S&P 500 returns. The highly significant leverage parameter γ indicates a pronounced asymmetry, showing stronger volatility reactions to negative shocks. Furthermore, the shape parameter ν close to 6 confirms that the Student-t distribution effectively captures the fat-tailed nature of the return distribution, which justifies the choice of this model specification.

Parameter	Estimate	Std. Error
μ	6.043e-04	9.9338e-05
ω	2.0661e-06	9.1117e-07
α	2.6752e-07	1.133e-02
γ	0.2331	2.86e-02
β	0.8657	1.2621e-02
ν (shape)	5.9706	0.5033

Table 5.6.3: Parameters estimated for the real data GJR-GARCH(1,1) Student-t model

Table 5.6.4 shows the results of the backtesting methods (statistics and p-values) for the Student-t GJR-GARCH(1,1) on real S&P500 returns.

Using a 5 % significance level, the DQ statistic (19.57) produces a p-value of 0.0121, indicating that the VaR forecasts undercover tail risk at 5 % (although at the 1 % level the test does not reject the null hypothesis of correct VaR estimation and independence). In contrast, the ES backtest K_t statistic (1.1947) returns a p-value of 0.2361, showing no evidence of miscoverage. The Portmanteau test performed on the violation residuals fails to reject the null hypothesis of linear independence at the 5% level.

For expectiles, the unconditional coverage statistic (3.568) produces a p-value of 0.059, implying that the 1 % expectile forecasts do not deviate significantly from the nominal coverage at the 5 % level. The Box–Pierce tests on the transformed series C_1, C_2 and C_3 show that only C_3 rejects independence ($p = 0.002$) at both the 1 % and 5 % levels, while C_1 ($p = 0.068$) and C_2 ($p = 0.962$) exhibit no serial dependence.

	Student-t GJR-GARCH(1,1) on real S&P 500 returns	
Backtest	**The Statistic**	**P-value**
VaR: DQ test	19.5657	1.211%
ES: K_t statistic	1.1947	23.61%
ES: BP(K_t violations)	6.5323	76.87 %
Expectile: UC test	3.568	5.9%
Expectile: BP(C_1)	17.306	6.8%
Expectile: BP(C_2)	3.648	96.2%
Expectile: BP(C_3)	28.363	0.2%

Table 5.6.4: The results of the backtesting methods for the different risk measures for the Student-t GJR-GARCH(1,1) on real S&P500 returns.

5.7 Conclusion

In this chapter, we demonstrate that replacing the basic Gaussian GARCH(1,1) with the asymmetric Student-t GJR-GARCH(1,1) yields markedly improved tail-risk forecasts at the 5 % level, reducing VaR under-coverage and producing accurate ES and expectile estimates. Extending from the symmetric Student-t GARCH(1,1) to the asymmetric Student-t GJR-GARCH(1,1) improved the estimation performance of the VaR and the expectile. These results underscore the importance of jointly modeling heavy-tailed shocks and leverage effects for reliable extreme risk measurement.

Chapter 6

Conclusion

The purpose of this thesis is to examine the backtesting of risk measures in dynamic financial models, with a particular focus on evaluating Value-at-Risk (VaR), Expected Shortfall (ES) and expectiles under a GARCH(1,1) and a GJR-GARCH(1,1) volatility framework. Our goal is to assess how well a dynamic model can predict extreme losses and estimate their risk measures and how reliable those predictions are when subjected to formal backtesting procedures. In this concluding chapter, we summarize the main findings of the study, reflect on the challenges encountered, and discuss the implications, limitations, and directions for future research.

Summary of Key Findings: The empirical results demonstrate that the GJR-GARCH(1,1) model with Student-t distributed innovations provides robust risk measures for financial returns. In particular, the model's predictions for expected shortfall and expectile show strong performance in backtesting. Both the ES and expectile backtests yield p-values above the 5% significance level, which means that we find no statistical evidence to reject the model at the conventional confidence threshold. This indicates that the GJR-GARCH(1,1) model with Student-t distributed innovations is capable of sufficiently capturing the tail risk of the returns data in terms of the frequency and magnitude of extreme losses. In contrast, the backtest for VaR proves to be more demanding: the dynamic quantile (DQ) test for VaR violations only yields a satisfactory result at the 1% significance level. In other words, at the 5% level the VaR backtest suggests a slight model misspecification (failing to meet the ideal coverage or independence criteria), whereas at the 1% level the model's VaR predictions can be considered acceptable. This outcome is not unexpected given the challenging nature of financial data and the strictness of VaR tests, financial return series are notoriously volatile and heavy-tailed, making it difficult for any single model to perfectly anticipate every VaR exceedance. However, the ability of our model to meet the VaR backtest criteria at a stricter confidence level, combined with its success in the ES and expectile tests at the standard level, shows the overall strength of the chosen GJR-GARCH approach with Student-t innovations in capturing risk.

Overview of Thesis Contributions: Each chapter of this thesis contributes to the investigation of backtesting risk measures in dynamic models. Chapter 1 introduces the research problem and context, highlighting the importance of accurate risk measurement for financial institutions and regulators. It indicates that the most used risk measure is the VaR, despite its non-coherence, and some regulation frameworks are replacing it with the ES, despite its non-elicitability. An alternative risk measure that is coherent and elicitable is the expectile, which is not yet used in regulation or in the industry, as expectiles are relatively new to risk management. Chapter 2 provides a review of the relevant literature and theoretical background. Risk measures, such as VaR, ES, and expectiles, and the properties of risk measures, such as coherence and elicitability, are defined.

In Chapter 3, we develop the methodological core of the study. We describe in detail the GARCH family of models used, that is, GARCH(1,1) and GJR-GARCH(1,1), with particular emphasis on the GJR-GARCH(1,1) specification and its ability to capture asymmetry in volatility (the leverage effect). Returns are simulated using a GARCH (1,1) model with normally distributed innovations, and risk measures are estimated for simulated returns. Chapter 3 also compares the forecasts of the risk measures at the 1% level and the following ordering is found: $|\text{Expectile}_t(1\%)| \leq |VaR_t(1\%)| \leq |ES_t(1\%)|$. Additionally, Chapter 4 outlines the backtesting approaches employed, including, among others, the DQ test for the VaR, the t-test on the violations residuals K_t of the ES and a Portmanteau test on those violation residuals. The backtesting methods used for the expectile test separately unconditional coverage with a Wald test on the identification function, and independence which is tested three times using different transformed processes derived from the identification function. The performance of those backtesting methods is consistently verified by bootstrap.

Chapter 5 presents the empirical study and its results. We describe the dataset of financial returns (daily S&P 500 returns) and the process of estimating the GARCH(1,1) and GJR-GARCH(1,1) models first for normally distributed innovations and then, since backtesting methods do not perform well for those models, for Student-t distributed innovations. The chapter then details the estimated risk measures and their backtests' results. It shows the improvement in the estimation of the risk measure and of the backtest results when we apply the GJR-GARCH models with Student-t distributed innovations to the S&P500 returns. In particular, the 1% VaR forecasts of the model occasionally underestimate actual losses, leading to slightly more VaR exceedances than the nominal rate of 5%, whereas the 1% ES forecasts are in line with observed tail losses. The expectile, which provides an alternative perspective on tail risk, is also backtested and found to be consistent with the data. Chapter 5's comparative analysis indicates that among the models considered, the GJR-GARCH(1,1) with Student-t innovations performed the best overall in terms of these backtest metrics.

Challenges with Real Data: A recurring theme throughout the research is the difficulty of working with real financial data in risk modeling and backtesting. Financial time series typically exhibit features like volatility clustering, fat tails, and occasional structural breaks (for example, sudden market crashes or regime shifts), all of which pose challenges for model-based prediction. In this thesis, we discuss these issues first-hand. Even with a flexible model such as GJR-GARCH(1,1) with Student-t distributed innovations, there are limits to capture every nuance of the data. For instance, during periods of extreme market turbulence, the model might lag in adjusting to a sudden increase in volatility, leading to risk measure forecasts that were temporarily off-mark. Moreover, backtesting results must be interpreted with caution: a failure to pass a test can occur simply due to random chance or sample peculiarities, while passing a test does not guarantee the model will continue to perform well under all future conditions. We are mindful of these nuances when drawing conclusions. Working with real data requires balancing statistical evidence with financial intuition, ensuring that models are not only mathematically sound, but also practically sensible. The challenges we face show the importance of robust risk management practices, the combination of quantitative models with expert judgment, and the continuous updating of models to reflect new information.

Limitations and Future Research: While the findings of this thesis are encouraging, they come with certain limitations that suggest directions for further study. One limitation is the scope of the models explored. We focus primarily on the GJR-GARCH(1,1) model with Student-t innovations due to its effectiveness in volatility forecasting, but many other modeling approaches could be examined within a similar backtesting framework. It would be valuable to compare the performance of the GJR-GARCH model against alternative volatility models such as EGARCH models, or even modern machine learning methods that can capture non-linear patterns in financial data. Additionally, exploring different distributional assumptions (such as a skewed-t distributed innovations or non-parametric density estimates for the innovations) might further improve the accuracy of risk forecasts by better capturing the extreme tails of the return distribution.

Another limitation lies in the empirical scope of the study. Our analysis is performed on a single dataset (a specific market index or asset over a given period), which means the conclusions drawn may not directly generalize to other markets or time frames. Different assets or longer sample periods could exhibit different behaviors, potentially requiring model recalibration or alternative approaches. Future research could address this by applying the backtest methodology to various datasets and comparing results.

Further research could focus on refining risk measures and their evaluation methods. Moreover, the expectile is still relatively new in practice and deserves deeper investigation. Additional studies could examine how expectile forecasts compare with VaR and ES across various scenarios, and work to improve the backtesting techniques for expectiles to enhance their reliability. Note that our backtesting of expected shortfall did not include a real test for independence, we only performed Portmanteau tests, which is just a test for linear independence and not independence in general, so future research could look at how to test for the independence of the estimated ES. Indeed, backtesting methods for the ES could be strengthened. Researchers might develop joint tests that evaluate VaR and ES simultaneously, or apply advanced scoring functions to directly judge the accuracy of tail-risk predictions.

Concluding Remarks: In conclusion, this thesis demonstrates the value of dynamic modeling in the backtesting of financial risk measures. Using an asymmetric GARCH model with heavy-tailed innovations, we produce risk forecasts that, for the most part, stand up to rigorous evaluation. The GJR-GARCH(1,1) model with Student-t innovations is particularly effective in capturing the distribution of losses, as evidenced by its strong performance in expected shortfall and expectile backtests. The minor shortfall observed in the VaR backtest highlights a well-known challenge: even the best models cannot fully eliminate uncertainty in extreme events. However, the methodologies and insights presented here contribute to a deeper understanding of risk measure backtesting. They reinforce the notion that combining advanced statistical models with thorough validation is essential for credible risk management. We hope that this work will serve as a useful reference for risk professionals and as a stepping stone for future research, ultimately helping to develop more resilient financial models and more effective risk management strategies.

Bibliography

Acerbi, C. and Szekely, B. (2014), 'Backtesting expected shortfall'. Preprint.

Armando, J., Lu, Y., Mailhot, M. and Meng, X. (2024), 'Backtesting expectile: Disentangling unconditional coverage and independence properties'.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999), 'Coherent measures of risk', *Mathematical Finance* .

Bellini, F., Klar, B., Müller, A. and Rosazza Gianin, E. (2014), 'Generalized quantiles as risk measures', *Insurance: Mathematics and Economics* **54**, 41–48.
URL: *https://www.sciencedirect.com/science/article/pii/S0167668713001698*

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**(3), 307–327.
URL: *https://ideas.repec.org/a/eee/econom/v31y1986i3p307-327.html*

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

Daouia, A., Girard, S. and Stupfler, G. (2018), 'Estimation of tail risk based on extreme expectiles', *Journal of the Royal Statistical Society Series B* **80**(2), 263–292.
URL: *https://ideas.repec.org/a/bla/jorssb/v80y2018i2p263-292.html*

Daouia, A., Girard, S. and Stupfler, G. (2020), 'Tail-expectile processes with applications to dynamic risk measurement', *Journal of Multivariate Analysis* **178**, 104626.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (2015), *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press.

Engle, R. F. (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica* **50**(4), 987–1007.
URL: *https://ideas.repec.org/a/ecm/emetrp/v50y1982i4p987-1007.html*

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.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *The Journal of Finance* **48**(5), 1779–1801.

Koenker, R. (1993), 'When are expectiles percentiles?', *Econometric Theory* **9**(3), 526–527.

Koenker, R. and Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

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

McNeil, A. J., Frey, R. and Embrechts, P. (2015), *Quantitative Risk Management*, 2 edn, Princeton University Press.

Newey, W. and Powell, J. (1987), 'Asymmetric least squares estimation and testing', *Econometrica* **55**(4), 819–847.

Spring, K. (2021), 'Backtesting the expected shortfall', *Junior Management Science (JUMS)* **6**(3), 590–636.
URL: *https://hdl.handle.net/10419/294965*

Tadese, M. and Drapeau, S. (2020), 'Relative bound and asymptotic comparison of expectile with respect to expected shortfall', *Insurance: Mathematics and Economics* **93**, 387–399.
URL: *https://www.sciencedirect.com/science/article/pii/S016766872030086X*

Yahoo Finance (2025), 'S&P 500 (ˆGSPC)', https://finance.yahoo.com/quote/ Data period: February 22, 2005–February 20, 2025.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**