Spring, Konstantin

**Article**

# Backtesting the Expected Shortfall

Junior Management Science (JUMS)

**Provided in Cooperation with:**
Junior Management Science e. V.

*Suggested Citation:* Spring, Konstantin (2021) : Backtesting the Expected Shortfall, Junior Management Science (JUMS), ISSN 2942-1861, Junior Management Science e. V., Planegg, Vol. 6, Iss. 3, pp. 590-636, https://doi.org/10.5282/jums/v6i3pp590-636

This Version is available at:
https://hdl.handle.net/10419/294965

Junior Management Science 6(3) (2021) 590-636

**Junior Management Science**

journal homepage: www.jums.academy

# Backtesting the Expected Shortfall

Konstantin Spring

*Universität Konstanz*

**Abstract**

Backtesting of risk measure estimates is an integral part for an effective risk management. With the growing importance of the Expected Shortfall (ES) to potentially replace the Value at Risk (VaR) as a primary measure for market risk this also calls for suitable backtesting solutions. Although a variety of approaches has been proposed in the past, there is still an on-going discussion whether the ES can be properly backtested. The thesis adds to this discussion in the following way. Five of the most promising backtests for the ES are implemented, compared based on theoretical properties like empirical size and power and tested against ES estimation models which are fitted to historical returns of the S&P 500. In addition, all backtests in scope are assessed against a set of criteria which reflect their practical applicability for both regulators and financial institutions. Results presented within this thesis confirm that backtesting the ES is indeed not much more complicated than backtesting the VaR. Backtesting ES might be conceptually less straight forward, but there are multiple promising approaches which allow for a reasonable validation of ES estimation models.

*Keywords:* Expected shortfall; backtesting; risk measures; statistical test.

## 1. Introduction

This thesis deals with the ability to backtest the *Expected Shortfall (ES)* risk measure. Indeed, backtesting the ES has been one of the most interesting topics in the field of financial risk management over the last two decades. Until today there is no consensus amongst researches on how to backtest the ES. Furthermore, there is still an ongoing and controversial debate on whether it is possible to backtest the ES at all. This thesis aims to shed light on the diverse literature on backtesting the ES. Moreover, it yields a comprehensive evaluation of multiple backtesting methodologies, suggested within the literature, also with regard to their practical applicability for financial institutions.

Within the financial industry, it is common practice to make use of so called *risk measures* in order to determine potential risks inherited in an investment portfolio. Risk measures are mappings of random variables, i.e. the *Profit and Loss (P&L) distribution* of a portfolio, into real-valued scalars, which represent the capital amount that needs to be held by a financial institution as a buffer against unexpected portfolio losses. Amongst practitioners, the ES together with the *Value at Risk (VaR)* are by far the two most relevant risk measures, which will both be defined in more detail in the course of this thesis.

Besides the estimation of these risk measures, it is fundamental, especially for regulators, to be able to ex-ante verify the accuracy of a risk forecasting model, to ensure that capital buffers are sufficient. This process is referred to as the *backtesting* of a certain risk measure. As outlined by Kratz, Lok, and McNeil (2018), a *backtest* is a statistical procedure which compares forecasts to actual realizations in order to judge the correctness of the implemented forecasting model. In the case of risk management, a set of ex-ante VaR or ES estimates is compared to realized portfolio returns in order to verify if the applied estimation model accurately forecasts the risk of the underlying portfolio. For regulators it is of particular interest to detect estimation models which underestimate the actual underlying risk, as an insufficient coverage against adverse scenarios might threaten the functioning of the whole financial system in times of market distress.

For many years the VaR has been the predominant risk measure in practice. Nevertheless, it is well known that the VaR does not belong to the class of *coherent* risk measures, defined by Artzner, Delbaen, Eber, and Heath (1999), as it fails to consistently account for portfolio diversification. Furthermore, the VaR only accounts for a predefined quantile of the P&L distri-

DOI: https://doi.org/10.5282/jums/v6i3pp590-636

bution but ignores the tail of extreme events beyond that quantile, which is for example noted in Danielsson et al. (2001). Thus, it appears to be consequential, that the Basel Committee of Banking Supervision in their "Fundamental Review of the Trading Book" (see Basel Committee (n.d.-a)), proposes to replace the VaR as a primary measure for market risk by the ES. Indeed, the ES makes up for the weaknesses of the VaR listed above, which will be outlined in the course of the following chapter. Furthermore, as for example argued in Emmer, Kratz, and Tasche (2015), most forecasting models for the VaR can easily be generalised for the estimation of the ES.

Nevertheless, there is one major concern about the ES both practitioners and researchers worry about - the ability to backtest the ES. Indeed, backtesting VaR is more or less intuitive and the respective literature is well developed.[1]. In comparison to the VaR, there is still an ongoing discussion on whether the ES is even backtestable at all. This was originally fueled by Gneiting (n.d.), who proved that the ES lacks a statistical property known as *elicitabilty*. As shortly summarized by Kratz et al. (2018), a risk measure is called elicitable, if it can be defined as the solution of a forecasting-error minimization problem. Further, Gneiting (n.d.) argues that the lack of this property makes it cumbersome or maybe even impossible to backtest the ES. This opinion is further supported by contributions like Chen (2014) or Carver (2013). Nevertheless another opinion starts to prevail in recent years. As an example, Acerbi and Szekely (2014) makes the point, that elicitability has actually nothing to do with backtesting at all. The concept of elicitability as well as any potential implications on the ability to backtest the ES will also be taken up in the course of this thesis.

Nevertheless and despite all doubts expressed within the literature, a variety of potential backtests for the ES has been suggested in the last two decades[2]. This thesis tries to provide insights into the diverse and still emerging literature by implementing some of the most promising approaches. More precisely, five different approaches are selected, based on Kratz et al. (2018), Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015), and McNeil and Frey (2000), which are evaluated in the course of this thesis. Furthermore, some adjustments to the original test version are proposed, in order to ensure both practical applicability as well as compliance with regulatory needs. All backtests are implemented in *Python* and compared according to classical measures like empirical power and size. Additionally, they are tested in the their judgement of actual forecasting models, which are fitted to financial returns of the S&P 500.

The remainder of this Master thesis is structured as follows. Chapter 2 is going to introduce the concept of risk measures, define both the VaR and the ES and outline their theoretical properties. Afterwards, chapter 3 focuses on the concept of backtesting in general, whereas chapter 4 introduces the selected approaches as well as adjustments, which will be made for the purpose of this thesis. Consecutively, chapter 5 conducts a simulation study to evaluate and compare empirical size and power of the applied tests, followed by an application of all backtests to real financial data in chapter 6. Finally, some concluding remarks on the results obtained within this thesis are presented in chapter 7.

[1] See for example Kupiec (1995) or Christoffersen (1998) for work on backtesting VaR.

[2] See for example Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015), Du and Escanciano (2017), Kratz et al. (2018), Wong (2008), Berkowitz (2001), Nolde and Ziegel (2017), McNeil and Frey (2000) and many more for approaches on backtesting the ES.

## 2. Risk Measures and their properties

The purpose of the following section is twofold. First, it gives an overview on the concept of risk measures and introduces some notation used throughout this thesis. Based on Artzner et al. (1999), Emmer et al. (2015) and Nolde and Ziegel (2017), desirable properties of risk measures are outlined. Second, the two most practically relevant risk measures, i.e. the Value at Risk (VaR) and the Expected Shortfall (ES) are formally defined. Furthermore, the section aims to give a theoretical foundation for the decision by regulators to replace the VaR by the ES as a primary measure for market risk. Thus, this section might be seen as a motivation why it is even necessary to pay attention to the backtestability of the ES.

### 2.1. Concept of Risk Measures

For the general definition of a risk measure, Artzner et al. (1999) is taken as a reference. Therefore, a position, e.g. an investment strategy or a portfolio of financial assets, is considered over a certain time horizon. The position might be described by $X$, its net worth at the end of the investment horizon, which is defined as a real-valued function on some set $\Omega$ of possible future outcomes. Moreover, let $\mathscr{X}$ be the set of all real-valued functions $X \in \mathscr{X}$, which represent the future net worth of a certain position. The formal definition of a risk measure is then given as follows.

**Definition 2.1** (Risk Measure)**.** *A risk measure $\rho$ is a mapping from $\mathscr{X}$ to $\mathbb{R}$, i.e.*

$$\rho : \mathscr{X} \to \mathbb{R}, \quad X \mapsto \rho(X). \tag{2.1}$$

Artzner et al. (1999) connects the concept of risk measurement to the particular question how close or far a certain position is from being accepted by either regulators or other stakeholders. More general speaking, $\rho(X)$, which is related to a certain end of period net worth $X \in \mathscr{X}$, is the additional capital amount that needs to be held such that the position is accepted by regulators. Correspondingly, a higher value of $\rho(X)$ is related to a riskier investment position. For instance, a value of $\rho(X) > 0$ signals that a financial institution needs to hold additional capital as a safety buffer such that the position is accepted by regulators. On the other hand, a position is accepted by regulators if $\rho(X) \leq 0$. As a natural extension, Artzner et al. (1999) thus defines the so called *acceptance set* $\mathscr{A}_\rho$, which contains all positions acceptable for regulators with respect to a certain risk measure.

**Definition 2.2** (Acceptance Set)**.** *The acceptance set related to a risk measure $\rho$ is denoted by $\mathscr{A}_\rho$ and defined through*

$$\mathscr{A}_\rho := \{X \in \mathscr{X} | \rho(X) \leq 0\} \tag{2.2}$$

Overall, the concept of risk measures facilitates the decision making of both financial institutions and regulators as it breaks down the risks inherited in a certain position to one single number. Furthermore, risk measure values ought to be easy to interpret as the amount of additional capital that needs to be held by an institution.
Obviously, a mapping $\rho$ from $\mathscr{X}$ into the real values should fulfil some desirable properties to be an adequate risk measure for practical needs. Nevertheless, the following subsection is first going to introduces both the VaR as well as the ES, before the subsection thereafter discusses their theoretical properties.

### 2.2. Value at Risk, Expected Shortfall and Spectral Risk Measures

There are several different conventions which are used throughout the literature for the definition of both the VaR as well as the ES. In order to maintain consistency, the following set-up is going to be used in the remainder of this thesis.
For the previous general introduction to risk measures, the future net worth of a position, $X$, was the variable of interest. Nevertheless, practitioners rather consider the Profit & Loss (P&L) distribution of a certain position. Moreover, for the case of risk management especially the estimation of a loss distribution is of particular interest. Consequently, I decided to rely on the following set of notations.

**Notation 2.3.** *The following notation is going to be used in the course of the thesis.*

- *Denote the unconditional return loss $L$ of a portfolio, as a real valued random variable on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Furthermore, let $F(y) := \mathbb{P}(L \leq y)$ be the cumulative distribution function (CDF) of $L$ and denote the respective probability density function (PDF) by $f(y)$.*
- *Denote the conditional return loss $\{L_t\}_{t\geq 0}$ of a portfolio as a real valued stochastic process on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, endowed with a filtration $\mathscr{F} := \{\mathscr{F}_t, t \geq 0\}$. The filtration is given by $\mathscr{F}_t := \sigma\{L_s, s \leq t\}$, for all $t \in \mathbb{N}$.*
- *For $t \in \mathbb{N}$, denote the conditional CDF of $L_t$ based on the set of past information $\mathscr{F}_{t-1}$ by $F_t(y) := \mathbb{P}(L_t \leq y|\mathscr{F}_{t-1})$. Respectively, denote the corresponding conditional PDF by $f_t$.*
- *For $t \in \mathbb{N}$, denote the conditional quantile of $L_t$ given the past information $\mathscr{F}_{t-1}$ and some level $\alpha \in (0,1)$ by $q_\alpha(L_t|\mathscr{F}_{t-1}) := F_t^{\leftarrow}(\alpha) = \inf\{y \in \mathbb{R} : F_t(y) \geq \alpha\}$, where $F_t^{\leftarrow}$ denotes the generalized inverse function of $F_t$.*

*A summary of the most important variables used throughout this thesis can be found in Appendix B attached to this thesis.*

In Notation 2.3, $L$ is defined as the random variable of return losses, nevertheless in some publications $L$ is prescribed to be the distribution of return losses, which does make a difference form a mathematical point of view. As noted by Nolde and Ziegel (2017), defining a risk measure $\rho$ on a space of return loss distributions instead of a space of random variables, only makes sense if the risk measure $\rho$ is law invariant. This means that two equivalent return loss distributions, i.e. $F^1(y) = F^2(y)$ for all $y \in \mathbb{R}$ and for some random variables $L^1$ and $L^2$, always results in the same risk measure value, i.e. $\rho(L^1) = \rho(L^2)$. Indeed, both the VaR and the ES satisfy the criteria of law-invariance, thus it is not crucial for the purpose of this thesis whether $L$ is defined as a random variable or as the related return loss distribution. Overall, I want to rely on the Notation 2.3, but I might at certain points mix up the notation, if this fits better into the commonly applied notion of a certain topic.
Especially in the area of market risk, it is well known that the financial environment and thus the risk inherited in a investment portfolio changes over time. Thus, modelling the unconditional loss $L$ of a portfolio can be seen as a rather naive approach. Consequently, the main focus of this thesis will be on conditional, time dependent, risk forecasts based on the estimated stochastic process of $\{L_t\}_{t\geq 0}$.
From a historical perspective both variance and standard deviation were the dominant risk measures in financial applications for a long period of time. The importance of the VaR started to raise when RiskMetrics announced to use the VaR as their standard measure of risk in 1996, as outlined in JP Morgan (1996). Shortly thereafter, the Basel Committee of Banking Supervision introduced the VaR as an industry standard for the measurement of market risk (see Basel Committee (n.d.-c)). In order to calculate the VaR one needs to consider a time horizon $\Delta$ as well as a confidence level $\alpha \in (0,1)$.[3] Given a return loss variable $L_t$, the VaR describes the loss over the considered time horizon $\Delta$, which is only exceeded in $1-\alpha$ percent of all cases. For the remainder of this thesis, a risk horizon of $\Delta = 1$ day is considered and thus $\Delta$ can be dropped in the following definition in order to simplify the notation.

**Definition 2.4** (Value at Risk)**.** *The Value at Risk (VaR) at time $t \geq 0$, $t \in \mathbb{N}$, at the confidence level $\alpha \in (0,1)$ based on some conditional return loss $L_t$ is defined as*

$$VaR_\alpha(L_t|\mathscr{F}_{t-1}) := q_\alpha(L_t|\mathscr{F}_{t-1}) = \inf\{y \in \mathbb{R} : F_t(y) \geq \alpha\} \,. \quad (2.3)$$

*For simplicity reasons, the short notation $VaR_{t,\alpha} = VaR_\alpha(L_t|\mathscr{F}_{t-1})$ is used.*

Analogous to Definition 2.4, one can define a time independent version $VaR_\alpha$ based on unconditional return loss $L$. The Basel Committee of Banking Supervision requires executives to calculate the VaR at a confidence level of $\alpha = 0.99$. Given any realistic portfolio and the conventions applied in Definition 2.4, it is thus reasonable to assume that the $VaR_{t,0.99}$ has a positive sign. The VaR is often criticized for two main reasons. First of all, it is insensitive to any extreme losses beyond the $\alpha$-quantile of the loss distribution. Secondly, it is not a sub-additive risk measure, i.e. it fails to consistently account for portfolio diversification. This will be outlined in more depth in subsection 2.3.

[3] Note that $\alpha$ is selected as a confidence level and not a significance level for the purpose of this thesis. Thus, $\alpha$ takes on values like 0.95 or 0.99 instead of 0.05 or 0.01.

[4]

As already shortly mentioned in the introduction, the ES overcomes both of these shortcomings and thus gains more and more attention in the measurement of market risk, as for example the Basel Committee of Banking Supervision suggests to replace the VaR by the ES under the Basel III framework[5]. Indeed, the ES appears to be a natural extension of the VaR. Whereas the VaR states the loss which is only exceeded at a certain confidence level $\alpha$, the ES goes one step further and asks for the average loss given an exceedance of the corresponding VaR threshold. More formally, given the notations introduced above, the ES is defined as follows.

**Definition 2.5** (Expected Shortfall)**.** *The Expected Shortfall (ES) at time $t \geq 0$, $t \in \mathbb{N}$, at the confidence level $\alpha \in (0,1)$ based on some conditional return loss $L_t$ with CDF $F_t$ is defined as*

$$\begin{aligned} ES_\alpha(L_t|\mathscr{F}_{t-1}) &:= \frac{1}{1-\alpha}\int_\alpha^1 q_p(L_t|\mathscr{F}_{t-1})dp \qquad (2.4)\\ &= \frac{1}{1-\alpha}\int_\alpha^1 VaR_{t,p}dp\,. \end{aligned}$$

*Moreover, if the CDF $F_t$ is continuous at the $\alpha$-quantile, the definition can be simplified, according to Acerbi and Tasche (2002), to*

$$ES_\alpha(L_t|\mathscr{F}_{t-1}) = \mathbb{E}[L_t|L_t \geq VaR_{t,\alpha}]\,. \qquad (2.5)$$

*Again the short notation $ES_{t,\alpha} = ES_\alpha(L_t|\mathscr{F}_{t-1})$ will be applied.*

The value of the ES again depends on the arbitrary confidence level $\alpha \in (0,1)$. As the conditional quantile $q_\alpha(L_t|\mathscr{F}_{t-1})$ is an increasing function in $\alpha$, it holds that $ES_{t,\alpha} \geq VaR_{t,\alpha}$. Kerkhof and Melenberg (2004) argues that the overall level of capital requirements, applying $VaR_{t,0.99}$, is more or less appropriate. Thus, in order to keep capital requirements on a similar level one should choose a lower confidence level for the ES compared to the VaR. Indeed, Basel Committee (n.d.-b) suggests the estimation of the ES at a level of $\alpha = 0.975$. If one assumes that the return loss follows a standard normal distribution, than both $VaR_{t,0.99}$ as well as $ES_{t,0.975}$ lead to similar values. Following Kerkhof and Melenberg (2004), one might denote the PDF and the CDF of the standard normal distribution by $\phi$ and $\Phi$, respectively. Then it holds [6],

$$\begin{aligned} VaR_{t,0.99} &= \Phi^{-1}(0.99) = 2.33 \approx 2.34 \qquad (2.6)\\ &= \frac{\phi(1.96)}{0.025}\\ &= \frac{\phi(\Phi^{-1}(0.975))}{1-0.975} = ES_{t,0.975}. \end{aligned}$$

Therefore, the overall magnitude of the capital buffer remains comparable. Nevertheless, if the loss distribution of a portfolio exhibits excess kurtosis compared to the standard normal distribution, this leads to a situation with $ES_{t,0.975} > VaR_{t,0.99}$. Consequently, the $ES_{t,0.975}$ measure captures the additional risk inherited in a heavy tailed loss distributions. Therefore, the ES is more sensitive to potential extreme losses in the underlying portfolio, which is beneficial from a risk management perspective. For the remainder of this thesis, the ES will always be considered at the confidence level of $\alpha = 0.975$, stipulated by the Basel Committee.

For a further classification of both the VaR and the ES the term of a *spectral risk measure*, which stems from Acerbi (2002), is introduced below. As for example described in Costanzino and Curran (2015), a spectral risk measure can be seen as a weighting of the VaR at different confidence levels given some spectrum $\psi$. According to Costanzino and Curran (2015), a so called *admissible risk spectrum* needs to fulfil the following properties.

**Definition 2.6** (Admissible risk spectrum)**.** *An integrable function $\psi \in L^1([0,1])$ is called an admissible risk spectrum, if*

*(i)* *$\psi$ is non-negative on $[0,1]$,*

*(ii)* *$\psi$ is non-decreasing on $[0,1]$,*

*(iii)* *$\|\psi\|_1 = 1$, where $\|\psi\|_1 = \int_0^1 |\psi(p)|dp$.*

Based on Definition 2.6, a spectral risk measure can be defined in the following way.

**Definition 2.7** (Spectral risk measure)**.** *Let $L_t$ be a conditional return loss with some cumulative distribution function $F_t$. Suppose the $\psi$ is an admissible risk spectrum, then the spectral risk measure $\mathscr{M}_\psi$ based on the risk spectrum $\psi$, at time $t > 0$, $t \in \mathbb{N}$, is defined by*

$$\mathscr{M}_{t,\psi} := \int_0^1 \psi(p) VaR_{t,p}dp. \qquad (2.7)$$

Given Definition 2.7, one can easily recognize that the VaR has no representation as a spectral measure, whereas the ES does belong to that class of risk measures. This is explained in the consecutive Lemma, following Costanzino and Curran (2015).

**Lemma 2.8.** *As outlined in Costanzino and Curran (2015) it holds:*

*(i)* *The VaR defined in Definition 2.4 is not a spectral risk measure.*

*(ii)* *The ES defined in Definition 2.5 is a spectral risk measure.*

*Proof.* (i) Define $\psi_{VaR}(p) := \delta_{\alpha,p}$, where $\delta$ denotes the Kronecker-Delta, then it holds

$$\mathscr{M}_{t,\psi_{VaR}} = \int_0^1 \delta_{\alpha,p} VaR_{t,p}dp = VaR_{t,\alpha}\,. \qquad (2.8)$$

But $\psi_{VaR}$ is not an admissible risk spectrum, as it violates conditions (ii) and (iii) from Definition 2.6.

[4] Note that I defined $L_t$ as the return loss and not as an absolute loss figure. Consequently, the resulting VaR is also on a return space, i.e. it needs to be multiplied with the corresponding position size to be interpreted as an absolute value in a certain currency.

[5] See for example Basel Committee (n.d.-a) and Basel Committee (n.d.-b).

[6] See chapter 6, formulas (6.2) and (6.3) for the calculation of VaR and ES under the assumption of normally distributed return losses.

(ii) Define $\psi_{ES}(p) := \frac{1}{1-\alpha}\mathbb{1}_{\{\alpha\leq p\leq 1\}}$, then it holds

$$\begin{aligned}\mathcal{M}_{t,\psi_{ES}} &= \int_0^1 \frac{1}{1-\alpha}\mathbb{1}_{\{\alpha\leq p\leq 1\}} VaR_{t,p}\,dp \qquad (2.9)\\ &= \frac{1}{1-\alpha}\int_\alpha^1 VaR_{t,p}\,dp\\ &= ES_{t,\alpha}.\end{aligned}$$

Furthermore, $\psi_{ES}$ fulfils all three properties of an admissible risk spectrum as defined in Definition 2.6.

□

Based on Lemma 2.8, one can rely on general results on the class of spectral risk measures to differentiate between the properties fulfilled by both the VaR and the ES.

### 2.3. Basic properties of Risk Measures

This subsection aims to give an overview on basic properties of risk measures, which are widely accepted throughout the literature. Furthermore, both the VaR and the ES are going to be evaluated based on the introduced properties.
As previously mentioned, Artzner et al. (1999) brought up the term of a *coherent risk measure*, which needs to fulfil a set of mathematical axioms. The notation for the following definition is based on Emmer et al. (2015)[7].

**Definition 2.9** (Coherent Risk Measure)**.** *A risk measure $\rho$ is called a coherent risk measure, if it satisfies the following properties:*

*(i) Homogeneity:*
*A risk measure $\rho$ is homogeneous, if for any return loss variable $L$ and $h\in\mathbb{R}$, $h\geq 0$ it holds that:*

$$\rho(hL) = h\rho(L)\,. \qquad (2.10)$$

*(ii) Monotonicity:*
*A risk measure $\rho$ is monotonic, if for any two return loss variables $L^1$ and $L^2$ it holds that:*

$$L^1\leq L^2 \Rightarrow \rho(L^1)\leq\rho(L^2)\,. \qquad (2.11)$$

*(iii) Translation invariance:*
*A risk measure $\rho$ is translation invariant, if for any return loss variable $L$ and for any $m\in\mathbb{R}$ it holds that:*

$$\rho(L-m) = \rho(L)-m\,. \qquad (2.12)$$

*(iv) Sub-additivity:*
*A risk measure $\rho$ is sub-additive, if for any return loss variables $L^1$ and $L^2$ it holds that:*

$$\rho(L^1+L^2)\leq\rho(L^1)+\rho(L^2) \qquad (2.13)$$

[7] Again different conventions are used for the definition of a coherent risk measure. Nevertheless, the notation used by Emmer et al. (2015) fits best into the previously introduced notation.

Indeed, all four axioms given in Definition 2.9 are reasonable properties risk measures should satisfy. Homogeneity states that the risk inherited in a portfolio should be scalable by the respective size of the portfolio.[8] Secondly, if a portfolio $\mathbb{P}$-almost surely generates a higher loss than another portfolio, i.e. $L^2\geq L^1$, then that portfolio should also be related with a higher riskiness, i.e. $\rho(L^2)\geq\rho(L^1)$, which is fulfilled if the risk measure is monotonic. Furthermore, if the portfolio loss is reduced by adding a risk-free capital amount $m$ to the portfolio, then the risk should decrease by exactly the amount of $m$, which is satisfied if the respective risk measure is translation invariant. Lastly, the risk inherited in one large portfolio consisting of multiple positions, should be at most as large as summing up the risks of the single positions due to diversification effects. This is fulfilled whenever a risk measure $\rho$ is sub-additive.
Acerbi (2002) even goes one step further and includes conditions (i)-(iv) in its general definition of a risk measure. Moreover, he argues that these conditions should be fundamental for every risk measure. For the particular case of both the VaR and the ES, Acerbi (2002) derives an essential result, which relates both the class of spectral and coherent risk measures.

**Theorem 2.10** (Acerbi (2002))**.** *Let $L_t$ be some conditional return loss variable with CDF $F_t$. Define $\mathcal{M}_{t,\psi}$ as*

$$\mathcal{M}_{t,\psi} := \int_0^1 \psi(p)VaR_{t,p}\,dp,$$

*where $\psi\in L^1([0,1])$. Then $\mathcal{M}_{t,\psi}$ is a coherent risk measure if and only if $\psi$ is an admissible risk spectrum according to Definition 2.6.*
*Moreover, a risk measure is coherent if and only if it has a representation as a spectral risk measure.*

*Proof.* See Theorem 2.5 and Theorem 4.1 in Acerbi (2002). □

As the VaR does not have a spectral representation, it is also not a coherent risk measure following Theorem 2.10. On the contrary, the ES does belong to the class of coherent risk measures. More precisely, the VaR fails to satisfy condition (iv) of a coherent risk measure, i.e. it is not sub-additive. One can construct counter-examples with $n\in\mathbb{N}$ loss distributions $L^1,\ldots,L^n$ such that $\rho\left(\sum_{i=1}^n L^i\right) > \sum_{i=1}^n \rho(L^i)$.[9] The lack of sub-additivity might be seen as a drawback of the VaR risk measure in practical applications. Indeed, portfolio diversification is a common method of reducing financial risks. Nevertheless, if the achieved diversification effects are not consistently reflected in the related VaR figures, this might yield wrong incentives for financial institutions.
Emmer et al. (2015) also list another property which appears to be essential for risk measures in practical applications. The *comonotonic additivity* of a risk measure can be seen as a complement to the sub-additivity condition stated in Definition 2.9. The property is based on the following two definitions.

[8] Note that in the presents of liquidity risk for large position sizes one might rather expect $\rho(hL)\geq h\rho(L)$ given that $h\gg 0$.

[9] See Hull (2015) Examples 12.5 and 12.6, for examples on the VaR where the sub-additivity condition is violated.

**Definition 2.11** (Comonotonicity)**.** *Let $X^1$ and $X^2$ be two real valued random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then $X^1$ and $X^2$ are called comonotonic, if there exists a real-valued random variable $\tilde{X}$ and two non-decreasing functions $f^1$ and $f^2$, such that*

$$X^1 = f^1(\tilde{X}) \text{ and } X^2 = f^2(\tilde{X}) \,. \tag{2.14}$$

**Definition 2.12** (Comonotonic Additivity)**.** *Let $L^1$ and $L^2$ be two comonotonic return loss variables, then a risk measure $\rho$ is called comonotonic additive, if it holds that*

$$\rho(L^1 + L^2) = \rho(L^1) + \rho(L^2) \,. \tag{2.15}$$

Indeed, this property has also a comprehensible intuition. If the loss distributions of two portfolios are comonotonic, then their losses are both driven by some common risk factor. Thus, merging both portfolios together should not lead to any diversification benefits as the risk stems from the same risk factor anyway. Actually, both the VaR as well as the ES satisfy the property of comonotonic additivity.[10]

Besides the above listed properties, the aspect of *robustness* is also often discussed in the context of risk measures. In a general sense, a risk measure is said to be robust if small changes in the underlying loss distribution are also related to small changes in the risk measure value. More formal definitions of robustness can for example be found in Cont, Deguest, and Scandolo (2008) or Emmer et al. (2015). Cont et al. (2008) evaluates how additional data points, added to the estimation sample for a return loss variable, effect overall risk measure values. Indeed, they conclude that robustness is not solely a matter of the selected risk measure, but does also depend on the type of implemented estimation model. Furthermore, they outline that the ES is in general less robust compared to the VaR. Nevertheless, as for example noted by Emmer et al. (2015), this is not overly surprising as the ES was introduced due to the insensitivity of the VaR with respect to observations beyond the $\alpha$-quantile. Thus, it is by definition more or less a logical implication that the ES is less robust compared to the VaR. Furthermore, Emmer et al. (2015) argues that in the context of risk management, extreme observations might be less related to measurement errors but rather to actual circumstances, which should indeed be reflected by the respective risk measure. Summing up, the ES is obviously less robust compared to the VaR, at least in the sense of Cont et al. (2008), but it is not clear whether this should be seen as a potential shortcoming of the ES, given that a certain degree of sensitivity towards extreme observations lies within the nature of a risk measure.

The properties presented in this section display just a small selection. Overall, there is far more literature on general properties of risk measures.[11] Nevertheless, as a result of this subsection, it can be seen that the VaR is not a coherent risk measure as it does not fulfil the axiom of sub-additivity. Furthermore, the VaR is only defined as a quantile of the return loss distribution, consequential it completely ignores all losses beyond the selected confidence level.

Concluding this subsection, it appears to be reasonable to favour the ES over the VaR, given that it is a coherent risk measure which furthermore satisfies the property of comonotonic additivity. In addition, also the lack of robustness should, at least per-se, not be seen as a disadvantage of the ES, given the previous argumentation. Nevertheless, for a long time the majority of both academics as well as practitioners expressed their reservations with regard to the ES, as they doubted the ability to backtest the ES. Just in recent time, triggered by the decision of the Basel Committee of Banking Supervision in 2013, researchers started to focus more on practical backtesting solutions instead of pointing towards potential conceptual limitations.

### 2.4. Elicitability

As shortly described in the introduction, the statistical property of elicitability is always closely related to any discussion on the backtestability of the ES. Thus, this subsection gives a brief summary on the elicitability of risk measures or functionals in a more general context.

Originally, the concept of elicitiability was introduced by Osband (1985). Overall, as for example stated by Gneiting (n.d.), elicitability is a useful tool for the evaluation of a series of forecasts based on ex-ante realizations. The concept makes use of a *strictly consistent scoring function S*, which needs to satisfy a set of criteria, and takes both forecasts and realizations as input parameters in order to assign them to a numerical score. A functional, i.e. a risk measure in this case, is said to be elicitable, if a suitable strictly consistent scoring function exists, which fulfils the properties listed in the course of this subsection.

Acerbi and Szekely (2014) briefly explains the main advantage of an elicitable risk measures. Let $\rho$ be some elicitable risk measure with strictly consistent scoring function $S$. Assume a risk manager estimates some return loss variable $\widehat{L}$, which results in a risk forecast $\rho(\widehat{L})$. Furthermore, let $L$ denote the true distribution of return losses. Then the true value of the risk measure, $\rho(L)$, can be determined by

$$\rho(L) = \arg\min_{\rho(\widehat{L})} \mathbb{E}[S(\rho(\widehat{L}), L)]. \tag{2.16}$$

In practice, a risk manager generates a set of forecasts $\{x_t = \rho(\widehat{L}) : t = 1, \ldots, T\}$ and observes a set of realizations $\{l_t, t = 1, \ldots, T\}$, which are based on the true data generating process $L$. Given both forecasts and realizations, one might try to minimize the mean-score,

$$\bar{S} = \frac{1}{T} \sum_{t=1}^{T} S(x_t, l_t), \tag{2.17}$$

in order to find a forecasting distribution $\widehat{L}$ such that the achieved risk estimates are close to the true values, i.e. $\rho(\widehat{L}) \approx \rho(L)$.

In the following a slightly more formal definition of elicitability is presented, based on Emmer et al. (2015) and Nolde and Ziegel (2017). For a more comprehensive and also technical

[10] See for example Emmer et al. (2015) for the definition of Expectiles, a coherent risk measure which does not satisfy comonotonic additivity.

[11] See for example Foellmer and Schied (2002) and its definition of a convex risk measure for another important contribution on risk measure properties.

review on the concept of elicitability see for example Gneiting (n.d.) or Fissler and Ziegel (2016). In line with common practice, $L$ denotes the distribution of return losses and not a random variable for the purpose of this definition. In general, the elicitability of a risk measure always depends on some set of probability distributions, i.e. loss distributions in the setting of this thesis. Denote the set of considered probability distributions by $\mathscr{P}$. Furthermore, similar as in Nolde and Ziegel (2017), let $\Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ be a vector of $k \geq 1$, $k \in \mathbb{N}$ risk measures based on some return loss distribution $L \in \mathscr{P}$.
The following formal definitions of a strictly consistent scoring function an the elicitability of risk measures follow Nolde and Ziegel (2017). Note that for the case of $k = 1$, i.e. $\Theta(L) = \rho^1(L)$ the elicitability of one single risk measure is evaluated, while for the case of $k > 1$ the joint elicitability of a vector of risk measures is considered.

**Definition 2.13** (Consistency)**.** *A scoring function $S : \mathbb{R}^k \times \mathbb{R} \to [0, \infty)$ is called*

*(i) consistent for $\Theta$ with respect to $\mathscr{P}$ if*

$$\mathbb{E}[S(\Theta(L), L)] \leq \mathbb{E}[S(x, L)] \,, \tag{2.18}$$

*for all $x \in \mathbb{R}^k$ with $x = (x^1, \dots, x^k) \neq \Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ and for all loss distributions $L \in \mathscr{P}$.*

*(ii) strictly consistent for $\Theta$ with respect to $\mathscr{P}$ if equation (2.18) holds with strict inequality, i.e.*

$$\mathbb{E}[S(\Theta(L), L)] < \mathbb{E}[S(x, L)] \,, \tag{2.19}$$

*for all $x \in \mathbb{R}^k$ with $x = (x^1, \dots, x^k) \neq \Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ and for all loss distributions $L \in \mathscr{P}$.*

**Definition 2.14** (Elicitability)**.** *A vector $\Theta$ of risk measures is called elicitable with respect to $\mathscr{P}$, if there exists a strictly consistent scoring function $S : \mathbb{R}^k \times \mathbb{R} \to [0, \infty)$ in the sense of Definition 2.13.*

As outlined above, elicitability can be seen as a helpful tool for the purpose of risk management as it yields a criterion for the estimation of an optimal forecast. Indeed, given a strictly consistent scoring function, different estimated loss distributions and the resulting risk forecasts can be compared with respect to the minimization problem stated in formula (2.16).
In the univariate case, i.e. for $k = 1$, the most prominent example might be the mean functional, which is elicitable with respect to all real valued probability distributions with finite second moment. Moreover, the squared error, i.e. $S(x, y) = (x - y)^2$, is a strictly consistent scoring function in this case. Thus, for example the estimation of a mean regression in a standard OLS framework, is achieved by minimizing the mean-squared-error, which corresponds to minimizing formula (2.17) in this particular case.
In order to remain within the scope of this thesis, only the major results regarding the elicitability of different risk measures are cited below. The VaR at confidence level $\alpha$ is elicitable with respect to the class of real-valued Borel-probability distributions, which have a unique $\alpha$-quantile. Moreover, the weighted absolute error function is a strictly consistent scoring function for the VaR, as shown in Thomson (1979) and Saerens (2000). On the contrary, as previously mentioned, Gneiting (n.d.) proved that the ES is not an elicitable risk measure. In more detail, the following proposition was stated by Gneiting (n.d.).

**Proposition 2.15** (Gneiting (n.d.), Theorem 11)**.** *The ES functional is not elicitable relative to any class $\mathscr{P}$ of real-valued probability distributions on the interval $I \subset \mathbb{R}$ that contains the measures with finite support, or the finite mixtures of the absolutely continuous distributions with compact support.*

Thus, whereas the VaR is elicitbable with respect to a reasonable set of probability distributions, the ES, at least standalone, is not.
Nevertheless, for the case of $k = 2$, recent contributions derived, that the vector $\tilde{\Theta} := (VaR_\alpha, ES_\alpha)$ is indeed elicitable following Definition 2.14. In the literature, this is often referred to as the conditional elicitability of the pair of ES and VaR.
The following result, proved by Fissler and Ziegel (2016), is cited below as is fundamental for the backtest by Bayer and Dimitriadis (2019), which will be introduced in the course of this thesis.

**Proposition 2.16** (Nolde and Ziegel (2017))**.** *All scoring functions of the form,*

$$\begin{aligned} S(x_1, x_2, y) \quad &= \mathbb{1}_{\{y > x_1\}} \\ &\quad (-G_1(x_1) + G_1(y) - G_2(x_2)(x_1 - y)) \\ &\quad + (1 - \alpha)(G_1(x_1) - G_2(x_2)(x_2 - x_1) + \mathscr{G}_2(x_2)) \,, \end{aligned} \tag{2.20}$$

*where $G_1$ is an increasing function, $\mathscr{G}_2' = G_2$ and $G_2$ is strictly increasing and strictly concave, are strictly consistent for $\tilde{\Theta} = (VaR_\alpha, ES_\alpha)$, $\alpha \in (0, 1)$, with respect to the class $\mathscr{P}'$. Moreover, $\mathscr{P}'$ are all real-valued probability distributions with finite mean, which have unique $\alpha$-quantiles and $G_1(X)$ is integrable for all random variables $X$ with distribution in $\mathscr{P}'$*

*Proof.* See Corollary 5.5 in Fissler and Ziegel (2016). □

Concluding this subsection, elicitability can be seen as a useful tool for the evaluation of risk measure forecasts. The ES itself does not fulfil this statistical property, whereas the VaR does. Nevertheless, one can find strictly consistent scoring functions for the pair of ES and VaR, or more precisely the ES together with the VaR is conditional elicitable. The consecutive section is going to turn to the main topic of this thesis, the backtesting of risk measures, especially of the ES. Thus, it will also evaluate the concern if the lack of elicitability of the ES is indeed a serious drawback for the development of backtesting approaches.

## 3. Backtesting of Risk Measures

For the estimation of risk measure values, there are two, potentially contrary objectives both regulators and financial institutions are mainly interested in. First of all, as already outlined

above, the estimation of risk figures is closely related to the capital buffer an institution needs to put aside. On the one hand, the regulator needs to assure a sufficient magnitude of capital reserves, such that banks and insurance companies stay solvent in times of financial distress. On the other hand, financial institutions try to keep their capital buffer at the lowest possible level in order to achieve a more efficient capital allocation. Given these two adverse targets, it is essential that risk measure forecasts are sufficiently backtested against actual P&L realizations.
Correspondingly, this section is structured as follows. Subsection 3.1 introduces the general idea of backtesting. Furthermore, I want to introduce a set of rather qualitative aspects, which I believe are of particular interest for any practitioner. Moreover, these aspects are related to potential issues for both the implementation as well as the execution of risk measure backtests in real-world scenarios. Subesection 3.2 gives a short overview of the backtesting framework of VaR forecasts, whereas the subsection thereafter aims to give an overview on the discussion, whether the ES is backtestable despite its lack of elicitability.

### 3.1. Definition and practical aspects

Loosely speaking, as mentioned by Kratz et al. (2018), a backtest is a statistical procedure which compares forecasts of risk measures to actual realizations in order to judge whether the forcasting model is accurate or not.
First of all, in order to prevent any possible confusions, one might further categorize the type of considered forecast. In line with common practice, Emmer et al. (2015) differentiates between three different types, which are listed below.

(i) *Point forecasts* for the value of a random variable or a probability distribution. These are often expressed by an conditional expectation, i.e. $\mathbb{E}[L_t|\mathscr{F}_{t-1}]$, where again $\mathscr{F}_{t-1}$ inherits all information up to time $t-1$.

(ii) *Interval forecasts*, resulting in an interval estimate in which the forecast is expected to lie with some probability level $p$. Both the VaR and the ES do belong to this class of forecasts. As an example, given a loss variable $L$ and the estimated interval $(VaR_\alpha, \infty)$, then the loss is expected to be within this interval with a probability of $1-\alpha$.

(iii) *Forecasts of the entire probability distribution*, like for example an estimate of the CDF or the PDF of a return loss distribution.

Therefore, both VaR and ES forecasts are in most cases categorized as interval forecasts.
For the purpose of backtesing, some notational aspects and assumptions are outlined in the following, which will be applied in the consecutive chapters.

**Notation/Assumptions 3.1.** *The following set of notations and assumptions in the context of backtesting is going to be used in the course of the thesis.*

- *Denote the backtesting horizon, i.e. the number of P&L observations used for the backtest, by $T \in \mathbb{N}$. The Basel Committee suggest a value of $T = 250$ days, which will be taken as a reference.*
- *Denote a vector of realized return losses over the backtesting horizon by $\vec{l} := \{l_t : t = 1,\dots,T\}$, where losses are assumed to be independently but not identically distributed. Correspondingly, the setting allows for time dependent return loss distributions.*
- *For any time $t \in \{1,\dots,T\}$ the realized loss $l_t$ is distributed according to some unknown return loss variable $L_t$, i.e $l_t \sim L_t$. Furthermore, the risk manager forecasts the return loss $\widehat{L}_t$ and the related distribution $F_t(\widehat{L}_t)$ at any point in time within the backtesting horizon. It is assumed, that both $L_t$ and $\widehat{L}_t$ follow a continuous and strictly increasing distribution for any considered time period.*
- *Denote a set of risk measure forecasts over the backtesting horizon by $\{\widehat{\rho}_t : t = 1,\dots,T\}$. For the particular case of the ES and the VaR, the estimated set of forecasts is denoted by $\{\widehat{ES}_{t,\alpha} : t = 1,\dots,T\}$ and $\{\widehat{VaR}_{t,\alpha} : t = 1,\dots,T\}$, respectively.*
- *Assume a risk manager requires a set of $k \in \mathbb{N}$ auxiliary variables at any point in time $t$ within the backtesting horizon, either for the estimation or the backtesting of the respective risk measure. Denote the set of auxiliary variables at time $t$ by, $(a_t^1,\dots,a_t^k) \in \mathbb{R}^k$.*

See Appendix B for a summary of the most important variables used within this thesis.
The following general definition of a backtest is motivated by Bayer and Dimitriadis (2019), nevertheless it is less strict in the sense that is allows for the inclusion of auxiliary variables in the backtest of risk measure forecasts.

**Definition 3.2** (Backtest)**.** *Given the notation introduced above, a backtest of the series of forecasts $\{\hat{\rho}_t : t = 1,\dots,T\}$ for the risk measure $\rho$, given a time series of $k \in \mathbb{N}$ auxiliary variables, based on the realized return loss series $\vec{l}$ is a function*

$$f : \mathbb{R}^T \times \mathbb{R}^{T\times k} \times \mathbb{R}^T \to \{0,1\}\,, \tag{3.1}$$

*which maps the series of forecasts, realized returns and potential auxiliary variables onto a test decisions.*

In an optimal scenario no auxiliary variables are needed in order to test risk measure forecasts on P&L realizations. As an example, all common VaR backtests do not rely on any further input parameters, nevertheless this is different for ES backtests, which often require auxiliary variables like volatility or VaR estimates.
Bayer and Dimitriadis (2019) puts a lot of emphasis on developing a standalone ES backtest, which does not require any additional input parameters. Indeed, this can be seen as a criteria for the conceptual soundness of the underlying backtest, nevertheless my main focus for this thesis is on the practical applicability of backtests for financial institutions. As an example,

the estimation of both ES and VaR is tightly connected. In general, if one estimates the ES, then one can easily also obtain the corresponding VaR forecast without any further assumptions or computational difficulties. On the contrary, a backtest which requires an estimation of the entire loss distribution should be seen critical. For any non-parametric estimation method for the ES, serious additional assumptions would be needed to come up with an explicit estimate for the CDF of the underlying loss variable.
Therefore, I do generally allow for auxiliary variables in ES backtesting approaches, nevertheless I will keep a close eye on the type of auxiliary variable and if this is related with any additional assumptions or effort for the risk manager.
Up to today, there is not one single ES backtest, which can be seen as an industry standard, but, as mentioned before, there is a variety of different approaches. In addition to Definition 3.2, I believe that any backtesting approach should fulfil some further criteria in order to qualify as a standard approach for the financial industry. Therefore, I derived the following set of qualitative properties, which are relevant in real-world applications.

**Proposition 3.3.** *Any backtest for a series of risk measure forecasts should take into account the following qualitative criteria in order to be suitable for the implementation by both regulators and financial institutions.*

*(i) Data intensity:*
*In the best case, a backtest should only require the actual forecasts $\hat{\rho}$ and observed realizations as input parameters. Any additional, auxiliary parameters should be easy to obtain without the need of further assumptions on the estimation model. Furthermore, regulators might have less auxiliary variables at hand as they need to rely on variables reported by financial institutions.*

*(ii) Computational effort:*
*Financial institutions in general need to apply backtests to multiple portfolios on a regular basis. Thus, a backtest which exhibits excessive computational effort might not be feasible for practical implementation. As an example, backtests which are based on a bootstrap procedure might be problematic.*

*(iii) Conceptual ease:*
*In practice, applied backtests need to be communicated with internal management, regulators and other stakeholders. Thus, the applied backtest should be conceptually sound and understandable. Furthermore, a backtest should be flexible, such that it can be aligned with changing internal or regulatory requirements.*

*(iv) Clear decision making:*
*A backtest should yield a clear decision (reject/do not reject) of a risk measure forecast at a given significance level. Optimally, one- and two-sided versions are available. Regulators care more about one-sided tests, which evaluate if the actual risk is underestimated, while from a modelling perspective two-sided tests might be more relevant.*

As a consequence, all implemented backtests are not just evaluated according to classical measures like empirical size or power, but also with respect to the qualitative criteria presented in Proposition 3.3. For the remainder of this thesis, the main focus will be on one-sided risk measure tests. In practice, this is by far the more relevant case, as regulators want to assure that financial institutions do not underestimate their actual risk and thus hold to low capital buffers. Furthermore, regulators do not care about overly conservative risk forecasts by financial institutions.
The consecutive subsection is shortly going to summarize the current backtesting framework for the VaR risk measure, which was brought up by Basel Committee (n.d.-d). Indeed, the backtesting methodology for the VaR can be seen as a starting point for the development of most relevant ES backtests. Thus, the VaR backtest introduced below will also be taken up again in the chapter thereafter, where all selected ES backtesting approaches are presented.

### 3.2. Backtesting Value at Risk

This subsection aims to outline the current industry standard for the backtesting of the VaR, the *traffic light test*, which was introduced Basel Committee (n.d.-d). Furthermore, this subsection presents how this practical approach can be related to more theoretical contributions like Christoffersen (1998). Indeed, the current applied traffic light test is also a one-sided test, which is tailored to regulator's requirements.
Assuming a continuous loss distribution $L_t$, for some $t > 0$, $t \in \mathbb{N}$, it holds that

$$\mathbb{P}(L_t \geq VaR_{\alpha,t}) = 1 - \alpha \,, \tag{3.2}$$

such that the probability of a violation of the VaR is given by $1-\alpha$. As for example proposed by Christoffersen (1998), one can define the *violation indicator* of the VaR at time $t$ as follows.

**Definition 3.4** (Violation indicator)**.** *For some $t \in \{1,\dots,T\}$ and some realized return loss $l_t \sim L_t$, the violation indicator of a VaR estimate depending on the confidence level $\alpha \in (0,1)$, is defined as,*

$$I_t : (0,1) \to \{0,1\} \qquad I_t(\alpha) := \mathbb{1}_{\{l_t \geq \widehat{VaR}_{\alpha,t}\}} = \begin{cases} 0 & \text{if } l_t < \widehat{VaR}_{\alpha,t} \\ 1 & \text{if } l_t \geq \widehat{VaR}_{\alpha,t} \end{cases} . \tag{3.3}$$

Given both the identity in (3.2) and the violation indicator in Definition 3.4, Christoffersen (1998) derived the following two conditions, which need to be satisfied by any series of correct VaR forecasts.

**Proposition 3.5** (VaR - Conditional Coverage)**.** *Consider a series of $T \in \mathbb{N}$ VaR forecasts and the violation indicator as defined in (3.3), then the VaR forecasts are accurate if and only if the following two conditions are fulfilled.*

*(i) Unconditional coverage:*
$\mathbb{E}[I_t(\alpha)] = 1-\alpha$ *for all* $t = 1,\dots,T$.

*(ii) Independence condition:*
$I_t(\alpha)$ *and* $I_s(\alpha)$ *are independent for all* $t \neq s$, *with* $t,s \in \{1,\dots,T\}$

*If both conditions (i) and (ii) are fulfilled, then the VaR forecasts have correct conditional coverage. In this case it holds*

$$I_t(\alpha) \overset{i.i.d}{\sim} Bernoulli(1-\alpha) \quad \text{for all } t \in \{1,\dots,T\}. \tag{3.4}$$

*Moreover, summing up the violation indicator for all t leads to*

$$\sum_{t=1}^{T} I_t(\alpha) \sim Binomial(T, 1-\alpha). \tag{3.5}$$

The Basel Committee prescribes a confidence level of $\alpha = 0.99$ as well as a backtesting period of $T = 250$, which roughly corresponds to one year if daily returns are considered. Thus, the expected number of VaR breaches is given by $\mathbb{E}\left[\sum_{t=1}^{250} I_t(0.99)\right] = 250 \cdot 0.01 = 2.5$.
For the one-sided Basel traffic light test, three colour zones are defined, which correspond to a green, a yellow and a red traffic light. Furthermore, the following set of hypothesis is evaluated in the VaR traffic light test,

$$H0: \quad \sum_{t=1}^{T} I_t(\alpha) \leq 1-\alpha \tag{3.6}$$

$$H1: \quad \sum_{t=1}^{T} I_t(\alpha) > 1-\alpha \tag{3.7}$$

Given that the forecasting model is correctly specified, the total number of violations follows a binomial distribution as outlined in (3.5). Respectively, a series of VaR estimates over a backtesting period of $T$ observation is rejected at a significance level of $\kappa$, whenever $F^{\text{binom}}(\sum_{t=1}^{T} I_t(\alpha), T, 1-\alpha) > 1-\kappa$. One should note that $F^{\text{binom}}(X,T,p)$ denotes the cumulative probability of a value X, related to a binomial distribution with $T$ trials and a success probability of $p$.
For the yellow traffic light a backtesting significance level of $\kappa = 0.05$ is applied, whereas a value of $\kappa = 0.0001$ is considered for the red traffic light. Thus, the *green zone* contains the number of violations, such that the cumulative probability of obtaining at most that many violations is below 95 %. The *yellow zone* contains the number of violations, in case the respective cumulative probability is between 95% and 99.99%. In case the cumulative probability of a certain number of violations is above 99.99%, the forecasting model is ranked in the *red zone*. If the backtest for a VaR estimation model exhibits a yellow or red traffic light, a multiplication factor is added upon the estimated VaR in order to increase the respective capital requirements and thus punish the financial institution, which understates the actual risk.[12]
Given the parameters suggested by Basel Committee (n.d.-d), i.e. $T = 250$ and $\alpha = 0.99$, the resulting traffic light zones for different numbers of violations are depicted in Table 1.

[12] See Basel Committee (n.d.-d) for more details on the procedure.

The traffic light test is a straightforward backtest, which detects financial institutions that underestimate their market risk figures and thus exhibit too many VaR violations. Nevertheless, the traffic light test has one shortcoming, as it only tests for the unconditional coverage of the VaR. Under the null hypothesis, it assumes that the occurrence of violations is independent. More precisely, it only accounts for the number of violations but not on their timing. In practice, if an estimation model is badly specified, violations might be clustered around certain events, which will not be detected by the traffic light test. As an example, Christoffersen (1998) suggests a backtest for conditional coverage of the VaR, which does also account for the independence condition stated in Proposition 3.5. Nevertheless, in practical applications the independence assumption is often checked separately in addition to the proposed traffic light test. Often, this is done by manual inspection of the occurrence of violations over time, as for example stated by Moldenhauer and Pitera (2018). Furthermore, the traffic light test does not need any auxiliary input variables in addition to VaR forecasts and realizations. The methodology is easy to understand and does not require any bootstrap procedure, such that it is also computationally efficient. Lastly, test decisions can easily be obtained at multiple significance levels of the binomial test.
Although the traffic light test exhibits theoretical shortcomings, as it is only an unconditional coverage test for the VaR, it is still the benchmark approach used within the financial industry. Moreover, the backtest satisfies all qualitative criteria outlined in Proposition 3.3. On the contrary, no comparable benchmark approach to backtest the ES, which is widely excepted amongst practitioners, has yet been agreed on.

### 3.3. Backtesting Expected Shortfall and the need for elicitability

The decision to introduce the ES as a primary measure of market risk was mainly based on the arguments outlined in chapter 2. Nevertheless, this inevitably calls for the need of some kind of standard backtesting procedure for the ES. Currently, the Basel Committee suggests to base capital requirements on the ES at level $\alpha = 0.975$, while still backtesting the related VaR figures at confidence levels of $\alpha = 0.99$ and $\alpha = 0.975$ by using the traffic light test outlined in the previous subsection (see Basel Committee (n.d.-b)). As for example argued by Costanzino and Curran (2018), this appears to be fairly insufficient. Furthermore, it reveals the difficulties of finding a suitable methodology, despite the multiple theoretical contributions on that topic within the last years. This subsection aims to give a short overview on the publications related to the backtesting of ES. Therefore, it also enters into the discussion, whether and how backtestability and elicitability should be related to each other.
The decision by the Basel Committee in 2013 to introduce the ES as a primary measure of market risk was often criticized, as Gneiting (n.d.), two years before, derived that the ES is not an elicitable functional. Publications like Chen (2014), thus conclude that the superior theoretical properties of the ES go hand in hand with the inability to derive suitable backtesting alternatives. Nevertheless, in the meantime contributions like Em-

| Basel Traffic Light Test - VaR | | |
|---|---|---|
| **Traffic Light** | **Number of Violations** | **Cumulative Probability** |
| **Green** | 0 | 8.11% |
| | 1 | 28.58% |
| | 2 | 54.32% |
| | 3 | 75.81% |
| | 4 | 89.22% |
| **Yellow** | 5 | 95.88% |
| | 6 | 98.63% |
| | 7 | 99.60% |
| | 8 | 99.89% |
| | 9 | 99.97% |
| **Red** | $> 10$ | $> 99.99\%$ |

**Table 1:** Traffic light zones for the VaR backtest suggested by Basel Committee (n.d.-d) given a backtesting horizon of $T = 250$ observations and a VaR confidence level of $\alpha = 0.99$.
Source: Basel Committee (n.d.-d)

mer et al. (2015) or Acerbi and Szekely (2014) start to prevail, which argue that elicitability should not be connected with the ability to backtest a risk measure. Furthermore, also in earlier years contributions like Kerkhof and Melenberg (2004) argue that backtesting the ES is indeed not any more difficult than backtesting the VaR.

One might take the VaR as an example. As pointed out in subsection 2.4, the VaR is a elicitable risk measure, thus one can compare different forecasting models with respect to the minimization problem stated in formula (2.16). Nevertheless, a backtesting procedure does not compare multiple models, but needs to judge the accuracy of one single approach. Consequently, most VaR backtests are based on the violation indicator defined in Definition 3.4 rather than anything that has to do with elicitability. This is in line with Acerbi and Szekely (2014), which states that the property of elicitability might be useful for model selection, but not for the case of backtesting. Similar as for the VaR, also most suggested approaches to backtest the ES are not based on the concept of elicitability.[13] Therefore, the existence of a variety of potential ES backtests is another indicator that elicitability is not a necessary condition to be able to backtest a risk measure.

Furthermore, I want to add another perspective to that discussion. As for example argued in Cont et al. (2008), the risk estimation procedure might be divided into two single steps. First of all, every estimation model requires either an explicit, via a parametric approach, or an implicit, via a non-parametric approach, estimation of the return loss distribution of the underlying portfolio. Secondly, both the VaR or the ES are only a deterministic mapping from the estimated model into the real-values. If a risk manager does a proper job in modelling the underlying loss distribution, then both ES and VaR figures will be accurate. It might be more difficult to derive an untainted framework for the backtesting of the ES as it is possible for the VaR. Nevertheless, the ES is just a functional of the $\alpha$-tail of return losses. Thus, any methodology to evaluate the appropriateness of the estimated tail distribution can be a reasonable backtesting tool, although it might not be a conceptually ideal backtest for the ES. My point is that just because it might be difficult to backtest the ES directly does not mean one can not judge whether a risk manager does a decent job in estimating ES forecasts. One can take the approach by Kratz et al. (2018), which will also be implemented for the purpose of this thesis, as an example. Kratz et al. (2018) approximates the ES given in Definition 2.5 by a Riemann-sum using VaR forecasts at different confidence levels between $\alpha$ and one. Therefore, Kratz et al. (2018) can rely on existing techniques to backtest the multiple VaR values in order to implicitly evaluate the ES forecast. This displays an indirect approach of backtesting the ES, nevertheless it might still be appropriate to evaluate ES forecasts irrespective the lack of elicitability. Overall, I belief that the lack of elicitiability of the ES should not be seen as an argument against the regulatory decision made by Basel Committee (n.d.-a).

In the following, there are three more comments I want to make on the general classifications of existing ES backtests. First of all, as already argued at the beginning of this chapter, most proposed ES backtests require additional input variables. Thus, existing approaches might be categorized according to the type of auxiliary input variables they require. As an example multiple tests rely on the VaR (Acerbi and Szekely (2014), Kratz et al. (2018)), the volatility (McNeil and Frey (2000), Nolde and Ziegel (2017)), the cumulative violation process $\int_\alpha^1 I_t(p)dp$ (Costanzino and Curran (2015), Du and Escanciano (2017)) or even the whole return loss distribution (Berkowitz (2001), Kerkhof and Melenberg (2004), Wong (2008)) as an input parameter in addition to the ES. As argued above, from a practical perspective this might be seen more or less problematic depending on the type of auxiliary variable.

[13] The ES backtests proposed by Fissler and Ziegel (2016) and Bayer and Dimitriadis (2019) can be seen as an exception as they are both based on the concept of conditional elicitability, previously introduced. Indeed, the later one will be implemented in the following chapter for the purpose of this thesis.

Secondly, there is one strand in the ES backtesting literature, which does not focus on point or interval forecast of risk measures, but rather on the backtesting of the entire estimated return loss distribution. Most of these approaches are based on *realized p-values*, which estimate the probability of observing ex-post losses based on the predicted return loss distribution. Diebold, Gunther, and Tay (1998) firstly introduced the idea of evaluating density forecasts based on realized p-values. Afterwards, both Berkowitz (2001) as well as Kerkhof and Melenberg (2004) derived ES backtests based on the concept of realized p-values. In the light of my previous argumentation, both approaches also avoid any difficulties of directly backtesting the ES. Nevertheless, both tests require an explicit estimate for the entire loss distribution, such that they are not feasible in many real-world applications. Thus, these backtests are not going to be considered in the course of this thesis.
As a third aspect, one can again differentiate between unconditional and conditional coverage tests for the ES. Given the usual parameter value of $\alpha = 0.975$, there are in general three aspects which need to be considered for a ES backtest. First of all, the number of violations exceeding the $VaR_{0.975}$ threshold, secondly the magnitude of any violation and lastly again the independence of violations beyond the $VaR_{0.975}$ value. A conditional coverage test needs to satisfy all three criteria, while an unconditional coverage test of the ES only takes into account the first two criteria. Du and Escanciano (2017) is, up to my knowledge, the only conditional coverage test for the ES that has been suggested up to now, all other approaches do not take into account the independence of VaR violations.
The following chapter is going to turn to the concrete ES backtesting approaches, which are considered within this thesis.

## 4. Evaluated backtests for the Expected Shortfall

This chapter presents all five backtests which are implemented in *Python* for the purpose of this thesis, as well as any adjustments which I made compared to the original approaches. As outlined in the previous chapter, there is a variety of backtests for the ES, which have been suggested within the last two decades. Although any choice of five different backtests is somehow arbitrary up to a certain extend, I tried to select five approaches with respect to the following objective.
*The main objective of this thesis is to find an one-sided, unconditional coverage, backtest for the Expected Shortfall (ES) at the confidence level of $\alpha = 0.975$, which should be practically applicable with respect to the criteria stated in Proposition 3.3.*

- First of all, I selected the *multinomial backtest* from Kratz et al. (2018), as it appears to be a natural extension to the binomial test for the VaR.
- Secondly, I chose the so called *"Test 2"* from Acerbi and Szekely (2014), which is probably the most prominent ES backtest and often used as a benchmark in other recent contributions.
- The third approach I selected is the *intercept ES regression (ESR) backtest* from Bayer and Dimitriadis (2019), as it takes a novel view on the ES backtesting by introducing a regression framework. Furthermore, it is the only approach presented within this thesis which is somehow related to the concept of elicitability.
- As a fourth approach the *Z-test* from Costanzino and Curran (2015) is implemented, which exploits nice distributional properties and can easily be generalized to backtest any spectral risk measure.
- The last approach I selected, is up to my knowledge the first ES backtest that has been suggested amongst researchers. Nevertheless, the *residuals bootstrap test* from McNeil and Frey (2000) is still widely used given its very intuitive concept.

### 4.1. Multinomial backtest from Kratz et al. (2018)

This section is first going to introduce the *multinomial backtest* brought up by Kratz et al. (2018). Whereas the original version is formulated as a two-sided test, I will propose some slight adjustments to the test decision in order to make it a 'de-facto' one-sided test, in line with the objective stated above.

#### 4.1.1. Original approach - Kratz et al. (2018)

The idea of the multinomial backtest by Kratz et al. (2018) is relatively straightforward. As shortly mentioned in the previous chapter, one might approximate the ES by a Riemann-sum like for example in the following way,

$$ES_{t,\alpha}(L_t) \approx \frac{1}{4}[VaR_{t,\alpha} + VaR_{t,0.75\alpha+0.25} + VaR_{t,0.5\alpha+0.5} + VaR_{t,0.25\alpha+0.75}]\,. \tag{4.1}$$

Kratz et al. (2018) suggests, to implicitly backtest the ES by deriving a multinomial test for the $N = 4$ VaR figures. In order to increase the approximation accuracy, one might for example also choose $N = 8$ or $N = 16$ VaR values, i.e. quantiles of the distribution of $L_t$.
In the following, some notational aspects are fixed. Given $N \in \mathbb{N}$ VaR levels which are considered for the approximation of the ES in (4.1), the respective VaR confidence levels are defined as,

$$\alpha_j := \alpha + \frac{j-1}{N}(1-\alpha), \quad j = 1,\dots,N\,, \tag{4.2}$$

where $\alpha$ is some reference confidence level, like for example 0.975. Due to technical reasons, Kratz et al. (2018) sets $\alpha_0 := 0$ and $\alpha_{N+1} := 1$.
For any backtesting horizon $T \in \mathbb{N}$, if the respective series of VaR estimates at confidence level $\alpha_j$ for $j \in \{1,\dots,N\}$ has correct unconditional coverage, then according to Proposition 3.5 it holds

$$\sum_{t=1}^{T} I_t(\alpha_j) \sim Binomial(T, 1-\alpha_j). \tag{4.3}$$

Therefore, one can simultaneously test VaR estimates at all $N$ considered confidence levels by employing a multinomial distribution. The multinomial distribution is denoted by $MN(n,(p_0,\dots,p_N))$, where each of the $n$ trials results in

$N+1$ outcomes distributed according to the vector of success probabilities $(p_0,\dots,p_N)$.
Therefore, Kratz et al. (2018) defines the series $\{X_t\}_{t=1,\dots,T}$, by

$$X_t := \sum_{j=1}^{N} I_t(\alpha_j). \tag{4.4}$$

Moreover, $X_t$ counts the number of breached VaR levels at time $t$. Furthermore, Kratz et al. (2018) uses a slight adjustment in order to consider all observations $t$ within the backtesting horizon simultaneously. They define the so called *cell counts* by,

$$O_j = \sum_{t=1}^{T} \mathbb{1}_{\{X_t=j\}} \quad \text{for all } j=0,\dots,N\,. \tag{4.5}$$

One should note, that any violation of $VaR_{t,\alpha_j}$ is automatically also a violation of $VaR_{t,\alpha_{j-1}}$, as by definition $VaR_{t,\alpha_j} \geq VaR_{t,\alpha_{j-1}}$. Thus, $O_j$ counts the number of observations over the backtesting horizon which breach the first $j$ VaR thresholds up to the confidence level $\alpha_j$, but do not breach the threshold at confidence level $\alpha_{j+1}$. Given correct unconditional coverage of the VaR at all considered confidence levels $\alpha_j$, the random vector $(O_0,\dots,O_N)$ is thus distributed according to the following multinomial distribution,

$$(O_0,\dots,O_N) \sim MN(T,(\alpha_1-\alpha_0,\dots,\alpha_{N+1}-\alpha_N))\,. \tag{4.6}$$

Consider the general case, where $(O_0,\dots,O_N) \sim MN(T,(\theta_1-\theta_0,\dots,\theta_{N+1}-\theta_N))$, for some arbitrary parameters $0=\theta_0<\theta_1<\ldots<\theta_N<\theta_{N+1}=1$. The formal null and alternative hypothesis according to Kratz et al. (2018) are then given by

$$\begin{aligned} H0: &\quad \theta_j = \alpha_j \quad \text{for all } j\in\{1,\dots,N\} \\ H1: &\quad \theta_j \neq \alpha_j \quad \text{for at least one } j\in\{1,\dots,N\}\,. \end{aligned} \tag{4.7}$$

Indeed, Kratz et al. (2018) evaluates three different multinomial tests for multiple amounts of VaR thresholds $N$. Following their conclusion, a $\chi^2$ test based on Nass (1959) and a value of $N=8$ VaR approximation levels yields the best results and will be presented below.
Moreover, the *Nass test* test proposed by Nass (1959) is an adjustment of the standard *Pearson* $\chi^2$ *test* introduced by Pearson (1900). As noted by Kratz et al. (2018), the Nass test is superior if cell probabilities are low, which is also the case in the considered scenario.
The test is based on the test statistic of a standard Pearson $\chi^2$ test, depending on the choice of considered VaR levels $N$ and the observed cell counts $O_j$,

$$Z_N := \sum_{j=0}^{N} \frac{(O_j - T(\alpha_{j+1}-\alpha_j))^2}{T(\alpha_{j+1}-\alpha_j)} \overset{d}{\underset{H0}{\sim}} \chi^2_N\,. \tag{4.8}$$

The Nass test consecutively uses an adjustment factor $c$ in the following way,

$$c\cdot Z_N \overset{d}{\underset{H0}{\sim}} \chi^2_\nu, \quad \text{with } c := \frac{2\mathbb{E}[Z_N]}{Var(Z_N)} \text{ and } \nu := c\mathbb{E}[Z_N]\,, \tag{4.9}$$

where $\mathbb{E}[Z_N]=N$ and $Var(Z_N)=2N-\frac{N^2+4N+1}{T}+\frac{1}{T}\sum_{j=0}^{N}\frac{1}{\alpha_{j+1}-\alpha_j}$.
[14] Given a significance level of $\kappa$ for the backtest, the null hypothesis of the two-sided test version by Kratz et al. (2018), given in (4.7), is rejected whenever $c\cdot Z_N > \chi^2_\nu(1-\kappa)$.
Given the two-sided hypothesis stated in (4.7), the multinomial ES backtest proposed by Kratz et al. (2018) does not only reject estimation models where the true risk is underestimated, but also overly conservative forecasting models. This is problematic with respect to the objective of this thesis stated above. Indeed, regulators only want to punish the underestimation of the actual risk, but not any conservative estimation approach.
For a potential example of a conservative ES estimation model, which would be rejected in the two-sided multinomial ES backtest, one might consider the following rather extreme scenario. Given a backtesting period of $T=250$ observations, the multinomial ES backtest would reject an estimation model where non of the realized return losses $l_t$ breaches any of the estimated $N$ VaR thresholds, i.e.

$$\begin{aligned} l_t < \widehat{VaR}_{t,\alpha_j} \qquad & \text{for all } t=1,\dots,T \\ & \text{and for all } j=1,\dots,N. \end{aligned} \tag{4.10}$$

Correspondingly, one obtains cell counts of $O_0=250$ and $O_j=0$ for all $j=1,\dots,N$, which leads to a rejection in the two-sided multinomial backtest.
Nevertheless, by design the two-sided version proposed by Kratz et al. (2018) only has few room to reject conservative ES estimation models. Thus, on the conservative side, the multinomial ES backtest only detects cases where a risk manager extremely overestimated the underlying risk. Especially, for rather short backtesting periods like $T=250$, the multinomial backtest will only reject extreme examples, like the one depicted above. Nevertheless, I want to add a slight adjustment to the test decision in order to make it indeed a de-facto one-sided test for any reasonable number of backtesting observations $T$.

#### 4.1.2. Adjusted approach - De-facto one-sided test

In order to be able to compare different backtests with respect to the objective stated above, I want to propose a de-facto one sided test, with the following null and alternative hypothesis,

$$\begin{aligned} H0: &\quad \theta_j \geq \alpha_j \quad \text{for all } j\in\{1,\dots,N\} \\ H1: &\quad \theta_j < \alpha_j \quad \text{for at least one } j\in\{1,\dots,N\}\,. \end{aligned} \tag{4.11}$$

For the de-facto one sided test I add an additional criteria in the form of a *conservatism indicator* to the test decision. The conservatism indicator is defined as

$$\mathbb{1}_{\text{conservatism}} := \begin{cases} 0 & \text{if } \exists j\in\{1,\dots,N\}: \quad \sum_{t=1}^{T} I_t(\alpha_j) > 1-\alpha_j \\ 1 & \text{if } \forall j\in\{1,\dots,N\}: \quad \sum_{t=1}^{T} I_t(\alpha_j) \leq 1-\alpha_j \end{cases}, \tag{4.12}$$

where $I_t(\alpha)$ is again the violation indicator as defined in (3.2). Moreover, the conservatism indicator is equal to one, if for all considered $N$ VaR levels the number of observed violations is

[14] See Nass (1959) for more details on the methodology of the proposed test.

smaller or equal than the expected number of violations, given the model is correctly specified. In this case the ES estimation model is definitely on the conservative side and should therefore not be rejected in a one-sided test. Accordingly, the null hypothesis in (4.11) is rejected if $c \cdot Z_N > \chi^2_\nu(1-\kappa)$ and $\mathbb{1}_{\text{conservatism}} = 0$.

As previously outlined, only extremely conservative ES estimation models are rejected by the two-sided approach and this only in case a rather long backtesting period is considered. The proposed adjustment of the test decision additionally also rules out the rejection of those, obviously too conservative, models. Therefore, it turns the two-sided test proposed by Kratz et al. (2018) into a de-facto one-sided version at least for any reasonable backtesting horizon and all relevant significance levels of the backtest, like $\kappa = 0.05$ or $\kappa = 0.0001$.

Definitely, the proposed version might not be a conceptual ideal backtest, in the sense that the stipulated significance level $\kappa$ might not perfectly coincide with the actual significance level of the one-sided test. Nevertheless, given the argumentation above, the impact of the proposed adjustment is expected to be rather marginal. More importantly, the adjusted multinomial backtest can also be compared to the other one-sided ES backtests, which are presented in the following subsections. With respect to the criteria stated in Proposition 3.2, the backtest requires the vector of VaR estimates, $(\widehat{VaR}_{t,\alpha_1}, \ldots, \widehat{VaR}_{t,\alpha_N})$, and P&L realizations as input variables. Although it does not backtest the ES directly, the related VaR figures in general can be easily obtained. Furthermore, given the multinomial distribution of the test statistic, no bootstrap procedure is required and the methodology is an intuitive generalization of the binomial test used for the VaR. Although, there is no straightforward one-sided version of the original test methodology proposed by Kratz et al. (2018), the backtest might still be highly relevant for practical applications given its easy concept and the requirement of both few input parameters and few computational effort.

### 4.2. "Test 2" from Acerbi and Szekely (2014)

The second test evaluated within this thesis stems from Acerbi and Szekely (2014) and is often referred to as *"Test 2"* within related contributions. Indeed, it is the second out of three backtests proposed by Acerbi and Szekely (2014). This subsection aims to introduce the test methodology and shortly discuss the need of a bootstrap algorithm.

The backtest suggested by Acerbi and Szekely (2014) is based on the following unconditional expectation

$$ES_{t,\alpha}(L_t) = \mathbb{E}\left[\frac{L_t I_t(\alpha)}{1-\alpha}\right], \tag{4.13}$$

which is a correct specification of the ES, as the loss distribution $L_t$ is assumed to be continuous. Depending on the vector of observed losses, $\vec{l}$, Acerbi and Szekely (2014)[15] defines the test statistic,

$$Z := \left(\sum_{t=1}^{T} \frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right) - 1\,. \tag{4.14}$$

Indeed, Acerbi and Szekely (2014) proposes a one-sided test, which ought to detect whether the estimated risk, $\widehat{ES}_{t,\alpha}$, understates the actual risk given by $ES_{t,\alpha}$. With respect to the test statistic $Z$ defined in (4.14), the following expectations can be derived, under the assumptions that the ES estimation model is firstly correctly specified or secondly underestimates the actual risk.

**Proposition 4.1** (Acerbi and Szekely (2014), Proposition A.3)**.** *Given the test statistic $Z$ defined in equation (4.14) it holds,*

*(i)* $\mathbb{E}[Z] = 0$*, given that the ES estimation model is correctly specified, and*

*(ii)* $\mathbb{E}[Z] > 0$*, given that the ES estimation model underestimates the actual underlying risk.*

*Proof.* (i) Under the assumption that the estimate $\widehat{ES}_{t,\alpha}$ is correctly specified for all $t$ within the backtesting horizon, the identity (4.13) yields,

$$\begin{aligned}
\widehat{ES}_{t,\alpha} &= \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\right] \\
&\Leftrightarrow \quad \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] = 1 \\
&\Leftrightarrow \quad \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] - 1 = 0\,,
\end{aligned} \tag{4.15}$$

for all $t \in \{1, \ldots, T\}$. Furthermore, for the test statistic $Z$ it holds,

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\left(\sum_{t=1}^{T} \frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right) - 1\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] - 1 \\
&\overset{(4.15)}{=} 0\,.
\end{aligned} \tag{4.16}$$

(ii) Given that the estimated ES model underestimates the true risk, the actual ES values $ES_{t,\alpha}$ are larger or equal to the estimated risk figures $\widehat{ES}_{t,\alpha}$. Furthermore, there exists some $t$ within the backtesting horizon, such that $ES_{t,\alpha} > \widehat{ES}_{t,\alpha}$. Therefore under the assumption of risk

[15] Note that the notation is slightly different in this thesis compared to Acerbi and Szekely (2014) given the different sign conventions used for the definition of the ES.

underestimation it holds,

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}\left[\left(\sum_{t=1}^{T}\frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right)-1\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\right]}_{=ES_{t,\alpha}}\frac{1}{\widehat{ES}_{t,\alpha}}-1 \\ &= \frac{1}{T}\sum_{t=0}^{T}\underbrace{\frac{ES_{t,\alpha}}{\widehat{ES}_{t,\alpha}}}_{\geq 1}-1 \\ &> 0\,, \end{aligned} \tag{4.17}$$

whereas the last inequality holds under the assumption that the ES model underestimates the actual risk.

□

Given the expectations derived in Proposition 4.1, one can derive the following null and alternative hypothesis for the Test 2 inspired by Acerbi and Szekely (2014)

$$\begin{aligned} H0: &\quad Z \leq 0 \\ H1: &\quad Z > 0\,. \end{aligned} \tag{4.18}$$

Indeed, the true distribution of the test statistic under the null hypothesis is unknown, thus one can not exploit any distributional properties to decide whether a realization of the test statistic is indeed significantly larger than zero. Acerbi and Szekely (2014) suggests the use of a simulation to determine the the p-value of the backtest related to some vector of realizations $\vec{l}$. Therefore, one needs to store all estimated return loss variables $\widehat{L}_t$ in order to simulate the distribution of $Z$ under the null hypothesis in the following way.

- Calculate the value of the test statistic related to the observed losses, i.e. $Z(\vec{l})$.
- Simulate $M$ independent bootstrap trials for every point in time within the backtesting horizon from the estimated loss distributions, i.e simulate $l_t^i \sim \widehat{L}_t$ for $i = 1,\ldots,M$ and $t = 1,\ldots,T$.
- Calculate the value of the test statistic related to any bootstrap trial, i.e. $Z^i = Z(\vec{l^i})$, where $\vec{l^i} = \{l_1^i,\ldots,l_T^i\}$ for $i = 1,\ldots,M$.
- Estimate the p-value related to the vector of observed losses $\vec{l}$ given by,

$$p = \frac{1}{M}\sum_{j=1}^{M}\mathbb{1}_{\{Z^i > Z(\vec{l})\}}\,. \tag{4.19}$$

Given the estimated p-value and a significance level $\kappa$ for the backtest, the null hypothesis is rejected whenever $p < \kappa$.
The simulation procedure suggested by Acerbi and Szekely (2014) is necessary as the distribution of the test statistic under the null hypothesis is unknown. Moreover, the distribution in general also depends on the fitted model for the underlying portfolio, i.e. on the estimation of $\widehat{L}_t$ and $\widehat{ES}_{t,\alpha}$, respectively. This might be seen as a drawback of the outlined backtest, with respect to the criteria stated in Proposition 3.3, for two particular reasons. First of all, the bootstrap procedure is related to an increased computational effort. Secondly, the simulation requires the risk manager to store the estimated return loss $\widehat{L}_t$ for all observations within the backtesting horizon.
One reason why the test outlined above is still widely applied, is that Acerbi and Szekely (2014) recognizes that critical values for the test decision are indeed stable for different applied estimation models. Acerbi and Szekely (2014) analyses the distribution of the test statistic given that the estimated loss $\widehat{L}_t$ follows a standard normal distribution as well as multiple t-distributions with varying degrees of freedom. Given the backtest significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$, stipulated by the Basel Committee, Acerbi and Szekely (2014) reports the following critical values based on their simulation study, as outlined in Table 2.
Indeed, critical values are comparable for estimated loss distributions which follow a standard normal or a t-distributions with at least $\nu = 5$ degrees of freedom. Acerbi and Szekely (2014) argues that a t-distribution with $\nu = 3$ degrees of freedom corresponds to a extremely heavy tailed loss distribution which is rather uncommon for actual portfolio losses. Furthermore, they argue that the backtest would be more penalizing in that case and thus still reject any model which underestimates the actual risk. Overall, they conclude that critical values for the backtesting significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$ are fairly stable at values of 0.7 and 1.8, respectively. Following their argumentation, it is therefore not necessary to save estimated return losses $\widehat{L}_t$ and apply a bootstrap, as realized test statistic values can immediately be compared to the proposed critical values.
In order to complement the analysis by Acerbi and Szekely (2014) on the robustness of critical values, I fitted two different ES estimation models on log-return losses of the S&P 500 and simulated the respective distributions of the test statistic under the null hypothesis. The simulation is based on the bootstrap procedure outlined above and $M = 10000$ simulation trials. The respective analysis is deferred to appendix A attached to this thesis.
Based on the simulated critical values depicted in Table 2 as well as on my analysis results outlined in appendix A, I believe that critical values are reasonably stable at a backtesting confidence level of $\kappa = 0.05$. Nevertheless, for more extreme quantiles of the test statistic distribution, like example $\kappa = 0.0001$, the usage of fixed critical values might lead to inaccurate test decisions.
Concluding, depending on the underlying portfolio and the type of estimation model a risk manager who wants to apply the Test 2 from Acerbi and Szekely (2014) faces the following trade-off. On the one hand, using fixed critical values leads to a superior computational performance and fewer input variables on the other hand the abandonment of a bootstrap procedure might lead to imprecise decisions in certain situations. Taking into

| | **Critical values for different backtest significance levels** | |
|---|---|---|
| Estimated return loss distribution | 5 % | 0.01 % |
| $t_3$ | 0.82 | 4.4 |
| $t_5$ | 0.74 | 2.0 |
| $t_{10}$ | 0.71 | 1.9 |
| $t_{100}$ | 0.70 | 1.8 |
| standard Normal | 0.70 | 1.8 |

**Table 2:** Simulation based critical values of the test statistic $Z$ at both backtest significance levels required by the Basel Committee, for estimation models based on the standard normal distribution and multiple standard t-distributions with varying degrees of freedom, $t_\nu$.

Source: Acerbi and Szekely (2014)

account this potential trade-off, I decide to apply Test 2 proposed by Acerbi and Szekely (2014) with a fixed critical value of $Z^* = 0.70$ given a backtesting significance level of $\kappa = 0.05$. In scenarios with a backtesting significance level of $\kappa = 0.0001$, which corresponds to the red zone in the VaR traffic light test, I decide to rely on a bootstrap procedure, as critical values show a high divergence for this rather extreme quantile of simulated test statistic distributions. Moreover, any bootstrap test decision of the Test 2 will be based on $M = 1000$ simulation trials in the consecutive chapters.

### 4.3. Intercept ES regression backtest from Bayer and Dimitriadis (2019)

This subsection presents an adjusted version of the ES backtest derived in Bayer and Dimitriadis (2019). Their approach is based on a regression framework, which takes a novel view on the backtesting issue of the ES. Indeed, the approach is based on the results regarding the conditional elicitability of ES, derived by Fissler and Ziegel (2016) and outlined in Proposition 2.16 within this thesis. Similar to the idea of fitting a quantile regression for the VaR, Bayer and Dimitriadis (2019) proposes a framework to estimate a regression for the ES of return losses distributed by $L_t$ on estimated ES figures $\widehat{ES}_{t,\alpha}$ as explanatory variables. Moreover, taking into account the different sign conventions within this thesis, they propose to estimate the following regression,

$$-ES_{t,\alpha} = -ES_\alpha(L_t|\mathscr{F}_{t-1}) = \gamma_0 - \gamma_1 \widehat{ES}_{t,\alpha} \,. \tag{4.20}$$

Bayer and Dimitriadis (2019) outlines two different test specifications a two-sided version, which is labelled as the *ESR (ES regression) backtetst* and a one-sided *intercept ESR backtest*. In the following, an adjusted version of the one-sided intercept ESR backtest will be presented. Therefore, Bayer and Dimitriadis (2019) sets the slope parameter in (4.20) equal to one and obtains

$$-ES_{t,\alpha} = -ES_\alpha(L_t|\mathscr{F}_{t-1}) = \gamma_0 - \widehat{ES}_{t,\alpha} \,. \tag{4.21}$$

Given the ES forecasting model is perfectly accurate, i.e. $ES_{t,\alpha} = \widehat{ES}_{t,\alpha}$ the estimated intercept parameter $\gamma_0$ will equal to 0. In case of an overly conservative forecasting model, one will obtain $\gamma_0 > 0$. On the contrary, an estimation model which understates the actual risk leads to a situation where $\widehat{ES}_{t,\alpha} < ES_{t,\alpha}$ for a reasonable number of observations within the backtesting period, such that an estimated intercept of $\gamma_0 < 0$ is to be expected. Therefore, formally the following one-sided null and alternative hypotheses are going to be evaluated

$$\begin{aligned} H0: &\quad \gamma_0 \geq 0 \\ H1: &\quad \gamma_0 < 0 \,. \end{aligned} \tag{4.22}$$

#### 4.3.1. Regression estimation based on conditional elicitability

As outlined by Bayer and Dimitriadis (2019), one of the main difficulties is to consistently estimate a regression for the ES of a series of return losses $l_t \sim L_t$ on ES forecasts $\widehat{ES}_{t,\alpha}$. As mentioned in Proposition 2.15 within this thesis, the ES itself is not an elicitable functional, thus there exists no strictly consistent scoring function in the sense of Definition 2.13. Therefore, there is no potential objective function for any *Maximum-Likelihood (ML) estimation* procedure, in order to directly estimate $\gamma_0$ in regression (4.21).

The solution proposed by Bayer and Dimitriadis (2019) is to exploit the conditional elicitability of the vector $(VaR_{t,\alpha}, ES_{t,\alpha})$ in order to simultaniously estimate two regression equations for the quantile and for the ES of the return losses $l_t \sim L_t$. Formally, the regression system is given by,

$$\begin{aligned} -l_t &= \beta_0 - \widehat{VaR}_{t,\alpha} + \varepsilon_t^q \,, \\ -l_t &= \gamma_0 - \widehat{ES}_{t,\alpha} + \varepsilon_t^e \,, \end{aligned} \tag{4.23}$$

where $q_\alpha(\varepsilon_t^q|\mathscr{F}_{t-1}) = 0$ and $ES_\alpha(\varepsilon_t^e|\mathscr{F}_{t-1}) = 0$ almost surely. As conditional VaR and ES forecasts are considered at any time $t$ based on the $\sigma$-Algebra $\mathscr{F}_{t-1}$, the conditions proposed to the error terms are equivalent to,

$$\begin{aligned} -VaR_\alpha(L_t|\mathscr{F}_{t-1}) &= \beta_0 - \widehat{VaR}_{t,\alpha} \,, \\ -ES_\alpha(L_t|\mathscr{F}_{t-1}) &= \gamma_0 - \widehat{ES}_{t,\alpha} \,. \end{aligned} \tag{4.24}$$

Consequently, one can use realized return losses $l_t$ as well as $\widehat{VaR}_{t,\alpha}$ and $\widehat{ES}_{t,\alpha}$ estimates over the backtesting horizon in order to fit the regression system (4.23) by making use of a suitable strictly consistent scoring function as outlined in Proposition 2.16.

Indeed, Bayer and Dimitriadis (2019) puts a major focus on developing a standalone backtest for the ES. Therefore, they replace $\widehat{VaR}_{t,\alpha}$ in the regression system (4.23) by forecasts of the ES, $\widehat{ES}_{t,\alpha}$. As a consequence, they only require ES forecasts and realized losses as an input, but end up with a certain degree of model misspecification.[16] Nevertheless, the main objective of this thesis is to develop a practically applicable backtest. As previously argued, adding VaR forecasts as additional input parameters for the backtest is not considered as a major drawback in this regard. Thus, in order to simplify the methodology of the applied backtest I propose to estimate the regression system as outlined in (4.23) and therefore deviate from the approach suggested by Bayer and Dimitriadis (2019). For the regression estimation, the following strictly consistent scoring function for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$ is considered in line with Bayer and Dimitriadis (2019).[17]

**Proposition 4.2.** *Let $t$ be any observation within the backtesting horizon, i.e. $t \in \{1,\dots,T\}$. Denote the VaR and ES forecast at time $t$ by $\widehat{VaR}_{t,\alpha}$ and $\widehat{ES}_{t,\alpha}$, respectively. Furthermore, denote the realized return loss by $l_t$ and define the vector of regression parameters by $\theta := (\beta_0, \gamma_0)$. Then the scoring function $S$ at time $t$ defined through,*

$$\begin{aligned} &S(\widehat{VaR}_{t,\alpha}, \widehat{ES}_{t,\alpha}, l_t, \theta) := \\ &\frac{1}{\gamma_0 - \widehat{ES}_{t,\alpha}} \cdot \Bigg( (\gamma_0 - \widehat{ES}_{t,\alpha}) - (\beta_0 - \widehat{VaR}_{t,\alpha}) \\ &+ \frac{\mathbb{1}_{\{l_t > \beta_0 - \widehat{VaR}_{t,\alpha}\}} \cdot (\beta_0 - \widehat{VaR}_{t,\alpha} - l_t)}{1-\alpha} \Bigg) \\ &- \log(-(\gamma_0 - \widehat{ES}_{t,\alpha})) \,, \end{aligned} \tag{4.25}$$

*is strictly consistent for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$ in the sense of Definition 2.13.*

*Proof.* Using the notation of Proposition 2.16, define

$$\begin{aligned} x_{1,t} &:= (\gamma_0 - \widehat{ES}_{t,\alpha}) \,, \\ x_{2,t} &:= (\beta_0 - \widehat{VaR}_{t,\alpha}) \,, \\ y_t &:= l_t \,, \\ G_1(x) &:= 0 \,, \\ \mathscr{G}_2(x) &:= -\log(-x) \,, \end{aligned}$$

for $t$ within the backtesting horizon. Furthermore, it holds $G_2(x) = \mathscr{G}_2'(x) = -1/x$, such that $G_1$ is an increasing function and $G_2$ is both strictly concave and strictly increasing. Moreover, according to the result of Proposition 2.16, the following scoring function is strictly consistent for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$.

$$\begin{aligned} &S(x_{1,t}, x_{2,t}, y_t) = \\ &\mathbb{1}_{\{y_t > x_{1,t}\}}(-G_1(x_{1,t}) + G_1(y_t) - G_2(x_{2,t})(x_{1,t} - y_t)) \\ &+(1-\alpha)(G_1(x_{1,t}) - G_2(x_{2,t})(x_{2,t} - x_{1,t})) + \mathscr{G}_2(x_{2,t}) \\ &= \mathbb{1}_{\{y_t > x_{1,t}\}} \left( -\frac{1}{-x_{2,t}} \cdot (x_{1,t} - y_t) \right) \\ &+(1-\alpha) \left( -\frac{1}{-x_{2,t}} (x_{2,t} - x_{1,t}) \right) - \log(-(x_{2,t})) \\ &= \frac{1}{x_{2,t}} \left( (x_{2,t} - x_{1,t}) + \frac{\mathbb{1}_{\{y_t > x_{1,t}\}}(x_{1,t} - y_t)}{1-\alpha} \right) \\ &- \log(-x_{2,t}) \end{aligned} \tag{4.26}$$

Inserting $x_{1,t}$, $x_{2,t}$ and $y_t$ into (4.26) yields the desired scoring function outlined in (4.25). □

Following the estimation procedure described by Bayer and Dimitriadis (2019), the ML-estimator of the regression parameter $\theta = (\beta_0, \gamma_0)$ related to the regression system (4.23) is given by,

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} -S(\widehat{VaR}_{t,\alpha}, \widehat{ES}_{t,\alpha}, l_t, \theta) \,, \tag{4.27}$$

where the strictly consistent scoring function $S$ defined in formula (4.25) is used.[18] Following the regression estimation, one can test the resulting parameter value $\gamma_0$ with respect to the hypotheses stated in (4.22).

### 4.3.2. Bootstrap test decision

Bayer and Dimitriadis (2019) suggests to apply a Wald-type test statistic based on some consistent covariance estimator $\hat{\Omega}$. Both Bayer and Dimitriadis (2019) and Dimitriadis and Bayer (2019) can be taken as a reference for the methodology and the implementation of a consistent asymptotic covariance estimator $\hat{\Omega}$, which is applied for their proposed test decision. Nevertheless, both the theory behind the estimator itself as well as the implementation procedure is highly advanced and might thus be burdensome for the practical implementation by financial institutions.

Therefore, I propose to rely on a bootstrap procedure. Although this does increase the computational effort, it greatly simplifies the methodology behind the applied backtest.

The idea is to estimate $\gamma_0$ related to a set of $T$ VaR and ES forecasts and a set of actual return loss realizations $\vec{l}$. In order to obtain the distribution of $\gamma_0$ under the assumption that the model is correctly specified, $M$ bootstrap simulation trials are applied. The overall procedure is as follows.

- Estimate the values of $\gamma_0$ and $\beta_0$ related to the observed losses and the given risk estimates, i.e. estimate $\gamma_0(\vec{l})$ and $\beta_0(\vec{l})$.

[16] See Bayer and Dimitriadis (2019) chapter 2.4 for asymptotic theory on the estimation procedure under model misspecification.

[17] Again, note that the scoring function slightly differs compared to Bayer and Dimitriadis (2019) given the divergent sign conventions.

[18] Dimitriadis and Bayer (2019) derives both consistency and asymptotic normality of the ML-estimator outlined above in case the parametric model given in (4.23) is correctly specified.

- In order to be able to simulate under the assumption of a correct estimation model, adjust the risk estimates by $\widehat{VaR}^{\text{adj}}_{t,\alpha} := \widehat{VaR}_{t,\alpha} - \beta_0(\vec{l})$ and $\widehat{ES}^{\text{adj}}_{t,\alpha} := \widehat{ES}_{t,\alpha} - \gamma_0(\vec{l})$.
- For any bootstrap trial $i \in \{1,\dots,M\}$, draw $T$ triples $(\widehat{VaR}^{\text{adj}}_{t,\alpha}, \widehat{ES}^{\text{adj}}_{t,\alpha}, l_t)$ with replacement from the actual sample. Estimate the regression parameter $\gamma_0^i$ based on the respective bootstrap sample for $i = 1,\dots,M$.
- Estimate the p-value related to the vector of observed losses $\vec{l}$ given by,

$$p = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{\gamma_0^i < \gamma_0(\vec{l})\}} \,. \tag{4.28}$$

Again, given some significance level for the backtest $\kappa$, the null hypothesis stated in (4.22) is rejected whenever $p < \kappa$. Moreover, in the following chapters, I will conduct the intercept ESR backtest based on $M = 100$ bootstrap simulations.[19]
Concluding, the intercept ESR backtest proposed within this thesis differs from the original version derived by Bayer and Dimitriadis (2019) with respect to the following two aspects. First of all it uses both VaR and ES forecasts as input variables and thus avoids model misspecification. Secondly, the test decision is based on a bootstrap procedure. Therefore, no cumbersome derivation and no further assumptions are required in order to come up with a consistent asymptotic covariance estimator. With respect to the criteria stated in Proposition 3.3, especially the bootstrap decision might be seen as a drawback of the outlined backtest, as it is related to high computation times. Furthermore, the overall test methodology is rather complex and requires a decent understanding of the concept of conditional elicitability. Nevertheless, the testing framework is highly interesting, as it provides a new and fundamentally different view on the issue of backtesting the ES compared to all previously introduced approaches.

### 4.4. Z-Test from Costanzino and Curran (2015)

This section is going to present the ES backtest proposed by Costanzino and Curran (2015). Moreover, the approach can be applied for any risk measure which belongs to the class of spectral risk measures according to definition Definition 2.7. The following subsection is first going to introduce the methodology of the backtest outlined by Costanzino and Curran (2015) and Costanzino and Curran (2018), whereas the sub-section thereafter is going to present a slightly adjusted, approximative version which requires less input variables.

[19] Note that a value of only $M = 100$ simulation trials is applied, as a ML estimation needs to be conducted in every simulation trial, which leads to a fast increase of computational time depending on the choice of $M$. Nevertheless, this choice of $M$ still leads to satisfying results as outlined in the following chapters.

#### 4.4.1. Original approach - Costanzino and Curran (2015)

As an extension of the violation indicator $I_t(\alpha)$ defined for the VaR, Costanzino and Curran (2015) suggests to define the so called *spectral risk measure violation rate* as follows.

**Definition 4.3** (Spectral risk measure violation rate)**.** *Let $\psi$ be an admissible risk spectrum in the sense of Definition 2.6, then for any $t$ within the backtesting horizon, $X^t_\psi \in [0,1]$ is defined as,*

$$X^t_\psi := \int_0^1 \psi(p) I_t(p) dp \,. \tag{4.29}$$

*Moreover, for any backtesting horizon $T$, define the spectral risk measure violation rate, $X^T_\psi \in [0,1]$ for an admissible risk spectrum $\psi$ as,*

$$X^T_\psi := \frac{1}{T}\sum_{t=1}^{T} X^t_\psi = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \psi(p) I_t(p) dp \,. \tag{4.30}$$

Compared to the violation indicator $I_t(\alpha)$ for the VaR, the spectral risk measure violation rate measures the exceedance of any VaR level, where the respective confidence level lies within the support of $\psi$. Therefore, in case of the ES, the spectral risk measure violation rate takes into account both the amount and the magnitude of losses beyond the VaR threshold with confidence level $\alpha$. Whereas $I_t(\alpha)$ can only take on the binary values 0 or 1, the spectral risk measure violation rate takes on continuous values in the interval $[0,1]$. Given that the underlying spectral risk measure is correctly specified, $\{X^t_\psi\}_{t=1}^T$ are i.i.d. distributed and furthermore, $\mathbb{P}[L_t \geq VaR_{t,p}] = 1-p$ for all $p \in \text{supp}\ \psi$ and $t \in \{1,\dots,T\}$. Given the model is accurately estimated, Costanzino and Curran (2015) derives the following distributional properties of the spectral risk measure violation rate.

**Proposition 4.4** (Costanzino and Curran (2015))**.** *Given a correctly specified spectral risk measure based on some admissible risk spectrum $\psi$, then the mean and variance of the spectral risk measure violation rate $X^T_\psi$ are given by,*

$$\mu_\psi := \mathbb{E}[X^T_\psi] = \int_0^1 \psi(p)(1-p)dp \,, \tag{4.31}$$

$$\sigma^2_\psi := Var[X^T_\psi] = \frac{1}{T}\left(2\int_0^1\int_0^p \psi(p)\psi(q)(1-q)dqdp - \mu^2_\psi\right) \,. \tag{4.32}$$

*Moreover, under the assumption that the spectral risk measure is correctly specified, the spectral risk measure violation rate is asymptotically normally distributed with,*

$$Z^T_\psi := \frac{X^T_\psi - \mu_\psi}{\sigma_\psi} \xrightarrow[T\to\infty]{d} N(0,1) \,. \tag{4.33}$$

*Proof.* (i) The derivation of both moments of the spectral risk measure violation rate $\mu_\psi$ and $\sigma_\psi$ follows according to Proposition 3.4 in Costanzino and Curran (2015) taking into account the different conventions within this thesis. Note for example, that $\mathbb{P}[L_t \geq VaR_{t,p}] = 1-p$ given the notation used within this thesis, which explains the difference in (4.31) and (4.32) compared to Costanzino and Curran (2015).

(ii) As derived in Lemma 3.5 within Costanzino and Curran (2015), the asymptotic distribution of $X_{\psi}^{T}$ follows by the *Lindeberg-Levy central limit theorem* as both $\mu_{\psi}$ given in (4.31) and $\sigma_{\psi}$ given in (4.32) are bounded.

□

For the particular case of backtesting the ES, one can construct a one-sided Z-test by exploiting the distributional properties of $Z_{\psi_{ES}}^{T}$. As pointed out in Lemma 2.8 within this thesis, the admissible risk spectrum related to the ES is given by $\psi_{ES}(p) = \frac{1}{1-a}\mathbb{1}_{\{\alpha \leq p \leq 1\}}$. Applying $\psi_{ES}$ to the previous derivations leads to the consecutively listed results, following both Costanzino and Curran (2015) and Costanzino and Curran (2018).[20]

- According to Costanzino and Curran (2018), the spectral risk measure violation rate for the ES can be calculated by,

$$\begin{aligned} X_{ES_\alpha}^{t} &= \int_0^1 \psi_{ES}(p) I_t(p) dp = \frac{1}{1-\alpha}\int_\alpha^1 I_t(p)dp \\ &= \left(1 - \frac{F_t(-L_t)}{1-\alpha}\right) I_t(\alpha)\,, \qquad (4.34) \\ X_{ES_\alpha}^{T} &= \frac{1}{T}\sum_{t=1}^{T} X_{ES_\alpha}^{t} = \frac{1}{T\cdot(1-\alpha)}\sum_{t=1}^{T}\int_\alpha^1 I_t(p)dp \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{F_t(-L_t)}{1-\alpha}\right) I_t(\alpha)\,. \qquad (4.35) \end{aligned}$$

  Note that in equation (4.34) and (4.35), the term $1 - F_t(-L_t)/(1-\alpha)$ accounts for the severity of a VaR breach beyond the confidence level of $\alpha$.

- Given that the ES is correctly specified, the following moments of the ES violation rate are derived in Costanzino and Curran (2015),

$$\begin{aligned} \mu_{ES_\alpha} &:= \mathbb{E}[X_{ES_\alpha}^{T}] = \frac{1-\alpha}{2}\,, \qquad (4.36) \\ \sigma_{ES_\alpha}^{2} &:= Var[X_{ES_\alpha}^{T}] = \\ &\quad \frac{1-\alpha}{T}\left(\frac{4-3(1-\alpha)}{12}\right)\,. \qquad (4.37) \end{aligned}$$

- Given the two previous listings and Proposition 4.4, the following distribution of the ES violation rate can be defined,

$$\begin{aligned} \lim_{T\to\infty} Z_{ES_\alpha}^{T} &:= \frac{X_{ES_\alpha}^{T} - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} \qquad (4.38) \\ &= \sqrt{3T}\left(\frac{2X_{ES_\alpha}^{T} - \alpha}{\sqrt{\alpha(4-3\alpha)}}\right) \sim N(0,1)\,. \end{aligned}$$

[20] Again note that some minor adjustments need to be made compared to Costanzino and Curran (2015) and Costanzino and Curran (2018) given slightly different conventions within this thesis.

Based on (4.35) and (4.38), one can calculate the realized value of both $X_{ES_\alpha}^{T}(\vec{l})$ and $Z_{ES_\alpha^T}(\vec{l})$, related to some set of return loss observations $\vec{l}$. Obviously, a high value of $Z_{ES_\alpha}^{T}(\vec{l})$ also indicates a high value of the ES violation rate and thus a potential underestimation of the actual risk. In order to test the one-sided hypothesis that the modelled risk $\widehat{ES}_{t,\alpha}$ underestimates the actual risk given by $ES_{t,\alpha}$, Costanzino and Curran (2018) suggests a similar procedure as for the VaR traffic light. More precisely, Costanzino and Curran (2018) argues that the realized ES violation rate, i.e. $X_{ES_\alpha}^{T}(\vec{l})$, can be seen as the equivalent to the number of observed violations in the VaR traffic light test. The higher the realized ES violation rate, the higher is the degree of risk underestimation.[21]

Denote the realized ES violation rate by $X_{ES_\alpha}^{T}(\vec{l}) = x$. Then the underlying estimation model is rejected in the one-sided test, at a significance level of $\kappa$, if the cumulative probability of $X_{ES_\alpha}^{T}$ being lower or equal to $x$ is larger than $1-\kappa$. This is indeed a similar test decision as outlined for the VaR traffic light test in chapter 3.2 within this thesis. Formally the estimation model is rejected whenever,

$$\mathbb{P}[X_{ES_\alpha}^{T} \leq x] > 1 - \kappa\,. \qquad (4.39)$$

In order to achieve a concrete test decision, on can standardize $X_{ES_\alpha}^{T}$ by $\mu_{ES_\alpha}$ and $\sigma_{ES_\alpha}$ and exploit the distributional properties outlined in (4.38). Applying the approximate distribution of $Z_{ES_\alpha}^{T}$ under the assumption that the model is correctly specified, the estimation model in the one-sided test, related to some set of realized losses $\vec{l}$, is rejected whenever,

$$Z_{ES_\alpha}^{T}(\vec{l}) = \frac{x - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} > \Phi^{-1}(1-\kappa)\,, \qquad (4.40)$$

where $\Phi$ denotes the CDF of the standard normal distribution. It should be noted that for a finite backtesting horizon, $Z_{ES_\alpha}^{T}$ is only approximately standard normally distributed under the assumption of a correctly specified risk measure forecast. This is in contrast to the VaR traffic light test, where observed violations follow a binomial distribution under a correctly specified model. Thus, for the VaR there is no need to employ an approximate normal distribution induced by the central limit theorem. Costanzino and Curran (2018) puts some further attention to the finite sample distribution of $Z_{ES_\alpha}^{T}$, but this goes beyond the scope of this thesis. Therefore, I rely on the assumption that critical values obtained by the standard normal distribution are reasonable approximations for a backtesting horizon like $T = 250$. In more detail, this means that critical values of $Z_{ES_\alpha}^{T,*} = 1.64$ and $Z_{ES_\alpha}^{T,*} = 3.72$ are going to be used within this thesis for the backtest significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$ stipulated by the Basel Committee.

With respect to the criteria stated in Proposition 3.3, there is one major drawback of the ES backtest as it is proposed

[21] Note that Costanzino and Curran (2018) uses a different scaling for the ES violation rate compared to the original version by Costanzino and Curran (2015), which is also used within this thesis. This has no impact of the overall test methodology, but the achieved ES violation rates differ in their magnitude compared to Costanzino and Curran (2018).

by Costanzino and Curran (2015) and Costanzino and Curran (2018). Namely, for the calculation of the ES violation rate given in (4.35), which functions as a basis for the test decision, the risk manager needs to estimate the cumulative distribution function of returns $-L_t$ for any date within the backtesting horizon. This is especially problematic in case the risk manager applies a non-parametric ES forecasting model like a historical simulation. In such a situation, the backtest might not be feasible for the risk manager, as estimating the CDF of returns would be related to imposing further restrictions to the ES estimation model. Therefore, I will propose an approximate alternative to the backtest proposed by Costanzino and Curran (2015), which does not require the explicit derivation of any CDF function.

#### 4.4.2. Adjusted approach - Approximate Z-test

Similar to the multinomial backtest proposed by Kratz et al. (2018), I suggest to approximate the integral $\int_\alpha^1 I_t(p)dp$ by a Riemann-sum using $N \in \mathbb{N}$ VaR figures with confidence levels between $\alpha$ and one. First of all, recall notation (4.2), i.e. $\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha)$ for $j = 1,\dots,N$, for some reference confidence level $\alpha$. Moreover, I propose to use the following approximation,

$$\int_\alpha^1 I_t(p)dp \approx \frac{1-\alpha}{N}\left(\sum_{j=1}^{N} I_t(\alpha_j)\right) . \tag{4.41}$$

This leads to an approximation of the ES violation ratio given in (4.35), which does not require any explicit estimation of a CDF function,

$$X_{ES_\alpha}^{T,\text{approx}} := \frac{1}{T(1-\alpha)}\sum_{t=1}^{T}\frac{(1-\alpha)}{N}\left(\sum_{j=1}^{N} I_t(\alpha_j)\right) , \tag{4.42}$$

for some $N \in \mathbb{N}$ which determines the approximation accuracy. It should be noted, that for the estimation of an realization $X_{ES_\alpha}^{T,\text{approx}}(\vec{l})$, one needs to estimate a vector of VaR figures, $(\widehat{VaR}_{t,\alpha_1},\dots,\widehat{VaR}_{t,\alpha_N})$. Nevertheless, in general this can be obtained from both parametric and non-parametric estimations models for the ES without any further difficulties. Furthermore, I define the approximative test statistic equivalent to (4.38) as,

$$Z_{ES_\alpha}^{T,\text{approx}} := \frac{X_{ES_\alpha}^{T,\text{approx}} - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} . \tag{4.43}$$

Additionally, the approximative approach also relies on the same one-sided test decision. The estimation model for the ES is rejected based on some return loss realizations $\vec{l}$, whenever

$$Z_{ES_\alpha}^{T,\text{approx}}(\vec{l}) > \Phi^{-1}(1-\kappa) , \tag{4.44}$$

where $\kappa$ is again the significance level of the backtest. In order to limit the degree of misspecification due to an imprecise approximation, I propose a value of $N = 8$ for the calculation of $X_{ES_\alpha}^{T,\text{approx}}$ and $Z_{ES_\alpha}^{T,\text{approx}}$.

Overall, both versions presented within this section are indirect backtests for the ES. Whereas the original version from Costanzino and Curran (2015) requires the entire tail distribution of returns as an input parameter, the approximative version needs a vector of VaR estimates as an input variable. Similar as the multinomial approach by Kratz et al. (2018), both versions do not need actual ES forecasts as an input parameter. In case of a parametric ES estimation model, the original version brought up by Costanzino and Curran (2015) can easily be applied, given that the risk manager anyway has an explicit estimation of the return loss CDF at hand. On the other hand, for any non-parametric approach the approximative version suggested in the last subsection can be conducted, which although might be related to a higher degree of misspecification in the ES backtest. In the light of the further criteria stated in Proposition 3.3, both versions are computationally efficient and can easily be modified to a two-sided version. Moreover, I believe the underlying concept of generalizing the violation indicator for the VaR to the class of spectral risk measures is relatively intuitive. Furthermore, this allows for a high degree of flexibility as the approach can also be transferred to other spectral risk measures.

### 4.5. Residuals Bootstrap Test from McNeil and Frey (2000)

The *residuals bootstrap test* derived by McNeil and Frey (2000) is up to my knowledge the first backtest for the ES suggested within the literature. Nevertheless, it is still widely used as for example pointed out in Bayer and Dimitriadis (2019). Furthermore, it functions as a basis for different backtesting approaches suggested within the literature later on.[22] This subsection is first going to introduce the original version by McNeil and Frey (2000). Consecutively, a slight addition to the test decision will be made for the purpose of this thesis.

The test by McNeil and Frey (2000) is based on the so called *exceedance residuals*. In addition to VaR and ES forecasts, assume the risk manager also has a set of estimates for the conditional volatility of return losses at hand, denoted by $\{\widehat{\sigma}_t : t = 1,\dots,T\}$. For a set of realized return losses $\vec{l}$, the exceedance residual for any time $t$ within the backtesting horizon is then given by,

$$r_t := \frac{l_t - \widehat{ES}_{t,\alpha}}{\widehat{\sigma}_t} I_t(\alpha) . \tag{4.45}$$

If one assumes that the applied risk estimation model is correctly specified, it holds

$$\mathbb{E}[r_t] = \mathbb{E}\left[\frac{L_t - ES_{t,\alpha}}{\sigma_t}\middle| L_t \geq VaR_{t,\alpha}\right] = 0 , \tag{4.46}$$

as the distribution of $L_t$ is assumed to be continuous and strictly increasing, such that $ES_{t,\alpha} = \mathbb{E}[L_t | L_t \geq VaR_{t,\alpha}]$. On the contrary, if the applied estimation model underestimates the actual risk, i.e. $ES_{t,\alpha} > \widehat{ES}_{t,\alpha}$, then the expectation will be larger than zero. In order to estimate the expected value of exceedance residuals over the backtesting horizon, McNeil and Frey (2000)

[22] As an example the *Test 1* in Acerbi and Szekely (2014) can be seen as an adjusted version of the residuals bootstrap test proposed by McNeil and Frey (2000).

suggests to consider their mean value over the backtesting period given by,

$$\bar{r} := \frac{1}{\sum_{t=1}^{T} I_t(\alpha)} \sum_{t=1}^{T} r_t \, . \tag{4.47}$$

Thus, for the one-sided test version the following null and alternative hypothesis are going to be considered,

$$\begin{aligned} H0: &\quad \bar{r} \leq 0 \\ H1: &\quad \bar{r} > 0 \, . \end{aligned} \tag{4.48}$$

The actual test decision proposed by McNeil and Frey (2000) follows a bootstrap procedure, which is described in detail in Efron and Tibshirani (1994) on page 224. In short, the bootstrap based on a set of forecasts for VaR, ES and $\sigma$ as well as realized return losses $\vec{l}$ is outlined below.

- Calculate the value of $\bar{r}$ related to the vector of observed losses and the risk estimates, i.e. calculate $\bar{r}(\vec{l})$.
- In order to simulate $\bar{r}$ under the $H0$ proceed in the following way. For any violation of $\widehat{VaR}_{t,\alpha}$, calculate the adjusted exceedance residual given by $r_t^{\text{adj}} = (r_t - \bar{r}(\vec{l}))\mathbb{1}_{\{r_t \neq 0\}}$.
- Set up the sample of adjusted residuals as $S := \{r_t^{\text{adj}} : r_t \neq 0\}$. If $x$ observations of $\vec{l}$ breach the respective $\widehat{VaR}_{t,\alpha}$ level, then $|S| = x$, where $|S|$ denotes the cardinality of $S$. For each of the $M$ bootstrap trials, draw $|S|$ observations with replacement from $S$, and calculate $\bar{r}^i$ for each trial $i \in \{1, \ldots, M\}$.
- Estimate the p-value related to the vector of observed loss $\vec{l}$ given by,

$$p = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{\bar{r}^i > \bar{r}(\vec{l})\}}. \tag{4.49}$$

Again the null hypothesis of the one-sided bootstrap test suggested by McNeil and Frey (2000) is rejected, whenever $p < \kappa$ for some predefined significance level $\kappa$ for the backtest. For the purpose of this thesis, the respective bootstrap test decisions are going to be based on $M = 1000$ simulation trials in the subsequent chapters.

There are two aspects which are often criticized about the test proposed by McNeil and Frey (2000). First of all, the backtest requires an estimation of the volatility of return losses, which might not be feasible in certain scenarios. Indeed, as argued by Acerbi and Szekely (2014) and Bayer and Dimitriadis (2019) one can simply drop the volatility of return losses in definition (4.45) and use exceedance residuals which are not standardized by the volatility. Indeed, this can easily be implemented as it does not require any further adjustments of the overall test methodology. The second conceptional shortcoming of the backtest proposed by McNeil and Frey (2000), is that it only accounts for the magnitude of violations beyond the $VaR_{t,\alpha}$ threshold, but not for the overall amount of violations. As outlined in section 3.3 within this thesis, both would be necessary to account for a correct unconditional coverage of the ES. In other words, Acerbi and Szekely (2014) argues that the outlined test backtests the ES conditional on correctly specified VaR estimates. In order to handle this issue, I propose to apply the residual backtest from McNeil and Frey (2000) in combination with the VaR traffic light test introduced in section 3.2. Therefore, I propose the following adjusted one-sided null and alternative hypothesis.

$$\begin{aligned} H0: &\quad \bar{r} \leq 0 \quad \text{and} \quad \sum_{t=1}^{T} I_t(\alpha) \leq 1 - \alpha \\ H1: &\quad \bar{r} > 0 \quad \text{or} \quad \sum_{t=1}^{T} I_t(\alpha) > 1 - \alpha \end{aligned} \tag{4.50}$$

Correspondingly, the $H0$ in the combined backtest is rejected if either $p < \kappa$, where $p$ is derived according to (4.49) or if $F^{binom}\left(\sum_{t=1}^{T} I_t(\alpha), T, 1-\alpha\right) > 1 - \kappa$. Therefore, I will use the *combined ES residuals backtest* based on the set of hypotheses (4.50) as it validates correct unconditional coverage of ES forecasts. Moreover, the test decision displays some kind of "worst-of-logic", as the null hypothesis is rejected, if either the original ES residuals backtest or the VaR traffic light test is rejected at a certain significance level $\kappa$.

With respect to the desirable properties of a backtest outlined in Proposition 3.3, the backtest described above is very intuitive from a conceptional point of view. Furthermore, it only requires realized P&L realizations as well as VaR and ES forecasts as input variables, given that the input of volatility forecasts is optional. The alleged shortcoming, that the backtest from McNeil and Frey (2000) only accounts for the magnitude but not the amount of VaR violations can easily be mitigated by linking the backtest to the VaR traffic light test. The only apparent disadvantage of the outlined test is, that the test decision again involves a simulation procedure.

Summing up this chapter, all five proposed ES backtest yield a one-sided test versions, in order to detect estimation models which underestimate the actual risk of an underlying portfolio. Furthermore, all introduced backtests are applicable for commonly used parametric and non-parametric ES estimation models, in a sense that all required data inputs can be obtained without inducing further assumptions. Overall, two of the five proposed tests are based on a simulation decision. Furthermore, for rather extreme backtest significance levels like $\kappa = 0.0001$, one should additionally also rely on the bootstrap version of Test 2 from Acerbi and Szekely (2014).

With respect to the practical aspects listed in Proposition 3.3, I propose the following qualitative judgement of all five overall testing methodologies based on the arguments outlined within this chapter. More precisely, I assign grades to all backtests in scope, with respect to the fulfilment of the four aspects listed in Proposition 3.3, namely *data intensity*, *computational effort*, *conceptual ease* and *clear decision making*. The grades are given by **(++)**, **(+)** for excellent or good results, **(o)** for an average result, and **(-)** or even **(- -)** if a certain aspect might be problematic for practical implementations. The proposed

judgement is outlined in Tables 3-7 below, together with a short explanation for any ES backtesting methodology.

- **Multinomial backtest - Original and de-facto one-sided version:**

All input variables for the multinomial backtesting methodology can easily be obtained for every parametric and non-parametric ES estimation approach. The backtest is computationally efficient as it does not rely on a bootstrap procedure and the methodology is a straightforward extension of the VaR traffic light test. Some adjustments to the test decision are needed in order to develop a one-sided test version, nevertheless this is not expected to have a major impact on the overall test performance.

- **Test 2:**

If the Test 2 is carried out with fixed critical values, it only requires VaR and ES forecasts in addition to P&L realizations, furthermore in this case no bootstrap is required. The overall methodology is rather intuitive and can also be adjusted to reflect a two-sided hypothesis. Especially for rather extreme backtesting confidence levels $\kappa$, an applicant faces the trade off between potentially inaccurate test decisions and a computationally costly bootstrap test version. Furthermore, in case a bootstrap is applied, additional input variables need to be stored.

- **Intercept ESR backtest:**

Input parameters for the intercept ESR backtest can be easily obtained for any ES estimation model, furthermore a standalone ES backtest version is proposed within Bayer and Dimitriadis (2019). On the contrary, the test methodology, especially the regression estimation based on the concept of conditional elicitabilty is rather complex. Test decisions can either be based on a Wald-test, which requires the cumbersome estimation of a consistent asymptotic covariance estimator as outlined in Bayer and Dimitriadis (2019), or on a bootstrap procedure, which is applied in the context of this thesis. Nevertheless, the bootstrap procedure is related to a rather extreme computational effort as a ML estimation needs to be conducted in every simulation trial. Both one and two-sided versions of the intercept ESR backtest can easily be conducted.

- **Z-Test - Original and approximative version:**

The original version requires the explicit estimation of the CDF of returns, which is not feasible for non-parametric ES forecasting models. This drawback can be avoided by applying the approximate version, which I suggested for the purpose of this thesis. Asymptotic distributional properties of the test statistic can be employed. Thus, no simulation procedure is required. Furthermore, the overall concept is rather intuitive. Both one-sided and two-sided test decisions can easily be obtained, nevertheless critical values for finite backtesting horizons are based on asymptotic distributional properties.

- **Combined ES residual backtest:**

The test methodology is very intuitive, furthermore only ES and VaR forecasts are required as mandatory input parameters in addition to P&L realizations. The test can be carried carried out as a one-sided or a two-sided version. Nevertheless, the test decisions is based on a simulation approach, which requires computational effort.

Overall, the concerns regarding the backtestability of the ES, which are for example stated in Carver (2013) and quoted within earlier stages of this thesis, appear to be exaggerated. Indeed, there are multiple promising backtest for the ES, some of which are listed above and are further evaluated in the consecutive chapters. With respect to practical applicability, especially the multinomial backtesting methodology as well as the combined ES residuals backtest need to be highlighted, as they exhibit the highest compliance with the criteria stated in Proposition 3.3. Nevertheless, all outlined ES backtesting versions qualify for a practical implementation.

## 5. Simulation study - Comparison of empirical size and power

This chapter aims to compare all five considered backtests as well as any adjusted version according to their empirical size and power. Within the literature both measures are commonly taken as a reference for the performance of statistical tests. For instance, analyses on both of these measures can be found in Acerbi and Szekely (2014), Bayer and Dimitriadis (2019) and Kratz et al. (2018), for the ES backtest presented within the respective article. On the other hand, up to the best of my knowledge, non of these two measures has been analysed yet for the Z-test introduced by Costanzino and Curran (2015). I will propose one single methodology for comparing simulated size and power of all ES backtests, which are in scope of this thesis. Thus, I will complement already existing analyses, in order to make a more founded judgement on the conceptual soundness of each of the evaluated approaches.

For the conducted simulation study, both versions of the multinomial backtest based on Kratz et al. (2018) as well as both version of the Z-test introduced by Costanzino and Curran (2015) are evaluated, in order to examine how the proposed adjustments impact both size and power of the original backtests. Furthermore, both test components of the combined ES residuals backtest are evaluated separately, to analyse how they contribute to the overall test decision. All other backtests are conducted as outlined in the preceding chapter.

The simulation study is structured as follows. Subsection 5.1 formally defines both size and power and introduces the set up of the analysis thereafter. Subsection 5.2 compares the size of all ES backtests, whereas subsection 5.3 compares their simulated power for different degrees of model misspecification. The chapter is concluded by a short judgement of the obtained size and power values for all backtests in scope.

### 5.1. Theory on empirical size and power

As for example outlined by Kratz et al. (2018), both size and power of an ES backtest are defined in the following way.

| Data intensity | Computational effort | Conceptual ease | Clear decision making |
|---|---|---|---|
| ( + ) | ( + + ) | ( + + ) | ( o ) |

**Table 3:** Assessment of multinomial ES backtest with respect to the criteria states in Proposition 3.3.

| Data intensity | Computational effort | Conceptual ease | Clear decision making |
|---|---|---|---|
| ( o ) | ( + ) | ( + ) | ( + ) |

**Table 4:** Assessment of the Test 2 with respect to the criteria states in Proposition 3.3.

| Data intensity | Computational effort | Conceptual ease | Clear decision making |
|---|---|---|---|
| ( + + ) | ( - ) | ( - ) | ( + + ) |

**Table 5:** Assessment of the intercept ESR backtest with respect to the criteria states in Proposition 3.3.

| Data intensity | Computational effort | Conceptual ease | Clear decision making |
|---|---|---|---|
| ( o ) | ( + + ) | ( + ) | ( + ) |

**Table 6:** Assessment of the Z-test with respect to the criteria states in Proposition 3.3.

| Data intensity | Computational effort | Conceptual ease | Clear decision making |
|---|---|---|---|
| ( + ) | ( o ) | ( + + ) | ( + + ) |

**Table 7:** Assessment of the combined ES residuals backtest with respect to the criteria states in Proposition 3.3.

**Definition 5.1** (Size/Power of ES backtest)**.** *Consider a backtest for the ES, then*

- *the size of the backtest is defined as* $\gamma := \mathbb{P}(\textit{reject } H0 | H0 \textit{ true})$. *Moreover, the size of the backtest corresponds to the respective type I error* $\gamma$.
- *the power of the backtest is defined as* $1-\beta := 1 - \mathbb{P}(\textit{accept } H0 | H0 \textit{ false})$. *Moreover, the power of the backtest corresponds to one minus the respective type II error* $\beta$.

On the one hand, any reasonable backtest for the ES should have a small size, which ought to be around the significance level $\kappa$ applied in the backtesting procedure. As argued by Kratz et al. (2018), size values below $\kappa$ are per se not problematic as they yield an even lower type I error. On the other hand, for a well functioning ES backtest, high power values, possibly close to one, are to be expected. Put in different words, a suitable ES backtest should yield a low probability of rejecting a correctly specified ES estimation model, while still being able to detect misspecified models with a high probability. Thus, in order to qualify for any practical implementation, an ES backtest should be as powerful as possible given a reasonably low size.

Within the literature, it is common practice to determine both size and power empirically by determining rejection rates in a simulation study. As mentioned above, respective simulation analyses can for example be found in Acerbi and Szekely (2014), Kratz et al. (2018) and Bayer and Dimitriadis (2019). The approach which will be applied within thesis is closest to the one outlined by Kratz et al. (2018). Sticking with the previously applied notation, a risk manager calculates an ES forecast based on some conditional return loss estimation $\widehat{L}_t$, whereas the true return loss is distributed according to $L_t$. Again both $\widehat{L}_t$ and $L_t$ are based on the information available up to $t-1$, which is given by the $\sigma$-Algebra $\mathscr{F}_{t-1}$. The difference of the proposed setting compared to Kratz et al. (2018) is that time dependent risk forecasts are considered. This is necessary, as the methodology of the intercept ESR backtest from Bayer and Dimitriadis (2019) is based on the assumption of conditional, time dependent, risk forecasts. Thus, applying the intercept ESR backtest to a series of time-independent risk forecasts, $ES_\alpha$ for all $t \in \{1, \dots, T\}$, would not lead to any meaningful test decision.

For both size and power simulations different location-scale distributions, like the normal or the t-distribution, are applied for both $\widehat{L}_t$ and $L_t$, for observations $t$ within the backtesting

horizon. Bayer and Dimitriadis (2019) uses a more sophisticated approach and fits ES estimation models to real financial data, nevertheless for both size and power simulations it is not essential that the assumed return loss distributions perfectly fit some underlying return series. Therefore, relying on commonly used location-scale distributions simplifies the set up of the applied analysis. In order to determine the size of the ES backtests, rejection rates over $MC = 1000$ test decisions are simulated in a scenario where $\widehat{L}_t = L_t$, i.e. the risk manager perfectly estimates the underlying risk. On the other hand, for the power calculations, rejection rates are based on $MC = 1000$ simulated test decisions in situations where $\widehat{L}_t \neq L_t$. More precisely, different scenarios are evaluated where the risk manager underestimates the actual risk of the underlying portfolio.

For the analysis conducted within this chapter, backtesting horizons of $T = 250$, 500 and 1000 observations are considered. Indeed, many authors argue that the backtesting period of 250 observations stipulated by the Basel Committee is not sufficient to backtest the ES. As an example, Kratz et al. (2018) considers a maximum backtesting period of $T = 2000$ observations, while Bayer and Dimitriadis (2019) even takes into account up to $T = 5000$ observations. Nevertheless, it should be noted that backtesting periods with $T > 1000$ are often not feasible in actual applications. Considering daily returns, 1000 observations correspond to a time series of roughly four years. Even in this case, it might be challenging for the risk manager to collect risk forecasts and return losses over such a long time frame. Furthermore, the conducted analysis takes into account a confidence level of $\alpha = 0.975$ for the ES and a significance level of $\kappa = 0.05$ for the backtests, which corresponds to the yellow traffic light zone in the methodology of the Basel Committee.

### 5.2. Empirical size of backtests

In order to determine the size of the backtest, I simulate $MC = 1000$ test decisions under the assumption that the risk manager correctly estimates the underlying risk, i.e. $\widehat{L}_t = L_t$ for all $t \in \{1, \dots, T\}$. Moreover, for every observation $t$ within the backtesting horizon I set $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim N(\mu_t, \sigma_t)$, respectively. For the procedure, at first 250 return loss observations $l_{-249}, \dots, l_0$ are drawn from the standard normal distribution. Consecutively, both $\mu_1$ and $\sigma_1$ are calculated as the sample mean and sample standard deviation of the preceding 250 observations. Furthermore, $\widehat{ES}_{1,\alpha}$ is calculated according to the theoretical value based on the distribution $N(\mu_1, \sigma_1)$. More precisely, the ES forecast based on the normal distribution is given by,

$$\widehat{ES}_{1,\alpha} = \frac{1}{1-\alpha}\sigma_1 \phi(\Phi^{-1}(\alpha)) + \mu_1 \,, \tag{5.1}$$

where again $\phi$ and $\Phi$ denote the PDF and the CDF of the standard normal distribution, respectively. Furthermore, also all auxiliary variables, needed for any of the outlined backtests, are calculated based on $N(\mu_1, \sigma_1)$. In the following, the actual observed loss is drawn from the same distribution, i.e $l_1 \sim N(\mu_1, \sigma_1)$. This procedure is afterwards carried out for any observation within the backtesting period, whereas both $\mu_t$ and $\sigma_t$ are always calculated on the rolling sample of the previous 250 observations, i.e. $l_{t-250}, \dots, l_{t-1}$. In the end, the decision of all evaluated backtests is based on the obtained vectors of $\{\widehat{ES}_{t,\alpha} : t = 1, \dots, T\}$ and $\{l_t : t = 1, \dots T\}$ as well as on potentially required auxiliary forecasts. Overall, $MC = 1000$ test decisions are simulated based on the outlined procedure and the respective rejection rate is calculated.

The resulting rejection rates for all considered backtests are depicted in Table 8. As previously mentioned, the ES is estimated at a confidence level of $\alpha = 0.975$. Moreover, a backtesting significance level of $\kappa = 0.05$ and backtesting horizons of $T = 250$, 500 and 1000 observations are considered. For visualisation purposes, the colouring scheme is taken from Kratz et al. (2018). Rejection rates in the size simulation below 6% yield good results and are thus coloured in green. Poor empirical size values are highlighted in orange if the simulated rejection rate exceeds 9% and in red given an even higher rejection rate above 12 %.

Overall, most of the evaluated backtests do show decent rejection rates in the size analysis over the considered backtesting horizons. Both multinomial backtests exhibit a rejection rate close to the significance level of 5% for all three backtesting time frames. As to be expected, the difference in rejection rates between both test versions is rather marginal. Furthermore, it is reasonable that rejection rates of the de-facto one-sided version are slightly below those of the original approach, as the adjusted backtest additionally rules out the rejection of simulation trials which display a conservative risk estimation. Overall, both multinomial backtests possess excellent size properties. The Test 2 from Acerbi and Szekely (2014) achieves low rejection rates, which are even below the significance level applied in the simulation study. Interestingly, it appears that rejection rates even converge to zero given an increase in the backtesting horizon. As argued before, this should per se not be seen as a drawback of the test from Acerbi and Szekely (2014). Nevertheless, if the Test 2 reveals a low power in the consecutive subsection, this might be an indicator for a poor balancing of both size and power in the respective test framework. The simulation results for the intercept ESR backtest are in line with expectations, with rejection rates below 7% in all three simulations. Furthermore, the empirical size of the intercept ESR backtest appears to be rather stable across different backtesting horizons, close to the applied significance level $\kappa$. On the contrary, the empirical size of both, the original Z-test from Costanzino and Curran (2015) as well as the proposed approximative version, exceeds a value of 10% for all evaluated scenarios. It is even more problematic that the simulated rejection rates further increase given an increase in the backtesting period. For a value of $T = 1000$ the original Z-test rejects the correctly specified ES estimation model in almost every fourth test decision. Surprisingly, the empirical size of the approximative Z-test seems to be slightly superior compared to the original version. Nevertheless, both test versions exhibit unsatisfactory high rejection rates in the size analysis, which hints towards a potential misspecification in the underlying framework of the ES backtest. On possible reason might be, that critical values based on the asymptotic distribution of the test statistic might not lead to accurate test decisions for finite backtesting horizons like $T = 250$, as shortly

**Empirical size - Rejection rates of evaluated ES backtests**

| | Backtesting period - T | | |
|---|---|---|---|
| **ES backtesting approach** | 250 | 500 | 1000 |
| Multinomial backtest - Original version | **0.040** | **0.040** | **0.051** |
| Multinomial backtest - One-sided version | **0.037** | **0.040** | **0.050** |
| Test 2 | **0.042** | **0.012** | **0.000** |
| Intercept ESR backtest | 0.063 | **0.043** | 0.063 |
| Z-test - Original version | **0.117** | **0.171** | **0.240** |
| Z-test - Approximative version | **0.101** | **0.117** | **0.171** |
| Combined ES residuals backtest | **0.098** | **0.110** | 0.085 |
| (ES residuals backtest) | **(0.056)** | **(0.030)** | **(0.024)** |
| (VaR traffic light test) | **(0.044)** | (0.083) | (0.062) |

**Table 8:** Empirical size - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t = L_t \sim N(\mu_t, \sigma_t)$ for all $t \in \{1, \ldots, T\}$.

outlined in chapter 4.4. The combined ES residuals backtest also shows slightly too high rejection rates ranging between 8% and 11%. It should be noted, that the test decision of the combined test is composed of both the VaR traffic light test as well as the original ES residuals backtest proposed by McNeil and Frey (2000). Both of these two single test decisions show reasonable size values in all three simulation scenarios. As outlined in chapter 4.5, the combined ES residual backtest is rejected if any of the two applied test components is rejected. Therefore, given the worst-of-logic applied for the combined test decision, it is not surprising that the combined ES residuals backtest is slightly oversized. One might decrease the significance levels for each of the two single test components in order to calibrate the combined ES residuals backtests to a size value closer to 5%. Given that the size values of the combined test are still within an acceptable range, no rescaling is applied and the backtest will be carried out as described in the previous chapter.

Concluding, both multinomial backtests based on Kratz et al. (2018) and the intercept ESR backtest from Bayer and Dimitriadis (2019) show decent size properties in line with expectations. Size values close to zero, observed for the Test 2 from Acerbi and Szekely (2014), might be seen as a bonus of the test framework, if the test still reveals decent power values. Nevertheless, this needs to be evaluated in the following subsection, as overly low size figures might also go hand in hand with a low power in detecting misspecified ES estimation models. The combined ES residuals backtest is slightly oversized. Nevertheless, this can be explained given the underlying test methodology and furthermore one might mitigate too high rejection rates by calibrating the single components of the test decision. Only the size values of the Z-test proposed by Costanzino and Curran (2015) deviate from expectations. It might be difficult for regulators to rely on an ES backtest which rejects a correctly specified ES estimation model in up to every fourth scenario depending on the selected testing parameters.

Furthermore, also no clear pattern can be observed across all backtests regarding the relation of empirical size values and the choice of the backtesting horizon. Thus, it is not obvious that an increased backtesting period automatically leads to improved size properties based on the conducted simulations.

### 5.3. Empirical power of backtests

This subsection aims to evaluate the empirical power of all considered ES backtests, i.e. their ability to detect certain levels of model misspecification. As I decided to base my analysis on one-sided test decisions, only scenarios are analysed where the risk manager underestimates the true underlying risk given by $L_t$. For all single power scenarios, again 250 initial loss observations are drawn from the standard normal distribution. Again both $\mu_t$ and $\sigma_t$ are estimated over the previous 250 observations for every observation $t$ within the backtesting horizon. Furthermore, the risk manager again estimates the underlying risk $\widehat{ES}_{t,\alpha}$ based on the respective normal distribution, i.e. $\widehat{L}_t \sim N(\mu_t, \sigma_t)$. The difference compared to the previous size analysis is that the true distribution $L_t$ deviates from the normal distribution. More precisely, observed return losses $l_t$, for $t \in \{1, \ldots, T\}$, are drawn from different location-scale distributions, whereas location and scale parameters again depend on both $\mu_t$ and $\sigma_t$. Moreover, four different scenarios are analysed, which can be grouped into the following two categories of model misspecification.

(i) **Misspecified tail behaviour and conditional variance**
For the first category, I assume that the true return loss follows a t-distribution with location parameter $\mu_t$ and scale parameter $\sigma_t$. Moreover, I consider two scenarios with $\nu = 3$ and $\nu = 5$ degrees of freedom. Thus, the observed return losses $l_t$ are drawn from $L_t \sim t_\nu(\mu_t, \sigma_t)$. As the t-distribution possesses fatter tails compared to the normal distribution, this leads to a scenario where the risk manager underestimates the underlying risk, due to a misspecification of the distribution tail. The tail

behaviour of the t-distribution converges to that of the normal distribution for increasing values of $\nu$. Therefore, the degree of risk underestimation is larger for the case of $\nu = 3$ compared to $\nu = 5$.
Secondly, the variance of $L_t \sim t_\nu(\mu_t, \sigma_t)$ is given by $Var(L_t) = \sigma_t^2 \frac{\nu}{\nu-2}$, whenever $\nu > 2$. Taking into account that the risk manager assumes return losses to be normally distributed, this leads to

$$Var(L_t) = \sigma_t^2 \frac{\nu}{\nu-2} > \sigma_t^2 = Var(\widehat{L}_t), \quad \text{for } \nu > 2. \tag{5.2}$$

Therefore, the risk manager additionally also underestimates the true conditional variance of return losses. Again, the degree of risk underestimation due to a misspecified variance is larger for the case of $\nu = 3$ compared to $\nu = 5$.
As previously argued, both the number of violations beyond the respective VaR threshold as well as their magnitude with respect to the estimated ES value, need to be considered for correct unconditional coverage of the ES. The misspecified tail behaviour will become apparent in the outer tail of the return loss distribution and thus primarily impact the magnitude of observed VaR violations. Additionally, due to the misspecified conditional variance, the overall level of riskiness is underestimated, which will result in a rather high number of violations beyond the forecasted VaR threshold. Therefore, for scenarios from the first category, not just ES estimates, but also the related VaR forecasts, are inaccurate.
For illustration purposes, Figure 1 depicts the PDF of the forecasted return loss distribution $\widehat{L}_t$ as well as the PDFs of both considered t-distributions with parameter values of $\mu_t = 0$ and $\sigma_t = 1$.

In order to complement the two selected scenarios, I additionally analyse how rejection rates evolve given that the number of considered degrees of freedom further increases beyond a value of $\nu = 5$.

(ii) **Misspecified tail behaviour**
For the second category, I analyse two scenarios where the risk manager only fails to correctly account for the distribution tail of return losses. Nevertheless, compared to the first category I evaluate scenarios where the risk manager correctly models both the conditional mean and variance of the underlying distributions. Thus, the level of risk underestimation in the second category is lower compared to the first category. More precisely, the underestimation of the actual risk will primarily become apparent in the magnitude of violations beyond the estimated VaR but not in their overall number. Thus, the foreasted VaR, for scenarios from the second category, might be roughly accurate, whereas the related ES figure still underestimates the actual underlying risk. Therefore, this makes it also more challenging for any ES backtest to detect a misspecified model, which stems from this second category.
For the first scenario in this category, I simulate observed return losses from a t-distribution with $\nu = 3$ degrees of freedom, which is standardized to have a conditional mean of $\mathbb{E}[L_t] = \mu_t$ and more importantly a conditional variance of $Var(L_t) = \sigma_t^2$. This is achieved by simulating observed return losses from $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$. Correspondingly, the risk manager takes into account the correct conditional variance, but does not account for the fat tails of the t-distribution.
For the second scenario in this category, I simulate return losses from a *skewed normal distribution*, which is again standardized such that both conditional mean and variance are correctly specified. In addition to location and scale parameters, the skewed normal distribution requires an additional shape input parameter $\lambda$ to determine the skewness of the distribution. The properties of the skewed normal distribution are for example outlined in Azzalini and Valle (1996). Given location and scale parameters of $\mu$ and $\sigma$ and a shape parameter of $\lambda$, the PDF of the skewed normal distribution is given by,

$$\phi^{\text{skewed}}(x) := \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right), \tag{5.3}$$

where again $\phi$ and $\Phi$ denote the PDF and the CDF of the standard normal distribution. Moreover, I select a shape parameter of $\lambda = 2$ and standardize both the location and the scale parameter such that $\mathbb{E}[L_t] = \mu_t$ and $Var(L_t) = \sigma_t^2$.[23] The resulting return loss distribution of $L_t$ is right-skewed, which is a realistic property of a financial time series of return losses. Furthermore, the distribution possesses over-kurtosis compared to the standard normal distribution. Thus, the risk manager underestimates the underlying risk, as he fails to correctly estimate both skewness and kurtosis of the return loss distribution. Figure 2 depicts the PDF of the estimated return loss $\widehat{L}_t$ as well as the PDFs of both assumed data generating processes within this category. Again for illustration purposes, location and scale parameters of $\mu_t = 0$ and $\sigma_t = 1$ are considered.

At first, the more severe cases of risk underestimation are evaluated, where both tail behaviour and conditional variance are wrongly estimated. In the first power analysis, observed losses are drawn from $L_t \sim t_3(\mu_t, \sigma_t)$, while the risk manager bases his ES forecasts on the estimate $\widehat{L}_t \sim N(\mu_t, \sigma_t)$. The respective simulated rejection rates are depicted in Table 9. The colour scheme for the power analysis is again motivated by Kratz et al. (2018). Nevertheless, the respective thresholds are taken at stricter values compared to their suggestion. Power values are marked in green given a high rejection rate above 80 %. On the other hand, poor power values are marked in orange given a

[23] For $\lambda = 2$ this is achieved by setting the scale parameter of the skewed normal distribution equal to $\frac{\sigma_t}{\sqrt{1-1.6/\pi}}$ and the location parameter equal to $\mu_t - \frac{\sigma_t}{\sqrt{1-1.6/\pi}}\frac{2}{\sqrt{5}}\sqrt{\frac{2}{\pi}}$.

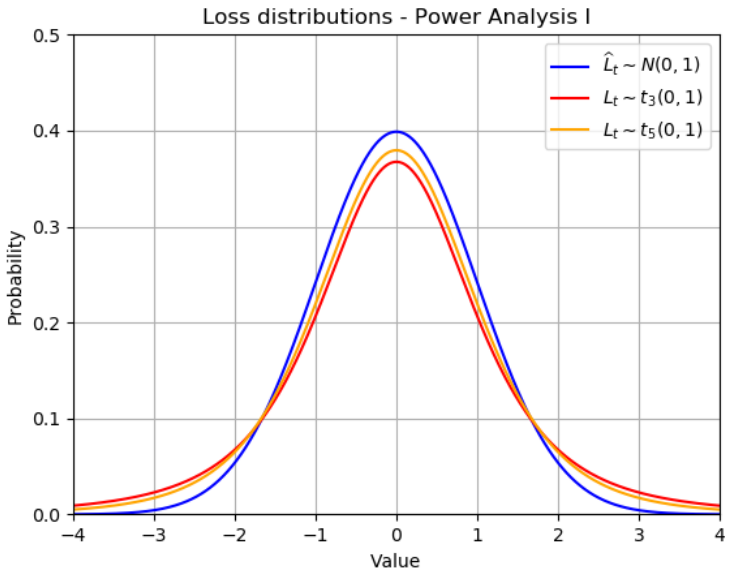


**Figure 1:** Power analysis - Comparison of PDFs corresponding to multiple loss distributions $L_t$ to misspecified distribution $\widehat{L}_t$ with different variance and tail behaviour.

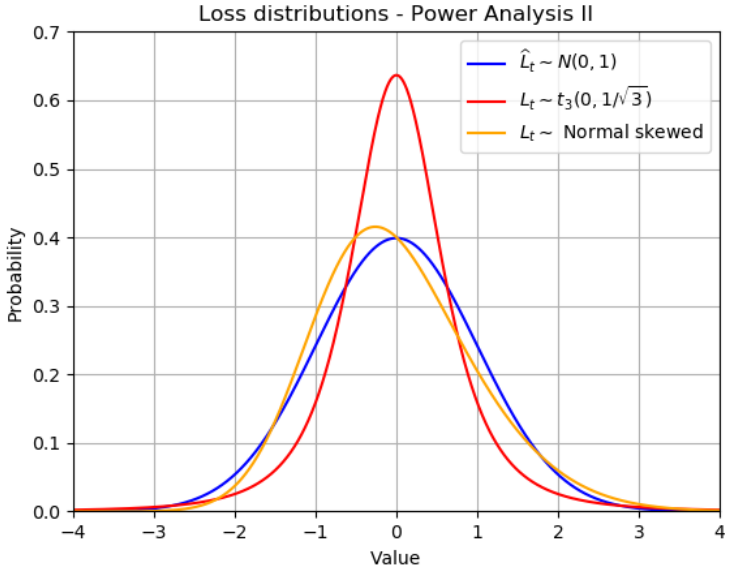


**Figure 2:** Power analysis - Comparison of PDFs corresponding to multiple loss distributions $L_t$ to misspecified distribution $\widehat{L}_t$ with different tail behaviour but same variance.

rejection rate below 40 % and in red if the simulated rejection rate even lies below 20 %. For the power analysis, rejection rates are again based on $MC = 1000$ simulated test decisions. Furthermore, an ES confidence level of $\alpha = 0.975$ and a backtest significance level of $\kappa = 0.05$ are applied for backtesting periods of $T = 250$, 500 and 1000 observations.

First of all, it can be noticed that all ES backtests in scope yield excellent power values in Table 9, as all simulated rejection rates are above 78 %. There is only one rejection rate that is not labelled in green, however only with a value slightly below 80 %. Moreover, this rejection rate belongs to the intercept ESR backtest in case of $T = 1000$ backtesting observations. It is also worth mentioning that rejection rates of the intercept ESR backtest decrease given an increase in the backtesting horizon $T$, which is rather counter-intuitive. Nevertheless, the intercept ESR backtest still exhibits decent power values in all evaluated cases. On the contrary, rejection rates for all other considered ES backtests increase in $T$. Furthermore, for both $T = 500$ and

**Empirical power $L_t \sim t_3(\mu_t, \sigma_t)$ - Rejection rates of evaluated ES backtests**

| | Backtesting period - T | | |
|---|---|---|---|
| **ES backtesting approach** | 250 | 500 | 1000 |
| Multinomial backtest - Original version | **0.988** | **1.000** | **1.000** |
| Multinomial backtest - One-sided version | **0.988** | **1.000** | **1.000** |
| Test 2 | **0.992** | **1.000** | **1.000** |
| Intercept ESR backtest | **0.914** | **0.888** | 0.789 |
| Z-test - Original version | **0.950** | **0.975** | **0.990** |
| Z-test - Approximative version | **0.995** | **1.000** | **1.000** |
| Combined ES residuals backtest | **0.987** | **1.000** | **1.000** |
| (ES residuals backtest) | **(0.881)** | **(0.975)** | **(0.997)** |
| (VaR traffic light test) | **(0.971)** | **(1.000)** | **(1.000)** |

**Table 9:** Empirical power, i.e. rejection rates of evaluated ES backtests. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t)$ for all $t \in \{1, \dots, T\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Backtesting horizons of $T = 250$, 500 and 1000 observations are considered.

$T = 1000$ observations, a majority of the backtests in scope reject the wrongly specified risk forecast at a perfect rate of 100%. It is in line with expectations, that power values increase given an increase in $T$, as a larger backtesting period also involves more misspecified risk forecasts into the respective test decision.

For the multinomial backtest there is no difference in rejection rates between the original and the de-facto one-sided version. As outlined in the previous chapter, the only difference between both multinomial tests is, that the adjusted approach additionally rules out the rejection of any conservative estimation model. Nevertheless, for the power analysis, only scenarios where the actual risk is underestimated are considered. Therefore, it is reasonable that rejection rates of both multinomial test versions coincide for any power scenario. For the Z-test proposed by Costanzino and Curran (2015) it can be noted that the approximative version, proposed within this thesis, yields slightly higher rejection rates compared to the original approach. Overall, all evaluated backtests show decent results in detecting the risk underestimation in a scenario where $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t)$.

In order to stepwise increase the challenge, I secondly consider a scenario where $L_t \sim t_5(\mu_t, \sigma_t)$. As outlined above, increasing the degrees of freedom decreases the degree of misspecification of both the distribution tail and the conditional variance. Thus, it is more difficult to detect, that the risk manager still underestimates the actual risk. The respective rejection rates for this scenario are depicted in Table 10 below. Both the colour scheme and the selected backtesting parameters are equivalent to the previous table.

Similar as in the previous scenario, all considered ES backtests exhibit excellent rejection rates, most of them again above 80 %. Nevertheless, the overall level of rejection rates is slightly lower compared to the previous simulation, which is in line with expectations given the decreased level of model misspecification. As an example, for the simulation of $L_t \sim t_3(\mu_t, \sigma_t)$, multiple backtests achieved a perfect rejection rate already at a backtesting horizon of $T = 500$ observations, whereas for this second scenario, using $L_t \sim t_5(\mu_t, \sigma_t)$, perfect rejection rates of 100 % are only achieved for the maximum considered backtesting time frame of $T = 1000$ observation. Thus, a risk manager needs to consider a longer time series of both risk forecasts and P&L realizations, to achieve the same backtesting accuracy as in the previous scenario. Furthermore, all ES backtests analysed in Table 10 generally, show increasing rejection rates given an increase in the backtesting horizon $T$. One can also notice, that the original Z-test proposed by Costanzino and Curran (2015) is again slightly out-performed by its approximative version, which was suggested within this thesis.

Although there are some minor differences between the ES backtests in scope, all of them show excellent empirical power values in both considered scenarios, where the risk manager misspecifies both the tail behaviour and the conditional variance. This is also in line with previous power analyses conducted for example by Acerbi and Szekely (2014). In order to evaluate how rejection rates further develop if the true return loss distribution approaches the normal distribution, I outline some additional analysis in the following. Moreover, I simulate $MC = 100$ test decisions for each scenario $L_t \sim t_\nu(\mu_t, \sigma_t)$, whereas $\nu$ takes on values in $\{3, \dots, 25\}$, and the estimated return loss $\widehat{L}_t$ is again assumed to be normally distributed. Furthermore, I choose the same backtesting parameters as in the previous scenarios, but fix a backtesting horizon of $T = 500$ observations. Thus, rejection rates are simulated for the same type of model misspecification as outlined in the previous two scenarios, but with a decreasing level of estimation inaccuracy. The resulting plot of rejection rates, depending on the considered degrees of freedom, for all considered backtests, is depicted in Figure 3[24] .

[24]Note that both the original two-sided and the de-facto one sided multino-

**Empirical power $L_t \sim t_5(\mu_t, \sigma_t)$ - Rejection rates of evaluated ES backtests**

| | Backtesting period - T | | |
|---|---|---|---|
| **ES backtesting approach** | 250 | 500 | 1000 |
| Multinomial backtest - Original version | **0.836** | **0.956** | **1.000** |
| Multinomial backtest - One-sided version | **0.836** | **0.956** | **1.000** |
| Test 2 | **0.900** | **0.957** | **0.995** |
| Intercept ESR backtest | **0.847** | **0.963** | **0.995** |
| Z-test - Original version | **0.817** | **0.862** | **0.884** |
| Z-test - Approximative version | **0.950** | **0.995** | **1.000** |
| Combined ES residuals backtest | **0.895** | **0.992** | **1.000** |
| (ES residuals backtest) | (0.655) | **(0.895)** | **(0.994)** |
| (VaR traffic light test) | **(0.801)** | **(0.971)** | **(1.000)** |

**Table 10:** Empirical power, i.e. rejection rates of evaluated ES backtests. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_5(\mu_t, \sigma_t)$ for all $t \in \{1, \ldots, T\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Backtesting horizons of $T = 250$, 500 and 1000 observations are considered.

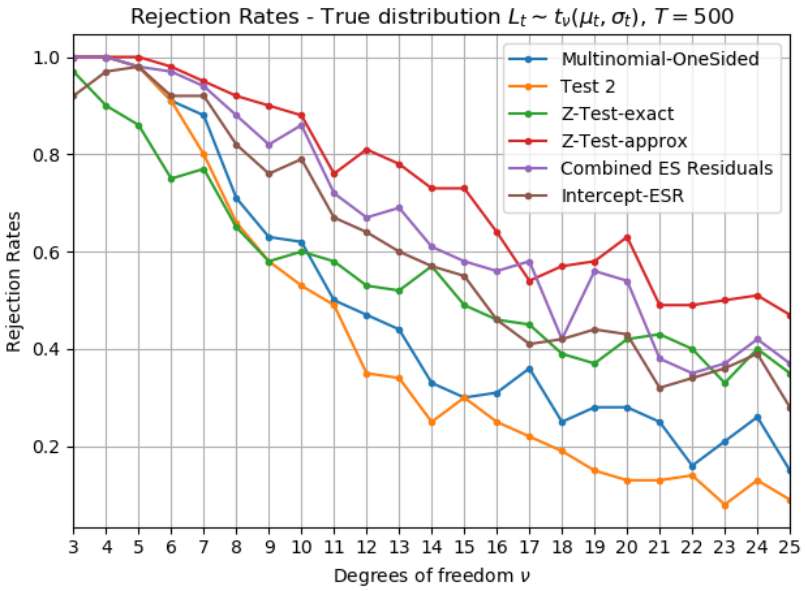


**Figure 3:** Rejection rates for all considered ES backtests depending on degree of model misspecification. Moreover, $MC = 100$ test decisions are simulated for each scenario $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_\nu(\mu_t, \sigma_t)$, whereas $\nu$ takes on values in $\{3, \ldots, 25\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtest set at $\kappa = 0.05$, based on a backtesting horizon of $T = 500$ observations.

Overall, the results depicted in Figure 3 are in line with expectation, as rejection rates decrease given an increase in degrees of freedom. Furthermore, up to a value of $\nu = 10$ all considered ES backtests reject the underestimated ES forecast in more than every second test decisions. For larger values of $\nu$ around 20, both applied Z-tests, the combined ES residuals backtest as well as the intercept ESR backtest display slightly higher rejection rates compared to both, the multinomial backtest as well as the Test 2. Nevertheless, for high values of $\nu$, minor differences between the single backtests should not be exaggerated, given the rather low degree of model inaccuracy. Overall, the power of all ES backtests in scope is deemed appropriate for scenarios from the first category of model misspecification.

In the following the two selected scenarios from the second category are analysed, where only the risk in the distribution tail is underestimated, but the conditional first two moments are correctly specified, i.e. $\mathbb{E}[\widehat{L}_t] = \mathbb{E}[L_t]$ and $Var(\widehat{L}_t) = Var(L_t)$ for

mial backtest achieve exactly the same rejection rates in the conducted analysis. Thus, only the one-sided version is included in Figure 3.

all observations $t$ within the backtesting period. Thus, the underestimation of the actual risk only becomes perceivable in the outer tail of the return loss distribution. As previously argued, this primarily affects the magnitude of violations beyond the respective VaR threshold, but not there overall number. Therefore, it is more challenging for any ES backtest to detect a misspecified model from this second considered category.

Table 11 depicts the simulated rejection rates for the scenario where the true return loss variable follows a standardized t-distribution with $\nu = 3$ degrees of freedom.

Again rejection rates are based on $MC = 1000$ simulated test decisions given an ES confidence level of $\alpha = 0.975$ and a significance level of $\kappa = 0.05$ for the considered backtests. The colouring scheme is equivalent to the previous tables.

Overall, it can be recognized that power values are by far worse given that $L_t$ follows a standardized t-distribution compared to the previous simulations, where the variance of the t-distribution was not adjusted. For the backtesting period of $T = 250$ days, stipulated by the Basel Committee, non of the evaluated backtests manages to reject the misspecified ES estimation model in at least half of all simulated decisions. For most backtests in scope, rejection rates tend to increase given an increase in the considered backtesting horizon.

Nevertheless, at least for two out of the five overall testing methodologies, the power results are not even satisfying for the maximum backtesting period of $T = 1000$ days. Both, the original Z-test proposed by Costanzino and Curran (2015) as well as the related approximative version, reject the underestimated ES estimate in less than one third of all test decisions for any considered backtesting period. Although rejection rates for both Z-tests slightly increase in $T$, both tests are not able to consistently detect this type of model misspecification. Furthermore, results obtained for the Test 2 from Acerbi and Szekely (2014) are even more problematic, as all simulated rejection rates for this test are labelled in red, with values even below 20 %. The rejection rates even further decrease given an increase in the backtesting time frame. Thus, the Test 2 does not appear to have any power to detect this kind of risk underestimation. Taking into account the size simulations in the previous subsection, the Test 2 exhibits a potential imbalanced relation between both size and power. Indeed, extremely low size values might come at the cost of a low power in detecting certain types of model misspecification.

On the contrary, at least for the maximum considered backtesting period, both multinomial backtest versions as well as the intercept ESR backtest achieve rejection rates above 50 %. Moreover, the rejection rate for the combined ES residuals backtest is even labelled in green with an excellent value of 83.7 %, given $T = 1000$ backtesting observations are taken into account. As the risk manager correctly estimates the conditional variance of return losses, it is not surprising that the VaR traffic light test produces rejection rates close to zero, as the misspecification only gets perceivable in the outer distribution tail. Still, in combination with the original ES residuals backtest proposed by McNeil and Frey (2000), the combined test shows the best empirical power values amongst all evaluated approaches in this simulation scenario.

For the last considered power simulation, true return losses are simulated from a skewed normal distribution, which is standardized such that both conditional mean and variance coincide with the respective figures of the estimated return loss variable $\widehat{L}_t$. Whereas $\widehat{L}_t$ models a symmetric return loss distribution, $L_t$ is right-skewed and also possesses a slightly higher kurtosis. The simulated rejection rates are depicted in Table 12 below. All backtesting parameters are set equivalent to the previous table.

First of all, it can be noticed that power figures are on a slightly higher level compared to the previous scenario. This is in line with expectations, as the true return loss distribution $L_t$ is both skewed and fat tailed, which induces a slightly higher degree of misspecification in the distribution tails compared to the previous simulation. For this last simulated power scenario, it becomes most apparent, that an increasing backtesting horizon is related to higher empirical power values. While most simulated rejection rates do not exceed 50 % for $T = 250$ backtesting observation, the majority of the backtesting approaches in scope does a decent job in detecting the misspecified model in case a backtesting horizon of $T = 1000$ is considered. Especially, the intercept ESR backtest, the approximative Z-test as well as the combined ES residuals backtest need to be highlighted with excellent rejection rates above 90 %, given the maximum considered backtesting period. As an exception, again the Test 2 from Acerbi and Szekely (2014) is not able to consistently detect this kind of risk underestimation. Indeed, for any of the considered backtesting time frames, the Test 2 detects the wrongly specified estimation model only in about 40 % of all simulated test decisions.

Concluding this subsection, all considered backtests in scope achieve excellent empirical power values in scenarios where the risk manager underestimates both the conditional variance as well as the distribution tail of return losses. On the contrary, it is far more difficult for any ES backtest to detect scenarios where the first two moments of return losses are correctly estimated but only the distribution tail is underestimated. Nevertheless, for these last two scenarios from the second category, differences between the considered ES backtests regarding their empirical power become most apparent. The Test 2 from Acerbi and Szekely (2014) performs worst in these two simulations, as it only exhibits few power in detecting the respective misspecified estimation models. As previously argued, this might be related to an imbalanced relation of both empirical size and power observed for the Test 2. The most consistent power values across both scenarios form the second category are achieved by, both multinomial backtests, the intercept ESR backtest as well as the combined ES residuals backtest, with decent rejection rates, at least for the maximum backtesting horizon of $T = 1000$ observations. Overall, it should also be noted that the approximative Z-test slightly outperforms the original test version proposed by Costanzino and Curran (2015). Although this appears to be surprising, this trend can be observed throughout all four conducted power simulations. Furthermore, in line with expectations, rejection rates of both multinomial test version coincide in all conducted power analyses.

Compared to the size analysis in the previous subsection, one

**Empirical power $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$**
**- Rejection rates of evaluated ES backtests**

| | Backtesting period - T | | |
|---|---|---|---|
| **ES backtesting approach** | 250 | 500 | 1000 |
| Multinomial backtest - Original version | **0.149** | **0.287** | 0.616 |
| Multinomial backtest - One-sided version | **0.149** | **0.287** | 0.616 |
| Test 2 | **0.152** | **0.080** | **0.028** |
| Intercept ESR backtest | **0.207** | **0.273** | 0.509 |
| Z-test - Original version | **0.176** | **0.231** | **0.293** |
| Z-test - Approximative version | **0.204** | **0.233** | **0.324** |
| Combined ES residuals backtest | 0.414 | 0.547 | **0.837** |
| (ES residuals backtest) | (0.408) | (0.543) | **(0.837)** |
| (VaR traffic light test) | **(0.018)** | **(0.028)** | **(0.015)** |

**Table 11:** Empirical power, true dist. $t_3$ standardized - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$ for all $t \in \{1, \ldots, T\}$.

**Empirical power $L_t \sim$ skewed normal -**
**Rejection rates of evaluated ES backtests**

| | Backtesting period - T | | |
|---|---|---|---|
| **ES backtesting approach** | 250 | 500 | 1000 |
| Multinomial backtest - Original version | **0.308** | 0.488 | 0.755 |
| Multinomial backtest - One-sided version | **0.308** | 0.488 | 0.755 |
| Test 2 | **0.396** | **0.388** | 0.410 |
| Intercept ESR backtest | 0.467 | 0.701 | **0.906** |
| Z-test - Original version | 0.442 | 0.594 | 0.603 |
| Z-test - Approximative version | 0.594 | **0.833** | **0.968** |
| Combined ES residuals backtest | 0.457 | 0.747 | **0.927** |
| (ES residuals backtest) | **(0.179)** | **(0.358)** | (0.625) |
| (VaR traffic light test) | **(0.349)** | (0.633) | **(0.821)** |

**Table 12:** Empirical power, true distribution skewed normal - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t$ follows a skewed normal distribution, which is standardized such that first two moments of $\widehat{L}_t$ and $L_t$ coincide for all $t \in \{1, \ldots, T\}$.

can also observe an obvious impact of the backtesting time frame on power values. Especially, for both scenarios from the second category ES backtests are generally more powerful given a larger backtesting horizon. Moreover, for the last two scenarios, the backtesting period of $T = 250$ observations, suggested by the Basel Committee, does not appear to be sufficient in order to consistently detect misspecified ES estimation models.

Summing up this chapter, I propose the following qualitative judgement of all ES backtesting methodologies in scope according to their results in all conducted size and power simulations. All backtests are graded according to both their empirical size and power values. In each category **(++)** corresponds to an excellent performance, followed by **(+)**, which still characterises an above average results. Average power or size results are labelled by **(o)**, whereas poor performance is marked with a **(-)** or even with a **(- -)** if it displays a serious issue for any practical implementation. The judgement of the backtests in scope as well as a short explanation is outlined below.

- **Multinomial backtest - Original and de-facto one-sided version :**
  **Size: (++)** Excellent empirical size for both the original and the de-facto one-sided version. Furthermore, the de-facto one-sided version avoids the rejection of any

conservative ES estimation model.
**Power: (+)** Excellent empirical power for scenarios from first category. Still decent power for scenarios from second category, given a sufficient backtesting horizon. As to be expected, there are no differences in power values between both multinomial test versions.

- **Test 2:**
  **Size: (+)** Very low empirical size, which is slightly below backtesting significance level and further decreases given an increase in $T$.
  **Power: (-)** Excellent empirical power for scenarios from first category. On the contrary, low power in detecting misspecified models from the second category. Overall, the obtained results hint towards a slight misalignment of both size and power.

- **Intercept ESR backtest:**
  **Size: (+)** Decent empirical size, which is nevertheless slightly above backtesting significance level.
  **Power: (+)** Excellent empirical power for scenarios from first category. Still decent power for scenarios from second category, given a sufficient backtesting horizon.

- **Z-test - Original and approximative version:**
  **Size: (- -)** Both the original as well as the approximate version are oversized. Furthermore, empirical size further increases with increasing values of $T$. High size values indicate a potential misspecification in the underlying test framework.
  **Power: (+)** Excellent power for scenarios from first category, but both versions can not consistently detect misspecified models from second category. Overall, approximative version yields slightly superior power results compared to the original approach.

- **Combined ES residuals backtest:**
  **Size: (o)** Backtest is slightly oversized, but might be adjusted by calibration of single test components.
  **Power: (++)** Excellent empirical power for scenarios from first category and also for the scenarios from the second category, given a sufficient backtesting horizon $T$.

Therefore, according to both empirical size and power, the de-facto one-sided version of the multinomial backtest yields the best results followed by both the intercept ESR backtest and the combined ES residuals backtest.
In the consecutive chapter, all backtests in scope are tested on reasonable ES estimation models, which are fitted to log-return losses of the S&P 500.

## 6. Application of backtests to real data

The object of this chapter is to apply the relevant ES backtests outlined in the fourth chapter of this thesis to actual financial time series. Thus, this chapter aims to evaluate whether the ES backtests in scope yield reasonable results in practical scenarios and therefore qualify for the use by both financial institutions and regulators. For this purpose, univariate price data of the S&P 500 index is considered over two distinct time periods, which function as the respective backtesting horizons in this context. The consideration of two different time periods of S&P 500 data allows for an evaluation of the introduced ES backtests in different market environments. The evaluated data set is further described in the consecutive subsection. Afterwards, five ES estimation models, with different degrees of complexity and forecasting accuracy, are estimated for every day within each of the two considered time frames of S&P 500 data. The respective risk models are further outlined in subsection 6.2. In the subsection thereafter, the estimated ES forecasts are evaluated based on all ES backtesting methodologies introduced in the course of this thesis. Similar as for the VaR traffic light approach, backtesting significance levels of both $\kappa = 0.05$ and $\kappa = 0.0001$ are considered in order to rank the ES estimation models with respect to the traffic light system prescribed by the Basel Committee.

### 6.1. Financial data - S&P 500 index

For the analysis conducted within this chapter, daily closing prices of the S&P 500 index are taken from the *Thomson Reuters Datastream* database for the time period from 01.01.2006 to 01.10.2019. In order to remain consistent with the previous notation, the S&P 500 price data is transformed to a time series of log return losses. More precisely, if $P_t$ and $P_{t-1}$ denote the prices of the S&P 500 at times $t$ and $t-1$, the respective log return loss is defined as,

$$l_t := -(\ln(P_t) - \ln(P_{t-1})). \tag{6.1}$$

Public holidays without any trading activities, i.e. with a related log return loss of exactly zero, are removed from the overall sample, as they would potentially bias the obtained backtesting results.
As shortly mentioned above, the overall sample is divided into two distinct time periods. The first one is labelled as the *crisis period* and contains a backtesting horizon of exactly 500 trading days, which spans from 27.12.2007 to 21.12.2009. All ES estimates, $\widehat{ES}_{t,0.975}$, are forecasted based on the rolling window of the previous 250 log return losses.[25] The crisis period includes the peak of the financial crisis in autumn 2008. Thus, this highly volatile market regime induces a major challenge for any of the applied risk estimation models.
The second considered backtesting period within this chapter contains the most recent S&P 500 data available at the start of the processing period for this thesis. Respectively, the period is labelled as *recent period* for the remainder of this thesis and includes 500 observations from 27.09.2017 to 23.09.2019. Again, all ES forecasts are estimated based on the rolling window of the previous 250 log return losses. [26] Compared to the crisis

[25] Given the rolling estimation window of 250 trading days, S&P 500 market data is considered starting from 28.12.2006, prior to the beginning of the crisis period, for estimation purposes.

[26] For the recent period, S&P 500 market data used for estimation purposes dates back until 28.09.2016.

period, the recent period displays a time frame with rather calm market conditions.

The differences in the volatility of the S&P 500 between the two considered backtesting periods can be observed in Figure 4, which depicts the log return losses over the respective time frames. Additionally, Figure 4 reveals that log return losses of the S&P 500 exhibit heteroskedasticity, which is a well known pattern of financial time series. Moreover, volatility clusters can be observed within both considered periods.

Another feature of log-return series is that they are often leptokurtic, i.e. they possess over-kurtosis compared to the standard normal distribution. In order to assess if log-return losses of the S&P 500 display leptokurtic patterns, Figure 5 depicts the QQ-plots of log return losses of the S&P 500 in percent against the standard normal distribution for both considered time periods. Especially in the crisis period, it is obvious that the considered log return losses posses fatter tails compared to the standard normal distribution, which can be observed on the left subplot of Figure 5. Even in the recent period, in a rather calm market environment, log return losses are still slightly leptokurtic. Nevertheless, distribution tails are by far thinner in the recent period compared to the crisis period.

Additionally, the differences in the distribution tails for both time windows of log return losses can be observed in Figure 6, which shows the respective histograms of log return losses. There are two further aspects which can be gathered from Figure 6. First of all, log return losses in both subplots of Figure 6 are centred around zero, which will be relevant for the subsection thereafter. Indeed, several of the applied ES estimation models are based on the assumption that average daily log return losses do not differ from zero. Secondly, log return losses over both considered backtesting time frames are slightly skewed to the right. As mentioned before, this is also a typical pattern of return loss series.

As outlined within this subsection, log return losses of the S&P 500 over both considered time frames exhibit typical patterns of financial time series. Overall, I decided to evaluate the ES backtests in scope on two distinct backtesting periods, as they display different market environments. On the one hand, in the recent period most ES estimation models are expected to work properly. Correspondingly, most ES estimation models should also pass any ES backtest over the recent period. On the other hand, it is far more challenging to make accurate ES forecasts in times of financial distress. Therefore, a majority of the ES estimation models might fail on the introduced backtests within the crisis period. Overall, regulators need to assure sufficient capital buffers irrespective of the underlying market environment. In a crisis period, a reasonable ES backtest needs to be penalizing and reject risk estimation models which do not quickly adopt to financial distress. On the contrary, in calm financial markets ES backtests should reward a sufficient risk coverage and not force a financial institution to hold unnecessarily high safety buffers.

Furthermore, the length of both considered backtesting horizons, i.e. exactly 500 observations, is motivated by the simulation results obtained within the previous chapter. As previously noticed, a backtesting horizon of 250 observations might be too short in order to consistently detect the risk underestimation inherited in an ES estimation model. On the other hand, excessively large backtesting windows might be problematic in practical application as they require the storage and moreover the availability of risk forecasts and return data over a long time frame. Thus, I believe the chosen backtesting horizon of $T = 500$ days gives a good trade off between backtesting power and data requirements, such that the conducted analysis displays a realistic set-up for an application by financial institutions or regulators.

The following subsection is first going to shortly summarize all five considered ES estimation models before the actual backtesting results can be discussed.

### 6.2. ES estimation models

This subsection presents all five applied ES estimation models, which are fitted to the S&P 500 log return loss data. All implemented ES estimation models differ with respect to their complexity and thus also yield different accuracies in modelling the underlying risk. Therefore, it is also interesting to observe, whether the applied ES backtests are able to discriminate between the different types of estimation models, in the following subsection.

The first, rather naive, approach corresponds to the ES estimation procedure applied within the size and power simulation studies. Thus, ES forecasts are based on a normal distribution which is fitted to the previous 250 observations. The second, third and fourth approach are also parametric ES estimation models, which apply econometric techniques in order to account for time varying conditional volatility of log return losses. Moreover, the respective econometric models all belong to the class of *Generalized Autoregressive Conditional Heteroskedasticity (GARCH)* processes, which were introduced by Bollerslev (1986). The three respective estimation models differ in their ability to capture the fat tails and the right-skewness of log return losses. The fifth ES estimation model is a *Filtered Historical Simulation (FHS)* with volatility updating, whereas conditional volatility is modelled with the RiskMetrics methodology suggested by JP Morgan (1996). Compared to all other considered models, the applied FHS it the only non-parametric approach and therefore requires no distributional assumption. As a consequence, comparable FHS methodologies are often applied by practitioners. In the following, the estimation procedure for all five approaches is shortly outlined.

(I) **ES-norm**
This first ES estimation model is labelled as *ES-norm*. Similar to the previous chapter, a normal distribution is fitted to the rolling window of the previous 250 S&P 500 log-return losses, such that $\widehat{L}_t \sim N(\widehat{\mu}_t, \widehat{\sigma}_t)$ for any observation $t$ within the respective backtesting horizon. Moreover, the corresponding VaR and ES estimates can easily

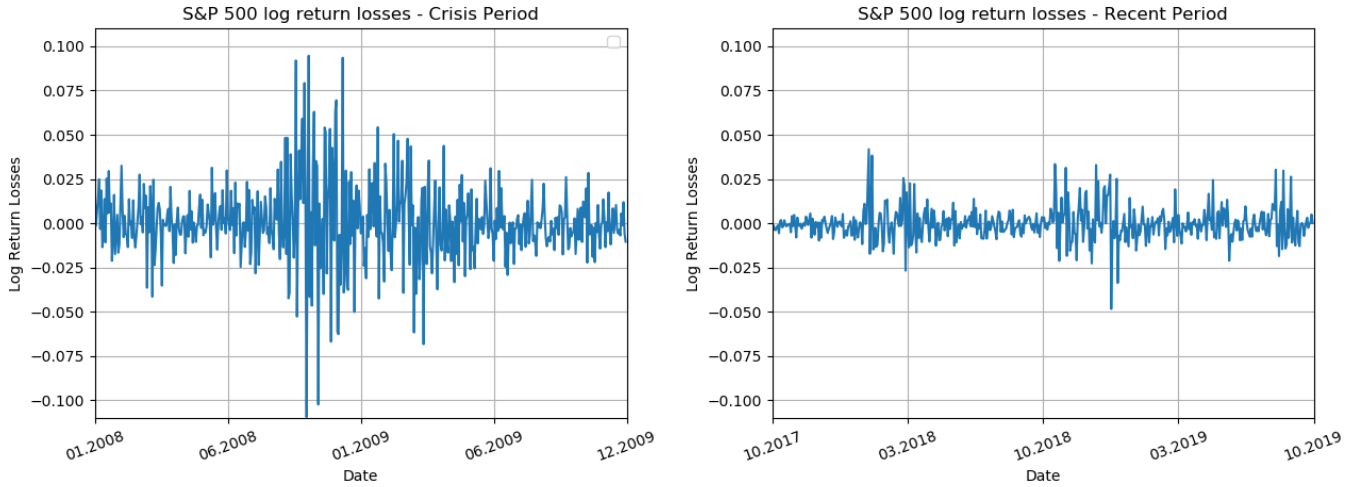


**Figure 4:** Plot of S&P 500 log return losses. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

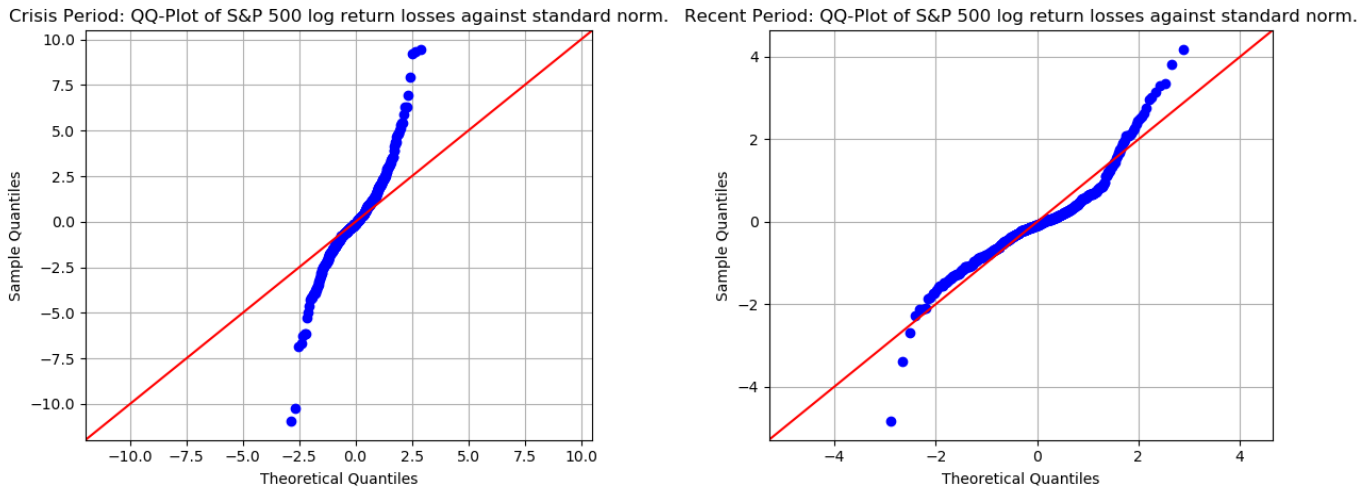


**Figure 5:** QQ-Plot of S&P 500 log return losses in percent against the standard normal distribution. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

be calculated as,[27]

$$\widehat{VaR}_{t,\alpha} = \widehat{\mu}_t + \widehat{\sigma}_t \Phi^{-1}(\alpha), \tag{6.2}$$

$$\widehat{ES}_{t,\alpha} = \widehat{\mu}_t + \widehat{\sigma}_t \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}, \tag{6.3}$$

where again $\phi$ and $\Phi$ denote the PDF and the CDF of the standard normal distribution, respectively. The ES-norm estimation model is a popular approach because of its simplicity. Nevertheless, risk estimates in the ES-norm model only slowly adopt to any changes in the volatility environment. Furthermore, especially in the evaluated crisis period, the distribution tails of S&P 500 log return losses heavily deviate from those of a standard normal distribution, which questions the assumption of normally distributed log return losses used in ES-norm. The resulting ES estimates and the respective S&P 500 log return losses, for both the crisis and the recent period, are depicted in Figure 7.

(II) **ES-GARCH-norm**

The second ES estimation model is labelled as *ES-GARCH-norm*. Compared to the first approach, conditional volatility is fitted to a GARCH(1,1) model in order to better capture the dynamics of volatility. For the second, as well as the third and fourth approach I assume that the daily conditional mean of log return losses

[27] See McNeil, Frey, and Embrechts (2015) pages 65 and 70 for a derivation of both formulas.

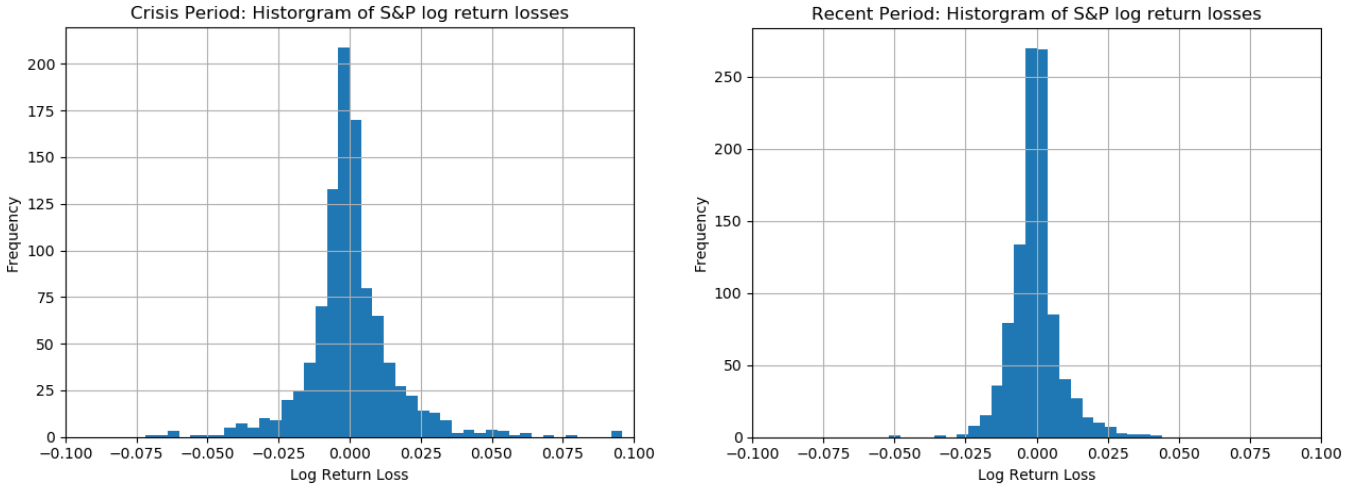


**Figure 6:** Histogram of S&P 500 log return losses. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

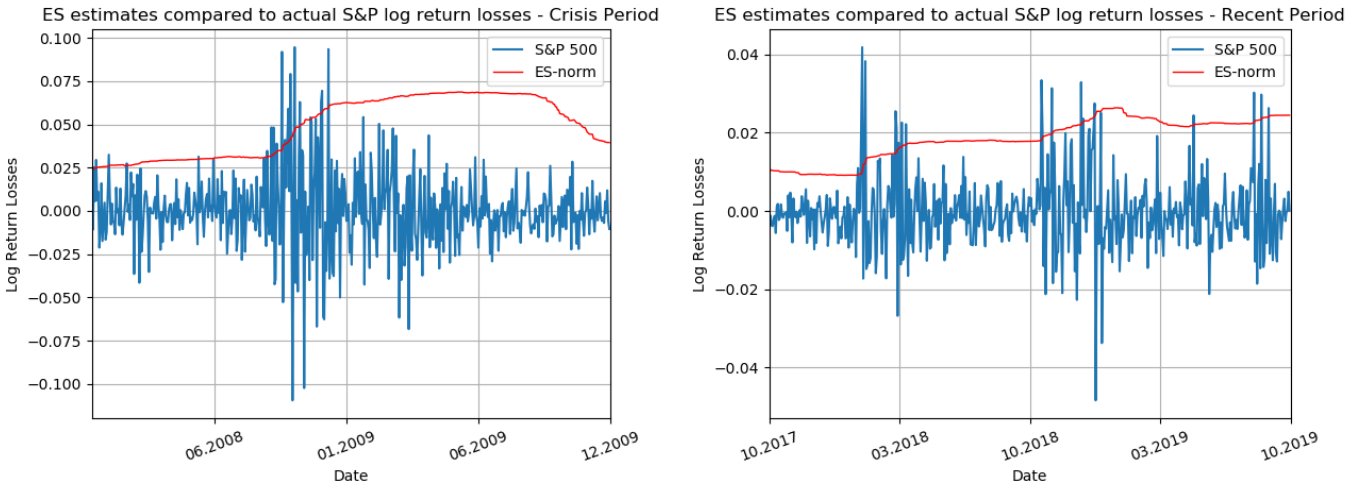


**Figure 7:** ES estimations based on fitted normal and t-distribution. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

is zero, which is a reasonable assumption given the argumentation in the previous subsection. Moreover, log return losses at time $t$ are specified by,

$$l_t = \sigma_t \varepsilon_t, \tag{6.4}$$

where $\sigma_t$ again denotes the conditional volatility of log-return losses and $\varepsilon_t$ is an innovation term. For the considered GARCH(1,1) model, innovations are assumed to be i.i.d. normally distributed, i.e. $\varepsilon_t \overset{i.i.d}{\sim} N(0,1)$. Furthermore, the volatility process is given by,

$$\sigma_t^2 = \omega + \alpha l_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{6.5}$$

Thus, the volatility at any time $t$ within the backtesting horizon depends on both the squared log return loss and the squared volatility from the previous time step.[28] For every observation $t$ within the backtesting horizon, the parameter vector $(\omega, \alpha, \beta)$ is fitted to the rolling window of the previous 250 log return losses, which is done in *Python* by applying a Maximum Likelihood (ML) estimation. Consecutively, one step ahead volatility forecasts $\widehat{\sigma}_t$ are calculated from the obtained parameters based on the volatility specification in (6.5). As outlined in McNeil et al. (2015) page 133, the respective VaR and ES

[28] As outlined above, the theory on a GARCH(1,1) model with normally distributed innovations $\varepsilon_t$ was originally derived by Bollerslev (1986).

forecasts can be calculated as,

$$\widehat{VaR}_{t,\alpha} = \widehat{\sigma}_t q_\alpha(\varepsilon_t) = \widehat{\sigma}_t VaR_{t,\alpha}(\varepsilon_t), \quad (6.6)$$

$$\widehat{ES}_{t,\alpha} = \widehat{\sigma} ES_{t,\alpha}(\varepsilon_t). \quad (6.7)$$

Given the standard normal distribution of innovations, $VaR_{t,\alpha}(\varepsilon_t)$ and $ES_{t,\alpha}(\varepsilon_t)$ can be calculated according to formulas (6.2) and (6.3).
Overall, ES estimations obtained by the ES-GARCH-norm model are far more flexible to incorporate the current market environment compared to the previous rather static ES-norm approach. The resulting ES forecasts are depicted in Figure 8 for both the crisis and the recent period, together with the forecasts of all applied GARCH-type estimation models.

(III) **ES-GARCH-t**
The third ES estimation model, which is labelled as *ES-GARCH-t*, is closely related to the previous estimation approach. Again, log return losses are modelled by equation (6.4) and the conditional volatility is specified as in equation (6.5). In comparison to the previous ES estimation model, innovations $\varepsilon_t$ are assumed to follow an i.i.d. standard t-distribution with $\nu$ degrees of freedom, i.e. $\varepsilon_t \overset{i.i.d.}{\sim} t_\nu$.
In this case not only the GARCH(1,1) parameters need to be estimated, but additionally also the degrees of freedom of the innovation term. Thus, for every $t$ in the backtesting horizon the vector $(\nu, \omega, \alpha, \beta)$ is estimated in a ML procedure based on the rolling window of the previous 250 log return loss observations. Again, one step ahead volatility forecasts $\widehat{\sigma}_t$ are calculated based on the specified volatility process and the obtained GARCH(1,1) parameters. In the following, the VaR and ES forecasts can again be calculated according to formulas (6.6) and (6.7). In this case, it needs to be noted that innovations follow a standard t-distribution with $\nu$ degrees of freedom, such that the overall ES and VaR estimates in the ES-GARCH-t model are given by,

$$\widehat{VaR}_{t,\alpha} = \widehat{\sigma}_t VaR_{t,\alpha}(\varepsilon_t) = \widehat{\sigma}_t t_\nu^{-1}(\alpha), \quad (6.8)$$

$$\widehat{ES}_{t,\alpha} = \widehat{\sigma}_t ES_{t,\alpha}(\varepsilon_t) = \widehat{\sigma}_t \frac{g_\nu(t_\nu^{-1}(\alpha))}{1-\alpha}\left(\frac{\nu + (t_\nu^{-1}(\alpha))^2}{\nu - 1}\right), \quad (6.9)$$

where $g_\nu$ and $t_\nu$ denote the PDF and the CDF of the standard t-distribution, respectively.[29]
The resulting ES forecasts for the ES-GARCH-t model are also depicted in Figure 8, for both the crisis and the recent period. From a theoretical point of view, the ES-GARCH-t approach is superior in accounting for the fat tails of the S&P 500 log return losses compared to the ES-GARCH-norm model. More generally speaking, Angelidis, Benos, and Degiannakis (2004) for example concludes, that GARCH models with heavy tailed innovation distributions should be preferred as they better capture the leptokurtic behaviour of financial data.

(IV) **ES-EGARCH-t**
The fourth considered ES estimation model is labelled as *ES-EGARCH-t*. Compared to the previous two models, the ES-EGARCH-t uses an extension of the classical GARCH framework in order to model conditional volatility. Moreover an *exponential GARCH (EGARCH)* is applied, which was originally introduced by Nelson (1991). For the selected EGARCH(1,1) model specification, log return losses are again specified as in formula (6.4), whereas innovations are assumed to follow an i.i.d. standard t-distribution with $\nu$ degrees of freedom, i.e. $\varepsilon_t \overset{i.i.d}{\sim} t_\nu$. In comparison to the previous two models, the conditional volatility process is defined through,

$$h_t = \omega + \alpha \frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}} + \gamma\left(\left|\frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right| - \mathbb{E}\left[\left|\frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right|\right]\right) + \beta h_{t-1}, \quad (6.10)$$

where $h_t = \ln(\sigma_t^2)$. Similar as for the previous models, all relevant parameters of the EGARCH specification are fitted for every $t$ within the backtesting horizon on the rolling window of the preceding 250 log return losses. In this case the parameter vector $(\nu, \omega, \alpha, \gamma, \beta)$ is fitted in every ML estimation. Afterwards, one step ahead volatility forecasts $\widehat{\sigma}_t$ based on equation (6.10) and the obtained parameter values are calculated. Moreover, the respective VaR and ES forecasts for the ES-EGARCH-t model are obtained analogous to equations (6.8) and (6.9), as innovations are assumed to follow an i.i.d standard t-distribution with $\nu$ degrees of freedom.
From a theoretical perspective, the EGARCH model is able to capture the so called *leverage effect*, which is often observed in financial time series and which is the main cause for right-skewed log returns loss series. In practice, volatility often increases more following a high return loss compared to a gain of the same magnitude. In the GARCH methodology the past log return losses have a symmetric impact on conditional volatility, which can be recognized from equation (6.5). The EGARCH methodology on the contrary, allows to model an asymmetric impact depending on the sign of observed log return losses. Amongst researchers, the EGARCH(1,1) specifications is widely used in various applications to capture the behaviour of conditional volatility.
The ES estimates of the ES-EGARCH-t model, for both the crisis and the recent period, are also depicted in Figure 8.

[29] See McNeil et al. (2015) pages 66 and 71 for a detailed description of the calculation of $VaR_{t,\alpha}(\varepsilon_t)$ and $ES_{t,\alpha}(\varepsilon_t)$ given that $\varepsilon_t \overset{i.i.d}{\sim} t_\nu$.

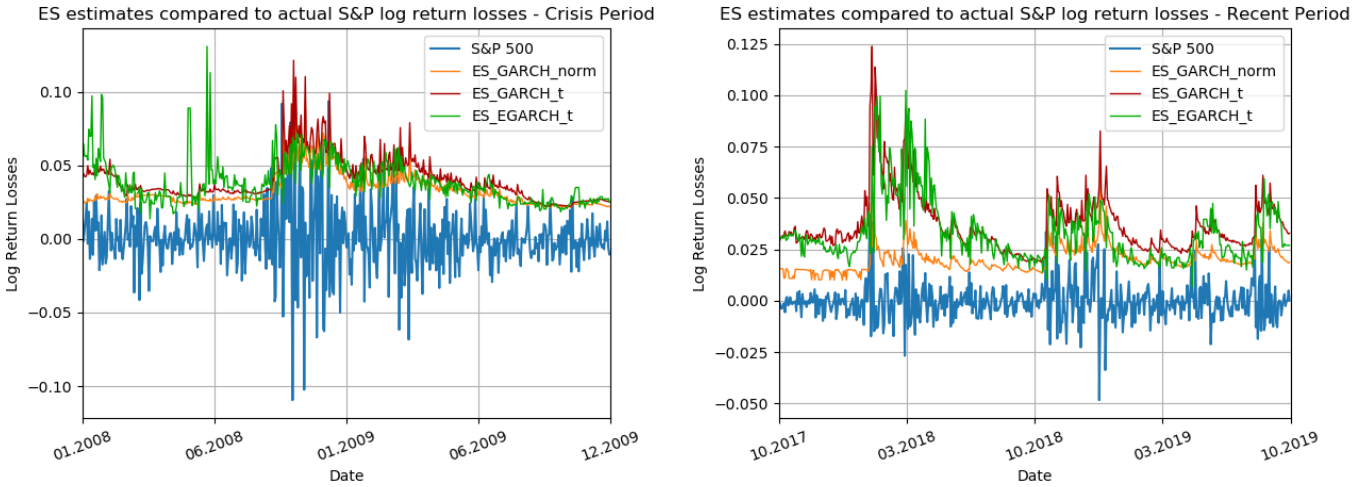


**Figure 8:** ES estimations based on GARCH-type specifications of conditional volatility. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

Overall, all three GARCH-type ES forecasts show similar patterns over time. Indeed, both approaches based on t-distributed innovations show slightly more conservative risk forecasts compared to the ES-GARCH-norm model, which might be due to a superior ability to capture the fat tails of log return losses. This becomes perceptible as risk estimates of the ES-GARCH-norm model are in general on a lower level compared to the other two approaches. Furthermore, there are barely any differences in the forecasts of both the ES-GARCH-t and the ES-EGARCH-t model, which might be caused by the rather low skewness of the underlying S&P 500 data. Indeed, during the peak of the crisis period, the ES-GARCH-t model might even yield a slightly superior coverage of the underlying risk.

(V) **ES-FHS-RiskMetrics**

The fifth ES estimation model is labelled as *ES-FHS-RiskMetrics*. Compared to the previous four models, it yields a non-parametric approach and is thus especially popular amongst practitioners. It makes use of a *Filtered Historical Simulation (FHS)* based on volatility forecasts given by the RiskMetrics approach outlined in JP Morgan (1996). Moreover, in the RiskMetrics methodology volatility is estimated by the use of an exponentially weighted moving average (EWMA) scheme. For every observation $t$ within the backtesting horizon, the respective volatility is given by,

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)l_{t-1}^2, \tag{6.11}$$

where $\lambda$ is the decay factor, which calibrates the weights in the EWMA scheme. In line with the recommendation given by JP Morgan (1996), I set the decay factor to a value of $\lambda = 0.94$. For every observation $t$ within the backtesting horizon, again the previous 250 return loss observations are used for estimation purposes. In detail, one can calculated the volatility of all 250 preceding observations, i.e. $(\sigma_{t-250},\ldots,\sigma_{t-1})$, as well as a forecast for the volatility at time $t$, i.e. $\widehat{\sigma}_t$, based on equation (6.11). For every time $t$ within the backtesting horizon, the sample $S_{\text{FHS},t}$ functions as a basis for the FHS and is defined as,

$$\begin{aligned} S_{\text{FHS},t} &:= \left\{ l_s \frac{\widehat{\sigma}_t}{\sigma_s}, \text{ for } s = t-250,\ldots,t-1 \right\} \\ &:= \{ l_s^{\text{scaled}}, \text{ for } s = t-250,\ldots,t-1 \}. \end{aligned} \tag{6.12}$$

Thus, $S_{\text{FHS},t}$ contains all log return losses over the preceding estimation period, which are scaled according to changes in the conditional volatility over time. If volatility was lower at time $s < t$ compared to the value forecasted for time $t$, the respective log return loss is up-scaled in order to capture the increased riskiness inherited in the financial market. The same also holds true vice-versa if volatility decreases over time. In order to calculate both VaR and ES forecasts in the ES-FHS-RiskMetrics approach, empirical quantiles of $S_{\text{FHS},t}$ are exploited in the following way,

$$\widehat{VaR}_{t,\alpha} = \tilde{q}_\alpha(S_{\text{FHS},t}), \tag{6.13}$$

$$\widehat{ES}_{t,\alpha} = \frac{1}{\sum_{s=1}^{250} \mathbb{1}_{\{l_{t-s}^{\text{scaled}} > \tilde{q}_\alpha(S_{\text{FHS},t})\}}} \sum_{t=1}^{250} l_{t-s}^{\text{scaled}} \mathbb{1}_{\{l_{t-s}^{\text{scaled}} > \tilde{q}_\alpha(S_{\text{FHS},t})\}}, \tag{6.14}$$

where $\tilde{q}_\alpha(S_{\text{FHS},t})$ denotes the empirical $\alpha$-quantile of $S_{\text{FHS},t}$.[30] The resulting ES estimates of the ES-FHS-RiskMetrics approach for both considered periods of

[30] More details on the methodology of a FHS can for example be found in Hull (2015), chapter 13, or in McNeil et al. (2015), chapter 9.2.

S&P 500 log return losses are depicted in Figure 9 below. Compared to the previously stated GARCH methodologies, the ES-FHS-RiskMetrics approach yields slightly less volatile risk estimates. Nevertheless, the magnitude of obtained ES forecasts still appears to be sufficient at a first glance, to cover the underlying risk in both considered periods.

As outlined within this subsection, all five introduced forecasting models differ with respect to their complexity and methodology and thus also yield a different quality of obtained ES estimates for the underlying series of S&P 500 market data. Nevertheless, apart from the rather naive ES-norm model, all other estimation approaches might potentially also be implemented within financial institutions in a similar manner. Therefore, the judgement on the obtained ES forecasts displays a realistic example of use for the ES backtests in scope.

### 6.3. Backtesting results

This subsection is going to evaluate all estimated ES figures for both considered time periods of S&P 500 log return losses based on all outlined ES backtests. Furthermore, in line with the main objective of this thesis, a one-sided test design will be applied in order to detect estimation models which underestimate the true risk inherited in the S&P 500 index.

For the purpose of this subsection, only the de-facto one-sided version of the multinomial testing framework will be applied given the selected one-sided backtesting design. This can be justified, as both multinomial versions displayed almost identical size and power figures in the previous chapter, while the one-sided version additionally rules out the rejection of any conservative ES estimation model. Furthermore, the approximate Z-test proposed within this thesis will be applied instead of the original approach outlined by Costanzino and Curran (2015), as the approximative version exhibited slightly superior power and size values in the preceding simulation chapter. In addition, the approximative version can also be applied to any of the outlined ES estimation models given that it does not require explicit parametric estimates of the distribution of return losses, as the original version does. All other backtesting approaches, i.e. the Test 2 from Acerbi and Szekely (2014), the intercept ESR backtest motivated by Bayer and Dimitriadis (2019) and the combined ES residuals backtest based on McNeil and Frey (2000), are implemented as described in chapter 4 of this thesis.

For both backtesting periods of the S&P 500 index, all five estimated ES models are backtested at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Both significance levels determine the traffic light system prescribed by the Basel Committee for the VaR traffic light test, as outlined in section 3.2. In the same fashion, I also want to group each of the five ES estimation models into the same colour scheme with respect to a certain ES backtest. Thus, if an estimation model is rejected at both considered significance levels it is assigned a red traffic light. An ES estimation model is labelled in yellow if it is rejected at $\kappa = 0.05$, but passes the respective ES backtest at a significance level of $\kappa = 0.0001$. An ES estimation model which passes the respective ES backtest at both considered significance levels is labelled in green.

In the following, subsection 6.3.1 presents the backtesting results over the crisis period, while subsection 6.3.2 thereafter summarizes the respective results over the recent period.

#### 6.3.1. Crisis period

As outlined before, the crisis period contains the peak of the financial crisis in 2008, and thus exhibits an environment of financial distress which imposes a major challenge to any ES estimation model. An overview, of the performance of all five ES estimation models is depicted in Figure 10, which shows scatter plots of S&P 500 log return losses together with both estimated ES and VaR figures for each of the considered ES estimation models. Moreover, a log return loss $l_t$ is marked in orange if it exceeds the respective $\widehat{VaR}_{t,0.975}$ threshold and it is marked in red if it additionally also violates the respective $\widehat{ES}_{t,0.975}$ estimate.

For all five estimation models, the majority of violations of both the VaR and the ES can be observed between June 2008 and January 2009. Indeed, the S&P 500 suffered the largest losses in the crisis period within this timespan, driven by the default of the Lehman Brothers at the 15.09.2008. The date when Lehman Brothers declared bankruptcy is market with dotted black lines in Figure 10. At a first glance, the ES-norm estimation model shows the worst results in this timespan with respect to both the number and the magnitude of observed outliers. This is not surprising, as both VaR and ES forecasts in the ES-norm model only slowly react to changes in the market environment. Around this peak of the financial crisis, all three GARCH-typ models exhibit similar patterns, whereas the ES-EGARCH-t model apparently displays slightly more severe violations compared to the other two approaches. From a first impression, the non-parametric ES-FHS-RiskMetrics approach yields the lowest degree of risk underestimation during the crisis period, with few violations, which additionally also appear to be of rather small magnitude.

In order to achieve a better understanding of the quality of the obtained risk estimates, I want to analyse some further statistics before turning to the actual ES backtesting results. As outlined within chapter 3, two aspects of ES estimates need to be taken into account in order to verify the correct unconditional coverage of an ES estimation model. First of all, the number of violations beyond $\widehat{VaR}_{t,0.975}$ over the backtesting horizon and secondly their magnitude with respect to the estimated $\widehat{ES}_{t,0.975}$. As outlined in Proposition 3.5, the first aspect is equivalent to the condition $\mathbb{E}[I_t(\alpha)] = 1-\alpha$ for every observation $t$ within the backtesting horizon. Therefore, over the crisis period the expected number of VaR violation of an accurate estimation model is given by,

$$\begin{aligned}\mathbb{E}[\#\text{ of VaR violations}] &= \sum_{t=1}^{500} \mathbb{E}[I_t(\alpha)] \\ &= 500 \cdot 0.025 = 12.5. \quad (6.15)\end{aligned}$$

With respect to the second aspect listed above, the magnitude of outliers beyond the respective VaR threshold needs to be taken

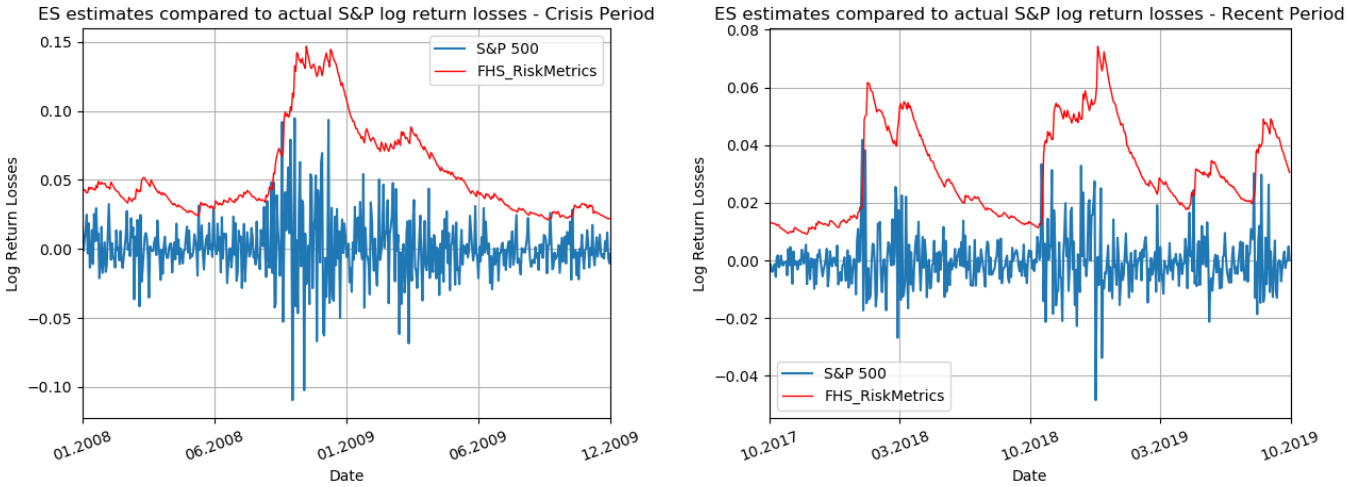


**Figure 9:** ES estimations - Filtered Historical Simulation with volatility updating based on RiskMetrics. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

**Crisis period - Violations of $\widehat{VaR}_{t,0.975}$ and $\widehat{ES}_{t,0.975}$**

| ES estimation model | # of VaR violations | # of ES violations | Average exceedance of ES in % |
|---|---|---|---|
| ES-norm | 28 | 21 | +24.44 |
| ES-GARCH-norm | 47 | 30 | +18.19 |
| ES-GARCH-t | 32 | 6 | +8.57 |
| ES-EGARCH-t | 34 | 14 | +13.52 |
| ES-FHS-RiskMetrics | 10 | 6 | +7.14 |

**Table 13:** Observed number of violations beyond $\widehat{VaR}_{t,0.975}$ and beyond $\widehat{ES}_{t,0.975}$, for all considered estimation models within the crisis period of the S&P 500. Furthermore, the average exceedance of VaR violations beyond the ES in percent is depicted.

into account. For an accurate ES estimation model, the average value of all VaR violations should be fairly close to the estimated ES figure. If on average VaR violations exceed the respective ES forecast, this hints towards a potential risk underestimation. For a first assessment of this second aspect, I define the *average exceedance of the ES* as,

$$\frac{1}{\sum_{t=1}^{500} I_t(\alpha)} \sum_{t=1}^{500} \frac{l_t - \widehat{ES}_{t,\alpha}}{\widehat{ES}_{t,\alpha}} I_t(\alpha) \tag{6.16}$$

In the following, both the observed number of VaR and ES violations as well as the average exceedance of the ES are listed in Table 13, for all ES estimation models within the crisis period. As depicted in Table 13, only the ES-FHS-RiskMetrics model exhibits less VaR violations than expected over the crisis period. All other applied estimation approaches fail to correctly model the conditional 0.975-quantile of log return losses during the period of high financial distress. Indeed, all four other estimation models display more than twice as many VaR outliers as expected. As previously noted, the magnitude of violations is the largest for the ES-norm model where 21 out of 28 VaR violations also lie beyond the estimated ES forecast, which leads to an average exceedance of the ES of 24.44 %. With respect to the magnitude of outliers, also all other ES estimation models underestimate the underlying risk during the crisis period. The least degree of risk underestimation in the distribution tail beyond the conditional 0.975-quantile can be observed for both the ES-FHS-RiskMetrics and the ES-GARCH-t model, with average exceedance values of the ES below 10 %. Based on this first assessment, I expect that the ES-FHS-RiskMetrics model achieves the best backtesting results in the crisis period. All other ES estimation approaches display weak results, especially in forecasting the conditional 0.975-quantile of log return losses given by $\widehat{VaR}_{t,0.975}$.

The actual backtesting results for all considered ES backtests over the crisis period of S&P 500 log return losses are listed in Table 14. Overall, the achieved backtesting results are in line with expectations given the preceding argumentation. Furthermore, there are no major differences regarding the traffic light classification of ES estimation models between the applied backtests. The ES-FHS-RiskMetrics approach passes all conducted backtests at both considered significance levels and is thus universally labelled in green over the crisis period. Nevertheless, achieved p-values for the ES-FHS-RiskMetrics model show a high variation between roughly 9% and 80%. Although

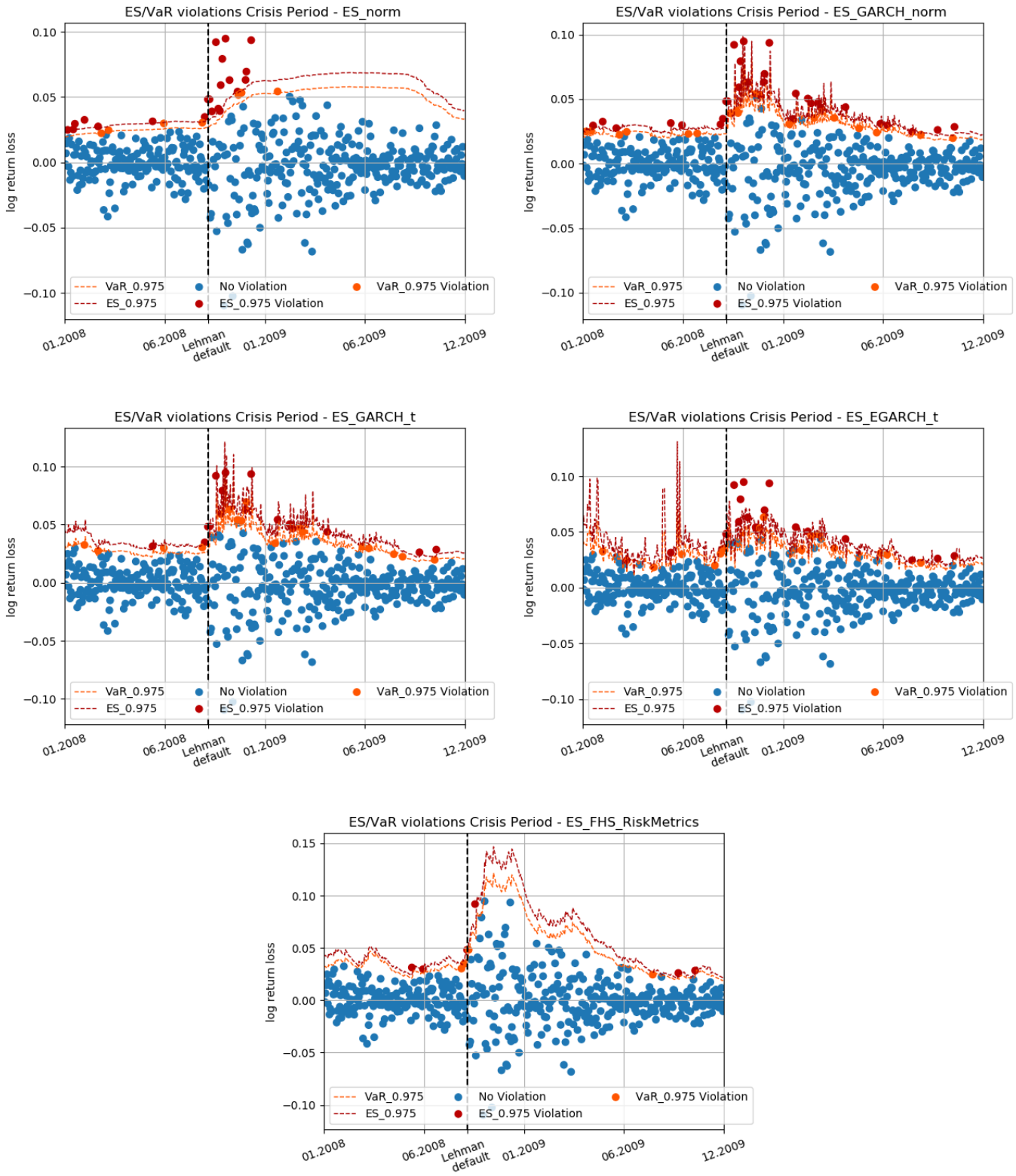


**Figure 10:** Scatter plots with observed log return losses from crisis period of the S&P 500. Violations beyond $\widehat{VaR}_{t,0.975}$ are marked in orange and violations beyond $\widehat{ES}_{t,0.975}$ in red for all five considered ES estimation models.

**Crisis Period - ES backtesting results**

| | | Estimation model | | | | |
|---|---|---|---|---|---|---|
| **ES backtest** | Significance Level | ES-norm | ES-GARCH-norm | ES-GARCH-t | ES-EGARCH-t | ES-FHS-RiskMetrics |
| Multinomial one-sided | 0.05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | 0.0964 |
| | traffic light | **red** | **red** | **red** | **red** | **green** |
| Test 2 | 0.05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | p-value* | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | 0.804 |
| | traffic light | **red** | **red** | **red** | **red** | **green** |
| Intercept ESR | 0.05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✓ | ✗ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | 0.0100 | $< 0.0001$ | 0.1300 |
| | traffic light | **red** | **red** | **yellow** | **red** | **green** |
| Z-test approximative version | 0.05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | 0.3786 |
| | traffic light | **red** | **red** | **red** | **red** | **green** |
| Combined ES residuals | 0.05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | 0.1160 |
| | traffic light | **red** | **red** | **red** | **red** | **green** |

**Table 14:** Backtesting results of all five implemented ES estimation models with respect to all ES backtests in scope over the crisis period of S&P log-return losses. Backtests are conducted at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Results are labelled with an ✗if the H0 in the one-sided test is rejected an with an ✓if the H0 can not be rejected.
*p-values for the Test 2 correspond to the bootstrap test version applied only for $\kappa = 0.0001$.

**Recent period - Violations beyond $\widehat{VaR}_{t,0.975}$ and $\widehat{ES}_{t,0.975}$**

| ES estimation model | # of violations of $\widehat{VaR}_{t,0.975}$ | # of violations of $\widehat{ES}_{t,0.975}$ | Average exceedance of ES in % |
|---|---|---|---|
| ES-norm | 28 | 19 | +39.57 |
| ES-GARCH-norm | 19 | 11 | +24.21 |
| ES-GARCH-t | 15 | 1 | -4.98 |
| ES-EGARCH-t | 19 | 6 | +11.26 |
| ES-FHS-RiskMetrics | 16 | 6 | +7.20 |

**Table 15:** Observed number of violations beyond $\widehat{VaR}_{t,0.975}$ and beyond $\widehat{ES}_{t,0.975}$ as well as average exceedance of VaR violations beyond the ES in percent, for all considered estimation models within the recent period.

this has no impact on the assigned traffic light, test decisions might diverge if further backtesting significance levels are taken into account.

On the contrary, all other implemented ES estimation models are rejected at all applied backtests at a significance level of $\kappa = 0.05$, and at almost all applied backtest at the more extreme significance level of $\kappa = 0.0001$. Indeed, only the intercept ESR backtest does not reject the ES-GARCH-t estimation model at the more extreme significance level. Therefore, the ES-GARCH-t estimation model achieves four red and one yellow traffic light, while the three remaining ES estimation models are assigned a red label throughout all backtests in the crisis period. This minor difference in assigned traffic lights might be justified with the results listed in Table 13, as the ES-GARCH-t estimation model apparently displays a lower degree of risk underestimation compared to the ES-norm, the ES-GARCH-norm and the ES-EGARCH-t approache. From a theoretical perspective, it is also reasonable that the ES-GARCH-t approach outperforms both estimation models based on normality assumptions, due to its superior ability to capture the highly leptokurtic patter of S&P 500 log return losses in the crisis period.

It should be noted that for the rather extreme value of $\kappa = 0.0001$ results of all applied bootstrap test decisions might potentially change depending on the respective simulation output. For the more extreme significance threshold, the Test 2[31], the combined ES residuals backtest as well as the intercept ESR backtest are based on a simulation procedure. Given the extreme significance threshold, the output of single simulations trials might decide over the achieved test decision unless a very large amount of simulation trials $M$ is considered. As results depicted in Table 14 are overall plausible, I stick with values of $M = 1000$ for both the Test 2 and the combined ES residuals backtest as well as $M = 100$ for the intercept ESR backtest following the argumentation given chapter 4. Furthermore, the selected number of simulation trials assures that computational times stay within an acceptable range for practical applications. Nevertheless, one should be aware of this issue, whenever a bootstrap decision is applied at rather extreme significance levels.

[31] As outlined within chapter 4.2 the Test 2 from Acerbi and Szekely (2014) is conducted with fixed critical values for a significance level of $\kappa = 0.05$, whereas a bootstrap decision is applied for a value of $\kappa = 0.0001$.

6.3.2. Recent period

This subsection presents the backtesting results over the most recent period of S&P 500 log return losses available at the start of the processing period for this thesis. As previously outlined the recent period depicts a rather calm market environment compared to the previously analysed crisis period. Thus, also the implemented ES estimation models are expected to show superior results over this second backtesting period.

A first impression of the accuracy of ES estimates can be gathered from Figure 11, which depicts the scatter plots of S&P 500 log return losses together with the respective VaR and ES estimates. Again all observed violations are marked in the same fashion as in Figure 10. Similar as in the crisis period, the ES-norm model is not able to capture the rather moderate peaks in log-return losses, due to its limited reactivity to changing market conditions. Again the patterns for all applied GARCH models look similar at a first glance. Nevertheless, apparently the ES-GARCH-t model yields the lowest amount of outliers amongst these three models. The non-parametric ES-FHS-RiskMetrics model again produces slightly less volatile ES estimates compared. Still, both the number and the magnitude of violations appears to be comparable to the implemented GARCH methodologies.

In order to get an even better idea of the performance of all implemented estimation models, Table 15 again lists some additional statistics regarding the number and the magnitude of observed violations. In the recent period, all estimation models exceed the expected number of VaR violations of 12.5 derived in formula (6.15). Nevertheless, four out of five estimation models range within a value of 15 to 19 observed VaR violations, which is still reasonably low. The ES-norm model on the contrary again realizes more than twice as many outliers as expected. Regarding the magnitude of VaR violations, both the ES-norm and the ES-GARCH-norm model stand out with an average exceedance of the ES of +39.57 % and +24.21 %. Both values indicate towards a rather high degree of risk underestimation. On the other hand, the ES-EGARCH-t and the ES-FHS-RiskMetrics approach only show a moderate degree of risk underestimation in the distribution tail with an average exceedance of the ES of +11.26 % and +7.20% respectively, whereas the ES-GARCH-t model is even slightly on the conservative side with a value of -4.98%.

Based on this first assessment, it is to be expected that the ES-

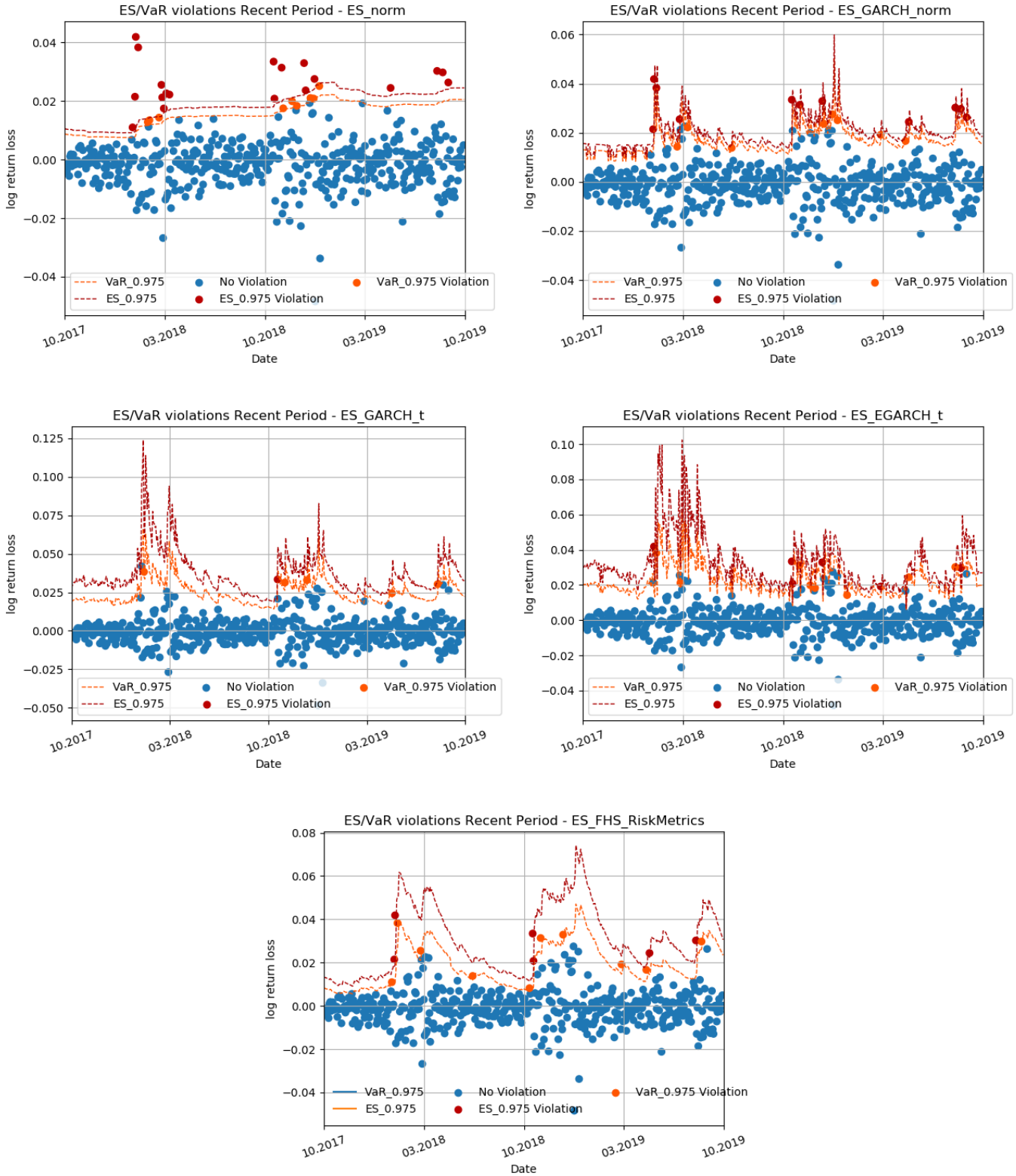


**Figure 11:** Scatter plots with observed log return losses from recent time period of the S&P 500. Violations beyond $\widehat{VaR}_{t,0.975}$ are marked in orange and violations beyond $\widehat{ES}_{t,0.975}$ in red for all five considered ES estimation models.

**Recent Period - ES backtesting results**

| | | Estimation model | | | | |
|---|---|---|---|---|---|---|
| **ES backtest** | Significance Level | ES-norm | ES-GARCH-norm | ES-GARCH-t | ES-EGARCH-t | ES-FHS-RiskMetrics |
| Multinomial one-sided | 0.05 | ✗ | ✗ | ✓ | ✓ | ✓ |
| | 0.0001 | ✗ | ✗ | ✓ | ✓ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | 0.6416 | 0.5853 | 0.3683 |
| | traffic light | **red** | **red** | **green** | **green** | **green** |
| Test 2 | 0.05 | ✗ | ✗ | ✓ | ✓ | ✓ |
| | 0.0001 | ✗ | ✗ | ✓ | ✓ | ✓ |
| | p-value* | $< 0.0001$ | $< 0.0001$ | 0.9840 | 0.1800 | 0.1970 |
| | traffic light | **red** | **red** | **green** | **green** | **green** |
| Intercept ESR | 0.05 | ✗ | ✗ | ✓ | ✗ | ✓ |
| | 0.0001 | ✗ | ✗ | ✓ | ✗ | ✓ |
| | p-value | $< 0.0001$ | $< 0.0001$ | 0.6700 | $< 0.0001$ | 0.0700 |
| | traffic light | **red** | **red** | **green** | **red** | **green** |
| Z-test approximative version | 0.05 | ✗ | ✗ | ✓ | ✓ | ✗ |
| | 0.0001 | ✗ | ✓ | ✓ | ✓ | ✓ |
| | p-value | $< 0.0001$ | 0.0003 | 0.9311 | 0.0775 | 0.0365 |
| | traffic light | **red** | **yellow** | **green** | **green** | **yellow** |
| Combined ES residuals | 0.05 | ✗ | ✗ | ✓ | ✓ | ✓ |
| | 0.0001 | ✗ | ✓ | ✓ | ✓ | ✓ |
| | p-value | $< 0.0001$ | 0.0110 | 0.8240 | 0.1910 | 0.1279 |
| | traffic light | **red** | **yellow** | **green** | **green** | **green** |

**Table 16:**
Backtesting results of all five implemented ES estimation models with respect to all ES backtests in scope over the recent period of S&P log-return losses. Backtests are conducted at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Results are labelled with an ✗if the H0 in the one-sided test is rejected an with an ✓if the H0 can not be rejected.
*p-values for the Test 2 correspond to the bootstrap test version applied only for $\kappa = 0.0001$.

GARCH-t model performs best in the conducted backtests over the recent period followed by both the ES-FHS-RiskMetrics and the ES-EGARCH-t model. It is interesting to observe, that the ES-GARCH-t model apparently yields slightly superior ES estimates compared to the more sophisticated ES-EGARCH-t approach. This might be explained as in the recent period of S&P 500 data, only a weak form of the leverage effect can be observed, as log return losses are only slightly right-skewed. Thus, the increased estimation uncertainty in the ES-EGARCH-t model, due to a higher number of estimation parameters, might outweigh the theoretical advantage to capture a potential leverage effect. The two remaining estimation approaches, i.e. the ES-norm and the ES-GARCH-norm model, again fail to consistently forecast the underlying risk in the S&P 500 index. Whereas the former one is not able to react to changing market conditions, the later one mainly fails due to its inability to account for the fat tails of the S&P 500 market data.

The results of the five implemented models with respect to the ES backtests in scope over the recent period of S&P 500 log return losses are listed in Table 16 below. Overall, the backtesting results over the recent period are again in line with expectations. Both the ES-norm as well as the ES-GARCH-norm model are rejected in the majority of all applied ES backtests. Whereas the former receives five red traffic lights, the latter ends up with three red and two yellow classifications. This is in line with expectations, as the ES-GARCH-norm model performs slightly better compared to the ES-norm approach, based on the results depicted in Table 15.

On the contrary, all three remaining estimation models are labelled in green for most of the considered backtesting methodologies. In line with previous argumentation, the ES-GARCH-t approach yields the best backtesting results over the recent period and achieves the optimum of five green traffic lights. Moreover, all p-values related to the ES-GARCH-t model are very high with values above 64 %. Furthermore, the decent backtesting results of both the ES-EGARCH-t and the ES-FHS-RiskMetrics are also reasonable, given the rather moderate degree of risk underestimation depicted in Table 15. The only slightly surprising result is that the intercept ESR backtests rejects the ES-EGARCH-t approach over the recent period not just at a significance level of $\kappa = 0.05$ but also at the more extreme threshold of $\kappa = 0.0001$. Nevertheless, the majority of all test decisions in the recent period coincides with the considerations based on the number and magnitude of violations.

Concluding this chapter, the outlined backtesting set-up displays a realistic example for a real-world application. The applied backtesting period of $T = 500$ observations is a reasonable choice, with respect to the length of the required input time series. Furthermore, ES backtest versions are applied which yield a one-sided test decision compliant with regulatory needs and which are applicable for all common ES estimation models. Overall, all five implemented ES backtests reveal reasonable results over both considered time periods of S&P 500 data. Moreover, the achieved backtesting results are in line with expectations taking into account previous considerations regarding the underlying market conditions. Besides the naive ES-norm estimation model, all other backtested forecasts stem from realistic estimation models, which are indeed relevant for practical applications. Furthermore, all applied backtests are able to differentiate between varying degrees of modelling accuracy induced by the different estimation models fitted to the S&P 500 data. Overall, it is also straightforward to implement a traffic light system, stipulated by the Basel Committee, for all evaluated ES backtests. Furthermore, in the majority of all cases the traffic light assigned to a certain estimation model coincides across all conducted backtests. Although it should be noted, that p-values might substantially differ depending on the applied backtest, which might be relevant if a more precise classification of estimation models needs to be conducted compared to the traffic light approach. Furthermore, any bootstrap decisions, especially at extreme significance levels like $\kappa = 0.0001$, up to a certain extend depends on a random factor based on the nature of the simulation procedure. Thus, at rather extreme significance levels bootstrap decisions are subject to potential changes depending on the simulation output. Apart from the last two mentioned aspects, there are indeed no major issues which might hinder the implementation of the outlined ES backtesting versions in real world scenarios.

## 7. Conclusion

This thesis revealed that backtesting the ES is indeed not much more complicated than backtesting the VaR, despite all doubts and reservations expressed within the literature. Definitely, backtesting the ES somehow requires a higher degree of creativity and there will probably never be any conceptual fully untainted backtest for the ES compared to available approaches to backtest the VaR. Nevertheless, there are several promising backtesting approaches, which have been proposed within the last two decades. Overall, five of them are presented in the course of this thesis, which originally stem from Kratz et al. (2018), Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015) and McNeil and Frey (2000). The main objective of this thesis was to outline and analyse one-sided, unconditional ES backtests, which take into account the practical aspects listed in Proposition 3.3. Therefore, some adjusted versions of the original backtesting approaches are proposed within chapter 4 of this thesis, which facilitate the application of the related methodologies in real-world scenarios. The conducted analysis is up to the best of my knowledge one of the most comprehensive evaluations of multiple ES backtests under one single framework.

In the empirical size and power simulations in chapter 5 of this thesis, the Z-test from Costanzino and Curran (2015) as well as the Test 2 from Acerbi and Szekely (2014) disclosed some potential miss-alignments in their underlying framework. Furthermore, the intercept ESR backtest motivated by Bayer and Dimitriadis (2019) might be the most challenging approach for any practical implementation due to a rather high degree of complexity and computational effort. Nevertheless, all ES backtests in scope proofed to be suitable in realistic backtesting scenarios, as outlined in chapter 6 of this thesis. With respect to the traffic light system required by the Basel Committee, all conducted ES backtests displayed reasonable results. Overall, especially

the de-facto one sided multinomial backtest based on Kratz et al. (2018) and the combined ES residual backtest motivated by McNeil and Frey (2000) need to be highlighted, due to their high practical relevance and the excellent results they exhibit throughout the course of this thesis.

Nevertheless, this thesis also revealed some open issues which are still subject to further research. As an example, substantial differences in the achieved p-values can be observed across the applied backtesting methodologies in certain scenarios within the preceding chapter. Whereas, this has no major impact on the assigned traffic lights, it might become more relevant if estimation models ought to be categorized on a finer grid. Furthermore, the choice of the respective backtest might also become relevant when rather extreme backtesting significance levels like $\kappa = 0.0001$ are evaluated, as bootstrap decisions might be subject to potential changes in the test outcome, due to the stochastic nature of the applied simulation decision. Thus, although all considered ES backtests exhibit overall reasonable results, it is still not clear which approach is the most appropriate one, as minor differences in the test outcomes might occur depending on the underlying setting. In line with other contributions, the power analysis within this thesis revealed that a backtesting horizon of $T = 250$ observations, which is currently stipulated by the Basel Committee, is not sufficient in order to consistently detect certain levels of risk underestimation in ES forecasting models. Within the preceding chapter, a backtesting horizon of $T = 500$ observations was applied, which appears to yield a good trade-off between backtesting power and data intensity. Nevertheless, some further considerations might be taken into account to determine an optimal backtesting time frame for the application of ES backtests. As one further aspect, there are still few contributions, like Du and Escanciano (2017), which focus on the development of a conditional ES backtest. Although this might not be of particular interest for practical applications, this is still a relevant aspect from a theoretical point of view.

Concluding this thesis, there is definitely further research that needs to be done in order to agree on an industry standard to backtest the ES. Nevertheless, I believe that difficulties in backtesting the ES should not any more be seen as a legit argument to favour the VaR as a primary risk measure for market risk, based on the results outlined within this thesis.

## References


Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, *26*(7), 1505 - 1518.

Acerbi, C., & Szekely, B. (2014). Back-testing expected shortfall. *Risk Magazine*.

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487 - 1503.

Angelidis, T., Benos, A., & Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Statistical Methodology*, *1*(1), 105 - 128. doi: https://doi.org/10.1016/j.stamet.2004.08.00

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203-228. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00068 doi: 10.1111/1467-9965.00068

Azzalini, A., & Valle, A. D. (1996, 12). The multivariate skew-normal distribution. *Biometrika*, *83*(4), 715-726. Retrieved from https://doi.org/10.1093/biomet/83.4.715 doi: 10.1093/biomet/83.4.715

Basel Committee. (n.d.-a). Fundamental review of the trading book: A revised market risk framework.

Basel Committee. (n.d.-b). Minimum capital requirements for market risk.

Basel Committee. (n.d.-c). Overview of the amendment to the capital accord to incorporate market risks.

Basel Committee. (n.d.-d). Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements.

Bayer, S., & Dimitriadis, T. (2019). *Regression based expected shortfall backtesting.* (Working paper)

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4), 465–474. Retrieved from http://www.jstor.org/stable/1392281

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307 - 327. Retrieved from http://www.sciencedirect.com/science/article/pii/0304407686900631 doi: https://doi.org/10.1016/0304-4076(86)90063-1

Carver, L. (2013). Mooted VaR substitute cannot be backtested, says top quant. *Risk*, *26*.

Chen, J. M. (2014). Measuring market risks under the Basel Accord: VaR, stressed VaR and expected shortfall. *Aestimatio: IEB international Journal of Finance*, *8*, 184-201.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841-62. Retrieved from https://EconPapers.repec.org/RePEc:ier:iecrev:v:39:y:1998:i:4:p:841-62

Cont, R., Deguest, R., & Scandolo, G. (2008, 01). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, *10*, 593-606. doi: 10.2139/ssrn.1086698

Costanzino, N., & Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation*, *9*, 21-31. doi: 10.21314/JRMV.2015.131

Costanzino, N., & Curran, M. (2018, 01). A simple traffic light approach to backtesting expected shortfall. *Risks*, *6*, 2. doi: 10.3390/risks6010002

Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., & Shin, H. (2001, 08). An academic response to Basel II. *Financial Markets Group Special Papers*(130).

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883. Retrieved from http://www.jstor.org/stable/2527342

Dimitriadis, T., & Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, *13*(1), 1823–1871.

Du, Z., & Escanciano, J. C. (2017). Backtesting expected shortfall: Accounting for tail risk. *Management Science*, *63*(4), 940-958. Retrieved from https://doi.org/10.1287/mnsc.2015.2342 doi: 10.1287/mnsc.2015.2342

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall.

Emmer, S., Kratz, M., & Tasche, D. (2015). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, *18*, 31-60. doi: 10.21314/JOR.2015.318

Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *Annals of Statistics*, *44*(4), 1680–1707.

Foellmer, H., & Schied, A. (2002, 09). Convex measures of risk and trading constraints. *Finance and Stochastics*, *6*, 429-447. doi: 10.1007/s007800200072

Gneiting, T. (n.d.). Making and evaluating point forecasts. , *106*, 746-762.

Hull, J. C. (2015). *Risk management and financial institutions* (Fourth Edition ed.). Wiley.

JP Morgan. (1996). *Riskmetrics, t.m.* (Tech. Rep.).

Kerkhof, J., & Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking & Finance*, *28*(8), 1845-1865.

Kratz, M., Lok, Y. H., & McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, *88*(C), 393-407. Retrieved from https://ideas.repec.org/a/eee/jbfina/v88y2018icp393-407.html doi: 10.1016/j.jbankfin.2018.0

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

McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, *7*(3-4), 271-300. Retrieved from https://EconPapers.repec.org/RePEc:eee:empfin:v:7:y:2000:i:3-4:p:271-300

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools* (Revised Edition ed.). Princeton University Press.

Moldenhauer, F., & Pitera, M. (2018, 08). Backtesting expected shortfall: a simple recipe? *SSRN Electronic Journal*.

Nass, C. A. G. (1959). The $x^2$ test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika*, *46*(3-4), 365-385. Retrieved from https://app.dimensions.ai/details/publication/pub.1059416861 doi: 10.1093/biomet/46.3-4.365

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370.

Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, *11*(4), 1833–1874. Retrieved from https://doi.org/10.1214/17-AOAS1041 doi: 10.1214/17-AOAS1041

Osband, K. (1985). *Providing incentives for better cost forecasting* (Unpublished doctoral dissertation). University of California.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157-175. doi: 10.1080/14786440009463897

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE transactions on neural networks*, *11*, 1263-71.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, *20*(3), 360 - 380.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, *32*(7), 1404 - 1415. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378426607003627 doi: https://doi.org/10.1016/j.jbankfin.2007.11.012