Backtesting Expected Shortfall

Carsten Frommhold

Carsten Frommhold
January 2018
carsten.frommhold@tu-dortmund.de

Abstract

With the Fundamental Review of the Trading Book, the Basel Committee on Banking Supervision proposed to replace the Value-at-Risk as the risk measure with the Expected Shortfall in order to calculate the regulatory capital. Since then, both risk measures have to be calculated but model validation and backtesting continues to be based on the Value-at-Risk. In the last years, an academic and practical debate ensued on whether and how the Expected Shortfall is backtestable. One central question is whether the property of elicitability, which is not fulfilled by the Expected Shortfall, is necessary to construct suitable backtests. Over the years, there have been a number of different backtest proposals. In this thesis, six selected backtests are examined in an exemplary manner and integrated into a theoretical framework. One backtest, namely a proposal of Acerbi and Szekely (2017), stands out. The reasons for this are, on the one hand, its ease of implementing. On the other hand, the small sensitivity to Value-at-Risk predictions, if only the Expected Shortfall should be backtested. Based on this test, a traffic light approach similar to the previous Value-at-Risk backtesting framework can be recommended.

Keywords: Value-at-Risk, Expected Shortfall, Elicitability, Backtesting, Model Validation, Regulatory Capital

Contents

List of Figures

List of Tables

List of Abbreviations

AR	Autoregressive
ARMA	Autoregressive Moving Average
AS_1	The first backtest of Acerbi and Szekely
AS_2	The second backtest of Acerbi and Szekely
AS_3	The third backtest of Acerbi and Szekely
AS_4	The fourth backtest of Acerbi and Szekely
BCBS	Basel Committee on Banking Supervision
cdf	Cumulative distribution function
EGARCH	Exponential General Autoregressive Conditional Heteroscedasticity
CP_1	The first backtest of Corbetta and Peri
CP_2	The first backtest of Corbetta and Peri
ES	Expected Shortfall
et al.	Et alia (Latin), and others
FRTB	Fundamental Review of the Trading Book
FX	Foreign Exchange
GARCH	General Autoregressive Conditional Heteroscedasticity
i.e.	Id est (Latin), that is to say
iid	Independent and identically distributed
MA	Moving Average
P&L	Profit and loss
SD	Standard deviation
TAR	Threshold Autoregressive
VaR	Value-at-Risk
vs.	versus

List of Symbols

α	Confidence level
χ^2_k	Chi-squared distribution with k degrees of freedom
γ	Shape Parameter for the skewness
$\frac{\delta F(x)}{\delta x}$	Derivation of F with respect to x
$\mathbb{E}(X)$	Expected value of X
$\mathbb{E}_F(X)$	Expected value of X if $X \sim F$
ε_t	Innovation at time t in an ARMA or GARCH process
ES_α	Expected Shortfall for a significance level α
F	Actual cdf
$\mathcal{F}$	Class of distributions
ϕ	Density function of $N(0,1)$
Φ	Cdf of $N(0,1)$
$I = (A < B)$	Indicator variable which is 1 if and only if A is smaller than B
$I(y,x)$	Identification function
$I_x(a,b)$	Regularized incomplete beta function
H_0	Null hypothesis
H_1	Alternative hypothesis
k	Kurtosis
λ	Coverage level
$N(\mu,\sigma)$	Normal distribution with mean μ and standard deviation σ
μ	Mean
$O(x)$	Landau-symbol
P	Predicted cdf
p	Significance level
s	Skewness
ρ	Risk measure
σ	Standard deviation
$S(y,x)$	Scoring function
$SN(\mu,\omega,\gamma)$	Skewed normal distribution with mean μ, standard deviation scale ω and skewness parameter γ
$ST(\mu,\omega,v,\gamma)$	Skewed t distribution with mean μ, standard deviation scale ω, v degrees of freedom and skewness parameter γ
T	Period length
$T(\mu,\omega,v)$	T distribution with mean μ, standard deviation scale ω and v degrees of freedom
t	Time
v	Degrees of freedom
VaR_α	Value-at-Risk for a significance level α
X	Random variable/ Profit and loss variable
X_t	Profit and loss variable at time t
(X_t)	Realized time series of a profit and loss variable, $t = 1,..,T$

$(x)^+$	Maximum of $\{x, 0\}$
$(x)^-$	Minimum of $\{x, 0\}$
y	Risk measure forecast
y_t	Risk measure forecast at time t
(y_t)	Vector of the risk measure forecasts for $t = 1, .., T$
$Z(y, x)$	Backtest function
ω	standard deviation scale
$A \in B$	A is an element of B
$A \subseteq B$	A is a subset of B
$A < B$	A is smaller than B
$A \leq B$	A is smaller or equal than B
$\lfloor x \rfloor$	Floor function
$\xrightarrow{D}$	Convergence in distribution
%	Percent
g^{-1}	Inverse function of g

1 Introduction

The question of how market risk can be measured is one of the central topics in financial mathematics. This issue is becoming particularly important with regard to financial institutions such as banks and insurance companies. They should operate a good risk management so that customers can be sure that future cash flows will continue to be secured. For many years, the equity capital ratio for banks has been determined on the basis of the so-called Value-at-Risk (**VaR**). This amount multiplied by a factor for a larger buffer must be covered at least by equity. A bank makes daily forecasts on the probability distribution of future profits and losses and has to deposit an equity capital which is large enough on the basis of these forecasts. The forecast must be validated retrospectively ("backtesting"). A very simple framework for the **VaR** backtesting has been established in 1996. This has been criticized again over the years, as losses below the predicted **VaR** are always counted equally, regardless of their magnitude. Between 2012 and 2016, there were a number of papers issued by the Basel Committee on Banking Supervision (BCBS) to review the minimum capital requirements. Part of these requirements is to replace the **VaR** with the Expected Shortfall (**ES**) as the risk measure. The **ES** describes the expected losses in that case the **VaR** is exceeded. This decision has opened a very interesting and controversial discussion in recent years. In particular, it deals with the difficulties in finding a suitable method for backtesting the **ES**.

The aim of this thesis is to provide an insight into the subject of risk measures and backtesting. The first step is to present desired characteristics of risk measures and compare the **VaR** with the **ES**. In addition, the decision of the BCBS is to be evaluated. The focus is on the question to what extend the **ES** can be backtested. There has long been disagreement in the academic debate. In addition to a formal classification, six different backtests are presented. Their advantages and especially disadvantages are exemplified in this thesis. Finally, the question of whether and how to construct a similarly simple backtesting framework as the previous one is discussed.

The structure of the thesis

Chapter 2 gives an overview of the mathematical background of risk measures. The **VaR** and the **ES** are presented and classified in the context of banking regulation. Afterwards, the reader is introduced to the concept of elicitability, which the **ES** does not fulfill. This has sparked a wide debate.

In chapter 3 the backtests of Acerbi and Szekely (2014) and Corbetta and Peri (2016) are presented and discussed. It also addresses the question of whether

or for what elicitability is actually necessary. Afterwards, the latest work of Acerbi and Szekely (2017) will be discussed, trying to create a theoretical framework on backtesting **ES**. For the first time, they give a concrete definition of what exactly backtestability means. Based on a concept of so-called ridge backtests, a new proposal is presented. The chapter ends with a brief overview of further backtest proposals.

Chapter 4 briefly discusses how returns can be modeled for the following analysis. Iid modeled returns are also discussed as the use of ARMA and GARCH models.

In chapter 5, the backtests are checked for significance and power in various settings in order to find out which of the suggestions are suitable for the practice and which reveal too much weaknesses. In particular, it deals with the question of what influences have an effect on the power. This also involves the influence of the **VaR** prediction on the power of the **ES** backtests.

The discussion and conclusion completes the thesis.

2 Background

2.1 Risk, risk measures and their properties

The term risk is initially not associated with a clear definition. Roughly speaking, risk means first of all that there is uncertainty about future outcomes and that these are not clearly predictable. One method is to capture the randomness using a distribution function. Any parameters of the distribution must then be estimated. However, this thesis does not aim at estimation theory. Rather, the focus is on so-called risk measures, which are measures that allocate exactly one real value to a random variable or to its cdf.[1] The questions and principles presented here can basically be applied to all forms of real random variables. However, due to regulatory requirements, they are currently used mostly in finance, especifically at banks. Therefore, in this thesis returns are always considered unless it is explicitly stated. Specifically, one-day returns are considered that is, the distribution of the gains and losses of an investment or a portfolio within one day. So the random variables considered here represent *profit and loss variables* (P&L variables). In this thesis, a risk measure has always a positive value.

In this context, there are a number of desirable properties that a risk measure should meet. They are grouped under the concept of *coherence*: A risk measure ρ is called *coherent* if the following axioms are fulfilled for all P&L variables X, X_1 and X_2 (see Emmer et. al., 2015):

Translation invariance

$$\rho(X + c) = \rho(X) - c \quad \forall \quad c \in R$$

Homogeneity

$$\rho(hX) = h\rho(X) \quad \forall \quad h \geq 0$$

Subadditivity

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

Monotonicity

$$X_2 \leq X_1 \quad \Rightarrow \quad \rho(X_1) \leq \rho(X_2)$$

Translation invariance means that a risk-free position prevents the risk measure by that amount. This is desirable because a security must only be deposited for risk positions of banks, as one will see later. Positive homogeneity means that the risk increases in a linear relation to the size of the risk position.

[1] Note that when merging different references it is not easy to ensure consistency in the notation. At this point it is initially easier for understanding to consider the risk measure of random variables directly, $\rho(X)$. Below, in order to avoid deviating too much from the notation in the sources used, a risk measure can also be related to a cdf F, $\rho(F)$.

If one invests the double amount in the same risky portfolio, the risk doubles as well. Subadditivity means that a combination of two investments should not pose a higher risk than the sum of the two risks for each individual investment. This makes sense in the context of portfolio theory. Diversification effects should reduce, but at least not increase the overall risk. Monotonicity means that if an investment leads to a higher payoff than an alternative in every scenario, the risk measure should be lower.

2.1.1 Value-at-Risk

Given a confidence level α, the **VaR** for a P&L variable X is the threshold, such that the probability of the P&L variable to be less than this threshold is α. If the cdf F, which represents the P&L distribution of X, the **VaR** is the threshold, such that the probability for a higher loss than the **VaR** is α, where $\alpha \in (0, 1)$:

$$VaR_\alpha(X) = -\inf\{x|F(x) \geqslant \alpha\}.$$

Sometimes one speaks of the quantile to the confidence level α, q_α. One point of criticism of the **VaR** is that it makes no statement about what happens when it is exceeded. Different stress scenarios with very different characteristics from mild to catastrophic may have the same **VaR**. Therefore, the **ES** is presented as another risk measure.

2.1.2 Expected Shortfall

During the search for a coherent risk measure which should be an improvement of the **VaR** and which should also describe the extent in case of an exceedance, the **ES** was created. The basic idea was published by Rappoport (1993) for the first time, Artzner et. al. (1999 and 2001) have evolved this thought. Acerbi and Tasche (2002) then finalized the definition for different cases. More precisely, they distinguish between continuous and non-continuous cdfs.

For general distributions and given a confidence level α, the **ES** for a P&L variable X is defined as follows:

$$\begin{aligned} ES_\alpha(X) &= -\frac{1}{\alpha}\mathbb{E}[(X(X+VaR_\alpha(X) < 0) - VaR_\alpha(X)(\alpha - (X+VaR_\alpha(X) < 0)] \\ &= VaR_\alpha(X) - \frac{1}{\alpha}\mathbb{E}[(X + VaR_\alpha(X)(X + VaR_\alpha(X) < 0)]. \end{aligned} \tag{1}$$

For continuous cdfs, the second part of the first equation is zero and one gets:

$$ES_\alpha(X) = -\frac{1}{\alpha}\mathbb{E}(X(X + VaR_\alpha(X) < 0)]. \tag{2}$$

The non-continuous case is not executed here in detail since it does not play a major role in the context of returns. Nevertheless, this definition will help

to design a certain backtest, as will be seen below. In case of a continuous cdf, the **ES** is equivalent to the mean of the tail conditional distribution:

$$\mathbb{E}(X|X < VaR_\alpha(X)). \tag{3}$$

It is also possible to express the **ES** as the mean of the **VaR** on the interval $(0, \alpha]$ (see Acerbi and Tasche (2002) for details):

$$ES_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha VaR_u(X)du.$$

Closed forms of ES

For some distributions, the **ES** can be derived in a closed form. This can be very helpful, especially for implementing. To give an example, the **ES** of a normal distributed P&L variable with mean μ and variance σ^2 has the following closed form:

$$ES(X) = \mu - \sigma\frac{\phi(\Phi^{-1}(\alpha))}{\alpha},$$

where ϕ denotes the density function for a standard normal distribution and Φ denotes the cdf for a standard normal distribution. For a detailed derivation, see Bernadi (2013). Below, further distributions such as the t distribution or the skewed normal distribution are discussed. Also, for these distributions exist closed forms for the **ES**. See Broda and Paolella (2011) and Bernadi (2013) for details.

2.2 VaR vs. ES

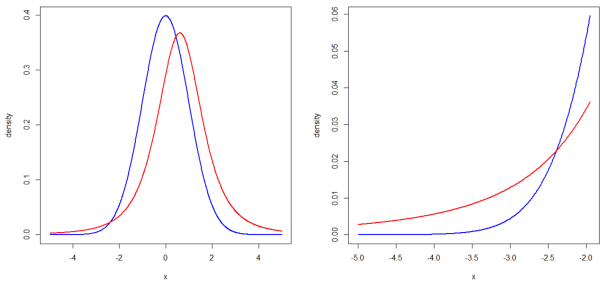

Figure 1: Two different densitiy functions with the same **VaR** but different **ES**.

As indicated above, the **VaR** makes no statement about the possible manifestations in the stress case, that is, if it has been exceeded. This becomes clear directly from an example, see figure 1. Both densitiy functions belong to the same **VaR**, but to different **ES**. The blue density function represents

	A	**B**	**A+B**
initial value	98.9	98.9	197.8
VaR 5	8.9	8.9	27.8
ES 5	20.9	20.9	27.8

Table 1: **VaR** and **ES** for each bond and a portfolio. Source: Acerbi et. al. (2001)

a standard normal distribution. The red one represents a t distribution with 5 degrees of freedom which is shiftet such that the the **VaR**s coinside. It is 1.96 for both but the **ES** is 2.33 for the normal distribution and 2.91 for the shifted t distribution.

VaR and **ES** fulfill translation invariance, homogeneity and monotonicity, but differ in subadditivity. The **ES** is always subadditive. There are a number of different proofs for this. For an overview see Embrechts and Wang (2015). In order to show that the **VaR** is generally not subadditive one can produce many examples. Here, a simple example taken from Acerbi et. al. (2001) is presented.

Suppose that two bonds, A and B are available as an investment opportunity. In case of success, the payout for each bond at the end of the term is 100. Each bond has a default risk with two different characteristics. With a probability of 3%, the recovery rate gives a payout of 70. With a probability of 2%, the payout is 90. However, by correlation, at most one of the two bonds can fail. Table 1 shows the corresponding values. The initial value is given by the expected value of a bond which is

$$0.95 * 100 + 0.03 * 70 + 0.02 * 90 = 98.9.$$

For the P&L distribution, the **VaR** at a confidence level $\alpha = 0.05$ is 8.9. The **ES** is

$$\frac{3}{5}(98.9 - 70) + \frac{2}{5}(98.9 - 90) = 20.9.$$

For a portfolio of both bonds, the **VaR** is

$$2 * 98.9 - (100 + 70) = 27.8,$$

since at most one bond can fail. The associated stress scenario describes the case in which a bond fails with a recovery value of only 70. Since there are no other scenarios in the tail, the **ES** is equal to the **VaR**.

In this example, the **VaR** violates subadditivity. The **VaR** of the portfolio is greater than the sum of the **VaR**s for each bond. Note that this could lead to fatal investment decisions. Regarding the **VaR**, it would be better to invest two bonds A or two bonds B than to invest in a portfolio with each bond. The diversification benefit is not reflected in this risk measure.

2.3 Classification in the regulatory system

The significance of the desired properties has already been suggested above in terms of risk measures with regard to financial institutions such as banks and insurance companies. Insurance companies' equity capital must be large enough to cover losses in the amount of the **VaR**. As a rule, insurance contracts are designed for many years, and the insurer must be able to meet these requirements even in stress scenarios. The project of the European Union, which has elaborated and reformed these capital requirements, is known as "Solvency II".

Banking, on the other hand, is moving much faster. There are daily incoming and outgoing payments and the assets are usually invested in shorter-term portfolios of shares, bonds and options. The BCBS sets guidelines for the resulting risks. Roughly speaking, a banks' equity capital must cover the risk of a portfolio, i.e. even in the event of high losses, the banks' customers must be able to claim their credit. Especially in this financial context, the question arises of how risk can be measured.

Previously, the **VaR** is used to measure the risk of a portfolio. More precisely, it is the **VaR** of a 10-day-return of the portfolio of a bank, which can be approximated by the 1-day-return scaled with $\sqrt{10}$. The amount of equity that a bank has to hold depends on the performance of the **VaR** forecast, which is checked by a backtest. This is described in more detail below.

In May 2012, the BCBS started the fundamental review of the trading book (FRTB). After publishing some discussion papers, the new capital adequacy standard was finalised (BCBS, 2016). It is planned to implement the new requirements by at least 2019. In the FRTB, the **VaR** with a confidence level $\alpha = 0.01$ is replaced by the **ES** with a confidence level of $\alpha = 0.025$. The BCBS argue as follows:

> *"Use of ES will help to ensure a more prudent capture of 'tail risk' and capital adequacy during periods of significant financial market stress."*

2.3.1 Backtesting VaR

Consider a model with confidence level α and significance level p, which leads to a **VaR** prediction ($VaR_{\alpha,t}$). A good model with regard to the **VaR** should meet the following criteria:

- Within one period of length T, $T\alpha$ exceedances are expected. The actual number of exceedances should not be significantly larger.

- The exceedances should be independent of each other, that is, not directly following each other.

Let $I_{\alpha,t} = (X_t < VaR_{\alpha,t})$ be the indicator variable, which is 1 if and only if one observes an exceedance with a loss higher than the **VaR**. X_t is the value of the P&L variable at time t.

2.3.2 The Binomial test

Under the null hypothesis, the number of returns that fall below the VaR are binomially distributed with parameters as above: This leads directly to the simplest type of backtest, the binomial test. The number of exceedances (denoted as N_T) is binomially distributed with probability α: $N_T = \sum_{t=1}^{T} I_t \sim \text{Bin}(T, \alpha)$. At a confidence level of $\alpha = 0.025$ and a significance level of $p = 0.05$, the prediction with respect to the **VaR** should be rejected at the latest when the probability for the observed number of exceedances under the null hypothesis is below the significance level:

$$P(\#\text{Exceedances} \geq N_T) < p.$$

2.3.3 The Kupiec test

Under the null hypothesis, the expected value of the *proportion of failure* $\frac{N_T}{T}$ is $1 - \alpha$. Kupiec (1995) proposed to use the likelihood ratio to test this assumption. He shows that the test statistic L is Chi-squared distributed with 1 degree of freedom. Say that p is the expected proportion of failure, then:

$$L = -2ln(\frac{(1-p)^{T-N_T} p^{N_T}}{(1-\frac{N_T}{T})^{T-N_T}(\frac{N_T}{T})^{N_T}}) \quad \sim \quad \chi^2_1.$$

The null hypothesis is rejected if L falls into the critical regions of the Chi-squared distribution. To test the assumption of independent exceedances, Christoffersen (1998) proposed an extension of the Kupiec test. More specifically, it is tested whether the likelihood for an exceedance is independent of the outcome of the previous day. The technical details will not be discussed here.

2.3.4 The traffic light approach

The simple binomial test is too weak to use it readily for the deposit of equity. If the null hypothesis is rejected only below a significance level p, bad predictions with weak p-values $<< 1$ would still be accepted. In addition, the

Zone	Number of exceedances	Scaling factor	Cumultative probability
Green	0	3	8.11%
	1	3	28.58%
	2	3	54.32%
	3	3	75.81%
	4	3	89.22%
Yellow	5	3.4	95.88%
	6	3.5	98.63%
	7	3.65	99.60%
	8	3.75	99.89%
	9	3.85	99.97%
Red	10	4	99.99+%

Table 2: Depending on the number of **VaR** exceedances, the prediction is divided into three different color zones (green, yellow and red). The numbers of exceedances for which the cumulative probability is below 95% fall into the green zone. For values between 95% and 99.9% the yellow zone is provided. Everything above falls into the red zone.

question remains open as to exactly what happens once the test is rejected. The BCBS (1996) has set a very simple framework based on the binomial distribution. The confidence level α is prescribed to be 0.01, every quarter a backtest has to be performed for the last 250 days. Based on the numbers of exceedances, the prediction of the bank is classified into one of three zones, namely green, yellow and red.

The system is shown in table 2. A bank already has to have equity in three times of the **VaR** anyway. That compensates for the fact that the burden of proof is not primarily on the banks' side here. For example, four exceedances are initially accepted in a year, even though that number speaks against the model. The cumulative probability of four exceedances in one year is 89.22%. For five and more exceedances, the bank is punished by an additional capital charge due to a higher factor with which the capital charge is calculated.

However, this system has some weaknesses since the actual deviation between the prediction and the true value does not necessarily affect the number of exceedances and thus also the zone. This is because the **VaR** backtest lacks a certain property called *sharpness*, which is presented later.

The traffic light could turn green, although the risk was underestimated. Vice versa, one could create examples where the true and the predicted **VaR** are very close to each other but the bank is punished by a high multiplicator. This would be the case, for example, if there are many **VaR** exceedances but they are only above the prediction. In this backtest, it does not matter whether the exceedances end up in the neighborhood of the predicted **VaR** or far away. It only counts the number.

2.4 Elicitability and the debate on the ES

At the latest following the decision of the BCBS to replace the **VaR** with the **ES**, a broad academic and practical debate has broken out. It is not completed until today. On the one hand there is the question of whether the **ES** is backtestable at all. On the other hand, the question arises whether one can find a similarly simple framework for backtesting the **ES** as it is described above for the **VaR**.

In order to understand this debate, one first needs the following definition. Following Gneiting (2011), a risk measure ρ is called *$\mathcal{F}$-elicitable*, if there exists a *scoring function* $S_\rho(y, x)$ such that the risk measure can be expressed as the minimizer of the expectation:

$$\rho(F) = \underset{y}{\operatorname{argmin}} \mathbb{E}_F[S_\rho(y, X)] \quad \forall F \in \mathcal{F},$$

where F is a cdf and $\mathcal{F}$ is a class of cdfs. A scoring function maps (y, x) to $[0, \infty)$ and is ≥ 0 with equality if $x = y$, continuous in x and the partial $\frac{\delta S_\rho(y,x)}{\delta x}$ exists and is continuous in x whenever $x \neq y$. y is the prediction of the risk measure, x stands for the outcome of the P&L variable.

To give an example, the scoring function for the mean is given by

$$S_\mu(y, x) = (y - x)^2.$$

The scoring function for the median (which is the quantile for $\alpha = \frac{1}{2}$, denotes as $q_{1/2}$) is given by

$$S_{q_{1/2}}(y, x) = |x - y|.$$

For the **VaR** to the significance level α, there exists a scoring function as well:

$$S_{q_\alpha}(y, x) = ((y \geq x) - \alpha)(y - x).$$

For a proof, see Wimmerstedt (2015).

Given a scoring function, it is directly possible to create a backtest. Let the risk measure forecasts of a model be $(y_1, .., y_T)$. Given a real time series $(X_1, ..., X_T)$, the test statistic is given by:

$$Z = \frac{1}{T} \sum_{t=t}^{T} S_\rho(y_t, X_t).$$

Given a predicted distribution P_t of the random variables X_t, one can derive the distribution of the test statistic. This can be done analytically (if possible) or via Monte Carlo simulations. Given a significance level, one can determine critial regions in which the predicted distribution is rejected. Below one will find some examples of this. However, the explanatory power of this depends

on the concrete scoring function. It is not always certain that this will enable reliable statements to be made. This is further discussed in chapter 3.

Model selection

The scoring function also makes it possible, to compare different models. Given two models 1 and 2 with risk measures forecasts $(y^1) = (y_1^1, .., y_T^1)$ and $(y^2) = (y_1^2, .., y_T^2)$ and the realized time series $(X_t) = (X_1, ..., X_T)$, one can calculate the test statistics $Z(y^1, X)$ and $Z(y^2, X)$. The one with the lowest value for Z can be seen as the "best". This is interesting, for example, when comparing a standard approach of a regulator to an internal model of a bank.

The ES is not elicitable - Has an impasse been reached?

In 2011, Gneiting proved that the **ES** is not elicitable. In 2013, Carver concluded, that the **ES** is not backtestable. But does the result of Gneiting really mean that it is not possible to backtest a modeled **ES**? One gets a first idea, if one looks at the **VaR** backtest. The backtesting method with the simple binomial test is based not on the scoring function but on the fact that the number of exceedances under the null hypothesis is binomially distributed. The property of elicitability is not used here at all.

In the following years, however, there were a number of proposals how the **ES** could be backtested. Some of them will be presented below. This is done chronologically. The proposals of Acerbi and Szekely (2014) and Corbetta and Peri (2016) are presented in detail. After that, based on the work of Acerbi and Szekely (2017), a concrete definition of "backtestability" is given. Based on the concept of so-called *ridge backtests*, another proposal is presented. After all there is a short overview of other suggestions.

Note that at the current time, backtesting for banks is still based on the **VaR** prediction, given a confidence level $\alpha = 0.01$. The traffic light approach is used to determine the multipler on the **ES** prediction (BCBS, 2016).

3 ES Backtests

3.1 Acerbi and Szekely (2014) - Three proposals

Acerbi and Szekely participated in the debate surrounding the backtesting of the **ES** in 2014, presenting three different backtests.[2] They apply their backtests directly to an analytical example of scaled t distributed returns and determine the power of these methods to discover wrong **ES** predictions.

Their core argument is that elicitability cannot make any statement as to whether a risk measure is not backtestable. As mentioned above, the scoring function makes it possible to compare different models for predicting a risk. However, this answers only the question of whether a model should be preferred to another. A concrete model cannot yet be validated by the scoring function, as no further statements can be derived from the value initially. As mentioned above, backtesting **VaR** is based on the binomial distribution and not on the scoring function. The fact that the **VaR** is elicitable has no influence on the test. In the following the three backtests are presented. Their significance and power are analyzed in a separate setting later.

3.1.1 Test 1

For the first test (AS_1), it is assumed that the **VaR** is known and also been backtested, for example with the methods mentioned above. In addition, the cdf is considered to be continuous. To derive the test statistic, recall equation (3). It follows that

$$\mathbb{E}(\frac{X}{ES_\alpha} + 1 | (X + VaR_\alpha) < 0) = 0.$$

This leads to the following test statistic:

$$Z_1 = \frac{\frac{\sum_{t=1}^{T} X_t I_t}{ES_{\alpha,t}}}{N_T} + 1,$$

where I_t is the indicator variable which is one if and only if a loss is higher than the predicted **VaR**: $I_t = (X_t < -VaR_{\alpha,t})$. N_T is the number of **VaR** exceedances. If there is confidence about the **VaR**, the expected value of the the observed **VaR** exceedances is equal to the respective **ES**. So, to calculate the test statistic, one needs the time series (X_t), the predicted **VaR** (VaR_t) and the predicted **ES** (ES_t) under the null hypothesis H_0 which is given by:

$$H_0 : \quad P_t^\alpha = F_t^\alpha \quad \forall t,$$

[2] As long as a concrete definition of what backtestable means is not given, one should actually be careful with the expression "backtest". However, in order not to make things too complicated at this point, the proposals presented in this thesis are always called "backtest".

where P_t is the assumed cdf and F_t the actual cdf for every time t and $P_t^\alpha(x) = \min(1, \frac{P_t(x)}{\alpha})$. That means that the distribution in the tail is predicted correctely. The alternative hypothesis is given by:

$$H_1 : ES_{\alpha,t}^F \geq ES_{\alpha,t} \quad \forall t \quad \text{and} > \text{for some } t;$$
$$VaR_{\alpha,t}^F = VaR_{\alpha,t} \quad \forall t.$$

Using the characteristics of conditional expectations , one can show that the expected value of Z_1 under H_0 is zero, and lower than zero under H_1:

$$\mathbb{E}_P(Z_1) = 0;$$
$$\mathbb{E}_F(Z_1) < 0.$$

3.1.2 Test 2

In the setting of the second test (AS_2), the **VaR** is assumed to be predicted as well, but not backtested. Again, the test statistic consists of the time series X_t, the predicted **VaR** VaR_t, the predicted **ES** ES_t under H_0 but also the confidence level α. The derivation is as follows:

Recall from equation (2), that for continuous cdfs, one can write the **ES** as follows:

$$ES_\alpha = -\mathbb{E}(\frac{X_t I_t}{\alpha}).$$

A change of this expression leads to the second test statistic:

$$Z_2 = \sum_{t=1}^{T} \frac{X_t I_t}{T\alpha ES_{\alpha,t}} + 1.$$

The intuition is very similar as above. The "empirical" **ES** is normalized. For non-continuous cdfs, this backtest can be extended for a different expression of I_t. See Acerbi and Tasche (2001) for details. Note that the following equation holds:

$$Z_2 = 1 - (1 - Z_1)\frac{\sum_{t=1}^{T} I_t}{T\alpha}.$$

This backtest is not only sensitive to the magnitude of **VaR**-exceedances but also on the frequency. Again, under the null hypothesis, the tail distribution is predicted correctly:

$$H_0 : \quad P_t^\alpha = F_t^\alpha \quad \forall t,$$

where $P_t^\alpha(x) = \min(1, \frac{P_t(x)}{\alpha})$.

The difference to AS_1 is that the alternative hypothesis also includes an un-

derestimation of the **VaR**:

$$H_1 : ES^F_{\alpha,t} \geq ES_{\alpha,t} \quad \forall t \quad \text{and} > \text{for some } t;$$
$$VaR^F_{\alpha,t} \geq VaR_{\alpha,t} \quad \forall t.$$

Again, using the characteristics of conditional expectations, one can show that $\mathbb{E}_P(Z) = 0$ and $\mathbb{E}_F(Z) < 0$.

3.1.3 Test 3

For the third test (AS_3) not only the pair $(VaR, ES)_t$ and the confidence level α, but the entire distribution is built in. The idea is, that $U_t = P_t(X_t)$ are identical, independent and uniform distributed on the interval $[0, 1]$, if the model is correct. If the risk would be underestimated, for example if the actual cdf would be broader, the values of U_t would not be uniform distributed but there would be a lot of smaller values. Acerbi and Szekely convert this into the following test statistic:

$$Z_3 = -\frac{1}{T}\sum_{t=1}^{T} \frac{ES^{(T)}_\alpha(P_t^{-1}(U))}{\mathbb{E}_W[ES^{(T)}_\alpha(P_t^{-1}(V))]} + 1,$$

where

$$ES^{(N)}_\alpha((Y_t)) = -\frac{1}{\lfloor N\alpha \rfloor}\sum_{i=1}^{\lfloor N\alpha \rfloor} Y_{1:N}$$

is the average of the $\lfloor N\alpha \rfloor$ lowest variables since $Y_{1:N}$ denotes the vector (Y_t) sorted by increasing value, $U = (U_1, .., U_T)$, W denotes the cdf of a uniform distribution on the interval $[0, 1]$ and V is a vector of T iid returns of which are W-distributed. $ES^{(N)}_\alpha$ can again be interpreted as empirical **ES**, whereby there is no statement about the **VaR**. The denominator that should normalize the counter can be calculated analytically:

$$\mathbb{E}_W[ES^{(T)}_\alpha(P_t^{-1}(V))] = -\frac{T}{\lfloor T\alpha \rfloor}\int_0^1 I_{1-p}(T - \lfloor T\alpha \rfloor, \lfloor T\alpha \rfloor)P_t^{-1}(p)dp,$$

where $I_x(a, b)$ is a *regularized incomplete beta function*. This function and its derivation is not discussed here in detail.

The hypotheses are much stronger than above and include the entire distribution. The null hypothesis is given by:

$$P_t = F_t \ \forall t.$$

The expected value of the test statistic given H_0 is zero and lower than zero

given H_1:

$$\mathbb{E}_P(Z_1) = 0;$$
$$\mathbb{E}_F(Z_1) < 0.$$

The alternative hypothesis is given by:

$$P_t \geq F_t \ \forall t \text{ and } > \text{for some } t.$$

The meaning of the null hypotheses

The informative value of the backtests is implicit in the null and alternative hypotheses. The alternative hypothesis of AS_1 includes a true **VaR** prediction. For a correct **VaR** but underestimated **ES**, the expected value of the test statistic is negative. This is a pretty strong limitation that cannot be used in many examples.

For AS_3, a statement about the expected value of the test statistics can only be made assuming the stochastic dominance of the entire distribution P against F. The backtest does not reject the null hypothesis until the strong counterhypothesis can be assumed. However, this includes a much greater statement than just about the **ES**. This strong wall for the rejection of the null hypothesis leads to the fact that this backtest can only be used conditionally: If this backtest does not decline, the **ES** nevertheless might be significantly underestimated.

Joint elicitability of ES and VaR

Although the **ES** is not elicitable, the pair (VaR, ES) ist jointly elicitable. Acerbi and Szekely propose a family of scoring functions. Fissler et. al. (2015) expand the framework. This makes a comparison of different predictions possible again, as mentioned above. However, both risk measures and not only the **ES** crop up. The joint elicitability and its' effects is discussed further below.

3.2 Corbetta and Peri (2016) - Counting the exceedances

Corbetta and Peri (2016) make a proposal for backtesting risk measures that docks on the simple **VaR** binomial test. They criticize that for many backtests (as well as those of Acerbi and Szekely) complex simulations are necessary (see section "implementing"). They make two suggestions for backtests based

on the number of exceedances above a desired risk measure. They use the poisson binomial distribution respectively an asymptotic convergence result, to determine the rejection areas.

3.2.1 Test 1

The construction of the first backtest of Corbetta and Peri (CP_1) is as follows: For every *law invariant* risk measure, one can define the *coverage level*. This describes the probability with which the P&L variable assumes a value less than that of the risk measure:

$$\lambda_P = P(-\rho(P)),$$

where law invariant means that a risk measure assigns the same value to P&L variables with the same distribution.

If (X_t) and the associated risk measure forecast (y_t) are predicted one day in advance, the indicator variables $I_t = (X_t < -y_t)$ are not identically distributed in general. A key assumption is that the exceedances are independent. This has to be tested in a complete framework as well. Let $\lambda_t^0 = P_t(-y_t)$, then I_t is Bernoulli distributed with probability λ_t^0 under H_0. The number of exceedances is Poisson binomial distributed:

$$Z = \sum_{t=1}^{T} I_t \sim \text{Poiss.Bin}((\lambda_t^0)_t).$$

If the number of exceedances lies in the critical areas of the Poisson binomial distribution, the assumption about the distribution must be rejected. Note that if one takes the **VaR** as the risk measure, this backtest ist equivalent to the simple binomial test as described above. One possibility would be to introduce a traffic light approach analogous to the previous **VaR** backtest, which determines a factor for the deposit of the regulatory capital based on the number of exceedances.

3.2.2 Test 2

Corbetta and Peri derived the following asymptotic result for a further test statistic (CP_2):

$$Z = \frac{\sum_{t=1}^{T}(I_t - \lambda_t^0)}{\sqrt{\sum_{t=1}^{T}\lambda_t^0(1-\lambda_t^0)}} \xrightarrow{D} N(0,1).$$

If this test statistic is in the critical ranges of the standard normal distribution, then the assumption about the risk measure must be rejected.

CP_1 rejects the null hypothesis if and only if significantly more exceedances are observed below the risk measure. Regarding the CP_2, in contrast to the backtests of Acerbi and Szekely one cannot easily make a general statement about the direction in which the expected value of the test statistic moves in case of underestimation.

Model Selection

Since the **ES** is not elicitable, one cannot use a scoring function to compare different models. Inspired by the result of Fissler et. al. (2015), Corbetta and Peri propose a universal method for measuring the magnitude of the exceedances. Let $(P, y)_t$ be the predicted distributions and risk measures with the associated coverage levels $(\lambda^0)_t$, then the magnitude of the exceptions is defined as follows:

$$M((P,y)_t) = \sum_{t=1}^{T} (X_t + y_t)^+ g(\lambda_t^0) + (-y_t - X_t)^+ g(1 - \lambda_t^0),$$

where $g : (0,1) \to (0,1)$ is monotone. For the purposes of risk management, g should be increasing such that underestimation $(-y_t - X_t)^+$ is penalized higher than overestimation $(X_t + y_t)^+$. Note that if the **VaR** is used as a risk measure and $g(u) = u$, M is the sum of the scoring functions for the **VaR**.

An advantage of the approach of Corbetta and Peri is that different risk measures can be used at different days. In this thesis it is limited to a constant **ES** prediction.

3.3 Implementing

AS_1 and AS_2 are very easy to implement, as the test statistic consists only of the forecast (VaR_t, ES_t) and the realized returns (X_t). One does not need a distribution P_t (at least not for the test statistic itself). For the calculation of the test statistic of AS_3, however, the distribution P_t of each daily forecast is needed, so it must remain stored until the time of the backtest. This can certainly be a decisive factor in more complex models. The same applies to the backtests of Corbetta and Peri. The advantage of these backtests is the knowledge of the asymptotic distribution of the test statistics. Using the Poisson binomial distribution for CP_1 or the standard normal distribution for CP_2, a critical region can be determined directly.

For the backtests of Acerbi and Szekely different assumptions under H_0 also show different asymptotic distributions of the test statistics. In order to be able to evaluate an actual observation in retrospect, the distribution of the

backtest must be approximated using a Monte Carlo simulation. This happens according to the following pattern:

- Simulation of M time series X_t under H_0
- Calculation of the test statistic Z_i depending on the forecast (VaR_t, ES_t) for AS_1 and AS_2 respectively P_t for AS_3 and the respective time series X_t.

For a specific observation (X_t) with associated test statistic Z, the p-value of this time series can now be read directly from the Monte Carlo distribution of the test statistic under H_0:

$$P(ES_t^F \leq ES_t^P) = \frac{\sum_{i=1}^{M}(Z_i < Z)}{M}.$$

If this value lies below the desired significance level p, the respective null hypothesis must be rejected. It is also possible to compute the critical regions directly if one takes the $\lfloor M * p \rfloor$ smallest value for a significance level p. Acerbi and Szekely argue that this value of AS_2 is almost constant for different distributions. A simulation could thus be avoided. This will be discussed further in the discussion.

Which of these advantages and disadvantages outweighs in practice depends on the actual application. While avoiding the Monte Carlo simulation for Corbetta and Peri, saving P_t can potentially consume a lot of disk space. AS_3 is the most impractical one in terms of implementation. One needs both the whole forecast P_t and a Monte Carlo simulation. For simple standard distributions, the effort is still kept within limits. For more complex models such as the GARCH model, which will be explained later, the calculation of the denominator in the test statistic can be very time-consuming for each point in time.

3.4 Acerbi and Szekely (2017) - A formal framework and a fourth backtest

Although the academic debate has been going on for a few years, there has long been no concrete definition of what "backtestable" actually means mathematically. Acerbi and Szekely (2017) closed this gap and proposed a formal framework for backtesting.

In chapter 2 the concept of elicitability was introduced. Recall that a risk measure ρ is called *$\mathcal{F}$-elicitable*, if there exists a *scoring function* $S_\rho(y, x)$ such that it can be expressed as the minimizer of the expectation:

$$y(F) = \underset{y}{\operatorname{argmin}} \mathbb{E}_F[S_\rho(y, X)] \quad \forall F \in \mathcal{F}.$$

Recall that the **ES** is not elicitable (Gneiting, 2011) as well as the variance (Lambert et. al., 2008). A scoring function is not unique but is for example invariant to transformations.

The fact that the scoring functions can be used to compare different models in retrospect has already been mentioned above. However, the elicitability does not serve to validate a model since the significance of the scoring function is limited. An example of this is the scoring function for the mean. It is always positive, so it makes no distinction between underestimation and overestimation of the true value. However, in the context of financial supervision, this is precisely what is desired since underestimation of risks should be avoided.

3.4.1 A formal definition of backtestability

First, Acerbi and Szekely propose the following definition: A risk measure ρ is called $\mathcal{F}$-*identifiable*, if there exists an *identification function* $I_\rho(y, x)$, such that

$$\mathbb{E}_F[I_\rho(y, X)] = 0 \quad \Leftrightarrow \quad y \in \rho(F) \quad \forall F \in \mathcal{F}.$$

Based on this concept, they provide a formal definition of backtestability: A risk measure ρ is called $\mathcal{F}$-*backtestable*, if there exists a *backtest function* $Z_\rho(y, x)$ such that:

$$\mathbb{E}_F[Z_\rho(y, X)] = 0 \quad \Leftrightarrow \quad y = \rho(F) \quad \forall F \in \mathcal{F},$$

which is strictly increasing in the prediction y, $\forall F \in \mathcal{F}$:

$$\mathbb{E}_F[Z_\rho(y_1, X)] < \mathbb{E}_F[Z_\rho(y_2, X)] \quad \text{if } y_1 < y_2.$$

This definition of the backtestability is very intuitive and eliminates the potential weakness of the scoring function that over- and underestimation cannot be distinguished. The expected value of the backtest function for underestimating the risk is negative, and the expected value for overestimating is positive. Similar to the model selection via scoring functions, one can now also use the backtest functions. In this way, empirical evidence can be used to draw the right conclusions. Negative values indicate underestimation, positive values indicate overestimation. In particular, underestimation is much more important in practice. Note that backtestability is an extension of identifiability in the sense that the expected value of the identification function hat to be strictly increasing in the prediction y. This can restrict the class of cdfs $\mathcal{F}$. Given a backtest function, one can construct the following test statistic:

$$Z = \frac{1}{T} \sum_{i=1}^{T} Z_\rho(y_t, X_t).$$

The distribution of Z can then be determined via Monte Carlo simulations as described above. The authors describe various identification and backtest functions for the mean, the median, the **VaR** and expectiles.[3] There are limitations to the class of distribution functions for quantiles, i.e. the median or the **VaR**. The distribution functions must be continuous in this quantile as well as strictly increasing. However, this is usually the case for returns and is therefore not discussed here in detail.

3.4.2 The connection between elicitability and backtestability

Acerbi and Szekely prove that if ρ is $\mathcal{F}$-backtestable with a backtest function $Z_\rho(y, x)$, then it is also $\mathcal{F}'$ elicitable with a y convex function

$$S_\rho(y, x) = \int^y Z_\rho(t, x)dt,$$

where $\mathcal{F}'$ is the class of distributions $\mathcal{F}$ for which $S_\rho(y, X)$ is integrable. Vice versa, elicitability implies backtestability.

If ρ is $\mathcal{F}$-elicitable with scoring function $S_\rho(y, x)$ which is striclty convex and continuously differentiable in y, then it is also $\mathcal{F}'$-backtestable with $\mathcal{F} \subseteq \mathcal{F}'$ and a backtest function

$$Z_\rho(y, x) = \frac{\delta S_\rho(y, x)}{\delta y}.$$

Since the **ES** is not elicitable it is not backtestable in the sense of the given definiton.

This can also be derived even more easily: In simple terms, the above definition of backtestability means that a risk measure can only be backtested if one finds a backtest function with an expected value of zero, which is only dependent on the risk measure and the P&L variable itself. It follows that the **ES** cannot be backtested according to this definition since it is dependend on the **VaR**.

The authors do not stop here, however, but rather propose an extended definition that is fulfilled. This leads to the so-called *ridge backtests*. Before that, the concept of *sharpness* is introduced.

3.4.3 Sharpness

A backtest function $Z_\rho(y, x)$ of a backtestable risk measure ρ is called *sharp* if it is strictly decreasing in the value $\rho(F)$ of the risk measure in the sense

[3] Expectiles are further risk measures and appear in the current debate as well. However, the decision of the BCBS does not result in them being described here.

that

$$\mathbb{E}_{F_1}[Z_\rho(y,X)] > \mathbb{E}_{F_2}[Z_\rho(y,X)] \quad \Leftrightarrow \quad \rho(F_1) < \rho(F_2) \;\; \forall F_1, F_2 \in \mathcal{F} \;\; \forall y.$$

This definition is also intuitive and desirable: If a prediction y is given, the absolute value of the expected value of the backtest function should increase the further away from the actual value of the prediction it is.

Note that, this is already a backtest function, which increases in y and has an expected value of 0 for $y = \rho(F)$. The property of positive (negative) expected values for an over- (under-)estimation is therefore retained. The above definition is thus also goal-oriented. To give an example, let $y = 5$ be a prediction and F_1 and F_2 two distributions with risk measures $\rho(F_1) = 4$ and $\rho(F_2) = 4.5$. One gets:

$$\mathbb{E}_{F_2}[Z_\rho(y,X)] < \mathbb{E}_{F_1}[Z_\rho(x,X)] < 0.$$

For $\rho(F_1) = 6$ and $\rho(F_2) = 6.5$ one gets:

$$0 < \mathbb{E}_{F_2}[Z_\rho(y,X)] < \mathbb{E}_{F_1}[Z_\rho(x,X)].$$

It follows that

$$\mathbb{E}_{F_1}[Z_\rho(y,X)] = \mathbb{E}_{F_2}[Z_\rho(y,X)] \;\text{ if }\; \rho(F_1) = \rho(F_2).$$

That means that $\mathbb{E}_F[Z_\rho(y,X)]$ depends only on $\rho(F)$. Therefore one can define the concept of sharpness equivalent to above: A backtest function $Z_\rho(y,x)$ of a backtestable risk measure ρ is called *sharp* if for all $F \in \mathcal{F}$ the expectation

$$\mathbb{E}_F[Z_\rho(y,X)] = \psi(y, \rho(F))$$

is a function which is increasing in y and decreasing in $\rho(F)$. Because of that, for any value z of the expected value there is only one compatible value

$$\rho(F) = \psi^{-1}(y,x)$$

of $\rho(F)$, where ψ^{-1} is the inverse function in the second argument of ψ.

A sharp backtest will not only check whether a prediction fits a real observation, but will also quantify the difference between observation and prediction.

They show that the backtest function for the **VaR** is not sharp. A backtest of a quantile is only able to testify whether or not the prediction fits the observation. If it does not, it does not provide a statement as to how far away the prediction is from the actual value. This causes problems like described in chapter 2.

For non-sharp backtests, even for a relatively safe containment of the expected value of the backtest function (for example, by a high number of observations), no conclusions can be drawn about the position of the actual value. In addition, there are also possible errors when measuring the expected value. The authors question whether with these findings any non-sharp backtests should be used at all.

3.4.4 Ridge backtests

As already mentioned above, the **ES** is not elicitable. In chapter 2 it is also mentioned, that there are findings that the pair (VaR,ES) is jointly elicitable. This becomes clear if one recalls the definition of elicitability since there exist terms with an expected value of 0 which include both the **VaR** and the **ES**, such as

$$\mathbb{E}[(X + ES)(X + VaR < 0)] = 0.$$

As already mentioned, one is looking for a function that depends only on the P&L variable and the risk measure. However, there are measures which inevitably depend on further dimensions. For example, the variance is dependent on the mean. The **ES** is dependent on the **VaR**. In general, there exists a condition like

$$\mathbb{E}[I(\rho_2(X), \rho_1(X), X)] = 0.$$

This means that both predictions must be tested simultaneously. The independent risk measure can of course be tested first and then accepted as given before the second dependent risk measure is tested. This is exactly the strategy which has been adopted for the derivation of AS_1. However, it remains unanswered how sensitive the backtest of the second risk measure reacts to potential errors when testing the first risk measure and vice versa.

To give a solution, the authors provide the following definition: A risk measure ρ_2 admits a *ridge* $F-$ backtest

$$Z_{\rho_2}(y_2, y_1, x) = h(y_2) - vS_{\rho_1}(y_1, x)$$

if it can be expressed (up to a strictly monotonic function g) as the minimum of the expected $\mathcal{F}$-scoring function S_{ρ_1} of an elicitable auxiliary risk measure ρ_1:

$$\begin{cases} \rho_2(F) = g(\min\limits_y \mathbb{E}_F[S_{\rho_1}(y, X)]) \\ \rho_1(F) = \operatorname{argmin}\limits_y \mathbb{E}_F[S_{\rho_1}(y, X)] \end{cases} \qquad \forall F \in \mathcal{F},$$

where $v \in \{-1; +1\}$ is the sign of the derivative of g and $h(x) = vg^{-1}(x)$.

This definition yields the following proposition:

Let Z_{ρ_2} be a ridge $\mathcal{F}$-backtest for ρ_2 with auxiliary risk measure ρ_1 as defined above. Then:

- the expected value of the backtest function is zero in the correct predictions for ρ_2 and ρ_1:

$$\mathbb{E}_F(Z_{\rho_2}(\rho_2(F), \rho_1(F), X)) = 0 \quad \forall F \in \mathcal{F};$$

- Z_{ρ_2} acts as a backtest for ρ_2 with a one-sided bias depending only on y_1 in the sense that

$$\mathbb{E}_F(Z_{\rho_2}(y_2, y_1, X)) = (h(y_2) - h(\rho_2(F))) - v\mathbb{E}_F(S_{\rho_1}(y_1, X) - S_{\rho_1}(\rho_1(F), X)),$$

  so that

$$\begin{cases} \mathbb{E}_F(Z_{\rho_2}(y_2, y_1, X)) \leq (h(y_2) - h(\rho_2(F))) & \text{if } \ v > 0 \\ \mathbb{E}_F(Z_{\rho_2}(y_2, y_1, X)) \geq (h(y_2) - h(\rho_2(F))) & \text{if } \ v < 0 \end{cases} \quad \forall y_1;$$

- If $\mathbb{E}_F(S_{\rho_1}(y_1, X), X)$ is continuously differentiable in $y_1 = \rho_1(F)$ for $F \in \mathcal{F}' \subseteq \mathcal{F}$, then

$$\mathbb{E}_F(Z_{\rho_2}(y_2, y_1, X) = (h(y_2) - h(\rho_2(F))) + O(y_1 - \rho_1(F))^2;$$

- Z_{ρ_2} is sharp for ρ_2 up to terms $O(y_1 - \rho_1(F))^2$.

Acerbi and Szekely depict this sentence with a hike on the ridge of a mountain.

> *"When you climb the ridge of a mountain, if you lose the way on either side of the edge, you can be sure of one thing: that you'll find yourself below where you should be."*

The same applies to the expected value of a backtest function for ρ_2 as compared to a false step with regard to the prediction of ρ_1. To understand that, suppose that $v > 0$ as it is for the **ES** as one will see below. Say that the risk of $\rho_2(F)$ is underestimated, $y_2 < \rho_2(F)$. This implies that the expected value of the backtest function is negative:

$$\mathbb{E}_F(Z_{\rho_2}(y_2, y_1, X)) < 0.$$

An imperfect prediction y_1 will make the expected value more negative, for both under- and overestimation of ρ_1. However, the proposition says that the impact is in a small scope for small misjudgements around the actual value.

3.4.5 The derivation of a ridge ES backtest

Based on the findings above, Acerbi and Szekely propose another backtest, namely a ridge **ES** backtest. If one recalls AS_2, one observes that the backtest function is given by

$$Z_2(e, v, x) = \alpha e + x(x + v < 0).$$

This follows from the assumption of continuous cdfs, where $E[Z_2(ES, VaR, X)] = 0$ (see equation (2)). But if one recalls the definition for general distributions from equation (1), one gets the following backtest function:

$$Z(e, v, x) = \alpha(e - v) + (x + v)(x + v < 0).$$

To understand that this is a ridge backtest, consider the following: Following Uryasev and Rockafellar (2002), the **ES** can be represented by the following equations which fulfill the prerequisites for the construction of a ridge backtest.

$$q_\alpha = \underset{y}{\operatorname{argmin}}\{-y + \frac{1}{\alpha}E[E(X - y)^-]\};$$

$$ES_\alpha = \min_y\{-y + \frac{1}{\alpha}E[E(X - y)^-]\}.$$

Note that $g(x) = x$ and therefore $v = 1$. Now recall that

$$S_{q_\alpha} = \alpha(x - y)^+ + (1 - \alpha)(x - y)^-$$

is a scoring function for q_α. If one transforms this function with α and $-x$, one gets

$$S'_{q_\alpha} = -y + \frac{1}{\alpha}(x - y)^-,$$

which is also a scoring function for q_α.

Let $v = -y$, then it follows that $S_v(y, x)' = v + 1/(\alpha)(x + v)^-$. The addition of $h(e) = e$ leads to:

$$Z = h - 1 * S_v = e - v - \frac{1}{\alpha}(x + v)^-.$$

Multiplication with α leads to:

$$Z = \alpha(e - v) + (x + v)^-.$$

This is precisely the backtest function as defined above. So, the test statistic

for the fourth test of Acerbi and Szekely (AS_4) is given by:

$$Z_4 = \sum_{t=1}^{T} \frac{\alpha(ES_{\alpha,t} - VaR_{\alpha,t}) + (X_t + VaR_{\alpha,t})(X_t + VaR_{\alpha,t} < 0)}{T\alpha ES_{\alpha,t}} + 1,$$

where $(VaR_\alpha, ES_\alpha)_t$ is the forecast. The null hypothesis is the same as for AS_2, namely the correct prediction in the tail. The alternative hypothesis is given by

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

The **VaR** must also be tested, but this does not have to be so strict. The advantage over AS_2 becomes clear in the examples in chapter 5.

However, what is not mentioned in the paper is the influence on the distribution and the critical values of the asymptotic distribution of the backtest function. A general statement cannot be made without further investigations. The simulations in this thesis will show that although the expected value of the test statistics of AS_4 for a misjudgement of the **VaR** has shifted to negative as described above, but the span of the left part of the asymptotic distribution becomes smaller in case of a **VaR** overestimation. This also opens up potential sources of fraud, which will be further investigated in chapter 5.

3.5 Further backtest suggestions

In the past few years, there have been a number of further publications and proposals on how to backtest **ES** predictions.

Based on the representation of the **ES** as the integrated **VaR**, Emmer et. al. (2015) propose to approximate the **ES** as

$$ES_\alpha = \frac{1}{\alpha}\int_0^\alpha VaR_u du \approx \frac{1}{4}(VaR_{\frac{\alpha}{4}} + VaR_{\frac{\alpha}{2}} + VaR_{\frac{3\alpha}{4}} + VaR_\alpha).$$

If all four quantiles are backtested and accepted, the **ES** prediction is accepted as well. Wimmerstedt (2015) shows that the numbers of **VaR** exceedances for different α are not independent. The BCBS (2016) propose to perform a backtest based on the **VaR** for a confidence level $\alpha = 0.025$ and $\alpha = 0.001$. If one approximates the **ES** as the mean of this two **VaR**, one gets a considerably lower significance than required. To give an example, for an iid $N(0,1)$ time series with length of 250 days, the acceptance rate for a correct **ES** prediction with significance level $p = 0.05$ is only 0.86.

The proposal of Du and Escanciano (2016) is based on the idea of Emmer et. al. (2015). Recall that $I_{\alpha,t} = (X_t < -VaR_{\alpha,t})$ is the indicator variable,

which is 1 if the **VaR** for a confidence level α is breached. They define the *cumulative violation process* as

$$H_t(\alpha) = \frac{1}{\alpha}\int_0^\alpha I_t(u)du = ... = \frac{1}{\alpha}(\alpha - P_t(X_t))(X_t < -VaR_{\alpha,t}).$$

They show that the expected value of $H_t(\alpha)$ is $\frac{\alpha}{2}$ and the variance is $\alpha(\frac{1}{3} - \frac{\alpha}{4})$. Based on that, they test if the mean of $(H_t(\alpha))$ for $t = 1, .., T$ fits to $\frac{\alpha}{2}$ (*unconditional test*) and wheather the random variables $(H_t(\alpha))$ are uncorrelated (*conditional test*). Note that this is not a test of the **ES** but a test of the entire tail distribution. An extension of the unconditional test is given by Löser et. al. (2016). By reformulating the null hypothesis, they give a new test statistic which should be easier to implement and easier to extend to a multivariate framework. Again, the entire tail distribution is tested.

Righi and Ceretta (2013) pursue another idea. They consider the conditional distribution in the tail:

$$P(X \leq \cdot | X \leq -VaR).$$

They define a *dispersion measure* which is the SD around the **ES** in the tail. Based on that, one can check daily if a **VaR** exceedance is significantly higher than the **ES**. A similar approach has already been presented by McNeil and Frey (2000), who use the SD of the complete distribution and not only the conditional one.

Note that all methods which were presented are "traditional" backtests in the sense that they return an answer "yes" or "no" for a model validation at a given significance level. As already seen above, elicitability, namely the existence of a scoring function, makes it possible to compare different models. This is called *comparative backtesting*. Nolde and Ziegel (2017) distinguish between the two approaches and argue that for comparative backtesting, elicitability is necessary. They argue that the comparison between a standard approach and an internal model should be part of the regulatory framework. They make concrete suggestions for the joint scoring function of **VaR** and **ES**, based on simulated and empirical data.

In this thesis, the backtests of Acerbi and Szekely (2014, 2017) and Corbetta and Peri (2016) are presented and checked in different settings. The reason for this is that they are easy to handle compared to other backtests. In addition, as already seen, they are very well placed in a theoretical framework.

4 Modeling Returns

So far, one has gained insight into the question of what constitutes a risk measure and which properties it desirably should fulfill. To review the **ES** backtests presented in chapter 3 in a theoretical framework, first the question of how to model returns and thus make predictions must be addressed.

4.1 Standard distributions

The first, simple to model returns is using the normal distribution. The density function is symmetrical about the expected value and can be varied via the variance. Common tests reject the normal distribution hypothesis for most financial market data (see Sheikh and Qiao, 2010). This is partly due to the symmetry, on the other hand due to fat tails. It will be included here anyway, as the backtests should provide reliable results, at least for simple settings.

The normal distribution is defined by its first two moments. In general, the first four moments of a distribution are of particular interest since this makes it possible to model more realistic distributions. For a random variable X, they are given as follows:

- The mean $\mu(X) = \mathbb{E}(X)$
- The variance $\sigma^2 = \mathbb{E}((X - \mu)^2)$
- The skewness $s = \mathbb{E}((\frac{X-\mu}{\sigma})^3)$
- The kurtosis $k = \mathbb{E}((\frac{X-\mu}{\sigma})^4)$

The first moment is intuitive, it is the expected value of a random variable. The variance of a random variable is the squared deviation from its expected value. The skewness describes the nature and expression of the asymmetry of a distribution. The kurtosis is a measure of the steepness of a distribution.

To simulate more fat tails, one can use the t distribution with v of freedom. Note that with an increasing number of degrees of freedom, the t distribution converges towards the normal distribution. The lower the number of v, the higher the kurtosis for the t distribution. However, it is symmetrical, the skewness is zero. In the context of risk measures, skewed distributions are interesting and empirically better suited for modeling. Besides, returns are often leptokurtic, meaning they have a stronger kurtosis than the normal distribution. As an example, the daily returns of Beiersdorf AG are shown in Figure 2. The skewness can be introduced into the two named distributions. This results with the so-called skewed normal distribution and skewed t

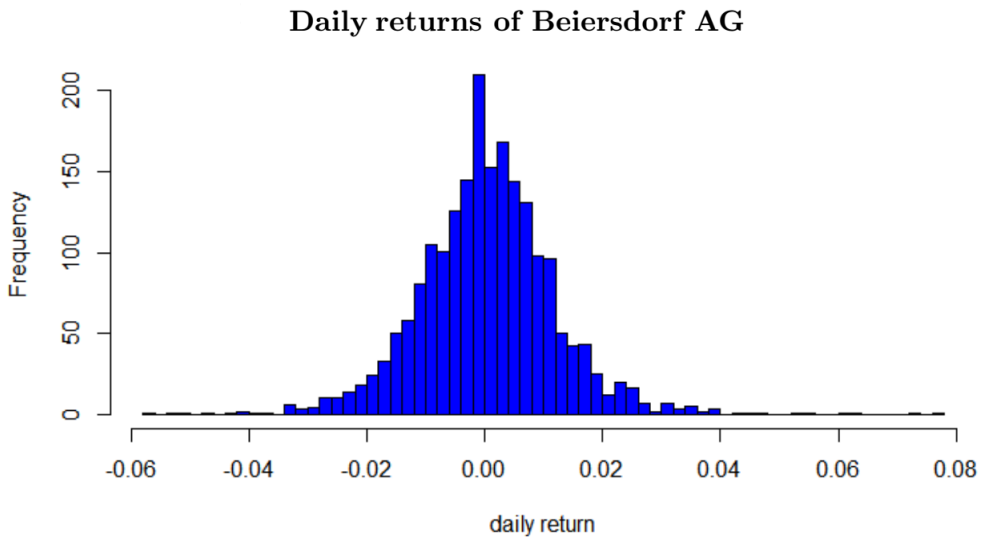

Figure 2: The daily returns of Beiersdorf AG from the beginning of 2010 until the end of 2017. The empirical mean is 0.00044, the empirical SD is 0.0118, the empirical skewness is 0.352 and the empirical kurtosis is 6.637. In particular, fat tails are observed.

distribution. There are several approaches in the literature to model the skewness. For the skewed normal distribution as suggested by Azallini (1985), the density is given by

$$f(x) = 2\phi(x)\Phi(\gamma x),$$

where ϕ is the density function of the standard normal distribution, Φ is its' cdf and γ is a shape parameter. For $\gamma > 0$ (< 0), the distribution is right (left) skewed.

The density of a skewed t distribution with v degrees of freedom proposed by Fernandez and Steel (1998) is given by:

$$f(x) = \begin{cases} \frac{2}{\gamma + \frac{1}{\gamma}} g(\gamma x) & x < 0 \\ \frac{2}{\gamma + \frac{1}{\gamma}} g(\frac{x}{\gamma}) & x \geq 0, \end{cases}$$

where g is the density of a t distribution with v degrees of freedom and γ is a shape parameter. For $\gamma > 1$ (< 1), the distribution is right (left) skewed. Note that modeling the skewness shifts the mean for both distributions. During the analysis in this thesis, a re-shift is executed. The other moments are not affected by this. Some examples are shown in figure 3.

This leads to a first modeling possibility for the returns. Based on historical data, the necessary parameters can be estimated. This raises the question of

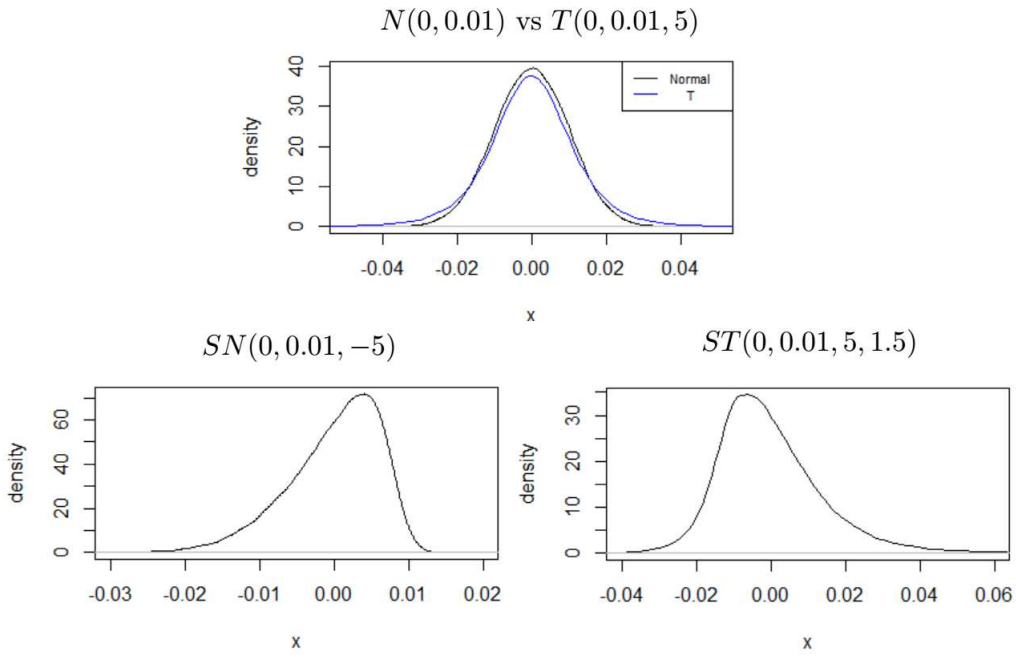

Figure 3: Density functions for a normal distribution, a t distribution, a left skewed normal distribution with shape parameter -5 and a right skewed t distribution with shape parameter 1.5. The SD/ SD scale is set to 0.01 to get more realstic values for daily returns.

whether the entire historical data is used to estimate, if the historical data is weighted according to how far historical values are in the past or not, or whether one uses only a fixed number of past data for the estimation ("moving window"). This question is not part of the present thesis.

4.2 ARMA and GARCH processes

As a rule, the hypothesis of identically distributed returns must be rejected for financial time series (see Sheikh and Qiao, 2010). A more realistic approach to model financial time series are so-called ARMA and GARCH processes. These are stochastic processes with which one can model temporally varying expected values, volatilities and autoregressions. The models presented below are taken from the illustration of Pfaff (2016).

Auto regression (AR) and Moving average (MA)

If one or more values from past events influence the current return, this is called *autoregression* (AR). If one or more past innovations influence the current return, this is called a *moving average* (MA). The influence of the last p returns and the influence of the last q innovations is modeled. Combining these two models one speaks of an ARMA(p, q) model. Note that an

ARMA$(p, 0)$ model is an AR model and an ARMA$(0, q)$ model is an MA model. The process is given as follows:

$$X_t = \mu + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t,$$

where μ, α_i and β_i are real numbers and the so-called *innovations* (ε_t) are iid with expected value 0 and variance 1. For example, an ARMA$(1, 0)$ process with $\alpha = 1$ is a simple random walk. The estimation of the parameters α_i and β_i as well as the question about the optimal number of p and q is again part of the estimation theory and should not be further elaborated here.

Generalized autoregressive conditional heteroscedasticity (GARCH)

In reality, it can be observed that the volatility of financial market data is not constant but itself fluctuates over time. Large shocks often lead to high volatility in the following months, which only fades away over time. As an example one can see the daily returns of the share of Volkswagen AG from the beginning of 2007 until the end of october 2017 in figure 4. This phenomenon can be well modeled with the so-called GARCH models (*generalized autoregressive conditional heteroscedasticity*).

A GARCH (q, p) time series is defined as follows:

$$X_t = \sigma_t * \varepsilon_t;$$
$$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2,$$

where (ε_t) are iid as above. The parameters ω, α_i and β_i are positive, so that the conditional variance is positive. The sum of the parameters α_i and β_i must be smaler than 1 to obtain stationarity of the time series: $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$.

Todays' volatility depends on past returns and past volatilities. There exist a number of extensions, such as the EGARCH model or the TGARCH model, where negative returns lead to higher volatility. Thus, asymmetries can be modeled. Of course the presented models can also be combined. The result is an ARMA-GARCH process. The models used in the analysis in chapter 5 are listed once again there in detail.

4.3 Forecasting VaR and ES

An ARMA-GARCH model provides both a forecast for the expected value and for the variance. Depending on which distribution one has chosen for the innovations ε_t, a distribution results for the forecast, P_t. From this distribution the daily predicted **VaR** and **ES** can be calculated. There exist other alternatives like the historical simulation, the filtered historical simulation or the exponential weighted moving average approach for forecasting. These should not be part of this thesis. For an example, see figure 5 where the **VaR** and the **ES** for the daily returns of Lufthansa AG is shown. The forecast is based on a GARCH$(1, 1)$ model with standard normal distributed innovations.

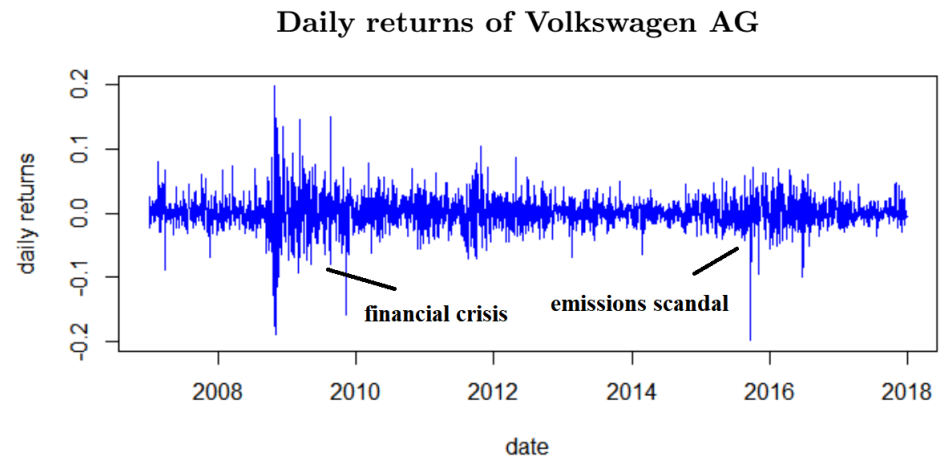

Figure 4: The daily returns of Volkswagen AG from the beginning of 2007 until the end of 2017. One can observe two very volatile phases, which belong to the financial crisis in 2008/2009 and to the emissions scandal in 2015. After the shocks, the returns are initially much more volatile. This effect decreases over time.

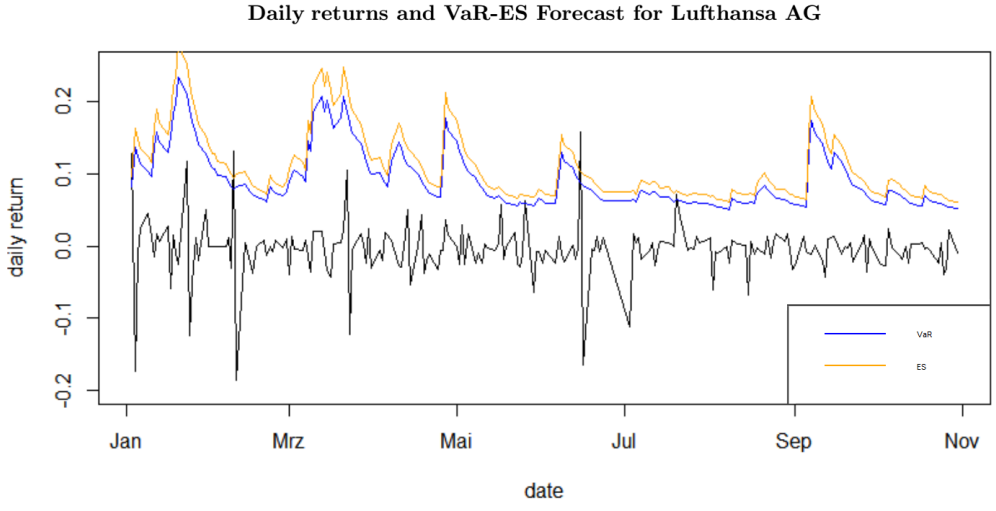

Figure 5: The daily returns and VaR and ES forecasts of Lufthansa AG from the beginning of 2017 until the end of october 2017. The returns are modeled as a GARCH$(1,1)$ process. The data of year 2016 was used to estimate the parameters α and β.

5 The Practice Test

In this section, the performance of the presented backtests is checked in a theoretical framework. The advantages and disadvantages of the presented backtests will be demonstrated by means of examples.

5.1 Standard distributions in the iid setup

At first the simplified assumption is made, that the returns are iid. In reality, autocorrelations and time-dependent volatilities occur in particular. This will be discussed later. However, the results in a simpler framework often point in the right direction. In this subchapter, it is first assumed that the returns X_t are iid and belong to a certain cdf F. The period of one or two banking years is considered, that is $T = 250$ or $T = 500$ days.

For the analysis, the four distributions discribed in chapter 4 are used. For the sake of simplicity, it is always assumed here that the expected value of the returns is zero. The relationship between the expected value and the **ES** is linear. If the expectation value shifts the **ES** moves in the same direction with the same size.

	H_0 is correct	H_0 is false
H_0 is rejected	Type 1 error	
H_0 is accepted		Type 2 error

Table 3: Type 1 and type 2 error.

In the following, the significance and the power of the backtests are derived under different assumptions and actual realizations. The significance is the probability of a type 1 error which describes the case in which a correct null hypothesis is rejected. The statistical power is the probability that the null hypothesis is rejected, when in fact the alternative hypothesis applies. The significance can usually be specified. However, a statement about the power can only be made if the actual distribution is known. Since one does not know this in reality, one can only test different distributions against each other to get an idea.

Since there are no analytical results for the backtests, the significance and power which means the acceptance and rejection rates have to be approximated by Monte Carlo simulations.

5.1.1 Significance - Correct ES predictions should be accepted

The first point of the analysis is the question with which probability a backtest will confirm a true prediction. For the standard distributions, the probability should not be below the significance level indicated in the backtest. The significance of the backtests of Acerbi and Szekely should lie at the desired level of significance. Since the distribution of the respective test statistic must be simulated under the assumed distribution P, even with a correctly predicted underlying distribution, the probability of a falsely rejection is p.

This is not mandatory for the backtests of Corbetta and Peri. The problem here is the CP_2 in particular. On the one hand, a rejection of the null hypothesis must be tested in both directions. On the other hand, use is made of the convergence statement, which, however, only applies to a very large number of days and thus also to exceedances.

$T = 250$

F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
$N(0, 0.01)$	0.946	0.953	0.951	0.950	0.902	0.965
$T(0, 0.01, 5)$	0.950	0.950	0.952	0.952	0.936	0.936
$SN(0, 0.01, 5)$	0.953	0.944	0.938	0.936	0.891	0.957
$SN(0, 0.01, -5)$	0.938	0.927	0.957	0.951	0.899	0.963
$ST(0, 0.01, 5, 1.5)$	0.956	0.956	0.952	0.945	0.937	0.937
$ST(0, 0.01, 5, 0.5)$	0.946	0.938	0.934	0.934	0.948	0.948

$T = 500$

F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
$N(0, 0.01)$	0.949	0.953	0.946	0.946	0.941	0.965
$T(0, 0.01, 5)$	0.951	0.948	0.950	0.950	0.933	0.955
$SN(0, 0.01, 5)$	0.950	0.950	0.947	0.934	0.930	0.954
$SN(0, 0.01, -5)$	0.927	0.960	0.955	0.957	0.927	0.957
$ST(0, 0.01, 5, 1.5)$	0.939	0.944	0.941	0.941	0.934	0.957
$ST(0, 0.01, 5, 0.5)$	0.931	0.944	0.949	0.942	0.942	0.962

Table 4: Acceptance rate of the backtests for different distributions, each based on 10^5 simulations.

Table 4 shows the significances for different distributions. It can be seen that the significance for the test of Acerbi and Szekely is in the range of $1 - p = 0.95$. This also applies to CP_2. The acceptance rate of CP_1 depends on the distribution and can be increased by a longer backtesting period.

5.1.2 Significance for a fixed number of VaR exceedances

As already mentioned in chapter 2, if the null hypothesis is correct, the number of **VaR** exceedances is binomially distributed with parameters T and α. Ideally, the significance of the backtests should not depend on the number of

exceedances. To check this, proceed analogously to the investigations done by Wimmerstedt (2015): Instead of random simulations of (X_t), a time series with exactly n **VaR** exceedances are simulated. Note that the analysis for AS_1 on a time series of 250 (500) days is only interesting up to 4 (9) exceedances, as for 5 (10) and more exceedances, the null hypothesis whould be rejected by a simple binomial test with $\alpha = 0.025$ to the significance $p = 0.05$.

To model a fixed number of **VaR** exceedances, n quantiles between 0 and α are simulated and the corresponding values of the distribution are calculated. Then, $T - n$ quantiles between α and 1 and the corresponding values are simulated. For each n, the significance is calculated using M simulations. The critical values are derived only once in a purely random setting. As an example, figure 6 shows the significance of the backtests against the **VaR** exceedances for a t distribution.

Recall that the test statistic of AS_1 tests the "empirical **ES**" vs. the predicted **ES**. This becomes statistically more secure as the number of **VaR** exceedances increases. The other backtests of Acerbi and Szekely behave inversely, the significance decreases as the number of exceedances increases. Although AS_2 and AS_3 tend to lose significance even with fewer exceedances, they are more stable for a higher number.

For the backtests of Corbetta and Peri a similar result can be observed, the significance decreases with the number of exceedances. This is intuitive: it counts the number of **ES** exceedances. Of course, as the number of **VaR** exceedances increases, so does the expected number of **ES** exceedances. For other distributions and for a longer period of 500 days, one can observe similar results.

For the backtests of Acerbi and Szekely, one way would be to first observe the number of **VaR** exceedances and calculate the critical values by Monte Carlo simulations with just that number of exceedances after. The significance should then be p regardless of the number of exceedances. However, this is not possible in the ARMA or GARCH setting, as it is not not possible to model paths with a fixed number of exceedances. So one should keep in mind the results in a setting with standard distributions.

5.1.3 Power - False ES predictions should be rejected

Next, the power of the backtests will be analyzed. It should give an impression of what affects the power. Is it just the difference between the **ES** prediction and the actual **ES**? Or do the **VaR** prediction, the type of distribution, the period length or the number of **VaR** exceedances play a role as well? To get a first impression, see table 5. Let the daily returns be predicted as t distributed with SD scale 0.01 and $v = 5$ degrees of freedom. The values for the **VaR** and

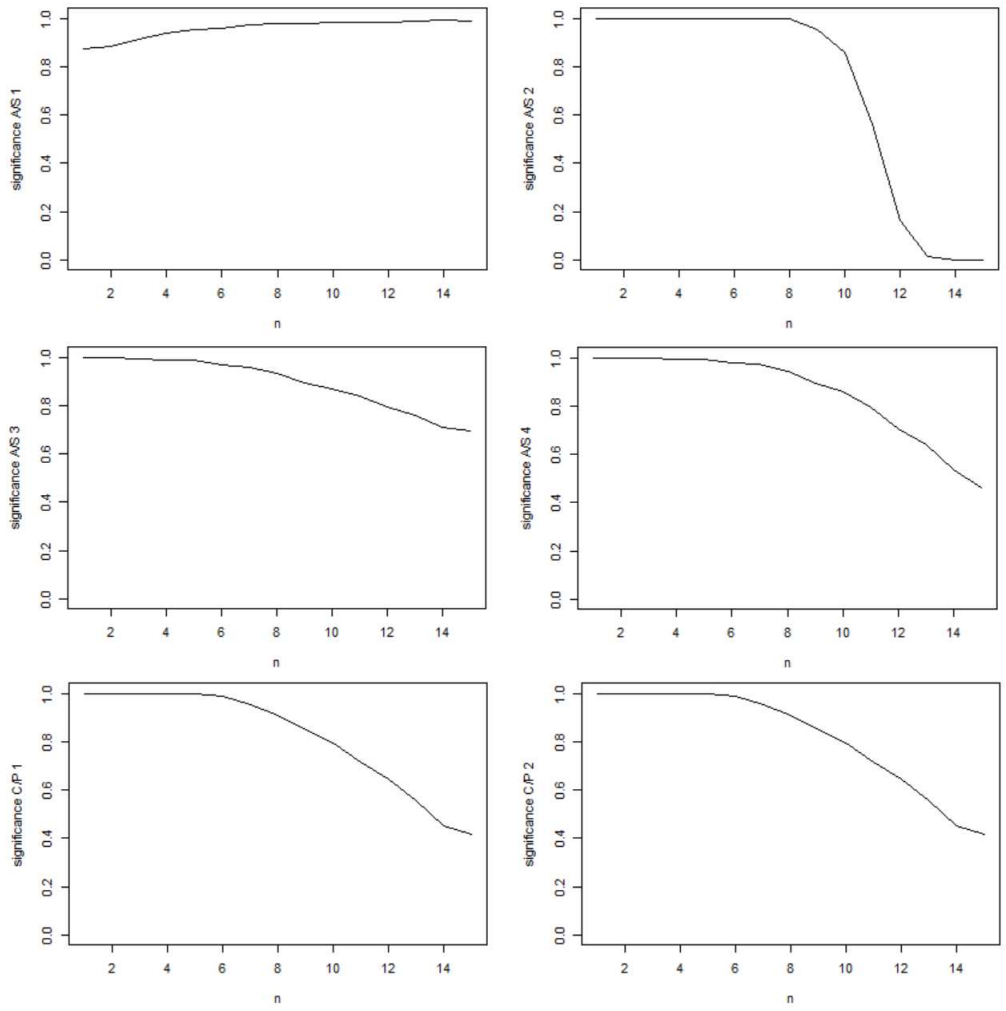

Figure 6: The significance of the backtests, depending on the number of **VaR** exceedances. The returns are iid t distributed with 5 degrees of freedom and SD scale of 0.01. The length of the time series is $T = 250$. The confidence level is given by $\alpha = 0.025$, the desired significance level is $p = 0.05$. For each fixed number of exceedances as well as for the calculation of the critical values, 10^5 simulations were made.

$T = 250$

F	VaR_F	ES_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
$T(0, 0.01, 4)$	0.028	0.040	0.091	0.207	0.186	0.205	0.183	0.183
$T(0, 0.01, 3)$	0.032	0.050	0.276	0.615	0.595	0.637	0.554	0.554
$ST(0, 0.005, 5, 0.3)$	0.037	0.054	0.234	0.933	0.791	0.889	0.861	0.861
$SN(0, 0.03, -5)$	0.043	0.054	0.008	1.000	0.932	0.995	0.995	0.995

$T = 500$

F	VaR_F	ES_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
$T(0, 0.01, 4)$	0.028	0.040	0.159	0.304	0.302	0.314	0.278	0.177
$T(0, 0.01, 3)$	0.032	0.050	0.450	0.822	0.811	0.840	0.766	0.634
$ST(0, 0.005, 5, 0.3)$	0.037	0.054	0.442	0.996	0.963	0.987	0.979	0.959
$SN(0, 0.03, -5)$	0.043	0.054	0.037	1.000	1.000	1.000	1.000	1.000

Table 5: Power of the backtests for different distributions F, each based on 10^5 simulations.

the **ES** are 0.025 and 0.035. Different actual distributions provide different power for the backtests. One gets a first idea that the length of the period but also the type of distribution has an impact on the power. Compare the alternatives 3 and 4 for example, which have a similar **ES** but different power for some backtests. This should be analyzed more detailed below.

5.1.4 Underestimated SD

In this setting, the power of the backtests will be examined if the SD in the prediction was underestimated. Note that this underestimates both the **VaR** and the **ES**. Both the symmetrical distributions and the skewed distributions are tested. For visualization purposes, the power of the backtests for a normal distribution and a t distribution is shown in figure 7.

One can observe that AS_1 shows a weak performance in every setting. Its power is dominated by all other backtests. In addition, the rejection rate increases for an overestimated SD scale and a period length of $T = 250$ days. However, a conservative prediction should not be rejected. For normal distributed returns, the performance of all backtests except AS_1 is very similar. These can only be improved slightly by a longer backtesting period. In case of $T = 500$, the rejection rate of CP_2 shows the same behavior as the AS_1 for $T = 250$. This makes sense since the approximation is tested in both directions. In case of the iid assumption and the standard distributions, the test statistic thus moves in the same direction as the prediction of the SD.

An interesting result is that the power of the backtests for fat tails differ significantly from each other. To get an idea, take a look at the power for a t distribution with 5 degrees of freedom. Here, the power of AS_2 is the best

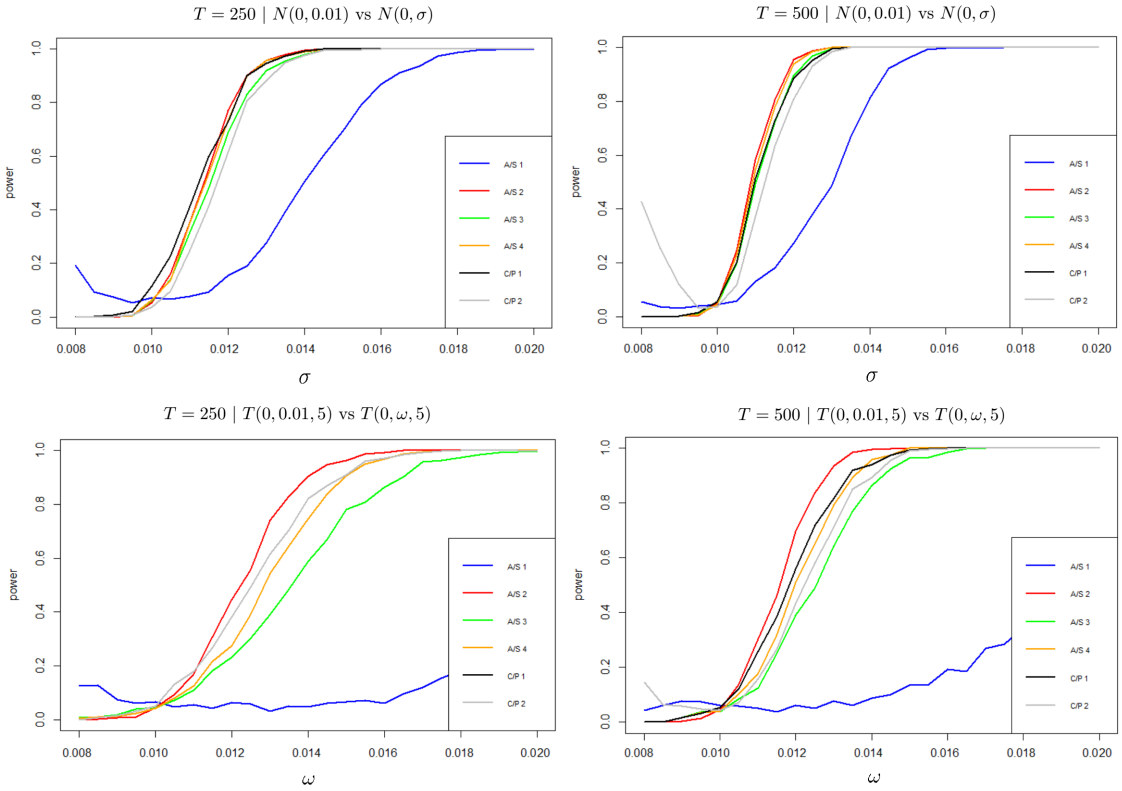

Figure 7: Power of the backtests, depending on the true value of the SD scale. To calculate the power, the interval $[0.008, 0.02]$ is approximated by 240 points. For every point, the power of the backtests is calculated via 10^4 simulations. For the calculation of the critical values, 10^5 simulations were made.

one. The power of the AS_1 is very weak and is still below 0.3 when all other backtests are already at 1. In addition, the power of all backtests increases slower than in the normal distribution case. This is against the intuition since with fat tails one would expect that the estimation of the actual **ES** is easier. Note that if one compares a predicted normal distribution with an actual t distribution, the power of AS_3 and AS_4 dominate all others. This two backtests recognize the fat tails best. Here, the power of all backtests can be significantly improved by extending the backtesting period.

More precisely, the power of course depends only on the structure in the left tail. If one compares the power for a left skewed t distribution with the ones for a right skewed t distribution (figure 8), one observes the same as above. For a left skewed distribution, AS_2 is the best one as well as for a right skewed distribution. However, the advantage is clearer in the first case. Again, the power can be increased by increasing the backtesting period. Note that, for a skewed normal distribution, one gets similar results.

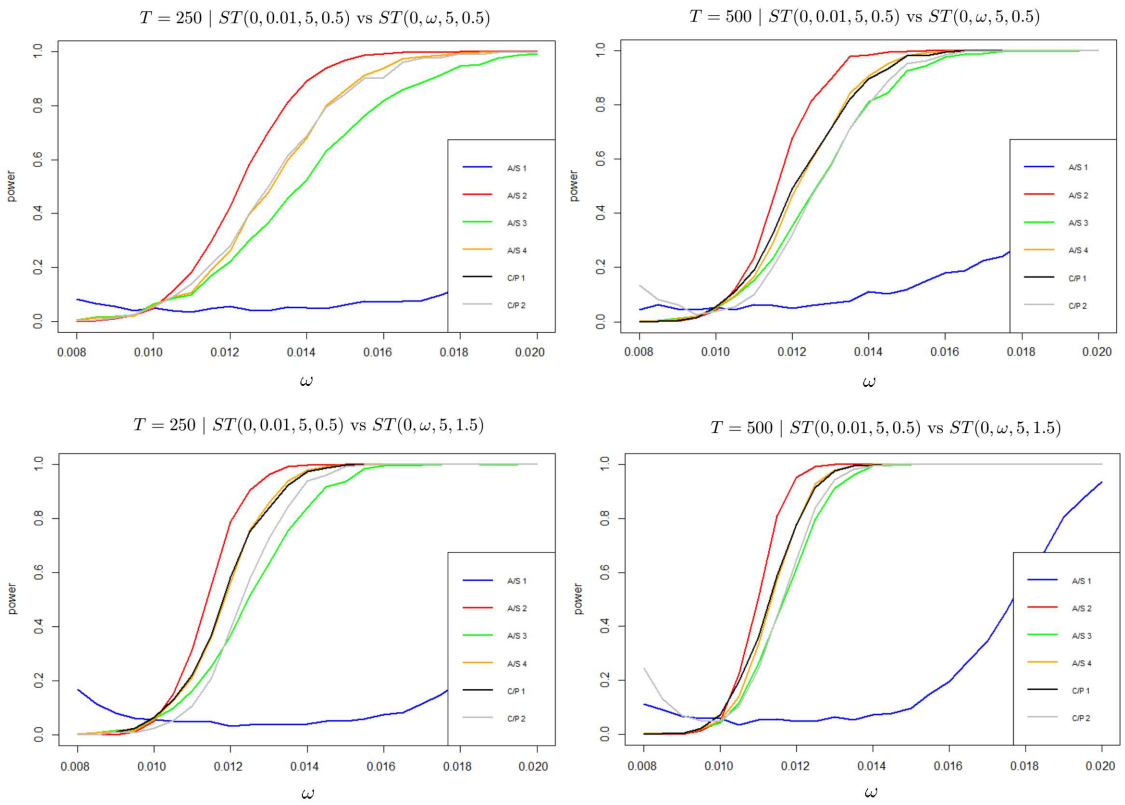

Figure 8: Power of the backtests, depending on the true value of the SD scale. To calculate the power, the interval $[0.008, 0.02]$ is approximated by 240 points. For every point, the power of the backtests is calculated via 10^4 simulations. For the calculation of the critical values, 10^5 simulations were made.

5.1.5 Power for a fixed number of VaR exceedances

Analogous to the significance, the question arises as to whether and to what extent the number of **VaR** exceedances has an influence on the power of the backtests. Note that in this case, the predicted **VaR** is meant.

Consider the following example: Say that the returns under F are iid and normal distributed with a SD of 0.0125. Given a confidence level of $\alpha = 0.025$, this leads to a **VaR** of 0.025 and an **ES** of 0.029. The prediction P is based on the assumption, that the distribution is a normal one but the SD is 0.01. The **VaR** as well as the **ES** are underestimated. The prediction has values of 0.0196 and 0.0234. Analogously as above, one can model a fixed number of **VaR** exceedances and calculate the power of the backtests. The resulsts for a backtesting period length of 250 days are shown in figure 9. Note that one gets similar results for different distribution types.

AS_1 shows the same weakness as described above, so it is no longer considered in this example. Note that in this example, the expected number of **VaR** exceedances is 12. For $n = 12$, AS_2 has the highest power with 0.962.

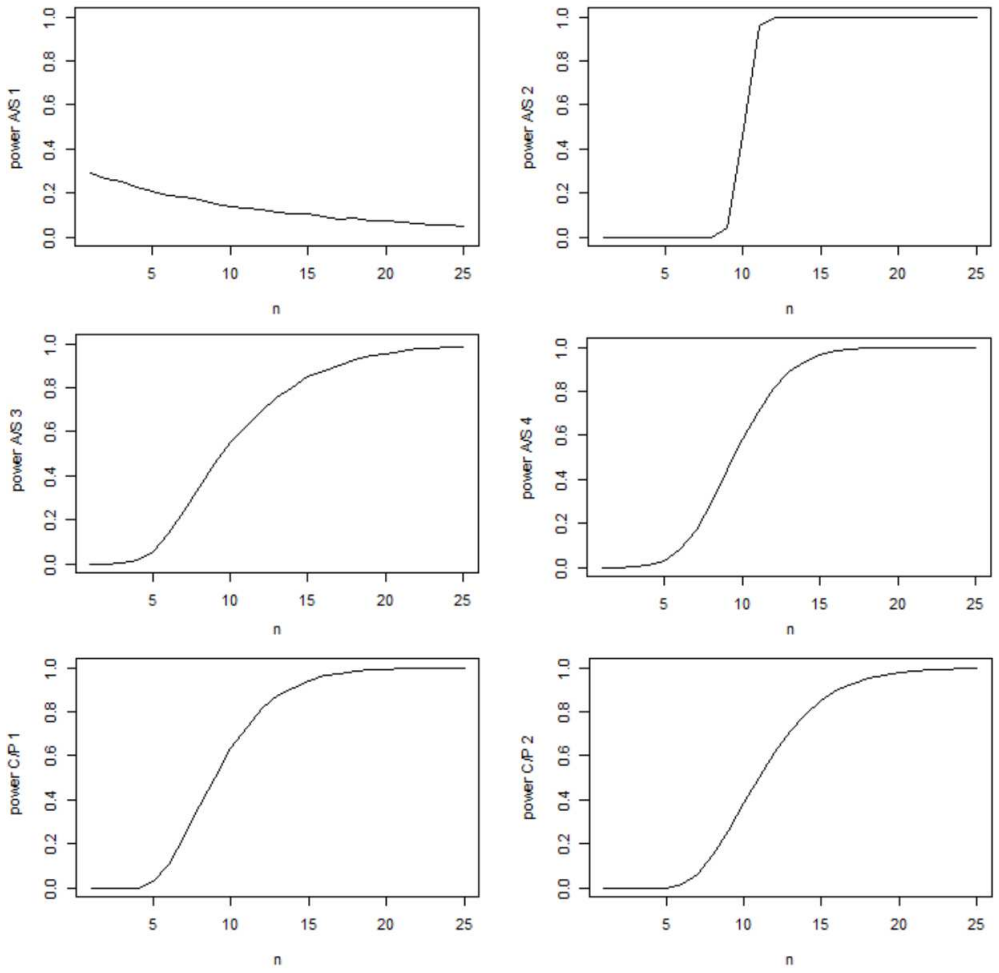

Figure 9: Power of the backtests, depending on the number of **VaR** exceedances. To calculate the critical values and the power, each 10^5 simulations were made.

It increases very fast and is already 1 for 13 and more exceedances. The other backtests are rising relatively constantly. But for a small number of exceedances (lower than 11), all backtests show a better power than AS_2.

One can get a premonition at this point. In case that the **VaR** prediction is correct, there is only a small number of exceedances. Will the power of the backtests be worse, despite an underestimated **ES**? Is one moving towards a tricky tradeoff? This question will be examined in the next subchapter.

5.2 What influence does the VaR have?

Since the **VaR** prediction influences some of the test statistics, it is an interesting task to figure out how they affect the significance and the power. The question arises as to whether the **VaR** should be completely eliminated as a risk measure or whether the **ES** and the **VaR** should be tested simultaneously. If the **ES** is the only critical risk measure, the **VaR** should at best have

no major impact on the results of the backtests. Specifically, this means that potential prediction errors in terms of **VaR** should be as small as possible to the statements about the **ES**, if the **ES** is predicted correctly. On the other hand, an underestimation of the **ES** should be discovered by the test, even with a correct **VaR** prediction.

However, at the beginning of the chapter one has already seen that the nature of the actual distribution and also the **VaR** prediction can have an impact on the power. This will be examined here with some examples.

5.2.1 Correct ES but wrong VaR prediction

Consider the following example: Say that the returns are iid and t distributed with 3 degrees of freedom and a SD scale of 0.01. For $\alpha = 0.025$, this leads to a **VaR** of $VaR_F = 0.0318$ and an **ES** of $ES_F = 0.0504$. In addition, the bank works with a very simple distribution for the forecast: In the tail, the returns are predicted to be uniform distributed on the interval $[-ES_F - (ES_F - VaR_P), -VaR_P]$ where VaR_P is the prediction for the **VaR** which is different to the true **VaR**. With probability $1 - \alpha$, the distribution of the returns is modeled as a uniform distribution on the intervall $[-VaR_P, VaR_P + \frac{2\alpha}{1-\alpha} ES_F]$. Note that in this case, the expected value for the returns is zero and the predicted **ES** matches the actual **ES**. Despite the false precition for the **VaR**, the significance of the backtests should be high. The null hypothesis should not be rejected. Table 6 shows the acceptance rates for different **VaR** predictions.

In this example, the backtests of Corbetta and Peri give the best results. This is clear since the **ES** was correctly predicted and the assumptions is rejected only for a significant number of **ES** violations. AS_1 and AS_2 are strongly dependent on the **VaR**. AS_3 only gives a better result than AS_4 for a strong underestimation of the **VaR**. One can observe the relative stability of the AS_3 and AS_4, both against underestimation and overestimation of the **VaR**. Note that even for a correct prediction in terms of **ES** and **VaR**, the significance of the backtests of Acerbi and Szekely is not at the desired level p. It can therefore be seen that differences in the structure of the actual and the predicted distribution also have an influence.

For a backtesting period of $T = 500$, the power decreases in some parts. This applies to AS_1 and an overestimation of the **VaR**. If the **VaR** is underestimated, the power for AS_2 decreases, as well as for AS_4 with a strong overestimation or underestimation. The influence of the characteristics of the distributions becomes clearer if one changes the distribution type of F and P. Table 7 shows the results in this case. The different **VaR** are chosen so that the difference between actual **VaR** and predicted **VaR** are the same as above.

Due to the structure of the actual distribution, there are many more **ES** exceedances, the backtests of Corbetta and Peri become weaker. In contrast, the backtests of Acerbi and Szekely are mostly getting better.

5.2.2 Correct VaR but wrong ES prediction

Once the **ES** is used to calculate the regulatory capital, the **VaR** should at best have no impact on the prediction rejection rate if the **ES** was underestimated. If it were like that, banks with an underestimated **ES** in the same amount but with different **VaR** estimates would have to accept potentially different capital surcharges. In the interests of fair competition, this should not happen.

Consider the following example. A bank models the returns normally distributed with mean 0 and SD 0.01. This leads to a daily **VaR** prediction of 0.0196 and a daily **ES** prediction of 0.0233. In fact, the daily returns are t distributed with SD scale of 0.01. Table 8 shows the power of the backtest for different settings. These differ in the number of degrees of freedom and whether or not the true **VaR** is the same as the predicted. This can be easily simulated with a mean shift equal to the difference between the predcited **VaR** and the **ES** depending on the number of degrees of freedom.

What one can observe is that for similar **ES** the **VaR** prediction has an impact on the power of all backtests (!). If the **VaR** has been determined correctly, AS_1 always shows the best power. That is intuitive again, regarding the test statistics and the assumption that the **VaR** has been tested correctly. Now take a look at the difference between the T5-Shifted and the T9 distribution. Both have an **ES** of almost the same size, namely 0.0291 and 0.0288. However, the **VaR** is also underestimated for the T9 distribution. Desirable would be a power in the same order of magnitude for both settings, but one does not observe that. The power of the backtests of Corbetta and Peri is almost doubling. Regarding the backtests of Acerbi and Szekely, removing the correct **VaR** prediction decreases the power of AS_1. Also AS_2 is very sensitive compared to the **VaR**, the power almost triples. Only AS_3 and 4 are reasonably stable for both settings. Although the underestimated **VaR** also increases the power for these two backtests, it only increases by about a quarter for AS_4, and only about 15% for AS_3. If one compares the T10 Shift with T20, one can observe similar effects. It is already very close to the actual **ES**, but most of the backtests are still very sensitive to the **VaR**. AS_3 and 4 are stable in their power.

$T = 250$

VaR_P	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.02	0.998	0.439	0.905	0.858	0.987	0.997
0.025	0.977	0.756	0.876	0.894	0.989	0.998
0.03	0.889	0.905	0.850	0.896	0.987	0.997
0.032	0.828	0.927	0.831	0.884	0.987	0.997
0.035	0.682	0.956	0.806	0.870	0.986	0.996
0.04	0.404	0.977	0.755	0.800	0.987	0.997

$T = 500$

VaR_P	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.02	0.999	0.145	0.914	0.800	0.998	0.892
0.025	0.987	0.618	0.881	0.879	0.999	0.890
0.03	0.891	0.891	0.848	0.888	0.998	0.887
0.0318	0.811	0.933	0.831	0.881	0.999	0.893
0.035	0.598	0.964	0.802	0.856	0.998	0.895
0.04	0.253	0.986	0.739	0.754	0.998	0.892

Table 6: Acceptance rate for the correct **ES** prediction and different **VaR** predictions, each based on 10^5 simulations.

$T = 250$

VaR_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.0436	1.000	0.000	1.000	0.314	0.622	0.791
0.0386	1.000	0.004	1.000	0.887	0.619	0.799
0.0336	1.000	0.668	1.000	0.997	0.616	0.793
0.0318	0.981	0.957	1.000	0.988	0.617	0.791
0.0286	0.997	0.977	1.000	0.994	0.617	0.799
0.0236	0.989	0.979	0.999	0.988	0.616	0.794

$T = 500$

VaR_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.0436	1.000	0.000	1.000	0.179	0.567	0.715
0.0386	1.000	0.000	1.000	0.879	0.560	0.706
0.0336	1.000	0.443	1.000	0.995	0.570	0.710
0.0318	1.000	0.973	1.000	0.996	0.566	0.710
0.0286	0.999	0.981	1.000	0.996	0.559	0.708
0.0236	0.995	0.983	0.999	0.989	0.564	0.705

Table 7: Acceptance rate for the correct **ES** prediction and different **VaR** predictions for reversed distributions, each based on 10^5 simulations.

$T = 250$

F	VaR_F	ES_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
T3	0.0318	0.0504	0.982	0.996	0.998	0.999	0.996	0.987
T3-Shift	0.0196	0.0382	0.930	0.414	0.823	0.820	0.516	0.339
T5	0.0257	0.0352	0.785	0.897	0.954	0.959	0.920	0.839
T5-Shift	0.0196	**0.0291**	**0.717**	**0.189**	**0.604**	**0.573**	**0.365**	**0.215**
T9	0.0226	**0.0288**	**0.406**	**0.553**	**0.692**	**0.704**	**0.655**	**0.478**
T10	0.0223	0.0282	0.373	0.489	0.640	0.650	0.602	0.422
T10-Shift	0.0196	**0.0255**	**0.366**	**0.089**	**0.319**	**0.285**	**0.233**	**0.111**
T15	0.0213	0.0264	0.215	0.311	0.420	0.425	0.427	0.255
T15-Shift	0.0196	0.0246	0.219	0.075	0.213	0.188	0.183	0.082
T20	0.0209	**0.0256**	**0.155**	**0.210**	**0.300**	**0.299**	**0.311**	**0.167**
T20-Shift	0.0196	0.0243	0.169	0.066	0.158	0.144	0.155	0.069

$T = 500$

F	VaR_F	ES_F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
T3	0.0318	0.0504	0.999	1.000	1.000	1.000	1.000	0.999
T3-Shift	0.0196	0.0382	0.994	0.562	0.943	0.948	0.608	0.467
T5	0.0257	0.0352	0.968	0.990	0.997	0.997	0.985	0.971
T5-Shift	0.0196	**0.0291**	**0.923**	**0.252**	**0.794**	**0.772**	**0.415**	**0.289**
T9	0.0226	**0.0288**	**0.664**	**0.785**	**0.894**	**0.898**	**0.772**	**0.669**
T10	0.0223	0.0282	0.598	0.709	0.839	0.840	0.710	0.583
T10-Shift	0.0196	**0.0255**	**0.549**	**0.115**	**0.435**	**0.398**	**0.219**	**0.134**
T15	0.0213	0.0264	0.353	0.448	0.602	0.598	0.464	0.337
T15-Shift	0.0196	0.0246	0.341	0.069	0.270	0.236	0.149	0.077
T20	0.0209	**0.0256**	**0.246**	**0.310**	**0.432**	**0.425**	**0.338**	**0.221**
T20-Shift	0.0196	0.0243	0.239	0.067	0.206	0.181	0.126	0.066

Table 8: **VaR** and **ES** for $\alpha = 0.025$ and the power of the backtests for different t distributions. The length of the time series is $T = 250$. The null hypothesis is that the returns are $N(0, 0.01)$ distributed. The calculation is based on each 10^5 simulations.

5.2.3 An example of deliberate deception

A dependency on **VaR** directly results in the ability for the bank to influence the power of a backtest and thus to reduce regulatory capital. Consider the following example: Suppose the daily returns of a banks' individual portfolio are iid and given by the t distribution with 3 degrees of freedom and a SD scale of 0.01. Of course one cannot know the actual distribution in reality. However, one can assume that the bank would accept this distribution through years of observation and various goodness of fit tests. The associated **VaR** and **ES** are 0.032 and 0.05. Now the bank wants to reduce the **ES** and models the returns with a simple distribution similar to the one above. The tail is modeled as a uniform distribution on the intervall $[-ES_F - (ES_F - VaR_F), -VaR_F]$ and with probability $1 - \alpha$, the returns are modeled as a uniform distribution on $[-VaR_P, VaR_P + \frac{2\alpha}{1-\alpha} ES_F]$. The values for the **VaR** and **ES** are deliberately changed. The **ES** forecast drops to 0.045, the **VaR** forecast rises to 0.04. Table 9 shows the acceptance rate for the backtests. In this example, this is the type 2 error and has a positive effect on the bank. Here one can see the weakness of AS_2. The strong sensitivity to the **VaR** opens up possibilities of manipulation, regardless of the length of the time series. After all, AS_3 and AS_4 reject with a rate of about 50%, which can be increased by a longer observation period.

$T = 250$

AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.130	0.978	0.502	0.516	0.966	0.992

$T = 500$

AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
0.040	0.970	0.420	0.350	0.992	0.946

Table 9: Acceptance rate for the manipulation example, each based on 10^5 simulations.

Particularly noteworthy is the result that AS_4 is also sensitive to manipulations. It is true that the expected value of the test statistics for a bank is shifting into a negative direction rather than the desired direction as predicted for ridge backtests. For the rejection of the null hypothesis, however, the behavior in the left tail of the asymptotic distribution is decisive. AS_4 does not completely eliminate the weaknesses of the **ES** in terms of its dependence on the **VaR**.

In this example, AS_1 convinces. Due to the overestimation of the **VaR**, only the empirical mean of a part of the tail is included in the test statistics. Regarding the weaknesses discovered so far, the question remains, whether one can produce examples in which it also shows a high acceptance rate. The

backtests of Corbetta and Peri are also not reliable here as there are not enough exceedances over the intentionally downsized **ES**.

5.3 Non-identical returns

Finally, the performance of the backtests will be worked out in a setting with daily-changing volatility. For this purpose, the returns are simulated with different GARCH models. First of all, however, the influence will be shown by a simple example to get an idea what happens if one deducts from the iid assumption. This effect should be included for further discussion.

5.3.1 Another example of underestimated SD

Consider the following two examples. In the first setup, the returns are modeled to be normal and iid with a mean 0 and SD 0.015. In fact, the SD is 1.2 times higher. The power of the backtests can be easily calculated as described in the previous chapters.

In the second setup, returns also follow a normal distribution but with a different SD for every day. Say that for every day, the predicted SD is given randomly, uniform distributed on the intervall $[0.01, 0.02]$. Again, in fact the SD is 1.2 times higher than the predicted one. Similar to the first setup, the power of the backtests is calculated.

Table 10 shows the power in both setups and for 250 and 500 days. Although the SD was always underestimated by the same factor in both cases, one can observe a better power for the backtests of Acerbi and Szekely if the returns are iid. For AS_2 and AS_4, however, this effect is significantly lower. With an observation period of 500 days, the power for both backtests and in both settings is above 90%. Since the backtests of Corbetta and Peri only count the number of **ES**-exceedances, their power is more stable in this example.

$T = 250$

	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
Set 1	0.161	0.768	0.690	0.765	0.779	0.627
Set 2	0.057	0.692	0.387	0.671	0.776	0.621

$T = 500$

	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
Set 1	0.271	0.949	0.903	0.940	0.897	0.828
Set 2	0.154	0.913	0.740	0.904	0.912	0.844

Table 10: Different power for iid distributed returns, each based on 10^5 simulations.

5.3.2 An example of autoregressive processes

The last analysis is guided by the models used by Du and Escanciano (2016) for the analysis of their proposed backtests. An AR(1)-GARCH(1,1) model is given and the power is calculated for different models with which the actual dependencies of the returns are simulated.

The null hypothesis is that the returns are given by the following model:

$$X_t = 0.05X_{t-1} + v_t, \quad v_t = \sigma_t\varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$
$$\sigma_t^2 = 0.05 + 0.1v_{t-1}^2 + 0.85\sigma_{t-1}^2$$

The alternative hypotheses to be tested are given as follows:

A_1 - A TAR model: $X_t = \alpha_t X_{t-1} + v_t, \quad v_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = 0.04 + 0.1v_{t-1}^2 + 0.89\sigma_{t-1}^2, \quad \alpha_t = 0.7(v_{t-1} \leq -2)$;

A_2 - A GARCH in mean model: $X_t = -0.5\sigma_t^2 + v_t, \quad v_t = \sigma_t\varepsilon_t, \quad \sigma_t = 0.01 + 0.29v_{t-1}^2 + 0.7\sigma_{t-1}^2$;

A_3 - An AR(1)-ARCH(2) model: $X_t = 0.05X_{t-1} + v_t, \quad v_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = 0.1 + 0.1v_{t-1}^2 + 0.8v_{t-2}^2$;

A_4 - An AR(1)-EGARCH(1,1) model: $X_t = 0.05X_{t-1}+v_t, \quad v_t = \sigma_t\varepsilon_t, \quad ln\sigma_t^2 = 0.01 + 0.9ln\sigma_{t-1}^2 + 0.3(|\varepsilon_{t-1}| - \sqrt{(}\frac{2}{\pi})) - 0.8\varepsilon_{t-1}$;

A_5 - An AR(1)-GARCH(1,1) model with mixed normal innovations: $X_t = 0.05X_{t-1}+v_t, \quad v_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = 0.05+0.1v_t^2+0.85\sigma_{t-1}^2 \quad \varepsilon_t \sim [0.6N(1, \sqrt{(}2))+ 0.4N(-1.5, \sqrt{(}0.75))]/\sqrt{(}3)$.

In figure 10, one path of each of the null hypothesis and one path of each alternative hypothesis are shown as examples. For A_1 and A_2, the conditional mean is incorrectly specified as well as the conditional variance. In the TAR model particularly negative returns still have a higher impact on future returns. This leads to a left skewed distribution and a fat left tail. The situation is similar to A_2. A high rejection rate is expected. For A_3 and A_4, the conditional mean is correct but the conditional variance is not. Regarding A_3, the dependency of the last two innovations makes the distribution of returns somewhat more "stable". Both, the **VaR** and the **ES** are overestimated under H_0. The EGARCH model leads to fatter tails, one expects a high rejection rate. For A_5, the model and the parameters are the same, only the distribution of the innovations is different. In this case, both the **VaR** and the **ES** are overestimated. A rejection rate close to zero is expected.

The results are shown in table 11. If one models paths under H_0 and the forecast is done under H_0 as well, all backtests are close to the desired significance level of 0.05. An exception is CP_1 for a time series of 250 days. For A_3

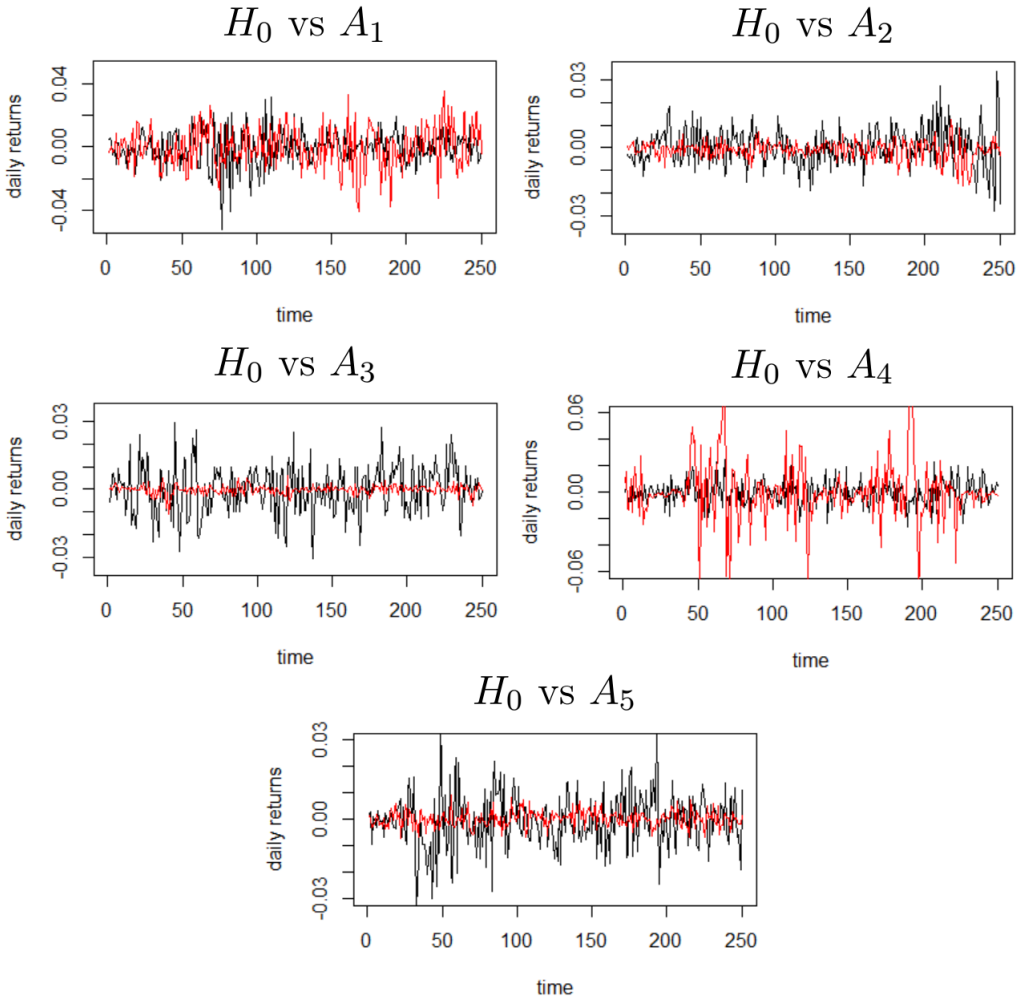

Figure 10: Examples of daily returns for a comparison between the null hypothesis and the presented alternatives. The path resulting from the H_0 model is black, the path resulting from the alternatives is red. Note that the models above are scaled with 0.01, so that the daily returns appear in a realistic scale.

$T = 250$

F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
H_0	0.052	0.044	0.070	0.032	0.140	0.050
A_1	0.422	0.594	0.818	0.680	0.768	0.632
A_2	0.274	0.080	0.332	0.092	0.126	0.092
A_3	0.850	0.002	0.024	0.006	0.002	0.002
A_4	0.961	0.967	0.999	0.997	0.990	0.964
A_5	0.964	0.000	0.000	0.000	0.000	0.000

$T = 500$

F	AS_1	AS_2	AS_3	AS_4	CP_1	CP_2
H_0	0.056	0.044	0.024	0.032	0.058	0.028
A_1	0.830	0.902	0.940	0.954	0.896	0.854
A_2	0.258	0.048	0.541	0.122	0.128	0.231
A_3	0.721	0.000	0.022	0.001	0.000	0.720
A_4	0.995	1.000	1.000	1.000	1.000	1.000
A_5	0.941	0.000	0.000	0.000	0.000	0.998

Table 11: Rejection rates for the null hypothesis and the alternatives, each based on 10^5 simulations.

and A_5, one observes low rejection rates as expected. But in both cases, the underestimation of the **VaR** leads to a high rejection rate for AS_1 and CP_2 for 500 days. The problems with this backtests have already been discussed above. For A_4, one observes a high rejection close to 1 for all backtests. Regarding A_2, the rejection rate is the highest for AS_3. Note that large shocks then continue to have a high effect on future returns, but only rarely occur. However, as the rare events and their consequences should be well predicted, this result is remarkable. Regarding the TAR model, all backtests perform well for $T = 500$. For a backtest period of 250 days, AS_3 has advantages again.

In the previous example of underestimated SD with different variances, the performance of AS_3 is not better than the others. Therefore, it is difficult to make a general statement here, it depends on the concrete situation. Moreover, the results cannot be easily compared with those of Du and Escanciano. Although it seems as if the backtests of Acerbi and Szekely promise a higher power here, this example does not include that the variables for each model have to be measured first. Du and Escanciano tested the impact of this as well.

6 Discussion and Conclusion

In this thesis different backtests which should validate **ES** predictions were discussed. First, there was a brief overview of desired characteristics for risk measures, then the **VaR** and the **ES** were presented and compared. The **VaR** makes no statement about the amount of losses if it is exceeded. In addition, it violates a mathematical property that is called subadditivity. The **ES** fulfills these requirements and is thus a general coherent risk measure.

The decision of the BCBS to replace the **VaR** at the 1% level with the the **ES** at the 2.5% level has triggered a broad academic and practical debate on the trade-off between coherence and elicitability. Due to the fact that the **ES** is not elicitable, it was initially assumed not to be backtestable. Nevertheless, there were a number of different proposals of **ES** backtests. In this thesis the backtests of Acerbi and Szekely (2014, 2017) and Corbetta and Peri (2016) were worked out. The work of Acerbi and Szekely (2017) confirms that the **ES** is not backtestable due to its **VaR** dependency, at least not according to the proposed definition. At the same time, however, they present a backtest that should minimize this dependency. All presented backtests were checked in various settings for significance and power.

The backtests of Corbetta and Peri proved to be unsuitable. For CP_2 this is mainly due to the analytical result with its' very low convergence rate. The backtests are heavily dependent on the differences in the structure of the tail distributions of the assumption and the actual one, as seen in chapter 5. One can construct further examples in which a correct **ES** prediction is often exceeded or an underestimated **ES** is rarely exceeded. The number of exceedances is not an adequate measure for assessing the quality of a prediction. Due to the strong dependence on the type of distribution, the proposal of the model selection should also be disregarded, because under- and overestimation is weighted by the coverage level.

The performance of the backtests of Acerbi and Szekely varies greatly in different settings. This is due to the dependence of the **ES** on the **VaR** in particular. The newest proposal by Acerbi and Szekely minimizes this dependency. Even if it fulfils the familiar property of a ridge backtest, there is still a dependence on the **VaR** prediction that leads in the wrong direction with regard to the rejection rate of a wrong **ES** prediction. Nevertheless, this dependency remains the smallest, as is shown for the example in chapter 5. This also minimizes the ability of banks to minimize capital appreciation through clever forecasting constructions. It also fulfills the property of sharpness which means that the magnitude of the test statistics can be used to draw a conclusion about the distance between the **ES** prediction and the actual value. Regarding the backtests presented here, it is to be preferred.

Replacement of the VaR - The right decison?

Lazar and Zhang (2017) study the model risk of **ES** and compare it to the model risk of **VaR**. They divide model risk into identification error, estimation error and specification error. The impact of the last two errors is calculated in a theoretical as well as in an empirical framework. Since the true model is unknown in reality the sources of risk cannot be divided into its' pieces. They propose a backtesting-based correction methodology of the **ES**. The corrected **ES** is given by

$$ES^{cor}(\hat{\theta}, \alpha, C) = ES(\hat{\theta}, \alpha) + C,$$

where $\hat{\theta}$ are the estimated parameter values of the chosen model and C is the minimum addition such that the corrected **ES** passes a certain backtest. This is an extension of the investigation by Boucher et. al. (2014), who did that for the **VaR**, namely the simple binomial test. This approach is performed for the unconditional and conditional backtest of Du Escanciano (2016) and AS_2. Different forms of investments like equity, bonds, commodity and FX are tested. Based on the correction parameter C, one can choose between different models. The main result is that the mean of the additon on the **ES** is much lower than that for the **VaR**. This fits to the investigations of Acerbi and Szekely in 2014. AS_2 already shows a generally higher power than the binomial test in their examples. These investigations support the decision of the BCBS.

Another traffic light approach

As part of the academic and practical debate on the **ES**, the question arises to what extend a similarly simple backtesting framework for the **ES** can be developed as for the **VaR**. Again, the question arises to what extend the **VaR** should be backtested as well. Acerbi and Szekely (2014) argue, that the critical values for AS_2 display remarkable stability across different distributions. They say that the critical value of -0.7 and -1.8 is a reliable border for a p-value smaller than 0.05 and 0.001. Based on the examples of this thesis, one can confirm the critical values. Note that these values belong to a backtesting period of 250 days. For 500 days, values of -0.5 and -0.1 would be suitable. Based on these values, one could establish a traffic light approach similar to the previous one.

According to the investigations in the context of this thesis, no general statement about the critical values can be made for AS_4. The critical values for a given confidence level α differ for different distributions with fat tails or skewness. Regarding this test, simulations are unavoidable.

Moldenhauer and Pitera (2017) propose another backtest framework which

ends up in a traffic light approach. Let $\rho_{1,t}$ be the **VaR** forecast for every day t and let X_t be the returns. They call $y_t = X_t + \rho_{2,t}$ the "secured position", where ρ_2 is the predicted **ES**. Let $V@R_T^\alpha(x) = -x_{\lfloor T\alpha \rfloor}$ be the empirical estimator of the **VaR** for a sample X and $\hat{ES}_T^\alpha(X)$ the empirical estimator of the **ES** for a significance level α:

$$\hat{ES}_T^\alpha(X) = -\frac{\sum_{t=1}^T X_t(X_t - V@R_T^\alpha(X) < 0)}{\sum_{t=1}^T 1(X_t - V@R_T^\alpha(X) < 0)}.$$

Based on that, they propose the following traffic light approach: For $T = 250$ and the "secured vector" Y, the prediction is in

- the green zone if $\hat{ES}_T^\alpha(Y) \leq 0$ for $\alpha = 0.04$;
- the yellow zone if it does not fall into the green zone but for $\alpha = 0.1$ one gets $\hat{ES}_T^\alpha(Y) \leq 0$;
- the red zone if it does not fall into one of the first zones.

They perform their approach for market data as well as for simulated data with different models and examine the consistency of the **VaR** backtest and the traffic light approach based on the magnitude of AS_2. The results of both approaches are in a very similar area. Based on the results of this thesis - especially with regard to the **VaR** sensitivity - it would be necessary to recompare the approach of Moldenhauer and Pitera with AS_4 to decide which of the two approaches would be the most reliable. In addition, the analysis and comparison should be done again with a backtesting period of 500 days, because it can increase the power as one saw in the examples.

Not all open questions regarding the decision of the BCBS have been answered within thesis. However, based on the findings, one can support and recommend a traffic light approach based on AS_4. Note that the **VaR** forecast also needs to be roughly validated for this test. So the question remains how to bring these two risk measures together in one backtesting framework. Although AS_3 in the examples here often shows a similarly strong performance, it is, in contrast to AS_4, not recommended without further investigations. Due to the complex structure, it is very difficult to make general statements, for example on the sensitivity against the **VaR**. In addition, not the entire tail distribution should be validated. Since the number of **ES** exceedances depends on the actual distribution, the main object of **ES** backtesting should be the magnitude of **VaR** exceedances as it is for AS_4. To show that the risk was modeled adequately, a good risk management should not only perform the backtest, but also present further results such as goodness of fit tests. This further analysis is also requested in the FRTB. The size of this buffer in a new traffic light approach needs to be investigated in more complex frameworks with different portfolios. If it is sufficiently large, one can certainly

support the decision of the BCBS. It is not easy but it is possible to validate **ES** predictions.

References

[1] Acerbi, C., Nordio, C., Sirtori, C. (2001): Expected shortfall as a tool for financial risk management. Working Paper.

[2] Acerbi, C., Tasche, D. (2002): On the coherence of Expected Shortfall. Journal of Banking and Finance (26), pp. 1487-1503.

[3] Acerbi, C., Szekely, B. (2014): Backtesting Expected Shortfall. MSCI Inc.

[4] Acerbi, C., Szekely, B. (2017): General properties of backtestable statistics. MSCI Inc.

[5] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1997): Thinking coherently. Risk (10), pp. 68–71.

[6] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999): Coherent measures of risk. Mathematical Finance (9), pp. 203–228.

[7] Azzalini, A. (1985): A class of distributions which includes the normal ones. Scandinavian Journal of Statistics (12), pp. 171-178.

[8] Basel Committee on Banking Supervision (1996): Supervisory Framework for the use of "Backtesting" in conjunction with the internal models approach to market risk capital requirements.

[9] Basel Committee on Banking Supervision (2012): Fundamental review of the trading book - consultative document.

[10] Basel Committee on Banking Supervision (2016): Minimum capital requirements for market risk.

[11] Bernadi, M. (2013): Risk measures for skew normal mixtures. Statistics and Probability Letters (83), pp.1819-1824.

[12] Boucher, C. M., Danielsson, J., Kouontchou, P. S., Maillet, B. B. (2014): Risk models-at-risk. Journal of Banking and Finance (44), pp. 72-92.

[13] Broda, S. A., Paolella, M. S. (2011): Expected shortfall for distributions in finance. In: Statistical Tools for Finance and Insurance, pp. 57-99, Heidelberg, Berlin.

[14] Carver, L. (2011): Mooted var substitute cannot be back-tested, says top quant. Risk magazine.

[15] Corbetta, J., Peri, I. (2016): A New Approach to Backtesting and Risk Model Selection. Ecole des Ponts ParisTech, University of Greenwich.

[16] Christoffersen, P. (1998): Evaluating Interval Forecasts. International Economic Review (39), pp. 841-862.

[17] Du, Z., Escanciano, J. C. (2016): Backtesting expected shortfall - accounting for tail risk. Management Science.

[18] Embrechts, P., Wang, R. (2015): Seven Proofs for the Subadditivity of Expected Shortfall. Dependence Modeling (3), pp. 126-240.

[19] Emmer, S., Kratz, M., Tasche, D. (2013): What Is the Best Risk Measure in Practice? A Comparison of Standard Measures. Journal of Risk 18(2), pp. 31-60.

[20] Fernandez, C., Steel, M. F. J. (1998): On Bayesian modeling of fat tails and skewness, J. Am. Statist. Assoc. (93), pp. 359–371.

[21] Fissler, T., Ziegel, J. F., Gneiting, T. (2015): Expected Shortfall is jointly elicitable with Value at Risk – Implications for backtesting. Risk magazine.

[22] Gneiting, T. (2011): Making and Evaluating Point Forecasts. Journal of the American Statistical Association (106), pp. 746–762.

[23] Kupiec, P.H. (1995): Techniques for Verifying the Accuracy of Risk Measurement Models. The Journal of Derivates (3), 73-84.

[24] Lambert, N., Pennock, D.M., Shoham, Y. (2008): Eliciting Properties of Probability Distributions, Proceedings of the 9th ACM Conference on Electronic Commerce, EC 08.

[25] Lazar, E., Zhang, N. (2017): Model Risk of Expected Shortfall. Discussion Paper.

[26] Löser, R., Wied, D., Ziggel, D. (2016): New Backtests for Unconditional Coverage of the Expected Shortfall. Discussion Paper, SFB 823.

[27] 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), pp. 271-300.

[28] Moldenhauer, F., Pitera, M. (2017): Backtesting Expected Shortfall: Is it really that hard? Preliminary Draft.

[29] Nolde, N., Ziegel, J.F. (2017): Elicitability and backtesting: Perspectives for banking regulation. Annals of Applied Statistics (11), pp. 1833-1874.

[30] Pfaff, B. (2016): Financial Risk Modelling and Portfolio Optimization with R (2nd edition). John Wiley & Sons Ltd. Chichester, United Kingdom.

[31] Rappoport, P. (1993): A new Approach, Average Shortfall. Technical report, J.P. Morgan.

[32] Righi, M. B., Ceretta, P. S. (2013): Individual and flexible expected shortfall backtesting. Journal of Risk Model Validation, 7(3), pp. 3–20.

[33] Rockafeller, R.T., Uryasev, S. (2002): Conditional Value-at-Risk for general loss distributions. Journal of Banking and Finance, 26 (7), pp. 1443-1471.

[34] Sheikh, A.Z., Qiao, H. (2010): Non-Normality of Market Returns: A Framework for Asset Allocation Decision Making. Journal of Alternative Investments, Vol. 12 (3), pp. 8-35.

[35] Wimmerstedt, L. (2015): Backtesting Expected Shortfall: the design and implementation of different backtests (Master Thesis). KTH, School of Engineering Sciences (SCI), Mathematics (Dept.), Mathematical Statistics, Stockholm.