Backtesting Expected Shortfall: the design and implementation of different backtests

Lisa Wimmerstedt

Abstract

In recent years, the question of whether Expected Shortfall is possible to backtest has been a hot topic after the findings of Gneiting in 2011 that Expected Shortfall lacks a mathematical property called elicitability. However, new research has indicated that backtesting of Expected Shortfall is in fact possible and that it does not have to be very difficult. The purpose of this thesis is to show that Expected Shortfall is in fact backtestable by providing six different examples of how a backtest could be designed without exploiting the property of elicitability. The different approaches are tested and their performances are compared against each other. The material can be seen as guidance on how to think in the initial steps of the implementation of an Expected Shortfall backtest in practice.

Keywords: Expected Shortfall, Backtests, Value-at-Risk, Elicitability

Acknowledgements

I would like to express my gratitude to my supervisor and associate professor Filip Lindskog at the Royal Institute of Technology for his contribution to the Master's Thesis through constructive discussions, support and encouragement. I would also like to thank Bengt Pramborg, Leonhard Skoog and Gustaf Jarder at Market & Counterparty Risk at Swedbank for their valuable comments.

Stockholm, August 2015

Lisa Wimmerstedt

Contents

1 Introduction **1**

2 Background **5**
2.1 The mathematical properties of risk measures 5
2.2 Value-at-Risk 7
2.3 Expected Shortfall 8
2.4 Parametric values of VaR and Expected Shortfall 8
2.5 Elicitability 11
2.6 Backtesting VaR 16
2.7 Conclusion 19

3 The design of different Expected Shortfall backtests **21**
3.1 Wong's saddlepoint technique 22
3.2 Righi and Ceretta's truncated distribution 28
3.3 Emmer, Kratz and Tasche's quantile approximation 30
3.4 Acerbi and Szekely's unparametric models 33
3.5 Conclusion 37

4 The ability to accept true Expected Shortfall predictions **39**
4.1 Methodology 39
4.2 Results 40
4.3 Conclusion 43

5 The ability to reject false Expected Shortfall predictions **45**
5.1 Methodology 46
5.2 The overall ability to reject 50
5.3 The ability to reject with a fixed number of exceedances . . . 52
5.4 Conclusion 54

6 Implementing the methods in practice **55**
6.1 Choosing a model internally 56
6.2 The general method 58
6.3 Conclusion 65

7 Conclusion **67**
7.1 Finding methods without elicitability 68
7.2 Performance 68
7.3 Implementation 70
7.4 The difficulties in designing a simple backtest 70
7.5 The backtestability of Expected Shortfall 72

Chapter 1

Introduction

Following recent financial crises and the increased complexity of financial markets, quantifying risk has become a more important matter. Supervisors increase the control of banks to make sure they have enough capital to survive in bad markets. While risk is associated with probabilities about the future, one usually uses risk measures to estimate the total risk exposure. A risk measure summarises the total risk of an entity into one single number. While this is beneficial in many respects, it opens up a debate regarding what risk measures that are appropriate to use and how one can test their performance. Risk measures are used for internal control as well as in the supervision of banks by the Basel Committee of Banking Supervision.

Value-at-Risk (VaR) is the most frequently used risk measure. VaR measures a threshold loss over a time period that will not be exceeded with a given level of confidence. If we estimate the 99 % 1-day VaR of a bank to be 10 million then we can say that we are 99 % confident that within the next day, the bank will not lose more than 10 millions. One of the main reasons for its popularity as a risk measure is that the concept is easy to understand without having any deeper knowledge about risk. Furthermore, VaR is very easy to validate or backtest in the sense that after having experienced a number of losses, it is possible to go back and compare the predicted risk with the actual risk. If we claim that a bank has a 99 % 1-day VaR of 10 million then we can expect that one day out of 100, the bank will have losses exceeding this value. If we find that in the past 100 days, the bank has had 20 losses exceeding 10 million then something is most likely wrong with the VaR estimation.

In the current regulatory framework by the Basel Committee, banks are required to report its 99 % VaR on a daily basis. The VaR numbers reported specifies the amount of capital required to maintain this level of risk in the bank, also called the capital charge. To ensure that VaR estimates

made by banks are reported correctly, the numbers are backtested against the realised losses counting the number of exceedances, that is the number of days a bank's losses exceeded VaR during the past year. If the number of exceedances are too many then the bank will be punished with higher capital charge.

The main criticism of VaR has been that it fails to capture tail risk. This means that VaR specifies the value that the loss will be exceeding in a bad day, but it does not specify by how much the loss will be exceeding VaR. In other words, the measure does not take into account what happens beyond the threshold level. Figure 1.1 shows two different distributions with the same 95 % VaR that can illustrate this issue. The two probability distributions have the same 95 % VaR of 1.65 but should not be seen as equally risky since the losses, defined by the left tail of the distribution, are different for the two different distributions.

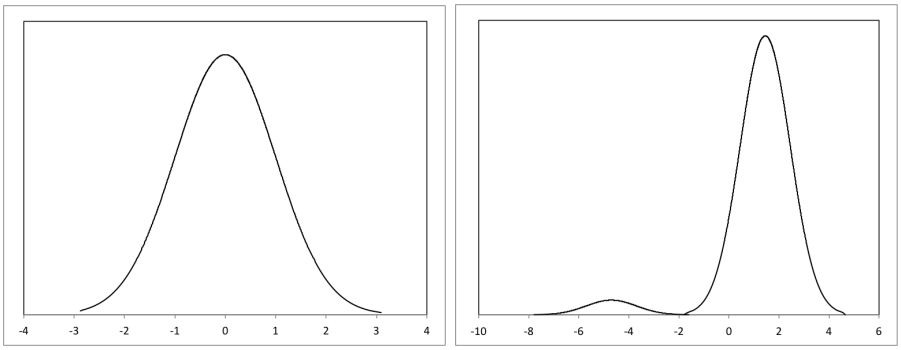

Figure 1.1: Shows two return distributions with the same 95 % VaR of 1.65. We see that even though VaR is the same, the right plot is more risky.

Furthermore, VaR lacks a mathematical property called subadditivity. In short, this means that VaR for two combined portfolios can be larger than VaR for the sum of the two portfolios independently. This implies that diversification could increase risk, a contradiction to standard beliefs in finance. Therefore, it is not a desired property for a risk measure. The failure to capture tail risk and the lack of subadditivity has lead to the increased adoption of another risk measure called Expected Shortfall. The risk measure was presented as a response to the criticism of VaR. In short, Expected Shortfall measures the expected loss in the tail of the distribution. That is, the expected loss on the days when the loss exceeds VaR. Since the risk measure takes into account what happens in the tail of the distribution, it captures tail risk well. Furthermore, Expected Shortfall is subadditive solving also this issue associated with VaR. In figure 1.1, the left graph would have an Expected Shortfall of 2.1 while the right graph would have an Ex-

pected Shortfall of 4.7. Hence, in contrary to VaR, this risk measure would capture the fact that the right scenario is much more risky.

While VaR is still the most important risk measure today, a change is expected. Following the criticism of VaR, supervisors have proposed to replace the 99 % VaR with a 97.5 % Expected Shortfall as the official risk measure in the calculations of capital requirements. The purpose behind this is that tail risk matters and should therefore be accounted for. The discussion around this can be found in the Fundamental Review of the Trading Book by The Basel Committee (2013).

While Expected Shortfall solves some of the issues related to VaR, there is one drawback that prevents a full transition from VaR to Expected Shortfall. As explained above, it is very straightforward to backtest VaR by counting the number of exceedances. However, when it comes to Expected Shortfall, there are several questions outstanding on how the risk measure should be backtested. In 2011, Gneiting published a paper showing that Expected Shortfall lacked a mathematical property called elicitability that VaR had and that this could be necessary for the backtestability of the risk measure. Following his findings, many people were convinced that it was not possible to backtest Expected Shortfall at all. If so, this would imply that if supervisors change their main risk measure from VaR to Expected Shortfall then they lose the possibility to evaluate the risk reported and punish banks that report too low risk. When The Basel Committee (2013) proposed to replace VaR with Expected Shortfall they concluded that backtesting would still have to be done on VaR even though the capital would be based on Expected Shortfall estimates. After this proposal from the Basel Committee, new research has indicated that backtesting Expected Shortfall may in fact be possible and that it does not have to be very difficult.

The purpose of this thesis will be to show that it is possible to backtest Expected Shortfall and describe in detail how this can be done. We will do this by presenting six methods that can be used for backtesting Expected Shortfall that do not exploit the property of elicitability. We will show that the methods work in practice by doing controlled simulations from known distributions and investigate in what scenarios the methods accept true Expected Shortfall predictions and when they reject false predictions. We will show that all methods can be used as a backtest of Expected Shortfall but that some methods perform better than others. We will also advise on which methods that are appropriate to implement in a bank both in terms of performance and complexity.

The material presented here can be seen as guidance on how to think in the initial process of the implementation of an Expected Shortfall backtest.

If the proposed transition from VaR to Expected Shortfall will take place within the next years then all banks will be faced with a situation where backtesting Expected Shortfall will be necessary for internal validation and perhaps eventually also for regulatory control. The question of how a backtest of Expected Shortfall can be designed is therefore of great interest.

We will start the next chapter by proper definitions of risk measures, their properties and how backtests of VaR are done today before we move on to the next chapters where we go into the potential methods and their performance.

Chapter 2

Background

This chapter will give an overview of the mathematical properties of risk measures, give formal definitions of VaR and Expected Shortfall as well as discuss their mathematical properties. Furthermore, the concept of elicitability will be explained in detail and we will describe how the backtesting of VaR is done today.

2.1 The mathematical properties of risk measures

There are many ways in which one could define risk in just one number. Standard deviation is perhaps the most fundamental measure that could be used for quantifying risk. However, risk is mainly concerned with losses and standard deviation measures deviations both up and down. There are several properties we wish to have in a good risk measure. One fundamental criterion is that we want to be able to interpret the risk measure as the buffer capital needed to maintain that level of risk. Hence, risk should be denominated in a monetary unit. Following will be a short description of the six fundamental properties that we look for in a good risk measure. The properties are important for understanding the academic debate about the differences between VaR and Expected Shortfall. The properties are cited from Hult et al. (2012). We let X be a portfolio value today, R_0 be the risk-free return and $\rho(X)$ be our risk measure of portfolio X. Furthermore we let c be an amount held in cash.

- **Translation invariance**.

$$\rho(X + cR_0) = \rho(X) - c$$

This means that having cash reduces risk by the same amount. This follows automatically from our definition of a risk measure as the buffer capital needed to maintain a certain level of risk. Having cash equal to

the risk held in a portfolio $c = \rho(X)$ means that the total risk equals zero.

- **Monotonicity**.

$$X_2 \leq X_1, \quad \text{implies that} \quad \rho(X_1) \leq \rho(X_2)$$

This means that if we know that the value of one portfolio will always be larger than the value of another portfolio, then the portfolio with higher guaranteed value will always be less risky.

- **Convexity**.

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

In essence, this means that diversification and investing in different assets should never increase the risk but it may decrease it.

- **Normalization.** This means that having no position imposes no risk. Hence, we have that $\rho(0) = 0$.

- **Positive homogeneity**.

$$\rho(\lambda X) = \lambda\rho(X) \quad \text{for all} \quad \lambda \geq 0$$

In other words, to double the capital means to double the risk.

- **Subadditivity**.

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

Two combined portfolios should never be more risky than the sum of the risk of the two portfolios separately.

When a risk measure satisfy the properties of translation invariance, monotonicity, positive homogeneity and subadditivity it is called a *coherent measure of risk*. Normalization is usually not a problem when defining a risk measure. Furthermore, convexity and positive homogeneity together imply subadditivity.

2.2 Value-at-Risk

We are now going to give a formal definition of VaR. We let V_0 be a portfolio value today and V_1 be the portfolio value one day from now. Furthermore, we let R_0 be the percentage return on a risk-free asset. When people talk about VaR, the most frequent use is that of a 99 % VaR in the sense that it is the loss that will not be exceeded with 99 % confidence. However, in mathematical terms we deal with the loss in a small part of the distribution. Hence, in mathematical terms a 99 % VaR is referred to as the 1 % worst loss of the return distribution. We will therefore denote a 99 % VaR with $\text{VaR}_{1\%}$ and say that a 99 % VaR has $\alpha = 0.01$. We define VaR for a portfolio with net gain $X = V_1 - V_0 R_0$ at a level α as

$$\text{VaR}_\alpha(X) = \min\{m : P(L \leq m) \geq 1 - \alpha\}, \tag{2.1}$$

where L is the discounted portfolio loss $L = -X/R_0$. We assume a future profit-and-loss (P&L) distribution function P. If P is continuous and strictly increasing, we can also define VaR as

$$\text{VaR}_\alpha(X) = -P^{-1}(\alpha). \tag{2.2}$$

While VaR satisfy the properties of translation invariance, monotonicity and positive homogeneity, it is not subadditive. Hence, VaR is not a coherent measure of risk. It is straightforward to show that VaR is not subadditive by Example 2.1, taken from Acerbi et al. (2001).

Example 2.1 *We assume that we have two bonds X_1 and X_2. The bonds have default probability 3 % with recovery rate 70 % and default probability 2 % with a recovery rate of 90 %. The bonds cannot both default. This could be the case if they are corporate bonds competing in the same market so one will benefit from the other's default. The numbers are shown in table 2.1.*

Probability	X_1	X_2	$X_1 + X_2$
3 %	70	100	170
3 %	100	70	170
2 %	90	100	190
2 %	100	90	190
90 %	100	100	200

Table 2.1: The table illustrates an example showing that VaR is not subadditive.

We can calculate the initial value of each bond as 98.9 and the value of the bonds together as 197.8. We calculate the 95 % VaR for each bond by ordering the returns and take a look at the one with a cumulative probability

of 5 % (in this case 90). Hence, for each bond the 95 % VaR is 8.9. The 95 % VaR for the two bonds together is 27.8. Hence, VaR for the two bonds together is larger than VaR of the sum of the two bonds independently. This shows that VaR is not subadditive.

2.3 Expected Shortfall

The idea of Expected Shortfall was first introduced in Rappoport (1993). Artzner et al. (1997, 1999) formally developed the concept. We define Expected Shortfall as

$$\mathrm{ES}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)du. \tag{2.3}$$

Expected Shortfall inherits the properties of translation invariance, monotonicity and positive homogeneity from VaR. Furthermore, is it also subadditive. Hence, Expected Shortfall is a coherent measure of risk.

2.4 Parametric values of VaR and Expected Shortfall

We will now show how VaR and Expected Shortfall can be calculated for some standard distributions. We will do this for the normal distribution with mean 0 and standard deviation σ and the location-scale Student's t distribution with degrees of freedom ν, location parameter 0 and scale parameter σ. We start by proper definitions of the distributions before we show how the risk measures can be calculated.

2.4.1 The normal distribution

We start by defining a random variable X that follows a normal distribution with mean 0 and standard deviation σ. We can write this as

$$X = \sigma Y \tag{2.4}$$

where Y is a standard normal variable. We can write this directly as $Y \sim N(0,1)$ and $X \sim N(0,\sigma)$.

2.4.2 Student's t distribution

We now assume that X follows a location-scale Student's t distribution. This means that we can write the random variable X as a function of a

random variable T that follows a standard Student's t distribution with ν degrees of freedom.

$$X = \mu + \sigma T.$$

We say that the distribution has location parameter μ and scale parameter σ. σ does not denote the standard deviation of X but is called the scaling. Instead, we have that

$$\begin{aligned} \mathbb{E}(X) &= \mu \qquad \text{for } \nu > 1, \\ \text{Var}(X) &= \frac{\nu}{\nu - 2}\sigma^2 \qquad \text{for } \nu > 2. \end{aligned}$$

We will write this as $X \sim t_\nu(\mu, \sigma)$. In the analysis we will always assume that $\mu = 0$. This means that we get

$$X = \sigma T. \tag{2.5}$$

The probability density of X is given by

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

2.4.3 VaR

We now want to find the analytical expression of VaR and Expected Shortfall for the distributions given above. Since both the normal distribution and the Student's t distribution have continuous and increasing probability functions, we have by definition (2.2) that VaR is

$$\text{VaR}_\alpha(X) = -F^{-1}(\alpha). \tag{2.6}$$

We start by assuming that X follows a standard normal distribution according to equation (2.4). We can then calculate VaR as

$$\text{VaR}_\alpha(X) = -\sigma\Phi^{-1}(\alpha) = \sigma\Phi^{-1}(1 - \alpha), \tag{2.7}$$

where $\Phi(x)$ is the standard normal cumulative probability function.

We now assume that X is Student's t distributed with parameters ν and σ according to equation (2.5). We can then write VaR as

$$\text{VaR}_\alpha(X) = -\sigma t_\nu^{-1}(\alpha) = \sigma t_\nu^{-1}(1 - \alpha), \tag{2.8}$$

where $t_\nu(x)$ is the standard Student's t cumulative probability function.

2.4.4 Expected Shortfall

We now move on to Expected Shortfall. By definition (2.3) we have that

$$\mathrm{ES}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)du \tag{2.9}$$

We start by assuming that X is a standard normal variable according to equation (2.4). This means that we know $\mathrm{VaR}_\alpha(X) = \sigma\Phi^{-1}(1-\alpha)$. We can write this as

$$\begin{aligned}\mathrm{ES}_\alpha(X) &= \frac{\sigma}{\alpha}\int_0^\alpha \Phi^{-1}(1-u)du \\ &= \frac{\sigma}{\alpha}\int_{1-\alpha}^1 \Phi^{-1}(u)du\end{aligned}$$

We do a change of variables and set $q = \Phi^{-1}(u)$. We get

$$\begin{aligned}\mathrm{ES}_\alpha(X) &= \frac{\sigma}{\alpha}\int_{\Phi^{-1}(1-\alpha)}^\infty q\phi(q)dq \\ &= \frac{\sigma}{\alpha}\int_{\Phi^{-1}(1-\alpha)}^\infty q\frac{1}{\sqrt{2\pi}}\exp^{-q^2/2}dq \\ &= -\frac{\sigma}{\alpha}\Big[\frac{1}{\sqrt{2\pi}}\exp^{-q^2/2}\Big]_{\Phi^{-1}(1-\alpha)}^\infty \\ &= -\frac{\sigma}{\alpha}\Big[\frac{1}{\sqrt{2\pi}}\exp^{-q^2/2}\Big]_{\Phi^{-1}(1-\alpha)}^\infty \\ &= \sigma\frac{\phi(\Phi^{-1}(1-\alpha))}{\alpha},\end{aligned}$$

where, as above, $\phi(x)$ is the standard normal density function and $\Phi(x)$ is the standard normal cumulative distribution function. We can do the same calculation assuming X follows a Student's t distribution with parameters ν and σ according to equation (2.5). The calculations can be found in McNeil et al. (2015). Expected Shortfall can be written as

$$\mathrm{ES}_\alpha(X) = \sigma\frac{g_\nu(t_\nu^{-1}(1-\alpha))}{\alpha}\frac{\nu+(t_\nu^{-1}(\alpha))^2}{\nu-1}, \tag{2.10}$$

where $t_\nu(x)$ is the cumulative probability function of the standard Student's t distribution and $g_\nu(x)$ is the probability density function of the same distribution.

Table 2.2 shows Expected Shortfall and VaR for different levels using different parametric assumptions. All estimates are with $\sigma = 1$.

	VaR			Expected Shortfall		
	95 %	97.5 %	99%	95 %	97.5 %	99%
$t_3(0,1)$	2.35	3.18	4.54	3.87	5.04	7.00
$t_6(0,1)$	1.94	2.45	3.14	2.71	3.26	4.03
$t_9(0,1)$	1.83	2.26	2.82	2.45	2.88	3.46
$t_{12}(0,1)$	1.78	2.18	2.68	2.34	2.73	3.22
$t_{15}(0,1)$	1.75	2.13	2.60	2.28	2.64	3.10
$N(0,1)$	1.64	1.96	2.33	2.06	2.34	2.67

Table 2.2: The table shows some values of VaR and Expected Shortfall for some underlying distributions where $N(0,1)$ denotes the standard normal distribution and $t_\nu(0,1)$ denotes the Student's t distribution with degrees of freedom ν, $\mu = 0$ and $\sigma = 1$.

Table 2.2 gives some guidance on how Expected Shortfall and VaR correspond to each other for different distributional assumptions. It is interesting to note that for the normal distribution, 99 % VaR and 97.5 % Expected Shortfall are almost the same. This means that if returns are normally distributed then a transition from a 99 % VaR to a 97.5 % Expected Shortfall would not increase capital charges. However, if returns are Student's t distributed then the proposed transition would increase capital requirements.

2.5 Elicitability

The concept of elicitability was introduced by Osband (1985) and further developed by Lambert et al. (2008). This mathematical property is important for the evaluation of forecasting performance. In general, a law invariant risk measure takes a probability distribution and transforms it into a single-valued point forecast. Hence, backtesting a risk measure is the same as evaluating forecasting performance. This means that in order to backtest a risk measure we must also look at mathematical properties that are important for evaluating forecasts. In 2011, Gneiting showed that Expected Shortfall lacks the mathematical property called elicitability. This section will define elicitabiliy and explain why it is a problem that Expected Shortfall is not elicitable.

2.5.1 Definition

In the evaluation of forecasts we want to compare forecasted estimates with observed data. We say that we have forecasts that we call x and verifying observations that we call y. We now want to compare the forecasts to the verifying observations to see if the forecasts were any good. To do this, we introduce a scoring function $S(x, y)$ that we want to use to evaluate the

performance of x given some values on y. Examples of scoring functions are squared errors where $S(x, y) = (x - y)^2$ and absolute errors where $S(x, y) = |x - y|$. Depending on the type of forecast made, different scoring functions should be used in the evaluation. For example, when forecasting the mean, squared errors is the most natural scoring function to use. This can be seen from the fact that we can define the mean in terms of that particular scoring function. We can show that

$$\mathbb{E}[Y] = \underset{x}{\operatorname{argmin}}\, \mathbb{E}[(x - Y)^2]. \tag{2.11}$$

To prove (2.11), we want to minimise the expected value $\mathbb{E}[(x - Y)^2]$ with respect to x. We start by writing

$$\begin{aligned}\mathbb{E}[(x - Y)^2] &= \mathbb{E}[x^2 - 2xY + Y^2] \\ &= x^2 - 2x\mathbb{E}[Y] + \mathbb{E}[Y^2]\end{aligned}$$

We minimise this with respect to x by taking the first derivative equal to zero and solving for x. We get that

$$\frac{d}{dx}(\mathbb{E}[Y^2] - 2x\mathbb{E}[Y] + x^2) = -2\mathbb{E}[Y] + 2x$$

We set this equal to zero and get

$$-2\mathbb{E}[Y] + 2x = 0,$$

which can be rewritten as

$$x = \mathbb{E}[Y].$$

For example, take Y to be equally distributed on the set $(y_1, y_2, .., y_N)$. Then

$$\mathbb{E}[Y] = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

which is the sample mean.

A forecasting statistic, such as the mean, that can be expressed in terms of a mimised value of a scoring function is said to have the mathematical property called elicitability. We say that ψ is elicitable if it is the minimised value of some scoring function $S(x, y)$ according to

$$\psi = \underset{x}{\operatorname{argmin}}\, \mathbb{E}[S(x, Y)], \tag{2.12}$$

where Y is the distribution representing verified observations. The distribution can be empirical, parametric or simulated. Furthermore, for elicitability

to hold, the scoring function has to be strictly consistent. The scoring function is defined by Gneiting as a mapping S : I x I $\to [0, \infty)$ where $I = (0, \infty)$. A functional is defined as a mapping $F \to \mathrm{T}(F) \subseteq$ I. Consistency implies that

$$\mathbb{E}[S(t, Y)] \leq \mathbb{E}[S(x, Y)], \tag{2.13}$$

for all F, all $t \in T(F)$ and all $x \in$ I. Strict consistency implies consistency and that equality in (2.13) means that $x \in \mathrm{T}(F)$.

Intuitively we can say that elicitability is a property such that the functional can be estimated with a generalised regression. Furthermore, as mentioned above, the scoring function is appropriate to evaluate the performance of some prediction.

2.5.2 The elicitability of VaR

We can show that $\mathrm{VaR}_\alpha(Y)$ is elicitable through the scoring function

$$S(x, y) = (\mathbb{1}_{(x \geq y)} - \alpha)(x - y). \tag{2.14}$$

According to (2.12) this is true if we can show that

$$\mathrm{VaR}_\alpha(Y) = \underset{x}{\operatorname{argmin}}\, \mathbb{E}[(\mathbb{1}_{(x \geq Y)} - \alpha)(x - Y)]. \tag{2.15}$$

Hence, if we minimise $\mathbb{E}[(\mathbb{1}_{(x \geq Y)} - \alpha)(x - Y)]$ and show that we get $\mathrm{VaR}_\alpha(Y)$ as the minimiser, this proves that VaR is elicitable through its scoring function (2.14). We use $\mathbb{1}_{(x \geq y)} = \theta(x - y)$ where $\theta(x)$ is the Heaviside step function equal to one when $x \geq 0$ and zero otherwise. We can write (2.14) as

$$S(x, y) = (\theta(x - y) - \alpha)(x - y).$$

From this we get

$$\mathbb{E}[S(x, Y)] = \mathbb{E}[(\theta(x - Y) - \alpha)(x - Y)].$$

We can write this as

$$\begin{aligned}\mathbb{E}[(\theta(x - Y) - \alpha)(x - Y)] &= \int (\theta(x - y) - \alpha)(x - y) f_Y(y) dy \\ &= (1 - \alpha) \int_{-\infty}^{x} (x - y) f_Y(y) dy - \alpha \int_{x}^{\infty} (x - y) f_Y(y) dy.\end{aligned}$$

We now want to take the first derivative of $\mathbb{E}[S(x, Y)]$, set it equal to 0 and solve for x. We want to calculate

$$\frac{d}{dx}\Big((1 - \alpha) \int_{-\infty}^{x} (x - y) f_Y(y) dy - \alpha \int_{x}^{\infty} (x - y) f_Y(y) dy\Big) \tag{2.16}$$

We take the derivative of the two terms in (2.16) independently. From the first term, by using Leibniz's rule, we get that

$$\begin{aligned}
&\frac{d}{dx}\Big((1-\alpha)\int_{-\infty}^{x}(x-y)f_Y(y)dy\Big)\\
&\quad=(1-\alpha)\Big(\int_{-\infty}^{x}f_Y(y)dy+(x-x)f_Y(y)-0f_Y(-\infty)(x+\infty)\Big)\\
&\quad=(1-\alpha)\int_{-\infty}^{x}f_Y(y)dy
\end{aligned}$$

Similarly for the second term, we get

$$\begin{aligned}
&\frac{d}{dx}\Big(-\alpha\int_{x}^{\infty}(x-y)f_Y(y)dy\Big)\\
&\quad=-\alpha\int_{x}^{\infty}f_Y(y)dy
\end{aligned}$$

We can now add the two terms together and get

$$\begin{aligned}
\frac{d}{dx}\mathbb{E}[S(x,Y)]&=(1-\alpha)\int_{-\infty}^{x}f_Y(y)dy-\alpha\int_{x}^{\infty}f_Y(y)dy\\
&=\int_{-\infty}^{x}f_Y(y)dy-\alpha
\end{aligned}$$

We set this equal to zero and find

$$\begin{aligned}
\alpha&=\int_{-\infty}^{x}f_Y(y)dy\\
x&=F_Y^{-1}(\alpha),
\end{aligned}$$

which defines $\mathrm{VaR}_\alpha(Y)$. Thus, we have proved that $\mathrm{VaR}_\alpha(Y)$ is elicitable through its scoring function (2.14).

2.5.3 The lack of elicitability and backtestability

Gneiting (2011) contributed to the academic debate by showing that Expected Shortfall is not elicitable. This means that it is not possible to find a scoring function $S(x,y)$ such that Expected Shortfall is defined as the forecast x given a distribution Y that minimises the scoring function $S(x,y)$. A scoring function is a natural tool in evaluating forecasts. Assume you wanted to evaluate the temperature forecasts for the coming week from three different weather institutes in a particular city. You noted the temperature of each day from the three institutes and then you take notes of the actual temperature. How do you then evaluate the three institutes? Most likely you take the error of each day, square it and sum it up over all days. The institute with the lowest sum of squared errors is the best at forecasting the temperature. In mathematical terms, you have minimised the scoring

function of the temperature predictions. What Gneiting showed was that this was not possible to do for Expected Shortfall since the scoring function does not exist. Following his findings, many others have interpreted this as evidence that it is not possible to backtest Expected Shortfall at all. This can be seen in for example Carver (2013). The paper by Gneiting changed the discussion of Expected Shortfall from how it could be backtested to a question of whether it was even possible to do so.

Not all people have interpreted Gneiting's findings as evidence that Expected Shortfall is not backtestable. One of the outstanding issues after his findings was that successful attempts of backtesting Expected Shortfall had been made before 2011. For example, Kerkhof and Melenberg (2004) found methods that performed better than comparable VaR backtests. Following Gneiting's findings, Emmer et al. (2013), showed that Expected Shortfall is in fact conditionally elicitable, consisting of two elicitable components. Backtesting can then be done by testing the two components separately. We let Y denote a random variable with a parametric or empirical distribution from which the estimates are drawn. They proposed using the following algorithm:

- Calculate the quantile as

$$\mathrm{VaR}_\alpha(Y) = \underset{x}{\operatorname{argmin}}\, \mathbb{E}[(\mathbb{1}_{(x \geq Y)} - \alpha)(x - Y)].$$

- Calculate $\mathrm{ES}_\alpha(Y) = \mathbb{E}[L|L \geq \mathrm{VaR}_\alpha]$, where $L = -Y$ is the loss, using the scoring function $\mathbb{E}_P[(x-Y)^2]$, with probabilities $P(A) = P(A|L \geq \mathrm{VaR}_\alpha(Y))$. This gives

$$\mathrm{ES}_\alpha(Y) = \underset{x}{\operatorname{argmin}}\, \mathbb{E}_P[(x - Y)^2)].$$

We know that VaR is elicitable. If we first confirm this, then what is left is simply a conditional expectation and expectations are always elicitable. In the same paper, Emmer et al. (2013) made a careful comparison of different measures and their mathematical properties. They concluded that Expected Shortfall is the most appropriate risk measure even though it is not elicitable. A similar discussion of the implications of different risk measures and its effect on regulation can be found in Chen (2014).

Acerbi and Szekely (2014) argued in a recent article that even without the conditional elicitability, Expected Shortfall is still backtestable. Elicitability is mainly a way to rank the forecasting performance of different models. While VaR is elicitable, this property is not exploited in a normal VaR backtest. This means that Expected Shortfall cannot be backtested through any scoring function but there is no reason why this could not be

done using another method. This means that if we can find a backtest that does not exploit the property of elicitability, there is no reason why that backtest would not work.

Much evidence in the last few years shows that it is possible to backtest Expected Shortfall. The literature presents a variety of methods that can be used. Some of them will be presented in the next chapter.

2.6 Backtesting VaR

We will now describe the mathematics behind a backtest of VaR. Backtesting VaR is straightforward by counting the number of exceedances. That is, counting the number of realised losses that exceeded the predicted VaR level. We define a potential exceedance in time t as

$$e_t = \mathbb{1}_{(L_t \geq \text{VaR}_\alpha(X))}, \tag{2.17}$$

where $L_t = -X_t$ is defined as the realised loss in a period t. $e_t = 1$ implies an exceedance in period t while $e_t = 0$ means no exceedance in time period t. Each potential exceedance is a Bernoulli distributed random variable with probability α. We let e_1, e_2, ... , e_T be all potential exceedances in a period of T days. We assume the random variables to be independent and identically distributed with a Bernoulli distribution. We will always assume that $T = 250$ since backtests are normally done with one year's data at hand. We let Y be the sum of the exceedances, that is the sum of T independent and identically distributed Bernoulli random variables with probability α. Since Y is the sum of independent Bernoulli random variables with the same probability, Y will follow a binomial distribution with parameters $n = T$ and probability $p = \alpha$. We get that

$$Y = \sum_{t=1}^{T} e_t \sim \text{Bin}(T, \alpha).$$

This means that the total number of exceedances in a given year is a binomial random variable with expected value given by the binomial distribution as $T\alpha$. A 99 % VaR has an α of 0.01. Since we have assumed $T = 250$, the expected number of exceedances in one year is 2.5.

With the knowledge that the series of exceedances follows a binomial distribution, it is possible to determine not only the expected value from one year's realised returns but also the probability of a particular number of exceedances. We can define the cumulative distribution function of a binomial variable as

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{2.18}$$

The cumulative probability is simply the probability that the number of exceedances is fewer or equal to the realised number of exceedances for a correct model. This can be used to calculate the confidence when rejecting VaR estimates with too many exceedances. We can explain this using an example of coin flips. We know that the probability of heads or tails is 0.5 for a fair coin. However, after a few flips it seems evident that this coin only shows heads. What is the probability that the coin is not fair after each time it has shown heads? After the first toss, the probability of heads given a fair coin is 0.5. After the second time it is 0.25. After the third, fourth and fifth time it is 0.125, 0.063 and 0.031 respectively. The cumulative probability is the probability that the number of heads in a row is this or fewer. That is, we take one minus the given probabilities. For three, four and five heads in a row it is 0.875, 0.938 and 0.969. This means that after five heads in a row we can say with 96.9 % confidence that the coin is not fair. We can apply the same reasoning to the number of VaR exceedances in a given year if we know the cumulative probability from (2.18). The cumulative probabilities are shown in table 2.3.

Number of exceedances	Cumulative probability
0	8.11
1	28.58
2	54.32
3	75.81
4	89.22
5	95.88
6	98.63
7	99.60
8	99.89
9	99.97
10	99.99

Table 2.3: The table shows the cumulative probabililities of a particular number of exceedances for a 99 % VaR using 250 days returns. In other words, the probability that the number of exceedances is equal to or lower than the number of exceedances given in the first column. The numbers are calculated from (2.18).

We see from table 2.3 that for more than four exceedances we can say with 95 % confidence that there is something wrong with the model since the probability that the number of exceedances is five or fewer is 95.88 %. If we want a confidence level of 99 % then VaR estimates with more than six exceedances should be rejected since the probability of seven or less exceedances is 99.60 %.

2.6.1 The Basel rules on backtesting VaR

We will continue by explaining the Basel rules on backtesting of VaR that apply to all banks. VaR estimates from the calculation of a 99 % VaR have to be reported on a daily basis for supervisors to be able to control that banks have the capital necessary to maintain a certain level of risk. The Basel Committee also requires that the number of VaR exceedances during the last 250 days is reported. Since it is expensive for banks to hold a large amount of capital, they would have an incentive to report too low risk estimates. Hence, the supervisors need some mechanism to increase the capital charge when there is suspicion that the risk estimates reported are too low. This issue is solved by applying an additional capital charge when the number of VaR exceedances during the last year are too many. In this setting, the cumulative probabilities from table 2.3 are of great help.

Zone	Number of exceedances	Factor	Cumulative probability
Green	0	0.00	8.11
	1	0.00	28.58
	2	0.00	54.32
	3	0.00	75.81
	4	0.00	89.22
Yellow	5	0.40	95.88
	6	0.50	98.63
	7	0.65	99.60
	8	0.75	99.89
	9	0.85	99.97
Red	10+	1.00	99.99

Table 2.4: Shows the zones from the Basel rules on backtesting of 99 % VaR. The number of VaR exceedances during the last 250 days determines if the VaR model is in the green, yellow or red zone. The yellow and red zone result in higher capital charge according to equation (2.19) with the additional factor m given in the table.

Table 2.4 shows the framework for backtesting VaR. The starting point is that the number of exceedances in the last 250 days are counted. The number of exceedances in the backtest divides the bank into three zones according to the cumulative probabilities from table 2.3. The bank is considered to be in the green zone if the number of exceedances cannot reject the VaR model with less than 95 % confidence. Hence, from the cumulative probabilities, this implies that the green zone includes up to four exceedances in a year. The yellow zone is defined as the zone where a bank has exceedances so that the model can be rejected with 95 % confidence but not rejected with 99.99 % confidence. We see from the cumulative probabilities in table

2.3 that this implies that between five and nine exceedances forces a bank into the yellow zone. The red zone is defined for the number of exceedances that implies that the VaR model can be rejected with 99.99 % confidence. By the cumulative probabilities this implies at least ten exceedances. The column that is called factor in table 2.4 determines how much the bank will be punished for having too many exceedances. Simplified, we can say that the capital charge is calculated as in (2.19) where m is the factor from table 2.4 and MRC_t is the market risk capital charge in time period t.

$$\text{MRC}_t = (3 + m)\text{VaR}_{t-1} \tag{2.19}$$

From table 2.4 we see that this implies that banks that are in the yellow or red zone will be punished with higher capital charge than banks that are in the green zone. The number of exceedances determines how much extra capital that is needed. By the cumulative probabilities, we see that the Basel Committee adds extra capital when the cumulative probability is higher than 95 %.

Example 2.2 *Assume that a bank has reported a 99 % VaR of 10 million during the last 250 days but has had seven losses larger than 10 million in the last year. According to table 2.4, the probability of six or less exceedances is 98.63 %. This means that the probability that the bank's VaR model is correct is only 1.37 % given the seven exceedances. The bank is in the yellow zone and will be punished for this with a higher capital charge. The additional factor corresponding to seven exceedances is 0.65 according to table 2.4. If the bank would have been in the green zone then the factor m in* (2.19) *had been 0 and the total capital charge would have been* $3 \times VaR_{1\%}$*, that is 30 million. However, since the bank is in the yellow zone with seven exceedances and* $m = 0.65$*, the capital charge is now equal to* $3.65 \times VaR_{1\%}$*, amounting to 36.5 million. Hence, the bank is punished with 6.5 million extra in capital requirements for having too many VaR exceedances during the last year.*

2.7 Conclusion

Expected Shortfall solves two of the main issues related to VaR. The risk measure is subadditive and captures tail risk. While we have shown that backtesting VaR is very straightforward and simple to implement in a regulatory framework, this is not the case for Expected Shortfall. After Gneiting (2011) showed that Expected Shortfall lacked the mathematical property of elicitability, the backtestability of Expected Shortfall has been questioned. However, following the findings of Emmer et al. (2013) and Acerbi and Szekely (2014), it seems like backtesting of Expected Shortfall can be done as long as the method does not exploit the property of elicitabilty.

The next chapter will introduce several methods that can be used in backtesting Expected Shortfall that do not exploit the property of elicitability. The methods take different approaches to solving the problem but all have in common that they do not rely on the use a scoring function.

Chapter 3

The design of different Expected Shortfall backtests

This chapter will describe different approaches to backtesting Expected Shortfall that have been presented in previous literature. The approaches will be explained in detail together with the underlying mechanisms. In total, we will examine the methods from four different papers published between 2008 and 2014 that all take different approaches to solving the problem. Since Expected Shortfall deals with losses in extreme situations, the number of observations that are present at the time of a backtest are usually only a few. The four methods that will be presented here all have a solution to this small sample problem that is associated with the backtesting of Expected Shortfall.

		Parametric assumption	
		Yes	No
Simulations	Yes	*Righi and Ceretta*	*Acerbi and Szekely*
	No	*Wong*	*Emmer, Kratz, and Tasche*

Table 3.1: Shows the fundamental properties of each method introduced in the chapter.

The four methods can be divided into two different category types. They are parametric or non-parametric and they either require simulations or not. These properties are fundamental if the methods are to be implemented in practice. The properties of each model are shown in table 3.1. The methods

are presented in chronological order by publication date. The chapter intends to give an intuition behind the methods and the important steps used to derive the methods rather than to give full proofs.

Before we go deeper into the four approaches, we should note that it is also possible to find several early proposals in the literature of methods to backtest Expected Shortfall. These methods have played an important role in the discussion of Expected Shortfall and its backtestability and should not be disregarded even though they will not be presented here. Some examples are McNeil and Frey (2000) who suggested what they call a residual approach, Berkowitz (2001) who proposed a method that is referred to as the Gaussian approach and there is also what is called the functional delta method proposed by Kerkhof and Melenberg (2004). According to the authors of the papers, all these methods are able to backtest Expected Shortfall under the right circumstances. However, the methods suffer from two drawbacks. They require parametric assumptions and they need large samples. The need for parametric assumption does not have to be an issue if VaR is calculated using a parametric distribution. However, it is important to be able to distinguish a bad model from a bad parametric assumption. The major drawback of the methods is the need for large samples, an unrealistic assumption in the backtesting of Expected Shortfall since the number of losses at hand are always just a few.

3.1 Wong's saddlepoint technique

Wong (2008) proposed the use of a parametric method to backtest Expected Shortfall. The goal of the method is to find the probability density function of Expected Shortfall in the sense that the sample Expected Shortfall is defined as the mean of a number of independent and identically distributed random variables representing the tail distribution. The density function is found by using an inversion formula on the moment generating function and approximate the integral by a saddlepoint technique based on the work of Lugannani and Rice (1980). The probability density function can then be used to determine the cumulative probability function and from there it is straightforward to use any outcome of Expected Shortfall to determine a significance level compared to the estimated value. This section will present the intuition behind the method, how it should be applied and present an example where the method is used. In the end, the method is just about applying a formula. However, there are two steps that need to be understood to get a sense of how the model works. The first step is how to arrive at the inversion formula to calculate the density of Expected Shortfall. The second step is the use of a saddlepoint technique to approximate the integral. Following will be an explanation of the two steps.

3.1.1 Finding the Inversion Integral

The sample Expected Shortfall can be seen as the mean of a number of independent and identically distributed random variables representing the losses larger than VaR. We let ES_N denote the sample Expected Shortfall from N exceedances. We can write this as

$$\mathrm{ES}_N = -\bar{X} = -\frac{1}{N}\sum_{i=1}^{N} X_i, \tag{3.1}$$

where X_i is returns exceeding VaR. Say that we know that returns are normally distributed. This means that every X_i in (3.1) is distributed as the left tail of a normal distribution. In other words, we know the probability density exactly. We assume that we have had in total four VaR exceedances during the last year. We can then assume that the observed Expected Shortfall is an equally weighted sum of four independent and identically distributed random variables. By finding the probability density function of this mean of random variables, it is possible to evaluate each realised Expected Shortfall outcome and its confidence level against the density function. This can be done by assuming that returns follow some known distribution.

We assume a known characteristic function of some random variable X, we call it $\varphi_X(t)$. The characteristic function of the random variable X is defined as $\varphi_X(t) = \mathbb{E}[e^{itX}]$. The probability density function can be calculated from the characteristic function by using the inversion formula given as

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt. \tag{3.2}$$

We now define a new random variable as the mean of the random variable X. We set

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{3.3}$$

We want to find the characteristic function of $\bar{X}$ given that we know the characteristic function of X. We have that

$$\begin{aligned}
\varphi_{\bar{X}}(t) &= \mathbb{E}[e^{it\bar{X}}] \\
&= \mathbb{E}[e^{\frac{1}{N}\sum_{i=1}^{N} X_i}] \\
&= \mathbb{E}[e^{\frac{1}{N}X_1} e^{\frac{1}{N}X_2} \cdots e^{\frac{1}{N}X_N}] \\
&= \mathbb{E}[e^{\frac{1}{N}X_1}]\mathbb{E}[e^{\frac{1}{N}X_2}] \cdots \mathbb{E}[e^{\frac{1}{N}X_N}] \\
&= (\varphi_X(\frac{t}{N}))^N.
\end{aligned} \tag{3.4}$$

So by knowing the characteristic function of X we also know the characteristic function of $\bar{X}$. We can now use this in equation (3.2) and find

$$f_{\bar{X}}(\bar{x}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it\bar{x}}(\varphi_X(\frac{t}{N}))^N dt.$$

By doing a change of variables we get that

$$f_{\bar{X}}(\bar{x}) = \frac{N}{2\pi}\int_{-\infty}^{\infty} e^{-itN\bar{x}}(\varphi_X(t))^N dt. \tag{3.5}$$

We set $\varphi_X(t) = M_X(it)$ where M is the moment generating function defined as $M_X(t) = \mathbb{E}[e^{tX}]$. Furthermore, we define the cumulant generating function as $K_X(t) = \ln M_X(t)$. This means we can write the integral (3.5) as

$$\begin{aligned} f_{\bar{X}}(\bar{x}) &= \frac{N}{2\pi}\int_{-\infty}^{\infty} e^{-itN\bar{x}}(M_X(it))^N dt \\ &= \frac{N}{2\pi}\int_{-\infty}^{\infty} e^{-itN\bar{x}}e^{NK(it)} dt \\ &= \frac{N}{2\pi}\int_{-\infty}^{\infty} e^{N[K(it)-it\bar{x}]} dt. \end{aligned} \tag{3.6}$$

By knowing the distribution and characteristic function of some random variable X we can use the inversion formula given by (3.6) to calculate the probability density function of the mean.

We now define returns $R_1, R_2, ..., R_T$ that are assumed to be independent and identically distributed from a continuous distribution $F(x)$ with density $f(x)$. Wong makes the assumption that the returns are Gaussian and we will follow by his example. It would be convenient to do the exercise assuming a Student's t distribution but then we face the problem that the moment generating function is not defined for this distribution. We then define a new return series consisting only of returns when the VaR level is exceeded. We call them $X_1, X_2, ..., X_N$ where N is the number of VaR exceedances in the return series. We start by defining the sample Expected Shortfall from the returns using the N VaR exceedances as

$$\mathrm{ES}_N = -\bar{X} = -\frac{1}{N}\sum_{t=1}^{N} X_t. \tag{3.7}$$

We assume the random variable R to be standard normally distributed. What we are really interested in is the distribution of X, that is, the tail of the distribution of R. This probability density function of X is simply a scaled version of the density function for R with a smaller interval. We have that

$$f_X(x) = \alpha^{-1}\phi(x)\mathbb{1}_{(x\leq -\mathrm{VaR}_\alpha(R))}, \tag{3.8}$$

where $\phi(x)$ is the standard normal density function. We now want to look for the moment generating function of the random variable X with the probability density function given by (3.8). For the random variable X we get that

$$M_X(t) = \int_{-\infty}^{q} e^{tX} \alpha^{-1} \phi(x) dx, \tag{3.9}$$

where $q = -\text{VaR}_\alpha(R) = \Phi^{-1}(\alpha)$. We can calculate this integral (3.9) as

$$\begin{aligned} M(t) &= \alpha^{-1} \int_{-\infty}^{q} e^{tX} \frac{1}{\sqrt{2\pi}} e^{x^2/2} dx \\ &= \alpha^{-1} e^{t^2/2} \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} e^{(x-t)^2/2} dx \\ &= \alpha^{-1} e^{t^2/2} \times \Phi(q-t). \end{aligned} \tag{3.10}$$

In the approximation of the integral (3.6) we will also need the derivatives of the moment generating function. It is straightforward to show that

$$\begin{aligned} M(t) &= \alpha^{-1} \exp(t^2/2) \times \Phi(q-t) \\ M'(t) &= t \times M(t) - \exp(qt) \times \alpha^{-1} \phi(q) \\ M''(t) &= t \times M'(t) + M(t) - \exp(qt) \times q\alpha^{-1} \phi(q) \\ M^{(m)}(t) &= t \times M(t)^{(m-1)}(t) + (m-1) M(t)^{(m-2)}(t) - \exp(qt) q^{m-1} \times \alpha^{-1} \phi(q) \end{aligned}$$

This means that if we are able to calculate the integral (3.6) with the moment generating function given by (3.10), we have found the probability density function of the mean of the tail. In order to do this we need to approximate the integral. This can be done using a saddlepoint technique that will be explained in the next section.

3.1.2 The saddlepoint technique

We are now going to illustrate how to use the saddlepoint technique in the approximation of integrals. We assume that we want to calculate an integral of the function $f(x)$. We assume that this function $f(x)$ is the exponential of some other function $h(x)$. This means that we have that $f(x) = \exp h(x)$. We now use Taylor expansion to approximate $h(x)$. We get that

$$h(x) \approx h(x_0) + (x - x_0) h'(x) + \frac{(x-x_0)^2}{2} h''(x).$$

This means that we can write

$$f(x) \approx \exp(h(x_0) + (x - x_0) h'(x) + \frac{(x-x_0)^2}{2} h''(x)).$$

We now choose x_0 to be a local maximum. Hence, we set $x_0 = \hat{x}$ defined by $h'(\hat{x}) = 0$ and $h''(x) \leq 0$. We get that

$$f(x) \approx \exp(h(\hat{x}) + \frac{(x-\hat{x})^2}{2} h''(\hat{x})).$$

We now want to find the integral of $f(x)$. We set

$$\int_{-\infty}^{\infty} f(x)dx \approx \int_{-\infty}^{\infty} \exp(h(\hat{x}) + \frac{(x-\hat{x})^2}{2} h''(\hat{x})).$$

We can write the right hand side as

$$\int_{-\infty}^{\infty} \exp(h(\hat{x}) + \frac{(x-\hat{x})^2}{2} h''(\hat{x})) = \exp(h(\hat{x})) \int_{-\infty}^{\infty} \exp(\frac{(x-\hat{x})^2}{2} h''(\hat{x})).$$

The integral is the same integral as a normal density with variance $-h''(\hat{x})$ and mean $\hat{x}$. Hence we can calculate this integral as

$$\int_{-\infty}^{\infty} f(x)dx \approx \exp(h(\hat{x}))\sqrt{-\frac{2\pi}{h''(x)}} = f(\hat{x})\sqrt{-\frac{2\pi}{h''(\hat{x})}}.$$

3.1.3 Wong's method

The intuition behind Wong's method is to use the saddlepoint technique to approximate the integral (3.6) and in that way find the probability density of $\bar{X}$. We start by looking for the saddlepoint of the integral (3.6). Using the notation above, we have that $f(x) = e^{N[K(it)-it\bar{x}]}$. Hence, $h(x) = N[K(it) - it\bar{x}]$. It is then straightforward to find the saddlepoint where $h'(x) = 0$ as

$$K'(\bar{\omega}) = \bar{x}. \tag{3.11}$$

Using the inversion formula and the saddlepoint technique, Lugannani and Rice (1980) showed that if we have the saddlepoint $\bar{\omega}$ we can define

$$\eta = \bar{\omega}\sqrt{NK''(\bar{\omega})}, \tag{3.12}$$

$$\varsigma = \text{sgn}(\bar{\omega})\sqrt{2N(\bar{\omega}\bar{x} - K(\bar{\omega}))}, \tag{3.13}$$

and from this calculate the probability

$$P(\bar{X} \leq \bar{x}) = \begin{cases} \Phi(\varsigma) - \phi(\varsigma)(\frac{1}{\eta} - \frac{1}{\varsigma} + O(N^{-3/2})), & \text{for } \bar{x} < q \\ 1, & \text{for } \bar{x} \geq q \end{cases} \tag{3.14}$$

where $\Phi(x)$ is the standard normal cumulative probability function and $\phi(x)$ is the standard normal density function. The proof is extensive and can be found in Daniels (1987). The null hypothesis is given by

$$H_0: \quad \text{ES}_N = \text{ES}_\alpha(R),$$

where ES_N denotes the sample Expected Shortfall and $\mathrm{ES}_\alpha(R)$ denotes the Expected Shortfall predicted from the normal distribution. The null is tested against the alternative

$$H_1: \quad \mathrm{ES}_N > \mathrm{ES}_\alpha(R).$$

With the moment generating function defined above in equation (3.10), we can get the saddlepoint by solving for t in the following expression

$$K'(t) = \frac{M'(t)}{M(t)} = t - \exp(qt - t^2/2)\frac{\phi(q)}{\Phi(q-t)} = \bar{x}. \tag{3.15}$$

We can then use the saddlepoint $\bar{\omega}$ to calculate η and ς and obtain the p-value stating the probability that the predicted Expected Shortfall is correct given the realised value on Expected Shortfall.

Example 3.1 *We assume that a bank has predicted that its P&L distribution follows a standard normal distribution. The bank is required to report its 97.5 % Expected Shortfall on a daily basis. We can easily determine VaR and hence the threshold value for calculating Expected Shortfall from the standard normal distribution. By* (2.2) *we have that* $VaR_{2.5\%}$ *is given by*

$$\mathrm{VaR}_{2.5\%}(X) = -\Phi^{-1}(0.025) = 1.96. \tag{3.16}$$

Furthermore, we can calculate Expected Shortfall as

$$\mathrm{ES}_{2.5\%}(X) = \frac{\phi(-1.96)}{0.025} = 2.34. \tag{3.17}$$

Based on the last years realised returns, the bank is now going to backtest its Expected Shortfall prediction of 2.34. We assume that during the last year, VaR was exceeded five times with returns equal to $(X_1, X_2, X_3, X_4, X_5)$*=(-2.39, -2.60, -1.99, -2.75, -2.48). Hence, the observed Expected Shortfall is 2.44 and* $\bar{X}$ *=-2.44.*

We now want to find the saddlepoint $\bar{\omega}$ *such that* (3.11) *is fulfilled. If we solve equation* (3.15) *we get a saddlepoint* $\bar{\omega}$ *equal to -0.7286. We now need to find* η *and* ς *to calculate* (3.12) *and* (3.13)*. For that purpose we first need to define* $K(\bar{\omega})$ *and* $K''(\bar{\omega})$*. By using* $K(\bar{\omega}) = \ln M(\bar{\omega})$*, with* $M(\bar{\omega})$ *from* (3.10)*, we get* $K(\bar{\omega})$ *=16 543. Furthermore, we can find* $K''(t)$ *by taking the second derivative of* $K(t)$ *as*

$$\begin{aligned} K'(t) &= \frac{d}{dt} lnM(t) = \frac{M'(t)}{M(t)} \\ K''(t) &= \frac{d}{dt}\frac{M'(t)}{M(t)} = \frac{M''(t)M(t) - (M'(t))^2}{(M(t))^2} \end{aligned} \tag{3.18}$$

In our example we find that $K''(\bar{\omega}) = 0.1741$. We can now take the numbers and plug them into (3.14)*. We find that our p-value is $P(\bar{X} \leq \bar{x}) = 0.2653$. For us to be able to reject Expected Shortfall as incorrect with 95 % significance we would have needed a p-value of at most 0.05. This means that the bank's predicted Expected Shortfall of 2.34 will pass the backtest.*

3.2 Righi and Ceretta's truncated distribution

Righi and Ceretta (2013) proposed a way to backtest Expected Shortfall that relies on the use of a truncated distribution. A truncated distribution is a conditional distribution, for example the conditional normal distribution, that exists only above or below a certain value. In this case, the truncated distribution is the distribution that only exists below the negative VaR level. The core of the method is that by using the truncated distribution it is possible to predict Expected Shortfall as the expected value of the truncated distribution and find the variance of the expected value of the truncated distribution. The variance can then define a dispersion value around Expected Shortfall. With the use of an expected value and a dispersion measure it is easy to define a standard test statistic according to

$$t_s = \frac{r - \mu}{\sigma} \tag{3.19}$$

where t_s denotes the test statistic, r the observed value, μ the expected value and σ the dispersion measure. However, standard test statistics usually need larger samples for convergence. To solve this issue, Righi and Ceretta proposed the use of Monte Carlo simulations. However, since the model is parametric, critical levels can be defined in advance by simulating from the predictive distribution. We will now describe how the method works and how to determine the critical levels in advance.

3.2.1 The Method

Log-returns can be modelled as a GARCH(p, q)-model

$$r_t = \mu_t + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \tag{3.20}$$

$$\sigma_t^2 = \omega + \sum_{p=1}^{P} a_p \varepsilon_{t-p}^2 + \sum_{q=1}^{Q} b_q \sigma_{t-q}^2, \tag{3.21}$$

where r_t is the log-return in time t from the random variable R. μ_t is the conditional mean, ε_t is the shock in return in day t, σ_t^2 is the conditional variance and z_t is white noise. ω, a_p and b_p are parameters that can be estimated from data. From this we can derive the expression of VaR and

Expected Shortfall as

$$\mathrm{VaR}_\alpha(R) = \mu + \sigma F^{-1}(\alpha) \tag{3.22}$$

$$\mathrm{ES}_\alpha(R) = \mu + \sigma E[z_{t+1} | z_{t+1} < F^{-1}(\alpha)] \tag{3.23}$$

where $F(z)$ is the distribution of z. Note here that VaR is a negative value compared to our previous definition. Furthermore, they propose a new measure that they call dispersion of Shortfall (SD) which is to be seen as a dispersion around the mean of the tail distribution. In other words, the dispersion of Expected Shortfall. This is defined as

$$\mathrm{SD}_\alpha(R) = (\sigma^2 \mathrm{Var}[z_{t+1} | z_{t+1} < F^{-1}(\alpha)])^{1/2}. \tag{3.24}$$

By knowing the mean value of the truncated distribution and a dispersion of the mean, it is possible to define a standard test statistic

$$BT_{t+1} = \frac{r_{t+1} - \mathrm{ES}_\alpha(R)}{\mathrm{SD}_\alpha(R)}. \tag{3.25}$$

This value can be estimated directly from observed data if $\mathrm{ES}_\alpha(R)$ and $\mathrm{SD}_\alpha(R)$ are known. This means that it will be easy to backtest Expected Shortfall as long as the dispersion is predicted together with Expected Shortfall. The dispersion can be determined by using the truncated distribution of the underlying parametric distribution of the returns.

To test for significance, Righi and Ceretta proposed simulations to determine a critical value. To make the simulations independent of the mean in (3.20) they replaced r_{t+1} in (3.25) with its GARCH-representation (3.20) and this becomes

$$BT = \frac{z_{t+1} - \mathbb{E}_{t+1}[z_{t+1} | z_{t+1} < F^{-1}(\alpha)]}{(\mathrm{Var}_{t+1}[z_{t+1} | z_{t+1} < F^{-1}(\alpha)])^{1/2}}. \tag{3.26}$$

By assuming that z follows a particular distribution it is possible to use a large number of simulations to determine a critical value. Since z is assumed to follow a given distribution, the critical value can be determined in advance. Righi and Ceretta proposed to simulate a critical value using (3.26) in the following steps

- Simulate N times M random variables u_{ij} from the distribution of z_t, where $i = 1, 2, ..., M$ and $j = 1, 2, ..., N$.
- For every $u_{ij} < \mathrm{VaR}_\alpha(R)$, calculate $B_{ij} = \frac{u_{ij} - \mathbb{E}[u_{ij} | u_{ij} < \mathrm{VaR}_\alpha(R)]}{(\mathrm{Var}[u_{ij} | u_{ij} < \mathrm{VaR}_\alpha(R)])^{1/2}}$
- Choose a significance level and and determine the critical value from the median of all B_{ij} critical values.

Example 3.2 *We assume that a bank uses a standard normal distribution to calculate risk. This means that the bank has a 97.5 % Expected Shortfall of 2.34 and* $VaR_{2.5\%}$ *of 1.96. The bank has now been forced to backtest the predicted Expected Shortfall. The bank has had six losses exceeding 1.96 in the last 250 days. The losses are (2.00, 2.54, 3.00, 2.41, 1.98). The observed Expected Shortfall from the five observations is therefore 2.39.*

In order to do the backtesting we first need to find the dispersion measure (3.24) *and calculate the critical value needed to determine if we will accept or reject the Expected Shortfall prediction of 2.34. The variance of a truncated normal distribution below a value Q is given by*

$$Var[X|X < Q] = [1 - Q\frac{\phi(Q)}{\Phi(Q)} - (\frac{\phi(Q)}{\Phi(Q)})^2] \tag{3.27}$$

In our case we have that $Q = -\mathrm{VaR}_\alpha = \Phi^{-1}(\alpha)$*. Hence we get that*

$$\begin{aligned} Var[X|X < \Phi^{-1}(\alpha)] &= [1 - \Phi^{-1}(\alpha)\frac{\phi(Q)}{\Phi(\Phi^{-1}(\alpha))} - (\frac{\phi(\Phi^{-1}(\alpha))}{\Phi(\Phi^{-1}(\alpha))})^2] \\ &= [1 - \Phi^{-1}(\alpha)\frac{\phi(\Phi^{-1}(\alpha)}{\alpha} - (\frac{\phi(\Phi^{-1}(\alpha)}{\alpha})^2] \end{aligned} \tag{3.28}$$

Plugging in $\alpha = 0.025$ *and calculating the dispersion measure from* (3.24) *we get that* $SD = 0.3416$*. We can now easily calculate the test statistic* (3.25) *as*

$$BT = \frac{-2.39 - (-2.34)}{0.3416} = -0.146$$

We need to do simulations of the test statistic (3.26) *using the algorithm described above. Since we assume returns to be normally distributed we let* z_t *be standard normally distributed and simulate* 10^7 *times from this distribution. However, we only calculate* BT_{t+1} *for* $z_{t+1} \leq -1.96$*. Using a 95 % confidence level, the critical value becomes -2.60. Since the test statistic is -0.146 and higher than the critical value of -2.60, we cannot reject the model and have to accept the predicted Expected Shortfall. The bank's Expected Shortfall prediction will pass the backtest.*

3.3 Emmer, Kratz and Tasche's quantile approximation

The method presented by Emmer et al. (2013) provides a simple way to backtest Expected Shortfall based on the approximation of several VaR levels. This method is a rough approximation compared to the other models but is by far the less complex one and the most likely model to be implemented

in practice due to its simplicity. The starting point of the method is that Expected Shortfall can be approximated with several VaR levels according to

$$\mathrm{ES}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)du \approx$$

$$\frac{1}{4}[\mathrm{VaR}_{0.25\alpha+0.0075}(X) + \mathrm{VaR}_{0.5\alpha+0.005}(X) + \mathrm{VaR}_{0.75\alpha+0.0025}(X) + \mathrm{VaR}_\alpha(X)].$$

Hence, if we assume that $\alpha = 0.05$ then

$$\mathrm{ES}_{5\%}(X) \approx$$

$$\frac{1}{4}[\mathrm{VaR}_{1.25\%}(X) + \mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{3.75\%}(X) + \mathrm{VaR}_{5\%}(X)]$$

That is, VaR 95 %, 96.25 %, 97.5 % and 98.75 % should be backtested jointly in order to backtest Expected Shortfall. If all these levels of VaR are successfully backtested then Expected Shortfall can be considered to be accurate as well. Emmer, Kratz, and Tasche do not specify why four levels of VaR should be used. Since we normally deal with a 97.5 % Expected Shortfall it would be more convenient to use five levels of VaR to get better quantiles. Hence, we can write

$$\mathrm{ES}_{2.5\%}(X) \approx \tag{3.29}$$

$$\frac{1}{5}[\mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{2.0\%}(X) + \mathrm{VaR}_{1.5\%}(X) + \mathrm{VaR}_{1.0\%}(X) + \mathrm{VaR}_{0.5\%}(X)]$$

In the Fundamental Review of the Trading Book by The Basel Committee (2013), supervisors propose that both the 99 % VaR and the 97.5 % should be backtested in the new framework. In some sense, this is an attempt to backtest Expected Shortfall in the same way as Emmer, Kratz, and Tasche propose. However, using just two levels of VaR may be considered too few to call it a backtest of Expected Shortfall.

The method allows for different interpretations since there is no guidance on how the different backtests should be added together. The starting point would be to assume that each VaR level is backtested in the same manner as the Basel backtest in table 2.4, counting the number of exceedances, assuming that we want to be 95 % sure when we reject a model. We approximate a 97.5 % Expected Shortfall as the sum of five different VaR levels as in equation (3.29). Hence, this method would allow us to backtest Expected Shortfall by making sure that VaR at 97.5 %, 98.0 %, 98.5 %, 99.0 % and 99.5 % all pass a backtest. We start by calculating the cumulative probabilities for a different number of VaR exceedances for the different VaR levels using the same methodology as in table 2.3 but for different VaR levels. As above, we assume 250 days in the backtesting which gives $T = 250$. The results can be seen in table 3.2.

	Cumulative Probability - VaR levels				
Number of exceedances	97.5 %	98.0 %	98.5 %	99.0 %	99.5 %
0	0.18	0.64	2.29	8.11	28.56
1	1.32	3.91	10.99	28.58	64.44
2	4.97	12.21	27.49	54.32	86.89
3	12.70	26.22	48.26	75.81	96.21
4	24.95	43.87	67.79	89.22	99.11
5	40.40	61.60	82.43	95.88	99.82
6	56.57	76.37	91.53	98.63	99.97
7	71.03	86.87	96.36	99.60	100.00
8	82.29	93.39	98.59	99.89	100.00
9	90.05	96.96	99.51	99.97	100.00
10	94.85	98.72	99.84	99.99	100.00
11	97.53	99.50	99.95	100.00	100.00
12	98.90	99.82	99.99	100.00	100.00

Table 3.2: Shows the cumulative probability for different number of exceedances for different VaR levels. The probabilities are calculated from equation (2.18) with $T = 250$, assuming 250 returns and probabilities given by the different $\text{VaR}_\alpha(X)$ levels.

We reject a VaR prediction if the cumulative probability is higher than 95 %. We take the 98.5 % VaR as an example. We see that for seven exceedances, the cumulative probability is 96.36 %. This means that if the 98.5 % VaR level is exceeded seven times in the last year then we can reject the VaR prediction with 96.36 % confidence. In other words, we allow maximum six exceedances not to reject the VaR prediction. From table 3.2 we see that in order to have 95 % probability for each VaR level we should not accept more than ten exceedances for VaR 97.5 %, eight for VaR 98.0 %, six for VaR 98.5 %, four for VaR 99.0 % and two for VaR 99.5 %. If any of these backtests fail then Expected Shortfall can be rejected. The maximum number of exceedances at each VaR level can be seen in table 3.3.

α	Maximum number of exceedances
0.025	10
0.020	8
0.015	6
0.010	4
0.005	2

Table 3.3: The table shows the maximum number of exceedances allowed in a backtest of different $\text{VaR}_\alpha(X)$ to be able to reject the VaR prediction with 95 % confidence assuming 250 returns.

Example 3.3 *We assume a bank that knows its VaR 97.5 % to be 1.96 and estimates that its Expected Shortfall at the same level is 2.34. Both estimates are from the standard normal distribution. This means that the bank also has a VaR 98 % of 2.05, VaR 98.5 % of 2.17, VaR 99 % of 2.33 and VaR 99.5 % of 2.58. At the time of backtesting, the bank has had seven losses exceeding VaR 97.5 %. The losses are (2.91, 1.98, 2.34, 2.50, 2.02, 2.39, 2.52). This means a realised Expected Shortfall of 2.38. Each VaR level should be backtested according to the Basel backtest and rejected at the 95 % confidence level. The maximum number of exceedances are those given by table 3.3. We compare the VaR levels to the losses and sum up the number of exceedances for each level in table 3.4.*

	Losses exceeding $\text{VaR}_\alpha(X)$							Total
Loss	2.91	1.98	2.34	2.50	2.02	2.39	2.52	
$\text{VaR}_{2.5\%}$ at 1.96	x	x	x	x	x	x	x	7
$\text{VaR}_{2.0\%}$ at 2.05	x	-	x	x	-	x	x	5
$\text{VaR}_{1.5\%}$ at 2.17	x	-	x	x	-	x	x	5
$\text{VaR}_{1.0\%}$ at 2.33	x	-	x	x	-	x	x	5
$\text{VaR}_{0.5\%}$ at 2.58	x	-	-	-	-	-	-	1

Table 3.4: The table illustrates the numbers in Example 3.3. For each loss, x marks that the loss exceeds the given $\text{VaR}_\alpha(X)$ and - means that it does not exceed the given $\text{VaR}_\alpha(X)$.

We see that $VaR_{1.0\%}(X)$ has five exceedances while according to table 3.3 only four exceedances are allowed not to reject the VaR prediction at this level. Hence, Expected Shortfall can be rejected since one of the VaR levels fails the backtest. The bank does not pass the backtest of Expected Shortfall.

3.4 Acerbi and Szekely's unparametric models

Acerbi and Szekely (2014) have proposed three different methods to backtest Expected Shortfall. The methods are all non-parametric but similar to Righi and Ceretta's methods in the sense that they define a test statistic and try for significance using simulations. However, the advantage with these methods is that no parametric assumption is needed. We will explain the intuition behind the three methods and give the definition of the different test statistics.

3.4.1 The first method

The first method exploits Expected Shortfall conditional on VaR. Expected Shortfall can be written as

$$\mathrm{ES}_\alpha(X) = -\mathbb{E}\Big[X | X + \mathrm{VaR}_\alpha < 0\Big], \tag{3.30}$$

where X is the random variable representing returns. We can rewrite (3.30) as

$$\mathbb{E}\Big[\frac{X}{\mathrm{ES}_\alpha(X)} + 1 | X + \mathrm{VaR}_\alpha(X) < 0\Big] = 0. \tag{3.31}$$

We define an indicator function $I_t = \mathbb{1}_{(X_t < -\mathrm{VaR}_\alpha(X))}$ that indicates a back-testing exceedance of VaR for a realised return X_t in period t. We set $N_T = \sum_{t=1}^{T} I_t$ as the number of exceedances. The test statistic based on (3.31) can then be written as

$$Z_1(\mathbf{X}) = \frac{\sum_{t=1}^{T}(X_t I_t / \mathrm{ES}_{\alpha,t})}{N_T} + 1, \tag{3.32}$$

where $\mathbf{X}$ denotes the vector of realised returns $(X_1, X_2, ..., X_T)$. We call the realised distribution of returns F_t and the predicted distribution of returns P_t. We write $P_t^{[\alpha]}$ for the conditonal distribution tail of the distribution of P_t below the quantile α. We can write this as $P_t^{[\alpha]}(x) = \min(1, P_t(x)/\alpha)$. From this we can define a null hypothesis

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

against the alternatives

$$\begin{aligned} H_1 : \quad & \widehat{\mathrm{ES}}_{\alpha,t}(X) \geq \mathrm{ES}_{\alpha,t}(X), \text{ for all } t \text{ and } > \text{for some } t \\ & \widehat{\mathrm{VaR}}_{\alpha,t}(X) = \mathrm{VaR}_{\alpha,t}(X), \text{ for all } t, \end{aligned}$$

where $\widehat{\mathrm{ES}}_{\alpha,t}(X)$ and $\widehat{\mathrm{VaR}}_{\alpha,t}(X)$ denotes the sample VaR and Expected Shortfall from the realised returns. Under the null the realised tail is assumed to be the same as the predicted tail of the return distribution. The alternative hypothesis rejects Expected Shortfall without rejecting VaR.

3.4.2 The second method

We can write Expected Shortfall as an unconditional expectation

$$\mathrm{ES}_\alpha(X) = -\mathbb{E}\Big[\frac{X_t I_t}{\alpha}\Big]. \tag{3.33}$$

From (3.33), Acerbi and Szekely propose the test statistic

$$Z_2(\mathbf{X}) = \sum_{t=1}^{T} \frac{X_t I_t}{T\alpha \mathrm{ES}_{\alpha,t}(X)} + 1, \tag{3.34}$$

with the following null hypothesis

$$H_0: \quad P_t^{[\alpha]} = F_t^{[\alpha]} \quad \forall t$$

against the alternative

$$H_1: \quad \widehat{\mathrm{ES}}_{\alpha,t}(X) \geq \mathrm{ES}_{\alpha,t}(X), \text{ for all } t \text{ and } > \text{for some } t \tag{3.35}$$

$$\widehat{\mathrm{VaR}}_{\alpha,t}(X) \geq \mathrm{VaR}_{\alpha,t}(X), \text{ for all } t. \tag{3.36}$$

The second model tests Expected Shortfall directly without first backtesting VaR as can be seen from the alternative hypothesis. It jointly rejects VaR and Expected Shortfall.

3.4.3 The third method

The third method presented by Acerbi and Szekely was inspired by an article published by Berkowitz (2001). The idea is that you test the entire return distribution and not just Expected Shortfall. As above, we assume a predictive distribution function P_t. Here, we need the assumption that P_t is continuous. Now we want to do a probability transformation and test if the observed ranks $U_t = P_t(X_t)$ are independent and uniformly distributed U(0,1). Say that we have predicted that the return distribution is normally distributed. This means that P_t is a standard normal distribution function. We then observe 250 realised returns X_t. If we take $P_t(X_t) = \Phi(X_t)$ then we expect to get 250 random variables uniformly distributed between 0 and 1. If we get many values close to zero then we suspect that the returns are not normally distributed. Acerbi and Szekely proposed that Expected Shortfall was estimated as

$$\widehat{\mathrm{ES}}_{\alpha}^{N}(Y) = -\frac{1}{[N\alpha]} \sum_{i}^{[N\alpha]} Y_{i:N}, \tag{3.37}$$

where N is the number of observed returns and $Y_{i:N}$ is ordered returns. Hence, Expected Shortfall is estimated by the average of the $N\alpha$ worst outcomes, rounded to the nearest lower integer. This is the same as in the definition of Expected Shortfall from an empirical distribution. The proposed test statistic to use is

$$Z_3(\mathbf{X}) = -\frac{1}{T} \sum_{t=1}^{T} \frac{\widehat{\mathrm{ES}_{\alpha}^{(T)}}(P_t^{-1}(U))}{\mathbb{E}_V[\mathrm{ES}_{\alpha}^{(T)}(P_t^{-1}(V))]} + 1, \tag{3.38}$$

where the denominator can be computer directly as

$$\mathbb{E}_V[\mathrm{ES}_\alpha^{(T)}(P_t^{-1}(V))] = -\frac{T}{[T\alpha]}\int_0^1 I_{1-p}(T-[T\alpha],[T\alpha])P_t^{-1}(p)dp. \quad (3.39)$$

$I_x(a,b)$ is a regularized incomplete beta function. In this case the entire distribution is tested under the null

$$H_0: \quad P_t = F_t \quad \forall t$$

against the alternative

$$H_1: \quad P_t \succcurlyeq F_t \quad \forall t$$

where $\succcurlyeq$ denotes weak stochastic dominance. Also in this case, Expected Shortfall is not backtested independently but jointly with other quantiles of the distribution.

3.4.4 Finding the significance

To test for significance in the three methods above, Acerbi and Szekely proposed simulations from the distribution under H_0. They proposed the following steps

- Simulate X_t^i from P_t for all t and $i = 1, 2, ..., M$
- For every i, compute $Z^i = Z(X^i)$. That is, compute the value of Z_1, Z_2 or Z_3 depending to the type of method applied, using the simulations from the previous step.
- Estimate the p-value as $p = \sum_{i=1}^{M}(Z^i < Z(x))/M$. Where $Z(x)$ denotes the observed value on Z_1, Z_2 or Z_3.

This can be done using for example 5000 simulations for each of the methods.

Example 3.4 *We illustrate an example of Acerbi and Szekely's first method. We assume that a bank predicts that their return distribution follows a standard normal distribution. Hence, the predicted VaR at 97.5 % is 1.96 and the predicted Expected Shortfall at the same level is 2.34. Last year's returns resulted in five losses exceeding VaR at (2.01, 2.90, 2.78, 2.41, 2.44) which gives a realised Expected Shortfall of 2.51. We have* $(X_1, X_2, X_3, X_4, X_5)$*=(-2.01, -2.90, -2.78, -2.41, -2.44) We now want to calculate the test statistic* (3.32) *with our values. We have that*

$$Z_1(\boldsymbol{X}) = \frac{\sum_{t=1}^{T}(X_t I_t/\mathrm{ES}_{\alpha,t})}{N_T} + 1 = -\frac{2.51}{2.34} + 1 = 0.01$$

To find the significance or the p-value we need to use simulations. Using 5000 simulations, we get a p-value of 0.13. Hence, we cannot reject the predicted Expected Shortfall of 2.34 and the bank will pass the backtest.

3.5 Conclusion

In this chapter, we have presented four different approaches that can be taken in the backtesting of Expected Shortfall. In all of the methods, the lack of elicitability is not a problem since the backtests do not rely on the use of a scoring function. The approximative method proposed by Emmer et al. (2013) relies on a generalisation of a standard VaR test and does therefore not suffer from the fact that the number of Expected Shortfall observations usually are few. The other three methods, Wong (2008), Righi and Ceretta (2013) and Acerbi and Szekely (2014), solve the problem of small samples by using two different approaches. They either use Monte Carlo simulations to determine the confidence level of the backtest or they use a parametric assumption to determine some kind of probability density of Expected Shortfall. The use of simulations to determine the significance in the backtests could be generealised to other types of backtests, for example one of the early Expected Shortfall backtests proposed by McNeil and Frey (2000), Berkowitz (2001) or Kerkhof and Melenberg (2004). Acerbi and Szekely's third method is an example of such a generalisation of the work of Berkowitz (2001).

From this chapter we can conclude that there are several methods in which Expected Shortfall can be backtested. However, we have yet not established whether the methods work. In the next three chapters we will test the properties of each method. The purpose of this is to try to answer some of the questions outstanding on the backtestability of Expected Shortfall. We would like to know if there is any method that can backtest Expected Shortfall accurately. We will do this by investigating three issues related to the backtesting methods; the ability to accept true predictions, the ability to reject false predictions and implementation. We will devote one chapter to the acceptance, one chapter to the rejection and one chapter to the implementation of the methods. We will start each chapter by stating some questions that will help us with the analysis.

Chapter 4

The ability to accept true Expected Shortfall predictions

One of the most important aspects of a backtest is that when a predicted Expected Shortfall is correct, the backtest should not reject this estimate. That is, if we predict Expected Shortfall from a certain distribution and then simulate exceedances from the tail of the same distribution we want the backtest to accept that prediction with high confidence. We now want to investigate if the methods defined in the previous chapter are able to do this. We will investigate this through answering two questions related to this issue:

- Which method gives the highest confidence in accepting true Expected Shortfall estimates?
- Does the acceptance performance depend on the number of VaR exceedances?

We will begin the next section by presenting the methodology that will be used before we move on to the results. The answers to the questions can be found in the final section of the chapter.

4.1 Methodology

If we believe that a bank has a 97.5 % Expected Shortfall of 2.34 and that its returns follow a standard normal distribution, 250 simulations drawn from the standard normal distribution should accept the Expected Shortfall estimate with high confidence. This is very straightforward to test. We want to simulate realised values on Expected Shortfall from the standard

normal distribution. We do this in two different ways to answer each of the two questions stated in the previous section.

4.1.1 Simulating a random number of exceedances

We start by simulating randomly 250 returns from a standard normal distribution and calculate the number of exceedances and the realised Expected Shortfall. This means that we compare every loss to our VaR and then calculate the number of exceedances and the realised Expected Shortfall as the mean of all the losses that exceed VaR. We do 10^5 simulations for each method to determine the acceptance rate, that is the proportion of times the Expected Shortfall prediction is accepted.

4.1.2 Simulating a fixed number of exceedances

In a second approach we want to see how the acceptance rate varies with the number of exceedances. This means that we determine how many exceedances we want and simulate them directly from the tail of the standard normal distribution. Let us assume that we want to investigate the behaviour for six exceedances. In this case we simulate uniformly six values of $\alpha \in (0, 0.025)$ and find the corresponding value for the standard normal distribution. We do this using 5000 simulations. Since we are more interested in the relative performance for different number of exceedances rather than the actual numbers, 5000 simulations is enough. By doing simulations were we control the number of exceedances we can see if the rate at which the method manages to accept a true model varies with the number of exceedances. We let N be the number of exceedances assumed. For each value on $N = 1, 2, ..., 10$ we do 5000 simulations where each simulation represents N draws from the standard normal tail below the value -1.96. Every backtest uses a confidence of 95 % meaning that a p-value below 0.05 implies a rejected Expected Shortfall prediction.

4.2 Results

We will now present the results of the simulations. We will to this first for the simulations with a random number of exceedances and then for the simulations with a fixed number of exceedances.

4.2.1 A random number of exceedances

We start by looking at the rate at which a true prediction is accepted for a random number of exceedances. The results are shown in table 4.1.

Method	Acceptance rate
Wong	0.9482
Righi and Ceretta	0.9995
Emmer, Kratz, and Tasche	0.7793
Acerbi and Szekely's first method	0.9423
Acerbi and Szekely's second method	0.9562
Acerbi and Szekely's third method	0.9528

Table 4.1: The table shows the outcome of 10^5 simulated backtests where the predicted 97.5 % Expected Shortfall is from the standard normal distribution and the simulated values are 250 returns from the same distribution. The acceptance rate defines the proportion of the simulations that accepted the true Expected Shortfall prediction.

We see that the method that accepts a true Expected Shortfall prediction with the highest confidence is Righi and Ceretta's method. It has close to perfect acceptance rate of 0.9995. The approximative method by Emmer, Kratz, and Tasche has the lowest acceptance rate of only 0.7793. The other four methods have acceptance rates around 0.95.

4.2.2 A fixed number of exceedances

We now take a look at the results were the acceptance rate can be found as a function of the number of exceedances. Figure 4.1 shows the outcome. The figure confirms that Righi and Ceretta's method has a high acceptance rate. Acerbi and Szekely's second and third methods have perfect acceptance up to nine exceedances but then it starts to decrease. On the other hand, Acerbi and Szekely's first method has an increasing rate of acceptance as the number of exceedances increase.

The only odd behaviour among the methods in that the method by Emmer, Kratz, and Tasche has a decreasing acceptance rate as the number of exceedances increase. This explains why the acceptance rate of Emmer, Kratz, and Tasche's method is so low in table 4.1. The reason for this may be that since the method is built on counting the number of exceedances, more exceedances at the primary VaR level increases the probability of many exceedances at a VaR level with lower α. We see that for seven exceedances, which is close to the expected number of exceedances of 6.25, the acceptance rate is below 0.8.

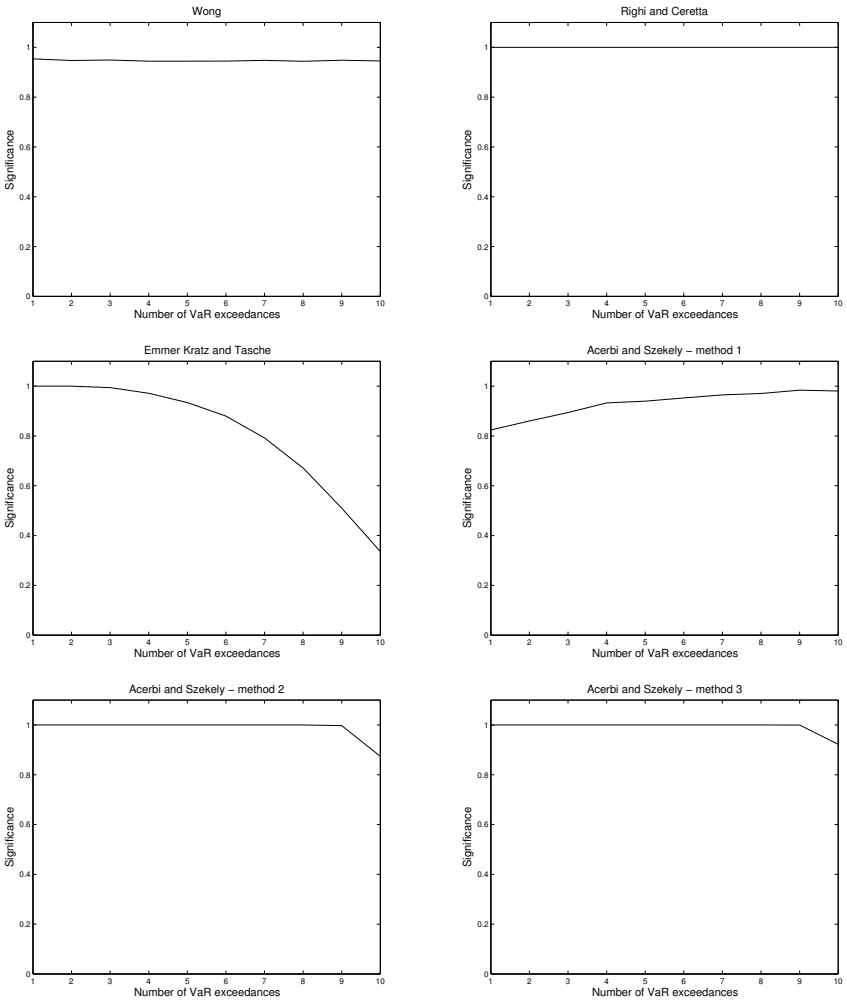

Figure 4.1: The figure shows how often the 97.5 % Expected Shortfall prediction of 2.34 from a standard normal distribution is accepted based on simulations of a particular number of exceedances from the same distribution. In total, 5000 simulations were used to determine the significance. Significance of 1 implies that all 5000 simulations accepted the prediction and 0 means that the Expected Shortfall prediction was never accepted. We can interpret this as the probability that a true Expected Shortfall prediction will be accepted.

4.3 Conclusion

Overall, the methods accept true Expected Shortfall predictions with high confidence meaning that in the sense of accepting true estimates all the backtests seem to work well. The method that has the highest confidence in accepting true Expected Shortfall predictions is Righi and Ceretta's method which accept almost 100 % of the simulations for all number of exceedances. Second up are Acerbi and Szekely's second and third method which have 100 % acceptance rate as long as the number of exceedances are below ten but overall has the same performance as Acerbi and Szekely's first method and Wong's method.

All methods except Wong and Righi and Ceretta show that the performance depends on the number of exceedances. Acerbi and Szekely's first method shows a higher rate of acceptance as the number of exceedances increase. In contrary, Emmer, Kratz, and Tasche and Acerbi and Szekely's second and third method show decreasing acceptance as the number of exceedances increase.

The method with the lowest acceptance rate and the only concern is Emmer, Kratz, and Tasche's approximative approach that shows that the acceptance rate decreases quickly as the number of exceedances increases. This affects the overall performance of the method in terms of the confidence in accepting true predictions.

Chapter 5

The ability to reject false Expected Shortfall predictions

We are now going to investigate the next important property of a backtest; each method's ability to reject wrong Expected Shortfall predictions. The purpose of this will be to see if the methods are able to reject too small predictions of Expected Shortfall if we observe that the realised Expected Shortfall is large compared to the prediction. If we have predicted an Expected Shortfall of 2 and then observe that the realised Expected Shortfall from 250 days data is 10, it is obvious that we want the backtest to reject the prediction. To investigate each method's ability to reject we will try to answer two questions related to this:

- Which method gives the highest confidence in rejecting false Expected Shortfall estimates?
- Does the rejection performance depend on the number of VaR exceedances?

In the next section we will explain the methodology that will be used to investigate each method's ability to reject wrong Expected Shortfall predictions. We will then devote one section to each of the questions above. The answers to each question can be found at the end of each section as well as in the conclusions in the last section of the chapter.

5.1 Methodology

We will now define the framework that will be used to investigate when the different methods reject wrong Expected Shortfall predictions. This is more complicated than finding the acceptance rate since it is not as straightforward to define an incorrect prediction as it is to define a correct prediction. As above, we will assume a bank that uses a standard normal distribution to calculate its risk measures. This means that we can use VaR and Expected Shortfall of the standard normal distribution from table 2.2. We assume here that the relevant risk measure for reporting and backtesting is the 97.5 % Expected Shortfall. Hence, the bank's capital requirement is based on the 97.5 % Expected Shortfall of 2.34. We will always assume that the predicted Expected Shortfall is 2.34 and change the underlying distribution from which we will simulate 250 realised returns.

5.1.1 Simulating realised returns

We simulate 250 returns from a given distribution that represents last year's realised returns. The distribution from which we simulate the returns will be the Student's t distribution. By varying the degrees of freedom it will be possible to change the tail of the distribution. For small values on the degrees of freedom, the realised Expected Shortfall from the 250 returns will get large. If we simulate returns from a Student's t distribution with three degrees of freedom then we expect an Expected Shortfall of 5.04 as can be seen from table 2.2. In this case, the predicted Expected Shortfall of 2.34 will be small in comparison to the simulated Expected Shortfall and we want the method to reject the bank's Expected Shortfall prediction. For large values on the degrees of freedom, the Student's t distribution converges to a normal distribution. Hence, for many degrees of freedom, we want the underlying Expected Shortfall to be accepted since the realised returns will be simulated from the same distribution as we calculated the predicted Expected Shortfall from. Furthermore, we can add another parameter to the simulations. By changing the standard deviation of the realised returns we get another dimension in which we can increase or decrease the simulated observed value on Expected Shortfall and see how far we can go before we find that the predicted Expected Shortfall is rejected.

5.1.2 Simulating a tail

There is one problem with the methodology described above. If we simulate 250 returns from a Student's t distribution with three degrees of freedom and count the number of losses larger than VaR then we expect to get a large number of exceedances. For a standard normal distribution the 97.5 % VaR is 1.96. The cumulative probability of -1.96 for a Student's t dis-

tribution with three degrees of freedom is 0.072. This means that the expected number of exceedances in this case is $250 \times 0.072 \approx 18$ compared to the expected number of exceedances when the VaR estimate is correct of $250 \times 0.025 = 6.25$. This means that as we increase the degrees of freedom we will also increase the number of exceedances. We will reach some point where we would easily be able to reject the model using a simple VaR backtest. However, we want to try the backtests for situations similar to those in figure 1.1, when VaR is correct but Expected Shortfall has been underestimated. To solve this issue we can simulate quantiles rather than returns. If we simulate 250 values on α, we can choose every $\alpha \leq 0.025$ and find the corresponding value from the chosen underlying distribution. Since we are only concerned about what happens in the tail of the distribution, we can assume that the right part of the the distribution, defined above VaR, is the same for all distributions. It only differs for $\alpha \leq 0.025$, the part of the distribution that defines Expected Shortfall. This will allow us to get the right number of exceedances and the same VaR while varying the risk in the tail of the distribution. This works because there are just two inputs needed in the methods above, the observed Expected Shortfall, defined as the mean of the losses exceeding VaR, and the number of exceedances. The value of $\alpha \leq 0.025$ is always larger in absolute terms for the Student's t distribution than for the normal distribution, which means that for $\sigma \geq 1$, the losses drawn from the tail will always be larger than VaR. However, as we decrease σ, we may have exceedances that do not exceed our VaR of 1.96 with this methodology. Such exceedances are removed from the analysis. Table 5.1 illustrates possible realised values on Expected Shortfall from different underlying distributions. They are based on 10^4 simulations of 250 returns.

	Empirical Expected Shortfall		
Distribution	Minimum	Maximum	Average
$t_3(0,1)$	3.18	10.09	4.61
$t_5(0,1)$	2.57	5.79	3.35
$t_{10}(0,1)$	2.23	4.14	2.74
$t_{30}(0,1)$	2.04	3.38	2.43
$t_{60}(0,1)$	2.00	3.22	2.36

Table 5.1: The table shows the outcome of 10^4 simulations of 250 returns. The tail of the distribution from which the exceedances are simulated from is given by the table.

5.1.3 Implementing the methodology

We will now give an example explaining how the analysis will be carried out.

Example 5.1 *A bank uses a standard normal distribution to predict its 97.5 % VaR and Expected Shortfall. By table 2.2 this means that the bank has predicted that its VaR 97.5 % is 1.96 and Expected Shortfall 97.5 % is 2.34. The bank knows that the backtesting of Expected Shortfall will be done according to Wong's method and want to investigate what will happen in the backtesting depending on the realised returns during the next 250 days. The bank will fail the Expected Shortfall backtest if Wong's model can reject the bank's Expected Shortfall estimate with 95 % confidence. After many years of risk management, the bank is sure that they report accurate VaR numbers. However, they are afraid that the tail risk is larger than what they predict. Therefore, they will simulate realised returns from the tail of a distribution and change some parameters to see how this would affect the backtest. They try three different tail distributions that represent potential losses exceeding VaR in the next year. They use the Student's t distribution with degrees of freedom* $\nu = 3$ *and* $\sigma = 1$*, which we call distribution 1. The Student's t distribution with* $\nu = 10$ *and* $\sigma = 0.9$ *which we call distribution 2 and the Student's t distribution with* $\nu = 55$ *and* $\sigma = 1$*, called distribution 3.*

For distribution 1 they simulate 250 quantiles between 0 and 1. The simulation gives seven quantiles less than or equal to $\alpha = 0.025$*. They are (0.021, 0.024, 0.022, 0.016, 0.002, 0.010, 0.025). The corresponding values for a Student's t distribution with* $\nu = 3$ *and* $\sigma = 1$ *is (-3.437, -3.227, -3.381, -3.770, -8.047, -4.574, -3.189). This means that if this distribution represents the losses during the next year, the bank would have a realised Expected Shortfall of 4.232. Using Wong's methodology we have that* $\bar{X}$*=-4.232 as in equation* (3.3)*. With* $\bar{X}$*=-4.232 and* N*=7, Wong's method gives a p-value of 0.000. The Expected Shortfall prediction is rejected and the bank would not pass the backtest.*

For distribution 2 we also simulate 250 quantiles between 0 and 1. The simulation gives four quantiles less than or equal to $\alpha = 0.025$*. They are (0.007, 0.011, 0.006, 0.005). The corresponding values for a Student's t distribution with* $\nu = 10$ *and* $\sigma = 0.9$ *is (-2.693, -2.453, -2.785, -2.863). This means that from this simulation we get a realised Expected Shortfall of 2.699. Using Wong's methodology we have that* $\bar{X}$*=-2.699 as in equation* (3.3)*. With* $\bar{X}$*=-2.699 and* N*=4, Wong's method gives a p-value of 0.033. Since the p-value is smaller than 0.05, we reject the Expected Shortfall prediction and the bank would fail the backtest.*

For distribution 3 we repeat the same procedure as above. The simulation gives five quantiles less than or equal to our $\alpha = 0.025$*. They are (0.005, 0.006, 0.010, 0.002, 0.023). The corresponding values for a Student's t distribution with* $\nu = 55$ *and* $\sigma = 1$ *is (-2.679, -2.693, -2.453, -2.785, -2.863). This means that from this simulation we get a realised Expected Shortfall*

of 2.548. Using Wong's methodology we have that $\bar{X}$=-2.548 as in equation (3.3)*. With $\bar{X}$=-2.548 and N=5, Wong's method gives a p-value of 0.0957. In this case, Expected Shortfall is not rejected and the bank would pass the backtest.*

We illustrate the results in figure 5.1.

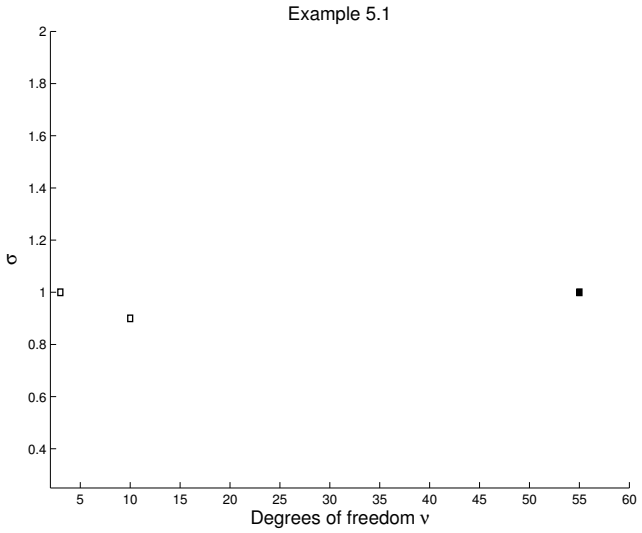

Figure 5.1: The figure illustrates the outcome of the three backtests simulated in Example 5.1. 250 realised returns are simulated with the tail distribution $t_\nu(0, \sigma)$ with parameters according to the figure. The predicted VaR and Expected Shortfall are assumed to be $N(0, 1)$. The simulations give a number of exceedances which can be used to calculate a realised Expected Shortfall. The realised Expected Shortfall is then backtested using Wong's method. A white rectangle represents a rejected backtest while a black rectangle represents an accepted backtest.

Example 5.1 illustrates the framework that will be used to determine how each model works in practice. The three distributions described above will in the full analysis be 27 903 different distributions where σ goes from 0.2 to 2 and ν goes from 2 to 100. This means that we can define regions of the graph in figure 5.1 that tell us where the prediction is rejected and where it is accepted based on the parameters of the underlying distribution. We will now use this framework to analyse the questions stated in the beginning of the chapter.

5.2 The overall ability to reject

We will now try to answer the first question stated in the beginning of this chapter: which method gives the highest confidence in rejecting false Expected Shortfall estimates? We will do this using the framework defined in the previous section.

5.2.1 Results

The results of the backtests can be seen in figure 5.2. We see that all methods have similar performance. It is evident that they are all one-sided tests rejecting the predicted Expected Shortfall when the observed Expected Shortfall is too large but never when it is too small. This can be seen from the fact that when ν is large and σ small, the realised Expected Shortfall is small and the figures are black in this area, representing sure acceptance of the prediction. As the realised Expected Shortfall becomes larger, either by a decrease of ν or an increase in σ, the likelihood of rejection increases as can be seen from the fact that the area becomes more white as we move in that direction. This means that the probability of acceptance decreases as the observed Expected Shortfall becomes large. Overall, we see that all the methods have high probabilities of accepting a true model and decreasing probability of acceptance as the observed Expected Shortfall gets larger. This is exactly the behaviour we want to see in a backtest of Expected Shortfall.

The method that seems to have the highest confidence in rejecting a false prediction is Acerbi and Szekely's first method. This can be seen from the fact that the white area in the graph describing the rejection rate of Acerbi and Szekely's first method is much larger than for the other methods. The white area represents a probability of rejection of at least 95 %. Acerbi and Szekely's second and third method do not show as good performance in terms of rejecting. The large medium grey area for Acerbi and Szekely's second and third methods in the graphs in figure 5.2 implies that the acceptance rate is above 50 % and below 95 % for a large part of the area were we would like to reject the predicted Expected Shortfall. This means that the best performance of the two methods is a 50 % chance of rejection.

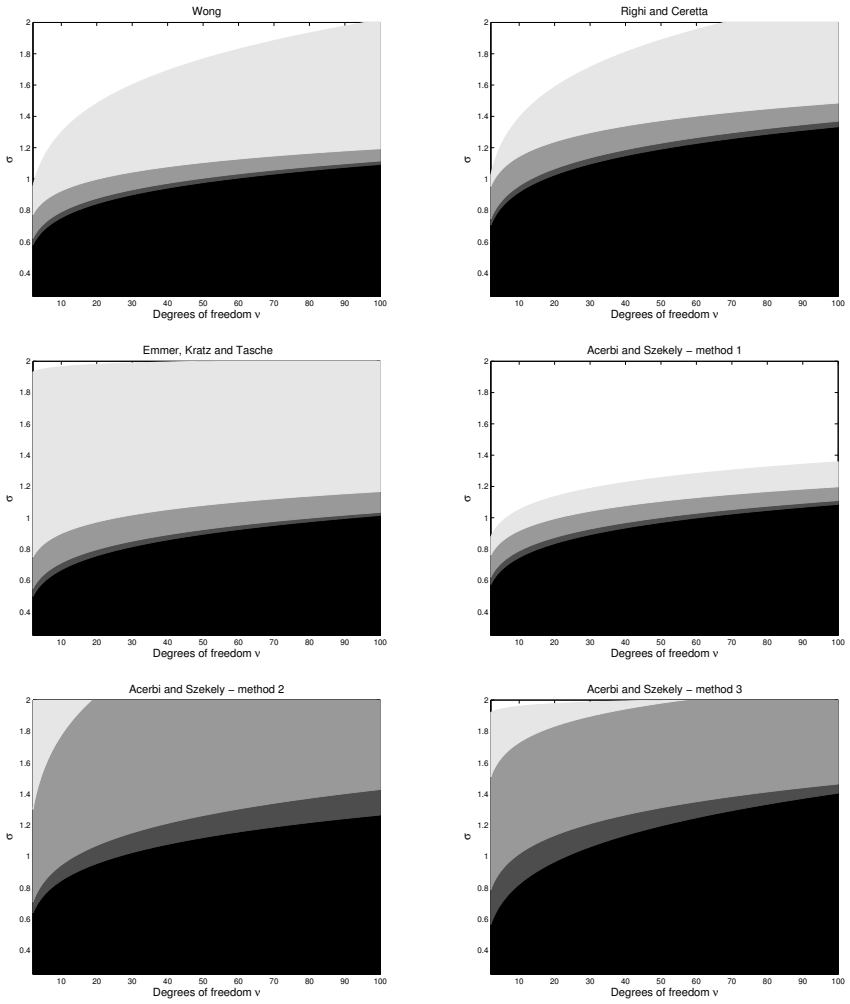

Figure 5.2: The figure illustrates the outcome of 27 903 simulated backtests. The predicted VaR and Expected Shortfall are assumed to be $N(0,1)$. 250 realised quantiles are then simulated with the tail distribution $t_\nu(0,\sigma)$ with parameters according to the figure. The simulations give a number of exceedances which can be used to calculate a realised Expected Shortfall. The figure shows in total five different areas where black illustrates a sure acceptance of Expected Shortfall in the backtest and white means a sure rejection of Expected Shortfall. A black area means 100 % acceptance rate. Dark grey has acceptance rate above 95 % but below 100 %. Medium grey has an acceptance rate somewhere between 95 % and 50 %. For light grey, the acceptance rate is lower than 50 % but higher than 5 %. The white area has an acceptance rate lower than 5 %.

5.2.2 Conclusion - which method gives the highest rejection confidence?

The method that has the highest confidence in rejecting false Expected Shortfall predictions is Acerbi and Szekely's first method. This can be seen by looking at figure 5.2. The methods that show the weakest performance in terms of rejecting are Acerbi and Szekely's second and third method.

5.3 The ability to reject with a fixed number of exceedances

We are now going to investigate the second question of the chapter: does the rejection performance depend on the number of VaR exceedances?

We are going to do the same simulations as above but we are now interested in the overall ability to reject wrong predictions as a function of the number of exceedances. We define the rejection rate as the proportion of the 27 903 backtests made in figure 5.2 that are rejected. A rejection rate of 1 would mean that the entire window defined in figure 5.2 is white while a rejection rate of 0 would mean that the entire window would be black and that none of the 27 903 backtests are rejected.

5.3.1 Results

Figure 5.3 shows the results. We see that Wong, Righi and Ceretta and Acerbi and Szekely's first methods all have stable rejection rates which can be seen from the fact that they are not dependent or shows very little dependence to the number of exceedances. However, the rejection rate of Righi and Ceretta is lower than for the other two methods. For Emmer, Kratz, and Tasche and Acerbi and Szekely's second and third method the rejection rate depends on the number of exceedances. Overall, the methods converge to the same rate of rejection as the number of exceedances comes closer to ten. At ten exceedances, all methods except Righi and Ceretta have a rejection rate around 0.6.

Emmer, Kratz, and Tasche's method has a rejection rate of 0 for less than two exceedances. This is easy to confirm from the definition of the method. We backtest VaR at different levels and then count the number of exceedances. We need at least two exceedances to be able to reject $\text{VaR}_{0.5\%}(X)$ according to table 3.3, so for less than two exceedances we can never reject any prediction and the rejection rate is 0 independent of the value on the realised Expected Shortfall.

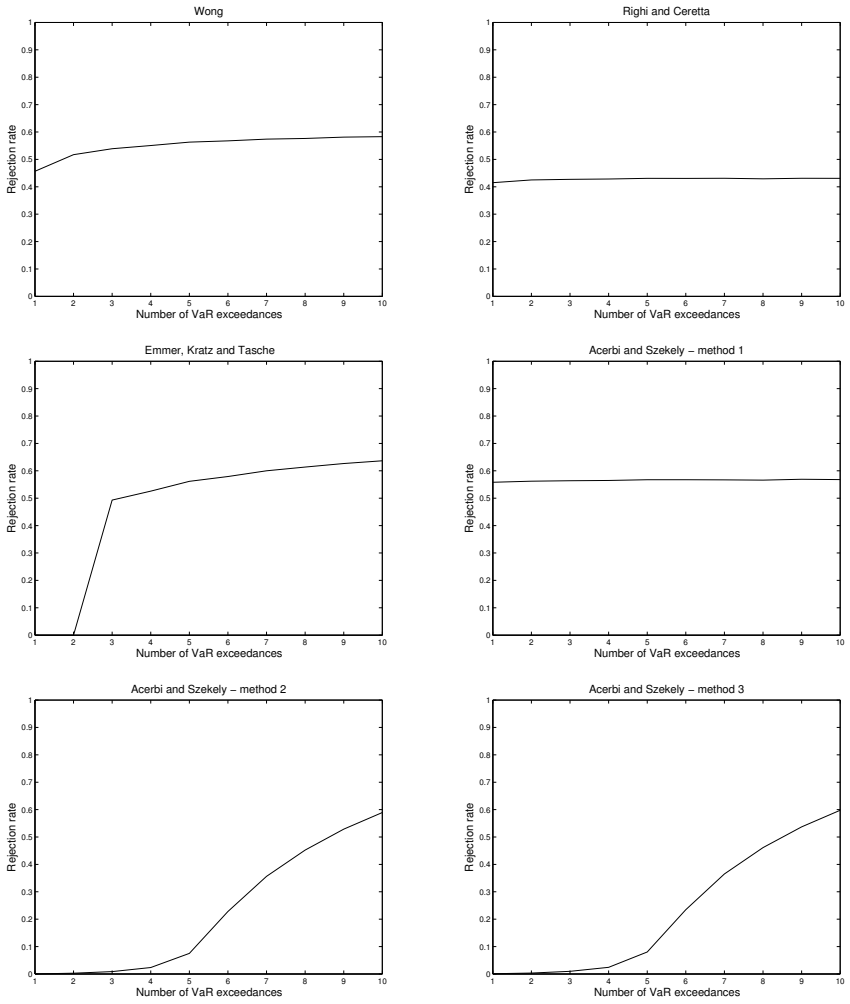

Figure 5.3: The figure illustrates the outcome of 27 903 simulated backtests. The predicted VaR and Expected Shortfall are assumed to be $N(0,1)$. 250 realised quantiles are then simulated with the tail distribution $t_\nu(0,\sigma)$ with parameters $\nu \in (2,60)$ and $\sigma \in (0.2,2.0)$. The rejection rate specifies how many of the 27 903 backtests made in figure 5.2 that rejects the predicted 97.5 % Expected Shortfall as a function of the number of VaR exceedances. 1 means that all 27 903 backtests shows a rejection of the predicted Expected Shortfall and 0 means that no simulation is a rejected backtest.

For Acerbi and Szekely's second and third method we see that the performance is highly dependent on the number of exceedances. For up to $N = 4$, almost all simulations accept the predicted Expected Shortfall. As the number of exceedances increase, the area of rejection increases and for $N = 10$, the method has more or less the same performance as the other methods. The explanation can be found by looking at the alternative hypothesis in

Acerbi and Szekely's second method as in equation (3.35) and (3.36). Expected Shortfall and VaR are tested jointly and rejected jointly. This means that the number of exceedances are important in the backtesting of Expected Shortfall since the backtesting of VaR is done at the same time. When the number of exceedances are few it is evident that VaR cannot be rejected and this seems to compensate for the large value on Expected Shortfall. As an independent Expected Shortfall backtest this method does not work very well in that sense. The same reasoning holds for Acerbi and Szekely's third method. However, instead of testing VaR and Expected Shortfall jointly, the entire distribution is tested.

5.3.2 Conclusion - does the number of exceedances matter?

We see that for Emmer, Kratz, and Tasche and Acerbi and Szekely's second and third methods, the ability to reject depends on the number of exceedances. The rejection rate increases with the number of exceedances. For the other three methods the rejection rate is more or less stable. For ten exceedances all methods except Righi and Ceretta have the same rejection rate.

5.4 Conclusion

Overall, we see that all methods are able to reject the predicted Expected Shortfall when it is small compared to the realised Expected Shortfall. Acerbi and Szekely's first method showed the strongest performance in rejecting. For a large number of exceedances, all methods have fairly similar behaviour in their ability to reject. However, the rejection performance of Emmer, Kratz, and Tasche and Acerbi and Szekely's second and third methods depends on the number of exceedances. For all three methods, the ability to reject increases with the number of exceedances.

There seems to be several models that work well in backtesting Expected Shortfall. All methods taken from the literature have a high accuracy in accepting a true Expected Shortfall prediction as was shown in the previous chapter and decreasing probability of accepting as the prediction becomes small in comparison to the realised outcome from 250 simulated returns as was shown in this chapter. This means that we have found four different approaches that do not exploit the property of elicitability and backtest Expected Shortfall.

We are now going to look at the complexity of the methods together with their performance to give some guidance on if and how the methods can be implemented in practice.

Chapter 6

Implementing the methods in practice

This chapter will analyse the methods in terms of their complexity and performance to see whether it is realistic to implement them in practice. Until now, we have presented six different methods and established that they work. However, we have not discussed their complexity or whether they can be implemented in practice. We can take two different approaches to this problem. The first is that banks would like to test their Expected Shortfall predictions internally. This means that the method may be used in the comparison between different Expected Shortfall predictions or used to validate risk models. In this case, the choice of method to use may differ between different banks depending on the type of VaR model that is applied. The second approach we could take is that of discussing if there is a backtest that works well independently of the type of VaR model that a bank uses. That is, is it possible to find a more general way to backtest Expected Shortfall that would work for example in a regulatory framework. We will try to answer two different questions:

- Based on the performance and complexity of each method, which method(s) can be implemented in a bank for internal validation?
- Based on the performance and complexity of each method, is there a method that can be implemented in a general framework for backtesting Expected Shortfall?

The analysis will be based on the findings in the previous chapters. We will devote one section to each of the questions. The answers can be found at the end of each section and in the concluding section of the chapter. We will start by addressing the problem of internal risk validation.

6.1 Choosing a model internally

We are now going to investigate the first question of the chapter: based on the performance and complexity of each method, which method(s) can be implemented in a bank for internal validation?

6.1.1 Methodology

We will take three different aspects into account. Firstly, we will do the analysis based on the type of VaR model that a bank applies. From the definition of VaR in equation (2.1), we have that VaR is taken as a quantile from a probability distribution function P_t representing the returns. There are three ways in which the return distribution is normally calculated. The distribution can be empirical consisting of a number of historical returns, usually 250 days returns, in which the method is referred to as historical simulation. The distribution can be parametric assuming that returns follow a known distribution or it can be based on Monte Carlo simulations where the distribution is determined by simulating a large number of scenarios assuming some underlying behaviour of the assets. The way in which the distribution that defines VaR and Expected Shortfall is calculated is of great importance in the discussion of the type of backtesting method that should be used.

The second aspect that we will take into account are the properties of the backtesting methods. That is, the properties specified in table 3.1. We have that the methods are parametric or non-parametric and require simulations or do not require simulations. We will see that these properties are fundamental when we discuss complexity in relation to the type of risk model a bank uses.

Finally, we will also take the performance of each method into account to say something about which methods that are appropriate to implement. We saw for example that Acerbi and Szekely's second and third methods showed weak performance in terms of rejection compared to Acerbi and Szekely's first method even though the methods have similar complexity.

6.1.2 Backtesting with Monte Carlo

For a bank that uses Monte Carlo simulations for their VaR and Expected Shortfall calculations, determining the significance using the same simulation outcomes in the backtesting as in the risk calculations is straightforward. On the other hand, methods with parametric assumptions will not work well since the outcome of the entire return distribution rarely follows one of the standard distributions. This means that it is possible to choose

from Acerbi and Szekely's methods or the method by Emmer, Kratz, and Tasche that does not require simulations at all. Righi and Ceretta's method is not appropriate due to the need of a parametric assumption. We saw that the performance of Acerbi and Szekely's second and third method was weak compared to the other methods in terms of the large number of exceedances needed in the implementation. Furthermore, Acerbi and Szekely's first method showed a stronger performance relative to the method by Emmer, Kratz, and Tasche. Therefore, banks that use Monte Carlo simulations to calculate their risk measures are recommended to apply Acerbi and Szekely's first method in the backtesting of Expected Shortfall.

6.1.3 Backtesting with a parametric method

When a bank uses a parametric model to calculate risk, backtesting should be done against that same distribution. In this case, all methods are applicable. A method that requires simulations will add extra complexity but it is simple to simulate from a known parametric distribution. We saw that Wong's method performed well both in terms of accepting true predictions and rejecting false predictions. Furthermore, by using Wong's method, simulations can be avoided. However, the method only works if the underlying assumption is that of a normal distribution. While Righi and Ceretta's method showed stable performance, it has the drawback of needing both a parametric assumption and simulations which makes it the most complex method to implement in practice. In relation to its complexity, it does not perform better than any other method. Together with Wong's method, Acerbi and Szekely's first method showed the most stable performance. Therefore, a bank that uses a normal distribution in a parametric method is recommended to apply Wong's method in backtesting Expected Shortfall. For other distributional assumptions, Acerbi and Szekely's first method can be applied.

6.1.4 Backtesting with historical simulation

The third type of VaR calculation, historical simulation, is perhaps the most difficult one to give advise on. This is because parametric assumptions are usually wrong and simulations will be drawn from a distribution consisting only of 250 returns. This means that it will be difficult to say something about significance in the tail of the distribution. That means that we are left with the approximative method presented by Emmer, Kratz, and Tasche. We saw from the analysis above that the performance of the approximative method was good in terms of rejections but had some issues with accepting a true prediction as the number of exceedances became large. However, the method's simplicity makes it a good choice. Therefore, banks that use historical simulation to calculate risk are advised to implement Emmer,

Kratz, and Tasche's approximative method in the backtesting of Expected Shortfall. The next section will elaborate more on this particular method.

6.1.5 Conclusion - internal implementation

Based on the performance and on the complexity of the methods, three different methods are recommended for implementation in banks. The recommendations are shown in table 6.1.

Recommended methods	
Risk model	
Monte Carlo	Acerbi and Szekely's first method
Parametric - normal distribution	Wong's method
Parametric - other distribution	Acerbi and Szekely's first method
Historical simulation	Emmer, Kratz, and Tasche's method

Table 6.1: The table shows the methods that are recommended to use in practice depending on the type of risk model that a bank applies.

For banks using a Monte Carlo method to calculate risk we propose the use of Acerbi and Szekely's first method for backtesting Expected Shortfall. For a bank using a parametric model to calculate risk we propose the use of Wong's method if the parametric assumption is that of a normal distribution and Acerbi and Szekely's first method in case of a different distributional assumption. For banks using historical simulations, that is an empirical distribution, we could rule out five of the six proposed methods and found that Emmer, Kratz, and Tasche's approximative method is the only one appropriate to use.

6.2 The general method

We are now going to investigate the second question related to the implementation of a backtest: based on the performance and complexity of each method, is there a method that can be implemented in a general framework for backtesting Expected Shortfall?

In the previous section we discussed the potential methods that could be used for banks with different VaR models in their internal validation of risk measures. However, we saw that depending on the model, the recommendations changed and the method that is appropriate to use in one case is not possible to use in another case. Furthermore, compared to a standard VaR backtest, the methods recommended are quite complicated. The probability that one of them would be implemented in a regulatory framework is

therefore not very high. Since backtesting can have huge effects on capital charges it is important that the backtesting procedure can be understood by all parts of the organisation that affect the risk. Furthermore, increased complexity means higher probability that things go wrong in the implementation.

The only method that is simple enough and that can be applied to all types of VaR models is the one proposed by Emmer, Kratz, and Tasche. This section will explain why it is so difficult to find the confidence level of the test proposed by Emmer, Kratz, and Tasche and the type of improvements that need to be done in order for the method to have a better performance.

6.2.1 The difficulties in finding a correct confidence level

We saw from our analysis that the approximative method by Emmer, Kratz, and Tasche had similar performance as the other methods when it came to rejecting false Expected Shortfall predictions. However, in terms of accepting true Expected Shortfall predictions, the performance was much worse. The overall probability of accepting a true Expected Shortfall prediction was just below 78 %.

As above we assume that Expected Shortfall is approximated using five different VaR levels as

$$\mathrm{ES}_{2.5\%}(X) \approx \tag{6.1}$$
$$\frac{1}{5}[\mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{2.0\%}(X) + \mathrm{VaR}_{1.5\%}(X) + \mathrm{VaR}_{1.0\%}(X) + \mathrm{VaR}_{0.5\%}(X)].$$

We let Y_1 denote the number of exceedances during the last 250 days for VaR 97.5 %. Y_2 denotes the number of exceedances for VaR 98 % and similarly up to Y_5 that denotes the number of exceedances for VaR 99.5 %. We know the distribution of each of these different random variables. We can write

$$\begin{aligned} Y_1 &\sim \mathrm{Bin}(T, \alpha_1) \\ Y_2 &\sim \mathrm{Bin}(T, \alpha_2) \\ Y_3 &\sim \mathrm{Bin}(T, \alpha_3) \\ Y_4 &\sim \mathrm{Bin}(T, \alpha_4) \\ Y_5 &\sim \mathrm{Bin}(T, \alpha_5) \end{aligned}$$

Where in this case $T = 250$ and $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ =(0.025, 0.02, 0.015, 0.01, 0.005). Using table 3.2 above, we designed the Expected Shortfall backtest of the approximative method so that each VaR backtest was rejected with 95 % confidence independently of the outcome of the other VaR

backtests. However, we saw from our results that the overall confidence was much lower than this. This is because the random variables $(Y_1, Y_2, ..., Y_5)$ are not independent. In fact, for Y_2, the maximum number of exceedances are those given by Y_1. This means we can write the relationship between $(Y_1, Y_2, ..., Y_5)$ as

$$\begin{aligned} Y_1 &\sim \text{Bin}(T, \alpha_1) \\ Y_2 &\sim \text{Bin}(Y_1, \frac{\alpha_2}{\alpha_1}) \\ Y_3 &\sim \text{Bin}(Y_2, \frac{\alpha_3}{\alpha_2}) \\ Y_4 &\sim \text{Bin}(Y_3, \frac{\alpha_4}{\alpha_3}) \\ Y_5 &\sim \text{Bin}(Y_4, \frac{\alpha_5}{\alpha_4}) \end{aligned}$$

The implications of this is that if Y_1 is high then the probability that Y_2 is high has increased as well since the number of trials in the binomial distribution of Y_2 has increased. We can specify the probabilities of Y_2 both in terms of Y_1 and independently of Y_1. We assume that ten losses exceeded VaR 97.5 % during the last year, that is $Y_1 = 10$. We are now going to see how this information affects the probability of Y_2. We have that $Y_1 = 10$ and that (α_1, α_2)=(0.025, 0.02). We write

$$Y_2 \sim \text{Bin}(250, 0.02) \tag{6.2}$$

$$\tilde{Y}_2 \sim \text{Bin}(10, 0.8) \tag{6.3}$$

The cumulative probabilities of Y_2 and $\tilde{Y}_2$ are shown in table 6.2. We see that while we can reject exceedances above eight for Y_2 with 95 % confidence, the cumulative probability for nine or more exceedances for $\tilde{Y}_2$ is only 89.26 %. This means that given $Y_1 = 10$, we should not reject VaR 98 % if we observe that $Y_2 = 9$. However, this is what we do when we design the backtest of each VaR level independently of each other. This explains why the acceptance rate gets low when we use the approximative method in this setting.

In the same manner, if Y_2 is high then the probability that Y_3 is high will also increase. This describes why $(Y_1, Y_2, ..., Y_5)$ are not independent random variables and why this affects the rate of acceptance as we could see from the analysis above. The best solution to this problem would be if we could find the distribution of the mean of the random variables of exceedances $(Y_1, Y_2, ..., Y_5)$ given as

$$\bar{Y} = \frac{1}{5} \sum_{i=1}^{5} Y_i$$

Summing binomial random variables is straightforward if they are independent and the probabilities of each binomial is the same. However, when the probabilities are different and they are not independent, this becomes a complex statistical problem.

	Cumulative probabilities	
Number of exceedances	Y_2	$\tilde{Y}_2$
0	0.0064	0.0000
1	0.0391	0.0000
2	0.1221	0.0001
3	0.2622	0.0009
4	0.4387	0.0064
5	0.6160	0.0328
6	0.7637	0.1209
7	0.8687	0.3222
8	0.9339	0.6242
9	0.9696	0.8926
10	0.9872	1.0000

Table 6.2: The table compares the cumulative probabilities of the two distributions given by (6.2) and (6.3). The probabilities are calculated from (2.18).

We will not attempt to give a solution to this problem here. Instead we will devote the next section to trying to find a confidence level empirically for the approximative backtest using only two VaR quantiles.

6.2.2 Empirical confidence levels

For simplicity, we assume that Expected Shortfall is approximated using only two VaR levels. We say that we approximate Expected Shortfall with the 97.5 % VaR and the 99 % VaR. Hence, we write

$$\mathrm{ES}_{2.5\%}(X) \approx \frac{1}{2}[\mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{1.0\%}(X)] \tag{6.4}$$

We let Y_1 denote the number of exceedances for VaR 97.5 % and Y_2 denote the number of exceedances for VaR 99 % in a backtest with 250 returns. This means that we can write

$$Y_1 \sim \mathrm{Bin}(250,0.025),$$
$$Y_2 \sim \mathrm{Bin}(250,0.01).$$

or using their dependence as

$$Y_1 \sim \text{Bin}(250,0.025),$$
$$Y_2 \sim \text{Bin}(Y_1,0.40).$$

We are now going to use the approximative method to see what happens when the true prediction is rejected. That is, we want to see if we reject at the VaR 97.5 % level, at the VaR 99 % level or at both levels. We will start by assuming that Expected Shortfall is approximated by two VaR levels as above in equation (6.4). We assume that the predicted Expected Shortfall is from a standard normal distribution with a 97.5 % Expected Shortfall of 2.34. VaR 97.5 % is given by 1.96 and VaR 99 % is given by 2.33. We now simulate 250 returns representing profits and losses from the last year from a standard normal distribution. That is, we take the realised distribution to be the same as the predicted distribution. We do this 100 times and calculate the number of exceedances for the two VaR levels. We illustrate each outcome as a dot in figure 6.1.

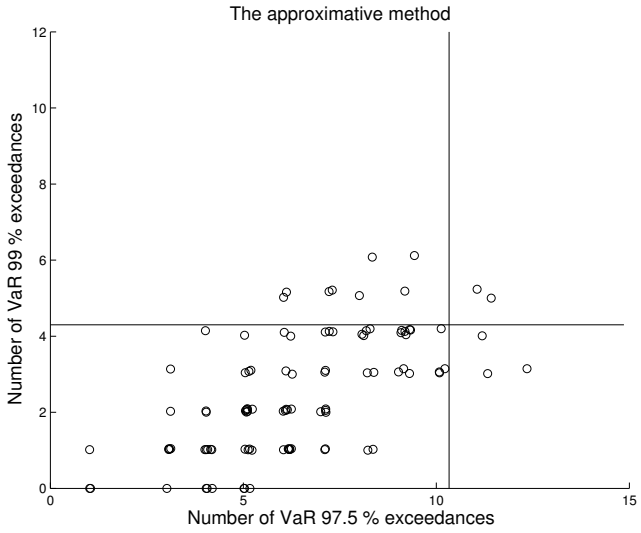

Figure 6.1: Shows the outcome of the approximative method using 100 simulations of 250 returns. Both the realised and predictive distribution is a standard normal distribution. The figure shows the number of exceedances at each of the two VaR levels given in equation (6.4). The limits for rejection are also shown in the figure. For VaR 97.5 %, exceedances above 10 are rejected while for VaR 99 % exceedances above 4 are rejected. The dots close to each other represent the same outcome but are shifted slightly from each other in order for us to see if there is one or several dots.

We see that for VaR 97.5 %, five dots are above the ten exceedances

needed to reject the prediction. For VaR 99 %, ten observations are above the four exceedances needed to reject the prediction at this level. Two outcomes are rejected both at the 97.5 % VaR level and at the 99 % VaR level. From this it seems like rejection happens more often for VaR 99 % than it does for VaR 97.5 % and not very often for the two levels jointly. However, 100 simulations are not enough to say something general about this.

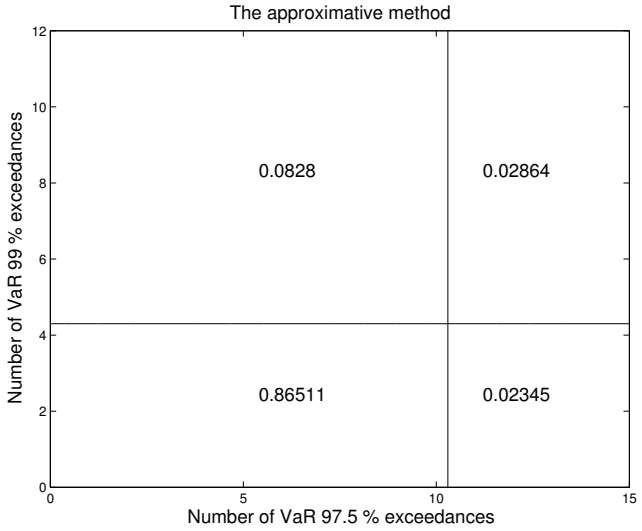

Figure 6.2: The figure shows the outcome of 10^5 simulations of 250 returns from the standard normal distribution, counting the number of exceedances for VaR 97.5 % and VaR 99 % according to Emmer, Kratz, and Tasche's approximative method. More than ten exceedances implies that the backtest of VaR 97.5 % is rejected and more than four exceedances for VaR 99 % implies that the predicted Expected Shortfall is rejected. The figure shows the proportion of all simulations that result in a backtest were VaR 97.5 % is rejected, VaR 99 % is rejected or both levels are rejected illustrated by the different areas of the figure.

Instead, we would like to know the overall probability of rejecting Expected Shortfall based on the fact that VaR 97.5 % is exceeded, that VaR 99 % is exceeded or that both the VaR levels are exceeded. We define four areas by the rectangles in figure 6.1. The first area is where the predicted Expected Shortfall is accepted, given for $x \leq 10$ and $y \leq 4$ in figure 6.1. The second area is where VaR 97.5 % is rejected but VaR 99 % is accepted, that is $x > 10$ and $y \leq 4$. The third area is where both VaR predictions are rejected by $x > 10$ and $y > 4$, and the fourth area where only VaR 99 % is rejected with $x \leq 10$ and $y > 4$. We do 10^5 simulations and determine the probability that the outcome ends up in one of the four different areas. We

illustrate the results in figure 6.2.

From figure 6.2 we see that the confidence level of this test is only 86.6 % as can be seen from the fact that this is the proportion of simulations that is below the rejection limit for both VaR levels. We also see that rejection happens more often at the VaR 99 % level than at the VaR 97.5 % level. The probability of rejecting only at the VaR 99 % level is above 8 %.

We generalise this assuming that Expected Shortfall is calculated as the mean of two VaR levels according to

$$\mathrm{ES}_{2.5\%}(X) \approx \frac{1}{2}[\mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{\alpha_2}(X)],$$

now changing α_2. We do the same analysis as above to see where rejection happens for a true prediction. We show the results in table 6.3.

	Probabilities - VaR levels			
VaR level	98.0 %	98.5 %	99.0 %	99.5 %
Rejected for VaR 97.5 %	0.0133	0.0183	0.0235	0.0288
Rejected for $\mathrm{VaR}_{\alpha_2}(X)$	0.0303	0.0543	0.0828	0.1122
Both levels rejected	0.0401	0.0324	0.0286	0.0233
Accepted	0.9164	0.8950	0.8651	0.8358

Table 6.3: The table shows the outcome of 10^5 simulations of 250 returns from the standard normal distribution, counting the number of exceedances for VaR 97.5 % and a different VaR level given by the table. More than ten exceedances implies that the backtest of VaR 97.5 % is rejected, more than eight exceedances means that VaR 98 % is rejected, more than six exceedances means that VaR 98.5 % is rejected, more than four exceedances implies that VaR 99 % is rejected and more than two exceedances for VaR 99.5 % implies rejection. The table illustrates the proportion of all simulations that result in a backtest were VaR 97.5 % is rejected, VaR at the higher level (lower α) is rejected or both levels are rejected as was illustrated by the different areas in figure 6.2.

We see that if we use two VaR levels to approximate Expected Shortfall, one of them being VaR 97.5 %, then the probability of acceptance decreases as we choose the other level further out in the tail. If we approximate VaR as the mean of VaR 97.5 % and VaR 99.5 % then we can expect that only 83 % of the true predictions will be accepted. On the other hand, if we choose to approximate Expected Shortfall with VaR 97.5 % and VaR 98 % we get a confidence level of almost 92 %.

6.2.3 Conclusion - general method

Due to the complexity of many of the methods presented in this thesis, the only method that is a candidate for a general backtest of Expected Shortfall is the approximative method presented by Emmer, Kratz, and Tasche. However, we saw from the previous chapters that this method fails in accepting true Expected Shortfall predictions with high confidence. There are other ways in which the different VaR backtests in the approximative method could be added together to give higher confidence. However, based on the dependency of the number of exceedances at each VaR level, we have seen that it is difficult to determine the confidence level of this method. In order to find a general method, more work has to be done.

6.3 Conclusion

We have found that many factors matter for the implementation of an Expected Shortfall backtest. Many of the methods presented in the previous chapters have high complexity in their need of parametric assumption or simulations to determine significance. We saw that depending on the type of model that is used to calculate risk, the parametric assumption or the need for simulations may not be a problem. We saw that several of the methods can be implemented in practice in a bank for internal validation of Expected Shortfall. The methods recommended for implementation were Wong's method, Acerbi and Szekely's first method and Emmer, Kratz, and Tasche's method.

We saw that it is difficult to find a general method that works independently of the type of risk model that is applied. A general method needs to be very simple. The approximative method by Emmer, Kratz, and Tasche is such a candidate due to its simplicity. However, in the way that we designed the backtesting for the approximative method, the confidence in accepting true Expected Shortfall predictions was too low. This means that more work has to be done in order to find the joint confidence of the independent VaR backtests for this method. If this issue is solved then it may be possible to find a good framework that backtests Expected Shortfall with high confidence that can be implemented only by small modifications of a VaR backtest.

Chapter 7

Conclusion

Expected Shortfall is a risk measure that is both subadditive and captures tail risk, solving two of the major issues related to VaR. However, while it is very easy to backtest VaR, there are still many issues outstanding on how Expected Shortfall should be backtested. In the Fundamental Review of the Trading Book, the Basel Committee propose a change in official risk measure from a 99 % VaR to a 97.5 % Expected Shortfall to take tail risk into account. However, since backtesting of Expected Shortfall is so difficult, they say that the official backtesting would still have to be done on VaR. Ever since Gneiting (2011) showed that Expected Shortfall lacks the mathematical property of elicitability, there has been a debate of whether it is possible to backtest Expected Shortfall at all.

The purpose of this thesis was to show that it is possible to backtest Expected Shortfall. This was done by presenting six methods from four different papers showing possible ways to design a backtest of Expected Shortfall without exploiting the property of elicitability. We evaluated the performance of the methods and could show that it is in fact possible to find backtests that not only work in theory but also in practice. In a final analysis, we looked at the properties and the complexity of the different methods to see whether it would be possible to implement them in a bank. On a more general level, we discussed the difficulties in finding an Expected Shortfall backtest with properties similar to a VaR backtest.

We will start by concluding what we discovered about elicitability and how a backtest can be designed without exploiting this property before we move on to conclusions of the analysis and the implications of this.

7.1 Finding methods without elicitability

Elicitability is a property such that a forecasting statistic can be expressed in terms of a scoring function. The scoring function can then be used to evaluate the forecasts against verified observations. As an example, we showed that the mean is elicitable through the scoring function squared errors $S(x, y) = (x - y)^2$. While VaR is elicitable, Expected Shortfall lacks this mathematical property. This implies that there is no scoring function that can be used in the backtesting of Expected Shortfall. However, it does not imply that Expected Shortfall is not backtestable. As long as it is possible to find a backtest that does not rely on the use of a scoring function, the lack of elicitability is irrelevant.

We presented the design of six different Expected Shortfall backtests that do not exploit the property of elicitability. The most simple example being the method proposed by Emmer et al. (2013) where Expected Shortfall is approximated by several VaR levels. We showed that we could write Expected Shortfall approximately as

$$\mathrm{ES}_{2.5\%}(X) \approx \tag{7.1}$$
$$\frac{1}{5}[\mathrm{VaR}_{2.5\%}(X) + \mathrm{VaR}_{2.0\%}(X) + \mathrm{VaR}_{1.5\%}(X) + \mathrm{VaR}_{1.0\%}(X) + \mathrm{VaR}_{0.5\%}(X)].$$

This means that backtesting of Expected Shortfall can be done by backtesting VaR at different levels. If all these VaR levels are found to be correct then it is possible to conclude that also Expected Shortfall must be correct.

7.2 Performance

We established that the six methods worked well as backtests by doing an analysis of their performance. We looked at three different aspects of the backtests. We wanted the backtests to accept true Expected Shortfall predictions, reject false Expected Shortfall predictions and to be able to do this using only a few VaR exceedances.

To investigate the behaviour of the methods we predicted a 97.5 % VaR and Expected Shortfall from a standard normal distribution. We simulated 250 realised returns from a Student's t distribution varying ν and σ to see when the Expected Shortfall prediction would be rejected. We illustrated the results as in figure 7.1. The black area in the figure means sure acceptance of the predicted Expected Shortfall while the white area means sure rejection. We see that when the realised Expected Shortfall comes from a Student's t distribution with many degrees of freedom and small σ, the backtest will surely accept the prediction as illustrated by the black area.

However, when the realised Expected Shortfall becomes large compared to the predicted value, either by a decrease in the degrees of freedom or an increase in σ, the probability that the prediction is accepted becomes small. This is illustrated by the white area in figure 7.1.

The implications from this is that when the observed Expected Shortfall from the last 250 days is large compared to the predicted Expected Shortfall, the backtest will reject the predicted Expected Shortfall with high confidence. However, when the observed Expected Shortfall is smaller than or of the same magnitude as the predicted Expected Shortfall, the backtest will accept the prediction.

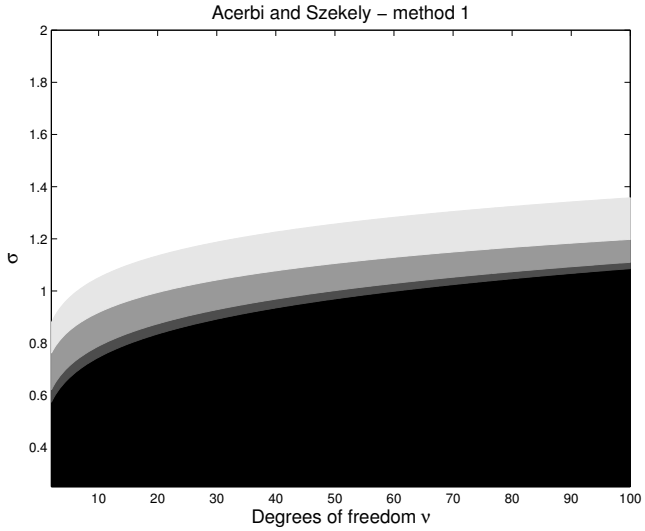

Figure 7.1: The figure illustrates the outcome of 27 903 simulated backtests. The predicted VaR and Expected Shortfall are assumed to be $N(0,1)$. 250 realised quantiles are then simulated with the tail distribution $t_\nu(0,\sigma)$ with parameters according to the figure. The simulations give a number of exceedances which can be used to calculate a realised Expected Shortfall. The figure shows in total five different areas where black illustrates a sure acceptance of Expected Shortfall in the backtest and white means a sure rejection of Expected Shortfall. A black area means 100 % acceptance rate. Dark grey has acceptance rate above 95 % but below 100 %. Medium grey has an acceptance rate somewhere between 95 % and 50 %. For light grey, the acceptance rate is lower than 50 % but higher than 5 %. The white area has an acceptance rate lower than 5 %.

We found that all methods showed this behaviour but that for some of the methods, the ability to reject was weak if the number of exceedances was

below nine. In terms of acceptance, we found that one of the methods had close to 100 % confidence in accepting true Expected Shortfall predictions. Overall, all methods managed to backtest Expected Shortfall.

We can conclude that it is possible to find several approaches in the backtesting of Expected Shortfall that all give very similar outcome in the sense that the backtests accept true Expected Shortfall predictions with high confidence and reject false Expected Shortfall predictions.

7.3 Implementation

A backtest that only works in theory is useless. Therefore, we also addressed the issue of implementing the backtests in practice. In this analysis we took both the performance and the complexity of the methods into account. Some recommendations were given on the type of backtesting method most appropriate to implement in practice. However, the recommendations were based on the type of risk model that the bank applies. We take Wong's method as an example. The method showed a good performance both in terms of accepting true Expected Shortfall predictions and rejecting false Expected Shortfall predictions for a small number of exceedances. However, it relies on the assumption that returns are normally distributed. This means that if a bank that uses a parametric model with a normal assumption to calculate risk wants to implement a backtest of Expected Shortfall, Wong's method is perfect. However, if the risk model is empirical or based on Monte Carlo simulations it is very unlikely that returns are normally distributed and Wong's method for backtesting Expected Shortfall is useless.

We found that for internal purposes, applying one of the methods presented in this thesis would be possible. However, the type of method to apply varied with the type of risk model the bank used. This means that backtesting Expected Shortfall is possible but there is no simple method that can always be applied.

7.4 The difficulties in designing a simple backtest

We found that it is difficult to find a general backtest of Expected Shortfall that is simple and works independently of the type of risk model that is used. That means that none of the methods presented here are simple and accurate enough to be implemented in a regulatory framework replacing the VaR backtest.

As a candidate, we presented the method by Emmer, Kratz, and Tasche that relies on the backtest of several VaR levels according to (7.1). We

introduced the random variables $(Y_1, Y_2, ..., Y_5)$ that represent the number of exceedances of the VaR levels in (7.1), where Y_1 denotes the number of exceedances for VaR 97.5 % and Y_5 the number of exceedances for VaR 99.5 %. We showed that we could write the random variables as

$$\begin{aligned} Y_1 &\sim \text{Bin}(T, \alpha_1) \\ Y_2 &\sim \text{Bin}(Y_1, \frac{\alpha_2}{\alpha_1}) \\ Y_3 &\sim \text{Bin}(Y_2, \frac{\alpha_3}{\alpha_2}) \\ Y_4 &\sim \text{Bin}(Y_3, \frac{\alpha_4}{\alpha_3}) \\ Y_5 &\sim \text{Bin}(Y_4, \frac{\alpha_5}{\alpha_4}) \end{aligned}$$

where in our case we would have $T = 250$ and $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$=(0.025, 0.02, 0.015, 0.01, 0.005).

This means that if we do a backtest for all VaR levels in (7.1), using the exceedances $(Y_1, Y_2, ..., Y_5)$, then we must account for their dependence in the confidence level of the test. We showed that this is a difficult task and a challenge in finding a general backtest of Expected Shortfall.

We investigated this dependency structure further by doing an analysis assuming that Expected Shortfall was approximated using only two VaR levels. That is, we assumed

$$\text{ES}_{2.5\%}(X) \approx \frac{1}{2}[\text{VaR}_{2.5\%}(X) + \text{VaR}_{\alpha_2}(X)].$$

By using a 95 % confidence level in each of the two VaR backtest independently, we found that using smaller values on α_2 decreased the total overall confidence in accepting true predictions. Assuming that Expected Shortfall is approximated with the mean of VaR 97.5 % and VaR 98 % gives a high confidence level. However, this is not a good approximation of Expected Shortfall. On the other hand, if we chose to approximate Expected Shortfall using the mean of VaR 97.5 % and VaR 99.5 %, then the Expected Shortfall approximation becomes more accurate but the overall confidence of the method decreased. We illustrate this in figure 7.2 where we show the outcome of two Expected Shortfall backtests. The left graph is based on a backtest of VaR 97.5 % and VaR 98 % while the second is based on a backtest of VaR 97.5 % and VaR 99.5 %. The backtest is rejected if VaR 97.5 % has more than ten exceedances or in the two different cases if VaR 98 % has more than eight exceedances or VaR 99.5 % more than two exceedances. When we use VaR 98 %, nine out of 100 observations are rejected. On the other hand, when using VaR 99.5 %, 19 observations out of 100 are rejected and the majority of them at the 99.5 % VaR level.

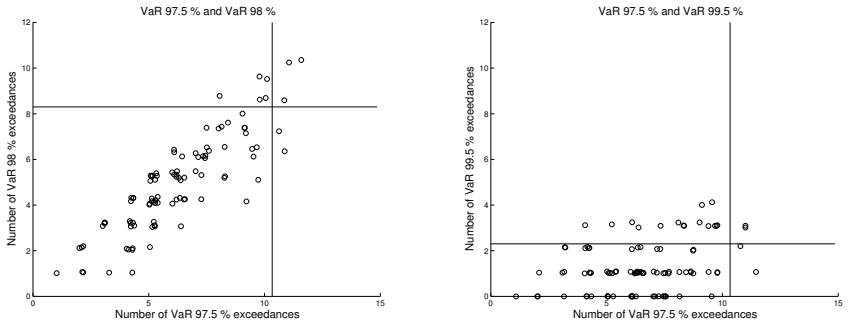

Figure 7.2: The figure describes the outcome of 100 backtests of the approximative method assuming that Expected Shortfall can be approximated with two different VaR levels. Expected Shortfall is rejected if the number of exceedances are more than what is given in the figure. That is more than ten for the 97.5 % VaR, more than eight for the 98 % VaR and more than two for the 99.5 % VaR.

This implies that for the approximative method there is a trade-off between approximating Expected Shortfall accurately by accounting for VaR levels far out in the tail and having high confidence in accepting true predictions. This is rather unfortunate since the purpose of Expected Shortfall is to capture tail risk and the approximative backtest only gives a high confidence when the tail risk is disregarded.

7.5 The backtestability of Expected Shortfall

After having defined the concept of elicitability, established that it is possible to find backtests that do not rely on the use of a scoring function and verified their performance we can conclude that it is in fact possible to backtest Expected Shortfall. We have presented examples of several methods that all take different approaches to solving the problem. The methods vary in complexity but nevertheless they all work as backtests of Expected Shortfall.

The implications are several. Firstly, if Expected Shortfall will become the official risk measure replacing VaR then backtesting of Expected Shortfall should also be mandatory. With this said, it is not at all an advise to replace the framework of VaR backtesting. The backtest of VaR is simple, well motivated by exact confidence levels and intuitive. This is not the case for many of the backtests presented here. However, for internal validation of Expected Shortfall the methods will work fine.

During the last years, the academic debate of Expected Shortfall backtesting has mainly been focused on whether it is possible to backtest Ex-

pected Shortfall at all. From this study, we would encourage more focus on providing methods in which backtesting can be done. We have seen that it is possible to backtest Expected Shortfall without exploiting the property of elicitability. We have also seen that it can be done in several different ways. However, we have not been able to provide a way that is accurate and general enough to be the preferred method in all situations. While we have concluded that backtesting of Expected Shortfall is possible we have also seen that many of the methods are much more complex than a normal VaR backtest. This means that in terms of backtesting, VaR is still the preferred risk measure. If supervisors want to continue the full transition from VaR to Expected Shortfall then there will probably be a need to find a framework that can backtest also Expected Shortfall in a similar manner as the VaR backtest does today. Based on the findings of this thesis, we believe that such a framework can be found with a starting point in the approximative method presented by Emmer, Kratz, and Tasche where Expected Shortfall is approximated with several VaR levels. Therefore, we encourage more work on these types of approximative methods.

Bibliography

C. Acerbi and B. Szekely. Backtesting expected shortfall. *Risk Magazine*, December 2014, 2014.

C. Acerbi, C. Nordio, and C. Sirtori. Expected shortfall as a tool for financial risk management. Working Paper, 2001.

P. Artzner, D. Heath, F. Delbaen, and JM Eber. Thinking coherently. *Risk*, 10:68–71, 1997.

P. Artzner, D. Heath, F. Delbaen, and JM Eber. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

J. Berkowitz. Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474, 2001.

L. Carver. Mooted var substitute cannot be back-tested, says top quant. Risk.net, 2013.

J. M. Chen. Measuring market risk under the basel accords: Var, stressed var, and expected shortfall. *Aestimaio, The IEB International Journal of Finance*, 8:184, 2014.

H. E. Daniels. Tail probability approximations. *International Statistical Review*, 55:37–48, 1987.

S. Emmer, M. Kratz, and D. Tasche. What is the best risk measure in practice? a comparison of standard measures. Working Paper, 2013.

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

H. Hult, F. Lindskog, O. Hammarlid, and C. J. Rehn. *Risk and portfolio analysis : principles and methods*. Springer, New York, 2012.

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

N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138. ACM, 2008.

R. Lugannani and S. Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, pages 475–490, 1980.

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

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management : concepts, techniques and tools*. Princeton Series in Finance, revised edition. edition, 2015.

K. H. Osband. *Providing Incentives for Better Cost Forecasting*. University of California, Berkeley, 1985.

P. Rappoport. A new approach: Average shortfall. Technical report, J.P. Morgan, 1993.

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

The Basel Committee. Fundamental review of the trading book: a revised market risk framework, 2013.

W. K. Wong. Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, 32(7):1404–1415, 2008.