Quantification of Model Risk

SOFIA SVED
Quantification of Model Risk

SOFIA SVED
Abstract

The awareness of model risk has increased due to the increased use of models to valuate financial instruments and their increasing complexity and regulators now require financial institutions to manage it. Despite this, there is still no industry or market standard when it comes to the quantification of model risk. The objective with this project is to find and implement a method that may be used to quantify model risk and evaluate it based on accuracy, efficiency and generalizability. Several approaches to model risk in the literature are explored in this thesis and it is concluded that existing methods are generally not efficient, accurate or generalizable. However, by combining two of the existing methods in the literature and using data on counterparty valuations, another method to quantify model risk can be constructed. This method is implemented and backtested and it is found to be accurate, general and more efficient than alternative approaches. Furthermore, this method may also serve in model validation as a mean to judge the quality of valuations and compare valuation models to each other.

One limitation of the method is that if there are few counterparties for a valuation model, say 1 or 2, the method used in this thesis is not suitable.
Kvantifiering av Modellrisk

Sammanfattning

Medvetenheten kring modellrisk har ökat på grund av ökad användning av modeller vid värdering av finansiella instrument samt deras ökande komplexitet. Dessutom begär nu regulatorer att institutioner ska beräkna samt redogöra för modellrisk. Trots detta finns ännu ingen bransch eller marknadsstandard när det kommer till hur modellrisk bör kvantifieras. Syftet med projektet är att hitta och implementera en metod som kan användas för att kvantifiera modellrisk samt utvärdera denna baserat på effektivitet, noggrannhet och generaliserbarhet. I den här uppsatsen har flera olika tillvägagångssätt i litteraturen för att kvantifiera modellrisk utvärderats med slutsatsen att befintliga metoder i allmänhet varken är effektiva, korrekta eller generaliserbara. Däremot, genom att kombinera två av de befintliga metoderna i litteraturen och använda data om motpartsvärderingar kan en annan metod för att kvantifiera modellrisken konstrueras. Denna metod implementeras och backtestas och den visar sig vara noggrann, generelariserbar och effektivare än de alternativa metoderna vilket var vad som eftersöktes. Vidare kan denna metod också tjäna i modellvalidering som ett medel för att bedöma hur väl värderingar från en viss modell överensstämmer med marknadens värderingar samt för att jämföra värderingsmodeller med varandra.

En begränsning som kan identifieras för denna metod är att om det finns få motparter till en värderingsmodell, såg 1 eller 2, är metoden som används i denna uppsats inte lämplig för att kvantifiera modellrisk.
Acknowledgements

I would like to thank my supervisor at KTH Royal Institute of Technology, Professor of mathematical statistics Boualem Djehiche for his support and guidance throughout the thesis. I would also like to thank the whole Model Validation & Quantitative Analysis team at Handelsbanken for proposing this interesting project and providing support, encouragement, data and resources to conduct it.

Stockholm, February 2019

Sofia Sved
# Contents

1 Introduction ........................................... 1
   1.1 Model Risk ........................................ 1
   1.2 Problem Description .............................. 2
   1.3 Data .............................................. 2
   1.4 Delimitations ..................................... 2
   1.5 Outline .......................................... 3

2 Literature Review ..................................... 4
   2.1 Benchmarking ...................................... 4
   2.2 The Worst-Case Approach ......................... 5
   2.3 The Bayesian Approach ............................ 7
   2.4 Hedging Error Approach to Model Risk .......... 11
   2.5 Regulations ....................................... 11
   2.6 Conclusion ...................................... 12

3 Risk Measurement ...................................... 15
   3.1 Risk Measures .................................... 15
      3.1.1 Value-at-Risk .................................. 16
      3.1.2 Expected Shortfall ............................ 16
   3.2 Risk Measurement Models ......................... 18
      3.2.1 Historical Simulation ......................... 18
      3.2.2 Parametric Models ............................ 20
      3.2.3 Extreme Value Theory ......................... 22

4 Evaluating Risk Measurement Models ............... 26
   4.1 Backtesting VaR .................................... 26
      4.1.1 Unconditional Coverage ....................... 27
      4.1.2 Independence .................................. 28
   4.2 Backtesting ES .................................... 29

5 Data Overview ......................................... 32
## Contents

6 Method

7 Results and Discussion

8 Summary and Conclusions

A Appendix

---

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>38</td>
</tr>
<tr>
<td>7 Results and Discussion</td>
<td>43</td>
</tr>
<tr>
<td>7.1 Risk Measurement Models</td>
<td>43</td>
</tr>
<tr>
<td>7.1.1 Historical Simulation</td>
<td>43</td>
</tr>
<tr>
<td>7.1.2 Parametric Models</td>
<td>45</td>
</tr>
<tr>
<td>7.1.3 Peaks over Threshold</td>
<td>47</td>
</tr>
<tr>
<td>7.1.4 Overview</td>
<td>52</td>
</tr>
<tr>
<td>7.2 Backtesting</td>
<td>52</td>
</tr>
<tr>
<td>7.2.1 Valuation Model 1</td>
<td>53</td>
</tr>
<tr>
<td>7.2.2 Valuation Model 2</td>
<td>55</td>
</tr>
<tr>
<td>7.2.3 Valuation Model 8</td>
<td>56</td>
</tr>
<tr>
<td>7.2.4 Overview</td>
<td>57</td>
</tr>
<tr>
<td>8 Summary and Conclusions</td>
<td>60</td>
</tr>
<tr>
<td>8.1 Further Research</td>
<td>60</td>
</tr>
<tr>
<td>A Appendix</td>
<td>65</td>
</tr>
<tr>
<td>A.1 Valuation Model 3</td>
<td>65</td>
</tr>
<tr>
<td>A.2 Valuation Model 4</td>
<td>68</td>
</tr>
<tr>
<td>A.3 Valuation Model 5</td>
<td>71</td>
</tr>
<tr>
<td>A.4 Valuation Model 6</td>
<td>74</td>
</tr>
<tr>
<td>A.5 Valuation Model 7</td>
<td>77</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Model Risk

Derman (1996) was one of the first to discuss model risk and how to validate models. The view is that the right model and the right value do not exist in practice, but wrong models and wrong values do exist and can be detected. We should commit ourselves to find models giving values as “little wrong” as possible, and then manage the residual unavoidable risk. This approach is thus about finding a model which is as realistic as possible. Derman identifies some assumptions and risks involved when using models to value financial securities, leading to model risk. In general, the following are sources of errors: incorrect model due to for example errors in input data or incorrect assumptions about relationships and dynamics, changed market conditions or correct model but inappropriate use. Derman (1996) further points out that one must remember that a model is a simplification of reality and thus a perfect output should not be expected, there will always be an error connected to its output.

Some years later, Rebonato (2003) introduced a new way of viewing model risk as “Model risk is the risk of occurrence of a significant difference between the mark-to-model value of a complex and/or illiquid instrument, and the price at which the same instrument is revealed to have traded in the market”. This means that market risk arises due to the range of different models or model calibrations which are used by market participants. This approach is sometimes referred to as the price approach while the view in Derman is referred to as the value approach. In this project we will take the price approach to model risk since the interest lies not in identifying the sources of the model risk and correct them but in identifying the risk of discrepancies between model and market valuations. This is more useful in practice while the value approach is of a more theoretical nature. As we will see later, this is also the view chosen by regulators.
1.2 Problem Description

The awareness of model risk has increased due to the increased use of models and their increasing complexity and regulators now require financial institutions to manage it. Despite this, there is still no industry or market standard when it comes to the definition and quantification of model risk. The aim of this thesis is to find and implement a suitable method that may be used to quantify model risk and evaluate it based on accuracy, efficiency and generalizability. The research question is therefore:

- What is the most suitable way to quantify model risk in terms of accuracy, efficiency and generalizability?

It is interesting for a financial institution to quantify model risk so that their exposure to risk is in line with their risk appetite. Furthermore, a measure of model risk may be used to compare different models in terms of, by taking the value approach, their conformance to the view of the market. This thesis has been done at Svenska Handelsbanken (SHB). For a risk averse institution such as SHB, a good valuation is a valuation that is in line with the view of the market which is how the quantification of model risk links to the quality of the valuation.

1.3 Data

When a financial institution sells a contract, it wants to hedge it in order to significantly lower the risk of selling the contract. This can be done in two different ways. Either they can try to replicate the contract by buying other instruments resulting in a payoff as close to the sold contract as possible, and how they replicate the contract depends on the model used to value it. Another alternative is to use back-to-back hedging which means that they buy the same contract as they sold from another market maker resulting in a perfect hedge. The latter is generally the case at SHB. From the back-to-back hedging the bid-ask spread of counterparties has been obtained and this paired with our valuation of the instrument and the valuation model which was used at SHB is what constitutes the data that we chose to use in this project. The data sets are linked to a representative selection of eight different valuation models used at SHB.

1.4 Delimitations

It is important to notice that plain vanilla options are not subject to large model risk. They are priced by the market and these prices are public, hence there is limited model risk involved since there is little uncertainty about the price and the valuation model. When pricing exotic options and structured products however, it becomes a more serious problem since market prices are not available and no industry standard exists when it comes to what model should be used for valuation. Hence only models valuing exotic options will be treated in this project.
Several methods to quantify model risk will be reviewed in this project but we delimit ourselves to implementing one method.

1.5 Outline

The thesis is structured as follows: As there is no industry standard when it comes to the quantification of model risk, Chapter 2 provides a literature review of methods to estimate model risk. This chapter will further account for the regulations concerning model risk and will be concluded with a discussion on what method is most suitable in this setting considering the regulations and our key evaluation points stated in the problem description. In Chapter 3, theory of risk measurement is presented as it is relevant for the application of, what was found in Chapter 2 to be, the most suitable approach to quantifying model risk. When using risk measurement models it is crucial that they are backtested in order to evaluate their accuracy. Hence different approaches to backtesting are reviewed in Chapter 4. Chapter 5 describes the data sets to be used as the characteristics of the data influence the details of the model risk method in ways which are described and discussed in Chapter 6. In Chapter 6 the model risk method that has been chosen is presented and motivated. Chapter 7 follows with results and also analysis and discussion of the same. Chapter 8 contains a summary and conclusion of the thesis followed by suggestions on further research.
Chapter 2

Literature Review

It was stated in the introduction that there is no industry standard on what method one should use to measure model risk, and hence, in this section some approaches to quantification of model risk found in the literature will be presented. We will continue by accounting for the regulations concerning model risk. In the light of the regulations and our key evaluation points, the suitability of the reviewed methods in this particular setting will be discussed.

2.1 Benchmarking

To measure model risk, outputs from possible models can be compared to a model which produce 'true' values of the outputs, a benchmark model. Clearly, to correctly specify the model risk such a benchmark model must be available which is not always the case, otherwise the approach would only lead to a relative estimation of the model risk. For some instruments good benchmark models are available but not used in practice as they are to complex or slow.

Hull and Suo (2002) defines model risk as the discrepancy relative to a benchmark model with respect to pricing and hedging in their article. They focus on the Model risk arising from the specification of the model, i.e. not parameter estimation risk for example. Their approach follows the steps below:

1. Assume that there is a 'true’ model which is more complex than the possible models and consider it’s outputs to be the true values of the exotic option.

2. Fit the true model to representative market data and in order to find the model parameters.

3. Compare the performances of the possible models with the 'true’ model in terms of pricing and hedging. It may be that market prices of vanilla options are needed to calibrate the possible models, in that case use vanilla values generated from the 'true’ model.
2.2 The Worst-Case Approach

Another approach to model risk is the worst-case, also called min-max approach. In this approach one considers a range of models and minimizes the loss encountered in the worst-case scenario. This approach to model risk has been developed by Cont (2006) and therefore we will now take a look at the methods proposed in Cont (2006) to quantify model risk.

A pricing model is defined as ”an arbitrage-free extrapolation rule extending the price system from market quoted instruments (liquid) to nonquoted ones (exotic)”. There are four questions to which the article wants to give quantitative answers:

1. How sensitive is the value of a given derivative to the choice of the pricing model?
2. Are some instruments more model-sensitive than others?
3. How large is the model uncertainty of a portfolio compared to it’s market risk?
4. Can one provision for model risk as one provision for market risk?

Even though all questions are interesting, The first question seems to be the most relevant for us. Further, it is stated that the article ”attempt to define a framework for quantifying model uncertainty in option pricing models”. Two methods are proposed, a convex measure based on a set of pricing models calibrated to the benchmark options and a coherent measure which relaxes the calibration requirement. Both methods lead to a decomposition of risk measures into a market value and a premium for model uncertainty.

Desirable qualities for a measure of model uncertainty are discussed, for example liquidly traded options should imply no model uncertainty and the model uncertainty of a portfolio should be in monetary units.

In order to understand the measures proposed by Cont, we first need to define the setting:

1. Benchmark instruments are options written on S with prices observed in the market. They have payoffs \((H_i)_{i \in I}\) and prices \(C_i^* \in [C_{i \text{bid}}, C_{i \text{ask}}]\).
2. A set of arbitrage-free pricing models \(Q\) consistent with the market prices of the benchmark instruments such that the asset price \((S_t)_{t \in [0,T]}\) is a martingale under each \(Q \in Q\) w.r.t its own history \(\mathcal{F}_t\) and \(\forall Q \in Q, \forall i \in I, E^Q[|H_i|] < \infty, E^Q[H_i] \in [C_{i \text{bid}}, C_{i \text{ask}}]\)

We may define

\[ \mathcal{C} = \{H \in \mathcal{F}_T, \sup_{Q \in Q} E^Q[|H|] < \infty\} \]
as the set of contingent claims with a well defined price in all models. We then consider a mapping \( \mu : \mathcal{C} \rightarrow [0, \infty] \) which represents the model uncertainty on the value of the contingent claim \( X \in \mathcal{C} \). A payoff \( X \) has a well-defined value in all the pricing models \( Q \in \mathbb{Q} \) and we may write the upper and lower price bounds in the following way:

\[
\bar{\pi} = \sup_{Q \in \mathbb{Q}} E_Q[X] \quad \underline{\pi} = \inf_{Q \in \mathbb{Q}} E_Q[X].
\]

Now, the coherent measure of model uncertainty on the value of a contingent claim \( X \) proposed by Cont (2006) is \( \mu_Q(X) = \bar{\pi} - \underline{\pi} \). Simply put it is the highest expectation of a payoff \( X \) under all pricing models minus the lowest expectation of a payoff \( X \) under all pricing models. Thus it is a span of the possible outcomes the models predicts for \( X \). The wider the span the more uncertainty linked to the value of \( X \) due to model uncertainty. If there is no model uncertainty \( \bar{\pi} = \underline{\pi} \) and each pricing models will assign a value of \( X \) in the interval \([\bar{\pi}, \underline{\pi}]\). Furthermore, for the benchmark derivatives, \( C^{bid}_i \leq \bar{\pi} \leq \underline{\pi} \leq C^{ask}_i \forall i \in I \). When valuing a derivative \( X \) at \( \pi_m(X) \), the model risk ratio \( MR(X) = \frac{\mu_Q(X)}{\pi_m(X)} \) may indicate if the model risk is a large part of the valuation and thus \( MR(X) \) may be used for model validation purposes.

In practice, the following would be the procedure to quantify model risk: the models in \( \mathbb{Q} \) are pricing models that are calibrated to the benchmark options, the focus should be only on the most interesting classes of models for the application (often market consensus on some standard models), we then compute the range \( \mu_Q(X) \) of values for \( X \) over all of the calibrated models.

It is pointed out that in existing works on model uncertainty it is assumed that all models agrees on the same space of possible scenarios and only differs in their probabilities for these scenarios. This is not assumed by the Cont approaches which is good since we have the same view. The disadvantage is that one would have to value the derivative of interest with all models, and not only the chosen one, in order to obtain a measure of the model risk.

We now continue with the proposed convex measure of model uncertainty. It is said to be an improvement in the way that we don’t need to calibrate various models to a set of benchmark instruments, hence it is more practical if we are dealing with a complex model or payoff of the benchmark derivatives. The setting is the same as before except for that we no longer require the pricing models \( Q \in \mathbb{Q} \) to produce the market prices of the benchmark instruments. Instead a penalization of the size of the valuation error is introduced:

\[
\pi^*(X) = \sup_{Q \in \mathbb{Q}} \{E_Q[X] - ||C^* - E_Q[H]||}\quad \pi_*(X) = \inf_{Q \in \mathbb{Q}} \{E_Q[X] + ||C^* - E_Q[H]||}\.
\]

The convex measure of model uncertainty is then, similarly to the coherent one, defined as

\[
\mu(X) = \pi^*(X) - \pi_*(X), \quad \forall X \in \mathcal{C}
\]

The practical procedure of computing this measure is the following: given a set of prices for the benchmark options, we first choose a pricing model \( Q_1 \) which can reproduce observed prices and
is easily calibrated to option prices. Such models are typically used for their ability to calibrate market prices easily but may not generate realistic dynamics for future market scenarios. We then add other pricing models $Q_2, \ldots$ with more realistic features but which may be more complex to calibrate.

In the end there is also a discussion about how to apply these measures as time evolves. It does not seem more sophisticated than simply recalibrating the models to the new benchmark prices. Furthermore, for the question about the measures sensitivity to market condition Cont has not yet explored this subject.

There are some obvious disadvantages when using this approach. First, one has to price derivatives using several different models and parameter sets which might take some time and it is not even sure that any traded values are in the resulting price range, as long as not all possible models are used. Furthermore, the results can be significantly biased by the introduction into the set of models of a model that produces, even occasionally, wild results.

### 2.3 The Bayesian Approach

Cont (2006) also reviews Bayesian model averaging as an approach to quantifying model risk. It is described as a way to incorporate model uncertainty into estimation procedures. However, many obstacles are identified by Cont when this method is applied on option pricing models. For example how to specify the priors on parameters and prior probabilities on the models considered, and thus it demand a lot from the user. In addition, it is computationally heavy due to the need for Markov Chain Monte Carlo (MCMC) algorithms. As the Bayesian approach was mentioned in Cont (2006), we will proceed by reviewing more literature on this subject. We will look at one book, Sekerke (2015), and one article, Droguette (2008), treating this subject.

Beginning with the book, it is said that "Bayesian methods employ basic tools of probability theory to assess alternative model specifications relative to each other. One can test whether the evidence in the data strongly favors a given model specification by computing a posterior model probability or a posterior odds ratio expressing the relative performance of any two models. In both cases, the quantity of primary interest is the marginal likelihood, or the probability of seeing the data conditional on the model.” We start by introducing the setting:

1. $x$ is the data set
2. $M_i$ is the candidate model
3. $\theta_i$ is the set of parameters belonging to model $i$

$p(M_i)$ is the prior model probabilities, $p(M_i|x)$ is the posterior model probabilities, $p(x|M_i)$ is the marginal likelihood of $M_i$ and $p(\theta_i|M_i)$ is the prior for the model parameters. Bayes’ theorem
yields that
\[ p(M_i|x) = \frac{p(x|M_i)p(M_i)}{p(x)} \propto p(x|M_i)p(M_i) \]
where the proportionality can be written because \( p(x) \) is independent of \( M_i \) and can be thought of as a normalizing factor to make sure that the area under the posterior density integrates to one. In a similar way we may write that
\[ p(\theta_i|x, M_i) = \frac{p(x|\theta_i, M_i)p(\theta_i|M_i)}{p(x|M_i)} \]
By integrating over \( \theta_i \) we obtain the integrated likelihood
\[ p(x|M_i) = \int_{\theta_i} p(x|\theta_i, M_i)p(\theta_i|M_i) \]
showing that evidence in favor of the model can come either from \( p(\theta_i|M_i) \) (ability of the model to capture useful prior information) or \( p(x|\theta_i, M_i) \) (how well the data follows the data-generating process). We now define Bayes factor, which can be used to compare two plausible models as
\[
\text{odds of event A occurring} = \frac{P(A)}{1 - P(A)} \Rightarrow p(M_i|x) = \frac{p(M_i|x)}{1 - p(M_i|x)} = \frac{p(M_i|x)p(M_i)}{p(M_j|x)p(M_j)} = BF \frac{p(M_i)}{p(M_j)}.
\]
The author says that Bayesian model averaging has the advantage over Bayes factor that it simplifies comparisons across sets of models and allow for multiple models to be assessed on the basis of prior and posterior probabilities. I will now briefly explain Bayesian model averaging. The law of total probabilities yields
\[ p(M_i|x) = \frac{p(x|M_i)p(M_i)}{\sum_M p(x|M_i)p(M_i)} \]
When the above has been computed for all \( M_i, p(\hat{y}) \), the weighted average forecast for \( y \) which is the model dependent quantity of interest can be obtained with
\[ p(\hat{y}) = \sum_M p(\hat{y}|x, M_i)p(M_i|x) \]
Which is not conditioned on any particular model, giving a range of expected values for the data. Note that this can also be seen as super positioning models according to a weighting scheme. In Montag and Persson (2015) the posterior model probabilities are approximated as functions of the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). The data set \( x \) typically consists of market prices for calibration and as we condition on a certain data set \( x \) instead of all possible, the results may differ if another data set is used. This is another
disadvantage of the Bayesian approach.

Moving on to Droguette (2008). It is stated that "This article describes a Bayesian methodology for the assessment of model uncertainties, where models are treated as sources of information on the unknown of interest. The general framework is then specialized for the case where models provide point estimates about a single-valued unknown, and where information about models are available in form of homogeneous and nonhomogeneous performance data (pairs of experimental observations and model predictions)." Next, the article talks about proposed approaches to quantitatively assess model uncertainty. Note that none of the approaches are applied to the derivatives pricing setting.

The framework of the Bayesian methodology for quantification of model uncertainty is then presented: We wish to assess $u_t$, the true value of the unknown $U$, and assume that we have a set of evidence $IM = u^*, D$ where $u^*$ is predictions from one or several models and $D$ is information about the models. What we want to achieve is a uncertainty range for $u_t$ which is characterized by a probability distribution $\pi(u)$. However, as this is a very difficult task we settle with $\pi(u|IM)$, the conditional probability distribution of $U$ given the evidence. This is quite similar to the min-max approach since we get a range of values but it takes it one step further giving a probability distribution of the values as well. It is said to be ‘a general Bayesian framework for dealing with model and parameter uncertainties in an integrated fashion’ where the available models are viewed as sources of information that can be used to estimate the unknown $U$. This can be done by applying Bayes’ theorem

$$\pi(u|IM) = \frac{L(u^*|D, u)\pi_0(u)}{\int_u L(u^*|D, u)\pi_0(u)du}$$

where $\pi_0(u)$ is the prior distribution and $L(u^*|D, u)$ is the likelihood function, the probability of observing evidence $u^*$ (model predictions) given that the true value is $u$ and $D$ is the information about the models. We now look into what kind of information we need to provide. The predictions of $U$ from models, $u^*$, can be of the forms probability distribution, point estimate or bounds of the unknown $U$ etc. Whereas information about the models, $D$, can be performance data which can be in the form of a set of model predictions and actual observations or other information. This article is using performance data. This framework is general in the sense that it allows for multiple dependent or independent models and different types of information from and about these models.

The question one poses after seeing the above equation is how do we formulate the likelihood function? This will be explored now. Consider the case of $n$ models $M_1, ..., M_n$ providing point estimates $x_1, ..., x_n$ about the quantity of interest $X$. Also, $n_1, ..., n_n$, number of performance data sets are available for the models, $D = D_1, ..., D_n$. We can write the posterior distribution of $X$
in the following way:

\[
\pi(x|x_1, \ldots, x_n, D) = \frac{L(x_1, \ldots, x_n, D|x) \pi_0(x)}{\int_x L(x_1, \ldots, x_n, D|x) \pi_0(x) dx}
\]

where

\[
L(x_1, \ldots, x_n, D|x) = L(x_1|x_2, \ldots, x_n, D, x) \ldots L(x_n|D, x) \times L(D_1|D_2, \ldots, D_n, x) \ldots L(D_n|x).
\]

It is assumed that the performance data sets are independent of \( x \) and hence we may write

\[
L(x_1, \ldots, x_n, D|x) = L(x_1|x_2, \ldots, x_n, D, x) \ldots L(x_n|D, x) \times L(D_1|D_2, \ldots, D_n) \ldots L(D_n).
\]

yielding

\[
\pi(x|x_1, \ldots, x_n, D) = \frac{L(x_1|x_2, \ldots, x_n, D, x) \ldots L(x_n|D, x) \pi_0(x)}{\int_x L(x_1|x_2, \ldots, x_n, D, x) \ldots L(x_n|D, x) \pi_0(x) dx}.
\]

Independence is now assumed between the models. This is not always the case and it is said that the case of multiple dependent models can be handled via a copula based approach but this was not discussed further in this article. The independence assumption leads to

\[
\pi(x|x_1, \ldots, x_n, D) = \frac{L(x_1|D, x) \ldots L(x_n|D, x) \pi_0(x)}{\int_x L(x_1|D, x) \ldots L(x_n|D, x) \pi_0(x) dx}.
\]

We now select a parametric form for the likelihood terms such that

\[
L(x_i|D_i, x) = \int_{\theta_i} L(x_i|\theta, x) \pi(\theta_i|D_i) d\theta_i.
\]

Here, \( L(x_i|\theta, x) \) is a parametric likelihood function characterized by a finite set of parameters \( \theta_i \) and to assess \( \theta_i \) we may use the past performance data of model \( M_i \) and Bayes’ theorem:

\[
\pi(\theta_i|D_i) = \frac{L(D_i|\theta, \theta_i) \pi_0(\theta_i)}{\int_{\theta_i} L(D_i|\theta, \theta_i) \pi_0(\theta_i) d\theta_i}
\]

Furthermore, different models can have different likelihood functions according to the nature and structure of the performance data sets. In this article the use of normal and log-normal distributions representing additive and multiplicative errors are explored. A basis for use of such distributions is the assumption that generally there are multiple factors contributing to the model errors and we often do not know the exact nature of these factors and their impact on model predictions. The combined effect leads to a behavior similar to what can be observed in experimental measurements. Finally, \( \pi(u|IM) \) yields a posterior uncertainty distribution of the quantity of interest \( U \) representing the model uncertainty and based on this distribution it is possible to quantify the model risk by computing a distribution based risk measure.

Homogeneous performance data set means that all performance data points are assumed to be realizations from the same distribution with parameter set \( \theta \). Both homogeneous and non-homogeneous performance data sets are considered in this article.
2.4 Hedging Error Approach to Model Risk

Detering and Packham (2014) propose to use Value-at-Risk (VaR) and Expected Shortfall (ES) type risk measures to quantify model risk as the potential losses arising from using misspecified models when pricing and hedging contingent claims. Their idea, is that since market risk is eliminated with a prefect hedge in a complete and frictionless market we may say that the observed losses (and profits) on a perfectly hedged position are due to use of a misspecified model when hedging. Thus they use hedging as a mean to separate losses from market risk and losses from model risk. The approach requires that we estimate the distribution of losses from hedging. The distributions of losses from hedging using different models are then probability weighted resulting in a unified loss distribution. The probability weights are computed as the probability of a model given the data which consists of vanilla option prices used to calibrate the model. They use AIC as a measure of goodness of fit as it penalizes models with many parameters (easier to calibrate correctly by simply adding more parameters). Given the unified loss distribution, value-at-risk (VaR) and expected shortfall (ES) are computed as measures of model risk. One advantage is that this allows for comparison between model risk and other kinds of risks such as market risk as VaR and ES are widely used. A disadvantage of the approach is discussed towards the end. The market model may be incomplete if for example jumps are included which would make the distinction between model and market loss distribution less clear.

2.5 Regulations

The EU commission has approved (October 2015) a draft Regulatory Technical Standards (RTS) for prudent valuation [RTS for prudent valuation (2015)] published by the European Banking Authority (EBA) which includes an article on model risk. The article says the following:

"Institutions shall estimate a model risk AVA for each valuation model (‘individual model risk AVA’) by considering valuation model risk which arises due to the potential existence of a range of different models or model calibrations, which are used by market participants, and the lack of a firm exit price for the specific product being valued. Institutions shall not consider valuation model risk which arises due to calibrations from market derived parameters, which shall be captured according to Article 9.”

AVA stands for Additional Valuation Adjustment and is the difference between the prudent value (i.e. quantile of price distribution) and the fair value (i.e. mean of price distribution) Furthermore two different valid approaches to calculating model risk AVA are described, the first one,

"Where possible, institutions shall calculate the model risk AVA by determining a range of plausible valuations produced from alternative appropriate modelling and calibration approaches. In this case, institutions shall estimate a point within the resulting range of valuations where they
are 90% confident they could exit the valuation exposure at that price or better.”

The second one is an expert-based approach with a list of requirements. Here we aim at finding a method for calculating the model risk \( \text{AVA} \) that falls under the first approach. I think that it is worth noting here that the regulations implies that plausible valuations and exit prices are equivalent which they are not necessarily. The exit prices proposed by different counterparties does not only depend on what model they used to value a derivative but also how they are positioned in the market which implies that exit prices differ from plausible valuations. Thus one has to make a choice here as the regulation is ambiguous, to estimate the 90% confidence point based on a range of plausible valuations or based on a range of plausible exit prices. Remembering that according to the Price approach to model risk, market risk arises due to the range of different models or model calibrations which are used by market participants, we choose to base it on a range of plausible valuations.

### 2.6 Conclusion

From the literature review above one may draw the conclusion that the quantification of model risk generally comprise valuating the same derivatives multiple times using a set of different models. In summary, the disadvantages with these methods are:

1. In some cases it might be difficult to decide which models to include if there is no market consensus on standard models in which case the set of models must be subjectively chosen by the modeller. This poses problems as the set of models have large influence on the model risk estimate.

2. In the case of many plausible model specifications these methods are time consuming.

3. By introducing a model which occasionally produces wild valuations the model risk measure can become biased.

These disadvantages are quite important, especially as accuracy, efficiency and generalizability are our key evaluation points. Therefore, if it is possible to estimate model risk without valuing the same derivative multiple times using different models it would be preferred. At the same time we want the method to estimate the 90% confidence point based on a range of plausible valuations thus satisfying the regulatory requirements in the section above. In the rest of this chapter we will explore if the methods used to estimate model risk in the literature or methods inspired by these can be applied in accordance with these aims.

### The Worst-Case Approach

The measures suggested by Cont fits well into the first approach in the regulations and is probably the most straight forward and easy to understand method that can be used to quantify model risk in accordance with the regulations. However, it is clearly not in line with our aims.
The Bayesian Approach

Bayes factor can be used to compare two models but it cannot be used as a measure of risk in the regulatory sense, if that is the aim as it is for us Bayesian model averaging seems more promising. Then it is possible to go from the \( p(\hat{y}) \) which gives us a range of expected values for the data to a model risk measure that is consistent with the regulations as the risk measure may then be computed as the 10% or 90% quantile. As pointed out by Cont there are several drawbacks with the Bayesian approach. The author of the book is mostly concerned by the problems related to what he calls “completeness of the model space”, warning that we might find ourselves with a model average which is either overwhelming to compute or with an incomplete model space. It becomes apparent that no methods exist that can be used to exhaust the universe of all possible model specifications. Additionally, the Bayesian approach requires, as many other approaches, that we set up all models we believe to be reasonable which again is an approach we consider to be erroneous. Therefore the Bayes approach from the book can be applied in accordance with the regulations but it does seem promising in the light of our evaluation points.

The framework of the article seems more promising as it can take performance data as input, which is something that is available to us. The general framework with multiple models and multiple performance data sets were presented above but our interest lies in the single model (the model we actually use) and single homogeneous performance data set (counterparty valuations). Section 3 and particularly the example application in section 3.3 describes this special case well. As the output from this Bayesian approach is an uncertainty distribution of the quantity of interest it would be possible to apply the risk measures used to quantify market risk mentioned above as they are distribution based. The issue is that the error terms \( E_i = x^*_i - x_i \), where \( x^*_i \) are the model’s estimate of \( x_i \) and \( x_i \) is the ’true’ observed value of the quantity \( X \) must, be assumed to be normally or log-normally distributed in order to stay in the same setting as the article. It might be possible to work with other distributions for the error terms but it would probably difficult the computation of the likelihood function and it might happen that we cannot find an appropriate distribution at all. However this approach lacks many of the drawbacks pointed out for the more classical Bayesian approach in the book. Thus, by viewing deviations obtained when using model \( X_i, i = 1, ..., 8 \) as our quantities of interest the Bayesian approach using performance data can be applied and produce a model risk measure in line with the regulations without requiring computation of many different models which complies with our aims.

The Benchmark Approach

The benchmark method do not fit into the first approach in the regulations, actually it does not even fit into the price approach described in 1.2 that we have chosen to take but leans more towards the value approach. As it is described in 2.1 it does not yield any kind of range of valuations. It is nevertheless possible to use some of the ideas from the benchmark approach and apply them to our setting in which we have access to data consisting of the valuations of other
market participants.

It is desirable to valuate instruments such that the resulting valuations are close to the middle of other market participants valuations. Therefore, as we have several counterparties for each model (min. 4), their valuations may be viewed as benchmarks and we can compare our valuations to theirs by computing the deviations between our valuations and theirs.

Recall that the regulations talk about a range of valuations, while we only have one real observation of counterparty valuations for each derivative and not a range. However, by assuming that all deviations observed for the model are plausible deviations for any derivative valued using this model we can create a sample space of plausible valuations. This assumption should not be too strong as the data includes several instruments with similar characteristics (as they are valuated using the same model) and several different counterparties. If we would have for example only one counterparty this would not be a good assumption. The sample space of plausible valuations could be obtained in the following way.

\[
(MID^i_{HB})_k - (MID^i_{CP})_k = (\text{deviation}^i)_k, \quad i = 1, ..., n, \quad k = 1, ..., 8
\]

\[
\Rightarrow (PV^j_{CP})_k = (MID^{n+1}_{HB})_k - (\text{deviation}^i)_k, \quad j = 1, ..., n
\]

where \(MID_{HB}\) refers to our mid valuation, \(MID_{CP}\) refers to the counterparty mid valuation and \(PV_{CP}\) refers to a plausible counterparty valuation. The index \(i\) refers to the observation number in the sample and \(k\) refers to the model used at HB for valuation of the instrument. Hence given a new valuation \(n + 1\) using model \(k\), \((PV^j_{CP})_k\) constitutes the range of size \(n\) of plausible exit prices at \(n + 1\). Note that if one can estimate the 90% confidence point of the deviations one has also estimated the 90% confidence point of the counterparty valuations, i.e. the plausible valuations. Thus it would be possible to use this benchmark inspired method to quantify model risk inside the regulations framework that satisfies our aims.

**Hedging Error Approach**

The approach by Detering and Packham (2014) is quite far from our setting as they use replication to hedge and values a contingent claim using several different models in order to quantify model risk. However there are parts of their method that can be of use to us. The benchmark inspired approach described above results in a range of deviations which can be viewed as a distribution of deviations for which we wish to find the 90% confidence point. Detering and Packham (2014) also have a distribution (loss from hedging) for which they compute the model risk using risk measures such as VaR and ES which are often used when computing market risk.
Chapter 3

Risk Measurement

In this section different measures of risk will be presented followed by how these may be estimated using different Risk Measurement Models (RMMs). Most methods and models presented here are often used when measuring market risk and hence literature on Market risk has been studied. In that kind of literature the distribution of interest is a profit and loss (P/L) distribution but it is not an issue that our interest lies in another kind of distribution than the P/L as the RMMs can be used for a wide range of different distributions with different properties.

3.1 Risk Measures

As our RMMs will be based on the distribution of a stochastic variable X we will in this section only consider risk measures $\rho$ for which $\rho(X)$ depends exclusively on the distribution of X. More formally a risk measure $\rho$ is distribution-based if $\rho(X) = \rho(Y)$ when $F_X = F_Y$.

Artzner et. al. (1999) proposed the theory of coherent risk measures which for the first time specified what a measure of financial risk should satisfy mathematically which resulted in the following set of ‘axioms of coherency’. Let $X_1$ and $X_2$ represent two stochastic variables and let $\rho$ be a measure of risk, $\rho$ is said to be coherent if the following properties are satisfied:

1. Monotonicity, $X_1 \leq X_2 \implies \rho(X_2) \leq \rho(X_1)$
2. Subadditivity, $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$
3. Positive homogeneity, $\rho(hX_1) = h\rho(X_1)$ for $h \geq 0$
4. Translational invariance, $\rho(X_1 + n) = \rho(X_1) - n$ for some certain amount $n$.

Other natural requirements for good risk measures are

1. Convexity, $\rho(\lambda X_1 + (1 - \lambda)X_2) \leq \lambda \rho(X_1) + (1 - \lambda)\rho(X_2)$ for any real number $\lambda \in [0, 1]$
2. Normalization, $\rho(0) = 0$
A risk measure is said to be convex if it satisfies the properties monotonicity, convexity and translation invariance. Furthermore, if a risk measure is coherent it is also convex but a convex risk measure is not necessarily coherent. [Hult et al. (2012)]

### 3.1.1 Value-at-Risk

Two commonly used distribution-based risk measures are Value-at-Risk (VaR) and Expected Shortfall (ES). Given a distribution, both measures focus on the worst \((100)p\%\) of the distribution in question where \(p \in [0, 1]\). If we want to have \(\alpha\) as the confidence level, \(p = 1 - \alpha\) and VaR of \(X\) at confidence level \(\alpha\) is

\[
VaR_p(X) = F_X^{-1}(1 - p).
\]

Hence it is simply the \(\alpha\)-quantile of the distribution meaning that losses greater than \(VaR_p(X)\) are suffered only with the small specified probability \(p\). This is the advantage with VaR as it means that we can estimate it for any distribution function. Values often used for \(p\) are \(p = 0.05, 0.01\) but here we will use \(p = 0.1\) as the 90\% confidence point is sought for. [Hult et al. (2012)]

VaR has the disadvantage of not being a coherent risk measure as it does not satisfy the subadditivity condition. However, according to Artzner et al. (1999), VaR is actually subadditive if we restrict the distribution to be elliptic (i.e. normal, \(t\)-dist etc). Although this information is not very useful as the distribution of financial data is rarely elliptic.

### 3.1.2 Expected Shortfall

VaR is a good measure of risk that work perfectly well in light of the regulations while ES may seem to be less in line with the regulations. However, Basel III proposes to use ES as a measure of risk when quantifying market risk instead of VaR due to its ability to better reflect tail risk. Hence ES will, in addition to VaR, be computed in this project at the levels 90\%. Note that these values cannot be directly compared to the VaR values at the same level as they say different things, they provide different interpretations of risk. ES is sometimes called other names in the literature such as Average VaR (AVaR), Conditional VaR (CVaR), Tail VaR (TVaR), or Tail Conditional Expectation (TCE). It is based on the \(\alpha\)-quantiles of the distribution of interest, more specifically it is the weighted average of the quantiles or the average of the worst \(100(p)\%\) of losses:

\[
ES_p(X) = \frac{1}{p} \int_0^p VaR_u(X) du.
\]

According to Lemma 2.13 in McNeil (2015), for an integrable distribution of \(X\) with continuous distribution function \(F_X\) and for any \(\alpha \in (0, 1)\) we have

\[
ES_p(X) = E[X \mid X \geq VaR_p(X)] du.
\]
Therefore, another way to describe it is that ES is the expected loss given that the loss falls in the worst (100)p% of the distribution. From Lemma 2.13 we can see that it is always the case that $ES_p(X) \geq VaR_p(X)$ as can be seen in Figure 3.1.

![Figure 3.1: VaR and ES for different values of $\alpha$.](image)

The main reason to use ES instead of VaR is that ES succeeds to reflect tail risk better than VaR in the way that it tells us what we can expect if a tail event occurs. VaR only tells us what we can lose if a tail event does not occur which is not optimal as it implies that two distributions can have the same VaR value but have widely different tails which means that they are not exposed to the same amount of risk (measuring market risk). Example 2.16 in McNeil et. al. (2015) further describes this. Furthermore, ES is a coherent measure of risk. This is another reason to why ES is considered to dominate VaR as a risk measure since the subadditivity property implicates that risks can be added together without underestimating the combined risk. ES is also convex which has certain advantages when it comes to optimisation. VaR is, as mentioned, not a coherent measure but it has the advantage over ES of being elicitable. The lack och elicibility is by some believed to prevent strict backtesting meaning that it is not possible to estimate the performance of ES on historical data. Although, as we will see later this is debatable and it would be more correct to say that the elicitation makes VaR easier to backtest than ES. [Hult et. al. (2012)]
3.2 Risk Measurement Models

3.2.1 Historical Simulation

The Historical Simulation (HS) method is a non-parametric method which allows us to not make any assumptions regarding the distribution of the underlying data, however when using this method we assume that the historical data is representative of the future outcomes. Hult et al. (2012) argues that while this is not necessarily a good assumption for historical stock prices $S_{-n}, ..., S_0$, it may be assumed for the historical returns $R_{-k} = S_{-k+1}/S_{-k}$ for $k = n - 1, ..., 1$. In the same way we can argue that this is a good assumption for the deviations even though it is not for the counterparty valuations. Some advantages when using the HS method are that it is simple to understand and implement and takes the fat tails of the loss distribution, commonly seen in financial data, into account.

We consider observations $x_1, ..., x_n$ from independent and identically distributed random variables $X_1, ..., X_n$ with a common unknown distribution function $F$. As $F$ is unknown it is not possible to compute quantities related to it such as mean, variance or quantiles, but given observations from it we can approximate $F$ as

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{x_k \leq x}$$

which is called the empirical distribution function (EDF) of $x_1, ..., x_n$. By the law of large numbers it holds that $F_n(x) \rightarrow F(x)$ a.s. as $n \rightarrow \infty$. Given the EDF it is possible to compute for example empirical quantiles which is of interest in this setting. The empirical quantile function is given as

$$F_n^{-1}(p) = \min\{x : F_n(x) \geq p\}.$$ 

Furthermore, it can be shown that if we order the samples such that $x_1 \geq ... \geq x_n$,

$$F_n^{-1}(p) = x_{[n(1-p)]+1}.$$ 

where $[ \ ]$ is the floor operator. This means that if $X$ is our distribution then $VaR_p(X) = F_X^{-1}(1 - p)$ and thus the empirical VaR is

$$\widehat{VaR}_p(X) = x_{[np]+1}$$

as $\widehat{VaR}_p(x)$ is simply the empirical $\alpha$-quantile of $X$. Since

$$ES_p(X) = \frac{1}{p} \int_0^p VaR_u(X)du,$$
the empirical ES estimator can easily be obtained by switching $VaR_p(X)$ to its empirical estimator $\hat{VaR}_p(x)$. This yields

$$\hat{ES}_p(X) = \frac{1}{p} \int_0^p x_{[n\mu]+1} du = \frac{1}{p} \left( \sum_{k=1}^{[np]} \frac{x_k}{n} + \left( p - \frac{[np]}{n} \right) x_{[np]+1} \right).$$

[Dowd (2005)]

**Non-Parametric Bootstrap**

Non-parametric bootstrapping can be used to improve the basic HS method as it will often result in more accurate estimates of VaR and ES but it can also be used to create confidence intervals for the risk estimates. It is a method in statistics belonging to the category resampling methods. Assume that data $y$ has been collected of some r.v. $Y$ with distribution $P_* \in P$ where $P$ is some family of distributions, general and possibly non-parametric. We wish to make inference about some property (estimand) $\tau = \tau(P_*)$, the estimand $\tau$ is estimated using some statistic $t(y)$. This $t(y)$ is an observation of the r.v. $t(Y)$ so if the data collection was repeated we would probably obtain another value of the estimator. To assess the uncertainty of the estimator we need to analyze the distribution $F_\Delta$ of $\Delta(Y)$ as the error $\Delta(y) = \tau - t(y)$ is a realization of the r.v. $\Delta(Y) = \tau - t(Y)$.

We assign an Empirical Distribution (ED) $\hat{P}_*$ to the data by giving equal weights $\left( \frac{1}{n} \right)$ to each observation $y_1, ..., y_n$, hence the ED function (EDF) associated with the data $y$ is $\hat{F}_n(z) = \hat{P}_*(Z \leq z) = \frac{1}{n} \sum_{i=1}^n 1_{y_i \leq z} = \text{fraction of } y_i \text{'s that are less than } z$. By the LLN, as $n \to \infty$, $\hat{F}_n(z) \to F_n(z)$.

The bootstrap algorithm:

1. Simulate $B$ new data sets $Y^*_b, b \in \{1, ..., B\}$ from $\hat{P}_*$, where $Y^*_b$ has the size of $y$ by drawing with replacement among the values $y_1, ..., y_n$.

2. calculate the values $t(Y^*_b)$ of the estimator.

3. By setting $\Delta^*_b = \hat{\tau} - t(Y^*_b)$ we obtain values being approximately distributed according to the error distribution which can be used for uncertainty analysis.

The only assumption for the bootstrap algorithm is that the samples are IID. In our case $t$ would be the empirical VaR or ES, thus more accurate estimates of these can be obtained as

$$\hat{VaR}_p(X) = \frac{1}{B} \sum_{b=1}^B \hat{VaR}_p(Y^*_b).$$
\[ \hat{ES}_p(X) = \frac{1}{B} \sum_{b=1}^{B} \hat{ES}_p(Y^*_b). \]

If we are also interested in confidence intervals for \( \hat{VaR}_p(X) \) or \( \hat{ES}_p(X) \) on a level \( q \), Hult et al.(2012) suggests that we compute them as

\[
I_{\hat{VaR},q} = [\hat{VaR}_p(X) + \Delta^*_B(\lceil B(1+q)/2 \rceil +1), \hat{VaR}_p(X) + \Delta^*_B(\lceil B(1-q)/2 \rceil +1)]
\]

\[
I_{\hat{ES},q} = [\hat{ES}_p(X) + \Delta^*_B(\lceil B(1+q)/2 \rceil +1), \hat{ES}_p(X) + \Delta^*_B(\lceil B(1-q)/2 \rceil +1)]
\]

where \( \Delta^*_1 \geq \ldots \geq \Delta^*_B \). Hult et al. (2012) further conclude that the bootstrap method performs rather well when used to compute confidence intervals (very similar to the exact confidence interval). Hence, if bootstrapping is already used to improve estimates it is the most convenient method to obtain the confidence intervals. In order to choose the necessary number of simulations, VaR and ES can be plotted together with their confidence interval and observe where it converge.[Skold (2005)]

### 3.2.2 Parametric Models

The main reason to choose the parametric method is that it provide us with a lot of information about risk measures and thus they will generally improve the estimates of the risk measures, given that the parametric model is correct, that is. Moreover, parametric methods are easy to practically implement but the disadvantage is that the risk measure estimates obtained are sensitive to errors due to inaccurate distribution assumptions (Dowd (2005)). Example 8.5 in Hult et. al. (2012) illustrates this by fitting three models to log-return data: an empirical distribution, a Normal distribution, and a Student’s \( t \)-distribution. The VaR values turn out quite similar while the ES values shows much larger differences between the models. This seems reasonable as ES depend to a higher degree on the shape of the tail in regions further out than VaR. The parameters of the distribution are estimated from past observations which means that this method also relies on the assumption that the past is a good approximation for the future.

The first step is to choose a parametric family \((F_{\theta})_{\theta \in \Theta}\) for the distribution. This may be done by plotting the data in various ways, for example using histograms and QQ plots. The second step is to estimate the parameters for the chosen parametric family. By solving for the log likelihood function \( \Psi(x, \theta) = \log f_{\theta}(x_i) \)

\[
\hat{\theta}(x) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \Psi(x_i, \theta)
\]

the maximum likelihood estimator \( \hat{\theta}(x) \) is obtained which is a natural choice of estimation technique as it maximizes the probability of the observed data. When using MLE it is assumed that
the samples are approximately IID. An alternative to the MLE is the least-squares estimation (LSE) minimizing the sum of the squares between the empirical quantiles and the parametric distribution quantiles. This is a particularly suitable approach if \( F_{\theta}(p) \) is a linear function of the parameter vector \( \theta = (\theta_0, ..., \theta_d)^T \) as this is a linear regression problem. Such is the case when the normal distribution is the parametric distribution and in the case of normal distribution the LSE will be the same as the MLE for a normal distribution. We will not search further for other methods to estimate the parameters as we will only use the normal and the \( t \)-distribution for which LS and ML methods are theoretically ideal and straightforward to implement. After estimating \( \hat{\theta}(x) \) it is necessary to validate the model by evaluating the fit to the historic data since this method is based on assumptions made by the modeler. This can be done using for example Q-Q plots which will be approximately linear if the data is generated by the proposed distribution. [Hult et. al. (2012)]

Two commonly used distributions are the normal distribution and the student’s \( t \) distribution, both elliptic distributions. The elliptic distribution generalize the normal distribution to symmetric distributions with heavier (ex \( t \)-distribution) or shorter tails than the normal distribution. Using these distributions comes with certain advantages, ES formulas are known only for a limited number of parametric distributions (e.g., elliptical, including normal, and generalized Pareto distributions). These and VaR formulas will now be presented for the normal and \( t \)-distribution starting with the normal. Given that \( X \) in normally distributed with scale and location parameters \( \mu \) and \( \sigma^2 \), then

\[
VaR_p(X) = \mu + \sigma \Phi^{-1}(1-p)
\]

where \( \Phi \) is the cumulative distribution function (cdf) of \( Z \). In the same way, by noting that \( X \) which is \( t \)-distributed with scale and location parameters \( \mu \) and \( \sigma^2 \) can be expressed as \( X = \mu \nu + \sigma Z \) where \( Z \) is standard \( t \)-distributed with degrees of freedom \( \nu > 1 \),

\[
VaR_p(X) = \mu + \sigma t_{\nu}^{-1}(1-p)
\]

where \( t_{\nu} \) is the cdf of \( Z \). Continuing with the expressions for calculating ES. Given the general expression for \( ES_p(X) \) and the formulas for VaR given above, it can be shown that, if \( X \) is assumed to be normally distributed as before then

\[
ES_p(X) = \mu + \frac{\sigma}{p} \phi(\Phi^{-1}(1-p))
\]

where \( \phi \) is the standard normal probability density function (pdf). If \( X \) is assumed to be \( t \)-distributed as before then

\[
ES_p(X) = \mu + \frac{\sigma}{p} \left( \frac{\nu + (t_{\nu}^{-1}(p))^2}{\nu - 1} \right) g_{\nu}(t_{\nu}^{-1}(1-p))
\]

where \( g_{\nu} \) is the pdf of \( Z \).
As mentioned earlier there are no exact formulas to compute ES for all distribution families, this is partly why we assume normal and \( t \)-distribution as parametric families. However, Dowd (2005) describes a way to approximate ES if it turns out that the data fits well to such distribution family. As said earlier, one of the advantages of VaR is that it can be computed at any level for any distribution. Also remember that ES is a probability-weighted average of tail losses. This means that ES can be estimated as an average of the VaRs in the tail. This can be done by dividing the tail into \( n \) (preferably large) parts with the same probability mass, then estimate VaR for each part. ES may now be approximated as the average of those VaRs. For example, if \( ES_{0.05} \) is sought for, we can calculate \( \text{VaR}_{0.045}, \text{VaR}_{0.04}, \text{VaR}_{0.035}, \text{VaR}_{0.03}, \text{VaR}_{0.025}, \text{VaR}_{0.02}, \text{VaR}_{0.015}, \text{VaR}_{0.01} \) and \( \text{VaR}_{0.005} \) and approximate \( ES_{0.05} \) as the average of these VaRs.

### 3.2.3 Extreme Value Theory

Extreme value theory (EVT) is a part of probability theory that treats extreme events. If the data in question has fat tails that ordinary probability distributions are unable to account for, EVT may yield better results as it provides theory to model extreme events which is exactly what the tails are. There are two different models for extreme events in EVT, the Block Maxima models which allows us to model the largest observations from large and identically distributed samples and the Peaks Over Threshold models (POT) that models observations exceeding some high threshold. According to McNeil et. al. (2015), POT is more used in practice as it more efficiently use the extreme event data which is often limited.

#### Block Maxima Method

The Generalized Extreme Value (GEV) distribution is necessary in the theory of extreme events, the distribution function of the standard GEV distribution is

\[
H_\xi(x) = \begin{cases} 
\exp\left(-\left(1 + \xi x\right)^{-1/\xi}\right) & \text{if } \xi \neq 0 \\
\exp(-e^{-x}) & \text{if } \xi = 0 
\end{cases}
\]

where \( 1 + \xi x > 0 \). To obtain a three parameter family with location parameter \( \mu \in \mathbb{R} \) and scale parameter \( \sigma > 0 \), \( H_{\xi,\mu,\sigma}(x) = H_\xi((x - \mu)/\sigma) \). The parameter \( \xi \) is called the shape parameter and characterizes a type of distribution belonging to the family of extreme value distributions.

\[
\begin{align*}
\xi > 0 & \quad \implies \text{Fréchet distribution} \\
\xi = 0 & \quad \implies \text{Gumbel distribution} \\
\xi < 0 & \quad \implies \text{Weibull distribution.}
\end{align*}
\]

Consider a sequence of IID r.v’s \( X_1, ..., X_n \) where \( n \geq 1 \) with finite variance from the loss distribution. According to the CLT, sums \( S_n = X_1 + ... + X_n \) normalized as \( (S_n - a_n)/b_n \) where \( a_n = nE(X_1) \) and \( b_n = \sqrt{n \text{var}(X_1)} \) converge in distribution to the standard normal distribution when \( n \to \infty \). In a similar way, EVT treats the convergence of maxima instead of
convergence of sums as CLT does. Denote by $M_n = \max(X_1, \ldots, X_n)$ the n-block maximum, the limiting distributions for appropriately normalized maxima as $n \to \infty$ belong to the GEV family. As

$$P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = F^n(x),$$

according to the CLT there exist real constants $d_n$ and $c_n > 0$ for all $n$ such that

$$\lim_{n \to \infty} P((M_n - d_n)/c_n \leq x) = \lim_{n \to \infty} F^n(c_n x + d_n) = H(x)$$

for some non-degenerate distribution function $H(x)$. If the equation above holds then $F$ is said to be in the maximum domain of attraction of $H$, $F \in MDA(H)$. Furthermore, the Fisher-Tippett-Gnedenko Theorem states that if $F \in MDA(H)$ then $H$ must be a GEV distribution. Most common continuous distributions in statistics are actually in $MDA(H_{\xi})$ for some $\xi$.

With this theory in mind, we shall now review the block maxima method in practice. Assuming that the sample is IID and from an unknown distribution $F$ in the domain of attraction of an EV distribution $H_{\xi}$. Then theory presented above says that the distribution of the n-block maximum $M_n$ can, for a large enough $n$ (such that the block maximas are independent), be approximated by a GEV distribution $H_{\xi,\mu,\sigma}$. We continue by dividing the data into $m$ blocks of size $n$, for example yearly blocks, resulting in data $M_{n1}, \ldots, M_{nm}$. This new data can now be fitted to a GEV distribution using for example maximum likelihood to estimate the parameters given that $\xi > -\frac{1}{2}$. It seems easy but unfortunately we have a trade-off problem here, if $n$ is high and thus the block maxima observations are independent the GEV distribution will be a better approximation of the data but there will also be fewer blocks/observations of block maximas and hence the ML estimation will be less accurate, and the other way around.

As can be seen the block maxima method only makes use of maximum losses of large blocks which is the reason to why McNeil (2015) writes that it is inefficient with the data. We will now look at the POT method that does not have this weakness as it makes use of all observations above a certain threshold.[McNeil et. al. (2015)]

**Peaks Over Threshold Method**

The asymptotic properties of the tails of a distribution function cannot be fully determined from a sample as sample size is always limited meaning that we cannot know what happens outside the tails. However if the samples are IID r.v’s with unknown distribution then the large values may be used to extrapolate the empirical tail outside the range of the sample, POT is one such extrapolation approach. It turns out that the tails, over some high threshold can be approximated by the Generalized Pareto Distribution (GPD). As it seems to be the case for us that tails are heavy and asymmetric, POT may provide better estimates of the risk measures than for example the normal and t-distribution as they are symmetric. Furthermore there are straightforward formulas for calculating VaR and ES when assigning GPDs to the tails which makes POT practical.
GPD is a suitable distribution to model exceedences over thresholds and has the following distribution function

\[ G_{\xi,\beta}(x) = \begin{cases} 
1 - (1 + \frac{\xi x}{\beta})^{-1/\xi} & \text{if } \xi \neq 0 \\
1 - \exp(-\frac{x}{\beta}) & \text{if } \xi = 0 
\end{cases} \]

where \( \xi \) is the shape and \( \beta \) the scale parameters. \( \beta > 0 \), and \( x \geq 0 \) when \( \xi \geq 0 \) and \( 0 \leq x \leq -\beta/\xi \) when \( \xi < 0 \). Furthermore,

\( \xi > 0 \implies \) ordinary Pareto distribution
\( \xi = 0 \implies \) Exponential distribution
\( \xi < 0 \implies \) short-tailed, Pareto type II distribution.

As \( G_{\xi,\beta} \in MDA(H_\xi) \) for all \( \xi \in \mathbb{R} \) the GPD is a natural model for the excess distribution over a high threshold in EVT.

If \( X \) is a r.v with distribution function \( F \) then the definition of the distribution function of the excess distribution over a threshold \( u \) is

\[ F_u(x) = P(X - u \leq x \mid X > u) = \frac{F(x + u) - F(u)}{1 - F(u)} \]

for \( 0 \leq x < x_F - u \) where \( x_F \leq \infty \) denotes the right endpoint of \( F \). The Pickand-Balkema-de Haan theorem states that there is a positive-measurable function \( \beta(u) \) such that

\[ \lim_{u \to x_F} \sup_{0 \leq x < x_F - u} | F_u(x) - G_{\xi,\beta(u)}(x) | = 0 \]

if and only if \( F \in MDA(H_\xi) \) where \( \xi \in \mathbb{R} \). This means that if the normalized maxima of a distribution converges to a GEV distribution then its excess distribution converges to the GPD.

We will now continue with how the POT method function. If we assume that the loss distribution \( F \in MDA(H_\xi) \), the Pickand-Balkema-de Haan theorem tells us that for a high enough threshold \( u \) the excess distribution can be modeled as a GPD, i.e. \( F_u(x) = G_{\xi,\beta}(x) \). Denote the data exceeding the threshold \( \hat{X}_1, ..., \hat{X}_{N_u} \) where \( N_u \) is the number of observations exceeding the threshold and \( Y_j = \hat{X}_j - u \) the amounts of excess losses. The data consisting of the amounts of excess losses \( Y_j, \ j = 1, ..., N_u \) may then be fitted to a GPD using for example maximum likelihood given that \( Y_1, ..., Y_{N_u} \) are IID. In the end we are interested in finding not only \( F_u(x) \) but also the risk measures VaR and ES. This is straight forward when using POT and their respective expressions are given as

\[ VaR_p = u + \frac{\beta}{\xi} \left( \left( \frac{n}{N_u}(1 - \alpha) \right)^{-\xi} - 1 \right) \]
ES_p = \frac{VaR_\alpha}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}

Where \( \alpha \) is the confidence level and \( n \) is the number of observations of the loss distribution. The calculation of \( ES_p \) is possible only if \( \xi < 1 \). The entire distribution of the data can then be expressed as a hybrid distribution of GPDs in the tails and some fitting distribution such as the empirical between the tails. [McNeil et. al. (2015)]

Hult et. al. (2012) argues that MLE is better than LSE when using the POT method as "LSE is a natural choice for parameter estimation if the tails of the model have significant influence on the estimates. Here we want all of the samples of excesses to fit nicely to the generalized Pareto distribution, not primarily the tails." and illustrates this with an example. Furthermore there is a trade-off problem as one must find a threshold large enough so that the excess distribution can be modeled as a GPD but with enough observations above the threshold for the parameter estimates of the GPD to be reliable and thus a valid question is, how can we choose the threshold \( u \)? In some problems it is clear what to set the threshold to i.e. what to consider as an extreme value. In this setting it is not, it is subjective. In the rest of this section we will discuss the choice of threshold.

One approach is to make QQ plots for a range of thresholds where the range of thresholds can be fixed values or quantiles. Another is to, for the same range of thresholds, plot the GPD parameter estimates with their confidence intervals for the different thresholds and identify a ’stable region’. Stability is reached when we are in the tail but going too far into the tail would lead to larger confidence intervals as there are too few observations there. While the QQ plots are informative regarding the fit of the tails to a GPD, drawing them for many values in somewhat inefficient. Thus the second approach is preferred and the choice may then be confirmed by studying the QQ plot at this threshold [McNeil et. al. (2015)]. Best practice is to not only consider the thresholds for one data set given a fixed window but to take other non-overlapping windows of the same length and verify that the chosen threshold is a good choice for them all, typically by drawing the GPD parameter estimates plots for the threshold range and a QQ plot for the chosen threshold. If one threshold is a good choice for all these data windows it can be assumed that this threshold is a good choice as time evolves as well.
Chapter 4

Evaluating Risk Measurement Models

The RMMs presented above are based on assumptions such as for example independence between observations and hence they need to be validated before they can be used in practice. The classical way to validate RMMs is by backtesting, i.e. estimating the risk measures with actual realized out-of-sample deviations and based on this inform about the credibilities of the risk measurement procedures. How backtests of VaR and ES may be performed will be treated in this chapter.

4.1 Backtesting VaR

When backtesting VaR there are two important properties to consider, unconditional coverage and independence. Unconditional coverage assures that the number of exceedences is not too high. Independence assures that exceedences are not too clustered together, if they are, the number of exceedences over for example a week could become unacceptably high. In general backtests are focused on unconditional coverage but independence is an important property as well. However, in this particular case when the observations in the data sets are not ordered according to time within the days, we should not put too much weight on the independence property as it will be difficult to reliably test it.

Define the indicator function $I_{t+1} = I_{X_{t+1} > \text{VaR}_p}$ as a VaR violation where $\text{VaR}_p$ denotes the VaR of the distribution $F_{X_{t+1}|F_t}$. Then by definition

$$E[I_{t+1} | F_t] = 1 - \alpha = p$$

implying that $I_{t+1}$ is a binomial variable with success probability $p$. Lemma 9.5 in McNeil (2015) shows that the sequence of these indicator variables is an process of IID binomial variables with success probability $p$.

**Lemma 9.5.** Let $(Y_t)_{t \in \mathbb{Z}}$ be a sequence of Bernoulli indicator variables adapted to a filtration
(\mathcal{F}_t)_{t \in \mathbb{Z}}$ and satisfying $E[I_{t+1} \mid \mathcal{F}_t] = p > 0$ for all $t$. Then $(Y_t)$ is a process of IID Bernoulli variables.

The independent Bernoulli property of the violations implies that:

1. The sums of violation indicators over different time periods are binomially distributed as $\sum_{t=1}^{n} I_{t+1} \sim B(n, 1 - \alpha)$.

2. Given that the violations occur at $t \in (T_1, ..., T_M)$, the spacings between violations $S_j = T_j - T_{j-1}$ are geometrically distributed and independent r.v’s with mean $1/p$ resulting in $P(S_j = k) = \alpha^{k-1}(1 - \alpha)$.

Under the null hypothesis that the method for estimating $VaR_t^p$ is correct, the sequence of $I_{t+1}$ should behave like an independent Bernoulli process. [McNeil et. al. (2015)]

### 4.1.1 Unconditional Coverage

The first property implied by the independent Bernoulli property, also called unconditional coverage, can be tested using a binomial test. There are several different binomial tests that would be possible to use, McNeil et. al. (2015) propose a two-sided Z-test. A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. This one has the statistic

$$Z_n = \frac{\sum_{t=1}^{n} I_{t+1} - np}{\sqrt{np(1-p)}}.$$ 

Assuming independence, the null hypothesis $H_0 : \sum_{t=1}^{n} I_{t+1} \sim B(n, 1 - \alpha)$ is rejected at 5% level if $|Z_n| > \Phi^{-1}(0.975)$ due to systematic overestimation or underestimation. A discrete Bernoulli process for rare events can be approximated by a continuous Poisson process and the discrete geometric distribution of the spacings between successes can be approximated as a continuous exponential distribution. This information can be used to test the second property implied by the independent Bernoulli property called independence. Thus, independence can be validated simply by making a Q-Q plot of the spacings against the exponential distribution.[McNeil et. al. (2015)]

Kupiec (1995) use a likelihood ratio test to test if the observed number of exceedences is equal to the number of exceedences expected if the process $I_t$ is IID Bernoulli distributed with success probability $p$. The likelihood function for Bernoulli($p$) distributed IID random variables is

$$L(p) = (1 - p)^{n-x}p^x$$

where n is the sample size and x is the number of observed exceedences and p is the probability of an exceedence on any one of the independent trials. The likelihood ratio (LR) test for testing the
null hypothesis \( H_0 : p = E[I_t] = p^* \), where \( p^* = x/n \) is the observed probability of exceedences and the MLE of \( p \).

\[
LR_{uc} = -2 \log[(1 - p)^{n-x}(p)^x] + 2 \log[(1 - \frac{x}{n})^{n-x}(\frac{x}{n})^x] = -2 \log \left( \frac{L(p)}{L(p^*)} \right).
\]

Under the null hypothesis, \( LR_{uc} \sim \chi^2_1 \) and thus the null hypothesis is rejected if \( LR_{uc} \) is larger than the critical value of \( \chi^2_1 \).

The two tests presented above are two-sided tests meaning that methods for estimating \( VaR_p^t \) are rejected both if \( VaR_p^t \) is underestimating the risk and if it is overestimating the risk. From the viewpoint of regulators it is acceptable to have a risk measure that yields a too conservative estimate of the risk, however it is better to have a risk measure expressing exactly what we want it to express, especially for an institution. Thus it is preferred to quantify the model risk using a model that is accepted in a two-sided test but it is acceptable to use a model passing only a one-sided test. The Basel Committee’s traffic light coverage test presented in

**Basel Committee on Banking Supervision (1996)** is a one-sided coverage test based on the binomial distribution and can be used for this purpose. There are three zones of acceptance, green yellow and red where the green zone is the best. We will only accept methods for which the test results falls in the green zone as these results suggests no problem with the RMM while results in the red zone almost certainly suggests a flawed RMM. Furthermore, the chance of erroneously accepting an inaccurate model (type II error) is very small in the green zone while it is much higher in the yellow and red zones.

Assuming independence between the exceedences, the null hypothesis of the traffic light coverage test is \( H_0 : \sum_{t=1}^n I_{t+1} \sim B(n, p^* \leq p) \) and the alternative hypothesis is \( H_1 : \sum_{t=1}^n I_{t+1} \sim B(n, p^* > p) \) where \( p^* \) in the realized success probability. This may be tested by computing the binomial cumulative probability of the observed number of exceedences, when the success probability is \( p \). Recall that the cumulative probability is the probability of obtaining a given number or fewer exceedences in a sample of \( n \) observations when the true coverage level is \((100)\alpha\%\). The green zone is defined by that the null hypothesis is rejected if the cumulative probability is larger than 95%.

### 4.1.2 Independence

The unconditional coverage test by **Kupiec (1995)** implicitly assumes that the exceedences are independent. Thus it is also necessary to test this assumption explicitly which is done in **Christoffersen (1998)** using, again, a LR test. Assume \( \pi_{ij} = P(I_t = j \mid I_{t+1} = i) \), this indicates that \( I_t \) is a first-order markov chain with the following transition probability matrix

\[
\Pi = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}
\]
where for example \( \pi_{01} \) is the probability that no exceedence is followed by an exceedence. The approximate likelihood function for the process is

\[
L(\Pi) = (1 - \pi_{01})^{n_{00}}(\pi_{01})^{n_{01}}(1 - \pi_{11})^{n_{10}}(\pi_{11})^{n_{11}}
\]

where \( n_{ij} \) denotes the number of observations with value i followed by j. The MLEs of the parameters are

\[
\hat{\pi}_{01} = \frac{n_{01}}{n_{00} + n_{01}}, \quad \hat{\pi}_{11} = \frac{n_{11}}{n_{10} + n_{11}}
\]

Note that if \( I_t \) is independent then \( \hat{\pi}_{01} = \hat{\pi}_{11} = \hat{\pi} = \frac{n_{01} + n_{11}}{n} \) and thus the null hypothesis is \( H_0 : \hat{\pi}_{01} = \hat{\pi}_{11} = \hat{\pi} \). The LR statistic is the following:

\[
LR_{\text{indep}} = -2 \log[(1 - \hat{\pi})^{n_{00} + n_{10}}(\hat{\pi})^{n_{01} + n_{11}}] + 2 \log[(1 - \hat{\pi}_{01})^{n_{00}}(\hat{\pi}_{01})^{n_{01}}(1 - \hat{\pi}_{11})^{n_{10}}(\hat{\pi}_{11})^{n_{11}}]
\]

\[
= -2 \log \left( \frac{L(\hat{\Pi})}{L(\hat{\Pi}_1)} \right) \sim \chi^2_1.
\]

Furthermore, Christoffersen (1998) not only test for unconditional coverage and independence using two separate tests, these test statistics can also be combined to form one single test of conditional coverage i.e. a joint test of coverage and independence. The null hypothesis \( H_0 : \hat{\pi}_{01} = \hat{\pi}_{11} = \hat{\pi} \) is tested with the statistic

\[
LR_{\text{cc}} = LR_{\text{uc}} + LR_{\text{indep}} \sim \chi^2_2.
\]

If we would reject a test we would also have to reject the underlying RMM.

Note that when VaR is validated based on the unconditional coverage and independence properties the size of the exceedences are not taken into account. While the magnitude of the exceedences may be of importance it is natural that they are not taken into account when validating VaR as it is a measure that does not take tail events into account.

### 4.2 Backtesting ES

"The Basel Committee on Banking Supervision (BCBS) is the primary global standard setter for the prudential regulation of banks and provides a forum for regular cooperation on banking supervisory matters. Its 45 members comprise central banks and bank supervisors from 28 jurisdictions. They come up with frameworks from a regulatory perspective to measure for example market risk. They introduced Basel II where VaR was viewed as best practice when it comes to market risk, ES was proposed but due to difficulties of backtesting it got rejected. Research published by Gneiting in 2011 concluded that ES was impossible to backtest due to lack of elicitability which is a mathematical property enabling straight forward backtesting through a scoring function. Yet Basel III later (2013) suggested that ES would replace VaR as risk measure
as VaR does not capture the tail risk that ES does. This lead to more research on how to backtest ES and resulted in several suggestions. Among the suggestions there is not yet a consensus concerning which backtesting method that should be used and hence we must consult literature on the subject.

Three articles comparing the performances and implementation restrictions of different ES backtesting methods has been reviewed, Engvall (2016), Wimmerstedt (2015) and Edberg and Kack (2017). Edberg and Kack (2017) have reviewed both Engvall (2016) and Wimmerstedt (2015) and thus we will mostly focus on this article when choosing a suitable ES backtest. They evaluate seven ES backtesting methods $Z_i, i = 1, \ldots, 7$ using both simulated and real market data. These will now be described briefly.

The backtest proposed by Acerbi and Szekely (2014) is based on simulation where one can choose between three different test statistics yielding three different backtests $Z_1, Z_2, Z_3$. $Z_1$ is based on the average of the VaR exceedences and assumes that VaR at the same level as the ES being tested is correct. This means that it would be wise to pair this test with a VaR backtest in order to avoid false rejections and false acceptances. This VaR test must be two-sided as both too high and too low estimates of VaR can cause such false rejections and acceptances. $Z_2$ is similar to $Z_1$ in the sense that it is also based on the average of the var exceedences but in addition to this it also penalizes by lowering the test statistic if there are more VaR exceedences than what can be expected. $Z_3$ is more complicated than the others but it also perform better on real data than $Z_1$ and $Z_2$ according to Edberg and Kack (2017).

Constanzino and Curran (2015) propose an asymptotically normally distributed Z-test that can be used for spectral risk measures such as ES. If we let the average number of VaR exceedences define the failure rate of VaR exceedences, they define a failure rate for spectral risk measures such as ES. The difference is that while failure is a discrete event for VaR it is continuous taking values between zero and one depending on the severity of the VaR exceedence for spectral risk measures. They show that this spectral measure failure rate is asymptotically normally distributed under the null hypothesis that the failure rates for the spectral risk measure are i.i.d and $P(X_t \leq VaR^p_t) = \nu \in (0, \alpha]$ which allows us to set up a formal test statistic $Z_4$. Edberg and Kack (2017) finds that $Z_4$ and $Z_2$ has equal performances but as $Z_2$ requires simulations and $Z_4$ does not, $Z_4$ is preferred.

Emmer, Kratz and Tasche (2015) has shown that even though ES is not elicitable, if it is conditioned on VaR which is elicitable it is, meaning that ES is conditionally elicitable. As ES can be approximated as a mean of different level VaRs,

$$ES_p(X) = \frac{1}{p} \int_0^p VaR_u(X) du \approx \frac{1}{4}[VaR_{0.25p}(X) + VaR_{0.5p}(X) + VaR_{0.75p}(X) + VaR_p(X)]$$
yielding that if \( p = 0.1 \) as is the case here,

\[
ES_p(X) = \frac{1}{p} \int_0^p Var_u(X) du \approx \frac{1}{4}[Var_{0.025}(X) + Var_{0.05}(X) + Var_{0.075}(X) + Var_{0.1}(X)].
\]

They propose to backtest ES with VaR backtesting methods. If the different level VaRs are all accepted we accept ES as well. This test will be called \( Z_5 \) but note that it differs from the test used by Edberg and Kack (2017) and Wimmerstedt (2015) as they choose to add a fifth VaR level to the approximation as it is judged more suitable for their situation. Earlier we have not used this formula to compute ES so in order to use this backtesting method we would have to compute new estimates of ES using this formula. However Engvall (2016) points out a difficulty with this method which is to determine the p-value of such backtest. It is possible to combine multiple tests but independence between them has to be assumed which is unrealistic in this setting. On the other hand, by assuming that the VaR exceedences at different levels are highly correlated, one may roughly estimate the combined p-value as the lowest p-value amongst them. We will not worry about the p-value of the test, as long as the VaR levels are accepted we will accept ES as well. Further it has the advantage of being the most simple and understandable method to backtest ES. Edberg and Kack (2017) finds that \( Z_5 \) generally perform better than \( Z_2 \) and \( Z_4 \) and a bit worse than \( Z_1 \) and \( Z_3 \) in terms of rejection rate, however, they still prefer \( Z_5 \) over the other tests due to its implementation benefits. Both Engvall (2016) and Wimmerstedt (2015) finds the method proposed by Emmer, Kratz and Tasche (2015) most promising.

Emmer, Kratz and Tasche (2015) are not strict about the number of VaR levels in the ES approximation or how they should be weighted. The four term approximation is only an example and their general recommendation is that this should be chosen for each specific case. The test denoted \( Z_6 \) is the same as \( Z_5 \) with an additional, higher level VaR added to the approximation of ES. Basel Committee on Banking Supervision (2013) suggests a backtest for ES which is basically the same as the test suggested by Emmer, Kratz and Tasche (2015) but with only two VaR levels in the ES approximation when \( p = 0.025 \) which is a common level in the case of market risk. This is denoted \( Z_7 \) in Edberg and Kack (2017) and they found it to have the worst performance of all methods considered.

To summarize, the study conducted by Edberg and Kack (2017) results in the recommendations to use \( Z_1 \) if tail distribution is estimated and VaR is known, \( Z_3 \) if distribution is estimated and VaR estimate uncertain and \( Z_6 \) if distribution is not estimated but discrete VaR values can be estimated.
Chapter 5
Data Overview

This section aims to get to know the data we have chosen to use in this project which consist of the deviations mentioned in the introduction. This will be done on a preliminary level where focus will be on visual tools such as histograms and time series. Note that what will be presented as time series here are not real time series as within the days, the time and order of the observations is unknown. When inspecting the time series of deviations for different Valuation Models (VM) it is found that the quality of the data has not always been good, possibly due to that the information regarding counterparty spreads has not always been saved for all counterparties. This results in few observations per day in older data, however for most models one certain day from which the number of observations increases can be identified. This date is not the same for all VMs but for two VMs (VM 1 and VM 2) out of the three VMs with most observations in total this date lies in November 2017 and hence we will assume that the data from 2018 and later is of relatively good quality for all models. By observing Figure 5.1, one can see that this assumption is quite good as the number of observations per day is fairly stable in 2018 for most VMs. As we continue, only these observations will be part of the data sets.
Furthermore, it was mentioned before that valuations and exit prices differ as exit prices are affected by the counterparties positions in the market. The counterparty has an idea about where bid-ask values lies in the market. If they want to be neutral i.e. not take on any risk, they place their bid-ask where the market seems to be. If they are willing to take on risk and believe that the instrument will be worth more or less in the future than the market seem to think, this belief will be reflected in their bid-ask prices. In general, a belief in increase will result in a higher bid and ask while belief in decrease will result in a lower bid and ask. The interpretation is that this shows up in our data as clusters of sparse peaks or clusters of sparse lows, as can be seen in Figure 5.2. Since we are interested in differences between valuations, these observations should be removed from the data set as they are corrupt due to speculation.

Figure 5.1: Number of observations per day in 2018.
Thus we would like to transform the data such that the counterparties positions in the market
doesn’t effect the time series of deviations. Several techniques to eliminate these observations
were tested such as different kinds of volatility normalization but what seems to yield the most
promising results is actually outliers elimination. More specifically outliers elimination where an
observation is viewed as an outlier if it is more than 3 standard deviations away from the mean of
the data set. The results together with the original observations can be seen in Figure 5.3 where,
indeed, the denser part of the time series have been captured.

Figure 5.2: Time series of the data sets from 2018.
Figure 5.3: Time series of the original data sets from 2018 together with the time series where outliers has been eliminated.

When observing the time series in Figure 5.3, one can see that the mean is approximately constant during the entire observation period whichever data set or part of the data set we choose to isolate (with a few short exceptions). The same can be observed for the standard deviation which seems to be fairly stable as well. Hence we can use the mean and standard deviation of the in-sample data to exclude the undesired observations in the out-of-sample data with the same method. This provides consistency.
Figure 5.4: Histograms of the data sets from 2018.

From the histograms in Figure 5.4 we observe that some distributions are bimodal while others are unimodal. From this one can draw the conclusion that ordinary distributions such as the Normal or \( t \)-distributions will definitely not be good fits for all VMs, in particular not the for the ones that are bi- or even multi-modal.

**Q-Q Plot**

In order to verify if a certain distribution is a good fit for the data one may use a quantile-quantile (Q-Q) plot. It is a plot of the sample quantiles of the empirical data against theoretical quantiles from a distribution of choice. If the chosen distribution is a good fit the blue in Figure 5.5 will take a linear shape coinciding with the red line, and if the shape is an S the empirical data has fatter tails than the chosen distribution. An inverted S shape indicates that the tails of the chosen distribution are fatter than the tails of our data. The S shape observed in Figure 5.5 is typical when doing a Q-Q plot of financial data against the standard normal distribution, as financial data often has fat tails.
Figure 5.5: Example of Q-Q plot of financial data against the standard normal distribution.
Chapter 6

Method

The method judged to be the most appropriate and will be used to quantify model risk is the benchmark inspired approach where the deviations between our valuations and counterparty valuations are used to create a sample space of plausible valuations for each VM. One of the RMMs presented in Chapter 3 can then be used to estimate the 90\% confidence point. This choice can be motivated by two of the key evaluation points from the problem description: efficiency and generalizability. The third evaluation point, accuracy, will be considered later as the backtests will be used to evaluate the accuracy of our method to quantify model risk.

The method chosen to be used in this thesis can determine the model risk of many different instruments at the same time if they are valuated using the same VM while when using any of the alternative methods to quantify model risk, each instrument has to be investigated in order to find its plausible VMs. It may be possible to group instruments when using the alternative methods as well if it is believed that they share the same set of plausible VMs but individual analyses of each instrument still has to be done in order to reach this conclusion. As these VMs might not be implemented already, computing the model risk becomes a very time consuming task. The method used in this project is not without effort as it requires that the user implement HS, POT, PN and Pt and backtest them to make a choice. Implementation of HS, PN and Pt is simple but when implementing POT the user has to choose a threshold which is a little bit more complicated. However it is not as time consuming nor as subjective as the choice of plausible VMs and hence our model may still be considered more efficient.

Our method is general in the sense that the same method, the one described in this chapter, may be applied given data from any VM. The alternative methods that can be used all depend on identification and implementation of plausible VMs for each instrument which is not in any way a general method. The generalizability of this method will be evaluated further after backtesting as the backtesting will tell us if it is possible to find a suitable RMM for data with different characteristics.
Risk Measurement Method Selection

Furthermore, one has to estimate a risk measure equivalent to the 90% confidence point, based on the distribution of the deviations. It has been seen in the chapter on RMMs how one can go from a distribution to a risk measure by computing the distribution based risk measures VaR and ES. There is no one risk measurement model that is considered “the best” in the literature, one must let the circumstances at hand determine which model to apply.

In light of the data, time series of the underlying data seems fairly stable without much fluctuations which was observed in section 4. This means that the past can be assumed to be a good approximation of the future in the case of model risk and thus HS may produce good estimates of the risk measures.

As most data sets are bi- or multi-modal or heavy tailed it is not believed that the Normal or $t$-distribution will be good approximations of the distributions. The data was fitted to a few less ordinary distribution families but none of these were satisfactory. Furthermore, as many distributions only take positive or non negative values one can try to divide the data into two parts where one is larger or equal to zero and the other is smaller than zero (and then takes the absolute value of this part) in order to have access to more distributions. While this may work, general techniques that applies to many time series are to prefer instead of detailed analysis of single series as there are many different models. Thus, although the Normal and $t$-distribution does not seem suitable when using the parametric model they will be used due to their simplicity.

It was seen that some of the data sets have heavy tails which indicates that some EVT method could yield better estimates of the risk measures than for example the normal distribution parametric method. We choose to use POT and not the block maxima method as it makes better use of the limited data than the block maxima method and also there are simple formulas that can be used to compute the risk measures $VaR_p$ and $ES_p$.

Some commonly used methods to estimate risk measures have not be treated here as they are not applicable. One such method is to use a time series model, typically GARCH and ARCH when dealing with financial time series, to describe the data generating process. They are popular as they capture the volatility clusters present in many financial time series which leads to more accurate risk estimates. In this setting our observations arrive at unknown times within the days and with varying time steps between them and thus cannot be described using any time series model. This is fine as model risk, the way we have chosen to approach it here, is based on data which is not primarily characterized by clusters. Thus using methods with advantages for data with volatility clusters seems unnecessary. Another method that can be used to estimate VaR and ES is Monte Carlo simulation. This is an advantageous approach when one deals with complex distributions and for example portfolios of correlated instruments and wish to make risk forecasts of one week or more and hence this computationally heavy method is not of interest to us either.

We will in this project use the RMMs HS, parametric Normal and $t$-distribution and POT and compare their risk estimates using backtesting which will provide support to choose one RMM
Weighting of Historical Data

A fixed window if historical data is used for all RMMs. As we saw in Figure 5.1, the data sets used in this project have different number of observations at different days meaning that the fixed windows will contain different number of observations if we decide to define the fixed window by number of days. We could also choose to define it by number of past observations to include in the fixed window. However, it makes little sense to include some observations from a certain day but not others and thus we will choose a number of days to include. It is difficult to decide on an appropriate length of sample data to include in this fixed window as it depends on the application. With the length of the fixed window we can make the risk measure depend more or less on older data and make it more or less responsive to major events. A short window will be more responsive to increased risk than a long window. This is also highly related to the weighting of the data included in the fixed window as we can adjust the weighting structure depending on what we want the risk measure to reflect.

In the HS method described in chapter 3 we assign equal weights to all dates in the fixed window. This means that older data influence the risk measure to the same extent as more recent data which could make the risk estimates unresponsive to extreme events. An alternative to improve the responsiveness of the risk measures to recent events in time is exponential age weighting where the weights assigned to the observation days decline exponentially until the last day in the window. The time series of the deviations in Figure 5.3 are mostly stable which implies that the past is assumed to be a fairly good approximation of the future in the case of model risk. Thus, equal weights is judged appropriate to use in this case.

When choosing the length of the fixed window one also has to keep in mind that a small sample size increases the possibility of estimation error, especially as we are interested in the tails where there are fewer observations. Thus a longer window is preferred to a shorter. Furthermore we have access to approximately 1 year of data and have to set some data aside for out of sample backtesting which impose a limitation for the window length. We will have to try different window lengths and base the decisions on the results of the backtesting. Note that the models have different sample sizes per day, some have very small and others have larger and thus we will not aim for a generic length for the group of models. The lengths 6 months, 3 months and 1 month will be considered and tested. \[\text{Dowd (2005)}\]

Now some final remarks. We are applying the RMMs on distributions where the sign does not matter in the sense that a positive sign of a deviation does not mean a good or bad outcome, positive and negative outcomes are viewed as equally bad.
A negative deviation means that, if HB is buying, it will have to pay more than expected. However, if HB is selling, a negative deviation is good as the counterparty is willing to pay more than expected.

A positive deviation on the other hand means that HB as a seller will sell at a lower price than they expected but as a buyer HB will buy at a lower price than they thought.

Thus, in order to obtain the 90% confidence point when HB need to sell to exit the valuation exposure the right tail risk measure needs to be added to the mid, if they need to buy to exit the valuation exposure the left tail risk measure needs to be added to the mid. Hence we are interested in estimating the risk measures in both right and left tails.

Furthermore, some of the methods that can be used to estimate risk measures, ex. Historical Simulation and Extreme Value Theory, assumes that the data is independent and identically distributed (IID) (parametric methods do not require this but MLE that we use to estimate parameters assumes IID data does). The most influential part of IID on the accuracy of the risk estimates is the identically distributed part. We may assume that a sample is identically distributed if the time series of these samples looks fairly similar during the period observed. As we have seen in chapter 5 the time series of the deviations are actually more or less similar during the observed period of time and thus, the assumption that the samples are approximately identically distributed is reasonable to make. If that would not be the case it does not mean that the methods assuming this property cannot be used but it can cause errors when using them which is something we must be aware of.

**Backtest procedure**

As several different RMMs are considered in this project we need to be able to compare and validate them before deciding on which one to use. The classic way to do this is by backtesting which was presented in detail earlier.

Starting with the VaR backtests. We first notice that due to the unknown order of the observations within the days the VaR independence test proposed in McNeil et. al. (2015) cannot be used in this project. While independence is good it is not a crucial part to accept a risk model, indeed it yields lower risk but if we look at Basel III, it is not required that a risk model pass an independence test. Thus one could say that firstly we accept models that are LR CC accepted, as the likelihood ratio independence test is the only independence test that is possible to do with the data at hand. If no such model exist we accept a two-sided coverage test accepted RMM, and lastly we accept models accepted by the traffic light test.

Furthermore it was found that the Z-test yields the same acceptances/rejections of RMMs as the LR unconditional coverage test for all models except one where the Z-test accepted more than the other test. This model happens to be one out of two models with significantly fewer
observations both in-sample and out-of-sample which provides an explanation to why these tests are not equivalent for this model. The distribution of the sum of the indicator variables is asymptotically normal with expected value $np$ and variance $np(1 - p)$, this is the base of the Z-test presented earlier. Hence for a model with few observations this might not be a reliable test. It is therefore concluded that the LR UC test dominate the Z-test as a two-sided test which makes the Z-test redundant. We now have the following levels on which we accept VaR models:

1. Likelihood ratio conditional coverage test
2. Likelihood ratio unconditional coverage test

Continuing with how to backtest ES. We are comparing the different RMMs HS, parametric and POT and it is desireable to use the same backtesting method for all risk models. Following the recommendation made by Edberg and Kack (2017) we should therefore use a VaR approximation approach. As this is not a thesis with main goal to compare ES backtests the default four term approximation will be used. The VaR levels will be backtested in the same ways as was stated above.
Chapter 7

Results and Discussion

The results for three of the eight Valuation Models (VM) will be presented in this chapter, VM 1, VM 2 and VM 8 as they are represented by different sizes of data sets and different number of observations per day. VM 1 has a large data set of approximately 200 observations per day, VM 2 has a medium size data set of 60 observations per day and VM 8 has a small data set of 3-5 observations per day. The results for the other five models may be found in Appendix.

7.1 Risk Measurement Models

In this section, an analysis of which RMM should be suitable for the different VMs will be presented, based on the in-sample data sets.

7.1.1 Historical Simulation

Only one decision has to be made when using the HS method, except for the window length and the weighting of data which has already been discussed. This is to choose the number of simulations N such that the risk measure estimates have converged. Convergence plots of the left and right tail VaR will be presented in this section in order to make this decision. For each window length we wish to find a value of N which is suitable for the right and left tail risk measure estimates. This is done by analyzing Figure 7.1. As the differences in suitable N values are small between the window lengths it is more practical to decide on a N which is generic across all window lengths. From Figure 7.1 it is decided that \( N = 500 \) is suitable as all estimates are stable by that number of simulations.
Figure 7.1: Convergence of risk measure with increasing number of simulations, VaR10 in the legends is the left tail VaR and VaR90 is the right tail VaR, for VM 1

Figure 7.2: Convergence of risk measure with increasing number of simulations, VaR10 in the legends is the left tail VaR and VaR90 is the right tail VaR, for VM 2
The value $N = 500$ seems suitable also for VM 2 by observing Figure 7.2. When looking at the rest of the convergence plots 7.3, A.1, A.5, A.9, A.13 and A.17 it actually turns out that $N = 500$ is suitable for all models. Hence we will, for all models and window lengths use $N = 500$.

Figure 7.3: Convergence of risk measure with increasing number of simulations, VaR10 in the legends is the left tail VaR and VaR90 is the right tail VaR, for VM 8

### 7.1.2 Parametric Models

Fitting the data to the Normal and $t$-distribution and estimating the parameters using MLE yields the QQ plots presented in this section. These serves to give us an idea about how well the in-sample data sets fits the Normal and $t$-distribution. Figure 7.4 tells us that the normal distribution should not be a good approximation for the data from VM 1 as the tails are too heavy. The left tail however seems go more towards the Normal distribution when the window length becomes shorter. The $t$-distribution does not seem to provide a good fit for the data at all as the data has much fatter tails than the $t$-distribution. As can be seen in Figure 5.4, the data set for VM 1 is unimodal with heavy tails which implies that these results are to be expected.
Figure 7.4: QQ plots of data against ML estimated Normal and \( t \)-distributions, for VM 1

Figure 7.5: QQ plots of data against ML estimated Normal and \( t \)-distributions, for VM 2

For VM 2, the Normal distribution actually provides a good fit for the left tail for all the
window length. The right tail is much too heavy for the Normal distribution and while the $t$-distribution seems to be more suitable for the left tail, especially for the shortest window length, it is not good for the left nor the right tail. Also here the tails of the sample data are too fat. In Figure 5.4 we can see that the data set for VM 2 is unimodal and right skewed which is in accordance with what we observe in the QQ plots in Figure 7.5.

![Figure 7.6: QQ plots of data against ML estimated Normal and $t$-distributions, for VM 8](image)

Both the Normal and $t$-distribution provides equally and quite good fits for the data and slightly better for the longest window length as can be observed in Figure 7.6. It also implies that the tails are equally fat unlike the tails of the previous two VMs. The data for VM 8 seems to have shorter tails than the previous two VMs as can be seen in Figure 5.4 where we can also observe that it does not have the bell shape that Normal and $t$-distributed data has. This is the reason to why we see that the sample quantiles deviate from the Normal and $t$-distribution theoretical quantiles in the beginning of the tails.

### 7.1.3 Peaks over Threshold

When applying the POT method one has to choose a threshold as was discussed earlier. The Figures 7.7, 7.9 and 7.11 in this section may be used to choose the threshold, one for each window length and tail. The QQ plots presented in this section are useful when evaluating the fit of the data to a GPD with the chosen threshold.
When looking at the formulas for VaR and ES in section 3.2.3, it becomes clear that the shape parameter has the largest influence on the risk measures, hence the thresholds should be chosen in a region for which primarily the shape parameter is stable. The chosen threshold will therefore be indicated with a vertical line only in the plots of the shape parameter for different thresholds and tails in the Figure 7.6, 7.9 and 7.11 below. It can be noted that sometimes there are several regions of thresholds that are stable, in that case, the one resulting in the best QQ plots has been chosen. Furthermore, the stability regions of the risk measures should also been taken into consideration when choosing the threshold.

When interpreting the QQ plots in this section, in general, if the sample quantiles in the end are below the red line the theoretical GPD distribution will overestimate the risk as it has heavier tails than the sample data. When the sample quantiles in the end are above the red line the theoretical GPD distribution will underestimate the risk as it has lighter tails than the samples.

Figure 7.7: POT parameters and risk measures for varying thresholds for VM 1
The choice of threshold is not trivial for the right tail as the shape parameter never really reach a stable region, not for any of the window lengths. This is something that can be observed for example if the data has been capped which it kind of has for VM 1 when the outliers were eliminated. For VM 8 the outlier elimination did not have the same capped effect and there we can find a stable region for the shape parameter. When looking at the QQ plots however GPD seems to provide a good fit given the chosen thresholds. The shape parameter for the left tail is much easier to find a stable region yielding nice looking QQ plots, although they indicate that the risk should be slightly overestimated when using the POT method for the left tail. Further it seems like the shortest window length is better when using the POT method.
Figure 7.9: POT parameters and risk measures for varying thresholds for VM 2

Figure 7.10: QQ plots of data against GPD with the thresholds of choice for VM 2

For the right tail as the shape parameter never really reach a stable region, not for any of the
window lengths, but again, the QQ plots look good. The QQ plots are acceptable for the left tail as well where in addition, the shape parameter has regions of stability. The longer window lengths seems to be better fitted by the GPD distribution, for both tails.

Figure 7.11: POT parameters and risk measures for varying thresholds for VM 8
Figure 7.12: QQ plots of data against GPD with the thresholds of choice for VM 8

For the data from VM 8 it is possible to find a stable region for the shape parameter both for the left and right tail and all window lengths. However, the resulting GPD distributions seems likely to often underestimate the risk measures at both tails when looking at Figure 7.12.

7.1.4 Overview

To summarize, VM 1 have heavy right and left tails yielding the hardly surprising result that both the Normal and \( t \)-distributions provide poor fits to the data while the GPD seems reasonable for both tails. VM 2 has a heavy right tail but not a heavy left tail. From the distribution analysis figures above conclusions were drawn that the left tail looks normally distributed and that the right tail is better approximated by a GPD. However, GDP seemed to provide a good fit for the left tail as well with more stable parameters than for the right tail. The data from VM 8 with its lighter tails may be acceptably fitted by the normal or \( t \)-distribution. From what has been seen in this section, the GPD should generally result in underestimated risk measures for this VM.

7.2 Backtesting

In the previous section it was discussed which RMMs should lead to correctly estimated risk measures. In this section we will continue by presenting visualizations of the model risk measures together with the out-of-sample data and presenting the results of the backtests which will
reveal how good the RMMs and therefore the model risk method actually performed on out-of-sample data sets. The backtest performances will be compared with what could be expected given the analysis of the in-sample data in section 7.1. As stated in the methodology, three different backtests have been evaluated corresponding to different "levels of accuracy" where the first is the best. When talking about backtest 1 it will refer to the LR conditional coverage test, backtest 2 is the LR unconditional coverage test and backtest 3 is the traffic light green zone test. Other notation in this section is the following: HS refers to Historical Simulation as before, PN refers to the parametric Normal distribution method, Pt refers to the parametric $t$-distribution method and POT simply refers to the POT method. Hence, for example HS 1 refers to HS backtested with backtest 1.

In the tables containing backtest results below, one will denote an acceptance and zero will denote a rejection.

### 7.2.1 Valuation Model 1

![Visualization of all model risk measures on the time series of data for VM 1.](image)

Figure 7.13: Visualization of all model risk measures on the time series of data for VM 1.

In Figure 7.13 we can see that the risk measures are quite flat for all window lengths which one would expect as the time series in Figure 5.3 is very stable. The estimates seem reasonable but one could guess that either the left tail risk measures will be overestimated by some of the methods or the right tail risk will be underestimated by some of the methods. By looking at the
backtesting results in Table 7.1 and 7.2 this guess is confirmed as the Pt method always underestimate the risk measures while the PN method almost always underestimate the risk measures for the right tail.

<table>
<thead>
<tr>
<th>VaR, Model 1</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.1: Acceptances and rejections of VaR models for VM 1.

<table>
<thead>
<tr>
<th>ES, Model 1</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.2: Acceptances and rejections of ES models for VM 1.

In table 7.1 and 7.2 it can be seen that the PN and Pt methods overestimates the risk in the right tail rather than underestimates it. This is different from what was predicted in the previous section as the data seemed more heavy-tailed than the normal and $t$-distribution. Also, VaR levels further out in the tails are all poorly estimated as no ES models are accepted by backtest 1 and 2, but in general, the same methods which overestimated VaR at 90 and 10% levels also overestimated VaR levels further out in the tails.
7.2.2 Valuation Model 2

Figure 7.14: Visualization of all model risk measures on the time series of data for VM 2.

Figure 7.14 reveal that for VM 2, the risk measures computed using the shorter window length of 30 days is a bit more flexible than for the longer window lengths where the risk measures seems equally flat. Furthermore the risk measures for the right tail seems to be more spread out than for the left tail.

<table>
<thead>
<tr>
<th>VaR, Model 2</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.3: Acceptances and rejections of VaR models for VM 2.
Table 7.4: Acceptances and rejections of ES models for VM 2.

The backtest results for VM 2 are given in Table 7.3 and 7.4. It is surprising that the PN method results in good risk estimates for the right tail as it was found in the previous section that the right tail was much too heavy for the normal distribution while the left tail seemed fine. The POT method yields good estimates of the risk measures which is consistent with the analysis of the QQ plots in Figure 7.10. As predicted, the Pt method gets rejected the most and is only accepted with backtest 3 for the left tail. From Table 7.4 one can see that the HS method manage to produce more correct risk measures than the other methods further into the tails which indicates that it dominates the POT method for VM 2.

### 7.2.3 Valuation Model 8

Figure 7.15: Visualization of all model risk measures on the time series of data for VM 8.
In Figure 7.15 it is observed that for all window lengths, the estimates of the risk measures are less flat which is reasonable as the time series for VM 8 is less stable than for the other two VMs.

<table>
<thead>
<tr>
<th>VaR, Model 8</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.5: Acceptances and rejections of VaR models for VM 8.

<table>
<thead>
<tr>
<th>ES, Model 8</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.6: Acceptances and rejections of ES models for VM 8.

The first thing that is noticed when observing the backtesting results in Table 7.5 and 7.6 is that POT actually does not overestimate the risk measures as was predicted. The PN and Pt models have better performances for this model than for the other two which makes sense as normal and $t$-distributions were found to be better fits for this model than the other two. POT clearly outperforms the other methods, also further into the tails as can be seen in Table 7.6.

### 7.2.4 Overview

Before making the final choices of methods, the advantages and disadvantages of the RMMs will be summarized, as it is possible that one has to make the choice between two equivalently performing models. HS is simple to implement and understand but POT, PN and Pt are faster to compute than HS as many new data sets have to be simulated. The parametric methods are also easy to implement but the disadvantage is that the risk measure estimates obtained are sensitive to inaccurate distribution assumptions, even though ES is more sensitive than VaR. POT has the same disadvantages as the risk estimates, especially ES but also VaR, are sensitive to the choice of the threshold which can be seen in for example Figure 7.11. Another disadvantage with POT is that it is not as easy to implement as the other methods. Knowing this the first choice should be HS, the second choice should be PN or Pt and the last choice should be POT.
Note that there is a trade-off between accuracy and simplicity when it comes to the choice of RMM. The simplest case would be to use the same model and window length for both left and right tail risk measures. The most accurate choice is to choose the RMMs accepted by firstly backtest 1, secondly by backtest 2 and thirdly by backtest 3. One has to decide if the simplicity is worth choosing a method for at least one of the tails which has been accepted for example by backtest 3 instead of backtest 1 or 2. That depends on if the one considers the independence property or an exact exceedence probability to be more important than simplicity. One could argue that for most VMs, the tails look different and hence it makes sense to use different models and window lengths when quantifying their respective risks. In table 7.7 below both the most accurate choices of methods and the most simple choices that would allow us to use the same method and window length for both tails are presented, for VaR.

<table>
<thead>
<tr>
<th></th>
<th>most accurate methods</th>
<th>backtest acc.</th>
<th>most simple methods</th>
<th>backtest acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 left tail</td>
<td>HS 90 days</td>
<td>2</td>
<td>HS 30 days</td>
<td>2</td>
</tr>
<tr>
<td>M1 right tail</td>
<td>PN 180 days</td>
<td>1</td>
<td>HS 30 days</td>
<td>3</td>
</tr>
<tr>
<td>M2 left tail</td>
<td>HS 30 days</td>
<td>1</td>
<td>HS 30 days</td>
<td>1</td>
</tr>
<tr>
<td>M2 right tail</td>
<td>HS 30 days</td>
<td>2</td>
<td>HS 30 days</td>
<td>2</td>
</tr>
<tr>
<td>M8 left tail</td>
<td>POT 30 days</td>
<td>1</td>
<td>POT 30 days</td>
<td>1</td>
</tr>
<tr>
<td>M8 right tail</td>
<td>POT 30 days</td>
<td>1</td>
<td>POT 30 days</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.7: Choices of RMMs used to estimate VaR

The most accurate methods in table 7.7 are not always the ones that one would think simply by looking at the in-sample data analysis plots earlier in this section. For VM 1, PN seemed terrible and POT was believed to be good, and still, the most accurate RMMs were found to be HS and PN. It was believed that POT would underestimate the risk for VM 8 but instead POT turns out to be the most accurate RMM for VM 8 accepted by backtest 1. Also, as the GPD shape parameter for VM 8 was stable in both tails, which it was not for VM 1 and 2, this might indicate that one should not get fooled by good QQ plots if the shape parameter does not have a stable region of thresholds.

Few RMMs are accepted by backtest 1 meaning that exceedences are rarely independent. Independence of the data is an assumption made when using all risk measure methods. If this is not fulfilled it could be the reason to why the exceedences does not fulfill independence either as dependence in the data could lead to bad estimates of the risk measures. Even though we have a test for independence of the exceedences we cannot know for certain if the exceedences are independent or not as our data is not properly ordered. Hence we should not put too much weight on the independence property. RMMs accepted by backtest 2 result in very good risk estimates as well.

The stability of the time series should have an influence on the length of the fixed window of data. Less stability should imply a shorter window of data in order to capture changes while more sta-
bility should implies that we can use a longer window since older data is similar to what can be observed today. As can be seen in Figure 5.3, VM 1 and VM 2 have stable time series while VM 8 has a less stable time series in comparison which should imply that a shorter window length should be accepted in the backtests for VM 8 and a longer should get more acceptances in the backtests for VM 1 and 2. Table 7.1, 7.2, 7.3, 7.4, 7.5 and 7.6 revels that this is not true for any of the models analyzed in this section. Furthermore, among the final RMM choices presented in Table 7.7, the shortest window length clearly dominates but for VM 1 which has a very stable time series, 90 days is best for the left tail and 180 days is best for the right tail. The conclusion that can be drawn is that the window length is not something that can be determined simply by looking at the time series and therefore it was necessary to backtest multiple window lengths for each RMM.

ES is, as mentioned before, not a suitable risk measure according to the regulations as they are formulated today. It is part of this study as it could be relevant in the future or internally and to investigate if it could be set up and successfully backtested which it could. I will therefore not present a table as the one above for ES, however, it is not difficult as much fewer methods are accepted in the ES backtests, and also, for VM 1, 2 and 8 the most simple choice is equivalent to the most accurate choice.
Chapter 8

Summary and Conclusions

The backtest results has showed that one can identify a RMM yielding accurate risk measures for all VMs and hence methods based on detailed analysis of the distribution of the data seems unnecessary as this would make the model risk method less generalizable. Further, as the backtest results showed that one could find accurate RMMs for all VMs, which are represented by data sets of different sizes, it can also be concluded that the model risk method is both accurate and generalizable.

To conclude, a model risk method has been implemented and evaluated using backtesting. The method is in accordance with the aims stated in the problem description as it is judged to be accurate, efficient and generalizable. It also satisfies the regulatory requirements on institutions when it comes to model risk quantification without being an expert-based approach. Furthermore it was mentioned in the introduction that a measure of model risk may also be used to compare different valuation models in terms of, by taking the value approach, their conformance to the view of the market. This may serve in model validation as a mean to judge the quality of valuations and compare valuation models to each other. One problem that can occur when using this method is that none of the RMMs are successfully backtested meaning that the model risk method is not accurate enough. This is not something that has been seen in this thesis but it is not an impossible scenario. Also, if there are few counterparties for a valuation model, say 1 or 2, our method is not suitable.

8.1 Further Research

One limiting factor was that the observations in each day cannot be ordered correctly according to time and surely, if they could, the spacing of observation times would not be constant. This makes it difficult to use methods seen in time series analysis often used as risk measure models. What one could do is to take the mean of the observations in each day and call this our new time series. The issue is that we would loose a lot of information by doing this as only six months of data is available in-sample. If one would take this approach the data first has to be scaled such
that there are only positive values in the data set, otherwise the positive and negative values may take each other out in the mean. This would allow us to use methods for time series. Furthermore, if the exact time point of the observations would be available the mean could be taken over shorter periods of time, for example an hour, which would allow for less information to be lost. But with this amount of data and as long as exact time point of the observations are not available, it is not recommended as too much information would be lost.

The model risk approach used here depend on that the outlier elimination remove what we want it to remove. In order for the outlier elimination to successfully eliminate what is believed to be the result of counterparties positions in the market, the mean and standard deviations has to continue to be fairly stable. As can be seen in Figure 8.1, the same outlier elimination method as was used on in-sample data manage to eliminate what we want to eliminate out-of-sample as well. What could be done, if one does not want to rely on that this yields good results in the future as well, is to estimate the mean and standard deviation in the end of every year or six months and use these as parameters in the outlier elimination the following year or six months. This is however less consistent and we cannot try it to see the results here as there is not enough data to do that.

![Figure 8.1: Time series of out-of-sample original data together with time series of outlier eliminated data.](image)

It is also possible that there exist more complex and reliable ways to eliminate the observations that we argue do not have a place in the data sets and remove through outlier elimination. If one analyze the data in further detail maybe one could find something that characterize these
observations making them easier to identify.

It would be interesting to try to implement the Bayesian approach using performance data presented in the literature study. In this project there was not enough time to do this but I believe that it could be a fruitful approach to the problem.
Bibliography


Appendix A

Appendix

A.1 Valuation Model 3

Figure A.1: Convergence of risk measure with increasing number of simulations
Figure A.2: POT parameters and risk measures for varying thresholds

(a) QQ plots of data against GPD with the thresholds of choice
(b) QQ plots of data against ML estimated normal and t-distributions

Figure A.3
Figure A.4: Visualization of all model risk measures on the time series of data.

Table A.1: Acceptances and rejections of VaR models for valuation model 3.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.2: Acceptances and rejections of ES models for valuation model 3.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
A.2 Valuation Model 4

Figure A.5: Convergence of risk measure with increasing number of simulations
Figure A.6: POT parameters and risk measures for varying thresholds

(a) QQ plots of data against GPD with the thresholds of choice
(b) QQ plots of data against ML estimated normal and t-distributions
Figure A.8: Visualization of all model risk measures on the time series of data.

Table A.3: Acceptances and rejections of VaR models for valuation model 4.

<table>
<thead>
<tr>
<th>VaR, Model 4</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.4: Acceptances and rejections of ES models for valuation model 4.

<table>
<thead>
<tr>
<th>ES, Model 4</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
A.3 Valuation Model 5

Figure A.9: Convergence of risk measure with increasing number of simulations
Figure A.10: POT parameters and risk measures for varying thresholds

(a) QQ plots of data against GPD with the thresholds of choice
(b) QQ plots of data against ML estimated normal and t-distributions

Figure A.11
Figure A.12: Visualization of all model risk measures on the time series of data.

### Table A.5: Acceptances and rejections of VaR models for valuation model 5.

<table>
<thead>
<tr>
<th></th>
<th>ES, Model 5</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.6: Acceptances and rejections of ES models for valuation model 5.

<table>
<thead>
<tr>
<th></th>
<th>ES, Model 5</th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
A.4 Valuation Model 6

Figure A.13: Convergence of risk measure with increasing number of simulations
Figure A.14: POT parameters and risk measures for varying thresholds

(a) QQ plots of data against GPD with the thresholds of choice
(b) QQ plots of data against ML estimated normal and t-distributions

Figure A.15
Figure A.16: Visualization of all model risk measures on the time series of data.

Table A.7: Acceptances and rejections of VaR models for valuation model 6.

<table>
<thead>
<tr>
<th></th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table A.8: Acceptances and rejections of ES models for valuation model 6.

<table>
<thead>
<tr>
<th></th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
A.5 Valuation Model 7

Figure A.17: Convergence of risk measure with increasing number of simulations
Figure A.18: POT parameters and risk measures for varying thresholds

(a) QQ plots of data against GPD with the thresholds of choice
(b) QQ plots of data against ML estimated normal and t-distributions

Figure A.19
Figure A.20: Visualization of all model risk measures on the time series of data.

Table A.9: Acceptances and rejections of VaR models for valuation model 7.

<table>
<thead>
<tr>
<th></th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.10: Acceptances and rejections of ES models for valuation model 7.

<table>
<thead>
<tr>
<th></th>
<th>HS 1</th>
<th>PN 1</th>
<th>Pt 1</th>
<th>POT 1</th>
<th>HS 2</th>
<th>PN 2</th>
<th>Pt 2</th>
<th>POT 2</th>
<th>HS 3</th>
<th>PN 3</th>
<th>Pt 3</th>
<th>POT 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>180 days, left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>180 days, right</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>