Robust Estimation In Operational Risk Modeling

Joris Chau

August, 2013

Robust Estimation In Operational Risk Modeling

by

Joris Chau

A Thesis submitted in Partial Fulfillment for the Degree of
Master of Science

In the
Department of Mathematics
Utrecht University

August, 2013

Supervisors Dr. Cristian Spitoni
Utrecht University

Dr. Diederik Fokkema
Ernst & Young

Co-reader Prof. dr. Roberto Fernández
Utrecht University

Joris Chau
Robust Estimation In Operational Risk Modeling
Master’s Thesis, August, 2013
Supervisors: Dr. Cristian Spitoni and Dr. Diederik Fokkema
Co-reader: Prof. dr. Roberto Fernández

Abstract

Under the Basel II framework, it is required for internationally active banks to allocate capital for operational risk, equal to the 1-in-1000 years worst operational loss event. The most widely accepted method to calculate the capital charge (also Value-at-Risk or VaR) is to use the Loss distribution approach, in which we model separate loss frequency and loss severity distributions, where the distribution parameters are estimated using classical estimation techniques, such as maximum likelihood. We show that, under a correctly specified severity distribution and iid loss data, we are able to estimate the capital charge quite accurately for both non-truncated and truncated loss data. This is done using Fast Fourier Transform methods, instead of the generally used Monte Carlo simulation methods, which are computationally much slower.

A major problem in practice is that the estimated capital charge under classical estimation techniques is highly sensitive to minor contamination of the loss data, which may result in large swings in the capital charge when a single or a few loss events are added to the database. To ensure stable capital charges, we introduce the robust statistics framework, where the idea is to sacrifice some efficiency at the exact model, in order to gain robustness against minor deviations of the model. By computing the influence function (IF), we show that under classical estimation methods, such as maximum likelihood, the capital charge is extremely sensitive to minor contamination of the assumed severity distribution. To mitigate the impact of contamination in the left tail of the severity distribution, which is found to be one of the major causes of the instability in the capital charge, we propose fitting a mixture of severity distributions to the loss data.

After this necessary extension of the model, we consider two robust estimation techniques. First, we study the optimal bias robust estimators (OBRE), which can be viewed as the robust version of maximum likelihood estimation. Due to their computational complexity, the optimal bias robust estimators may be difficult to implement in practice. This is why, secondly, we study the method of trimmed moments (MTM), which can be viewed as the robust version of the method of moments and may be more straightforward both analytically and computationally in computing stable capital charges.

In a simulation study, we show that using the introduced robust models, the estimated capital charges comply with the desired properties of high efficiency at the exact model and stability under contamination of the loss data, where the classical model generally fails.

Keywords: Operational risk ⋆ AMA ⋆ Value-at-Risk ⋆ LDA ⋆ FFT ⋆ Constrained maximum likelihood ⋆ Robust statistics ⋆ MLE ⋆ Influence function ⋆ Mixed severity ⋆ Optimal bias robust estimation ⋆ Method of trimmed moments

Acknowledgement

This thesis is written as the final part of the Master Mathematical Sciences at Utrecht University. I had the opportunity to write this thesis during an internship of six months at the Financial Services Risk Management department of Ernst & Young in Amsterdam.

First of all, I would like to express my gratitude to my thesis supervisor, Cristian Spitoni, for the useful comments, remarks and engagement through the process. Without his guidance and help this thesis would not have materialized.

Secondly, I would like to thank my Ernst & Young supervisor, Diederik Fokkema, for the possibility to carry out this project during my internship at Ernst & Young and for the effort and time spent in supporting me on the way. I would also like to thank Bernardo Kok for helping me get started on the project. And I would like to acknowledge Roberto Fernández for accepting the task of second reader.

Finally, I would like to thank all colleagues of the FS Risk department for an enjoyable time these last six months, with as highlight the ski trip to Saas Fee.

Joris Chau
August 2013, Amsterdam

Contents

Introduction 1

The global financial system has been shaken by numerous banking failures over the last two decades, and the risks that internationally active banks have had to deal with have become more complex and challenging. Since the 1990s, we have observed more than 100 loss events exceeding $\$100$ million in value and a number of loss events exceeding $\$1$ billion, due to unauthorized trading activities, tax noncompliance, internal fraudulent activities, natural disasters or terrorist attacks.

Such extreme large-scale losses have resulted in bankruptcies, mergers or substantial equity price declines of a number of large financial institutions and banks in particular. It is argued that the increase in the number and scale of such *operational* losses has to do with financial deregulation, globalization and advances in the information network over the last 20 years. Furthermore, we have seen the banking industry grown exponentially in size, and the scale and diversity of operations within an internationally active bank have grown with it, therefore increasing the risk of large-scale operational losses.

To assess this issue, over the last two decades there has been a growing need to model so-called *operational risk*. This is also encouraged by regulations, since it is nowadays mandatory for banks to allocate capital to cover most large-scale operational losses. The modeling of operational risk is still relatively new and currently there is much ongoing research regarding this subject.

Arguably one of the most important measures in modeling operational risk is *Value-at-Risk*, which estimates the worst possible loss event that can occur with a certain probability over a given time window (for instance one-thousand years). The main problem in estimating this *risk measure* for most banks is the lack of coherent historical operational loss data[1]. The regulations state that the allocated capital charge should correspond to a 1-in-1000 years worst possible loss event, which is a much larger time window than the span over which banks have been collecting operational loss data (e.g. five to ten years). This is why banks use probabilistic models to estimate the risk measures and the resulting operational risk capital charge. Due to the lack of historical loss data, it has been found that if we use *classical* estimation techniques, the estimated capital charge can become unreasonably sensitive to minor contamination of the loss data, resulting in large swings in the estimated capital charge, when a few extra loss observations enter the database. This is the main issue that will be addressed in this thesis.

1.1 Goal and outline of the thesis

The goal of this thesis is to assess the influence of minor contamination of the operational loss data on the estimated capital charge and to consider *robust* estimation methods, which may lead to more stability in the estimated capital charges when the loss data does not follow the assumed model distribution exactly.

We start by setting out techniques and estimation methods to produce a sound *classical* model to estimate the operational risk capital charge. We apply the Loss Distribution Approach (LDA), where the idea is to model the loss frequency and loss severity distributions separately, which are then combined into an aggregated loss distribution (also compound distribution). To accurately estimate the capital charge, which is given by the Value-at-Risk (VaR) measure of the aggregated loss distribution, we apply

[1] Currently most banks have been collecting data on internal operational loss events for a maximum of five to ten years.

an efficient method using Fourier transformations.

Furthermore, we address the issue of data truncation, since it is seen that in practice the operational loss data is subject to lower recording thresholds.

Next, we show that, under minor contamination of the loss data, the estimated VaR measures become very unstable, where contamination of the loss data is defined as a single or a few loss observations that to not follow the assumed model distribution exactly. To asses this issue, we introduce the robust statistics framework, where the idea is to sacrifice some efficiency at the exact model, in order to gain robustness against minor deviations of the model. By means of the *influence function*, it is seen that, under the classical model, the estimated capital charge is not only highly sensitive to large scale loss observations, but also to contamination due to very small loss events. To decrease the effect of contamination in the left tail of the severity distribution (small loss events), we propose fitting a mixture of severity distributions to the operational loss data.

After this extension of the classical model, we argue that in practice we may lack sufficient historical loss data to accurately fit a mixture of severity distributions, since we have to split up the -already scarce- loss data into a separate body and a tail region. Since it is found that the truncation of the loss data also mitigates the impact of small loss events on the estimated capital charge, it might be enough to fit a single severity distribution to the loss data using *robust* estimation methods.

We consider two robust estimation techniques. First, we study the optimal bias robust estimators (OBRE), which can be viewed as the robust version of maximum likelihood estimation. Due to their computational complexity, the optimal bias robust estimators may be difficult to implement in practice. This is why, secondly, we study the method of trimmed moments (MTM), which can be viewed as the robust version of the method of moments estimators and may be more straightforward both analytically and computationally in estimating stable capital charges.

We end the thesis with different simulation studies, in which we compare the behavior of the estimated capital charges under the introduced estimation methods. Using the robust models, we show that the estimated capital charges remain more stable under contamination of the loss data (in comparison to the classical model), while keeping a high level of efficiency when the loss data follows the assumed model distribution exactly.

1.2 Structure of the chapters

Below we summarize the material covered in each of the chapters:

- *Chapter 2: Operational Risk and Basel II*
  We give a short introduction to operational risk and the Basel II regulatory framework. Under Basel II, three approaches can be used to assess the regulatory capital charge. We focus on the Advances Measurement Approaches (AMA), which are described in further detail.

- *Chapter 3: Loss distribution approach*
  We describe how to estimate the operational risk capital charge (Value-at-Risk or VaR) under the Loss distribution approach, which is done by applying the FFT method. We discuss how to appropriately model the frequency of the loss data and consider several parametric distributions that can be used to model the severity of the loss data. Furthermore, we introduce the constrained maximum likelihood approach (CML), which can be used to appropriately address the issue of loss data truncation. Finally, the introduced techniques are combined into a *classical* model to estimate the operational risk capital charge for both non-truncated and truncated loss data from several loss severity distributions.

- *Chapter 4: Robust statistics framework and the Influence Function*
  We show that the estimated capital charge under the *classical* model is highly sensitive to minor

contamination of the loss data. We define the robust statistics framework and we introduce one of the most important tools in robust statistics: the *influence function* (IF). We derive the influence function for the maximum likelihood estimators and the constrained maximum likelihood estimators (which are the classical estimation methods) for the lognormal, log-gamma and Generalized Pareto severity distribution. Furthermore, we propose the concept of Δ-VaR, which extends the results of the derived influence functions to the estimated capital charge.
In order to reduce the impact of small loss events on the estimated capital charge, we propose fitting a mixture of severity distributions to the loss data and we introduce an efficient method using the FFT method to estimate the resulting capital charge. This method is described in further detail for an exponential body distribution and different tail severity distributions.

- *Chapter 5: Robust estimation methods*
  We introduce the optimal bias robust estimators (OBRE) and describe the general algorithm to compute the OBRE, which is then assessed in further detail for the lognormal, log-gamma and Generalized Pareto distribution. For each severity distribution we derive the OBRE IFs both for non-truncated and truncated loss data and compute the corresponding Δ-VaR figures. We also calculate the relative efficiency of the OBRE with respect to the maximum likelihood and constrained maximum likelihood estimators at the exact model.
  Next, we introduce the method of trimmed moments and describe the general estimation procedure for non-truncated loss data. We propose a new (iterative) estimation procedure for truncated loss data. For non-truncated and truncated loss data from the lognormal and log-gamma severity distribution, we derive the MTM IFs and corresponding Δ-VaR figures. We also calculate the relative efficiency of the MTM estimators with respect to the maximum likelihood and constrained maximum likelihood estimators at the exact model.

- *Chapter 6: Simulation Study*
  The goal of the simulation study is to assess the behavior of the VaR measures under the different estimation methods that we have examined throughout the thesis. After describing the general setup of the simulation study, we combine the introduced techniques into a final *robust* model that can be used in practice to estimate the operational risk capital charge for both non-truncated and truncated loss data from several severity distributions. In the simulation studies we assess the efficiency of the estimated VaR measures at the exact model and the stability of the estimated VaR measures under contamination of the loss data for the lognormal, log-gamma and Generalized Pareto distribution.

Operational Risk and Basel II 2

2.1 What is an operational loss?

All businesses, and banks in particular, are vulnerable to losses resulting from operational failures. Most losses are relatively small in magnitude, examples of such operational losses include accidental accounting errors, minor credit card fraud or equipment failure. More severe operational loss events could be for example tax noncompliance, unauthorized trading activities, major internal fraudulent activities or business disruptions due to external events (e.g. natural disasters). For the majority of the banks, before the 1990s, the latter events did not occur frequently. And if they did, banks were capable of sustaining the losses without major consequences. This is essentially because operations within the banking industry, until roughly 20 years ago, have been subject to numerous restrictions and regulations. Consequently keeping trading volumes relatively modest and the diversity of operations limited. Therefore, the risk of losses resulting from operational failures has been perceived as relatively minor and traditionally banks relied upon insurance protection and internal control mechanisms to manage such operational risks. Globalization, complex financial products and changes in information technology over the last 20 years, combined with a growing number of high-magnitude operational loss events worldwide, have increased the importance of operational risk for the banking industry. To give the reader an idea of high-magnitude historical operational losses, we list several well-known examples.

- *The Barings Bank (1995, U.K.)*: One of the most famous operational loss events is the bankruptcy of the Barings Bank (estimated loss of GBP 1.3 billion). This is alleged to have occurred because trader, Nick Leeson, took an enormous position in futures and options, significantly exceeding his trading limits without approval. This case has been discussed in many papers, books and by Nick Leeson himself. Being in charge of the trade and the back office enabled Leeson to hide his position and create an illusion of large profits. He was motivated by large bonuses and the desire for status within the bank. It could be argued that this loss occurred due to a lack of controls (i.e. inadequate separation of the front and back office duties; and the absence of an accounting system enabling the settlements department in London to reconcile trades with clients' orders made worldwide).

- *The Enron Scandal (2001, U.S.)*: The collapse of Enron Corporation has been one of the largest bankruptcies in U.S. history. On January 25, 2001, the stock price of Enron reached its peak at $81.39 per share, and began to drop. In November, the price per share fell below $10 and Enron announced $600 million in losses from 1997 to 2000. On December 2, when the share price finally hit zero, Enron filed for bankruptcy. The board of directors of Enron blamed the failure on poor information from the accountants and the management. An investigation into the case conducted by the Securities and Exchange Commission in 2002 suggested that Enron may have overstated its assets by upto $24 billion due to poor accounting practices. This is an example of losses due to legal liability in combination with fraudulent activities.

- *9/11, (2001, U.S. and worldwide)*: On September 11, 2001, New York's World Trade Center, and the Pentagon became the targets of large-scale terrorist attacks. On the morning of September 11, two American Airlines jets were hijacked and used to crash into the Twin Towers of the

World Trade Center, causing them to collapse about an hour later. Two other airplanes were hijacked and one hit the Pentagon; the other crashed in Pennsylvania. Apart from its devastating civilian loss (e.g. Cantor Fitzgerald lost 700 of its employees), this incident also resulted in tremendous property loss. The Bank of New York's losses alone were estimated at $140 million. The financial losses due to 9/11 have been reported to be the costliest insured property loss in history, with current estimates of $40 billion to $70 billion. Other consequences have been business disruptions of the affected financial service companies, and a tremendous economic and political impact worldwide. This is an example of losses due to the damage to physical assets, business disruptions, and losses inflicted by external events.

- *The Econimic Crisis (2008 - 2009, worldwide)*: Many events of the recent global economic crisis have their root causes in operational failures within financial firms: mortgage fraud, inadequate assessment of model risk, failure to implement and maintain adequate systems and controls, 'bonus culture' motivating high short-term sales regardless of the long-term consequences for the company and its clients and so on.

2.2 Defining operational risk

In the previous section we provided a few examples of high-magnitude operational loss events. Now, we want to formalize the notion of operational risk.

Until recently, it has been believed that banks are exposed to two main risks; *credit risk* (counterparty failure) and *market risk* (loss due to changes in market indicators, such as equity prices, interest rates and exchange rates). Operational risk has been regarded as a mere part of the other risks. There was no widely accepted definition of operational risk when the Basel Committee on Banking Supervision [1] (BCBS) began discussions on operational risk management at the end of the 1990s. Often, operational risk was defined as any risk not categorized as market or credit risk. In January 2001, the BCBS issued a proposal for a New Basel Capital Accord[2] (referred to as Basel II) where operational risk was formally defined as a new category of risk, in addition to market and credit risk. The formal definition that is currently widely accepted was initially proposed by the British Bankers Association (2001) and adopted by the BCBS in January 2001:

> "Operational risk is the risk of loss resulting from inadequate or failed internal processes, people or systems or from external events."

2.2.1 Basel II approaches to quantify operational risk

Under the Basel II framework, operational risk is subject to a regulatory capital charge. This regulatory capital, estimated separately by every bank, is designed to reflect the exposure of each individual bank to operational risk. The accord defines and sets detailed instructions on the capital assessment of operational risk and proposes several approaches that banks may consider to estimate the operational capital charge. The Basel II framework is based on a three-pillar concept:

- *Pillar I (Minimal capital requirements)*: This pillar requires an explicit minimum capital allocated for operational risk that can be calculated using different approaches.

[1] The Basel Committee on Banking Supervision is a committee of banking supervisory authorities that was established by the central bank governors of the Group of Ten countries in 1974. It provides a forum for regular cooperation on banking supervisory matters.

[2] Further details on the definition of operational risk under the Basel II framework can be found in [2].

- *Pillar II (Supervisory review process)*: This pillar focuses on the supervision of the banks' systems and capital adequacy by regulatory authorities.

- *Pillar III (Market discipline)*: The objective of this pillar is to establish market discipline through public disclosure of risk measures and other relevant information on risk management.

In this thesis we will only focus on Pillar I and consider probabilistic models to compute the regulatory capital charge. Under Basel II, three approaches can be used to assess the operational risk annual capital charge

1. *Basic Indicator Approach*: The Basic indicator approach (BIA) is the simplest approach. The operational risk capital charge under the BIA is calculated as a fixed percentage of the average over the previous three years of positive annual gross income, where the gross income is defined by the BCBS as net interest income plus net non interest income. The total capital charge K_{BIA} can be written as

$$K_{\text{BIA}} = \frac{\alpha}{n} \sum_{j=1}^{n} \text{GI}_j, \quad \text{with} \quad 0 \leq n \leq 3 \tag{2.1}$$

   with GI the gross income, n the number of the previous three years for which GI is positive and α a fixed percentage. α is currently set by the BCBS at 0.15.

2. *The Standardized Approach*: In the general Standardized approach (SA), banks' activities are divided into eight business lines, which are listed in Table 2.1. Within each business line, the gross income (GI) is taken as an indicator of the operational risk exposure. The capital charge for each business line is calculated by multiplying GI by a factor β_i, assigned to that business line. The total capital charge is then calculated as the three-year average of the maximum of: the summation of the capital charges across each business line and zero. The total capital charge K_{SA} can be expressed as

$$K_{\text{SA}} = \frac{1}{3} \sum_{j=1}^{3} \max \left\{ \sum_{i=1}^{8} \beta_i \text{GI}_i(j), 0 \right\} \tag{2.2}$$

i	Business line, BL(i)	β_i
1	Corporate finance	0.18
2	Trading and sales	0.18
3	Retail banking	0.12
4	Commercial banking	0.15
5	Payment and settlement	0.18
6	Agency services	0.15
7	Asset management	0.12
8	Retail brokerage	0.12

Tab. 2.1.: Basel II business lines (BL). $\beta_1, \ldots \beta_8$ are the business line factors used in the Basel II standardized approach.

3. *The Advanced Measurement Approaches*: In the Advanced Measurement Approaches (AMA) a bank can calculate the capital charge using an internally developed model subject to regulatory approval. A bank must demonstrate that its operational risk measure is evaluated for a one-year

holding period at a high[3] confidence level. The AMA are the most complex and advanced, as the resulting capital charge is the risk measure directly derived from the bank's internal loss data history and employs quantitative and qualitative aspects of the bank's risk measurement system for the assessment of the regulatory capital charge.

The BIA and the SA are technically not very sophisticated, and therefore easy to implement and use. They are often referred to as so-called *top-down approaches* in the sense that the capital charge is allocated according to a fixed proportion of income. The AMA are called *bottom-up approaches* in the sense that the capital charge is estimated from actual internal loss data. Internationally active banks with diverse business activities are mandatory to adopt the AMA, while smaller sized domestic banks are allowed to follow the BIA or the SA, at least at the primary stage of the implementation. Once a bank adopts a more advanced approach, it is not allowed to switch back to a simpler approach. The BCBS [3] expects that the AMA will be uniformly adopted by all banks in the near future.

2.3 AMA in detail

Many models have been suggested for modeling operational risk under the Basel II AMA. As noted in the previous section, two conceptual approaches are the top-down and bottom-up approaches. An ideal internal operational risk assessment procedure would be a balanced approach, and include both top-down and bottom-up elements into the analysis. Overviews of the models introduced below can be found in Chernobai, Rachev and Fabozzi [10]

2.3.1 Top-down approaches

Top-down approaches quantify operational risk without attempting to identify the events or causes of losses. That is, the losses are measured on a *macro*-level. The principal advantage of this approach is that little effort is required with collecting data and quantifying the operational risk capital charge. Several examples of the top-down models are:

- *Multifactor equity pricing models*: This approach (also referred to as multifactor models) assumes *market efficiency*, where the current asset price (stock price of the company) reflects all relevant information. Then the stock return process is assumed to be driven by many factors related to the market, credit and other non-operational risks. The residual term of this regression is treated as due to operational risk.

- *Capital asset pricing model (CAPM)*: In the CAPM the asset risk premium is quantified. This is the difference between expected return and risk-free return. The contributions from credit and market risks are measured and the operational risk is treated as the residual.

- *Scenario analysis and stress testing models*: These models, also called *expert judgment models* are estimated based on what-if scenarios generated with reference to expert opinion, external data, catastrophic events occurred in other banks, or imaginary high-magnitude events. Experts estimate the expected risk amounts and their associated probabilities of occurrence.

- *Risk indicator models*: These models link operational risk and exposure indicators such as gross income, volume of transactions, number of staff, etc. The BIA and SA are examples of a single indicator and multi-indicator model respectively.

[3] Typically the confidence level is set at 99.9%, which then corresponds to modeling the one-in-one-thousand years worst operational loss event

2.3.2 Bottom-up approaches

Bottom-up approaches try to quantify operational risk on a so-called *micro*-level on identified internal loss events. Information on these loss events is then incorporated into the overall capital charge calculation. The advantage of bottom-up approaches over top-down approaches lies in their ability to explain the mechanism of how and why operational risk is formed within an institution. Generally speaking, there are two types of bottom-up approaches: process based models and loss distribution approach (LDA) models. Below we give several examples of process based models:

- *Causal networks and Bayesian belief networks*: These are typically subjective models. For each banking activity, a tree of events that may lead to an operational loss is constructed. The probability of each event is specified by an expert. Bayesian networks account for causal dependencies enabling linkage of the operational conditions to the probability and severity of the losses. There is a view that these models are certainly useful for operational risk management, but not as models for quantification of the regulatory capital charge.

- *Multifactor causal models*: These models are based on regression of operational losses on a number of control factors (explanatory variables), such as the number of staff, the number of transactions, skill level of employees, etc. These factors are then used to predict future losses assuming that the factors are known for the next period of time.

- *Reliability models*: These models quantify the probability that a system will operate satisfactorily for a certain period of time. These are the models considered in operational research to study trustworthiness of system elements. This is relevant to many processes in operational risk, for example modeling the reliability of transaction processing systems. For calculations of the capital charge, it is not used as a standalone model, but rather as a part of other models.

The Loss Distribution Approach (LDA) is currently the most widely accepted model under the Basel II AMA to estimate the regulatory capital charge, it was also suggested by the Basel Committee in 2001 (see BCBS [2]). It is based on modeling frequencies and severities of loss events seperately. The aggregated annual loss distribution is then calculated as the compound distribution of frequency and severity, which are both modeled as random variables. Finally, the capital charge is estimated as the *Value-at-Risk*[4], which is basically a high quantile of the aggregated annual loss distribution. In the remainder of the thesis we follow the Loss Distribution Approach, since this model is most applied in practice and in the literature. In the following chapter we introduce the Loss Distribution Approach in detail and describe how to compute the regulatory capital charge. Furthermore, we discuss how to model the severity and frequency of the losses and how to account for the fact that the loss data is subject to lower recording thresholds.

[4] Value-at-Risk is defined in the next chapter

Loss distribution approach 3

3.1 What is Value-at-Risk?

In quantitative risk measurement we are often concerned with a question of the form: "What is the maximum amount that I can expect to lose with a certain probability over a given time horizon?" This can come from a financial manager's point of view or from a regulator's perspective. Where the regulator wants to ensure that banks hold sufficient quantities of reserves (i.e. regulatory capital) to cover most large-scale operational loss events. According to the BCBS [2] the following should be ensured:

> "Whatever approach is used, a bank must demonstrate that its operational risk measure meets a soundness standard comparable to that of the internal ratings based approach for credit risk, (i.e. comparable to a one year holding period and a 99.9 percent confidence interval)."

In risk management to assess this question we can estimate the *Value-at-Risk* (VaR) measure. VaR has a long history in both market and credit risk and also emerged as one of the most important risk measures in estimating the operational risk regulatory capital.

Definition 3.1. (Value-at-Risk) The VaR of a random variable X with loss distribution F at the α-th probability level, $\text{VaR}_\alpha(X)$, is defined as the α-th quantile of the distribution of X, i.e.

$$\text{VaR}_\alpha(X) = F^{-1}(\alpha) \tag{3.1}$$

Intuitively, VaR determines the worst possible loss with a given confidence level α and for a given time frame. For example, a one-year 99.9% VaR is the total amount of loss that may be exceeded only by 0.1% of all potential losses occuring over a one-year period. That is, we expect a possible loss to exceed 99.9% VaR only once every thousand years.

According to the Basel II framework the *regulatory* capital must correspond to the 99.9% VaR over a one-year holding period. *Economic* capital, is the amount that market forces imply for the risk, typically to calculate economic capital banks consider the 99.95-99.97% VaR over a one-year holding period. Hereafter, in the LDA, we compute regulatory capital as $\text{VaR}_{0.999}(X)$, where the distribution of X is the aggregated annual loss distribution, which is defined in the following section.

3.2 Loss Distribution model

As described earlier, in the LDA we model the frequencies and severities of the loss events as separate random variables. Next, we combine the frequency distribution and severity distribution into a compound distribution, which for a one-year period of loss events is referred to as the aggregated annual loss distribution.

Suppose that the severity of the loss events X are independent and identically distributed (IID) according to the severity loss distribution F. Suppose that the annual number of the loss events N are also random variables with values in $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ according to the frequency distribution Q (typically modeled by a counting process, such as a Poisson distribution). Annual losses may be expressed as the random sum

$$S_N = \sum_{j=1}^{N} X_j \tag{3.2}$$

Where the X_j are IID losses, independent of the random frequency N.
If we write $F^{(n)\star}$ as the n-fold convolution of F with itself:

$$F^{(n)\star}(x) = P\left(\sum_{j=1}^{n} X_j \leq x\right) \tag{3.3}$$

Then the cumulative distribution function of the annual loss distribution can be written as

$$Q \vee F(x) = P(S_N \leq x) = \begin{cases} \sum_{n=1}^{\infty} P(N=n)F^{(n)\star}(x) & \text{for} \quad x > 0 \\ P(N=0) & \text{for} \quad x = 0 \end{cases} \tag{3.4}$$

Where we use the notation $Q \vee F(x)$ for the compound loss distribution, with frequency distribution Q and severity distribution F.
The population mean and variance of the compound distribution can easily be found (by means of the tower property), due to assumed independence between the frequency and severity distributions:

$$\begin{aligned} \mathbb{E}[S_N] &= \mathbb{E}[N]\mathbb{E}[X_j] && (3.5) \\ \text{var}(S_N) &= \mathbb{E}[N]\text{var}(X_j) + \text{var}(N)\left(\mathbb{E}[X_j]\right)^2 && (3.6) \end{aligned}$$

Estimation of the annual loss distribution by modelling frequency and severity of losses is a well-known actuarial technique, see for example Klugman, Panjer and Willmot [23]. Typically, under the Basel II framework, we compute different annual loss distributions for different business lines (same as in the SA) and event types[1]. We will call such a combination of business line/event type a single *unit of measure*. In practice for each relevant unit of measure $i = 1, \ldots, k$ we compute different random annual losses $S_{N,i}$. Although the losses within a unit of measure are assumed to be independent, losses across different units of measure are generally not. The total annual loss for a bank with k units of measure is given by

$$L = \sum_{i=1}^{k} S_{N,i} \tag{3.7}$$

To determine the distribution of L, a dependence structure (such as a copula[2]) is estimated for the $S_{N,i}$'s, see for example Panjer [28].
In this form, the LDA is currently the most widely accepted framework for operational risk modeling under the Basel II AMA. For the sake of simplicity, in the remainder of the thesis we consider loss data from a single unit of measure. The contamination of the loss data across multiple units of measure might be a topic of future research.

3.3 Fast Fourier Transform method (FFT)

As we can see in Eq.(3.4) the distribution function of the compound loss distribution is nonlinear in X_j and N. Therefore, except for some simple scenarios, analytic expressions for the compound

[1] Under the Basel II framework we usually consider seven different event types, which can be characterized for each business line, the event types are given by: (1) internal fraud, (2) external fraud, (3) employment practices and workplace safety, (4) clients, products and business practices, (5) damage to physical assets, (6) business disruption and system failures, (7) execution, delivery and process management.

[2] Copula functions can be used to model multivariate dependence among random variables, in this way we can transform the marginal distributions of the random variables into the joint distribution.

loss distribution function do not exist. This is a classical problem in risk theory. With modern computer processing power, these distributions can be calculated virtually exactly using numerical algorithms. The method we will apply throughout the rest of the thesis to calculate the VaR measures (high quantiles of the compound distribution), is via Fourier inversion techniques. It is referred to as the Fast Fourier Transform (FFT) method. The FFT method is an efficient method to calculate compound distributions via the inversion of the characteristic function. The method has been known for many decades and originates from the signal processing field, but the application in operational risk modeling is only a recent development. In practice, financial risk managers regard the method as difficult and therefore rarely apply it. Typically, computationally intensive algorithms are preferred, such as Monte Carlo simulation or very crude approximations are made, using for instance the Single-Loss Approximation (SLA). In Appendix A.1 we describe the Monte Carlo method and the SLA and compare their performance to the FFT method. From the results it is clear that in practice the FFT method will always outperform both the Monte Carlo method and the SLA.
Below we describe the essential steps required for successful implementation of the FFT method in computing VaR of the aggregated annual loss distribution.

3.3.1 Discretization

To work with an FFT based algorithm, the possible severities X must be concentrated on a lattice $h\mathbb{N}_0 = \{0, h, 2h, 3h, \ldots\}$, where h is some strictly positive (small) bandwidth. Typically, we model a continuous severity distribution F for the severities and thus an initial discretization is required. We choose a suitably small bandwith $h > 0$ and replace F with a distribution $F_h = \{f_{h,j}\}_{j\in\mathbb{N}_0}$ on $h\mathbb{N}_0$. One method that we can use is the *rounding method*, where the severities are rounded to the closest integer multiple of h, that is

$$f_{h,j} = F\left(jh + \frac{h}{2}\right) - F\left(jh - \frac{h}{2}\right) \tag{3.8}$$

If we use the following notation:

$$q_j = P(N = j), \quad f_n = P(X_i = n), \quad g_n = P(S_N = n) \tag{3.9}$$

Where the frequency N has distribution Q, the severity X_i has distribution F and the compound loss S_N has distribution $Q \vee F$. Then the compound losses following the discrete version of $Q \vee F$ can be written as

$$g_n = \sum_{j=0}^{\infty} q_j f_n^{\star j} \tag{3.10}$$

where

$$f_n^{\star j} = \begin{cases} 1 & \text{if} \quad j = 0 \quad \text{and} \quad n = 0 \\ 0 & \text{if} \quad j = 0 \quad \text{and} \quad n \in \mathbb{N} \\ \sum_{i=0}^{n} f_{n-i}^{\star(j-1)} f_i & \text{if} \quad j \geq 0 \end{cases} \tag{3.11}$$

3.3.2 FFT based algorithm

The characteristic function[3] of the compound distribution S_N can be expressed in terms of the characteristic function of the severity X_1, since the severities X_i are IID and are also independent of the frequency N, that is

$$\phi_{S_N}(t) = \mathbb{E}[e^{itS_N}] = \mathbb{E}\left[\mathbb{E}\left[e^{it\sum_{i=1}^{N} X_i}|N\right]\right] = \mathbb{E}\left[(\phi_{X_1}(t))^N\right] = \mathcal{P}_N(\phi_{X_1}(t)) \tag{3.12}$$

[3] The characteristic function $\phi_X(t)$ for a random variable X, is defined as $\phi_X(t) = \mathbb{E}[\exp(itX)]$

Here $\mathcal{P}_N$ denotes the probability generating function[4] of N. For random variables that admit a density, such as S_N and X_1, the characteristic function is equal to the Fourier transform of the probability density function. We apply the discrete Fourier transform (DFT) and the inverse discrete Fourier Transform, with discretized severities on the lattice $\{0, h, 2h, \ldots, (M-1)h\}$ upto some truncation point M.

Definition 3.2. (Discrete Fourier Transform) For the discretized sequence $f_0, f_1, \ldots, f_{M-1}$ up to some truncation point M, the discrete Fourier transform

$$\widehat{f} = (\widehat{f_0}, \widehat{f_1}, \ldots, \widehat{f}_{M-1}) \tag{3.13}$$

is defined by

$$\widehat{f_j} = \sum_{k=0}^{M-1} f_k \exp\left(\frac{2\pi i}{M} kj\right), \quad j = 0, 1, \ldots, M-1 \tag{3.14}$$

Theorem 3.3.1. (Fourier Inversion Theorem) Given the DFT $\widehat{f}$, the original sequence f can be reconstructed by

$$f_j = \frac{1}{M} \sum_{k=0}^{M-1} \widehat{f_k} \exp\left(-\frac{2\pi i}{M} kj\right), \quad j = 0, 1, \ldots. M-1 \tag{3.15}$$

If M is a power of 2, the DFT and the inverse DFT can be computed efficiently using one of the various FFT algorithms (we refer to Embrechts and Frei [15] or Shevchenko [29] for further details).
If we consider the right-hand side of Eq.(3.12), the characteristic function of the severity distribution X_1 can be computed by the DFT of the discretized severities $\{f_j\}_{0 \leq j < M}$. Thus, we can take the inverse DFT of

$$\widehat{g^M} = \mathcal{P}_N(\widehat{f}) \tag{3.16}$$

as an approximation for $g = (g_0, g_1, \ldots, g_{M-1})$, which are the discretized compound loss probabilities upto the truncation point M.
If $\sum_{j=0}^{M-1} f_j = 1$, we lose no mass due to the truncation at the point M and the compound distribution is exactly evaluated on the cyclic group $\mathbb{Z}/M\mathbb{Z}$. Compound mass which lies at M and beyond will be *wrapped around* and will falsely appear in the range $0, \ldots, M-1$. For heavy-tailed severity distributions this wrap-around error, also referred to as the *aliasing* error, can become quite an issue.
Grübel and Hermesmeier [18] suggest to reduce the aliasing error by applying a tilting transformation on f in order to increase the tail decay[5].
The following tilting transformation is proposed: fix some $\theta > 0$ and set

$$E_\theta f = \left(e^{-\theta j} f_j\right)_{j=0,1,\ldots,M-1} \tag{3.17}$$

The tail of $E_\theta f$ decays at an exponential rate and thus potentially much faster than the tail of f. The operator E_θ commutes with convolution and therefore we have

$$Q \vee F = E_{-\theta}(Q \vee E_\theta F) \tag{3.18}$$

The process of constructing the compound distribution using the FFT method can be summarized as follows:

- Choose a bandwidth $h > 0$, truncation point $M \in \mathbb{N}$ and tilting parameter $\theta > 0$.

[4] For X a discrete random variable, taking values in the non-negative integers $\{0, 1, \ldots\}$, the probability generating function of X is defined as $\mathcal{P}_X(z) = \mathbb{E}[z^X] = \sum_{i=0}^{\infty} p(x) z^x$

[5] By tail decay, we mean the speed at which the probability density decreases to zero in the tails of the distribution.

- If F is a continuous severity distribution, discretize according to the rounding method and set $f = (f_0, f_1, \ldots, f_{M-1})$, which are the probability values of the severity distribution on the lattice $\{0, h, 2h, \ldots, (M-1)h\}$.

- By the exponential tilting transformation we compute $E_\theta f = \left(e^{-\theta j} f_j\right)_{j=0,1,\ldots,M-1}$.

- Compute the DFT $\widehat{E_\theta f}$ of the tilted sequence $E_\theta f$.

- Take the inverse DFT of $\mathcal{P}_N(\widehat{E_\theta f})$ and untilt by applying $E_{-\theta}$. We have now constructed $g = (g_0, g_1, \ldots, g_{M-1})$, which are the probability values of the annual loss distribution on the lattice $\{0, h, 2h, \ldots, (M-1)h\}$.

It is found that the tilting is of much practical value and can reduce the aliasing error to a minimum. Embrechts and Frei [15] show that the choice $M\theta \approx 20$ is reasonable in most cases, we will therefore choose $\theta = 20/M$.

The final step is to estimate VaR_α, which is the quantile of the compound distribution at the α-th probability level. Using the constructed sequence g, the α-th quantile is found by solving for M_0:

$$\sum_{i=0}^{M_0-1} g_i < \alpha < \sum_{i=0}^{M_0} g_i \tag{3.19}$$

and we put $\text{VaR}_\alpha = hM_0$. For given bandwidth h, if there is no aliasing error, the error of the resulting VaR_α is bounded by $\pm h$. The calculations are carried out numerically in *R* (version 2.15.2). Below we give an outline of the implementation in *R*, where we evaluate the compound distribution $\text{Pois}(\lambda) \vee \mathcal{LN}(\mu, \sigma)$, i.e. we model the annual loss frequencies by a Poisson distribution with intensity λ and the annual loss severities by a lognormal distribution with mean μ and standard deviation σ. The probability generating function of $N \sim \text{Pois}(\lambda)$ is given by

$$\mathcal{P}_N(z) = \exp(\lambda(z-1)) \tag{3.20}$$

The implementation in *R* can be summarized as follows

```
> #Discretize the severity distribution
> f <- discretize(plnorm(x, mu, sigma), from=0, to=h*M, by=h, method="rounding")
>
> #Apply the tilting transformation
> f <- exp(-theta * (0 : (M-1))) * f
>
> #Compute the DFT
> fhat <- fft(f, inverse=FALSE)
>
> #Define the pgf, take its inverse DFT and untilt
> P <- exp(lambda * (fhat - 1))
> g <- exp(theta * (0:(M-1))) * (1/M * fft(P, inverse=TRUE))
>
> #Calculate VaR
> if(sum(g) < alpha){ return(NULL)
        } else {
                M0 <- Solution Eq.(3.22)
                return(h * M0)
        }
```

3.4 Modeling the frequency distribution

In the previous example we considered modeling the frequency distribution by a (homogeneous) Poisson distribution. The Poisson distribution is a common choice in practice to model the annual loss frequencies, since it is easy to understand and fit to given loss frequency data[6]. Other parametric distributions that are often applied to model the loss frequency are the binomial and negative binomial distribution. A nice property is that the binomial distribution has its mean less than the variance; the negative binomial distribution has its mean larger than the variance; and the mean of the Poisson distribution is equal to the variance. This can be used in practice as a criterion to choose a suitable frequency distribution. In Table 3.1 below we list the probability mass functions and probability generating functions (which can be used in the implementation of the FFT method) of the Poisson, binomial and negative binomial distribution.

Distribution	$P(X = k)$	$\mathcal{P}_N(z)$
$\text{Pois}(\lambda)$	$\lambda^k e^{-\lambda}/k!$	$e^{\lambda(z-1))}$
$\text{Bin}(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	$((1-p) + pz)^n$
$\text{NegBin}(n, p)$	$\binom{n+k-1}{k} p^n (1-p)^k$	$\left(\frac{p}{1-(1-p)z}\right)^n$

Tab. 3.1.: Probability mass function and probability generating function of the Poisson, binomial and negative binomial distribution.

According to Carillo-Menéndez [7], the choice of frequency distribution has no serious impact on the estimated capital charge. The main difficulty in estimating accurate and stable operational risk capital charges is the modeling of the loss severities. Hereafter, we consistently model the annual loss frequencies by a homogeneous Poisson distribution, with fixed intensity parameter $\lambda = 25$ (corresponding to an average of 25 loss observations per year).

Remark 3.0.1. An important issue in practice is related to the recording of the data. In market and credit risk models, the impact of a relevant event is almost immediately reflected in the market and credit returns. In an ideal scenario, banks exactly know the severity of an operational loss at the moment it takes place, and would also record the loss at this moment. However, in practice this is impossible, because it takes time for the losses to accumulate after an event takes place. Therefore it may take days, months or even years for the full impact of a particular operational loss to be evaluated. Hence, generally there is a difference between the occurrence of an event and the time at which the incurred loss is being recorded. We will not concern ourselves with this problem in the rest of the thesis, and assume that there is no discrepancy in the occurrence and the recording of operational loss data.

[6] Suppose we have data $X_1, \ldots, X_n$ iid from a Poisson distribution, the maximum likelihood estimator of the single intensity parameter λ is then given by $\widehat{\lambda} = \overline{X}_n$.

3.5 Modeling the severity distribution

The main difficulty in estimating the operational risk capital charge under the LDA framework is the accurate modeling of the severity distribution. In this section, we argue that this is due to the lack of coherent historical loss data. Ideally, we would estimate the VaR measure empirically. But, estimating the compound loss 0.999-quantile using observed loss events only, is impossible in practice. This has to do with the insufficiency of the collected historical loss data, i.e. we wish to estimate the one-in-thousand years worst loss event, using only five to ten years of recorded operational loss data. Furthermore, an important property of the operational loss data is that it is very heavy-tailed. Heavy-tailedness[7] can be observed in nearly every category of operational losses, regardless of the event types or the business line, this property of operational loss data has been discussed in several articles, see Cope et al. [11] or Shevchenko [29]. To illustrate that empirical estimation of the VaR measure is difficult, we give a rough estimate of the number of data points that would be needed for such a calculation.

Assume that we observe losses $X_1, X_2, \ldots, X_n$ drawn from a heavy-tailed distribution with common density f. The quantile q_α at probability level α is empirically estimated as $\widehat{q}_\alpha = X_{(\lfloor n\alpha \rfloor + 1)}$, where $X_{(i)}$ is the i-th order statistic[8] of the data sample. According to Glasserman [17], the standard deviation of this empirical estimate is given by

$$\mathrm{sd}[\widehat{q}_\alpha] = \frac{\sqrt{\alpha(1-\alpha)}}{f(q_\alpha)\sqrt{n}} \tag{3.21}$$

Therefore, to calculate the quantile within a relative error $\epsilon = 2\mathrm{sd}[\widehat{q}_\alpha]/q_\alpha$, the number of observations that is approximately needed, is given by

$$n = \frac{4\alpha(1-\alpha)}{\epsilon^2 (f(q_\alpha) q_\alpha)^2} \tag{3.22}$$

Suppose that we observe operational loss data from the lognormal distribution $\mathcal{LN}(\mu = 10.95, \sigma = 1.75)$. Then using the above approximation, to achieve 10% accuracy ($\epsilon = 0.1$) in the 0.999 quantile estimate, we would require approximately $n = 107,943$ observations. For the log-gamma[9] distribution $\mathcal{LG}(a = 34.5, b = 3.5)$, with accuracy 10%, $n = 185,220$ observations are needed and for the Generalized Pareto distribution $\mathrm{GPD}(\xi = 0.65, \beta = 57500)$, with accuracy 10%, we would need $n = 172,685$ observations. Compare this to loss data from the exponential distribution with rate[10] $\lambda = 10^{-5}$, where we would need only around $n = 8,350$ data points to estimate the 0.999 quantile within 10% relative error. This example clearly shows that it is much more difficult to accurately estimate quantiles for heavy-tailed distributions than for light-tailed distributions using only observed loss data. For a more direct estimate of empirical data requirements consider Table 3.2 below:

[7] Operational risk losses are typically modeled by so-called *heavy-tailed* distributions. Heavy-tailed distributions are probability distributions that are not exponentially bounded.

[8] The i-th order statistic of a data sample is equal to its i-th smallest value.

[9] We define the lognormal, log-gamma and Generalized Pareto distribution in detail in the next section.

[10] For $\lambda = 10^{-5}$ the exponential distribution shows similar behavior as the lognormal, log-gamma and Generalized Pareto distribution (with given parameter values) on a large part of the body. Obviously, the behavior in the tails is very different.

		Width 95% CI for the 0.999 quantile			
	$q_{0.999}$	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
$\mathcal{LN}(10.95, 1.75)$	12,710,088	20,611,085	8,046,360	2,571,084	817,297
$\mathcal{L}G(34.5, 3.5)$	7,764,009	17,347,933	6,500,521	2,106,195	648,426
$\mathrm{GPD}(0.65, 57500)$	7,795,681	17,156,564	6,335,316	1,995,914	628,875

Tab. 3.2.: Estimated value of the 0.999 quantile for the lognormal, log-gamma and GPD and widths of a 95% confidence interval for the 0.999 quantile as a function of the number of data points collected.

In the calculations above it is seen that a large number of loss data points is needed to get accurate empirical estimates of high quantiles. To make it worse, this is only the amount of loss data we need to accurately estimate high quantiles of the severity distribution. And we do not wish to compute the 0.999 quantile of the loss severity distribution, we wish to calculate the 0.999 quantile of the compound loss distribution. To estimate this figure accurately, we would need to collect far more data than the largest operational risk database[11] has collected up to date. We conclude that we are unable to accurately estimate the VaR measure using empirical estimation. Therefore, we will resort to parametric models in modeling the severity of the operational loss data.

3.5.1 Parametric models

In this section we consider several heavy-tailed distributions that are able to appropriately capture the relevant features of operational loss data. We list the probability density functions and their maximum likelihood estimators. In practice, maximum likelihood estimation (MLE) is the most widely used method to fit parametric distributions under the LDA framework. Under the *classical* model, we will therefore consider maximum likelihood estimation to estimate the distribution parameters of the severity distribution.

- *The lognormal distribution*. This is one of the standard distributions used in insurance to model large claims and is commonly used to model operational loss severities. It is widely used, since its behavior is easy to understand (due to its relation to the normal distribution). However, the use of the lognormal distribution has been criticized for not being sufficiently heavy-tailed for the operational loss data. Its probability density function is given by

$$f(x|\mu,\sigma) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right) \tag{3.23}$$

  If $X_1, X_2, \dots, X_n$ are IID observations from a lognormal distribution, the maximum likelihood estimators are given by

$$\widehat{\mu} = \frac{1}{n}\sum_i \log(X_i), \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_i (\log(X_i) - \widehat{\mu})^2 \tag{3.24}$$

[11] At 31 December 2012, the ORX Data Consortium contained 299,672 operational loss events to a total value of € 151,559,050,244.

- *The log-gamma distribution*. This is also a common model for the operational loss severities. The log-gamma distribution is more flexible[12] than the lognormal distribution in modeling the heavy-tailed loss data and relates to the ordinary Gamma distribution. The probability density function of the log-gamma distribution is given by

$$f(x|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \log(x)^{\alpha-1} x^{-(\beta+1)} \tag{3.25}$$

With shape parameter $\alpha > 0$, rate parameter $\beta > 0$ and $\Gamma(\alpha)$ the Gamma function[13] evaluated at α. If $X_1, X_2, \dots, X_n$ are IID observations from a log-gamma distribution its log-likelihood function can be written as

$$\begin{aligned} \ell(\alpha, \beta|\boldsymbol{X}) &= n\alpha\log(\beta) - n\log(\Gamma(\alpha)) & (3.27) \\ & \quad -(\beta+1)\sum_i \log(X_i) + (\alpha-1)\sum_i \log(\log(X_i)) & (3.28) \end{aligned}$$

There is no closed form expression for the maximum likelihood estimators α and β. Thus, we can optimize the above equation numerically by means of the Newton-Raphson method[14].

- *The Generalized Pareto distribution*. Some authors have proposed extreme value theory (EVT) as the appropriate distributional form of the severity distribution, one that would properly account for the heavy tails in the data. In this approach losses above an appropriately chosen threshold are modeled by the Generalized Pareto distribution. An advantage of this methodology is that it relies on the well-established theoretical framework for the universal asymptotic behavior of probability distributions. On the other hand, there is an active debate on whether a naive application of EVT is appropriate for the quantification of operational risk. First, it is possible that the losses observed empirically do not correspond to the asymptotic regime (i.e. the extreme losses do not behave according to the extreme-value distributions). Furthermore, the value of the capital charge is often overestimated. This is due to the fact that the tail behavior of the GPD is in general much heavier than that of the lognormal and log-gamma distribution, so the resulting VaR measures will also be much larger. Nevertheless, in the literature the GPD is frequently used to model operational loss severities. The density of the GPD is given by

$$f(x|\xi, \beta) = \begin{cases} \frac{1}{\beta}(1+\xi x/\beta)^{-\frac{1}{\xi}-1} & \text{for} \quad \xi \neq 0 \\ \frac{1}{\beta}\exp(-x/\beta) & \text{for} \quad \xi = 0 \end{cases} \tag{3.29}$$

If $\xi > 0$, the GPD right tail is unbounded and the distribution is heavy-tailed, so that some moments do not exist. In particular, if $\xi \geq 1/m$, the m-th and higher moments do not exist. The analysis of operational risk data in Moscadelli [25] even suggests cases of $\xi \geq 1$ for some business lines, i.e. infinite mean distributions.
If $\xi \leq 0$, the GPD has a bounded right tail, with domain $x \in [0, -\beta/\xi]$. It seems that this case is not relevant to modeling operational risk as all reported results indicate a non-negative shape

[12] The log-gamma distribution is more flexible in the sense that the log-gamma distribution is able to capture heavier tail behavior than the lognormal distribution, see Zhou et al. [33].

[13] The gamma function is defined for all complex numbers except the negative integers and zero. For complex numbers with a positive real part, it is defined via an improper integral that converges according to

$$\Gamma(z) := \int_0^{\infty} t^{z-1} e^{-t}\, dt \tag{3.26}$$

[14] In numerical analysis, the Newton-Raphson method, named after Isaac Newton and Joseph Raphson, is a method for finding successively better approximations to the roots of a real-valued function, see Ypma [32] for a detailed explanation.

parameter.
Given IID observations $X_1, X_2, \ldots, X_n$ from the GPD the log-likelihood function is given by

$$\ell(\xi, \beta | \boldsymbol{X}) = -n \log(\beta) - \left(1 + \frac{1}{\xi}\right) \sum_i \log\left(1 + \xi \frac{X_i}{\beta}\right) \tag{3.30}$$

Again, we have no closed form expressions for the maximum likelihood estimators. Thus, the log-likelihood function can be maximized numerically in order to find the maximum likelihood estimators ξ and β.

- Alternative parametric distributions that can be used to model the severity are: the Weibull, the log-Weibull, g-and-h, alpha-stable, loglogistic, Burr, Inverse Gaussian. All these distributions belong to the class of heavy-tailed distributions. In this thesis we will use the three distributions described previously to model the operational loss severities. For each experiment in the following chapters, we will try to consider it separately for the case of the lognormal distribution, log-gamma distribution and Generalized Pareto distribution respectively.

3.6 Data truncation

In Remark 3.0.1 we noted that in an ideal situation, the data collection process results in all operational loss events being detected and timely recorded. We discussed that operational risk losses may take a long time to accumulate, which makes the recording of the operational losses difficult. If we assume that there is no discrepancy in the occurrence and the recording of the losses, we still have to decide which operational losses we record.
It is clear that it is practically impossible to record *every* occurred operational loss, since in theory an operational loss can be as small as $1. An important feature of operational risk loss data is therefore, that in practice, it is subject to lower recording thresholds, so that only data above a certain amount enters the database. For example, the largest operational risk database, the ORX consortium only records operational losses above € 25,000.
In this sense, the data available for the estimation appears to be left-truncated. Left-truncation of the data must be appropriately addressed in the estimation process, in order to determine a correct capital charge.

3.6.1 Frequency distribution of truncated losses

Let us first consider the effect of data truncation on the modeling of the frequency distribution. As noted before, in the remainder of the thesis we model the annual loss frequencies by a Poisson distribution. A useful property of the Poisson distribution is that its type is preserved in the case of loss truncation.

Proposition 3.6.1. Consider IID losses $X_1, X_2, \ldots, X_N$ with common severity distribution F independent of the random loss frequency N, where $N \sim \mathrm{Pois}(\lambda)$. Suppose that only losses above the threshold H are recorded. If we denote the frequency of the losses above the recording threshold H by N_H. Then

$$N_H \sim \mathrm{Pois}(\lambda(1 - F(H))) \tag{3.31}$$

Proof. The number of events above the truncation threshold H can be written as

$$N_H = I_1 + \ldots + I_N \tag{3.32}$$

where I_j are IID indicator random variables

$$I_j = \begin{cases} 1 & \text{if} \quad X_j > H \\ 0 & \text{if} \quad X_j \leq H \end{cases} \tag{3.33}$$

with probability generating function

$$\begin{aligned} \mathcal{P}_I(z) &= \mathbb{E}[z^I] && (3.34) \\ &= F(H) + z(1 - F(H)) && (3.35) \\ &= 1 + (1 - F(H))(z - 1) && (3.36) \end{aligned}$$

According to Eq.(3.12) the probability generating function of the number of events above the threshold H can be written as

$$\mathcal{P}_{N_H}(z) = \mathcal{P}_N\left(\mathcal{P}_I(z)\right) \tag{3.37}$$

We recall that the probability generating function of the Poisson distribution with intensity λ is given by $\mathcal{P}_N(z|\lambda) = \exp(\lambda(z-1))$. Therefore, we can write

$$\mathcal{P}_{N_H}(z|\lambda) = \exp(\lambda(1 - F(H))(z - 1)) = \mathcal{P}_N(z|\lambda(1 - F(H)) \tag{3.38}$$

Thus, it is found that both N and N_L have the same distribution type and if $N \sim \text{Pois}(\lambda)$ then $N_H \sim \text{Pois}(\lambda(1 - F(H))$. It can be checked directly that this relationship also holds for the binomial and negative binomial distribution (see Shevchenko [29]). □

We conclude that it is usually easy to modify the frequency distribution for truncated loss data.
In the next section we discuss how to appropriately model the severity distribution for truncated loss data.

3.6.2 Severity distribution of truncated losses

For the frequency distribution we can perform a simple transformation on the parameter estimates to account for the fact that the loss data is truncated.
For the severity distribution this is typically more difficult. Several articles regarding this subject have been written in the context of operational risk modeling, see Ergashev et al. [16] or Cavallo et al. [8]. There are various methods to account for the loss truncation in estimating the parameters of the severity distribution. A method that is widely used in practice is the *shifting approach*, mainly due to its simplicity. However, it is well-known that the estimated parameters are highly biased and can result in severe over- or underestimation of the capital charge.

Shifting approach Suppose we observe operational loss data $X_1, X_2, \ldots X_n$, truncated from below at the (known) truncation level $H > 0$. Under the shifting approach, the truncated loss data is shifted over the amount of the truncation threshold, resulting in the loss data sample $X_1 - H, X_2 - H, \ldots X_n - H$. The shifted sample is then treated as a non-truncated data sample[15] and the parameters are estimated accordingly, using for example (ordinary) maximum likelihood estimation.
To illustrate the capital bias induced by the shifting approach, we perform the following simulation experiment:

1. We draw loss data samples of size $n = 500$ from the lognormal distribution $\mathcal{LN}(\mu, \sigma)$, where the parameters range over the intervals $\mu \in [9.5, 12]$ and $\sigma \in [1, 2.5]$. Each sample is truncated at its 0.35 lower quantile.

[15] Although, there exist probability distributions for which this property holds, e.g. the exponential distribution (see Section 4.7.2). In general, the shifted sample is not the same as a non-truncated data sample.

2. We apply the shifting approach: we shift the loss data sample over the truncation threshold and compute the parameter estimates $(\widehat{\mu}, \widehat{\sigma})$ using ordinary maximum likelihood estimation.

3. We calculate the capital charge using the FFT method for the compound loss distribution $\mathrm{Pois}(\lambda = 25) \vee \mathcal{LN}(\widehat{\mu}, \widehat{\sigma})$. This is divided by the capital charge of the original non-truncated loss data sample, which is estimated by maximum likelihood.

This experiment is repeated 100 times, in order to get average results. See Fig.3.2 for the surface plot of the capital bias.

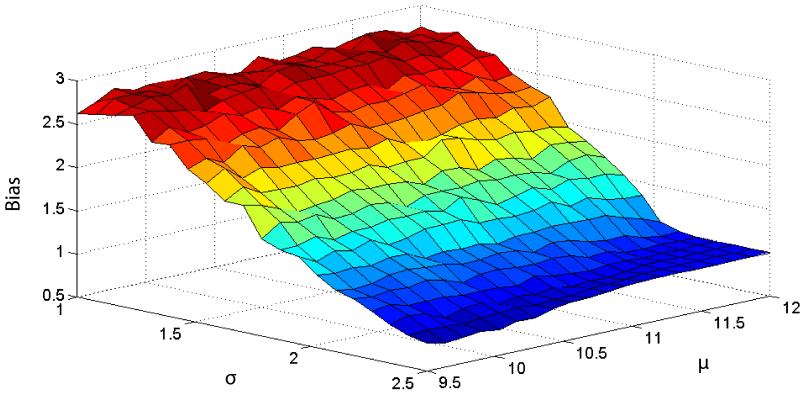


Fig. 3.1.: Capital bias under the shifting approach with threshold H at the lower 0.35 quantile of the loss data.

It is seen that for larger values of σ the capital charge is underestimated and as σ decreases the capital charge is (highly) overestimated. It is clear that this approach should not be used in practice, since it can give very inaccurate results in estimating the VaR measures. The method that we will use to account for the loss data truncation is the *constrained maximum likelihood approach* (CML).

Constrained maximum likelihood approach In this approach, we obtain the parameter estimates by maximizing the likelihood function of the conditional distribution. We perform a maximum likelihood estimation routine, but instead of using the unconditional probability density functions f, we use the conditional probability density function f_H, given that the loss data is truncated at H. In other words, let $X_1, X_2, \ldots, X_n$ be IID operational losses with common density f truncated at the recording threshold H, the conditional density and cumulative distribution of the truncated losses can be written as:

$$f_H(x|\theta) = \frac{f(x|\theta)}{1 - F(H|\theta)} \quad \text{for} \quad H < x \leq \infty \tag{3.39}$$

$$F_H(x|\theta) = \frac{F(x|\theta) - F(H|\theta)}{1 - F(H|\theta)} \quad \text{for} \quad H < x \leq \infty \tag{3.40}$$

Where F denotes the cumulative distribution of the non-truncated sample. Instead of maximizing the ordinary likelihood function, we maximize the likelihood function with respect to the truncated density according to

$$\arg\max_{\theta} L_H(\theta|\boldsymbol{X}) = \arg\max_{\theta} \prod_{i=1}^{n} f_H(X_i|\theta) \tag{3.41}$$

In practice, this method is not frequently applied, although it seems a very natural method to use. The reason for this is that analytic expressions of the likelihood equations can become quite complex and since, in general, there are no closed form expressions for the estimators, numerical algorithms have to be implemented in order to maximize the likelihood function for each different severity distribution. In Section 5.1 we introduce a numerical algorithm to compute the robust version of maximum likelihood estimators (so-called *optimal robust estimators*), for both non-truncated and truncated operational loss data. Under certain specifications of the parameters, the optimal robust estimators reduce exactly to the constrained maximum likelihood (CML) estimators. Thus, a numerical algorithm for the constrained maximum likelihood approach for different choices of the severity distribution is readily available to us.

We perform the same simulation experiment as for the shifting approach, only now we apply the constrained maximum likelihood function approach to compute parameter estimates $(\widehat{\mu}, \widehat{\sigma})$ for the truncated loss data samples. The routine is repeated 25 times, since the CML estimators take longer to compute than the MLE under the shifting approach. A surface plot of the capital bias can be found in Fig.3.2 below.

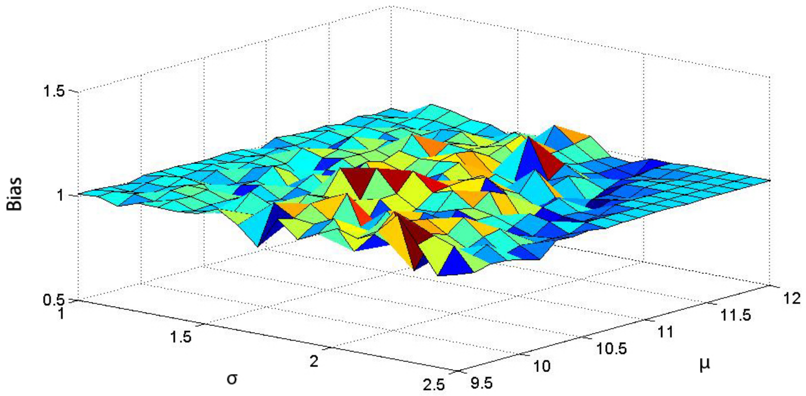


Fig. 3.2.: Capital bias under the constrained maximum likelihood approach with threshold H at the lower 0.35 quantile of the loss data.

Expectation-Maximization (EM) algorithm An other efficient method to estimate parameters of truncated loss data is the EM algorithm. The EM algorithm is aimed at estimating the unknown parameters by maximizing the expected likelihood function given the observed and missing data. In Appendix A.2 we give a detailed description of the EM algorithm and perform the same simulation experiment as for the shifting approach and CML approach. From the results in Appendix A.2, we conclude that the accuracy of the CML approach and EM algorithm in estimating the capital charges for truncated loss data is practically similar. As noted earlier, we choose to model the severity distribution of truncated losses by the CML approach, mainly because this approach relates directly to the robust estimators that we will derive in the following chapters.

3.7 Summary of the model so far

Combining everything we have introduced up till now, we can produce a sound *classical* model to estimate the operational risk capital charge.

Suppose we observe a data sample $x_1, x_2, \ldots, x_n$ of operational losses, corresponding to data collected over a number of m years from a single unit of measure. The goal is to estimate the VaR measures of the aggregated one-year loss distribution (compound loss distribution). Below we describe the estimation procedure informally, this is presented in a systematic way in Fig.3.7.

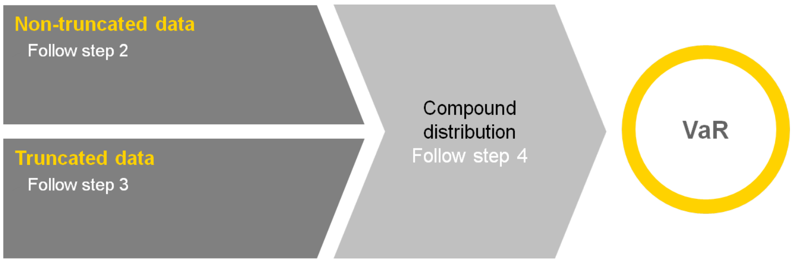


Fig. 3.3.: Overview steps in the *classical* model

1. We check whether the data is truncated or not, if the data is truncated (from below) we assume that the truncation level H is known. If the data is non-truncated go to step 2, else go to step 3.

2. The data is non-truncated.

   - We can fit the frequency distribution by for instance a Poisson distribution using maximum likelihood estimation on the m annual loss frequencies.

   - We can fit the severity distribution by for instance a lognormal, log-gamma or generalized Pareto distribution using maximum likelihood estimation on the n operational loss severities.

   Now go to step 4.

3. The data is truncated at truncation level H.

   - We can fit the frequency distribution by for instance a Poisson distribution according to Section 3.6.1 for the m truncated annual loss frequencies.

   - We can fit the severity distribution by for instance a lognormal, log-gamma or generalized Pareto distribution using the CML approach according to Section 3.6.2 for the n truncated operational loss severities.

   Now go to step 4.

4. Given the estimated severity and frequency distribution, by the LDA we construct the aggregate one-year loss distribution (compound loss distribution) and corresponding VaR measures using the FFT method, according to Section 3.3.

Simulation experiment To assess the accuracy of the VaR measures under the *classical* model, we perform an experiment with simulated loss data samples. The simulation experiment can be summarized as follows: we simulate data samples of size $n = 100$ (which is not an extraordinary low amount of data within a single unit of measure for a large internationally active bank) with severity distribution according to:

- A lognormal distribution $\mathcal{LN}(\mu = 10.95, \sigma = 1.75)$, with recording thresholds $H = 0$ (non-truncated loss data), $H = 10,000$, $H = 25,000$ and $H = 50,000$.

- A log-gamma distribution $\mathcal{LG}(a = 34.5, b = 3.5)$, with recording thresholds $H = 0$ (non-truncated loss data), $H = 10,000$, $H = 25,000$ and $H = 50,000$.

- A Generalized Pareto distribution $\mathrm{GPD}(\xi = 0.65, \beta = 57500)$, with recording thresholds $H = 0$ (non-truncated loss data), $H = 10,000$, $H = 25,000$ and $H = 50,000$.

For each loss data sample we estimate the parameters of the corresponding severity distribution, using MLE, when the severity loss data is non-truncated; and the CML approach, when the severity loss data is truncated. In the next step, we compute the VaR of the compound loss distribution using the FFT method with Poisson frequency distribution $\mathrm{Pois}(\lambda = 25)$. For each entry in Table 3.3 the results are the average VaR measures of 250 simulated loss data samples. The bias is computed by dividing the estimated VaR by VaR under the true parameters.

		Recording threshold H			
	true VaR	$H = 0$	$H = 10,000$	$H = 25,000$	$H = 50,000$
$\mathcal{LN}(\mu, \sigma)$	63,945,425	65,516,330	65,998,790	66,045,350	65,173,430
	Bias	1.02	1.03	1.03	1.02
$\mathcal{L}G(a, b)$	62,290,900	64,859,110	64,689,970	63,476,660	61,823,700
	Bias	1.04	1.04	1.02	0.99
$\mathrm{GPD}(\xi, \beta)$	67,916,625	73,816,280	73,220,820	73,447,940	70,592,800
	Bias	1.08	1.08	1.08	1.04

Tab. 3.3.: VaR estimates according to the classical model for both non-truncated and non-truncated loss data.

We conclude that, under a correctly specified severity distribution and IID loss data, using the methods introduced in this chapter we are able to estimate the VaR measures quite accurately. In the remainder of the thesis, by the *classical* model, we will refer to the model as described in this section.

Remark 3.0.2. In Table 3.3 the estimated VaR measures seem to be slightly biased upwards. Especially, for the Generalized Pareto distribution the estimated VaR measures differ by up to 8% from the *true* VaR. In Opdyke & Cavallo [26] it is shown that this upward bias is induced by Jensen's inequality and this bias increases as the tail becomes heavier. This also explains why the capital bias is most present for the Generalized Pareto distribution, which is the most heavy-tailed distribution.

4 Robust statistics framework and the Influence Function

4.1 Why robust procedures?

In general statistical procedures are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying generating mechanisms. Even in the most simple cases we make explicit or implicit assumptions about randomness, independence, distributional models (e.g. observations are identically and independently distributed). In practice we can never expect these beliefs to be *exactly* true, since they are mathematically convenient rationalizations of an often more complex reality.
In the previous chapter, we have shown that under a correctly specified model and IID operational loss observations, we are able to estimate the VaR measures quite accurately, even when the loss data is truncated. Such assumptions are usually made in the statistics literature in order to derive further theoretical results, which become highly complex without them. But, if we expect these assumptions never to be exactly true in practice, shouldn't we also investigate the behavior of the estimated VaR measures under minor deviations from the made assumptions?

Simulation experiment We consider the same simulation experiment as in Section 3.7, except that we manipulate the loss data in a way that it deviates slightly from the (model) assumptions. The simulation experiment can be summarized as follows: we simulate non-truncated data samples of size $n = 100$ from the lognormal distribution $\mathcal{LN}(10.95, 1.75)$, log-gamma distribution $\mathcal{LG}(34.5, 3.5)$ and Generalized Pareto distribution $\mathrm{GPD}(0.65, 57500)$. Now, we consider the following three cases of data manipulation:

1. For each loss data sample, we remove the largest in-sample loss observation.

2. For each loss data sample, we add a single loss equal to twice the largest in-sample loss observation.

3. For each loss data sample, we add a single absolute loss of $\$10$ to the sample.

After each manipulation procedure, we estimate the severity parameters of the new data sample via maximum likelihood estimation and compute the resulting VaR measures. For each entry in Table 4.1 the results are the average values of 250 estimated VaR measures. The bias is computed by dividing the estimated VaR by VaR under the true parameters.

		Data manipulation procedure		
	true VaR	1.	2.	3.
$\mathcal{LN}(10.95, 1.75)$	63,945,425	56,471,943	96,109,134	132,971,124
	Bias	0.88	1.50	2.08
$\mathcal{L}G(34.5, 3.5)$	62,290,900	53,700,174	115,543,428	462,752,895
	Bias	0.86	1.85	7.43
$\mathrm{GPD}(0.65, 57500)$	67,916,625	59,183,487	305,411,469	142,357,413
	Bias	0.87	4.50	2.10

Tab. 4.1.: VaR estimates according to the classical model for non-truncated loss data.

It is clear from the results in Table 4.1 that, although the *classical* model performs quite well under the usual textbook assumptions, when we introduce minor deviations at the assumed model the estimated VaR measures become almost useless. The VaR can differ up to a factor 4 for the Generalized Pareto distribution when a single extreme (large) loss is added to the loss data sample. Moreover, adding a single small loss event seems to cause an even greater catastrophe. For the log-gamma distribution we observe an increase of $\$400$ million in the capital charge, due to a new recorded loss of $\$10$.

The above example clearly has to do with long-tailnedness: lengthening the tails of the underlying severity distribution, (the right tail by adding large losses, but also the left tail by adding very small losses) explodes the variance of the severity distribution. Shortening the tails seems to have a similar effect, by decreasing the variance of the severity.

Some critics of the Basel II framework argue that the standards required for the calculation of regulatory capital are such that the amount of the capital charge may become arbitrarily large, leaving decreased availability of funds required for financial needs and investments. Furthermore, just a handful of new losses can result in large swings of the capital charge estimates.

In 2001, the Basel committee made the following recommendation:

> "Data will need to be collected and robust estimation techniques (for event impact, frequency and aggregate operational loss) will need to be developed."(BIS, 2001, Annex 6, pg. 26)

So what is meant with the notion of robustness? Since the middle of the 20th century, mathematicians have become increasingly aware that some of the most common statistical procedures are excessively sensitive to seemingly minor deviations from the model assumptions and many *robust* procedures have been proposed to remedy for this fact. The word *robust* is loaded with many (sometimes inconsistent) connotations. In this thesis we will use it in a relatively narrow sense: for our purposes, robustness signifies insensitivity to small deviations from the model assumptions. Primarily, we are concerned with so-called *distributional robustness*. That is, the shape of the true underlying distribution of the loss data deviates slightly from the assumed model. This is both the most important case and the best understood one. We will describe our definition of robustness in more detail in the following section.

4.2 The robust statistics framework

We are adopting what Huber [21] calls an *applied parametric viewpoint*. This can be summarized as follows: we have a parametric model, which hopefully is a good approximation to the true underlying generating mechanisms of the data, but we cannot and do not assume that it is *exactly* correct. It is a well-known fact that robustness is based on compromise, as stated by Anscombe (1960) with his insurance metaphor:

> "Sacrifice some efficiency at the exact model, in order to insure against accidents caused by deviations from the model."

Under the applied parametric viewpoint, a statistical procedure should possess the following desirable features, which we also employ as an informal definition of a *robust* procedure:

- *Efficiency*: It should have reasonably good efficiency at the assumed model, i.e. if the model assumptions are exactly true.

- *Stability*: It should be robust in the sense that small deviations from the model assumptions only have small impact on the performance of the model.

- *Breakdown*: Any larger deviations from the model assumptions should not cause a catastrophe.

In this context, a small deviation from the model is regarded as the occurrence of a small fraction of *outliers*[1] from the gross part of the data, where the gross part of the data does follow the assumed model exactly.

The above definition of (distributionally) robust procedures seems to be practically synonymous to *outlier resistant* procedures. Any reasonable, formal or informal, procedure for rejecting outliers will prevent the worst. We might therefore ask whether robust procedures are needed at all; perhaps a two-step approach would suffice:

1. Clean the data by applying some formal or informal rule for outlier rejection.

2. Use the classical estimation procedures on the remainder.

According to Huber [21] there are several reasons why these steps would not do the same job in a simpler way:

- It is rarely possible to separate the two steps cleanly. Due to the heavy-tailedness of the operational loss data, outliers can be difficult to recognize unless we have reliable, robust estimates for the parameters.

- Many classical rejection rules are unable to cope with multiple outliers. It can happen that a second outlier masks the first, so that none is rejected. An empirical study on this subject can be found in Hampel et al. [19].

[1] By outliers we will denote loss data points that do not follow the assumed model exactly.

- Outlier rejection procedures belong to the general class of robust estimation procedures. In the same empirical study by Hampel et al. [19] it is shown that the best outlier rejection procedures do not quite reach the performance of the best robust procedures. This is mainly because the latter can make a smooth transition between full acceptance and full rejection of an observation, whereas the outlier rejection procedures generally cannot.

In general, robust procedures do not aim at throwing away extreme (large òr small) observations. They focus on the behavior of the main part of the data that can be easily distorted by outliers. An important application of robust statistics, therefore, is to use it as a diagnostic technique for evaluating the sensitivity of the model to extreme loss events and to reveal their influence on the estimates.

Remark 4.0.3. In Chernobai and Rachev [9] it is noted that we should not blindly use the robust model instead of the classical, it would be better to use them as complements to each other. The results from both approaches are not expected to be the same, as they explain different phenomena of the original data. The robust model describes the behavior of the gross part of the data and an indication of highly influential data points, whereas the classical model gives a more conservative view of the behavior of all available data.

Arguably, the most important tool to measure distributional robustness is the *influence function*. The idea behind the influence function is to measure the stability of the estimators when the underlying distribution deviates slightly from the assumed model.
Before we introduce the influence function, we make several realistic assumptions on the estimation procedures. We recall that so far, in order to estimate the distribution parameters, we have only used ordinary maximum likelihood and the constrained maximum likelihood approach (which is also a version of maximum likelihood). Clearly, the assumptions that are made below are valid for maximum likelihood estimation.

- We assume that the considered estimators can be represented as a statistical functional T of the empirical distribution. In other terms, T depends on the sample $(x_1, x_2, \ldots, x_n)$ only through

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i<x} \tag{4.1}$$

  We write such a statistical functional as

$$T_n(x_1, \ldots, x_n) = T(F_n) \tag{4.2}$$

- We assume that $T_n = T(F_n)$ is (weakly) consistent at F, the true underlying distribution of the observations, in the sense that

$$T(F_n) \xrightarrow{P} T(F) \tag{4.3}$$

  Where $\xrightarrow{P}$ denotes *convergence in probability*[2].

- We assume that $T_n = T(F_n)$ is asymptotically normal, with asymptotic covariance matrix $A(F,T)$, which depends on F and T. This is true in most cases of practical interest. Mathematically, that is

$$\sqrt{n}(T_n - T(F)) \rightsquigarrow \mathcal{N}(0, A(F,T)) \tag{4.5}$$

[2] A sequence $\{X_n\}$ of random variables converges in probability towards X if for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0 \tag{4.4}$$

Where $\rightsquigarrow$ denotes *convergence in distribution*[3].

Furthermore, in order to construct minor deviations of the underlying distribution from the assumed model distribution we introduce the *contamination neighborhood*:

Definition 4.1. (Contamination neighborhood) The contamination neighborhood $\mathcal{P}_\epsilon(F_0) = \mathcal{P}_\epsilon$ of an assumed model distribution F_0 with contamination level $\epsilon geq 0$, is defined as

$$\mathcal{P}_\epsilon = \{F | F = (1-\epsilon)F_0 + \epsilon G, G \in \mathcal{M}\} \tag{4.7}$$

Where $\mathcal{M}$ is a family of (contaminating) parametric distributions.
Informally, under the contamination neighborhood, the fraction $(1-\epsilon)$ (i.e. the gross part) of the loss data follows the assumed model distribution exactly and we have a fraction ϵ of outliers following some contaminating distribution.

The above neighborhood is also called the *gross error model* and is typically considered in the literature[4]. In the remainder of the thesis we use the contamination neighborhood (gross error model) to construct the occurrence of a small fraction of outliers from the gross part of the data, where the gross part of the data does follow the assumed model distribution exactly.
We define the two most important characteristics in measuring *quantitative* robustness, originally defined in Huber [21]:

Definition 4.2. (Maximum bias and maximum variance) Let $\mathcal{P}_\epsilon(F_0)$ be the gross error model with model distribution F_0, then the *maximum bias* of T at contamination level ϵ is given by

$$b_1(\epsilon) = \sup_{F \in \mathcal{P}_\epsilon} |T(F) - T(F_0)| \tag{4.9}$$

and the *maximum variance* is given by

$$v_1(\epsilon) = \sup_{F \in \mathcal{P}_\epsilon} A(F, T) \tag{4.10}$$

Where $A(F, T)$ denotes the asymptotic covariance as in Eq.(4.5).

Using the established robust statistics framework, in the following section we introduce the influence function (IF).

4.3 The Influence Function

The influence function (IF) was originally introduced by Hampel (1968) in order to investigate the infinitesimal behavior of real-valued functionals such as $T(F_n)$. The IF is mainly a *heuristic tool*, with an important intuitive interpretation.

[3] A sequence $\{X_n\}$ of random variables converges in distribution (or weakly) to X if

$$\lim_{n \to \infty} F_n(x) = F(x) \tag{4.6}$$

for every $x \in \mathbb{R}$ at which F is continuous. (Here F_n and F denote the distribution functions of X_n and X respectively.)

[4] Other neighborhoods to construct minor deviations from the assumed model can be found in Huber [21]. We could for instance consider the *Lévy neighborhood*, defined as

$$\mathcal{P}_\epsilon = \{F | \forall t, F_0(t-\epsilon) - \epsilon \leq F(t) \leq F_0(t+\epsilon) + \epsilon\} \tag{4.8}$$

Definition 4.3. (Influence function) The influence function (IF) of the estimator T at F is given by

$$\text{IF}(x|F,T) = \lim_{\epsilon\to 0}\left[\frac{T[(1-\epsilon)F+\epsilon\delta_x]-T(F)}{\epsilon}\right] = \lim_{\epsilon\to 0}\left[\frac{T(F_\epsilon)-T(F)}{\epsilon}\right] \tag{4.11}$$

Where δ_x is the probability measure that puts mass 1 at the point x (Dirac measure).

The basic idea of differentiation of statistical functionals goes back to von Mises (1937) and Filippova (1962). The existence of the IF is an even weaker condition than Gâteaux differentiability[5], as is shown in Huber [21]. This makes the range of its applicability very large, as it can be calculated in all realistic situations without bothering about the regularity conditions. As already noted the importance of the IF lies in its heuristic interpretation: it describes the effect of infinitesimal contamination at the point x on the estimate, standardized by the mass of the contamination. The IF allows us to assess the relative influence of individual observations toward the value of an estimate. If the IF is unbounded, an outlier might cause trouble, since this means that a single loss observation can have an overriding influence on the estimate.
Below we show (informally) that the IF can also be used to calculate the asymptotic covariance matrix $A(F,T)$. The proof is adopted directly from Hampel et al. [19]

Proposition 4.3.1. Consider IID losses $X_1, X_2, \ldots, X_n$ with common severity distribution F. Under the assumptions in Section 4.2, the estimators $T_n = T(F_n)$ are asymptotically normal, according to Eq.(4.5), with asymptotic covariance matrix given by

$$A(F,T) = \int \text{IF}(x|F,T)^2 dF(x) \tag{4.13}$$

Proof. If T is sufficiently regular, it can be linearized near F in terms of the influence function: if G is near F, then the leading terms of the Taylor expansion are:

$$T(G) = T(F) + \int \text{IF}(x|F,T)d(G-F)(x) + \ldots \tag{4.14}$$

We note that

$$\int \text{IF}(x|F,T)dF(x) = 0 \tag{4.15}$$

and, for sufficiently large n, if we substitute the empirical distribution F_n for G in the above expansion, we obtain

$$\sqrt{n}(T(F_n)-T(F)) = \sqrt{n}\int \text{IF}(x|F,T)dF_n(x) + \ldots \tag{4.16}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \text{IF}(x_i|F,T) + \ldots \tag{4.17}$$

By the Central Limit Theorem, the leading term on the right-hand side is asymptotically normal with mean 0. In most cases of practical interest the remainder becomes negligible for $n \to \infty$, so $\sqrt{n}(T(F_n)-T(F))$ is then asymptotically normal with mean 0 and asymptotic variance

$$A(F,T) = \int \text{IF}(x|F,T)^2 dF(x) \tag{4.18}$$

□

[5] Following Huber [21], we say that a statistical functional T is Gâteaux differentiable at the distribution F, if there exists a linear functional $L = L_F$ such that, for all $G \in \mathcal{M}$ (with $\mathcal{M}$ a family of parametric distributions) we have

$$\lim_{t\to 0}\frac{T((1-t)F+tG)-T(F)}{t} = L_F(G-F) \tag{4.12}$$

Thus, the IF allows an immediate and simple, heuristic assessment of the asymptotic properties of an estimate, since it allows us to guess an explicit formula for the asymptotic variance (which then has to be proven rigorously by other means). In Section 5.1.1, we use this result to prove the optimality of the optimal bias robust estimators, in the sense that they minimize the asymptotic variance.

The IF is an asymptotic tool, there are several finite sample and/or difference quotient versions of the IF, the most important being the *empirical influence function (EIF)*, which we will use frequently in the following sections. When using actual data points we can replace F by its empirical distribution F_n, the EIF is then defined as

$$\mathrm{EIF}(x|F_n, T) = \lim_{\epsilon \to 0} \left[\frac{T[(1-\epsilon)F_n + \epsilon\delta_x] - T(F_n)}{\epsilon} \right] = \lim_{\epsilon \to 0} \left[\frac{T(F_{n,\epsilon}) - T(F_n)}{\epsilon} \right] \tag{4.19}$$

This is how the IF is often used in practice, with actual data, and typically $\epsilon = \frac{1}{n}$ to evaluate the effect of contamination of a single data point in x. The EIF often matches the IF almost identically, even for relatively small sample sizes, making the EIF a good practical tool for validating IF derivations or for approximating the IF.

4.3.1 Bias robustness

From the robustness point of view, there are several summary values that we can derive from the IF (see Hampel [19]). The most important being the *gross error sensitivity*, given by

$$\gamma^*(F, T) = \sup_x |\,\mathrm{IF}(x|F, T)| \tag{4.20}$$

The gross error sensitivity measures the worst (approximate) influence that a small amount of contamination of fixed size can have on the value of the estimator. It is desirable that γ^* is finite, i.e. the IF is bounded, in which case we say that T is *bias robust* (or B-robust) at F. Typically, putting a bound on γ^* is the first step in robustifying an estimator, and this will often conflict with the aim of asymptotic efficiency. In Section 5.1 we derive the optimal bias robust estimators, these estimators cannot be improved simultaneously with respect to the asymptotic covariance matrix $A(F, T)$ and the gross error sensitivity γ^*. On the other hand, there usually also exists a positive minimum of γ^* for consistent estimators, corresponding to the most bias robust estimators.

Our goal is to derive the IF for the estimation procedures in the classical model. That is, we wish to derive the IF for maximum likelihood estimation and the constrained maximum likelihood approach, (i.e. the truncated version of maximum likelihood). These are the only cases in the thesis, where we are able to derive the IF analytically, everywhere else we approximate the IF by means of the EIF.

In Huber [21] and Hampel [19] the IF is derived analytically for the general class of M-estimators and MLE in specific. In the following section we introduce the class of M-estimators, originally defined by Huber and derive the corresponding influence functions.

4.4 Influence function for M-estimators

In 1964, Huber [20] proposed generalizing maximum likelihood estimation to the broad class of M-estimators, which are obtained as the minima of sums of functions of the data. Well-known examples of M-estimators are the least-squares estimators and (obviously) the maximum likelihood estimators[6]. The definition of M-estimators was actually motivated by robust statistics, which contributed new types of M-estimators (e.g. the optimal bias robust estimators introduced in the next chapter).

[6] We note that the constrained maximum likelihood estimators, according to Section 3.6.2 are also M-estimators.

Definition 4.4. (M-estimators) Suppose that $X_1, X_2, \ldots, X_n$ are IID random variables with common distribution function F. Let $\widehat{\theta}$ be an estimator defined by a minimization problem of the form

$$\widehat{\theta} = \arg\min_{\theta} \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right) \tag{4.21}$$

With ρ some measurable function. Then $\widehat{\theta}$ is called a *maximum likelihood type*-estimate, or simply M-estimate. Often it is simpler to differentiate with respect to θ and solve for the root of the derivative. If ρ is differentiable in θ then $\sum \rho(x_i, \theta)$ is minimized by the solution of

$$\sum_{i=1}^{n} \phi(x_i, \theta) = 0 \tag{4.22}$$

Where $\phi(x, \theta) = \frac{\partial}{\partial \theta} \rho(x, \theta)$. When differentiation is possible, the M-estimator is said to be of *ϕ-type*. Otherwise, we say that the M-estimator is of *ρ-type*.

If we denote the parameter space by Θ. Then, in the single-parameter case $\Theta \subset \mathbb{R}$, by plugging in:

$$\rho(x, \theta) = -\log f(x, \theta) \tag{4.23}$$

Or

$$\phi(x, \theta) = \frac{\partial}{\partial \theta} \rho(x, \theta) = -\frac{1}{f(x, \theta)} \frac{\partial f(x, \theta)}{\delta \theta} \tag{4.24}$$

We get the ordinary maximum likelihood estimators, where $\phi(x, \theta)$ is the negative of the *score function*. In the multi-parameter case $\Theta \subset \mathbb{R}^k$, we have to find solutions of k separate equations. The multivariate version of Eq.(4.24) can be written as

$$\phi(x, \theta) = \left(\frac{\partial \log(f(x, \theta))}{\partial \theta_1}, \ldots, \frac{\partial \log(f(x, \theta))}{\partial \theta_k} \right)^T \tag{4.25}$$

Derivation of the IF for M-estimators For the class of M-estimators, we can derive analytical expressions for the general form of the IF. Following Huber [21] we give an (informal) derivation of the IF, which can then be used directly in the case of maximum likelihood and the constrained maximum likelihood approach.

Let us consider a data sample with common distribution function $F_\epsilon = (1 - \epsilon)F + \epsilon \delta_x$, where F is the assumed model distribution, ϵ the level of contamination and δ_x the Dirac measure with point mass in x. If we write Eq.(4.22) in functional form, the M-estimator $T_n = T(F_{\epsilon,n})$ is the solution of

$$\int \phi(x, T(F_\epsilon))\, dF_\epsilon(x) = 0 \tag{4.26}$$

We differentiate Eq.(4.26) with respect to ϵ. Under the assumption that $\phi(x, \theta)$ is continuously differentiable in θ, we can interchange integration and differentiation. And we can write

$$\int \phi(y, T(F))\, d(\delta_x - F) + \int \frac{\partial}{\partial \theta} [\phi(y, \theta)]_{T(F)}\, dF(x) \cdot \frac{\partial}{\partial \epsilon} [T((1 - \epsilon)F + \epsilon \delta_x)]_{\epsilon=0} = 0 \tag{4.27}$$

We note that the IF according to Eq.(4.11) can be rewritten as

$$\mathrm{IF}(x|F, T) = \lim_{\epsilon \to 0} \left[\frac{T[(1 - \epsilon)F + \epsilon \delta_x] - T(F)}{\epsilon} \right] = \frac{\partial}{\partial \epsilon} [T((1 - \epsilon)F + \epsilon \delta_x)]_{\epsilon=0} \tag{4.28}$$

Therefore, from Eq.(4.27) we obtain

$$\mathrm{IF}(x|F, T) = \frac{\phi(x, T(F))}{-\int \frac{\partial}{\partial \theta} \phi(x, T(F))\, dF(x)} \tag{4.29}$$

Furthermore, by means of Prop. 4.3.1, we obtain in a heuristic way that $\sqrt{n}(T_n - T(F))$ has asymptotic covariance matrix:

$$A(F,T) = \frac{\int \phi(x, T(F))^2 dF(x)}{\left[\int \frac{\partial}{\partial\theta}\phi(x, T(F))dF(x)\right]^2} \tag{4.30}$$

Influence function for MLE Since maximum likelihood estimation belongs to the class of M-estimators, we plug in the defining maximum likelihood equations. For the single-parameter case $\Theta \subset \mathbb{R}$, we have

$$\phi(x,\theta) = -\frac{1}{f(x,\theta)}\frac{\partial f(x,\theta)}{\partial\theta} \tag{4.31}$$

And taking its derivative yields

$$\frac{\partial}{\partial\theta}\phi(x,\theta) = \frac{-\partial^2 f(x,\theta)/\partial\theta^2 \cdot f(x,\theta) + [\partial f(x,\theta)/\partial\theta]^2}{f(x,\theta)^2} \tag{4.32}$$

According to Eq.(4.29), the maximum likelihood IF can be written as

$$\mathrm{IF}(x|\theta,T) = \frac{\frac{1}{f(x,\theta)}\frac{\partial f(x,\theta)}{\partial\theta}}{\int \frac{1}{f(y,\theta)}\left(\left[\frac{\partial f(y,\theta)}{\partial\theta}\right]^2 - \frac{\partial^2 f(y,\theta)}{\partial\theta^2}f(y,\theta)\right)dy} \tag{4.33}$$

For the multi-parameter case $\Theta \in \mathbb{R}^k$, if we write $\phi_\theta = (\phi_{\theta_1},\ldots,\phi_{\theta_k})^T = \left(\frac{\partial\log(f(x,\theta))}{\partial\theta_1},\ldots,\frac{\partial\log(f(x,\theta))}{\partial\theta_k}\right)^T$ the matrix form of the maximum likelihood IF, according to Eq.(4.29), becomes

$$\begin{aligned}
\mathrm{IF}(x|\theta,T) &= A(\theta)^{-1}\cdot\phi_\theta && (4.34)\\
&= \begin{bmatrix} -\int\frac{\partial\phi_{\theta_1}}{\partial\theta_1}dF(y) & \cdots & -\int\frac{\partial\phi_{\theta_1}}{\partial\theta_k}dF(y) \\ \vdots & \ddots & \\ -\int\frac{\partial\phi_{\theta_k}}{\partial\theta_1}dF(y) & & -\int\frac{\partial\phi_{\theta_k}}{\partial\theta_k}dF(y)\end{bmatrix}^{-1}\cdot\begin{bmatrix}\phi_{\theta_1}\\ \phi_{\theta_2}\\ \vdots\\ \phi_{\theta_k}\end{bmatrix} && (4.35)
\end{aligned}$$

In the above equation $A(\theta)$ corresponds to the negative of the Fisher information matrix[7]. The parameters are correlated when the cross-partial derivative terms are nonzero. In conclusion, it is seen that in the single-parameter case, bias robustness of the maximum likelihood estimators can be determined by evaluating whether the score function is bounded. In the multiple-parameter case, the cross-partial derivative terms also have to be taken into account for each parameter, to determine correlation between the parameters. In the following section, we derive analytic expressions of the maximum likelihood IF analytically for the lognormal, log-gamma and Generalized Pareto severity distribution.

4.5 Maximum likelihood IF for the lognormal, log-gamma and GPD

In this section we follow Opdyke and Cavallo [27], where the maximum likelihood IF is analytically derived for the lognormal, log-gamma and Generalized Pareto distribution.

[7] The Fisher information is the variance of the score, or the expected value of the observed information. If $\log(f(x,\theta))$ is twice differentiable with respect to θ, the Fisher information can be written as

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log(f(X,\theta))|\theta\right] \tag{4.36}$$

IF lognormal distribution First, we consider a lognormal severity distribution. The MLE IF can be computed using the general multi-parameter form in Eq.(4.34). We recall that the probability density function of the lognormal distribution is given by

$$f(x|\mu,\sigma) = \frac{1}{x\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2\right) \tag{4.37}$$

Now we compute

$$\phi_\theta = \begin{bmatrix}\phi_\mu \\ \phi_\sigma\end{bmatrix} = \begin{bmatrix}-\frac{1}{f(x|\mu,\sigma)}\left(\frac{\partial f(x|\mu,\sigma)}{\partial\mu}\right) \\ -\frac{1}{f(x|\mu,\sigma)}\left(\frac{\partial f(x|\mu,\sigma)}{\partial\sigma}\right)\end{bmatrix} = \begin{bmatrix}\frac{\mu-\log(x)}{\sigma^2} \\ \frac{1}{\sigma}-\frac{(\log(x)-\mu)^2}{\sigma^3}\end{bmatrix} \tag{4.38}$$

Furthermore, the coordinates of $A(\theta)$ are given by

$$-\int_0^\infty \frac{\partial\phi_\mu}{\partial\mu}\,dF(y) = -\int_0^\infty \frac{1}{\sigma^2}f(y)dy = -\frac{1}{\sigma^2} \tag{4.39}$$

$$-\int_0^\infty \frac{\partial\phi_\sigma}{\partial\sigma}\,dF(y) = -\int_0^\infty \left(\frac{3(\log(y)-\mu)^2}{\sigma^4}-\frac{1}{\sigma^2}\right)f(y)dy = -\frac{2}{\sigma^2} \tag{4.40}$$

$$-\int_0^\infty \frac{\partial\phi_\sigma}{\partial\mu}\,dF(y) = \int_0^\infty \left(\left[\frac{\log(y)-\mu}{\sigma^2}\right]\left[\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right]\right. \tag{4.41}$$

$$\left.-\left[\frac{\log(y)-\mu}{\sigma^2}\right]\left[\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right]\right)f(y)dy = 0 \tag{4.42}$$

$$-\int_0^\infty \frac{\partial\phi_\mu}{\partial\sigma}\,dF(y) = -\int_0^\infty \frac{\partial\phi_\sigma}{\partial\mu}\,dF(y) = 0 \tag{4.43}$$

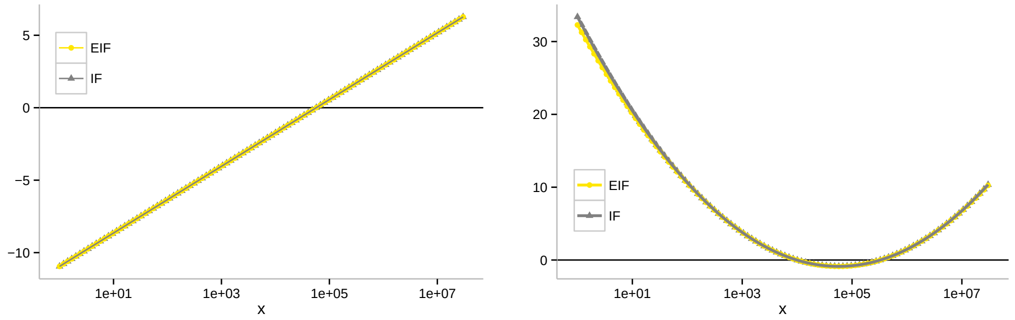
Since the off-diagonal entries are zero, we conclude that the parameters are uncorrelated. According to Eq.(4.34) the IF becomes

$$\mathrm{IF}(x|\theta,F) = A(\theta)^{-1}\cdot\phi_\theta \tag{4.44}$$

$$= \begin{bmatrix}-1/\sigma^2 & 0 \\ 0 & -2/\sigma^2\end{bmatrix}^{-1}\begin{bmatrix}\frac{\mu-\log(x)}{\sigma^2} \\ \frac{1}{\sigma}-\frac{(\log(x)-\mu)^2}{\sigma^3}\end{bmatrix} \tag{4.45}$$

$$= \begin{bmatrix}\log(x)-\mu \\ \frac{(\log(x)-\mu)^2-\sigma^2}{2\sigma}\end{bmatrix} \tag{4.46}$$

It is seen that the maximum likelihood estimators for μ (the sample mean of the $\log$-sample) and σ (the sample standard deviation of the $\log$-sample) are both *not* bias robust, since both coordinates of the IF diverge as $x \to \pm\infty$. With fixed parameters $\mu = 10.95$ and $\sigma = 1.75$, the MLE IFs are plotted in Fig.4.1 below. Furthermore, we compute the EIFs for $K = 1000$ loss data samples from a lognormal distribution of size $n = 250$, according to Eq.(4.19) with $\epsilon = 1/n$. It is seen that the average EIFs correspond to the IFs almost perfectly.



(a)IF μ

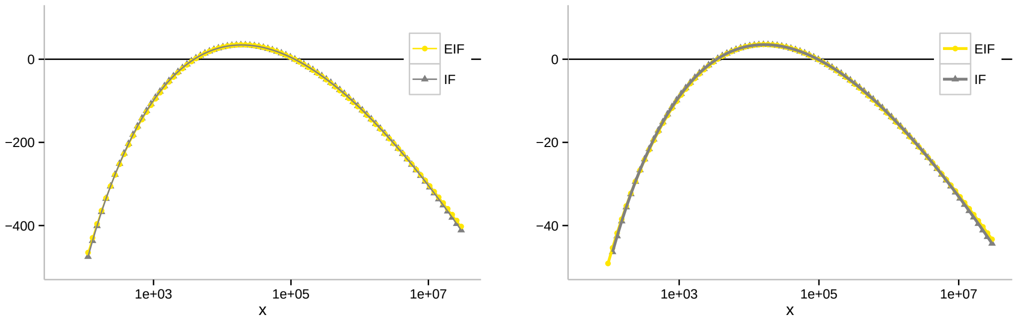
(b)IF σ

Fig. 4.1.: MLE IF and EIF of the lognormal distribution $\mathcal{LN}(10.95, 1.75)$.

IF log-gamma distribution The computations for the log-gamma distribution are analogous and can be found in Appendix B.1. The resulting analytical MLE IF is given by

$$\mathrm{IF}(x|\theta,T) = \begin{bmatrix} \frac{(a/b^2)[\log(b)+\log(\log(x))-\psi_0(a)]-(1/b)[\log(x)-(a/b)]}{\psi_1(a)(a/b^2)-(1/b^2)} \\ \frac{(1/b)[\log(b)+\log(\log(x))-\psi_0(a)]-\psi_1(a)[\log(x)-(a/b)]}{\psi_1(a)(a/b^2)-(1/b^2)} \end{bmatrix} \tag{4.47}$$

Where ψ_0 and ψ_1 denote the digamma and the trigamma function, which are the first- and second-order logarithmic derivatives of the gamma function $\psi_0(z) = \frac{\partial}{\partial z}\log(\Gamma(z))$ and $\psi_1(z) = \frac{\partial^2}{\partial z^2}\log(\Gamma(z))$. We note that both the IF of the shape parameter a and the IF of the rate parameter b diverge as $x \to \pm\infty$. Therefore, we conclude that both maximum likelihood estimators are not bias robust, since their IF is unbounded. With fixed parameters $a = 34.5$ and $b = 3.5$, the IF is plotted in Fig.4.2 below. In the same figures, we have plotted the average EIFs of $K = 1000$ loss data samples with size $n = 250$. Again, we find that the average EIF corresponds to the IF almost perfectly.



(a)IF a

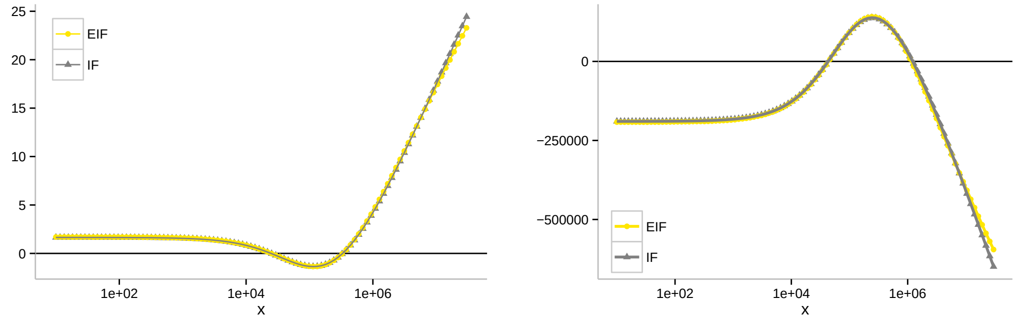
(b)IF b

Fig. 4.2.: MLE IF and EIF of the log-gamma distribution $\mathcal{LG}(34.5, 3.5)$.

IF Generalized Pareto distribution The computations for the Generalized Pareto distribution are analogous and can be found in Appendix B.2. The resulting MLE IF is given by

$$\mathrm{IF}(x|\theta,T) = \begin{bmatrix} (1+\xi)^2\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2 x}\right) + \frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right] + (1+\xi)\left[\frac{\beta-x}{\beta+\xi x}\right] \\ -(1+\xi)\beta\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2 x}\right) + \frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right] - 2(1+\xi)\beta\left[\frac{\beta-x}{\beta+\xi x}\right] \end{bmatrix} \tag{4.48}$$

It is more difficult to see directly that both the IF of the shape parameter ξ and the scale parameter β diverge as $x \to \pm\infty$. But from the plots in Fig. 4.3 it is clear that under the choice of parameters $\xi = 0.65$ and $\beta = 57500$, the resulting IFs are unbounded. Therefore, we again conclude that both maximum likelihood estimators for ξ and β are not bias robust. In the same figures, we have plotted the average EIFs of $K = 1000$ loss data samples with size $n = 250$. The EIFs correspond to the IFs very well, concluding that both the IF and EIF are correctly formulated.



(a)IF ξ

(b)IF β

Fig. 4.3.: MLE IF and EIF of the Generalized Pareto distribution $\text{GPD}(0.65, 57500)$.

As was seen in the previous chapter, the operational loss data is truncated in practice. In the following section, we derive the IF for the constrained maximum likelihood approach. Since, the constrained maximum likelihood (CML) approach is maximum likelihood estimation with conditional density functions, we apply the same methods as in the previous section, but instead of ordinary severity distributions, we consider the conditional severity distributions.

4.5.1 Constrained maximum likelihood IF for the lognormal, log-gamma and GPD

In this section we derive analytic expressions of the IF for the constrained maximum likelihood (CML) estimators. Essentially, we have to derive the MLE IFs for truncated severity distributions. The computations below are done following Opdyke and Cavallo [26], where the maximum likelihood IF is analytically derived for the truncated lognormal, log-gamma and Generalized Pareto distribution.
According to Eq.(3.39) and Eq.(3.40), we recall that if the loss data is truncated at H, the left-truncated probability density function g and cumulative distribution function G can be denoted by

$$g(x,\theta) = \frac{f(x,\theta)}{1-F(H,\theta)} \tag{4.49}$$

$$G(x,\theta) = 1-\frac{1-F(x,\theta)}{1-F(H,\theta)} \tag{4.50}$$

Where f and F denote the non-truncated probability density and cumulative distribution function. At this point we wish to plug in the defining truncated maximum likelihood equations in Eq.(4.29).
For the single-parameter case $\Theta \subset \mathbb{R}$, we have

$$\phi(x|H,\theta) = -\frac{1}{g(x,\theta)}\frac{\partial g(x,\theta)}{\partial\theta} = -\frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)} - \frac{\frac{\partial}{\partial\theta}F(H,\theta)}{1-F(H,\theta)} \tag{4.51}$$

And taking its derivative yields

$$\frac{\partial}{\partial\theta}\phi(x|H,\theta) = \frac{-(\frac{\partial^2}{\partial\theta^2}f(x,\theta))f(x,\theta)+(\frac{\partial}{\partial\theta}f(x,\theta))^2}{f(x,\theta)^2} \tag{4.52}$$

$$+\frac{-(\frac{\partial^2}{\partial\theta^2}F(H,\theta))(1-F(H,\theta))-(\frac{\partial}{\partial\theta}F(H,\theta))^2}{(1-F(H,\theta))^2} \tag{4.53}$$

Plugging these equations into Eq.(4.29), the left-truncated maximum likelihood IF can be written as

$$\mathrm{IF}(x|\theta,T,H) \quad = \quad -\phi(x|H,\theta)\left[\int_H^\infty \frac{\partial}{\partial\theta}\phi(x|H,\theta)\;dG(x)\right]^{-1} \tag{4.54}$$

With $\phi(x|H,\theta)$ and $\frac{\partial}{\partial\theta}\phi(x|H,\theta)$ according to the equations above. For the multi-parameter case $\Theta \subset \mathbb{R}^k$, with $\phi_\theta = (\phi_{\theta_1},\ldots,\phi_{\theta_k})^T = \left(\frac{\partial\log(g(x,\theta))}{\partial\theta_1},\ldots,\frac{\partial\log(g(x,\theta))}{\partial\theta_k}\right)^T$, the matrix form of the truncated maximum likelihood IF becomes

$$\mathrm{IF}(x|\theta,T,H) \quad = \quad A(\theta)^{-1}\cdot\phi_\theta \tag{4.55}$$

$$= \begin{bmatrix} -\int\frac{\partial\phi_{\theta_1}}{\partial\theta_1}dG(y) & \cdots & -\int\frac{\partial\phi_{\theta_1}}{\partial\theta_k}dG(y) \\ \vdots & \ddots & \\ -\int\frac{\partial\phi_{\theta_k}}{\partial\theta_1}dG(y) & & -\int\frac{\partial\phi_{\theta_k}}{\partial\theta_k}dG(y)\end{bmatrix}^{-1}\cdot\begin{bmatrix}\phi_{\theta_1}\\ \phi_{\theta_2}\\ \vdots\\ \phi_{\theta_k}\end{bmatrix} \tag{4.56}$$

CML IF lognormal distribution First, we consider the lognormal severity distribution. The calculations are analogous to the MLE IF, except that we consider the truncated version of the IF. We compute

$$\phi_\theta \quad = \quad \begin{bmatrix}\phi_\mu\\ \phi_\sigma\end{bmatrix} = \begin{bmatrix} -\frac{\partial f(x|\mu,\sigma)/\partial\mu}{f(x|\mu,\sigma)} - \frac{\partial F(H|\mu,\sigma)/\partial\mu}{1-F(H|\mu,\sigma)} \\ -\frac{\partial f(x|\mu,\sigma)/\partial\sigma}{f(x|\mu,\sigma)} - \frac{\partial F(H|\mu,\sigma)/\partial\sigma}{1-F(H|\mu,\sigma)}\end{bmatrix} \tag{4.57}$$

$$= \begin{bmatrix} -\left[\frac{\log(x)-\mu}{\sigma^2}\right] - \frac{\int_0^H\left[\frac{\log(y)-\mu}{\sigma^2}\right]f(y|\mu,\sigma)dy}{1-F(H|\mu,\sigma)} \\ -\left[\frac{(\log(x)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right] - \frac{\int_0^H\left[\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right]f(y|\mu,\sigma)dy}{1-F(H|\mu,\sigma)}\end{bmatrix} \tag{4.58}$$

Furthermore, the coordinates of $A(\theta)$ are given by

$$-\int_H^\infty\frac{\partial\phi_\mu}{\partial\mu}dG(y) \quad = \quad -\frac{1}{\sigma^2}+\frac{1}{(1-F(H|\mu,\sigma))^2}\left[\int_0^H\frac{\log(y)-\mu}{\sigma^2}f(y|\mu,\sigma)dy\right]^2 \tag{4.59}$$

$$+\frac{1}{1-F(H|\mu,\sigma)}\int_0^H\frac{(\log(y)-\mu)^2}{\sigma^4}-\frac{1}{\sigma^2}f(y|\mu,\sigma)dy \tag{4.60}$$

$$-\int_H^\infty\frac{\partial\phi_\sigma}{\partial\sigma}dG(y) \quad = \quad \frac{1}{1-F(H|\mu,\sigma)}\int_H^\infty\frac{3(\log(y)-\mu)^2}{\sigma^4}f(y|\mu,\sigma)dy \tag{4.61}$$

$$+\frac{1}{\sigma^2}+\frac{1}{(1-F(H|\mu,\sigma))^2}\left[\int_0^H\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma^2}f(y|\mu,\sigma)dy\right]^2 \tag{4.62}$$

$$+\frac{1}{1-F(H|\mu,\sigma)}\int_0^H\left(\left[\frac{1}{\sigma^2}-\frac{3(\log(y)-\mu)^2}{\sigma^4}\right]\right. \tag{4.63}$$

$$\left.+\left[\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right]^2\right)f(y)dy \tag{4.64}$$

$$-\int_H^\infty\frac{\partial\phi_\mu}{\partial\sigma}dG(y) \quad = \quad -\int_H^\infty\frac{\partial\phi_\sigma}{\partial\mu}dG(y) \tag{4.65}$$

$$= \quad \frac{1}{1-F(H|\mu,\sigma)}\int_0^H-\left[\frac{\log(y)-\mu}{\sigma^2}\right]\left[\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{3}{\sigma}\right]f(y)dy \tag{4.66}$$

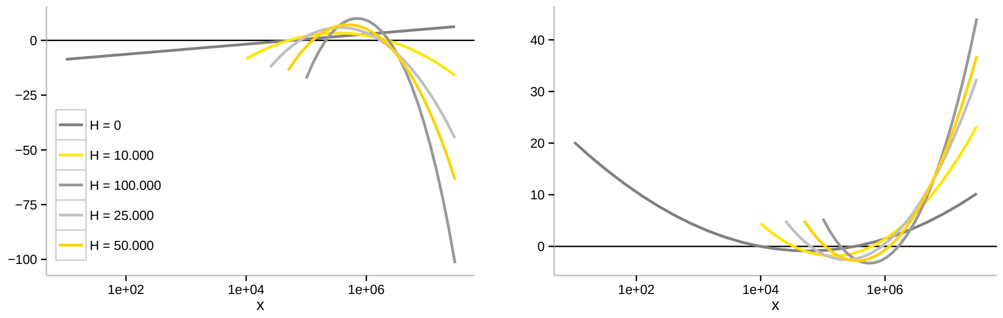
$$-\frac{1}{(1-F(H|\mu,\sigma))^2}\left[\int_0^H\frac{\log(y)-\mu}{\sigma^2}f(y)dy\right] \tag{4.67}$$

$$\cdot\left[\int_0^H\left(\frac{(\log(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right)f(y)dy\right] \tag{4.68}$$

We can find the IF by numerically solving:

$$\mathrm{IF}(x|\theta, T, H) = \begin{bmatrix} -\int_H^\infty \frac{\partial\phi_\mu}{\partial\mu} dG(y) & -\int_H^\infty \frac{\partial\phi_\mu}{\partial\sigma} dG(y) \\ -\int_H^\infty \frac{\partial\phi_\sigma}{\partial\mu} dG(y) & -\int_H^\infty \frac{\partial\phi_\sigma}{\partial\sigma} dG(y) \end{bmatrix}^{-1} \cdot \begin{bmatrix} \phi_\mu \\ \phi_\sigma \end{bmatrix} \tag{4.69}$$

We note that the off-diagonal coordinates of the negative Fisher information matrix have become nonzero, therefore introducing dependence between the parameters, where they were independent in the non-truncated case. The resulting IFs, for different truncation thresholds H, are presented in Fig.4.4.

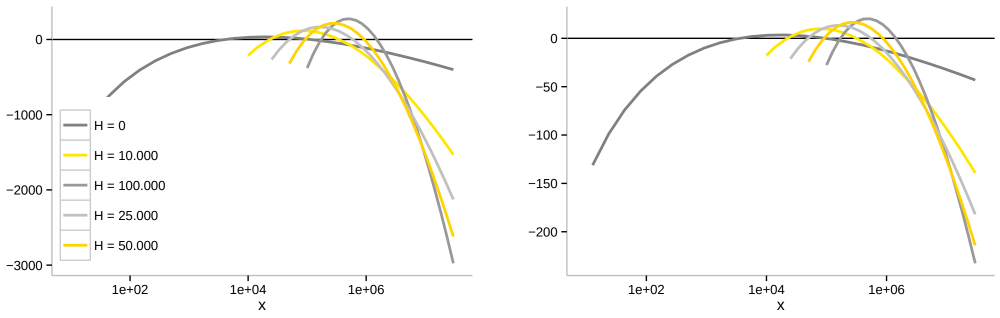


(a)IF μ (b)IF σ

Fig. 4.4.: CML IF of the lognormal distribution $\mathcal{LN}(10.95, 1.75)$.

Inspection of Fig.4.4 shows that both the CML estimators of μ and σ are not bias robust. Furthermore, the entire shape and direction of the IF for μ has changed. Consequently, the direction of the relationship between the two parameters has changed: previously large values of x would increase both μ and σ, but under truncation of the loss data, large values of x move μ and σ in opposite directions, i.e. contamination in the upper right tail now leads to arbitrary small μ and arbitrarily large σ. So, quite counterintuitively, a large loss value will actually decrease the value of the *location* parameter.

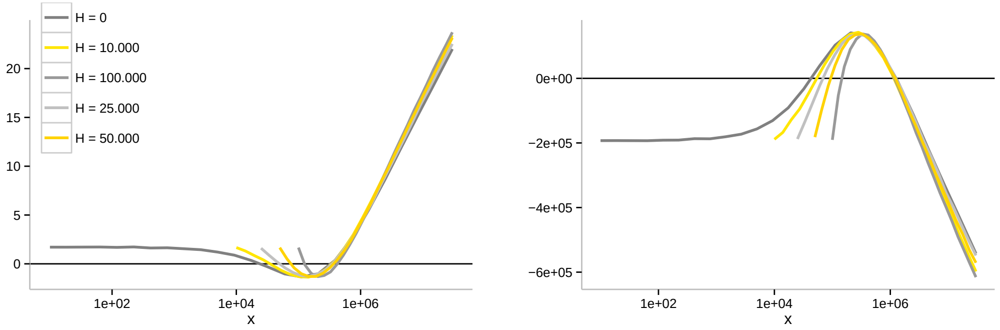
CML IF log-gamma distribution and GPD Next we derive analytic expressions of the CML IF for the log-gamma and Generalized Pareto distribution. Both calculations are tedious and not very informative, for the analytic calculations of the CML IF for the log-gamma and GPD we refer the reader to Appendix B.3 and B.4. Since we have no closed-form expressions for the CML IFs, we plot the resulting IFs here for different truncation thresholds H. See Fig.4.5 for the log-gamma CML IFs and Fig.4.6 for the Generalized Pareto CML IFs. We conclude that neither the CML estimators of the log-gamma distribution nor the CML estimators of the GPD are bias robust. The estimators of the GPD seem to be least influenced by the truncation of the loss data, since the truncated IFs of ξ and β are much more similar to the non-truncated IF than the truncated IFs of the lognormal and log-gamma distribution.



(a)IF a

(b)IF b

Fig. 4.5.: CML IF of the log-gamma distribution $\mathcal{L}G(34.5, 3.5)$.



(a)IF ξ

(b)IF β

Fig. 4.6.: CML IF of the Generalized Pareto distribution $\mathrm{GPD}(0.65, 57500)$.

4.6 Δ-VaR approximation

In the previous section we have derived the IFs for the maximum likelihood estimators of the severity distribution parameters and the IFs for the constrained maximum likelihood estimators. In this way, we have a clear view on the influence of contaminating loss data on the estimated parameters of the severity distribution. Our main interest, however, is the influence of contamination on the estimated VaR measures. To be more precise, if we add a contaminating loss observation to the data sample, we wish to know the impact on the resulting VaR measure. At this point, for loss data from the truncated lognormal distribution, if we add an extremely large or small loss to the data sample, we know that the location parameter will decrease and on the other hand the scale parameter will increase. It seems natural to translate this into the effect on the resulting VaR measure. To this purpose, we introduce the new concept of Δ-VaR: we use the derived expressions for the IFs to approximate the difference in the VaR measures, due to contamination in the point x.

By the definition of the IF in Eq.(4.11), we can heuristically approximate the estimator due to ϵ-contamination in the point x by

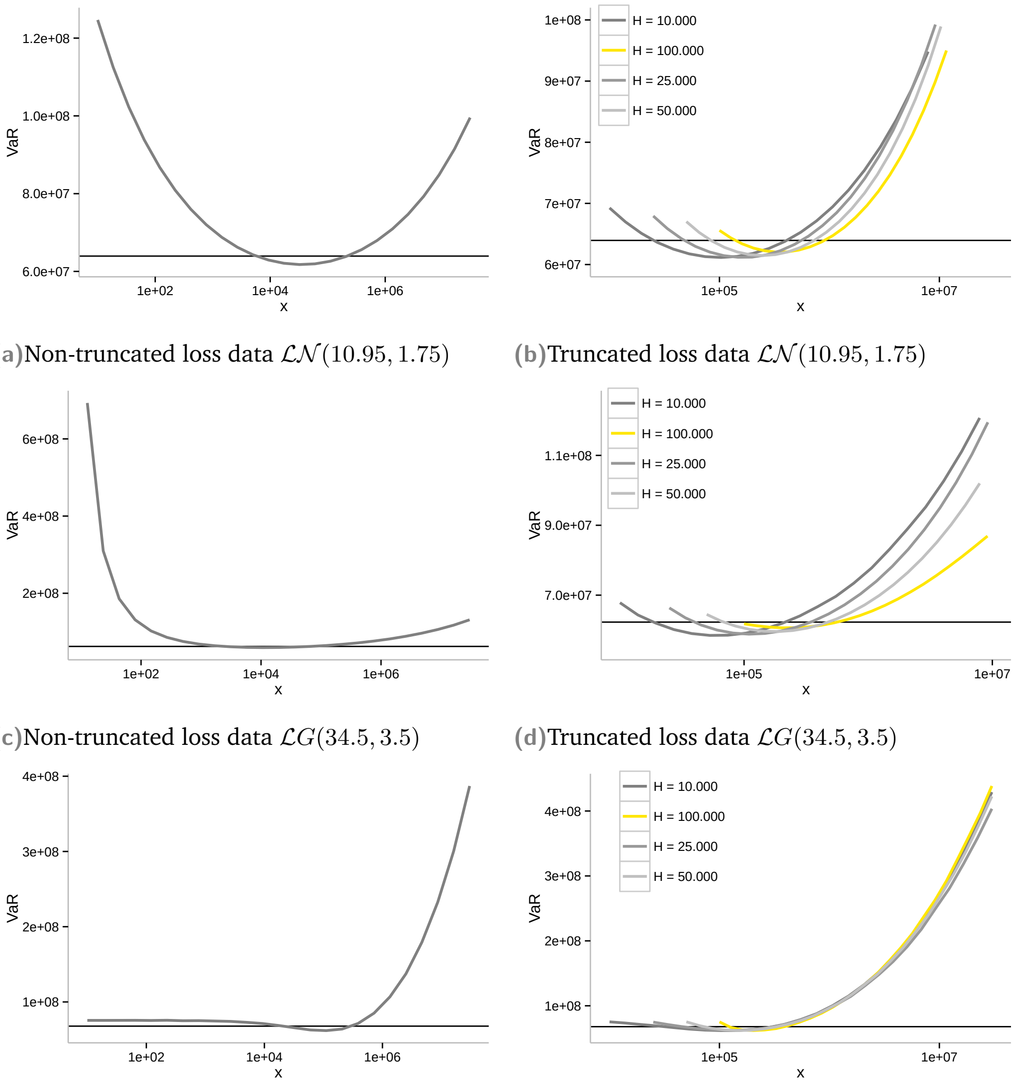
$$T(F_\epsilon) \approx T(F) + \mathrm{IF}(x|F,T) \cdot \epsilon \tag{4.70}$$

Since this is only an approximation, we are not concerned with any further theoretical justifications. If we choose $\epsilon = \frac{1}{n}$, we can write:

$$T\left(\left(1 - \frac{1}{n}\right) F + \frac{1}{n}\delta_x\right) \approx T(F) + \frac{\mathrm{IF}(x|F,T)}{n} \tag{4.71}$$

Informally, we approximate the value of the estimator T when the original data sample of size n is *contaminated* by a single loss in the point x.

Using the derived IFs in the previous section, we can easily approximate the difference in VaR measures under contamination in a single data point x for different sample sizes n. We plot the Δ-VaR approximation, under MLE for non-truncated loss data and the CML approach for truncated loss data. The results correspond to data samples of size $n = 100$, with (the usual) severities according to the lognormal $\mathcal{LN}(10.95, 1.75)$, log-gamma $\mathcal{LG}(34.5, 3.5)$ and Generalized Pareto distribution $\mathrm{GPD}(0.65, 57500)$ and fixed frequency distribution $\mathrm{Pois}(25)$.



(a)Non-truncated loss data $\mathcal{LN}(10.95, 1.75)$ (b)Truncated loss data $\mathcal{LN}(10.95, 1.75)$

(c)Non-truncated loss data $\mathcal{LG}(34.5, 3.5)$ (d)Truncated loss data $\mathcal{LG}(34.5, 3.5)$

(e)Non-truncated loss data $\mathrm{GPD}(0.65, 57500)$ (f)Truncated loss data $\mathrm{GPD}(0.65, 57500)$

Fig. 4.7.: Difference in VaR under contamination in x for maximum likelihood (left) and the constrained maximum likelihood approach (right).

We draw several conclusions from the figures above. First, both the VaR measures of the lognormal and log-gamma distribution are equally sensitive to extreme small losses as to extreme large losses. For the log-gamma distribution the capital charge may increase by a factor 10, when a single small loss is added to the data sample. This agrees with the results in Table 4.1, where it was found that the log-gamma distribution was the most sensitive to contamination due to small losses. As for the GPD severity, the VaR measures are far less sensitive to contamination in the left tail. However, the GPD is the most sensitive to contamination in the right tail, which can be explained by the fact that it is the most heavy-tailed distribution of the three. It is seen that the VaR measures increase by up to a factor 7, when a single (extreme) large loss is added to the data sample. This corresponds to the results in Table 4.1, where it was found that the GPD was the most sensitive to contamination due to extreme large losses.

Under truncation of the loss data, it is seen that for all severity distributions the sensitivity to small losses is heavily reduced. In this sense, we argue that a higher truncation threshold will lead to more stable capital charges overall. However, for all three severity distributions the capital charge is still influenced by contamination in the left tail. A new recorded loss just above the truncation threshold can still increase the capital charge by several millions.

4.7 Mixed severity distributions

The only article that studies the impact of small losses on the capital charge is Opdyke and Cavallo [27]. Opdyke and Cavallo, however, do not propose any solution to this problem, they only state that this issue might (for a major part) be responsible for the instability of the operational risk capital charge.

As was seen in the Δ-VaR figures, the truncation of the data heavily mitigates the impact of small losses on the VaR measures. However, the VaR may still increase with several millions, only due to a new loss observation just above the truncation threshold.

If we look at the distributions we use to fit the loss data, this is not an unexpected outcome. For heavy-tailed distributions, such as the lognormal distribution, the probability of an extremely small loss is similar to the probability of an extremely large loss. Therefore, adding such a small loss to our data sample, will $-$for the lognormal distribution$-$ greatly increase the scale parameter. Consequently, the capital charge increases by a similar amount as if an extremely large loss was added to the data sample.

In our opinion, the natural solution to this problem is to fit a mixture of severity distributions. In other terms, we wish to find a severity distribution with the properties of a light-tailed distribution at the left tail (i.e. the *body region*) and the properties of a heavy-tailed distribution at the right tail (i.e. the *tail region*). Fitting severity distributions to the body and the tail region separately is a method that is frequently used in practice, but is not considered in Opdyke and Cavallo [27].

Monte Carlo simulation for severity mixture As stated in the previous chapter, a commonly used method to estimate the operational risk capital charge in practice is by means of Monte Carlo simulation (as described in Appendix A.1). This method can easily be extended to model a mixture of severity distributions. Such a procedure can be described as follows:

1. Fix a body-tail threshold L according to some pre-specified measure, for example the 90%-quantile of the loss data.

2. Fit a light-tailed severity distribution to $[H, L)$, i.e. the *body region* of the data, (where the data sample might already be truncated at level H). Fit a heavy-tailed severity distribution to $[L, \infty)$, i.e. the *tail region* of the data.

3. We simulate a large amount K of loss data points, where $\frac{L-H}{K} \times 100\%$ of the data points is sampled from the fitted body distribution and $\frac{K-(L-H)}{K} \times 100\%$ of the data points is sampled from the fitted tail distribution. We compute the VaR of the simulated data using the Monte Carlo method as described in Appendix A.1.

Although the above process is easy to understand and implement, we would like the avoid the use of Monte Carlo simulation. In order to apply the FFT method, which gives accurate and computationally fast results for the VaR measures (which the Monte Carlo methods generally do not), we have to normalize the combined mixture of the fitted body and tail distribution. Using the resulting expression for the overall severity distribution, we can then apply the FFT method as usual.

4.7.1 Normalizing the severity mixture

Non-truncated severity loss data First, we consider the situation where the original severity loss sample is non-truncated, i.e. truncation threshold $H = 0$. The body-tail threshold is denote by L. Assume we wish to fit a light-tailed distribution to the body region $[0, L)$ according to distribution function F, with parameters $\theta \in \Theta$. Furthermore, we wish to fit a heavy-tailed distribution to the tail region $[L, \infty)$ according to distribution function G, with parameters $\xi \in \Xi$. Here Θ and Ξ denote the parameter space of θ and ξ respectively. Let the weight of the body be given by the average of the two distribution functions at the body-tail threshold L,

$$W_{\text{body}}(L|\theta, \xi) = \frac{F(L, \theta) + G(L, \xi)}{2} \tag{4.72}$$

Consequently, let the weight of the tail be given by the average of the two tail distribution functions at the threshold level L,

$$W_{\text{tail}}(L|\theta, \xi) = \frac{(1 - F(L, \theta)) + (1 - G(L, \xi))}{2} = 1 - W_{\text{body}} \tag{4.73}$$

It is clear that, if the body and tail distribution are equivalent, i.e. $F(L, \theta) = G(L, \xi)$, we have $W_{\text{body}} = F(H, \theta)$ and $W_{\text{tail}} = 1 - F(H, \theta)$. Since the body distribution is truncated to the right and the tail distribution is truncated to the left, their respective conditional probability density functions $f_L(x, \theta)$, $g_L(x, \xi)$ and cumulative distribution functions $F_L(x, \theta)$, $G_L(x, \xi)$ can be written as:

$$f_L(x, \theta) = \frac{f(x, \theta)}{F(L, \theta)} \quad \text{for} \quad 0 \le x < L \tag{4.74}$$

$$F_L(x, \theta) = \frac{F(x, \theta)}{F(L, \theta)} \quad \text{for} \quad 0 \le x < L \tag{4.75}$$

$$g_L(x, \xi) = \frac{g(x, \xi)}{1 - G(L, \xi)} \quad \text{for} \quad L \le x < \infty \tag{4.76}$$

$$G_L(x, \xi) = \frac{G(x, \xi) - G(L, \xi)}{1 - G(L, \xi)} \quad \text{for} \quad L \le x < \infty \tag{4.77}$$

We note that the body distribution now has its domain on $[0, L)$ and the tail distribution has its domain on $[L, \infty)$. The overall severity distribution function is then given by:

$$\mathcal{F}_L(x|\theta, \xi) = \begin{cases} W_{\text{body}} \cdot F_L(x, \theta) & \text{for} \quad 0 \le x < L \\ W_{\text{tail}} \cdot G_L(x, \xi) & \text{for} \quad L \le x < \infty \end{cases} \tag{4.78}$$

We check the normalization of $\mathcal{F}_L$:

$$\int_0^\infty d\mathcal{F}_L(x|\theta, \xi) = W_{\text{body}} \int_0^L dF_L(x, \theta) + W_{\text{tail}} \int_L^\infty dG_L(x, \xi) = W_{\text{body}} + W_{\text{tail}} = 1 \tag{4.79}$$

The advantage of this approach is that the weights of the body and tail distributions are not fixed by the threshold L, they also depend on the parameter values. Furthermore, we have now created a proper distribution function (with mass equal to 1), for which we can calculate VaR measures using the FFT method as usual.

Truncated severity loss data We also consider the scenario where the original severity loss sample is left-truncated at truncation threshold $H > 0$. We fit a light-tailed distribution to the body region $[H, L)$ according to the conditional distribution function

$$F_H(x|\theta) = \frac{F(x,\theta) - F(H,\theta)}{1 - F(H,\theta)} \tag{4.80}$$

Which yields the new weights

$$\begin{aligned} W_{\text{body}}(L|\theta,\xi,H) &= \frac{\frac{F(L,\theta)-F(H,\theta)}{1-F(H,\theta)} + G(L,\xi)}{2} && (4.81) \\ W_{\text{tail}}(L|\theta,\xi,H) &= \frac{\frac{1-F(L,\theta)}{1-F(H,\theta)} + (1 - G(L,\xi))}{2} && (4.82) \end{aligned}$$

The conditional tail density $g_L(x,\xi)$ and distribution function $G_L(x,\xi)$ remain the same. The truncated body density $f_{H,L}(x,\theta)$ and distribution function $F_{H,L}(x,\theta)$ become

$$\begin{aligned} f_{H,L}(x,\theta) &= \frac{f(x,\theta)}{F(L,\theta) - F(H,\theta)} \quad \text{for} \quad H \leq x \leq L && (4.83) \\ F_{H,L}(x,\theta) &= \frac{F(x,\theta) - F(H,\theta)}{F(L,\theta) - F(H,\theta)} \quad \text{for} \quad H \leq x \leq L && (4.84) \end{aligned}$$

The body distribution now has its domain on $[H, L)$ and the tail distribution has its domain on $[L, \infty)$. The overall severity distribution function is then given by

$$\mathcal{F}_{H,L}(x|\theta,\xi) = \begin{cases} W_{\text{body}} \cdot F_{H,L}(x,\theta) & \text{for} \quad H < x \leq L \\ W_{\text{tail}} \cdot G_L(x,\xi) & \text{for} \quad L < x < \infty \end{cases} \tag{4.85}$$

And again $\mathcal{F}_{H,L}$ integrates to 1, thus allowing us to use the FFT method to compute the VaR measures.

Outline FFT method for severity mixture Suppose the severity loss data is left-truncated at truncation level H. The process to estimate the capital charge, under mixed severity distributions, via the FFT method can be summarized as follows:

1. Fix the body-tail threshold at some pre-specified measure L, for example again the 90%-quantile of the loss data.

2. Fit a light-tailed severity distribution to $[H, L)$, i.e. the *body region* of the data. Fit a heavy-tailed severity distribution to $[L, \infty)$, i.e. the *tail region* of the data.

3. Given the estimated parameters, compute the body and tail weights and truncated distribution functions and combine them to the overall distribution function as described in this section.

4. Use the FFT method to compute VaR measures of the compound loss distribution.

4.7.2 Application: FFT method with mixed severity distributions

The goal is to find whether fitting a mixture of severity distributions is of use in practice. Hence, we wish to know whether the impact of (very) small losses on the estimated VaR measures is reduced when we fit a mixture of severity distributions. We compute the Δ-VaR figures under the new approach and compare these to the Δ-VaR figures in Section 4.6, which correspond to a single fitted severity distribution. We only consider the single-parameter (light-tailed) exponential distribution[8] for the body region. We approximate the IF of the truncated exponential distribution by its empirical counterpart, the EIF.

Exponential distribution for the body First, we consider fitting an exponential distribution to the body region $[0, L)$. The conditional probability density f_L and distribution function F_L are given by

$$f_L(x|\theta) = \frac{\exp\left(-x/\theta\right)}{\theta\left(1 - \exp\left(-L/\theta\right)\right)} \tag{4.88}$$

$$F_L(x|\theta) = \frac{1 - \exp(-x/\theta)}{1 - \exp(-L/\theta)} \tag{4.89}$$

The log-likelihood function of the right-truncated exponential distribution can be written as

$$\ell(\theta) = -\theta^{-1}\sum_{i=1}^{n} X_i - n\log(\theta) - n\log\left(1 - \exp\left(-\frac{L}{\theta}\right)\right) \tag{4.90}$$

Which is maximized for θ solving the equation

$$\overline{X}_n = \theta - \frac{L\exp(-L/\theta)}{1 - \exp(-L/\theta)} \tag{4.91}$$

We use a Newton-Raphson procedure (see Dixit [12]) to find solutions in θ of the above equation. After choosing the initial value θ_0, we put

$$\theta_{i+1} = \theta_i - \frac{g(\theta_i)}{g'(\theta_i)} \tag{4.92}$$

Where

$$g(\theta_i) = \theta_i - \frac{L\exp(-L/\theta_i)}{1 - \exp(-L/\theta_i)} - \overline{X}_n \tag{4.93}$$

And $g'(\theta_i)$ is the first derivative of $g(\theta_i)$ with respect to θ_i:

$$g'(\theta_i) = 1 - \frac{L^2\exp(-2L/\theta_i)}{\theta_i^2(1 - \exp(-L/\theta_i))^2} \tag{4.94}$$

In this way we have a quick numerical procedure of finding the maximum likelihood estimator $\widehat{\theta}$ for a right-truncated exponential distribution. As the initial parameter value, we choose the ordinary maximum likelihood estimator, such that $\theta_0 = \overline{X}_n$.

To approximate the IF of θ, we compute the EIF for $K = 1000$ loss data samples from a right-truncated exponential distribution of size $n = 250$ and take the average results. The EIFs for different body-tail thresholds L can be found in Fig. 4.8.

[8] We choose to parametrize the exponential distribution via the scale parameter θ, which is the reciprocal of the rate parameter λ. In this way, the probability density function f and cumulative distribution function F are given by

$$f(x,\theta) = \frac{1}{\theta}\exp(-x/\theta) \quad \text{for} \quad x \geq 0 \tag{4.86}$$

$$F(x,\theta) = 1 - \exp(-x/\theta) \quad \text{for} \quad x \geq 0 \tag{4.87}$$

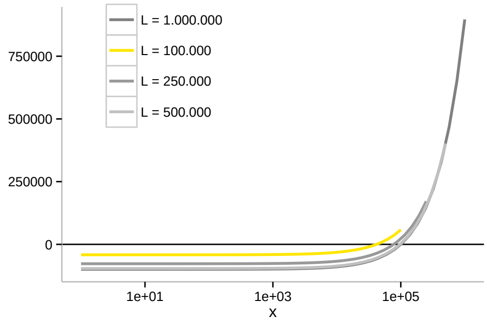


Fig. 4.8.: MLE EIF of the right-truncated exponential, with $\theta = 10^5$.

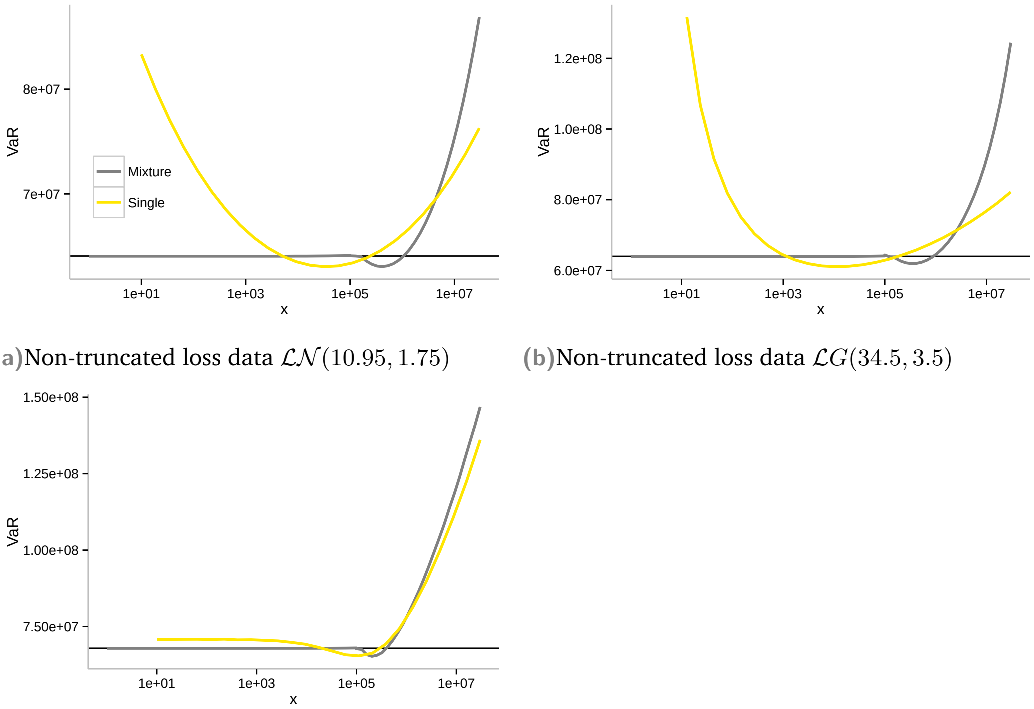
It is seen from Fig. 4.8 that the exponential distribution is far less influenced by small loss data points, as expected from a light-tailed distribution. Now, we also consider the scenario when the loss data is left-truncated at truncation level H. In this case, the shifting approach as described in Section 3.6.2 holds true: the left-truncated exponential distribution is equivalent to the shifted exponential distribution. Mathematically, we can write

$$\begin{aligned} f_H(x|\theta) &= \frac{f(x|\theta)}{1-F(H|\theta)} && (4.95)\\ &= \frac{\exp(-x/\theta)}{\theta(1-(1-\exp(-H/\theta)))} && (4.96)\\ &= \frac{1}{\theta}\exp\left(-\frac{x-H}{\theta}\right) = f(x-H|\theta) && (4.97) \end{aligned}$$

Therefore, if we wish to fit an exponential distribution to the body region $[H, L)$, we first shift the distribution to the left over the threshold H and then fit a right-truncated exponential distribution to $[0, L-H)$ using the Newton-Raphson procedure described earlier.

Δ-VaR approximation We compute the Δ-VaR approximations by means of Eq.(4.71) for a mixture of severity distributions, with an exponential distribution on the body region and respectively a lognormal, log-gamma and Generalized Pareto distribution on the tail region. The body-tail threshold is fixed at $L = 10^5$.

The results below correspond to the Δ-VaR approximations under mixed severity distributions for loss data samples of size $n = 250$ from a lognormal $\mathcal{LN}(10.95, 1.75)$, log-gamma $\mathcal{L}G(34.5, 3.5)$ and Generalized Pareto distribution $\mathrm{GPD}(0.65, 57500)$. These are compared to the Δ-VaR figures found in Section 4.6, which correspond to a single fitted severity distribution.



(a)Non-truncated loss data $\mathcal{LN}(10.95, 1.75)$

(b)Non-truncated loss data $\mathcal{LG}(34.5, 3.5)$

(c)Non-truncated loss data $\mathrm{GPD}(0.65, 57500)$

Fig. 4.9.: Difference in VaR under contamination in x for single and mixed severity distributions.

We draw several conclusions from the figures above. First of all, the sensitivity of the capital charge to losses below the body-tail threshold has been reduced to a minimum. However, there is also a major drawback in fitting a mixture of severities to the loss data. In the computations above, we have chosen loss data samples of size $n = 250$. In practice, banks often have to do the job with far less historical loss data[9] (for single units of measure). In order to fit mixed severity distributions according to the described method, we have to split up the -already scarce- loss data even further into a body and tail region. Since the distribution parameters are estimated using (far) less data points, as in the case of a single fitted severity distribution, they are likely to become more unstable. This is seen in Fig.4.9a and Fig.4.9b, in the sense that the capital charge becomes more sensitive to loss data points in the tail region, which is due to the fact that the parameters of the tail distribution need to be estimated using only the loss observations above the body-tail threshold. We argue that fitting a mixture of severity distributions reduces the impact of small losses on the estimated capital charge to completely. However, this reduction of the influence to losses in the left tail comes at the expense of a higher sensitivity to losses in the right tail.

If we recall the results from Section 4.6, we know that the impact of small loss observations on the capital charge is also mitigated by truncation of the loss data. Since, in practice the operational loss data will always be subject to lower truncation thresholds, it might be sufficient to apply the following more straightforward estimation procedure:

1. Fit a single (heavy-tailed) severity distribution to the truncated loss data, and instead of using the classical estimation procedures (i.e. maximum likelihood and the constrained maximum likelihood approach), we use *good* robust estimation procedures.

[9] Under the Basel II framework, large internationally active banks usually split the operational loss data into 8 business times 7 event types, resulting in 56 units of measures. Consequently, it is not unusual for single units of measure to have less than 50 operational loss observations.

2. Estimate VaR according to the FFT method as usual.

In this way, we do not have to split up the loss data into a separate body and tail region, (thus *not* increasing the sensitivity to loss observations in the right tail), while still reducing the impact of small losses on the capital charge to a minimum. In the next chapter, we study two robust estimation procedures: the optimal bias robust estimators (OBRE) and the method of trimmed moments. We compute the influence functions and Δ-VaR figures for these new estimation methods and compare their performance (at the exact model) to the classical estimation methods that we have been using so far.

5 Robust estimation methods

5.1 Optimal bias robust estimators (OBRE)

In the previous chapter we established non-robustness of the maximum likelihood estimators and constrained maximum likelihood estimators by computing their influence functions. Actually, it is not surprising that the maximum likelihood estimators turned out to be non-robust. In Section 4.2 we have asserted that in general to achieve robustness we have to sacrifice efficiency. It is well-known that the maximum likelihood estimators are the most efficient estimators, so according to the argument in Section 4.2, we also expect them to be the least robust estimators. On the other hand, the *most* robust estimators are expected to be the least efficient.

According to the robust statistics framework in Section 4.2, we wish to achieve *reasonably good* efficiency at the assumed model. Since optimal efficiency is impossible (this is equivalent to MLE), we wish to find robust estimators with *nearly* optimal efficiency.

In this regard, we consider the *optimal bias robust estimators* (OBRE). The aim of the OBRE is to find robust estimators such that the efficiency loss is minimal. They are designed to achieve highest possible efficiency, while remaining bias robust. Historically, according to Hampel et al. [19] there are two methods to optimal robust estimation. The first method is the *minimax approach*[1] by Huber [21]. According to the minimax approach, optimal robust estimators are found by minimizing the maximum bias (according to Def. 4.2) due to contamination.

The second approach is called the *infinitesimal approach* by Hampel (see Hampel et al. [19]). In the infinitesimal approach, we try to find M-estimators with bounded influence function, that have minimum asymptotic variance. In this thesis we choose to follow the infinitesimal approach, since it relies on the influence function and we have already established much of the needed results.

We consider the problem of finding the most efficient estimator under the condition that the gross-error sensitivity, according to Eq.(4.20), must not exceed the given bound c. Thus bounding the IF and making the estimator bias robust. In the multiparameter case, we have to clarify the term efficiency. In Hampel et al. [19] it is shown that under the usual partial ordering of covariance matrices, according to

$$A \leq B \Leftrightarrow B - A \quad \text{positive definite} \tag{5.1}$$

the estimator that minimizes the asymptotic variance under bounded IF, may not exist. If we consider instead the trace[2] of the asymptotic covariance matrix, it is shown that we can always find the minimizing solution. Following Hampel et al. [19], in this context we will therefore measure efficiency by the trace of the asymptotic covariance matrix. There are several optimal estimators depending on the way one chooses to bound the IF. We consider the *standardized* OBRE, which is proven to be numerically more stable than other variants in Alaiz and Victoria-Feser [1].

Definition standardized OBRE Suppose that $X_1, \ldots, X_n$ are IID random variables according to the severity distribution F, with common density f. We consider the parameter space $\Theta \subseteq \mathbb{R}^k$. Thus,

[1] We refer to Huber [21] for a detailed description of the minimax approach.

[2] The trace of a $k \times k$ square matrix A is defined to be the sum of the elements on the main diagonal (the diagonal from the upper left to the lower right) of A.

we wish to estimate k distribution parameters $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)$.
Following Dupuis and Field [13], the standardized OBRE is defined as the solution $\widehat{\theta}$ of

$$\sum_{i=1}^{n} \phi(x_i, \theta) = \sum_{i=1}^{n} [s(x_i, \theta) - a(\theta)] W_c(x_i, \theta) = 0 \tag{5.2}$$

with $s(x, \theta) = \frac{1}{f(x,\theta)} \frac{\partial f(x,\theta)}{\partial \theta}$ the score function and $W_c(x, \theta)$ the weight function defined by

$$W_c(x, \theta) = \min \left\{ 1, \frac{c}{\|A(\theta)[s(x, \theta) - a(\theta)]\|} \right\} \tag{5.3}$$

The $k \times k$ matrix $A(\theta)$ and k-dimensional vector $a(\theta)$ are defined implicitly by solving the equations

$$\mathbb{E}\left[\phi(x_i, \theta)\phi(x_i, \theta)^T\right] = A(\theta)^T A(\theta) \tag{5.4}$$

$$\mathbb{E}\left[\phi(x_i, \theta)\right] = 0 \tag{5.5}$$

It is seen from Eq.(5.2) that the OBRE is an M-estimator of ϕ-type as defined in Eq.(4.22). The ϕ-function tries to *reduce the influence* of loss data points outside the hypersphere with radius c, modifying the data points as little as possible. The function that transforms each point outside a hypersphere to its nearest point on it and leave those inside alone is the multidimensional Huber function, given by

$$h_c(z) = z \min \left\{ 1, \frac{c}{\|z\|} \right\}, \qquad z \in \mathbb{R}^k \tag{5.6}$$

In Fig.5.1 below we give a sketch of the Huber function $h_c(z)$, this figure can be found in Chapter 4 of Hampel et al. [19].

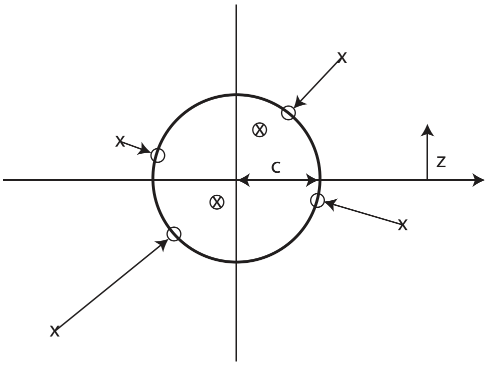


Fig. 5.1.: Sketch of the multidimensional Huber function $h_c(z)$.

The OBRE is then defined as the M-estimator with bounded ϕ-function as close as possible to the ordinary score function according to:

$$\phi(x, \theta) = A(\theta)^{-1} \cdot h_c\Big(A(\theta)[s(x, \theta) - a(\theta)]\Big) \tag{5.7}$$

The matrix $A(\theta)$ and the vector $a(\theta)$ can be viewed as Lagrange multipliers for the constraints resulting from a bounded IF and Fisher consistency, via Eq.(5.4) and Eq.(5.5). For further details we refer the reader to Hampel et al. [19].
The constant c is the bound on the IF and can be interpreted as the tuner between robustness and efficiency. For a lower c we gain robustness, but lose efficiency and for a higher c we gain efficiency, but lose robustness. If we choose $c = \infty$, we remove the bound on the IF and the OBRE reduce to the ordinary maximum likelihood estimators. Typically, researchers choose c to achieve between the 90 to 95 percent efficiency at the model, but this actual value for c depends on the model being implemented

and the data sample under analysis. In Section 5.1.4 we will compute the level of efficiency at the assumed model that is reached for different choices of the tuning parameter c.

5.1.1 Example: one-dimensional OBRE

In the single-parameter case $\Theta \subseteq \mathbb{R}$, we show that the OBRE minimizes the asymptotic variance. Since the OBRE belong to the class of M-estimators, according to Section 4.4 the IF and asymptotic variance of the OBRE can be written as:

$$\mathrm{IF}(x|F,T) = \frac{\phi(x,\theta_*)}{-\int \frac{\partial}{\partial\theta}[\phi(y,\theta)]_{\theta_*} dF(y)} \tag{5.8}$$

$$A(F,T) = \frac{\int \phi(x,\theta_*)^2 dF(x)}{\left(\int \frac{\partial}{\partial\theta}[\phi(y,\theta)]_{\theta_*} dF(y)\right)^2} \tag{5.9}$$

Where $\theta_* = T(F)$ belongs to the parameter space Θ and $\phi(x,\theta)$ is according to the definition of the OBRE in the previous section. In the single parameter case the weight function reduces to the one-dimensional Huber function. We use the notation $[h(x)]_{-b}^{b}$ for an arbitrary function $h(x)$ truncated at the levels b and $-b$:

$$[h(x)]_{-b}^{b} = \begin{cases} b & \text{if} \quad h(x) > b \\ h(x) & \text{if} \quad -b \leq h(x) \leq b \\ -b & \text{if} \quad -b > h(x) \end{cases} \tag{5.10}$$

Under certain regularity conditions, we can prove the optimality of the OBRE in the single parameter case. The proof is adopted directly from Hampel et al. [19].

Proposition 5.1.1. Let $b > 0$ fixed and put $F_* = F_{\theta_*}$ and $f_* = f_{\theta_*}$. Then there exists a real number a such that

$$\phi(x,\theta_*) = [s(x,\theta_*) - a]_{-b}^{b} \tag{5.11}$$

satisfies $\int \phi(x,\theta_*)dF_*(x) = 0$ and $d := \int \frac{\partial}{\partial\theta}[\phi(y,\theta)]_{\theta_*} dF_*(y) > 0$. Now $\phi(x,\theta_*)$ minimizes Eq.(5.9) among all mappings $\phi(x,\theta)$ such that

$$\int \phi(x,\theta)dF_*(x) = 0 \tag{5.12}$$

and

$$\gamma^*(F,T) = \sup_x \left| \frac{\phi(x,\theta)}{\int \frac{\partial}{\partial\theta}[\phi(y,\theta)]_{\theta_*} dF_*(y)} \right| < c := \frac{b}{d} \tag{5.13}$$

where $\gamma^*(F,T) = \sup_x |\,\mathrm{IF}(x|F,T)|$ denotes the gross-error sensitivity as defined in Eq.(4.20).

Proof. First, we show that such a value of a exists. By the dominated convergence theorem, $\int [s(x,\theta_*) - a]_{-b}^{b} dF_*(x)$ is a continuous function in a, and as $a \to \pm\infty$ this integral tends to $\pm b$. Hence there exists an a such that this integral becomes zero. Since we are interested in the optimality result we choose to skip the part which shows that $d > 0$, the proof is straightforward and we refer the interested reader to Hampel et al. [19].

Let us now prove the optimality and the uniqueness of $\phi(x,\theta_*)$. Consider any $\phi(x,\theta)$ which is measurable and satisfies Eq.(5.12) and Eq.(5.13). Without loss of generality we may assume that $\int \frac{\partial}{\partial\theta}[\phi(y,\theta)]_{\theta_*}\, dF_*(y) = d$, so we only have to minimize the numerator of Eq.(5.9). Using Eq.(5.12) we can write

$$\int [(s(x,\theta_*) - a) - \phi(x,\theta))]^2 dF_*(x) = \int (s(x,\theta_*) - a)^2 dF_*(x) - 2d + \int \phi(x,\theta)^2 dF_*(x) \tag{5.14}$$

It suffices to minimize the left member of the equation, which can be written as

$$\int_{\{\tilde{s}>b\}} (\tilde{s}(x) - \phi(x,\theta))^2 dF_*(x) + \int_{\{|\tilde{s}|\leq b\}} (\tilde{s}(x) - \phi(x,\theta))^2 dF_*(x) + \int_{\{\tilde{s}<-b\}} (\tilde{s}(x) - \phi(x,\theta))^2 dF_*(x) \tag{5.15}$$

where $\tilde{s}(x) = s(x,\theta_*) - a$. Because $|\phi| \leq b$ by Eq.(5.13), i.e. the bound on the IF, the above equation is minimized if and only if $\phi(x,\theta) = \phi(x,\theta)$ almost everywhere with respect to F_*. □

5.1.2 OBRE computation

To compute the OBRE, we follow an algorithm proposed by Alaiz and Victora-Feser [1] based on a Newton-Raphson procedure to find successively better numerical approximations of the estimators. In summary, we compute the matrix $A = A(\theta)$ and the vector $a = a(\theta)$ for given θ by solving Eq.(5.4) and Eq.(5.5). Secondly, we apply a Newton-Raphson step to solve Eq.(5.2) given the found matrix A and vector a. These steps are iterated until the parameter estimates for θ converge.
For an observed loss data sample $x_1, \dots, x_n$ and assumed severity distribution F, the OBRE can be found by the following procedure:

1. Fix precision threshold ϵ, and initial parameter values $\theta^{(0)}$ for the vector θ. Initially we set $a = 0$ and $A = J^{\frac{1}{2}}(\theta)^{-T}$ where
$$J(\theta) = \int s(x,\theta)s(x,\theta)^T \, dF(x) \tag{5.16}$$
is the Fisher information matrix.

2. Solve the following equations with respect to a and A:
$$\begin{aligned} A^T A &= M_2^{-1} &(5.17)\\ a &= \frac{\int s(x,\theta)W_c(x,\theta)\, dF(x)}{\int W_c(x,\theta)\, dF(x)} &(5.18)\end{aligned}$$
Where
$$M_k = \int [s(x,\theta) - a][s(x,\theta) - a]^T W_c(x,\theta)^k \, dF(x) \quad \text{for} \quad k = 1,2 \tag{5.19}$$

3. Compute M_1 and $\Delta\theta = M_1^{-1}\left(\frac{1}{n}\sum_{i=1}^n [s(x_i,\theta) - a]W_c(x_i,\theta)\right)$

4. If $\max_j |\frac{\Delta\theta_j}{\theta_j}| > \epsilon$, then we choose the new initial values $\theta \to \theta + \Delta\theta$ and return to Step 2, else stop.

According to Alaiz and Victoria-Feser [1] the algorithm is convergent provided that the initial parameter values $\theta^{(0)}$ are near the solution of the OBRE. In the first step, we can take the MLE as initial parameter values. An extended procedure could be: use the MLE as the starting point, next compute the OBRE with a high value for c (corresponding to a low level of robustness), then use these estimates as the starting point for an estimation procedure with a lower value of c (corresponding to a higher level of robustness). The choice of the initial values for the matrices A and a are the values that solve the equations for maximum likelihood estimation, where the ϕ-function reduces to the ordinary score function[3]. The integrals are all computed numerically in *R*. In step (2), the matrix A is taken as the square root of the Cholesky[4] decomposition of M_2^{-1}, therefore M_2^{-1} must be symmetric and

[3] For ordinary MLE, the expectation of the score function is zero, therefore initially we put $a = 0$. Furthermore, the asymptotic covariance matrix $A^T A$ is equal to the inverse Fisher information matrix.

[4] The Cholesky decomposition or Cholesky triangle is a decomposition of a symmetric, positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose $A = LL^T$. Every real-valued, symmetric, positive-definite matrix has a unique Cholesky decomposition.

positive-definite for A to exist. No major problems were encountered with the positive-definiteness of M_2^{-1} (in very few cases we are unable to complete the routine, due to the fact that M_2^{-1} is not positive-definite, our solution in these cases is to take the OBRE values that have been computed up to the point where the error is produced).

Remark 5.0.4. We note that the produced OBRE weights themselves contain valuable information. They indicate to which degree a particular data point deviates from the assumed severity distribution F. Therefore, they can be used for outlier detection or goodness-of-fit testing. If the OBRE gives low weights to many observations in the sample, this indicates a bad fit and we may consider choosing a different severity model distribution.

5.1.3 OBRE asymptotics

The OBRE are defined as M-estimators of ϕ-type, it can be shown that they are asymptotically normal provided that the score function ϕ is bounded and continuous. The exact derivations can be found in Dupuis and Field [13]. If we write $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ for the OBRE, we have

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_k \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}\right) \rightsquigarrow \mathcal{N}_k\left(0, V(\phi, F)\right) \tag{5.20}$$

Where the asymptotic covariance matrix $V(\phi, F_\theta)$ is given by

$$V(\phi, F) = M(\phi, F)^{-1} Q(\phi, F) M(\phi, F)^{-T} \tag{5.21}$$

With

$$M(\phi, F) = -\int \frac{\partial}{\partial\theta}\phi(x,\theta) dF(x) \tag{5.22}$$

And

$$Q(\phi, F) = \int \phi(x,\theta)\phi(x,\theta)^T dF(x) \tag{5.23}$$

For $c = \infty$, we have $\phi(x, \theta) = s(x, \theta)$ (i.e. the ordinary score function), in this case $V(\phi, F)$ reduces to the inverse of the Fisher information matrix, which is the asymptotic covariance matrix of the MLE. In general, the exact asymptotic covariance matrix is difficult to compute, due to the implicit definition of many terms. Therefore, in the calculations below we approximate the OBRE covariance matrix by Monte Carlo simulation, explained in more detail in Section 5.1.4 below.

5.1.4 OBRE for lognormal severity distribution

According to the estimation procedure in Section 5.1.2, for a particular severity distribution F we need to specify: the initial parameter values $\theta^{(0)}$, the score function $s(x, \theta)$ and inverse Fisher information matrix $A^T(\theta)A(\theta)$. In the OBRE algorithm, we choose the initial parameters $\theta^{(0)}$ equal to the MLE for non-truncated loss data; and equal to the CML estimators for truncated loss data. In Section 4.5 we have analytically derived the score functions and Fisher information matrices for all relevant severity distributions (i.e. lognormal, log-gamma and GPD). We can immediately use the derived results as the input for the OBRE algorithm.

Non-truncated severity loss data Let us first consider non-truncated loss data from the lognormal severity distribution. We put F the cumulative distribution function of the lognormal distribution with parameters $\theta = (\mu, \sigma)$. We specify the following initial values[5] and functions:

$$\begin{aligned}
\theta^{(0)} &= (\theta_\mu, \theta_\sigma) = (\widehat{\mu}_{\text{MLE}}, \widehat{\sigma}_{\text{MLE}}) && (5.24)\\
s(x,\theta) &= \begin{bmatrix} \frac{\log(x)-\theta_\mu}{\theta_\sigma^2} \\ -\frac{1}{\theta_\sigma} + \frac{(\log(x)-\theta_\mu)^2}{\theta_\sigma^3} \end{bmatrix} && (5.25)\\
A^T(\theta)A(\theta) &= \begin{bmatrix} 1/\theta_\sigma & 0 \\ 0 & 2/\theta_\sigma^2 \end{bmatrix}^{-1} && (5.26)
\end{aligned}$$

Where the score function is taken as the negative of Eq.(4.38) and the Fisher information matrix is specified by the negative of Eq.(4.39) to Eq.(4.43). No major convergence issues were encountered when running the algorithm. To illustrate this: if we choose initial values $\theta^{(0)}$ unequal to the MLE, for tuning $c = \infty$ the algorithm converges to the MLE in no more than two or three steps. The implementation of the described numerical algorithm in *R* for non-truncted loss data from the lognormal distribution can be found in Appendix C.1.

Truncated severity loss data For truncated loss data from a lognormal severity distribution, we only need to specify different initial values and functions. We put F and f the left-truncated lognormal density and cumulative distribution according to Eq.(3.39) and Eq.(3.40), with truncation threshold H and parameters $\theta = (\mu, \sigma)$. We set the initial parameter values $\theta^{(0)} = (\widehat{\mu}_{\text{CML}}, \widehat{\sigma}_{\text{CML}})$, which are the values of the constrained maximum likelihood approach according to Section 3.6.2. Furthermore, we specify the truncated score function and Fisher information as:

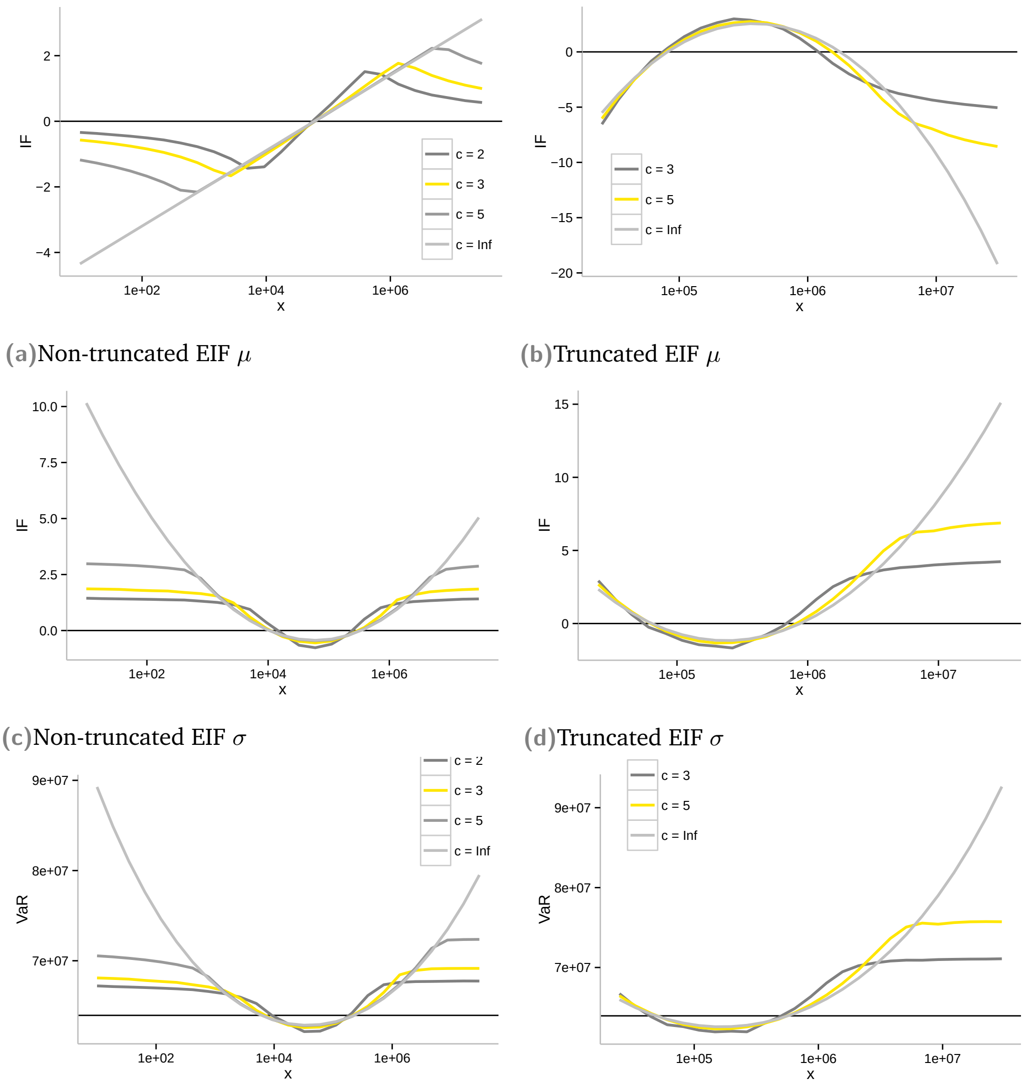
$$\begin{aligned}
s(x,\theta) &= \begin{bmatrix} \left[\frac{\log(x)-\theta_\mu}{\theta_\sigma^2}\right] + \frac{\int_0^H \left[\frac{\log(y)-\theta_\mu}{\theta_\sigma^2}\right] f(y|\theta_\mu,\theta_\sigma)dy}{1-F(H|\theta_\mu,\theta_\sigma)} \\ \left[\frac{(\log(x)-\theta_\mu)^2}{\theta_\sigma^3} - \frac{1}{\theta_\sigma}\right] + \frac{\int_0^H \left[\frac{(y-\theta_\mu)^2}{\theta_\sigma^3} - \frac{1}{\theta_\sigma}\right] f(y|\theta_\mu,\theta_\sigma)dy}{1-F(H|\theta_\mu,\theta_\sigma)} \end{bmatrix} && (5.27)\\
A^T(\theta)A(\theta) &= \begin{bmatrix} \int_H^\infty \frac{\partial\phi_{\theta_\mu}}{\partial\theta_\mu} dF(y|\theta_\mu,\theta_\sigma) & \int_H^\infty \frac{\partial\phi_{\theta_\mu}}{\partial\theta_\sigma} dF(y|\theta_\mu,\theta_\sigma) \\ \int_H^\infty \frac{\partial\phi_{\theta_\sigma}}{\partial\theta_\mu} dF(y|\theta_\mu,\theta_\sigma) & \int_H^\infty \frac{\partial\phi_{\theta_\sigma}}{\partial\theta_\sigma} dF(y|\theta_\mu,\theta_\sigma) \end{bmatrix}^{-1} && (5.28)
\end{aligned}$$

Where the truncated score function is the negative of Eq.(4.58) and the Fisher information matrix is specified by the negative of Eq.(4.59) to Eq.(4.68).

OBRE IFs and Δ-VaR for non-truncated and truncated loss data Due to the many implicit definitions in the OBRE estimation procedure, we are unable to derive analytic expressions for the OBRE IFs. Therefore, we compute the empirical version, the EIF according to Eq.(4.19). To approximate the OBRE IFs for non-truncated loss data, we compute the average EIFs for μ and σ, with $K = 100$ different non-truncated loss data samples of size $n = 250$ from a lognormal distribution. For different values of the tuning parameter c the approximated IFs of μ and σ can be found in Fig.5.2a and Fig.5.2c. Since our main interest is in the (in)stability of the VaR measures, we use the average EIFs to approximate the Δ-VaR according to Eq.(4.71). In Fig.5.2e we plot the difference in VaR due to contamination in the point x for a loss data sample of size $n = 100$, with frequency distribution $\text{Pois}(25)$.

To approximate the truncated OBRE IFs, we proceed in the same manner. We compute the average EIFs for $K = 100$ truncated loss data samples of size $n = 250$ from a lognormal distribution. For fixed truncation threshold $H = 25,000$ and different values of the tuning parameter c the approximated IFs of μ and σ are plotted in Fig.5.2b and Fig.5.2d. In Fig.5.2f we plot the corresponding Δ-VaR for loss data samples of size $n = 100$, with fixed frequency distribution $\text{Pois}(25)$.

[5] The precision threshold ϵ is chosen in *R* in such a way that it is machine specific . In our case, the precision threshold is fixed at $\epsilon = 0.000122$.



(a)Non-truncated EIF μ

(b)Truncated EIF μ

(c)Non-truncated EIF σ

(d)Truncated EIF σ

(e)Non-truncated Δ-VaR

(f)Truncated Δ-VaR

Fig. 5.2.: OBRE approximated IFs and Δ-VaR for lognormal severity distribution $\mathcal{LN}(10.95, 1.75)$.

From the EIF figures above, it is seen that the OBRE becomes more bias robust as the value of the tuning parameter c decreases. Furthermore, for $c = \infty$ we observe that the OBRE IFs correspond perfectly with the MLE IFs and CML IFs in Fig.4.1 and Fig.4.4 respectively.
The difference in the VaR measures under contamination is almost perfectly correlated with the EIF of the scale parameter σ. As the value of c decreases, the capital charge gains stability. For non-truncated loss data, we argue that we should fit a mixture of severity distributions as described in Section 4.7, in order to completely remove the impact of contamination in the left tail (small losses) on the estimated capital charge. For truncated loss data, the effect of small losses -just above the truncation threshold- on the capital charge can be reduced further by truncating the loss data at a higher recording level H, as was seen in Fig.4.7.

Relative efficiency for non-truncated and truncated loss data In practice, we have to justify the choice of the tuning parameter c. We definitely wish to achieve a high level of robustness, such that the capital charge remains stable under contamination of the loss data. On the other hand, we wish to maintain a certain level of efficiency at the exact model distribution, so that we have accurate estimates of the capital charge when the loss data does follow the assumed severity distribution exactly. For non-truncated loss data, we calculate the amount of efficiency we lose with respect to the MLE; and for truncated loss data, we calculate the amount of efficiency we lose with respect to the CML estimators at the exact severity distribution. Ideally, we would compute the asymptotic relative efficiency[6] (ARE) of the OBRE with respect to the MLE and CML estimators. But, since we have no closed-form expressions for the ARE of the OBRE, we resort to the relative efficiency (RE), which we compute by means of Monte Carlo simulation.

We generate $K = 100$ loss data samples of size $n = 1000$ from the lognormal distribution, truncated at the threshold H, (where $H = 0$ corresponds to non-truncated loss data). For each sample, we compute the CML estimators and the OBRE of the unknown parameters $\theta = (\mu, \sigma)$ for different values of c. We calculate the CML covariance matrix Σ_0 by replacing all its entries by the corresponding mean-squared errors and do the same for the OBRE covariance matrix Σ_1. By the Central Limit Theorem, Σ_0 converges in probability to the inverse Fisher information matrix. That is, if we write $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma})$ for the CML estimators, we have

$$\Sigma_0 = \begin{bmatrix} \frac{1}{K}\sum_i(\widehat{\mu}_i - \mu)^2 & \frac{1}{K}\sum_i(\widehat{\mu}_i - \mu)(\widehat{\sigma}_i - \sigma) \\ \frac{1}{K}\sum_i(\widehat{\mu}_i - \mu)(\widehat{\sigma}_i - \sigma) & \frac{1}{K}\sum_i(\widehat{\sigma}_i - \sigma)^2 \end{bmatrix} \overset{P}{\to} A^T A \tag{5.29}$$

And for the OBRE covariance matrix Σ_1, with the OBRE written as $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma})$, we have:

$$\Sigma_1 = \begin{bmatrix} \frac{1}{K}\sum_i(\tilde{\mu}_i - \mu)^2 & \frac{1}{K}\sum_i(\tilde{\mu}_i - \mu)(\tilde{\sigma}_i - \sigma) \\ \frac{1}{K}\sum_i(\tilde{\mu}_i - \mu)(\tilde{\sigma}_i - \sigma) & \frac{1}{K}\sum_i(\tilde{\sigma}_i - \sigma)^2 \end{bmatrix} \overset{P}{\to} V(\phi, F) \tag{5.30}$$

Where $V(\phi, F)$ denotes the asymptotic covariance matrix of the OBRE according to Eq.(5.21). In the next step, we calculate the relative efficieny (RE) of the K estimates, which in the two-parameter case is defined as

$$\text{RE}(\Sigma_0, \Sigma_1) = \sqrt{\frac{\det(\Sigma_0)}{\det(\Sigma_1)}} \tag{5.31}$$

We repeat this process 5 times and the resulting 5 REs are again averaged. In Table 5.1 below, the average results are presented for different values of the tuning parameter c.

		Tuning parameter				
		$c = \infty$	$c = 7$	$c = 5$	$c = 3$	$c = 2$
RE	$H = 0$	1	0.995	0.979	0.894	0.798
	$H = 10,000$	1	0.992	0.972	0.899	0.765
	$H = 25,000$	1	0.996	0.974	0.897	0.796
	$H = 50,000$	1	0.991	0.965	0.881	0.761
	$H = 100,000$	1	0.987	0.957	0.893	0.783

Tab. 5.1.: RE of the OBRE with respect to MLE for non-truncated loss data and the CML estimators for truncated loss data from $\mathcal{LN}(10.95, 1.75)$.

As expected we achieve the highest possible efficiency when we remove the bound on the IF, i.e. $c = \infty$, and the level of efficiency decreases for lower values of the tuning parameter c. The truncation

[6] The relative efficiency of two procedures is the ratio of their efficiencies. The efficiencies and the relative efficiency of two procedures theoretically depend on the sample size available for the given procedure, but it is often possible to use the *asymptotic relative efficiency* (defined as the limit of the relative efficiencies as the sample size grows) as the principal comparison measure.

of the loss data seems to have no significant impact in the relative efficiency that is achieved[7]. If we combine these results with the plots in Fig.5.2, we have a clear view on the amount of efficiency we have to sacrifice at the exact model and the level of stability we gain in the VaR measures for different values of c and H.

In the following section we study the OBRE for loss data from the log-gamma and Generalized Pareto severity distribution.

5.1.5 OBRE for log-gamma and Generalized Pareto severity distribution

For loss data from the log-gamma and Generalized Pareto severity distribution, the process is largely similar to that of the lognormal distribution. We skip the computations of the OBRE for non-truncated loss data and immediately derive the OBRE for truncated loss data. In this way, we can calculate the OBRE for non-truncated loss data as a particular case, with truncation level $H = 0$.

OBRE for log-gamma severity distribution Let us denote by f and F the density and cumulative distribution function of the left-truncated log-gamma distribution according to Eq.(3.39) and Eq.(3.40), with fixed truncation threshold H and parameter values $\theta = (a, b)$. We set the initial parameter values $\theta^{(0)} = (\widehat{a}_{\text{CML}}, \widehat{b}_{\text{CML}})$, which are the estimated values of the constrained maximum likelihood approach. Furthermore, we specify the truncated score function and Fisher information as:

$$s(x,\theta) = \begin{bmatrix} [\log(\theta_b) - \psi_0(\theta_a) + \log(x)] + \frac{\int_0^H [\log(\theta_b)-\psi_0(\theta_a)+\log(x)] f(y|\theta_a,\theta_b)dy}{1-F(H|\theta_a,\theta_b)} \\ \left[\frac{\theta_a}{\theta_b} - x\right] + \frac{\int_0^H \left[\frac{\theta_a}{\theta_b}-y\right] f(y|\theta_a,\theta_b)dy}{1-F(H|\theta_a,\theta_b)} \end{bmatrix} \tag{5.32}$$

$$A^T(\theta)A(\theta) = \begin{bmatrix} \int_H^\infty \frac{\partial\phi_{\theta_a}}{\partial\theta_a} dF(y|\theta_a,\theta_b) & \int_H^\infty \frac{\partial\phi_{\theta_a}}{\partial\theta_b} dF(y|\theta_a,\theta_b) \\ \int_H^\infty \frac{\partial\phi_{\theta_b}}{\partial\theta_a} dF(y|\theta_a,\theta_b) & \int_H^\infty \frac{\partial\phi_{\theta_b}}{\partial\theta_b} dF(y|\theta_a,\theta_b) \end{bmatrix}^{-1} \tag{5.33}$$

Where the truncated score function is the negative of Eq.(B.20) and the coordinates of the Fisher information matrix are specified by the negative of Eq.(B.21) to Eq.(B.28). For the choice $H = 0$, the above equations reduce to the input of the non-truncated log-gamma OBRE with:

$$\theta^{(0)} = (\theta_a, \theta_b) = (\widehat{a}_{\text{MLE}}, \widehat{b}_{\text{MLE}}) \tag{5.34}$$

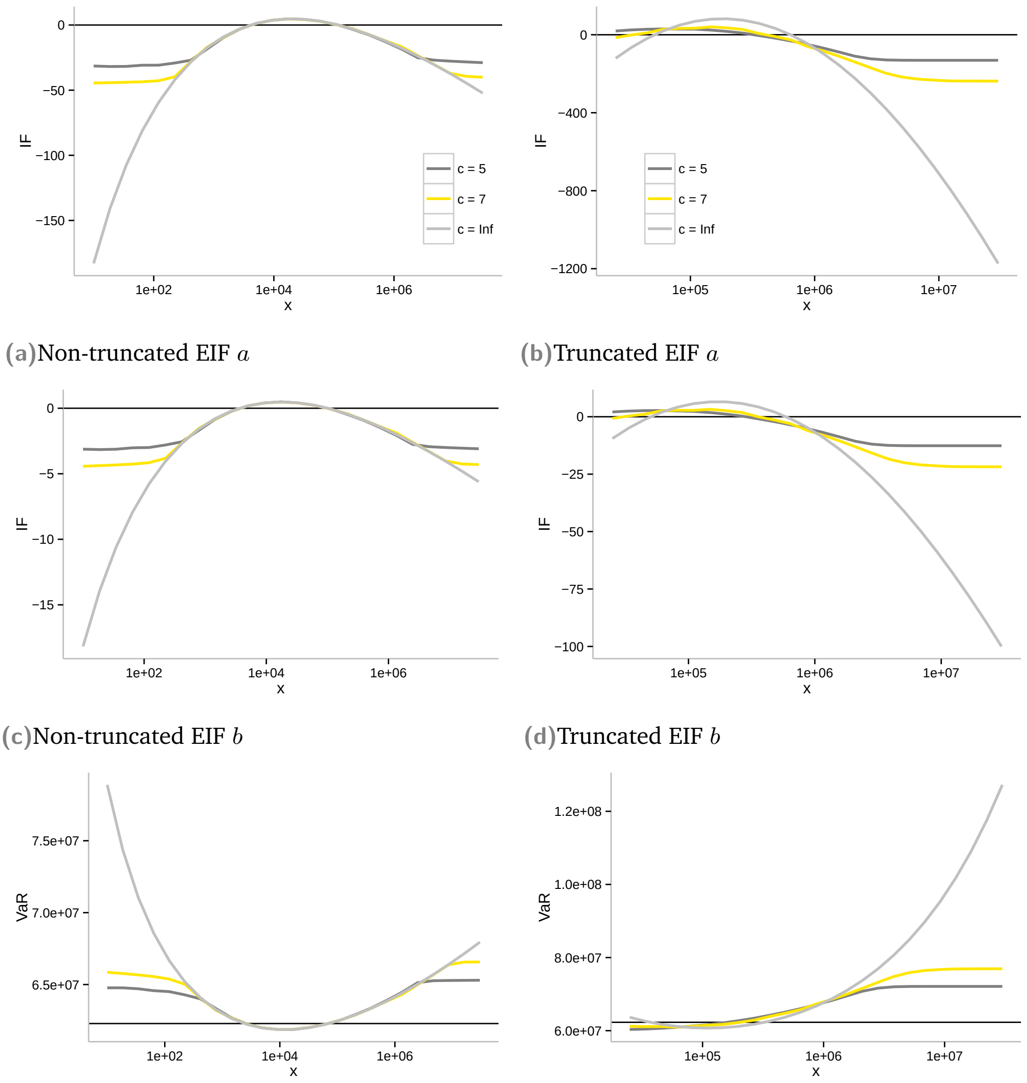
$$s(x,\theta) = \begin{bmatrix} \log(\theta_b) - \psi_0(\theta_a) + \log(x) \\ \theta_a/\theta_b - x \end{bmatrix} \tag{5.35}$$

$$A^T(\theta)A(\theta) = \begin{bmatrix} \psi_1(\theta_a) & -1/\theta_b \\ -1/\theta_b & \theta_a/\theta_b^2 \end{bmatrix}^{-1} \tag{5.36}$$

We compute the OBRE EIFs and corresponding Δ-VaR figures according to the procedures described in the previous section. For non-truncated loss data from the log-gamma distribution, the average EIFs of the shape parameter a and rate parameter b and the corresponding Δ-VaR for loss data samples of size $n = 100$, with fixed frequency distribution $\text{Pois}(25)$, can be found in Fig.5.3a, Fig.5.3c and Fig.5.3e respectively. For truncated loss data from the log-gamma distribution, with fixed threshold $H = 25000$, the plots can be found in Fig.5.3b, Fig.5.3d and Fig.5.3f respectively.

Furthermore, we compute the relative efficiency of the OBRE with respect to the MLE for non-truncated loss data and with respect to the CML estimators for truncated loss data in the same way as described in the previous section. In Table 5.2 below, the results are presented for different values of the tuning parameter c and truncation threshold H.

[7] We note that we simulate loss data samples of the same size for different truncation thresholds. In practice, if the truncation threshold H increases, the loss data sample shrinks and we would achieve a lower level of efficiency compared to lower values of H.



(a)Non-truncated EIF a

(b)Truncated EIF a

(c)Non-truncated EIF b

(d)Truncated EIF b

(e)Non-truncated Δ-VaR

(f)Truncated Δ-VaR

Fig. 5.3.: OBRE approximated IFs and Δ-VaR for log-gamma severity distribution $\mathcal{LG}(34.5, 3.5)$.

		Tuning parameter			
		$c = \infty$	$c = 7$	$c = 5$	$c = 3$
RE	$H = 0$	1	0.984	0.979	0.894
	$H = 10,000$	1	0.981	0.951	0.877
	$H = 25,000$	1	0.978	0.939	0.880
	$H = 50,000$	1	0.985	0.925	0.865
	$H = 100,000$	1	0.977	0.942	0.871

Tab. 5.2.: RE of the OBRE with respect to the MLE for non-truncated loss data and the CML estimators for truncated loss data from $\mathcal{LG}(34.5, 3.5)$.

From the Δ-VaR approximation in Fig.5.3f, we observe that for the log-gamma distribution, under truncation of the loss data, the sensitivity of the VaR measures to small losses (just above the truncation threshold) is completely removed. Thus, under these circumstances in practice, we argue that it suffices to fit a single severity distribution to the loss data and use the OBRE to estimate the distribution parameters. In this way, we do not have to split the loss data into a separate body and tail region, thereby making the estimation process more straightforward and less sensitive to contamination in the right tail, (compare this to the results in Fig.4.9b).

OBRE for Generalized Pareto severity distribution We denote by f and F the density and cumulative distribution function of the left-truncated Generalized Pareto distribution according to Eq.(3.39) and Eq.(3.40) respectively, with truncation threshold H and parameter values $\theta = (\xi, \beta)$. We set the initial parameter values $\theta^{(0)} = (\widehat{\xi}_{\text{CML}}, \widehat{\beta}_{\text{CML}})$, which are the estimated values of the constrained maximum likelihood approach. The truncated score function and the Fisher information matrix can be written as:

$$\begin{aligned} s(x,\theta) &= \begin{bmatrix} -\frac{1}{\theta_\beta}\left[\frac{\theta_\beta - x}{\theta_\beta + \theta_\xi x}\right] + \frac{\int_0^H -\frac{1}{\theta_\beta}\left[\frac{\theta_\beta - y}{\theta_\beta + \theta_\xi y}\right] f(y|\theta_\xi,\theta_\beta)dy}{1-F(H|\theta_\xi,\theta_\beta)} \\ \left[\left(\frac{-x(1+\theta_\xi)}{\theta_\beta\theta_\xi + \theta_\xi^2 x}\right) + \left(\frac{1}{\theta_\xi^2}\log\left(1+\frac{\theta_\xi x}{\theta_\beta}\right)\right)\right] + \frac{\int_0^H \left[\left(\frac{-y(1+\theta_\xi)}{\theta_\beta\theta_\xi + \theta_\xi^2 y}\right) + \left(\frac{1}{\theta_\xi^2}\log\left(1+\frac{\theta_\xi y}{\theta_\beta}\right)\right)\right] f(y|\theta_\xi,\theta_\beta)dy}{1-F(H|\theta_\xi,\theta_\beta)} \end{bmatrix} \\ A^T(\theta)A(\theta) &= \begin{bmatrix} \int_H^\infty \frac{\partial\phi_{\theta_\xi}}{\partial\theta_\xi} dF(y|\theta_\xi,\theta_\beta) & \int_H^\infty \frac{\partial\phi_{\theta_\xi}}{\partial\theta_\beta} dF(y|\theta_\xi,\theta_\beta) \\ \int_H^\infty \frac{\partial\phi_{\theta_\beta}}{\partial\theta_\xi} dF(y|\theta_\xi,\theta_\beta) & \int_H^\infty \frac{\partial\phi_{\theta_\beta}}{\partial\theta_\beta} dF(y|\theta_\xi,\theta_\beta) \end{bmatrix}^{-1} \end{aligned}$$

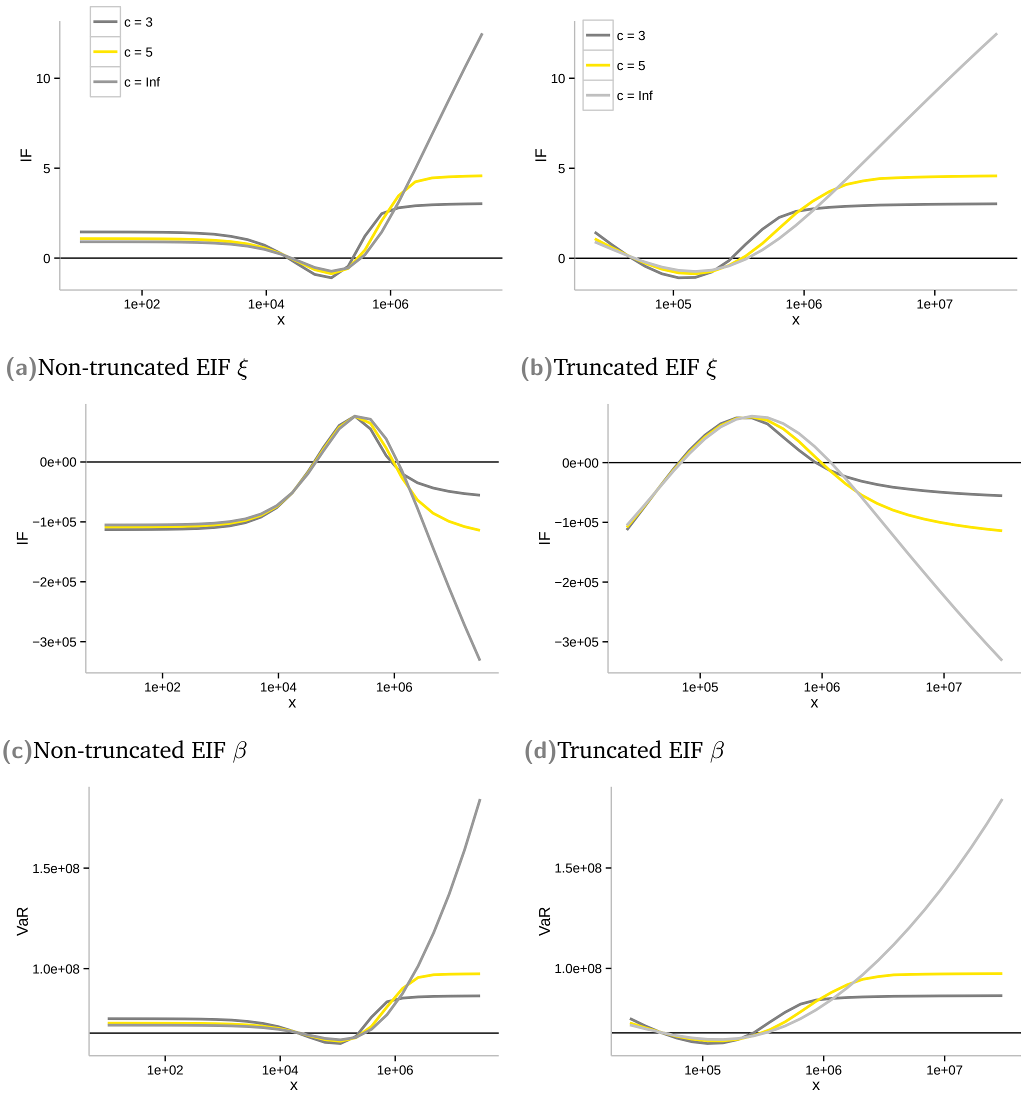
Where the truncated score function is the negative of Eq.(B.31) and the coordinates of the Fisher information matrix are given by the negative of Eq.(B.32) to Eq.(B.42).
We note that, for the choice $H = 0$, the above equations reduce to the input of the non-truncated GPD OBRE with:

$$\theta^{(0)} = (\theta_\xi, \theta_\beta) = (\widehat{\xi}_{\text{MLE}}, \widehat{\beta}_{\text{MLE}}) \tag{5.37}$$

$$s(x,\theta) = \begin{bmatrix} -\frac{1}{\theta_\beta}\left(\frac{\theta_\beta - x}{\theta_\beta + \theta_\xi x}\right) \\ \frac{-x(1+\theta_\xi)}{\theta_\beta\theta_\xi + \theta_\xi^2 x} + \frac{1}{\theta_\xi^2}\log\left(1+\frac{\theta_\xi x}{\theta_\beta}\right) \end{bmatrix} \tag{5.38}$$

$$A^T(\theta)A(\theta) = \begin{bmatrix} (1+\theta_\xi)^2 & -(1+\theta_\xi)\theta_\beta \\ -(1+\theta_\xi)\theta_\beta & 2(1+\theta_\xi)\theta_\beta^2 \end{bmatrix} \tag{5.39}$$

We compute the OBRE EIFs and corresponding Δ-VaR approximations according to the procedures described in the previous section. For non-truncated loss data from the GPD, the average EIFs of the shape parameter ξ and scale parameter β and the corresponding Δ-VaR for loss data samples of size $n = 100$, with fixed frequency distribution $\text{Pois}(25)$, can be found in Fig.5.4a, Fig.5.4c and Fig.5.4e. For truncated loss data from the GPD, with fixed threshold $H = 25000$, the resulting plots are presented in Fig.5.4b, Fig.5.4d and Fig.5.4f.
The relative efficiency of the OBRE with respect to the MLE for non-truncated loss data; and with respect to the CML estimators for truncated loss data can be found in Table 5.3 below.



(a)Non-truncated EIF ξ

(b)Truncated EIF ξ

(c)Non-truncated EIF β

(d)Truncated EIF β

(e)Non-truncated Δ-VaR

(f)Truncated Δ-VaR

Fig. 5.4.: OBRE approximated IFs and Δ-VaR for log-gamma severity distribution $\mathrm{GPD}(0.65, 57500)$.

		Tuning parameter			
		$c = \infty$	$c = 7$	$c = 5$	$c = 3$
RE	$H = 0$	1	0.959	0.910	0.795
	$H = 10,000$	1	0.939	0.899	0.783
	$H = 25,000$	1	0.966	0.918	0.810
	$H = 50,000$	1	0.961	0.862	0.778
	$H = 100,000$	1	0.954	0.877	0.789

Tab. 5.3.: RE of the OBRE with respect to the MLE for non-truncated loss data and the CML estimators for truncated loss data from $\mathrm{GPD}(0.65, 57500)$.

5.1.6 Concluding remarks

In this section we have shown how to compute the optimal bias robust estimators for both non-truncated and truncated loss data from the lognormal, log-gamma and Generalized Pareto severity distribution.

It is worth noting that after implementation of the numerical algorithm described in Section 5.1.2, for a particular severity distribution, we only need to specify: the initial parameter values, the score functions and the Fisher information matrices. In this thesis, we have derived analytic expressions of the scores and Fisher information ourselves. Of course, in practice one does not have to perform these calculations, and can simply extract the results from existing literature.

From the derived Δ-VaR figures, we conclude that the OBRE are able to significantly reduce the impact of contamination of the loss data on the estimated capital charges. Furthermore, by computing the RE of the OBRE we have seen that they tend to keep a *reasonably* high level of efficiency when the loss data follows the assumed severity distribution exactly. The next step is to assess the efficiency of the OBRE with respect to the estimation methods in the classical model (i.e. MLE and the CML approach), when the loss data does *not* follow the assumed model distribution exactly. This is the goal of the simulation studies in the following chapter.

Remark 5.0.5. By recalling the constrained maximum likelihood (CML) approach in Section 3.6.2, we argued that we could apply the OBRE numerical algorithm to compute the CML estimators for truncated loss data. At this point we can be more precise: throughout the thesis to compute the CML estimators, we use the OBRE numerical algorithm for truncated loss data, with tuning parameter $c = \infty$.

5.2 Method of Trimmed Moments (MTM)

In the previous section, we derived the optimal bias robust estimators for non-truncated and truncated loss data and examined their behavior for the usual heavy-tailed severity distributions. The aim of the OBRE is to find robust estimators, such that the efficiency loss is minimal. In this sense, in terms of performance at the assumed model distribution, it is *by definition* one of the best possible robust estimation procedures. However, if we look at the estimation procedure from a practical point of view, it also has several drawbacks. The most important one is its computational complexity and the lack of transparency. It is quite difficult to explain the definition of the OBRE (in Section 5.1) and the algorithm we use to compute the OBRE (in Section 5.1.2) to a non-mathematician. Avoidance of such complexities is very appealing in practice. An attempt to resolve the aforementioned issues was made by Brazauskas et al. [6], who introduced a general robust estimation procedure, called the *method of trimmed moments* (MTM). This method essentially works like the standard method of moments and thus is easy to understand and is fairly simple analytically and computationally.

In this section we study the MTM as a possible alternative to the OBRE. It may not achieve the high level of efficiency that the OBRE does, but it is easier to understand and implement which is an important property in practical applications. As for any robust estimation procedure we gain distributional robustness (i.e. resistance against outliers) by trading off some efficiency at the assumed model distribution. For the MTM this is done by specifying the trimming proportions, which can be easily specified and understood by the user. (We compare this to the tuning parameter c in the OBRE, where the gain in robustness due to a change in the value of c is not immediately clear).

In the following section, we describe the general outline of the MTM estimation process for loss data from the severity distribution F.

5.2.1 MTM estimation procedure

Suppose that we observe a loss data sample $x_1, x_2, \ldots, x_n$, with common severity distribution function F. We consider the parameter space $\Theta \subseteq \mathbb{R}^k$. Thus, we wish to estimate k distribution parameters $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$. By $x_{(i)}$ we denote the i-the order statistic of the sample and (a_i, b_i) are the i-th trimming proportions, where a_i denotes the fraction of observations that is trimmed in the left tail of the loss data and b_i denotes the fraction of observations that is trimmed in the right tail.
Following Brazauskas et al.[6] the MTM estimators can be found via the following estimation procedure:

1. For $j = 1, \cdots, k$, we specify constant trimming proportions (a_j, b_j) and we compute the sample trimmed moments $\widehat{\mu}_j$, given by

$$\widehat{\mu}_j = \frac{1}{n - na_j - nb_j} \sum_{i=\lfloor na_j \rfloor + 1}^{n - \lfloor nb_j \rfloor} h_j(x_{(i)}) \tag{5.40}$$

   Where $h_j : \mathbb{R} \to \mathbb{R}$ is a function chosen by the researcher. Note that, if we choose $h_j(x) = x^j$, we get the *ordinary* sample trimmed moments.

2. For $j = 1, \cdots, k$, we derive the corresponding population trimmed moments μ_j according to

$$\mu_j = \mu_j(\theta) = \frac{1}{1 - a_j - b_j} \int_{a_j}^{1-b_j} h_j(F^{-1}(u))du \tag{5.41}$$

   We note that choosing trimming proportions $a_j = b_j = 0$, gives $\mu_j = \mathbb{E}[h_j(X)]$ (which corresponds to the classical method of moments).

3. We solve the system of equations that follows by matching the sample trimmed moments with the population trimmed moments with respect to $\theta_1, \cdots, \theta_k$, so that

$$\begin{cases} \mu_1(\theta_1, \ldots, \theta_k) & = & \widehat{\mu}_1 \\ & \vdots & \\ \mu_k(\theta_1, \ldots, \theta_k) & = & \widehat{\mu}_k \end{cases} \tag{5.42}$$

4. The obtained solutions to the equations above are the MTM estimators of the parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$:

$$\begin{cases} \widehat{\theta}_1 & = & g_1(\widehat{\mu}_1, \ldots, \widehat{\mu}_k) \\ \vdots & & \\ \widehat{\theta}_k & = & g_k(\widehat{\mu}_1, \ldots, \widehat{\mu}_k) \end{cases} \tag{5.43}$$

To apply the estimation procedure we have to specify the functions $h_j(x)$ and the trimming proportions a_j and b_j. A natural choice is $h_j(x) = x^j$ for all $j = 1, \ldots, k$, which will lead to the matching of ordinary trimmed moments. For the choice of the trimming proportions a_j and b_j, there is no single answer as to how much robustness is needed, typically an argument is made on the basis of how much efficiency at the assumed model F we are willing to sacrifice to gain a certain level of robustness. If we choose $a_j = b_j = 0$ for all $j = 1, \ldots, k$, we get the classical method of moments. Since the trimming proportions correspond to the fraction of loss observations that are cut off from the tails, the MTM estimators will not be influenced by the proportion $a_* = \min\{a_1, \ldots, a_k\}$ of lowest loss observations and the proportion $b_* = \min\{b_1, \ldots, b_k\}$ of highest loss observations. In general, the MTM estimators will not quite reach the amount of efficiency that is achieved by the OBRE, this is mainly because the MTM estimators are unable to make a smooth transition between trimming (rejecting) and not trimming (accepting) a certain loss data point, whereas the OBRE is able to down weight outlying observations only partially.

5.2.2 MTM estimation for truncated loss data

With the estimation procedure introduced above we are able to compute the MTM estimators for non-truncated loss data from each of the three usual heavy-tailed severity distributions (lognormal, log-gamma and GPD). Now, we need to assess whether it also works for truncated loss data. If not, this method would be of little value in practice to compute the operational risk capital charge. To our knowledge, there is currently no literature available on applying the MTM to truncated loss data. But, as it turns out, we can extend the estimation procedure quite easily in order to cope with truncated loss data.

The natural way to compute the MTM estimators for an observed truncated loss data sample of $x_1, x_2, \ldots, x_n$ truncated at threshold H, would be to substitute the non-truncated severity distribution function F by its truncated version F_H, according to Eq.(3.40). This is what we have done each time throughout the thesis when dealing with truncated loss data. However, in this context the following problem arises: in order to solve the system of equations in step (3) of the MTM estimation procedure, we need analytic expressions for the population trimmed moments. If we are even able to derive analytic expressions for the population trimmed moments of the considered truncated severity distributions, it is likely that the resulting system of moment equations will be (very) difficult to solve. Hence, in order to avoid tedious analytic computations, we consider a different approach.

Let us consider a truncated loss data sample $x_1, x_2, \ldots, x_n$, truncated at threshold H. Suppose, that we have chosen the left trimming proportion a, in such a way that it coincides with the fraction of observations that is cut off from below, i.e. the fraction of observations below H. Then, by definition of the MTM, equating the population trimmed moments with the truncated sample moments, should exactly give us the MTM estimators for the truncated loss data sample with severity distribution F_H. It is clear that if we choose $a(\theta) = F(H, \theta)$, the left trimming proportion corresponds exactly with the fraction of observations below H. However, initially the true parameters θ are unknown, so we do not know what value to choose for the left trimming proportion $a(\theta)$. We propose an iterative procedure to get successively better approximations of the left trimming proportion $a(\theta)$ and the corresponding MTM estimators for truncated loss data. Below we describe this iterative procedure in more detail.

Outline MTM procedure for truncated loss data Consider an observed truncated loss data sample $x_1, x_2, \ldots, x_n$, with severity distribution F truncated at recording threshold H. We compute the MTM estimators of the truncated loss data sample according to the following procedure:

1. Fix a precision threshold ϵ, and specify initial parameter values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$. Furthermore, choose constant right trimming proportions $b_1 = \ldots = b_k := b$ and set the left trimming proportions to $a_1 = \ldots = a_k := F(H, \theta^{(0)})$.

2. We need to correct the right trimming proportions b for the fact that the loss data is already left-truncated. If we put n the size of the truncated loss data sample, then the expected size of the original non-truncated loss data sample is given by $m = n/(1 - F(H, \theta^{(0)}))$. We write $\tilde{b}_1 = \ldots = \tilde{b}_n = bm/n$ for the right trimming proportions under the non-truncated data sample.

3. For $j = 1, \ldots, k$, compute the sample trimmed moments according to

$$\widehat{\mu}_j = \frac{1}{n - n\tilde{b}_j} \sum_{i=1}^{n - \lfloor n\tilde{b}_j \rfloor} h_j(x_{(i)}) \tag{5.44}$$

Where $h_j : \mathbb{R} \to \mathbb{R}$ is a function chosen by the researcher. Note that we use the adjusted right trimming proportions $\tilde{b}_j$, since the loss data sample is already truncated from below.

4. For $j = 1, \dots, k$ we derive the corresponding population trimmed moments μ_j according to

$$\mu_j = \mu_j(\theta) = \frac{1}{1 - a_j - b_j} \int_{a_j}^{1-b_j} h_j(F^{-1}(u))du \tag{5.45}$$

   With left trimming proportion $a_j = F(H, \theta^{(0)})$ and the non-adjusted right trimming proportion $b_j = b$.

5. Follow Step (3) and Step (4) according to the ordinary MTM estimation procedure (as described in Section 5.2.1) to obtain the MTM estimators $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)$.

6. If $\max_j |\theta_j^{(0)} - \widehat{\theta}_j| > \epsilon$, then we choose the new initial values $\theta^{(0)} = \widehat{\theta}$ and return to Step (1), else stop.

The algorithm is convergent provided that the initial parameter values $\theta^{(0)}$ are close to the solution of the MTM. In the first step, we can take for instance the (shifted) MLE as initial parameter values. As will be evaluated below, the truncated MTM estimation procedure is generally much faster than the numerical algorithm of the OBRE. We note that if we wish to adjust a in order to trim more observations in the left tail, (thus reduce the influence of small losses on the parameter estimates), this is equivalent to specifying a higher truncation threshold H and compute the MTM estimators under this new condition. Essentially, this is equivalent to rejection of all data points below the truncation threshold H. In the following sections we will put the MTM estimation procedure into practice and compute the MTM estimators for loss data from the lognormal and log-gamma distribution. To keep the material within bounds, we chose to skip the computation of the MTM estimators for the Generalized Pareto distribution, since the results are largely similar to that of the OBRE. Furthermore, for real operational loss data, the GPD may have infinite moments, as described in Section 3.5.1. In which case we are unable to apply the method of trimmed moments.

Remark 5.0.6. In Brazauskas [4] it is shown that the MTM estimators for non-truncated loss data are asymptotically normal. Furthermore, for loss data from the lognormal distribution the exact asymptotic covariance matrix is derived. Under truncation of the loss data it is not immediately clear whether the asymptotic properties of the MTM estimators are still valid, since we introduce dependence between the parameter values θ and the trimming proportions $a(\theta)$. In the remainder of the thesis we make no use of the asymptotic properties of the MTM estimators (the relative efficiency is computed in the same way as for the OBRE, we use Monte Carlo simulation to approximate the asymptotic covariance matrices). Thus, we will not derive the asymptotic results for the MTM estimators under truncation of the loss data[8].

5.2.3 MTM estimators for lognormal severity distribution

In contrast to the OBRE, we do not have a single numerical algorithm that is able to produce estimators for each severity distribution F. We have to derive analytic expressions for the population trimmed moments for each particular severity distribution. On the other hand, the OBRE estimation procedure is generally quite slow. Once we have derived the MTM estimators for a specific severity distribution, the computation procedure will be very fast in comparison to the OBRE. Let us first consider non-truncated loss data from the lognormal severity distribution.

[8] For the asymptotic properties of the MTM estimators when the loss data is non-truncated, we refer the interested reader to Brazauskas [4].

Non-truncated severity loss data Suppose we observe a non-truncated loss data sample $x_1, x_2, \ldots, x_n$ IID from a lognormal severity distribution. We have unknown parameters $\theta = (\nu, \sigma)$, where we choose to write ν instead of the usual μ to avoid confusion between the parameter μ and the moment μ. To simplify computations we apply a log-transformation to the sample, such that -after the transformation- the random variables follow the normal (Gaussian) distribution. Furthermore, we specify the left and right trimming proportions $a_1 = a_2 := a$ and $b_1 = b_2 := b$; and the first and second order ordinary moment functions $h_1(x) = \log(x)$ and $h_2(x) = (\log(x))^2$, which results in a system of two moment equations.

Following the estimation procedure in Section 5.2.1, the sample trimmed moments are given by

$$\widehat{\mu}_1 = \frac{1}{n - na - nb} \sum_{i=\lfloor na \rfloor + 1}^{n - \lfloor nb \rfloor} \log(x_{(i)}) \tag{5.46}$$

$$\widehat{\mu}_2 = \frac{1}{n - na - nb} \sum_{i=\lfloor na \rfloor + 1}^{n - \lfloor nb \rfloor} \big(\log(x_{(i)})\big)^2 \tag{5.47}$$

The corresponding population trimmed moments are given by

$$\mu_1 = \mu_1(\theta) = \frac{1}{1-a-b} \int_a^{1-b} F^{-1}(u) du \tag{5.48}$$

$$\mu_2 = \mu_2(\theta) = \frac{1}{1-a-b} \int_a^{1-b} \big(F^{-1}(u)\big)^2 du \tag{5.49}$$

Where F is the non-truncated normal cumulative distribution function, with unknown parameters ν and σ. Because the normal distribution belongs to the class of location-scale families[9], the population trimmed moments can be written as

$$\mu_1(\nu, \sigma) = \frac{1}{1-a-b} \int_a^{1-b} F^{-1}(u) du = \nu + \sigma c_1 \tag{5.50}$$

$$\mu_2(\nu, \sigma) = \frac{1}{1-a-b} \int_a^{1-b} \big(F^{-1}(u)\big)^2 du = \nu^2 + 2\nu\sigma c_1 + \sigma^2 c_2 \tag{5.51}$$

Where

$$c_k = \frac{1}{1-a-b} \int_a^{1-b} \big(\Phi^{-1}(u)\big)^k du \tag{5.52}$$

With Φ the standard normal cumulative distribution function. In this way c_1 and c_2 are parameter free, (the integrals are solved numerically in *R*).

We solve the system of moment equations with respect to ν and σ, which leads to the MTM estimators:

$$\begin{cases} \widehat{\nu} = g_1(\widehat{\mu}_1, \widehat{\mu}_2) = \widehat{\mu}_1 - c_1 \widehat{\sigma} \\ \widehat{\sigma} = g_2(\widehat{\mu}_1, \widehat{\mu}_2) = \sqrt{(\widehat{\mu}_2 - \widehat{\mu}_1^2)/(c_2 - c_1^2)} \end{cases} \tag{5.53}$$

Our goal in applying the MTM estimators was reducing the computational complexity, it is worth noting that the computation of the MTM estimators according to the equations above, is definitely easier to implement and understand than the computation of the OBRE for the lognormal distribution.

Truncated severity loss data Suppose we observe a truncated loss data sample $x_1, \ldots, x_n$ from the lognormal distribution, truncated at the recording threshold H. Again we write $\theta = (\nu, \sigma)$ for the unknown location parameter ν and scale parameter σ. Following the proposed MTM estimation procedure in Section 5.2.2, we have to specify the right trimming proportions b and initial parameter

[9] A location-scale family is a family of probability distributions parametrized by a location parameter and a (non-negative) scale parameter. If X is a zero-mean, unit-variance member of the family, then every member Y of the family can be written as $\mu + \sigma X$, where μ and σ are the mean and standard deviation of Y.

values $\theta^{(0)}$. As initial parameter values $\theta^{(0)}$ we choose the shifted maximum likelihood estimators (according to Section 3.6.2). We encountered no major problems in running the algorithm, and in general the truncated MTM estimators are found in less then 10 iterative steps.

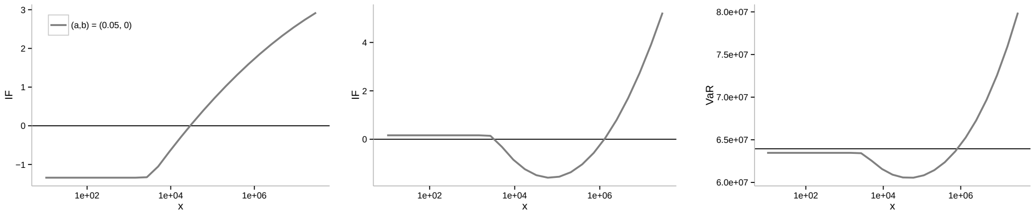
The implementation of the described numerical algorithm in *R* for truncated loss data from the lognormal distribution can be summarized as follows:

```
> # Define the MTM estimators, with trimming (a,b)
> MTM <- function(X, a, b)
            {
                    Y <- sort(X)
                    n <- length(X)
                    m <- n / (1 - a)
                    b.tilde <- b * m / n

                    # Compute the sample trimmed moments
                    mu1.hat <- 1 / (n - round(n * b.tilde)) * sum(Y[1 : (n - round(n * b.tilde))])
                      mu2.hat <- 1 / (n - round(n * b.tilde)) * sum((Y[1 : (n - round(n * b.tilde))])^2)

                    # Compute the population trimmed moments
                    F.inv <- function(x){qnorm(x)}
                    C1 <- 1 / (1 - a - b) * integrate(F.inv, lower = a, upper = b)$value
                    C2 <- 1 / (1 - a - b) * integrate(F.inv^2, lower = a, upper = b)$value

                    sigma.mtm <- sqrt((mu2.hat - (mu1.hat)^2) / (C2 - C1^2))
                    mu.mtm <- mu1.hat - C1 * sigma.mtm
                    return(mu.mtm, sigma.mtm)
            }
> # Implement the iterative procedure, with 'X' the truncated loss data, 'theta0' the
> # shifted maximum likelihood estimators, 'b' the right trimming
> # proportion and 'eps' the precision threshold.
> repeat{
                    a0 <- pnorm(H, theta0[1], theta0[2])
                    theta.hat <- MTM(X = X, a = a0, b = b)
                    if(abs(theta.hat[1] - theta0[1]) < eps[1] & abs(theta.hat[2] - theta0[2]) < eps[2]){
                       return(theta.hat)
                       break
                       }
                    else{
                    theta0 <- theta.hat
                       }
             }
```

MTM IFs and Δ-VaR for non-truncated and truncated loss data For non-truncated loss data, we approximate the MTM IFs by the average EIFs of $K = 10,000$ non-truncated loss data samples[10] of size $n = 250$ from the lognormal distribution. The Δ-VaR figures are constructed in the same way as for the OBRE; for loss data samples of size $n = 100$ and with frequency distribution $\mathrm{Pois}(25)$.

If we consider trimming proportions $a = b = 0$, we obtain the ordinary method of moment estimators, which for the lognormal distribution coincide with the maximum likelihood estimators. For $a + b \to 1$, we trim (nearly) all loss observations, which leads to estimators based on single quantiles (e.g. the median). These estimators are clearly the most robust, but consequently also the least efficient. A definite advantage in the MTM procedure compared to the OBRE procedure is that the trimming proportions are more flexible. In the OBRE process, we can only specify a single tuning parameter c, whereas in the MTM process we can specify (at least) two trimming proportions a and b. By choosing $a > 0$ and $b = 0$, this allows us for instance to trim only observations in the left tail, therefore directly gaining robustness against small losses, without cutting off observations in the right tail, thus not losing valuable information contained in the extreme large losses. For the choice $a = 0.05$ and $b = 0$, the approximate IFs and corresponding Δ-VaR can be found in Fig.5.5.

[10] Note that we are able to compute the average of $K = 10,000$ loss data samples, while in the case of the OBRE we considered $K = 100$, since the OBRE takes much longer to compute.

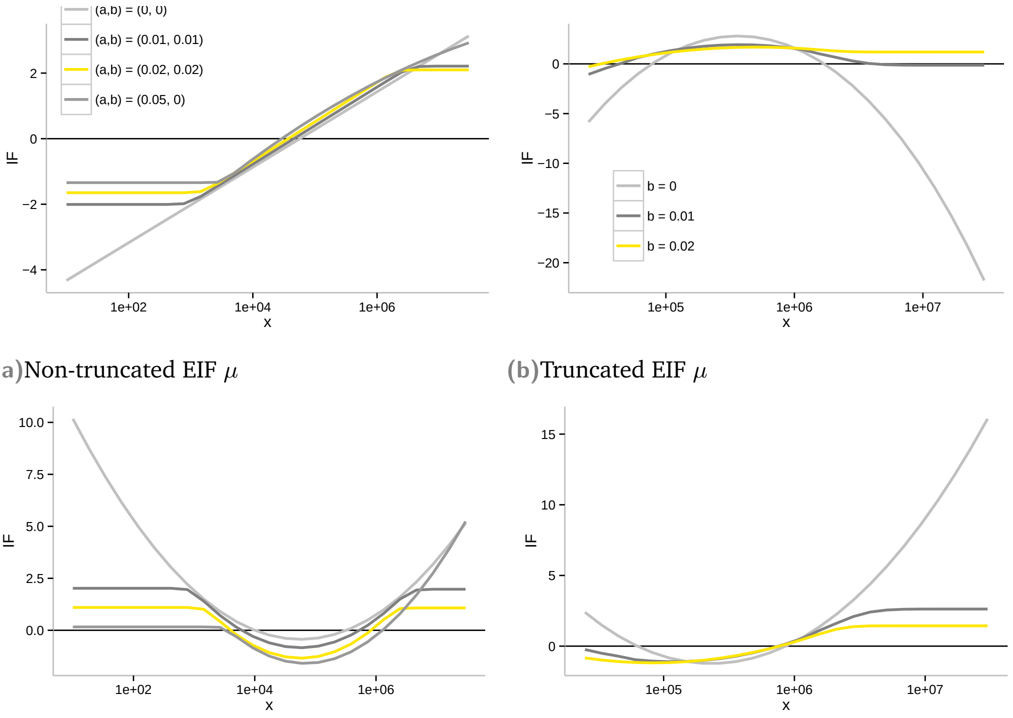


(a)EIF μ (b)EIF σ (c)Δ-VaR

Fig. 5.5.: MTM approximated IFs and Δ-VaR for lognormal severity distribution $\mathcal{LN}(10.95, 1.75)$, with trimming $(a, b) = (0.05, 0)$.

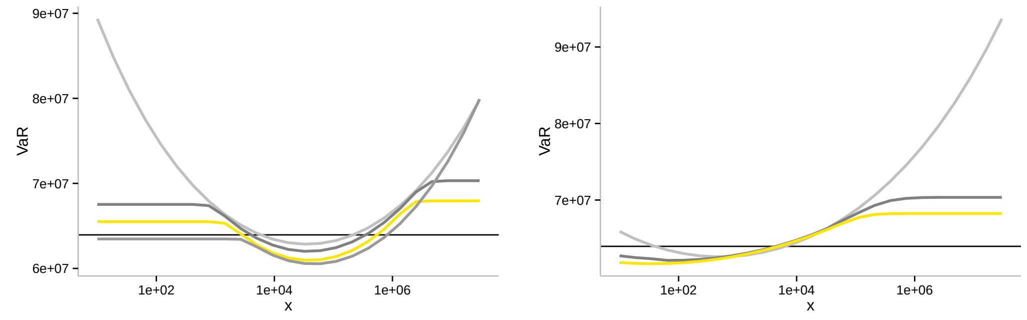
For several other non-symmetric and symmetric values of the trimming proportions (a, b) the approximated MTM IFs and corresponding Δ-VaR approximations are presented in Fig.5.6a, Fig.5.6c and Fig.5.7a respectively.

For truncated loss data, we approximate the MTM IFs by the average EIFs of $K = 1,000$ truncated loss data samples from the lognormal distribution of size $n = 250$, with fixed truncation threshold $H = 25,000$. For different values of the right trimming proportion b, the approximated IFs and corresponding Δ-VaR approximations can be found in Fig.5.6b, Fig.5.6d and Fig.5.7b.



(a)Non-truncated EIF μ (b)Truncated EIF μ

(c)Non-truncated EIF σ (d)Truncated EIF σ



(a)Non-truncated Δ-VaR (b)Truncated Δ-VaR

Fig. 5.7.: MTM approximated IFs and Δ-VaR for lognormal severity distribution $\mathcal{LN}(10.95, 1.75)$.

It is seen that the shape of the MTM IFs is largely similar to the shape of the OBRE IFs. The main difference is that the MTM IFs are somewhat less *smooth*, we argue that this is because of the fact that the OBRE procedure is able to down weight outlying observations only partially, while the MTM procedure can only completely accept or reject outlying observations.

The truncated MTM estimators adequately reduce the sensitivity of the capital charge to contamination of the loss data. The influence of small losses (just above the threshold H) on the capital charge is fully mitigated. The VaR measures are also seen to be less sensitive to contamination in the right tail. Thus, making the capital charge more stable overall.

By specifying different sets of trimming proportions we are able to achieve practically similar levels of robustness as for the OBRE. However, we expect the MTM estimators to achieve a lower amount of relative efficiency than the OBRE.

Relative efficiency for non-truncated and truncated loss data The relative efficiency (RE) of the MTM estimators with respect to the MLE is calculated in the same way as in Section 5.1.4: we generate $K = 10,000$ non-truncated loss data samples of size $n = 1000$ from the lognormal distribution. For each sample, we estimate the MLE and the MTM estimators of the unknown parameters for different values of the trimming proportions (a, b). Next we compute the MLE covariance matrix Σ_0 by replacing all its entries by the corresponding mean-squared errors and do the same for the MTM covariance matrix Σ_1. The relative efficiency of the K estimates is then given by

$$\text{RE}(\Sigma_0, \Sigma_1) = \sqrt{\frac{\det(\Sigma_0)}{\det(\Sigma_1)}} \tag{5.54}$$

This process is repeated 10 times and the resulting 10 REs are again averaged. In Fig.5.8 below, we give a surface plot of the average results with trimming proportions $0 \le a \le 0.25$ and $0 \le b \le 0.25$.

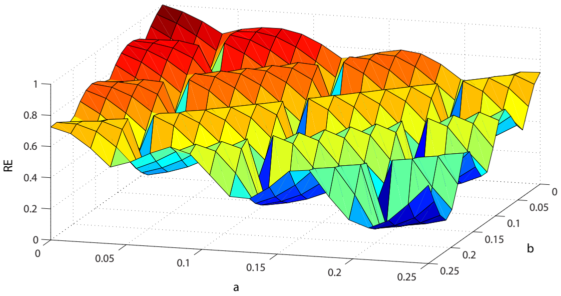


Fig. 5.8.: RE of the MTM estimators with respect to the MLE for non-truncated loss data from $\mathcal{LN}(10.95, 1.75)$.

As expected the relative efficiency decreases as the values of the trimming proportions (a, b) increase. We explain the observed behavior by a heuristic argument: in calculating the sample trimmed moments, we have to round off the number of lower and upper trimmed observations to the nearest integer. In this case the resulting system of equations will not match perfectly and we expect to achieve a lower level of efficiency, than in the case where the sample trimmed moments and the population trimmed moments match perfectly (i.e. when we do not have to round off the number of trimmed observations). In conclusion, this results in a *wave*-like surface, with the RE subsequently increasing and decreasing. For truncated loss data, we compute the RE of the MTM estimators with respect to the CML estimators in the same way as in Section 5.1.4, for $K = 10,000$ truncated loss data samples of size $n = 1000$ from the lognormal distribution. In Table 5.4 below, the RE is computed for different values of the threshold H and the trimming proportion b.

		Trimming proportion				
		$b = 0$	$b = 0.01$	$b = 0.02$	$b = 0.03$	$b = 0.05$
RE	$H = 10,000$	1	0.969	0.949	0.904	0.851
	$H = 25,000$	1	0.930	0.902	0.867	0.832
	$H = 50,000$	1	0.920	0.842	0.803	0.698
	$H = 100,000$	1	0.865	0.795	0.710	0.602

Tab. 5.4.: RE of the MTM estimators with respect to the CML estimators for truncated loss data from $\mathcal{LN}(10.95, 1.75)$.

In contrast to the relative efficiency of the OBRE for truncated loss data, it is seen that the MTM estimators become less efficient as the truncation threshold increases (for fixed trimming proportion b). This can be explained by the following argument: in step (2) of the estimation procedure in Section 5.2.2, we adjust the right trimming proportions b for the fact that the loss data is already left-truncated, according to $\tilde{b} = bm/n$ (with n the size of the observed truncated loss sample and m the expected size of the original non-truncated loss sample). As the truncation threshold increases, the ratio m/n becomes larger, since $n = 1000$ is fixed for each threshold H. Consequently, for a higher truncation threshold, the value $\tilde{b} = bm/n$ increases and we will actually trim more loss observations when computing the sample trimmed moments in step (3). Thus, the resulting MTM estimators are expected to become less efficient as the truncation threshold increases. On the other hand, since the absolute number of trimmed loss observations in the right tail increases, we will also expect them to achieve a higher level of robustness.

5.2.4 MTM estimators for log-gamma severity distribution

Finally, we consider loss data from the log-gamma severity distribution. The derivations are largely similar to previous section, except that no closed-form expressions are found for the MTM estimators. Let us first consider non-truncated loss data from the log-gamma distribution.

Non-truncated severity loss data Suppose we observe a non-truncated loss data sample $x_1, x_2, \ldots, x_n$ IID from the log-gamma distribution. We have unknown parameters $\theta = (\alpha, \beta)$, where we chose to write α and β instead of the usual parameters a and b avoid confusion between the parameters a and b and the trimming proportions a and b. Similar to the MTM estimation procedure for the lognormal distribution, we apply a log-transformation to the sample. The resulting random variables follow an ordinary Gamma distribution with shape parameter α and rate parameter β. Furthermore, we specify the left and right trimming proportions $a_1 = a_2 := a$ and $b_1 = b_2 := b$; and the first and second order ordinary moment functions[11] $h_1(x) = \log(x)$ and $h_2(x) = (\log(x))^2$. Following the estimation procedure in Section 5.2.1, the sample trimmed moments are given by

$$\widehat{\mu}_1 = \frac{1}{n - na - nb} \sum_{i=\lfloor na \rfloor+1}^{n-\lfloor nb \rfloor} \log(x_{(i)}) \tag{5.55}$$

$$\widehat{\mu}_2 = \frac{1}{n - na - nb} \sum_{i=\lfloor na \rfloor+1}^{n-\lfloor nb \rfloor} \left(\log(x_{(i)})\right)^2 \tag{5.56}$$

The corresponding population trimmed moments can be written as

$$\mu_1 = \mu_1(\theta) = \frac{1}{1-a-b} \int_a^{1-b} F^{-1}(u)du \tag{5.57}$$

$$\mu_2 = \mu_2(\theta) = \frac{1}{1-a-b} \int_a^{1-b} (F^{-1}(u))^2 du \tag{5.58}$$

Where F^{-1} denotes the quantile function of the Gamma distribution.
In Steinbrecher and Shaw [31] a power series expansion of the quantile function is derived for the standardized case with shape α unknown and rate $\beta = 1$ fixed. The quantile function for the general case is then also available through

$$F^{-1}(u) = \frac{F^{-1}_{\alpha,1}(u)}{\beta} \tag{5.59}$$

For the derivations of the quantile function we refer to Steinbrecher and Shaw [31]. We will not concern ourselves with the analytic expression and compute the quantile function numerically in *R*. It is then possible to solve the resulting system of equations for α and β due to Eq.(5.59) above.
For the first population trimmed moment we can write

$$\mu_1 = \frac{1}{1-a-b} \int_a^{1-b} F^{-1}(u)du \tag{5.60}$$

$$= \frac{1}{\beta} \cdot \frac{1}{1-a-b} \int_a^{1-b} F^{-1}_{\alpha,1}(u)du \tag{5.61}$$

$$:= \frac{1}{\beta} \cdot \delta_1(\alpha) \tag{5.62}$$

We note that the expression for $\delta_1(\alpha)$ depends only on the parameter α.
Similarly for the second population trimmed moment we have

$$\mu_2 = \frac{1}{1-a-b} \int_a^{1-b} (F^{-1}(u))^2 du \tag{5.63}$$

$$= \frac{1}{\beta^2} \cdot \frac{1}{1-a-b} \int_a^{1-b} \left(F^{-1}_{\alpha,1}(u)\right)^2 du \tag{5.64}$$

$$:= \frac{1}{\beta^2} \cdot \delta_2(\alpha) \tag{5.65}$$

Where $\delta_2(\alpha)$ only depends on the parameter α.
The next step is to solve the moment equations in terms of the parameters α and β. By matching $\widehat{\mu}_1$ and $\widehat{\mu}_2$ to μ_1 and μ_2, the MTM estimator $\widehat{\alpha}$ is found by solving

$$\frac{\widehat{\mu}_1^2}{\widehat{\mu}_2} - \frac{\delta_1(\widehat{\alpha})^2}{\delta_2(\widehat{\alpha})} = 0 \tag{5.66}$$

[11] In Kleefeld and Brazauskas [22] the MTM estimators for the Gamma distribution are computed using the moment functions $h_1(x) = \log(x)$ and $h_2(x) = \log(x)$, and different trimming proportions $(a_1, b_1) \neq (a_2, b_2)$. We choose to follow a different approach, such that we can choose a single set of trimming proportions (a, b).

Given the solution for $\widehat{\alpha}$, we find the MTM estimator $\widehat{\beta}$ through

$$\widehat{\beta} = \frac{\delta_1(\widehat{\alpha})}{\widehat{\mu}_1} \tag{5.67}$$

The implementation of the described estimation procedure in *R* is described in the next paragraph for truncated severity loss data.

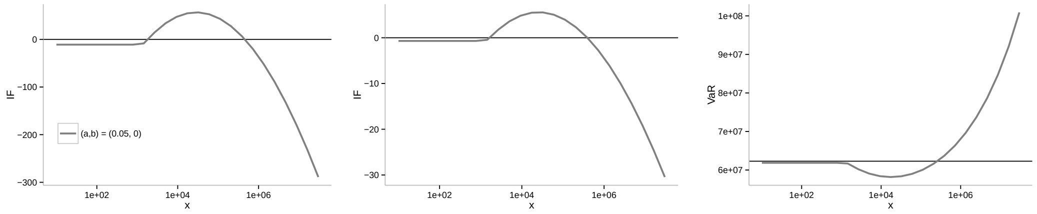
Truncated severity loss data Suppose we observe a truncated loss data sample $x_1, \ldots, x_n$ from the log-gamma distribution, truncated at the recording threshold H. Following the estimation procedure in Section 5.2.2, we have to specify the right trimming proportion b and initial parameter values $\theta^{(0)}$. As initial parameter values $\theta^{(0)}$ we again choose the shifted maximum likelihood estimators (according to Section 3.6.2).

The implementation of the described numerical algorithm in *R* for truncated loss data from the log-gamma distribution can be summarized as follows:

```
> # Define the MTM estimators for non-truncated data 'X', and trimming 'c(a,b)'
> MTM <- function(X, a, b){
              Y <- sort(X)
               n <- length(X)
               m <- n / (1 - a)
               b.tilde <- b * m / n
               # Compute the sample trimmed moments
               mu1.hat <- 1 / (n - round(n * b.tilde)) * sum(Y[1:(n - round(n * b.tilde))])
               mu2.hat <- 1 / (n - round(n * b.tilde)) * sum((Y[1:(n - round(n * b.tilde))])^2)
               # Compute the population trimmed moments and the MTM estimators
               f <- function(a){
                    F.inv <- function(x){qgamma(x, shape = a, rate = 1)}
                    delta1 <- 1 / (1 - a - b) * integrate(F.inv, lower = a, upper = 1-b)$value$
                    delta2 <- 1 / (1 - a - b) * integrate(F.inv^2, lower = a, upper = 1-b)$value$
                    return(delta1^2 / delta2 - mu1.hat^2 / mu2.hat)
                    }
               f.vec <- Vectorize(f, vectorize.args = "a")
               root <- uniroot.all(f.vec, lower = 0, upper = 10^5, n = 1)
               a.hat <- root[1]
               if(length(r) == 0){
                 return(NULL)
               }
               F.inv <- function(x){qgamma(x, shape = a.hat, rate = 1)}
               delta1.a.hat <- 1/(1 - a.hat - b) * integrate(F.inv, lower = a.hat, upper = 1-b)$value$
                b.hat <- delta1.a.hat / mu1.hat
               return(c(a.hat, b.hat))
          }
> # Implement the iterative procedure, with 'X' the truncated loss data, 'theta0' the
> # shifted maximum likelihood estimators, 'b' the right trimming
> # proportion and 'eps' the precision threshold.
> repeat{
              a0 <- pnorm(H, theta0[1], theta0[2])
              theta.hat <- MTM(X = X, a = a0, b = b)
              if(abs(theta.hat[1] - theta0[1]) < eps[1] & abs(theta.hat[2] - theta0[2]) < eps[2]){
                return(theta.hat)
                break
                }
              else{
              theta0 <- theta.hat
                }
               }
```

MTM IFs and Δ-VaR for non-truncated and truncated loss data We approximate the MTM IFs by computing the average EIFs for $K = 10,000$ non-truncated loss data samples from a log-gamma distribution of size $n = 250$. Using the approximated IFs we construct the Δ-VaR figures for loss data samples of size $n = 100$ and with frequency distribution $\mathrm{Pois}(25)$.

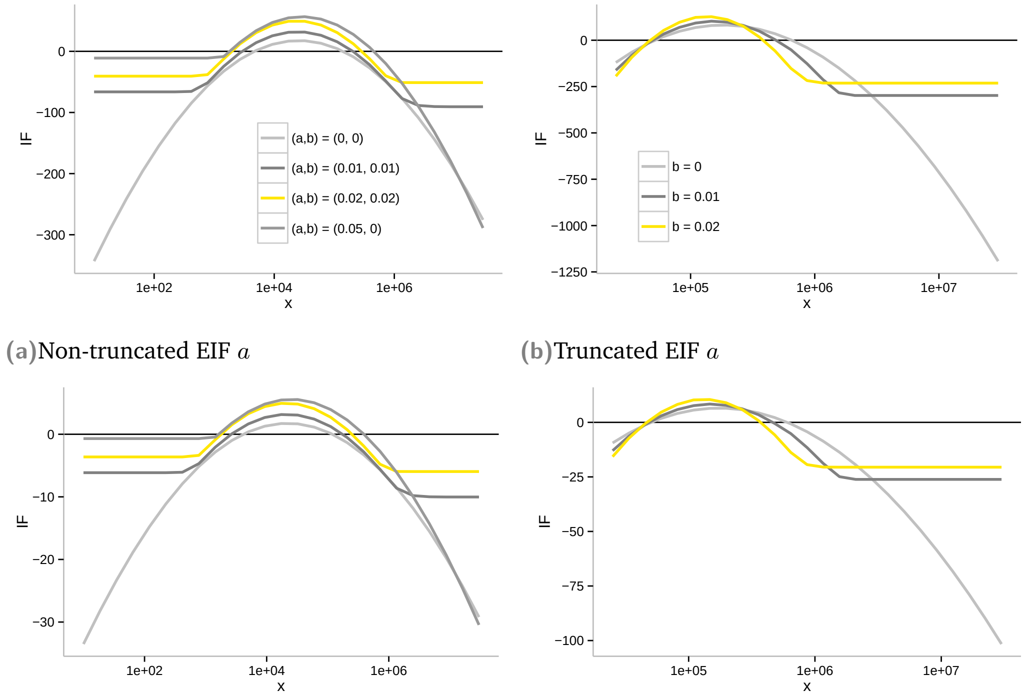
If we choose $a = b = 0$ we get the ordinary method of moments estimators, (which for the log-gamma distribution do *not* coincide with the MLE). By choosing $a > 0$ and $b = 0$, we are again able to trim only observations contained in the left tail, therefore directly gaining robustness against small losses. For the choice $a = 0.05$ and $b = 0$, we find the following approximate IFs and Δ-VaR measures.



(a)EIF a (b)EIF b (c)EIF b

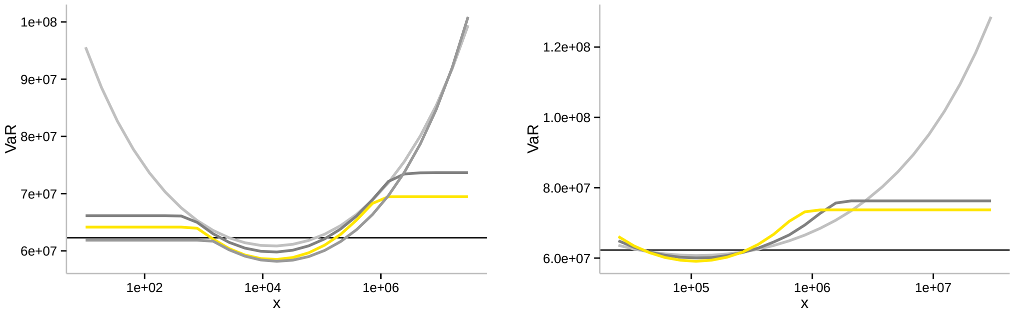
Fig. 5.9.: MTM approximated IFs and Δ-VaR for log-gamma severity distribution $\mathcal{LG}(34.5, 3.5)$, with trimming $(a, b) = (0.05, 0)$.

For several other non-symmetric and symmetric values of (a, b) the approximated MTM IFs and corresponding Δ-VaR approximations can be found in Fig. 5.10a, Fig.5.10c and Fig.5.11a respectively. For truncated loss data, we approximate the MTM IFs by computing the average EIFs for $K = 1,000$ loss data samples from a left-truncated log-gamma distribution of size $n = 250$, with fixed threshold $H = 25,000$. For different values of the right trimming proportion b, the approximated IFs and corresponding Δ-VaR approximations can be found in Fig.5.10b, Fig.5.10d and Fig.5.11b.



(a)Non-truncated EIF a (b)Truncated EIF a

(c)Non-truncated EIF b (d)Truncated EIF b



(a)Non-truncated Δ-VaR (b)Truncated Δ-VaR

Fig. 5.11.: MTM approximated IFs and Δ-VaR for log-gamma severity distribution $\mathcal{LG}(34.5, 3.5)$.

Relative efficiency for non-truncated and truncated loss data To compute the relative efficiency of the MTM estimators with respect to the MLE, we perform the same calculations as in the previous section for $K = 10,000$ non-truncated loss data samples of size $n = 1000$ from the log-gamma distribution. In Fig.5.12 below, we give the surface plot of the average results with trimming proportions $0 \leq a \leq 0.25$ and $0 \leq b \leq 0.25$.

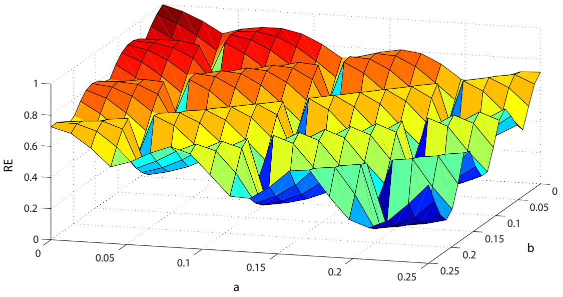


Fig. 5.12.: RE of the MTM estimators with respect to the MLE for non-truncated loss data from $\mathcal{LG}(34.5, 3.5)$.

We observe similar behavior as in Fig.5.8, the relative efficiency of the MTM estimators with respect to the MLE is seen to decrease symmetrically as the trimming proportions (a, b) increase, according to a *wave*-like surface.

In Table 5.5 below, we have computed the RE of the truncated MTM estimators with respect to the CML estimators for different values of the threshold H and the trimming proportion b.

		Trimming proportion				
		$b = 0$	$b = 0.01$	$b = 0.02$	$b = 0.03$	$b = 0.05$
RE	$H = 10,000$	1	0.949	0.893	0.864	0.765
	$H = 25,000$	1	0.909	0.839	0.779	0.654
	$H = 50,000$	1	0.834	0.725	0.675	0.526
	$H = 100,000$	1	0.763	0.679	0.600	0.508

Tab. 5.5.: RE of the MTM estimators with respect to the CML estimators for truncated loss data from $\mathcal{LG}(34.5, 3.5)$.

5.2.5 Concluding remarks

In this section we have shown how to compute parameter estimators via the method of trimmed moments (MTM) for both non-truncated and truncated loss data from the lognormal and log-gamma severity distribution. In order to cope with truncated loss data, we have extended the original MTM estimation procedure by proposing an iterative method. It is seen that, in general, the MTM estimation procedure is easier to understand and implement (e.g. compare the implementation in Section 5.2.3 and the implementation in Appendix C.1). Furthermore, for the considered severity distributions, the MTM estimation procedure is computationally much faster than the OBRE procedure, both for non-truncated and truncated loss data. For the considered trimming proportions, the computed RE of the MTM estimators tend to keep a level of efficiency similar to the computed RE of the OBRE. Since the OBRE are supposed to achieve an *optimal* level of efficiency, we expect the MTM estimators to be less robust than the OBRE for similar levels of efficiency. The trade-off between efficiency and robustness for the OBRE and the MTM estimators is assessed in the simulation studies in the following chapter.

Remark 5.0.7. As noted before, we chose to skip the derivation of the MTM estimators for the Generalized Pareto distribution, (since we may observe infinite moments when fitting the GPD). In particular, for the choice $\xi = 0.65$, the GPD has infinite second and higher moments. In Brazauskas and Kleefeld [5] it is shown that, under these conditions, it is still possible to derive the MTM estimators for non-truncated loss data. This is done by choosing two different sets of trimming proportions $(a_1, b_1) \neq (a_2, b_2)$ and considering only first order moment functions $h_1(x) = x$ and $h_2(x) = x$. The resulting system of equations then has a (unique) solution, since for $\xi < 1$ the first moments of the GPD *do* exist. (We note that in this case we can not compute the standard method of moment estimators, since we need the trimming proportions to be different).

Simulation Study 6

6.1 Summary of the material so far

The main goal of the simulation study is to assess the behavior of the VaR-measures under the different estimation procedures that we have examined throughout the thesis. In order to do this, it is constructive to give a short summary of the material and the introduced estimation methods so far.

In Chapter 3, we established the results needed to produce a *good* classical model. The VaR measures are computed under the LDA framework, in which we estimate the parameters of the severity distribution using maximum likelihood estimation for non-truncated loss data and using the constrained maximum likelihood approach for truncated loss data. In Section 3.7 it is shown that under a correctly specified severity distribution and independent and identically distributed (IID) loss data, we are able to estimate the VaR measures quite accurately.

In Chapter 4, we argued that the VaR measures are highly sensitive to minor contamination of the loss data. In order to ensure stable capital charges, we introduced the robust statistics framework. The idea is to sacrifice some efficiency at the exact severity distribution, in order to gain robustness against minor deviations of the model severity distribution. By computing the influence function of the MLE and the CML estimators, we have shown that the VaR measures are very sensitive to minor contamination due to both extreme large and small losses. To mitigate the impact of small losses on the VaR, we proposed fitting a mixture of severity distributions to the loss data. The main drawback of this method, is that we have to split the -usually already scarce- loss data even further into separate a body and a tail region. We found that, although the sensitivity of the VaR to small losses is reduced to a minimum, the VaR becomes much more sensitive to extreme large losses. Since it is found that the truncation of the loss data also mitigates the impact of small loss events on the VaR measures, we argue that it might be sufficient to fit a single severity distribution to the loss data using robust estimation techniques. In Chapter 5 we studied two robust estimation methods as a more straightforward procedure to gain stability in the VaR measures under contamination of the loss data. First, we studied the optimal bias robust estimators, which are designed to achieve highest possible efficiency, while remaining bias robust. Due to their computational complexity the optimal bias robust estimators may be difficult to implement in practice. This is why, secondly, we studied the method of trimmed moment estimators, which are more straightforward both analytically and computationally.

At this point, we wish to compare the behavior of the VaR measures under all of the introduced estimation methods. In the following section we give a description of the performed simulation studies.

We make the following important remark: under an ideal scenario, we would also assess the behavior of the VaR measures for *real* historical operational loss data. During the internship at EY, we have had access to the historical operational loss database of a large internationally active bank and the operational risk AMA models of several other large internationally active banks. However, due to confidentiality issues, we are unable to make any references to the AMA models or include any calculations performed with real operational loss data. This is why all experiments in this chapter are performed using simulated operational loss data only.

6.2 Description of the simulation study

In order to draw conclusions from the results of the simulation experiments, we must have a clear idea of what we want to achieve. By definition the robust estimation procedures follow the properties according to the robust statistics framework in Section 4.2. For the VaR measures to be applicable in practice, they should inherit the same properties, that is:

- *Efficiency*: VaR should have reasonably good efficiency at the assumed model, in the sense that the VaR is close to the *true* capital charge when the loss data sample follows the assumed model exactly.

- *Stability*: VaR should be robust in the sense that small deviations from the model (in the form of contamination due to a small fraction of outliers) only have minor impact on the VaR measure.

- *Breakdown*: Any larger deviations from the model, should not cause the VaR to become arbitrarily large.

In order to examine these properties we again assume the gross error model according to Eq.(4.7). We consider the following class of ϵ-contaminated distributions

$$F_\epsilon = (1-\epsilon)F + \epsilon G \tag{6.1}$$

with $\epsilon \geq 0$ is the level of contamination, F the assumed *true* model distribution and G a contaminating distribution defined below. We specify three types of ϵ-contamination:

1. *Random ϵ-contamination*: in this case, G randomly draws gross errors from the interval $[H+1, 3\times 10^7]$ on a logarithmic scale, with truncation threshold $H \geq 0$. Mathematically, G can be written as

   $$G = \exp(\mathrm{U}[\log(H+1), \log(3\times 10^7)]) \tag{6.2}$$

   Where $\mathrm{U}[a,b]$ denotes the uniform distribution on the interval $[a,b]$.
   Informally, we have replaced a fraction ϵ of the loss data by random points on a logarithmic scale.

2. *Left-tail ϵ-contamination*: in this case, G randomly draws gross errors from the interval $[H+1, H+1000]$. The contaminating distribution G can be written as

   $$G = \mathrm{U}[H+1, H+1000] \tag{6.3}$$

   Here, we replace a fraction ϵ of the loss data by random extreme small losses (left-tail gross errors), where we choose the upper bound of a small loss to be 1000 units above the truncation threshold H.

3. *Right-tail ϵ-contamination*: in this case G randomly draws gross errors from the interval $[1\times 10^7, 3\times 10^7]$ on a logarithmic scale. The contaminating distribution G can be written as

   $$G = \exp(\mathrm{U}[\log(1\times 10^7), \log(3\times 10^7)]) \tag{6.4}$$

   Here, we replace a fraction ϵ of the data by random extreme large losses (right-tail gross errors) on a logarithmic scale, where we have chosen the lower bound of a large loss to be 1×10^7 units.

We examine levels of ϵ-contamination, $0 \leq \epsilon \leq 0.05$. We argue that at these levels of contamination it is still reasonable to fit a particular parametric distribution, since for $\epsilon = 0.05$ we still have 95% of the loss data following the model distribution exactly. (In the performed simulation studies, the fits of the severity distributions to the loss data are also checked by means of QQ-plots).

The choice $\epsilon = 0$ allows us to assess the *efficiency* property, since the data sample now follows the assumed model distribution exactly, i.e. $F_\epsilon = F$. We note that this is similar to the calculations of the relative efficiency in the previous chapter, except that the calculations are extended to the VaR measures instead of the parameter estimates.

If we choose $\epsilon > 0$ we are able to examine the *stability* and *breakdown* property. Furthermore, we are interested in the small sample behavior, and the behavior under different levels of the truncation threshold H. Before we give the results of the performed simulation studies, we give a description of the definite model that is applied to estimate the VaR measures. We extend the model of Section 3.7 by combining the results presented in the previous chapters.

6.3 Description final model

Suppose we observe a data sample $x_1, x_2, \ldots, x_n$ of operational losses, corresponding to data collected over a number of m years from a single unit of measure. The goal is to estimate VaR measures of the aggregated one-year loss distribution (compound distribution).

Below we describe the estimation process informally, this is presented in a systematic way in Fig.6.1.

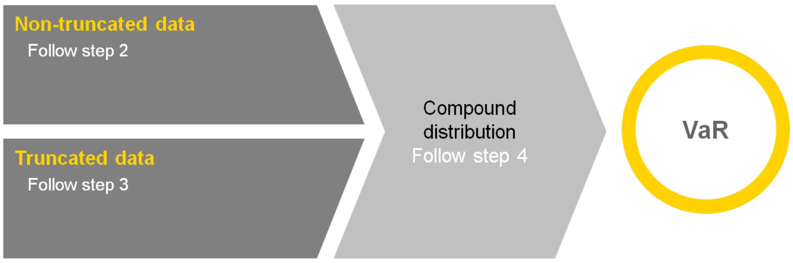


Fig. 6.1.: Overview steps in the model

1. We check whether the data is truncated or not, if the data is truncated (from below) we assume that the truncation level H is known. If the data is non-truncated go to step 2, else go to step 3.

2. The loss data is non-truncated. If we would fit a single heavy-tailed severity distribution to the loss data, we know that it is highly sensitive to small losses (left-tail gross errors). Therefore we wish to fit a body- and tail-distribution separately according to the methods described in Section 4.7

   a) Fix the body-tail threshold L according to some pre-specified measure

   b) Fit a right-truncated exponential distribution to the body region $[0, L)$ using maximum likelihood estimation as described in Section 4.7.2.

   c) Fit a left-truncated heavy-tailed distribution to the tail region $[L, \infty)$ using robust estimation methods (OBRE or MTM), in order to reduce the sensitivity of the VaR measures to extreme large losses (right-tail gross errors). The heavy-tailed distributions to consider are: the lognormal distribution, for the OBRE see Section 5.1.4 and for the MTM procedure see Section 5.2.3; the log-gamma distribution, for the OBRE see Section 5.1.5 and for the MTM procedure see Section 5.2.4; the Generalized Pareto distribution, for the OBRE see Section 5.1.5

d) We construct the overall severity distribution by combining the body and tail distribution under the estimated parameters, according to the method described in Section 4.7.1.

e) Fit a non-truncated Poisson distribution to the annual loss frequencies using ordinary maximum likelihood estimation.

Now go to step 4.

3. The loss data is truncated at recording threshold H. In this case it might be suitable to fit a single heavy-tailed severity distribution to the loss data, since the sensitivity of the capital charge to small losses is already mitigated by the truncation of the loss data. If we still wish to fit a separate body and tail distribution, we can follow the same procedure as in step (2), only now the body region is given by $[H, L)$. Otherwise,

   a) Fit a left-truncated heavy-tailed distribution to the full region $[H, \infty)$ using robust estimation methods (OBRE or MTM). The heavy-tailed distributions we may consider are again: the lognormal distribution, for the OBRE see Section 5.1.4 and for the MTM procedure see Section 5.2.3; the log-gamma distribution, for the OBRE see Section 5.1.5 and for the MTM procedure see Section 5.2.4; the Generalized Pareto distribution, for the OBRE see Section 5.1.5. The overall severity distribution is then given by the single left-truncated heavy-tailed severity distribution under the estimated parameters.

   b) Fit a left-truncated Poisson distribution to the one-year frequencies using maximum likelihood estimation as described in Section 3.6.1.

   Now go to step 4.

4. Given the estimated severity and frequency distribution, by the LDA we construct the aggregated annual loss distribution (compound distribution) and corresponding VaR measures using the FFT method, as described in Section 3.3.

6.4 Simulation study

As described in Section 6.2, we are interested in the efficiency, stability and breakdown properties of the VaR measures, which can be assessed by choosing different contamination levels ϵ. The simulation studies are performed for the three usual types of heavy-tailed severity distributions: the lognormal, log-gamma and Generalized Pareto distribution.

6.4.1 Simulation study for lognormal severity distribution

Non-truncated severity loss data According to the model in Section 6.3, for non-truncated severity loss data, we wish to fit a separate body- and tail-distribution to the loss data. Since, in practice it might be difficult to fit two separate distributions, due to scarcity of the operational loss data, we also consider fitting a single heavy-tailed severity distribution using robust estimation methods.
We perform the following simulation experiment: we draw $K = 250$ non-truncated loss data samples from the lognormal distribution $\mathcal{LN}(10.95, 1.75)$ of size $n = 500$. Next, we contaminate each loss data sample under different levels of random ϵ-contamination, with $\epsilon = \{0,\ 0.025,\ 0.05\}$; left-tail ϵ-contamination, with $\epsilon = 0.025$ and right-tail ϵ-contamination, with $\epsilon = 0.025$. We model the severity distribution according to the following methods:

1. A single non-truncated lognormal severity distribution, where the parameters are estimated using maximum likelihood estimation (MLE). We note that this is the approach applied under the classical model and corresponds to the results in Section 3.7.

2. A single non-truncated lognormal severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 3$ and $c = 2$.

3. A single non-truncated lognormal severity distribution, where the parameters are estimated using the method of trimmed moments, with (symmetric) trimming proportions $(a, b) = (0.02, 0.02)$ and $(a, b) = (0.05, 0.05)$.

4. A mixture of severity distributions, with a right-truncated exponential distribution for the body region $[0, 25000)$, where the paramaters are estimated using MLE (according to Section 4.7.2). And with a left-truncated lognormal distribution for the tail region $[25000, \infty)$, where the parameters are estimated using OBRE, with tuning $c = 3$ and $c = 2$.

5. A mixture of severity distributions, with a right-truncated exponential distribution for the body region $[0, 25000)$, where the parameters are estimated using MLE. And with a left-truncated lognormal distribution for the tail region $[25000, \infty)$, where the parameters are estimated using MTM, with trimming proportion $b = 0.03$ and $b = 0.05$.

We assume that the annual loss frequencies follow a Poisson distribution $\mathrm{Pois}(25)$, (which corresponds to an average of 25 loss observations per year). For the K simulated loss data samples, we estimate the VaR measures according to the FFT method and compute the average results. The bias is calculated by dividing the average results by the capital charge under the true parameters, which is given by $\text{VaR}_{0.999} = 63,945,425$ (this can also be found in Table 3.3). The average VaR measures and corresponding biases can be found in Table 6.1.

Truncated severity loss data We perform the same simulation experiment for truncated loss data from the lognormal distribution. In the case of truncated loss data, we argue that it is enough to fit a single heavy-tailed severity distribution to the loss data, since the impact of small losses on the VaR is already mitigated by the truncation of the loss data. We perform the following simulation experiment: we draw $K = 250$ truncated loss data samples of size $n = 500$ from the lognormal distribution $\mathcal{LN}(10.95, 1.75)$, truncated at $H = 25,000$ and $H = 50,000$. We contaminate each loss data sample in the same way as before, i.e. random ϵ-contamination, with $\epsilon = \{0,\ 0.025,\ 0.05\}$; left-tail ϵ-contamination, with $\epsilon = 0.025$ and right-tail ϵ-contamination, with $\epsilon = 0.025$. We model the severity distribution according to the following methods:

1. A left-truncated lognormal severity distribution, where the parameters are estimated using the constrained maximum likelihood approach (CML). This is the approach under the classical model, corresponding to the results in Section 3.7.

2. A left-truncated lognormal severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 3$ and $c = 2$.

3. A left-truncated lognormal severity distribution, where the parameters are estimated using the method of trimmed moments (MTM), with trimming proportion $b = 0.02$ and $b = 0.05$.

We model the annual loss frequencies by a Poisson distribution $\mathrm{Pois}(\lambda = 25)$. The VaR measure under the true parameters is again given by $\text{VaR}_{0.999} = 63,945,425$ and the bias is calculated in the same way as before. For the results, see Table 6.2.

		Random ϵ-contamination						Left-tail		Right-tail	
		$\epsilon = 0$		$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.025$		$\epsilon = 0.025$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
MLE		65,516,330	1.02	122,919,885	1.92	214,884,450	3.36	212,745,060	3.34	156,034,175	2.44
OBRE	$c = 3$	65,569,130	1.03	81,841,705	1.28	104,116,485	1.63	97,339,880	1.52	100,494,955	1.57
	$c = 2$	64,018,295	1.00	76,715,760	1.20	93,088,270	1.46	91,640,835	1.43	88,212,520	1.38
MTM	$(a, b) = (.02, .02)$	65,773,675	1.03	81,911,555	1.28	107,948,555	1.68	99,486,816	1.56	127,908,429	2.00
	$(a, b) = (.05, .05)$	66,193,050	1.04	78,526,085	1.23	95,254,060	1.48	83,442,183	1.30	98,814,903	1.55
Mix. OBRE	$c = 3$	46,649,900	0.73	54,271,305	0.85	63,946,740	1.00	47,686,100	0.75	122,330,230	1.91
	$c = 2$	40,233,710	0.63	46,036,815	0.72	53,585,070	0.84	40,717,985	0.64	88,080,410	1.38
Mix. MTM	$b = .03$	47,461,315	0.74	55,154,880	0.86	64,958,905	1.02	48,455,385	0.76	126,189,415	1.97
	$b = .05$	48,497,570	0.76	55,437,525	0.87	64,276,575	1.00	48,807,935	0.76	102,319,635	1.60

Tab. 6.1.: Average VaR measures and bias for non-truncated $\mathcal{LN}(10.95, 1.75)$ loss data of size $n = 500$, under different sources of ϵ-contamination.

		Random ϵ-contamination						Left-tail		Right-tail	
$H = 25,000$		$\epsilon = 0$		$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.025$		$\epsilon = 0.025$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		65,092,555	1.02	97,016,040	1.52	144,271,545	2.26	84,637,740	1.32	272,657,495	4.26
OBRE	$c = 3$	65,400,555	1.02	83,590,540	1.31	110,431,090	1.73	76,992,190	1.20	122,404,810	1.91
	$c = 2$	56,027,345	0.88	69,805,010	1.09	88,337,150	1.38	75,476,720	1.18	95,211,930	1.49
MTM	$b = 0.02$	65,962,050	1.03	86,246,655	1.35	117,848,720	1.84	80,723,555	1.26	143,850,465	2.49
	$b = 0.05$	67,397,385	1.05	83,638,280	1.31	106,620,415	1.67	78,894,805	1.23	110,701,305	1.73
$H = 50,000$		$\epsilon = 0$		$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.025$		$\epsilon = 0.025$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		64,959,620	1.02	96,498,105	1.51	141,985,745	2.22	85,099,080	1.33	236,422,230	3.70
OBRE	$c = 3$	64,388,060	1.00	83,518,600	1.31	111,287,770	1.74	73,523,010	1.15	120,680,395	1.89
	$c = 2$	48,842,035	0.76	60,415,190	0.94	75,145,785	1.18	71,746,565	1.12	79,651,385	1.26
MTM	$b = 0.02$	66,149,270	1.03	86,381,350	1.35	116,522,505	1.82	78,582,185	1.23	128,629,985	2.01
	$b = 0.05$	69,125,100	1.08	85,634,615	1.34	107,772,775	1.69	75,434,700	1.18	108,899,560	1.70

Tab. 6.2.: Average VaR measures and bias for truncated $\mathcal{LN}(10.95, 1.75)$ loss data of size $n = 500$, truncated at $H = 25,000$ and $H = 50,000$ under different sources of ϵ-contamination.

Conclusion results Table 6.1 and Table 6.2 Let us examine the results in Table 6.1 more closely. First of all, we observe that the amount of efficiency we lose at the exact model ($\epsilon = 0$) under the OBRE and MTM procedure is negligible in comparison to the MLE. We conclude that the estimated VaR measures under the OBRE and MTM procedure satisfy the *efficiency* property, since the estimated VaR remains close to the true capital charge, when the loss data samples follow the lognormal distribution exactly. In other words, if there is no contamination of the loss data and if we choose to apply the robust estimation methods (OBRE and MTM) instead of the MLE, the resulting VaR measures will not be significantly different.

If we look at the bias of the OBRE and MTM with mixed severities at the exact model ($\epsilon = 0$), it is seen that the VaR measures are in general underestimated. Thus, we argue that applying robust estimation methods in combination with mixed severity distributions, might sacrifice too much efficiency in order to gain robustness. This can be solved by choosing a lower tuning parameter c for the OBRE or a lower trimming proportion b for the MTM. On the other hand, it is seen that the VaR measures under mixed severity distributions remain very stable. For instance, the VaR measures under the OBRE, with mixed severities and tuning $c = 2$ increase by only $13,351,360$ as the loss data becomes highly contaminated ($\epsilon = 0.05$). Compare this to an increase in the VaR measures under the MLE of $148,368,120$, which is larger by more than a factor 10.

The results for left- and right-tail contamination seem to correspond well with our initial expectations. The bias for both left- and right-tail contamination is highly reduced under the OBRE and MTM, but it is seen that under left-tail contamination the VaR may still increase by upto 56 percent. The bias caused by left-tail contamination under the OBRE and MTM with mixed severities is reduced to a minimum. Furthermore, since we apply robust estimation methods, the VaR measures do not become highly sensitive to right-tail contamination, as was found in Section 4.7.

For the results in Table 6.2 we can draw similar conclusions. Under the exact model distribution ($\epsilon = 0$), both the OBRE and MTM keep a level of efficiency that is close to the CML estimators. Only for the OBRE with tuning $c = 2$ the VaR measures are somewhat underestimated.

As expected, the truncation of the loss data reduces the sensitivity of the VaR measures to left-tail contamination. In this sense, we can argue that we benefit from the fact that the loss data is truncated, making the capital charges more stable in general.

Overall, the results seem to be quite similar for the different truncation thresholds $H = 25,000$ and $H = 50,000$. We note that in the simulation study the size of the loss data samples are equal for both truncation thresholds. In practice, if the truncation threshold increases, the available loss data shrinks, thereby decreasing the stability of the VaR measures.

In conclusion, we argue that the robust estimation methods clearly outperform the classical estimation methods. The estimated VaR measures stay close to the *true* VaR when there is no contamination of the loss data and they remain much more stable when the loss data becomes increasingly contaminated, both for non-truncated loss data and truncated loss data.

In the following section we perform the same simulation experiments with loss data from the log-gamma distribution.

6.4.2 Simulation study for log-gamma severity distribution

Non-truncated severity loss data The performed simulation experiment is similar to the previous section. The experiment can is summarized as follows: we draw $K = 250$ non-truncated loss data samples from the log-gamma distribution $\mathcal{L}G(34.5, 3.5)$ of size $n = 500$. Next, we contaminate each sample with random ϵ-contamination, where $\epsilon = \{0,\ 0.01,\ 0.025\}$; with left-tail ϵ-contamination, where $\epsilon = 0.01$ and with right-tail ϵ-contamination, where $\epsilon = 0.01$. The severity distribution is modeled according to the following methods:

1. A single non-truncated log-gamma severity distribution, where the parameters are estimated using maximum likelihood estimation (MLE). We note that this is the approach applied under the classical model and corresponds to the results in Section 3.7.

2. A single non-truncated log-gamma severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 3$ and $c = 2.5$.

3. A single non-truncated log-gamma severity distribution, where the parameters are estimated using the method of trimmed moments (MTM), with (symmetric) trimming proportions $(a, b) = (0.01, 0.01)$ and $(a, b) = (0.03, 0.03)$.

4. A mixture of severity distributions, with a right-truncated exponential distribution for the body region $[0, 25000)$, where the parameters are estimated using MLE and with a left-truncated log-gamma distribution for the tail region $[25000, \infty)$, where the parameters are estimated using OBRE, with tuning $c = 3$ and $c = 2.5$.

5. A mixture of severity distributions, with a right-truncated exponential distribution for the body region $[0, 25000)$, where the parameters are estimated using MLE and with a left-truncated log-gamma distribution for the tail region $[25000, \infty)$, where the parameters are estimated using MTM estimation, with trimming proportion $b = 0.01$ and $b = 0.03$.

The frequency distribution is modeled by a Poisson distribution $\mathrm{Pois}(25)$. The VaR measure under the true parameters is given by VaR$_{0.999} = 62, 290, 900$ (this can also be found in Table 3.3). The average VaR measures and biases are calculated in the same way as before, see Table 6.3 for the results.

Truncated severity loss data For truncated loss data from the log-gamma distribution, similar to the case of a truncated lognormal distribution, we fit a single heavy-tailed severity distribution to the loss data, arguing that the sensitivity to small losses is already mitigated by the truncation of the data. We perform the following simulation experiment: we draw $K = 250$ truncated loss data samples from the log-gamma distribution $\mathcal{L}G(34.5, 3.5)$ of size $n = 500$, with truncation thresholds $H = 10, 000$ and $H = 25, 000$. We contaminate each loss data sample with random ϵ-contamination, where $\epsilon = \{0,\ 0.025,\ 0.05\}$; and with left- and right-tail ϵ-contamination, where $\epsilon = 0.025$. We model the severity distribution according to the same methods as in the case of the lognormal distribution:

1. A left-truncated log-gamma severity distribution, where the parameters are estimated using the constrained maximum likelihood approach (CML). This is the approach under the classical model, corresponding to the results in Section 3.7.

2. A left-truncated log-gamma severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 3$ and $c = 2.5$.

3. A left-truncated log-gamma severity distribution, where the parameters are estimated using the method of trimmed moments (MTM), with trimming proportion $b = 0.03$ and $b = 0.05$.

We model the annual loss frequencies by a Poisson distribution $\mathrm{Pois}(25)$. The VaR measure under the true parameters is again given by VaR$_{0.999} = 62, 290, 900$. The computed results of the average VaR measures and their corresponding biases can be found in Table 6.4.

		Random ϵ-contamination						Left-tail		Right-tail	
		$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.025$		$\epsilon = 0.01$		$\epsilon = 0.01$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
MLE		64,114,710	1.03	210,444,025	3.38	688,214,835	11.05	597,019,390	9.58	115,524,255	1.85
OBRE	$c = 3$	64,493,165	1.04	74,639,235	1.20	93,404,025	1.50	78,245,750	1.26	81,945,930	1.32
	$c = 2.5$	68,127,620	1.09	75,260,590	1.21	86,041,340	1.38	76,109,990	1.22	75,323,875	1.21
MTM	$(a, b) = (.01, .01)$	64,413,415	1.03	75,297,365	1.21	96,581,430	1.55	77,235,895	1.24	102,887,565	1.65
	$(a, b) = (.03, .03)$	64,811,725	1.04	73,216,605	1.18	87,732,260	1.41	73,746,420	1.18	85,615,860	1.37
Mix. OBRE	$c = 3$	43,351,495	0.69	48,825,700	0.78	60,325,100	0.97	43,226,425	0.69	83,129,035	1.33
	$c = 2.5$	40,469,935	0.65	46,934,525	0.75	56,673,100	0.91	41,028,020	0.66	73,643,350	1.18
Mix. MTM	$b = .01$	46,308,130	0.74	53,265,795	0.86	64,759,805	1.04	46,222,715	0.74	82,983,120	1.33
	$b = .03$	49,228,740	0.79	55,244,530	0.89	66,606,925	1.07	49,839,460	0.80	78,274,020	1.26

Tab. 6.3.: Average VaR measures and bias for non-truncated $\mathcal{LG}(34.5, 3.5)$ loss data of size $n = 500$, under different sources of ϵ-contamination.

		Random ϵ-contamination						Left-tail		Right-tail	
$H = 10,000$		$\epsilon = 0$		$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.025$		$\epsilon = 0.025$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		66,061,270	1.06	144,266,650	2.32	310,701,655	4.99	89,369,775	1.40	771,456,400	12.38
OBRE	$c = 3$	66,178,420	1.06	103,519,185	1.66	173,127,735	2.78	79,455,145	1.24	180,081,055	2.89
	$c = 2.5$	66,149,270	1.06	99,759,880	1.60	153,486,575	2.46	74,497,885	1.17	172,398,875	2.77
MTM	$b = 0.02$	68,614,590	1.10	105,698,615	1.70	175,078,860	2.81	83,051,870	1.30	177,916,915	2.86
	$b = 0.05$	69,771,075	1.12	101,381,335	1.63	154,348,920	2.48	81,514,400	1.27	147,948,515	2.34
$H = 25,000$		$\epsilon = 0$		$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.025$		$\epsilon = 0.025$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		62,281,835	1.00	126,143,490	2.03	234,888,060	3.77	89,267,970	1.40	403,239,540	6.47
OBRE	$c = 3$	64,647,715	1.04	100,383,195	1.61	162,417,970	2.61	71,552,250	1.12	160,274,015	2.58
	$c = 2.5$	59,173,125	0.95	92,971,230	1.49	134282,555	2.15	67,467,730	1.06	122,668,865	1.97
MTM	$b = 0.02$	66,980,925	1.08	99,165,220	1.59	151,728,555	2.43	78,992,705	1.24	140,698,800	2.26
	$b = 0.05$	71,824,060	1.15	99,974,545	1.60	141,525,010	2.27	73,765,945	1.14	124,156,725	1.99

Tab. 6.4.: Average VaR measures and bias for truncated $\mathcal{LG}(34.5, 3.5)$ loss data of size $n = 500$, truncated at $H = 10,000$ and $H = 25,000$ under different sources of ϵ-contamination.

Conclusion results Table 6.3 and Table 6.4 First, we examine the results in Table 6.3. We note that the estimated VaR measures under the MLE are much less stable as in the previous section, (with loss data from the lognormal distribution). As the loss data becomes increasingly contaminated the VaR under the MLE increases by upto a factor 11 ($\epsilon = 0.025$). It is seen that this spectacular increase can be attributed almost entirely to the contamination in the left-tail, since the VaR increases to $597,019,390$ under left-tail contamination, whereas it only increases to $115,524,255$ under right-tail contamination. Under these circumstances in practice, it is clear that one should not blindly apply classical estimation methods, such as maximum likelihood.

Let us now focus on the estimated VaR measures under the robust estimation methods. Similar to the previous section, both the OBRE and MTM keep a level of efficiency that is close to the MLE, when the loss data follows the log-gamma distribution exactly ($\epsilon = 0$). Only for the OBRE with tuning $c = 2.5$ the VaR measures are somewhat overestimated.

The OBRE and MTM with mixed severities generally underestimate the VaR at the exact model ($\epsilon = 0$). However, they also remain the most stable when the loss data becomes increasingly contaminated. This should come as no surprise, since the instability in the VaR measures is mainly caused by the left-tail contamination and the mixed severities are designed to capture exactly that.

Overall, the results in Table 6.3 give a clear view of the advantages of the robust estimation techniques over the classical estimation methods: although we have to sacrifice some efficiency at the exact model, we are able to produce *far* more stable results when the loss data becomes increasingly contaminated.

For the results in Table 6.4 we can see that the results are somewhat different. First of all, the induced bias under the CML estimators is heavily reduced in comparison to the MLE in Table 6.3. We argue that this is because of the fact that the estimated VaR measures are much less sensitive to left-tail contamination (due to the truncation of the loss data). However, the VaR measures have become much more sensitive to contamination in the right tail; for $H = 10,000$, we see that the VaR measures increase to $771,456,400$ under right-tail contamination. The robust estimation methods are seen to produce much more stable results. Especially in the case of right-tail contamination, the robust methods are seen to outperform the CML estimators by far.

In conclusion, both for non-truncated and truncated loss data the robust estimation methods are able to produce VaR measures close to the *true* VaR when the loss data follows the log-gamma distribution exactly and they produce much more stable VaR measures when the loss data becomes increasingly contaminated than the classical estimation methods are able to.

In the following section we perform the simulation experiments for loss data from the Generalized Pareto distribution.

6.4.3 Simulation study for Generalized Pareto severity distribution

Non-truncated and truncated severity loss data Finally, we consider loss data from the Generalized Pareto distribution. For non-truncated loss data, we perform the same simulation experiment as for the lognormal and log-gamma severity distribution: we draw $K = 250$ non-truncated loss data samples from the Generalized Pareto distribution $\mathrm{GPD}(0.65, 57500)$ of size $n = 500$. We contaminate each sample with random ϵ-contamination, where $\epsilon = \{0,\ 0.01,\ 0.025\}$ and with left- and right-tail contamination, where $\epsilon = 0.01$. For the Generalized Pareto distribution, we chose to skip the derivation of the MTM estimators in the previous chapter. Therefore, we will also not consider it here. The severity distribution is modeled according to the following methods:

1. A single non-truncated GPD severity distribution, where the parameters are estimated using maximum likelihood estimation (MLE). This is the approach applied under the classical model and corresponds to the results in Section 3.7.

2. A single non-truncated GPD severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 5$, $c = 3$ and $c = 2$.

3. A mixture of severity distributions, with a right-truncated exponential for the body region $[0, 25000)$, where the parameters are estimated using MLE and with a left-truncated GPD for the tail region $[25000, \infty)$, where the parameters are estimated using OBRE, with tuning $c = 5$, $c = 3$ and $c = 2$.

For truncated loss data from the Generalized Pareto distribution, we draw $K = 250$ truncated loss data samples from the distribution $\mathrm{GPD}(0.65, 57500)$ of size $n = 500$, with truncation thresholds $H = 25,000$ and $H = 50,000$. We contaminate each loss data sample in the same way as for non-truncated loss data from the GPD. We model the severity distribution according to:

1. A left-truncated GPD severity distribution, where the parameters are estimated using the constrained maximum likelihood approach (CML). This is the approach under the classical model, corresponding to the results in Section 3.7.

2. A left-truncated GPD severity distribution, where the parameters are estimated using optimal bias robust estimation (OBRE), with tuning parameter $c = 5$, $c = 3$ and $c = 2$.

For both the non-truncated and truncated loss data, the annual loss frequencies are modeled by a Poisson distribution $\mathrm{Pois}(25)$. The VaR measure under the true parameters is given by $\mathrm{VaR}_{0.999} = 67,916,625$. The average results of the K estimated VaR measures and their corresponding biases can be found in Table 6.5.

		Random ϵ-contamination						Left-tail		Right-tail	
$H = 0$		$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.025$		$\epsilon = 0.01$		$\epsilon = 0.01$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		73,686,635	1.08	98,632,380	1.45	148,105,760	2.18	91,869,250	1.35	304,995,460	4.49
OBRE	$c = 5$	73,816,215	1.09	94,547,915	1.39	138,569,640	2.04	84,855,980	1.25	214,506,160	3.16
	$c = 3$	74,305,110	1.09	92,457,475	1.36	130,436,185	1.92	82,918,715	1.22	156,668,820	2.31
	$c = 2$	75,467,205	1.11	93,019,355	1.37	130,426,120	1.92	81,792,865	1.20	122,655,445	1.81
Mix. OBRE	$c = 5$	62,541,215	0.92	76,801,120	1.13	98,677,260	1.45	64,850,665	0.95	175,168,015	2.58
	$c = 3$	64,073,735	0.94	74,621,800	1.10	90,055,295	1.33	64,113,610	0.94	124,353,020	1.83
	$c = 2$	62,532,305	0.92	73,531,260	1.08	84,928,085	1.25	64,794,345	0.95	107,588,525	1.58
$H = 25,000$		$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.025$		$\epsilon = 0.01$		$\epsilon = 0.01$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		71,626,170	1.05	122,682,450	1.81	239,095,230	3.52	94,216,375	1.39	279,203,265	4.11
OBRE	$c = 5$	73,262,475	1.08	104,624,025	1.54	179,869,580	2.65	84,213,580	1.24	210,694,000	3.10
	$c = 3$	76,961,775	1.13	100,081,355	1.47	150,251,090	2.21	81,663,340	1.20	155,596,815	2.29
	$c = 2$	80,129,500	1.18	99,005,830	1.46	138,450,950	2.04	79,963,400	1.18	125,460,445	1.85
$H = 50,000$		$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.025$		$\epsilon = 0.01$		$\epsilon = 0.01$	
		VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias	VaR	Bias
CML		70,224,880	1.03	129,226,680	1.90	264,021,450	3.89	91,938,495	1.35	293,993,975	4.33
OBRE	$c = 5$	70,791,545	1.04	107,284,045	1.58	195,008,825	2.87	87,829,995	1.29	151,717,775	2.23
	$c = 3$	70,748,370	1.04	95,931,440	1.41	147,037,110	2.16	82,407,105	1.21	116,195,530	1.71
	$c = 2$	67,180,190	0.99	86,492,340	1.27	121,826,540	1.79	78,862,465	1.16	98,988,285	1.46

Tab. 6.5.: Average VaR measures and bias for non-truncated and truncated $\mathrm{GPD}(0.65, 57500)$ loss data of size $n = 500$, truncated at $H = 0$, $H = 25,000$ and $H = 50,000$ under different sources of ϵ-contamination.

Conclusion results Table 6.5 Let us first examine the results in Table 6.5 for non-truncated loss data. Under the exact model distribution ($\epsilon = 0$), the OBRE keeps a level of efficiency close to the CML esimators. Since the CML estimators overestimate the *true* VaR themselves at the exact model (see also Remark 3.0.2), it is seen that the OBRE with mixed severities remain much closer to the *true* VaR as in the case of the lognormal and the log-gamma distribution.

It is found that the main cause of instability in the VaR measures is due to right-tail contamination, which corresponds to the results in the previous chapters, where it was shown that the GPD is much less sensitive to left-tail contamination than the lognormal and log-gamma distribution, but more sensitive to contamination in the right tail. Both the OBRE for a single fitted severity distribution and the OBRE with mixed severities are able to reduce the impact of right-tail contamination on the estimated VaR measures.

If we examine the results in Table 6.5 for truncated loss data, we conclude that the behavior of the VaR measures under contamination of the loss data is quite similar to the non-truncated case. It seems that the induced bias under the CML estimators and the OBRE becomes somewhat larger, as the truncation threshold H increases.

In conclusion, both for non-truncated and truncated loss data, the robust estimation methods are able to produce VaR measures close to the *true* VaR when the loss data follows the Generalized Pareto distribution exactly (sometimes they are seen to perform even better than the classical estimation methods, which are supposed to be *most* efficient). At the same time, they produce more stable VaR measures when the loss data becomes increasingly contaminated than the estimation methods under the classical model are able to.

Recalling the objectives stated in Section 6.2, using the introduced robust models, the estimated capital charges -for the lognormal, log-gamma and Generalized Pareto distribution- comply with the desired properties of reasonably high efficiency at the exact model and stability under minor contamination of the loss data.

7 Concluding remarks & Future Research

The goal of the thesis was to assess the behavior of the operational risk capital charge under contamination of the loss data, and to develop *robust* models in order to maintain stability in the capital charge, while keeping a high level of efficiency.
We summarize the key messages and results:

- It was shown that the FFT is an efficient method to calculate the aggregated loss distribution and the corresponding VaR measures. If it is used correctly, it outperforms standard methods (e.g. SLA and Monte Carlo simulation) in both accuracy and computation time. Since it is relatively new to the field of operational risk modeling, it is rarely used in practice. This may change in the near future, when the operational risk models used by the banks become more sophisticated. If the FFT method would not be available to us, we would have been unable to produce many of the simulation results. We argue that this may be a major reason there is little existing literature on the instability of the operational risk capital charges. Using the Monte Carlo simulation methods to estimate the operational risk capital charge, it would simply take too much time to make a thorough assessment of the behavior of the VaR measures under the different scenarios we have considered in this thesis.

- We have developed a *classical* model that utilizes the LDA framework to estimate the operational risk capital charge. It was shown that we are able to produce accurate results for relatively small loss data samples, when the severity distribution is correctly specified and the operational loss data is independent and identically distributed (IID). Furthermore, we have shown how to appropriately address the issue of loss data truncation. We applied the constrained maximum likelihood approach (CML) to estimate the severity distribution parameters for truncated loss data. It was seen that this approach produces far less biased results in estimated the capital charge than the shifting approach, which is a commonly used method in practice.

- It was seen that the operational risk capital charge is highly sensitive to minor contamination of the loss data. This has to do with the fact that the distributions that are used to model the loss severity are generally very heavy-tailed and a minor change in the parameter estimates can have a large impact on the resulting capital charge. We note that this phenomenon is much less present in market risk, where regulatory capital can also be calculated using VaR, since the distributions that are used there are usually much less heavy-tailed (e.g. the normal distribution). Furthermore, an important result in assessing the influence of minor contamination on the capital charge is that (especially for non-truncated loss data), very small losses can be just as devastating for the estimated VaR measures as extreme large losses. In practice, banks actually do a good job in suppressing this issue, (even if they are unaware of it), since the loss data enters the database only above certain thresholds and banks often use mixed severity distributions to model the operational loss data. In this sense, banks can use the arguments presented in this thesis as a justification for the models applied to estimated the operational risk capital charge.

- We have introduced two robust estimation methods: the optimal bias robust estimators (OBRE) and the method of trimmed moments (MTM). From a practical point of view, we have described in detail how to implement both estimation procedures (for both non-truncated and truncated loss data) and how to compute the estimators for several heavy-tailed severity distributions. The (theoretical) input for the numerical algorithms can be adopted directly from the thesis in order to produce results, if a bank decides to adopt the robust estimation techniques presented here.

- In conclusion, we have combined the introduced material into a single final *robust* model to estimate the operational risk capital charge. Following this model we have shown that the estimated VaR measures keep a high level of efficiency at the exact model distribution and it gains stability under minor contamination of the loss data, where the classical models (that are used in practice) generally fail.

7.1 Implications for future research

In practice, banks have to take into account several other factors to estimate the operational risk charge that we did not consider here. Below we list several interesting topics of research, extending the results derived so far.

- To meet the Basel II regulatory requirements for the AMA, a bank's internal model must include the use of internal data, (relevant) external data, expert opinions and factors reflecting the business environment and internal control systems. The most straightforward approach to combine multiple data sources is to give fixed weights to each different data source. Essentially, the overall severity distribution is then estimated as a mixture of severity distributions, with fixed weights. Under this *Ad-hoc combining* approach, we can use the introduced robust estimation methods to model the severity distributions for each single data source and combine them afterwards. A more sophisticated approach, that is recently introduced in the setting of operational risk, is to combine the multiple data sources using *Bayesian methods*. The idea is to model a prior severity distribution using for instance external data and/or expert opinions, then as new internal loss data enters the database, we update our prior beliefs into a posterior severity distribution (we refer to Shevchenko [29] for further details). In Shevchenko [29] it is shown that, as more internal loss data becomes available, the Bayesian estimators converge to the maximum likelihood estimators (of the internal loss data), which we have seen are not robust. Initially, we set out to include this subject in the thesis, but we decided to skip this in order to keep the material in proportions. For the lognormal severity distribution, following Shevchenko [29] one can find closed-form expressions for the Bayesian parameter estimators. Using the weights of the OBRE, we can modify the Bayesian estimators such that, when the loss data sample increases, they converge to the optimal bias robust estimators of the internal loss data (instead of the maximum likelihood estimators). For the log-gamma and GPD, no closed-form expressions are available and the Bayesian estimators are computed using Monte Carlo Markov Chains. In this case, we can robustify the Bayesian estimators according to some outlier rejection rule when the internal loss data enters the database. In conclusion, it might be interesting to study robust Bayesian estimators more extensively, since the banks that decide to implement robust estimation methods also need to know how to address the combination of multiple data sources.

- Throughout the thesis, we have only considered loss data from a single unit of measure. In practice, banks may have to deal with upto 56 different units of measure. As described in Section 3.2, there is usually a weak dependence structure across the units of measure, which can be modeled by using for instance Copula functions. We expect many of the results in the thesis to be equally valid when multiple units of measure are considered. It might be interesting to study the impact on the capital charge of contamination of the loss data across different units of measure.

Remark 7.0.8. We note that currently, there is an ongoing discussion on the subject of stress testing in an operational risk setting. In the previous context, contamination of the loss data across multiple units of measure, may be viewed as (fictional) stress scenarios in the market. Thus, using the derived results, we can assess the behavior of the estimated capital charge under different sources of contamination, i.e. stress scenarios.

- In the thesis we have only examined results for a single risk measure: Value-at-Risk (VaR). It is well-known that VaR is not a coherent risk measure[1]. Therefore, in practice often other risk measures are used. An example of a coherent risk measure is Conditional Value-at-Risk (CVaR) (also referred to as Expected Shortfall). CVaR is defined as

$$\text{CVaR}_\alpha(X) = \mathbb{E}[X|X \geq \text{VaR}_\alpha(X)] \tag{7.1}$$

  In general, for high quantiles (such as 0.999-VaR), ordinary VaR and CVaR will be very close to each other. Thus, we expect that the derived results for VaR will for the most part also hold true for other risk measures, such as CVaR. It is noted that currently, the formal Basel II regulatory requirement for the operational risk capital charge refers to the ordinary VaR (see BCBS [2]), but perhaps this will change in the near future.

- Our main focus has been on the influence function to assess the behavior of the parameter estimators under minor contamination of the loss data. An other important tool in robust statistics is the *breakdown point*, which describes the amount of contamination an estimator can sustain before it becomes useless. If we consider the parameter estimators under the method of trimmed moments, the computation of the breakdown point is quite straightforward. For trimming proportions (a_j, b_j) for $j = 1, \ldots, k$, the MTM estimators remain resistant against the proportion $a_\star = \min\{a_1, \ldots, a_k\}$ of lowest loss observations and the proportion $b_\star = \min\{b_1, \ldots, b_k\}$ of highest loss observations. For the OBRE it is not immediately clear how to compute the breakdown point under different choices of the tuning parameter c, since the OBRE do not cut off loss observations entirely, but assign lower weights to outlying observations.

- In the simulation studies of the previous chapter, we have specified the tuning parameters c and trimming proportions (a, b) manually beforehand. In practice, a bank should be able to justify these choices, therefore it might be interesting to study a data-driven process, where the tuning parameter or trimming proportions are determined by the loss data sample at hand. This line of research is also proposed in Opdyke and Cavallo [27] for an optimal choice of the tuning parameter c under the OBRE, and it is stated that research is currently underway.

[1] A coherent risk measure satisfies (among other properties) sub-additivity, VaR is typically not sub-additive, i.e. $\text{VaR}_\alpha(X+Y) \nleq \text{VaR}_\alpha(X) + \text{VaR}_\alpha(Y)$, that is the VaR of a sum may be larger than the sum of VaRs.

Appendix A

A

A.1 Methods for computing the compound distribution

We give a brief overview of several approaches to evaluate the compound loss distribution and compute the VaR measures. Below we will describe the Monte Carlo method and the Single-Loss Approximation (also called VaR closed-form approximation). We compare the performance of these two approaches to the FFT method as described in Section 3.3.

A.1.1 Monte Carlo Method

The simplest and most commonly used approach relies on Monte Carlo simulation of loss data samples. The algorithm can be outlined as follows:

1. Simulate N the number of events from the assumed frequency distribution Q
2. Simulate independent severities $X_1, X_2, \dots, X_N$ from the severity distribution F
3. Calculate $S_N = \sum_{i=1}^{N} X_i$
4. Repeat Step 1 to 3 for K a large number of times, depending on the accuracy we wish to achieve.

The obtained values $S_N^{(1)}, S_N^{(2)}, \dots S_N^{(K)}$ are samples from the compound distribution $Q \vee F$. VaR$_\alpha$ can then be obtained as the α-th empirical quantile of the compound loss distribution. Clearly, the precision of the approximation increases with a larger number of scenarios. This becomes particularly important for heavy-tailed loss distributions, because a large number of simulations is required to generate a sufficient number of tail events. In general when estimating a very high quantile such as 0.999-VaR, Monte Carlo Methods are very time-consuming and slow to converge to the *true* solution.

A.1.2 Single-Loss Approximation (SLA)

There are several well-known approximations for the compound loss distributions. These can be used with different degrees of success depending on the quantity to be calculated and distribution types. Even when the accuracy is not good, these approximations are still useful from the methodological point of view. Also, the VaR estimate derived from the SLA can successfully be used to set a truncation point M for the FFT method, that will subsequently determine the VaR more precisely.
Heavy-tailed distributions belong to the class of subexponential distributions, in which the maximum observation $M_n = \max\{X_1, X_2, \dots, X_n\}$ in a sample of size n determines the behavior of the entire sum $S_n = \sum_{i=1}^{n} X_i$, i.e. one very large observation can dominate the tail behavior of the whole aggregate process, which can mathematically be written as

$$P(S_n > x) \approx P(M_n > x) \tag{A.1}$$

By assumed independence of the X_i's, for heavy-tailed severity distributions, the following approximation holds

$$P(S_N > x) \approx \mathbb{E}[N](1 - F(x)) \tag{A.2}$$

Embrechts [14] shows the validity of this asymptotic result when N follows a Poisson, Binomial or Negative Binomial distribution. This approximation can be used to calculate high quantiles of the compound loss distribution according to

$$\text{VaR}_\alpha(X) \to F^{-1}\left(1 - \frac{1-\alpha}{\mathbb{E}[N]}\right) \quad \text{as} \quad \alpha \to 1 \tag{A.3}$$

A.1.3 Comparison of numerical methods

We wish to compare the previously described methods to the FFT method described in Section 3.3. We compute VaR measures for the following combinations of frequency and severity distribution

- $Pois(\lambda = 25) \vee \mathcal{LN}(\mu, \sigma)$, with fixed parameters of the lognormal distribution $\mu = 10.95$ and $\sigma = 1.75$.

- $Pois(\lambda = 25) \vee \mathcal{L}G(a, b)$, with fixed parameters of the log-gamma distribution $a = 34.5$ and $b = 3.5$

- $Pois(\lambda = 25) \vee GPD(\xi, \beta)$, with fixed parameters of the Generalized Pareto distribution $\xi = 0.65$ and $\beta = 57500$.

In Table A.1 below, we list the resulting VaR measures and corresponding computation time in *R*.
We know that the Monte Carlo method will eventually converge to the solution of the FFT method, since -if there is no aliasing error- the FFT method computes the *exact* solution (up to $\pm h$). Therefore, both in terms of accuracy and computation time, we will always prefer the FFT method over the Monte Carlo method. Furthermore, it is seen that the results from the SLA are not very inaccurate, especially for the GPD it is quite close to the solution of the Monte Carlo and FFT method. So, if one does not want to bother with the implementation of the FFT method, the SLA can be a good alternative, due to its simplicity.

		Severity distribution		
		$Pois \vee \mathcal{LN}$	$Pois \vee \mathcal{LG}$	$Pois \vee GPD$
MC	VaR	63,931,502	62,174,273	67,882,052
	time	12.3 min	16.9 min	15.9 min
	K	10^7	10^7	10^6
SLA	VaR	56,667,135	59,478,642	67,906,122
	time	0 s	0 s	0 s
FFT	VaR	63,945,500	62,291,000	67,916,500
	time	0.70 s	0.73 s	0.64 s
	h	500	500	500
	M	2^{18}	2^{18}	2^{18}

Tab. A.1.: Comparison of VaR estimates via the Monte Carlo method, SLA and FFT method.

A.2 Capital bias under the EM algorithm

The Expectation-Maximization (EM) algorithm, proposed by Dempster, Laird and Rubin (1977) is aimed at estimating the unknown parameters by maximizing the expected likelihood function given the observed *and* the missing data. The EM algorithm is a two-step iterative procedure. In the initial step, given an initial guess value $\theta^{(0)}$ for the unknown parameter set θ, the missing data values in the log-likelihood function are replaced by their expected values. This leads to the guess value for the expected complete log-likelihood function (expectation step), which is further maximized with respect to the parameter values (maximization step). The solution is then used as the initial guess in the next iteration of the algorithm, and the expectation step and the maximization step are repeated until the solution converges. The EM algorithm can thus be summarized as follows:

1. *Initial step*: Choose initial parameter values $\theta^{(0)}$. These can be used to estimate the initial guess value $m^{(0)}$ representing the number of missing observations.

2. *Expectation step* (E-step): Given $\theta^{(0)}$, calculate the expected log-likelihood function of the complete data. If we denote the observed data sample by $\boldsymbol{X}^+$, the missing observations by $\boldsymbol{X}^-$ and the complete data sample by $\boldsymbol{X}^\pm$, we wish to compute

$$\mathbb{E}_{\theta^{(0)}}\left[\ell(\theta|\boldsymbol{X}^\pm)|\boldsymbol{X}^+\right] = m^{(0)}\mathbb{E}_{\theta^{(0)}}\left[\log f_\theta(\boldsymbol{X}^-)\right] + \sum_j \log f_\theta(X_j^+) \tag{A.4}$$

3. *Maximization step* (M-step): Find the parameter set θ that maximizes the expected log-likelihood function from the previous step and set it equal to the guess value in the next step $\theta^{(1)}$. Mathematically we need to estimate

$$\theta^{(1)} = \arg\max_\theta \mathbb{E}_{\theta^{(0)}}\left[\ell(\theta|\boldsymbol{X}^\pm)|\boldsymbol{X}^+\right] \tag{A.5}$$

4. *Iteration*: Repeat the E- and M-step, until the sequence $\left\{\theta^{(n)}\right\}_{n>0}$ converges to the desired maximum likelihood estimates of the parameter of the distribution of the complete data sample.

Because in every round of the EM-algorithm the unknown parameters are replaced with the values that are closer to the true values, at every round the value of the likelihood function increases relative to the previous round. We perform the same simulation experiment as for the shifting approach and constrained maximum likelihood function approach, only now we apply the EM algorithm to compute parameter estimates $(\widehat{\mu}, \widehat{\sigma})$ for the truncated loss data samples. The routine is repeated 25 times in order to produce average results, see Fig.A.1 below for a surface plot of the capital bias.

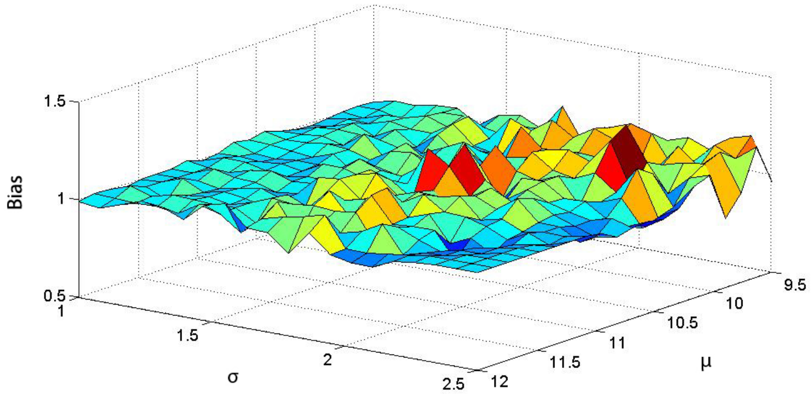


Fig. A.1.: Capital bias under the EM-algorithm with threshold H at the lower 0.35 quantile of the data

Appendix B

B

B.1 MLE IF log-gamma distribution

The MLE IF of the log-gamma distribution can be computed using the general multi-parameter form in Eq.4.34. We recall that the probability density function of the log-gamma function with shape parameter $a > 0$ and rate parameter $b > 0$ is given by

$$f(x|a,b) = \frac{b^a (\log(x))^{(a-1)}}{\Gamma(a) x^{(b+1)}} \tag{B.1}$$

Where $\Gamma(a)$ is the complete gamma function.
Now we compute

$$\phi_\theta = \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} = \begin{bmatrix} -\frac{1}{f(x|a,b)} \left(\frac{\partial f(x|a,b)}{\partial a}\right) \\ -\frac{1}{f(x|a,b)} \left(\frac{\partial f(x|a,b)}{\partial b}\right) \end{bmatrix} = \begin{bmatrix} -[\log(b) + \log(\log(x) - \psi_0(a)] \\ -[(a/b) - \log(x)] \end{bmatrix} \tag{B.2}$$

Furthermore, the coordinates of the Fisher information matrix are given by

$$\begin{aligned} -\int \frac{\partial \phi_a}{\partial a} &= -\int \psi_1(a) f(y) dy = -\psi_1(a) && \text{(B.3)} \\ -\int \frac{\partial \phi_b}{\partial b} &= -\int \frac{a}{b^2} f(y) dy = -\frac{a}{b^2} && \text{(B.4)} \\ -\int \frac{\partial \phi_b}{\partial a} &= -\int -\frac{1}{b} f(y) dy = \frac{1}{b} && \text{(B.5)} \\ -\int \frac{\partial \phi_a}{\partial b} &= -\int \frac{\partial \phi_b}{\partial a} = \frac{1}{b} && \text{(B.6)} \end{aligned}$$

Where ψ_0 and ψ_1 denote the digamma and the trigamma functions respectively, i.e. the first- and second-order logarithmic derivatives of the complete gamma function: $\psi_0(z) = \frac{\partial}{\partial z} \log(\Gamma(z))$ and $\psi_1(z) = \frac{\partial^2}{\partial z^2} \log(\Gamma(z))$. From the above equations we see that the off-diagonal entries of the Fisher information are zero, indicating that the parameters are uncorrelated.
According to Eq.4.34 the IF then becomes

$$\begin{aligned} \mathrm{IF}(x|\theta, F) &= A(\theta)^{-1} \cdot \phi_\theta && \text{(B.7)} \\ &= \begin{bmatrix} -\psi_1(a) & 1/b \\ 1/b & -a/b^2 \end{bmatrix}^{-1} \begin{bmatrix} -\log(b) - \log(\log(x)) + \psi_0(a) \\ -(a/b) + \log(x) \end{bmatrix} && \text{(B.8)} \\ &= \frac{1}{(a/b^2)\psi_1(a) - 1/b^2} \begin{bmatrix} -a/b^2 & -1/b \\ -1/b & -\psi_1(a) \end{bmatrix} \begin{bmatrix} -\log(b) - \log(\log(x)) + \psi_0(a) \\ -(a/b) + \log(x) \end{bmatrix} && \text{(B.9)} \\ &= \begin{bmatrix} \frac{(a/b^2)[\log(b)+\log(\log(x))-\psi_0(a)]-(1/b)[\log(x)-(a/b)]}{\psi_1(a)(a/b^2)-(1/b^2)} \\ \frac{(1/b)[\log(b)+\log(\log(x))-\psi_0(a)]-\psi_1(a)[\log(x)-(a/b)]}{\psi_1(a)(a/b^2)-(1/b^2)} \end{bmatrix} && \text{(B.10)} \end{aligned}$$

Both the IF of a and the IF of b diverge as $x \rightarrow +\infty$. Therefore, we conclude that the maximum likelihood estimators for a and b are *not* bias robust.

B.2 MLE IF Generalized Pareto distribution

We repeat the derivation process for the MLE IF of the Generalized Pareto distribution. We recall that the probability density function of the Generalized Pareto distribution with shape parameter $\xi > 0$ and scale parameter $\beta > 0$ is given by

$$f(x|\xi,\beta) = \frac{1}{\beta}\left[1+\xi\frac{x}{\beta}\right]^{(-\frac{1}{\xi}-1)} \tag{B.11}$$

We can compute

$$\phi_\theta = \begin{bmatrix}\phi_\xi \\ \phi_\beta\end{bmatrix} = \begin{bmatrix}-\frac{1}{f(x|\xi,\beta)}\left(\frac{\partial f(x|\xi,\beta)}{\partial\xi}\right) \\ -\frac{1}{f(x|\xi,\beta)}\left(\frac{\partial f(x|\xi,\beta)}{\partial\beta}\right)\end{bmatrix} = \begin{bmatrix}-\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2x}\right)+\frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right] \\ \frac{1}{\beta}\left[\frac{\beta-x}{\beta+\xi x}\right]\end{bmatrix} \tag{B.12}$$

Furthermore, the terms of the Fisher information matrix become

$$\begin{aligned}
-\int\frac{\partial\phi_\xi}{\partial\xi}dF(x) &= -\int\left[\frac{x\beta+2\xi x^2+\xi^2x^2}{(\beta\xi+\xi^2x)^2}+\frac{x}{(\beta+\xi x)\xi^2}\right. && \text{(B.13)}\\
&\qquad\left.-\frac{1}{\xi^3}2\log\left(1+\frac{\xi x}{\beta}\right)\right]f(x)dx && \text{(B.14)}\\
-\int\frac{\partial\phi_\beta}{\partial\beta}dF(x) &= -\int\left[\frac{1}{\beta^2}-\frac{x(1+\xi)(2\beta+\xi x)}{(\beta^2+\beta\xi x)^2}\right]f(x)dx && \text{(B.15)}\\
-\int\frac{\partial\phi_\xi}{\partial\beta}dF(x) &= -\int\frac{\partial\phi_\beta}{\partial\xi}dF(x) && \text{(B.16)}\\
&= -\int\left[\frac{x}{\beta\xi(\beta+\xi x)}-\frac{\xi x(1+\xi)}{(\beta\xi+\xi^2x)^2}\right]f(x)dx && \text{(B.17)}
\end{aligned}$$

The above integrals need to be solved numerically, however according to Smith [30] the inverse of the Fisher information matrix can be simplified to

$$A(\theta)^{-1} = \begin{bmatrix}(1+\xi)^2 & -(1+\xi)\beta \\ -(1+\xi)\beta & 2(1+\xi)\beta^2\end{bmatrix} \tag{B.18}$$

The IF is then given by

$$\mathrm{IF}(x|\theta,T) = A(\theta)^{-1}\phi_\theta = \begin{bmatrix}-(1+\xi)^2\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2x}\right)+\frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right]-(1+\xi)\left[\frac{\beta-x}{\beta+\xi x}\right] \\ (1+\xi)\beta\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2x}\right)+\frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right]+2(1+\xi)\beta\left[\frac{\beta-x}{\beta+\xi x}\right]\end{bmatrix}$$

Again both the IF of ξ and the IF of β diverge as $x \to \pm\infty$. Thus, the maximum likelihood estimators for ξ and β are *not* bias robust.

B.3 CML IF log-gamma distribution

We repeat the calculation of Section 4.5.1 for the truncated log-gamma distribution. We can compute

$$\begin{aligned}
\phi_\theta &= \begin{bmatrix}\phi_a \\ \phi_b\end{bmatrix} = \begin{bmatrix}-\frac{\partial f(x|a,b)/\partial a}{f(x|a,b)}-\frac{\partial F(H|a,b)/\partial a}{1-F(H|a,b)} \\ -\frac{\partial f(x|a,b)/\partial b}{f(x|a,b)}-\frac{\partial F(H|a,b)/\partial b}{1-F(H|a,b)}\end{bmatrix} && \text{(B.19)}\\
&= \begin{bmatrix}-[\log(b)+\log(\log(x))-\psi_0(a)]-\frac{\int_1^H[\log(b)+\log(\log(y))-\psi_0(a)]f(y|a,b)dy}{1-F(H|a,b)} \\ -\left[\frac{a}{b}-\log(x)\right]-\frac{\int_1^H\left[\frac{a}{b}-\log(y)\right]f(y|a,b)dy}{1-F(H|a,b)}\end{bmatrix} && \text{(B.20)}
\end{aligned}$$

And the coordinates of the Fisher information matrix are given by

$$-\int_H^\infty \frac{\partial \phi_a}{\partial a} dG(y) = -\psi_1(a) + \frac{\left[\int_1^H ([\log(b) + \log(\log(y)) - \psi_0(a)]^2 - \psi_1(a)) f(y|a,b)dy\right]^2}{(1 - F(H|a,b))^2} \tag{B.21}$$

$$+\frac{\int_1^H ([\log(b) + \log(\log(y)) - \psi_0(a)]^2 - \psi_1(a)) f(y|a,b)dy}{1 - F(H|a,b)} \tag{B.22}$$

$$-\int_H^\infty \frac{\partial \phi_b}{\partial b} dG(y) = -\frac{a}{b^2} + \frac{\left[\int_1^H \left[\frac{a}{b} - \log(y)\right] f(y|a,b)dy\right]^2}{(1 - F(H|a,b))^2} \tag{B.23}$$

$$+\frac{\int_1^H \left[\frac{a(a-1)}{b^2} - \frac{2a\log(y)}{b} + (\log(y))^2\right] f(y|a,b)dy}{1 - F(H|a,b)} \tag{B.24}$$

$$-\int_H^\infty \frac{\partial \phi_a}{\partial b} dG(y) = -\int_H^\infty \frac{\partial \phi_b}{\partial a} dG(y) \tag{B.25}$$

$$= \frac{1}{b} + \frac{1}{1 - F(H|a,b)} \int_1^H \left(\frac{1}{b} + [\log(b) + \log(\log(y)) - \psi_0(a)] \right. \tag{B.26}$$

$$\left. \cdot \left(\frac{a}{b} - \log(y) \right) \right) f(y|a,b)dy + \left[\int_1^H \left[\frac{a}{b} - \log(y) \right] f(y|a,b)dy \right] \tag{B.27}$$

$$\cdot \left[\int_1^H [\log(b) + \log(\log(y)) - \psi_0(a)] f(y|a,b)dy \right] \cdot (1 - F(H|a,b))^{-2} \tag{B.28}$$

We are now able to find the IF by solving numerically

$$\mathrm{IF}(x|\theta, T, H) = \begin{bmatrix} -\int_H^\infty \frac{\partial \phi_a}{\partial a} dG(y) & -\int_H^\infty \frac{\partial \phi_a}{\partial b} dG(y) \\ -\int_H^\infty \frac{\partial \phi_b}{\partial a} dG(y) & -\int_H^\infty \frac{\partial \phi_b}{\partial b} dG(y) \end{bmatrix}^{-1} \cdot \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} \tag{B.29}$$

We note that parameter correlation is indicated by nonzero off-diagonal terms. Inspection of the IFs for both a and b under different truncation thresholds show that the CML estimators of the log-gamma distribution are not bias robust.

B.4 CML IF Generalized Pareto distribution

We repeat the calculation of Section 4.5.1 for the truncated GPD. We can compute

$$\phi_\theta = \begin{bmatrix}\phi_\xi \\ \phi_\beta\end{bmatrix} = \begin{bmatrix}-\frac{\partial f(x)/\partial\xi}{f(x)} - \frac{\partial F(H)/\partial\xi}{1-F(H)} \\ -\frac{\partial f(x)/\partial\beta}{f(x)} - \frac{\partial F(H)/\partial\beta}{1-F(H)}\end{bmatrix} \tag{B.30}$$

$$= \begin{bmatrix} -\left[\left(\frac{-x(1+\xi)}{\beta\xi+\xi^2x}\right) + \left(\frac{1}{\xi^2}\log\left(1+\frac{\xi x}{\beta}\right)\right)\right] - \frac{\int_0^H\left[\left(\frac{-y(1+\xi)}{\beta\xi+\xi^2y}\right)+\left(\frac{1}{\xi^2}\log\left(1+\frac{\xi y}{\beta}\right)\right)\right]f(y)dy}{1-F(H)} \\ \frac{1}{\beta}\left[\frac{\beta-x}{\beta+\xi x}\right] - \frac{\int_0^H -\frac{1}{\beta}\left[\frac{\beta-y}{\beta+\xi y}\right]f(y)dy}{1-F(H)}\end{bmatrix} \tag{B.31}$$

And the coordinates of the Fisher information matrix are given by

$$-\int_H^\infty \frac{\partial\phi_\xi}{\partial\xi}dG(y) = -\frac{\int_H^\infty\left[\frac{y\beta+2\xi y^2+\xi^2y^2}{(\beta\xi+\xi^2y)^2} + \frac{y}{(\beta+\xi y)\xi^2} - \frac{2\log(1+\xi y/\beta)}{\xi^3}\right]f(y)dy}{1-F(H)} \tag{B.32}$$

$$+\frac{\left(\int_0^H\left[\left(\frac{-y(1+\xi)}{\beta\xi+\xi^2y}\right)+\left(\frac{1}{\xi^2}\log\left(1+\frac{\xi y}{\beta}\right)\right)\right]f(y)dy\right)^2}{(1-F(H))^2} \tag{B.33}$$

$$-\int_H^\infty \frac{\partial\phi_\beta}{\partial\beta}dG(y) = -\frac{\int_H^\infty\left[\frac{1}{\beta^2} - \frac{y(1+\xi)(2\beta+\xi y}{(\beta^2+\beta\xi y)^2}\right]f(y)dy}{1-F(H)} \tag{B.34}$$

$$+\frac{\left(\int_0^H -\frac{1}{\beta}\left[\frac{\beta-y}{\beta+\xi y}\right]f(y)dy\right)^2}{(1-F(H))^2} \tag{B.35}$$

$$+\frac{\int_0^H\left(\left[\frac{1}{\beta^2} - \frac{y(1+\xi)(2\beta+\xi y}{(\beta^2+\beta\xi y)^2}\right] + \frac{1}{\beta^2}\left[\frac{\beta-y}{\beta+\xi y}\right]^2\right)f(y)dy}{1-F(H)} \tag{B.36}$$

$$+(1-F(H))^{-1}\int_0^H\left(\left[\frac{y\beta+2\xi y^2+\xi^2y^2}{(\beta\xi+\xi^2y)^2} + \frac{y}{(\beta+\xi y)\xi^2} - \frac{2\log(1+\xi y/\beta)}{\xi^3}\right]\right. \tag{B.37}$$

$$\left.+\left[\left(\frac{-y(1+\xi)}{\beta\xi+\xi^2y}\right) + \frac{\log(1+\xi y/\beta)}{\xi^2}\right]^2\right)f(y)dy \tag{B.38}$$

$$-\int_H^\infty \frac{\partial\phi_\xi}{\partial\beta}dG(y) = -\int_H^\infty \frac{\partial\phi_\beta}{\partial\xi}dG(y) \tag{B.39}$$

$$= -\frac{\int_H^\infty\left[\frac{\xi y(1+\xi)}{(\beta\xi+\xi^2y)^2} - \frac{y}{\beta\xi(\beta+\xi y)}\right]f(y)dy}{1-F(H)} \tag{B.40}$$

$$+\frac{\left[\int_0^H\left[\left(\frac{-y(1+\xi)}{\beta\xi+\xi^2y}\right)+\left(\frac{1}{\xi^2}\log\left(1+\frac{\xi y}{\beta}\right)\right)\right]f(y)dy\right]\left[\int_0^H -\frac{1}{\beta}\left[\frac{\beta-y}{\beta+\xi y}\right]f(y)dy\right]}{(1-F(H))^2} \tag{B.41}$$

$$+\frac{\int_0^H\left(\left[-\frac{1}{\beta}\left(\frac{\beta-y}{\beta+\xi y}\right)\right]\left[\frac{-y(1+\xi)}{\beta\xi+\xi^2y} + \frac{\log(1+\xi y/\beta)}{\xi^2}\right]\left[\frac{\xi y(1+\xi)}{(\beta\xi+\xi^2y)^2} - \frac{y}{\beta\xi(\beta+\xi y)}\right]\right)f(y)dy}{1-F(H)} \tag{B.42}$$

We are now able to find the IF by solving numerically

$$IF_\theta(x|\theta,T,H) = \begin{bmatrix}-\int_H^\infty\frac{\partial\phi_\beta}{\partial\beta}dG(y) & -\int_H^\infty\frac{\partial\phi_\beta}{\partial\xi}dG(y) \\ -\int_H^\infty\frac{\partial\phi_\xi}{\partial\beta}dG(y) & -\int_H^\infty\frac{\partial\phi_\xi}{\partial\xi}dG(y)\end{bmatrix}^{-1}\cdot\begin{bmatrix}\phi_\beta \\ \phi_\xi\end{bmatrix} \tag{B.43}$$

We note that parameter correlation is indicated by nonzero off-diagonal terms. Inspection of the IFs for both β and ξ under different truncation thresholds show that the CML estimators of the Generalized Pareto distribution are not bias robust.

Appendix C

C

C.1 Implementation OBRE algorithm

For non-truncated loss data from the lognormal distribution, the numerical algorithm described in Section 5.1.2 is implemented in *R* according to

```
> #Define prob. density and score functions with parameter values ’theta’
> f <- function(x, theta){return(dnorm(x, theta[1], sqrt(theta[2])))}
> score <- function(x, theta){
                          s1 <- 1/theta[2] * (x - theta[1])
                          s2 <- -1/(2 * theta[2]) + 1/(2 * theta[2]^2) * (x - theta[1])^2
                          return(c(s1,s2))
                           }
> score.vec <- Vectorize(score, vectorize.args = "x")
> #Define weight functions with parameter values ’theta’, tuning ’c’, and Lagrange multipliers ’a’ and ’A’
> weight <- function(x, theta, c, a, A){
                          norm <- sqrt(sum((A %*% (score.vec(x, theta) - a))^2))
                          W <- min(1, c/norm)
                          return(W)
                         }
> weight.vec <- Vectorize(weight, vectorize.args = "x")
> #Define the functions needed to compute ’a’, ’A’, ’M1.matrix’ and ’M2.matrix’
> a.fun <- function(x, theta, c, a, A){
                         s <- score.vec(x, theta)
                         a11 <- s[1] * weight.vec(x, theta, c, a, A) * f(x, theta)
                         a12 <- s[2] * weight.vec(x, theta, c, a, A) * f(x,theta)
                         a2 <- weight.vec(x, theta, c, a, A) * f(x,theta)
                          return(c(a11, a12, a2))
                         }
> a.vec <- Vectorize(a.fun, vectorize.args = "x")
> M2 <- function(x, theta, c, a, A){
                          s <- (score.vec(x, theta) - a) %*% t(score.vec(x, theta) - a)
                          M2.matrix <- s * (weight.vec(x, theta, c, a, A))^2 * f(x, theta)
                          return(M2.matrix)
                           }
> M2.vec <- Vectorize(M2, vectorize.args = "x")
> M1 <- function(x, theta, c, a, A){
                          s <- (score.vec(x, theta) - a) %*% t(score.vec(x, theta) - a)
                          M1.matrix <- s * (weight.vec(x, theta, c, a, A)) * f(x, theta)
                          return(M1.matrix)
                           }
> M1.vec <- Vectorize(M1, vectorize.args = "x")
> #Define the OBRE algorithm, with initial parameters ’theta0’, loss data ’X’ and tuning ’c’
> OBRE.LN <- function(theta0, X, c){
                   #Step 1: Define initial value ’theta’ and compute initial values ’a’ and ’A’
                    theta <- theta0
                    tol <- .Machine$double.eps^0.25
                    a <- c(0,0)
                    J11 <- 1/(theta[2])
                    J12 <- 0
```

```
    J22 <- 1/(2*(theta[2])^2)
    J <- matrix(c(J11, J12, J12, J22), nrow = 2, ncol = 2)
    A <- chol(inv(J))

    repeat{
     #Step 2: Compute the new values of 'a' and 'A'
      a11.int <- function(x){a.vec(x, theta, c, a, A)[1]}
      a12.int <- function(x){a.vec(x, theta, c, a, A)[2]}
      a2.int <- function(x){a.vec(x, theta, c, a, A)[3]}
      a11 <- integrate(a11.int, lower = -Inf, upper = Inf)
      a12 <- integrate(a12.int, lower = -Inf, upper = Inf)
      a2 <- integrate(a2.int, lower = -Inf, upper = Inf)
      a.new <- c(a11$value, a12$value) / a2$value
      M2.11.int <- function(x){M2.vec(x, theta, c, a, A)[1,1]}
      M2.12.int <- function(x){M2.vec(x, theta, c, a, A)[1,2]}
      M2.21.int <- function(x){M2.vec(x, theta, c, a, A)[2,1]}
      M2.22.int <- function(x){M2.vec(x, theta, c, a, A)[2,2]}
      M2.matrix.11 <- integrate(M2.11.int, lower = -Inf, upper = Inf)
      M2.matrix.12 <- integrate(M2.12.int, lower = -Inf, upper = Inf)
      M2.matrix.21 <- integrate(M2.21.int, lower = -Inf, upper = Inf)
      M2.matrix.22 <- integrate(M2.22.int, lower = -Inf, upper = Inf)
      M2.matrix <- matrix(c(M2.matrix.11$value, M2.matrix.21$value,
                            M2.matrix.12$value, M2.matrix.22$value), nrow = 2, ncol = 2)
      if(!is.positive.definite(inv(M2.matrix))){
              break
      }
            A.new <- chol(inv(M2.matrix))

      #Step 3: Compute 'd.theta', using 'a.new' and 'A.new'
      M1.11.int <- function(x){M1.vec(x, theta, c, a.new, A.new)[1,1]}
      M1.12.int <- function(x){M1.vec(x, theta, c, a.new, A.new)[1,2]}
      M1.21.int <- function(x){M1.vec(x, theta, c, a.new, A.new)[2,1]}
      M1.22.int <- function(x){M1.vec(x, theta, c, a.new, A.new)[2,2]}
      M1.matrix.11 <- integrate(M1.11.int, lower = -Inf, upper = Inf)
      M1.matrix.12 <- integrate(M1.12.int, lower = -Inf, upper = Inf)
      M1.matrix.21 <- integrate(M1.21.int, lower = -Inf, upper = Inf)
      M1.matrix.22 <- integrate(M1.22.int, lower = -Inf, upper = Inf)
      M1.matrix <- matrix(c(M1.matrix.11$value, M1.matrix.21$value,
                            M1.matrix.12$value, M1.matrix.22$value), nrow = 2, ncol = 2)
      score.values <- score.vec(X, theta)
      weight.vales <- weight.vec(X, theta, c, a.new, A.new)
      s1 <- mean((score.values[1,] - a.new[1]) * weight.values)
      s2 <- mean((score.values[2,] - a.new[2]) * weight.values)
      d.theta <- inv(M1.matrix) %*% c(s1,s2)

      #Step 4: if difference below 'tol' then stop, else return to Step 2
      if((max(abs(d.theta/theta))) > tol){
              theta <- theta + d.theta
              a <- a.new
              A <- A.new
            } else{
                  break
                }
    }
    #Return the OBRE
    return(c(theta[1], sqrt(theta[2])))
}
```

Bibliography

[1]M. P. Alaiz and M. Victoria-Feser. *Modeling income distribution in Spain: a robust parametric approach*. Suntory, Toyota International Centres for Economics, and Related Disciplines, 1996 (cit. on pp. 51, 54).

[2]Basel Committee on Banking Supervision. *Consultative Document Operational Risk*. Bank of International Settlements. 2001 (cit. on pp. 6, 9, 11, 95).

[3]Basel Committee on Banking Supervision. *Observed range of practice in key elements of Advanced Measurement Approaches (AMA)*. Bank of International Settlements. 2009 (cit. on p. 8).

[4]V. Brazauskas. „Robust and efficient fitting of loss models: diagnostic tools and insights“. In: *North American Actuarial Journal* 13 (2009), pp. 356–369 (cit. on p. 66).

[5]V. Brazauskas and A. Kleefeld. „Robust and efficient fitting of the generalized Pareto distribution with actuarial applications in view“. In: *Insurance: Mathematics and Economics* 45 (2009), pp. 424–435 (cit. on p. 76).

[6]V. Brazauskas, B. L. Jones, and R. Zitikis. „Robust fitting of claim severity distributions and the method of trimmed moments“. In: *Journal of Statistical Planning and Inference* 139 (2009), pp. 2028–2043 (cit. on pp. 63, 64).

[7]S. Carillo-Menéndez. *Operational risk, presentation at International summer school on Risk measurement and Control*. Rome. 2005 (cit. on p. 16).

[8]A. Cavallo, B. Rosenthal, X. Wang, and J. Yan. „Treatment of the data collection threshold in operational risk: a case study using the lognormal distribution“. In: *Journal of Operational Risk* 7.1 (2012), pp. 3–38 (cit. on p. 21).

[9]A. S. Chernobai and S. T. Rachev. „Applying robust method to operational risk modeling“. In: *Journal of Operational Risk* 1 (2006), pp. 27–41 (cit. on p. 30).

[10]A. S. Chernobai, S. T. Rachev, F. J. Fabozzi, et al. *Operational risk: a guide to Basel II capital requirements, models, and analysis*. Vol. 180. John Wiley & Sons, 2008 (cit. on p. 8).

[11]E. W. Cope, G. Mignola, G. Antonini, and R. Ugoccioni. „Challenges and pitfalls in measuring operational risk from loss data“. In: *Journal of Operational Risk* 4.4 (2009), pp. 3–27 (cit. on p. 17).

[12]U. J. Dixit and P. N. Nasiri. „Estimation of parameters of a right truncated exponential distribution“. In: *Statistical Papers* 49 (2008), pp. 225–236 (cit. on p. 46).

[13]D.J. Dupuis and C.A. Field. „Robust estimation of extremes“. In: *Canadian Journal of Statistics* 26.2 (1998), pp. 199–215 (cit. on pp. 52, 55).

[14]P. Embrechts. *Modelling extremal events: for insurance and finance*. Vol. 33. Springer, 1997 (cit. on p. 98).

[15]P. Embrechts and M. Frei. „Panjer recursion versus FFT for compound distributions“. In: *Mathematical Methods of Operations Research* 3 (2009), pp. 497–508 (cit. on pp. 14, 15).

[16]B. Ergashev, K. Pavlikov, S. Uryasev, and E. Sekeris. „Estimation of Truncated Data Samples in Operational Risk Modeling“. In: *Available at SSRN 2193493* (2012) (cit. on p. 21).

[17]P. Glasserman. *Monte Carlo methods in financial engineering*. Vol. 53. Springer, 2004 (cit. on p. 17).

[18]R. Grübel and R. Hermesmeier. „Computation of compound distributions I: Aliasing errors and exponential tilting“. In: *Astin Bulletin* 29.2 (1999), pp. 197–214 (cit. on p. 14).

[19]F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 1986 (cit. on pp. 29, 30, 32, 33, 51–53).

[20]P. J. Huber. „Robust estimation of a location parameter“. In: *The Annals of Mathematical Statistics* 35 (1964), pp. 73–101 (cit. on p. 33).

[21]P. J. Huber. *Robust statistics*. Springer, 2011 (cit. on pp. 29, 31–34, 51).

[22]A. Kleefeld and V. Brazauskas. „A statistical application of the quantile mechanics approach: MTM estimators for the parameters of t and gamma distributions“. In: *European Journal of Applied Mathematics* 23 (2012), pp. 593–610 (cit. on p. 72).

[23]S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss models: from data to decisions*. Vol. 715. John Wiley & Sons, 2012 (cit. on p. 12).

[24]G. Mignola and R. Ugoccioni. „Sources of uncertainty in modeling operational risk losses“. In: *Journal of Operational Risk* 1.2 (2006), pp. 33–50.

[25]M. Moscadelli. „The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee“. In: *Available at SSRN 557214* (2004) (cit. on p. 19).

[26]J.D. Opdyke. *Estimating Operational Risk Capital with Greater Accuracy, Precision, and Robustness or How to Prevent Jensens Inequality from Inflating Your OpRisk Capital Estimates*. Bates White LLC. 2012 (cit. on pp. 25, 38).

[27]J.D. Opdyke and A. Cavallo. „Estimating operational risk capital: the challenges of truncation, the hazards of MLE, and the promise of robust statistics“. In: *Journal of Operational Risk* (2012), Forthcoming (cit. on pp. 35, 43, 95).

[28]H. H. Panjer. *Operational risk: modeling analytics*. Vol. 620. John Wiley & Sons, 2006 (cit. on p. 12).

[29]P. V. Shevchenko. *Modelling operational risk using Bayesian inference*. Springer, 2011 (cit. on pp. 14, 17, 21, 94).

[30]J. A. Smith. „Estimating the upper tail of flood frequency distributions“. In: *Water Resources Research* 23.8 (1987), pp. 1657–1666 (cit. on p. 102).

[31]G. Steinbrecher and W. T. Shaw. „Quantile mechanics“. In: *European Journal of Applied Mathematics* 19 (2008), pp. 87–112 (cit. on p. 72).

[32]T. J. Ypma. *Historical development of the Newton-Raphson method*. Vol. 37. 4. SIAM, 1995, pp. 531–551 (cit. on p. 19).

[33]X. Zhou, R. Giacometti, F.J. Fabozzi, and A. H. Tucker. „Bayesian estimation of truncated data with applications to operational risk measurement“. In: *Quantitative Finance* (2013), pp. 1–26 (cit. on p. 19).

List of Figures

List of Tables