Bayesian Inference Methods in Operational Risk

Author:
ERIK DAHLBERG, eada@kth.se

June 1, 2015
Master of Science Thesis Report
SA299X Degree Project in Mathematical Statistics, Second Cycle

Supervisors:
Filip Lindskog, KTH
Alexander Jöhnemark, Swedbank

Abstract

Under the Advanced Measurement Approach (AMA), banks must use four different sources of information to assess their operational risk capital requirement. The three main quantitative sources available to build the future loss distribution are internal loss data, external loss data and scenario analysis. The fourth source, business environment and internal control factors, is treated as an ex-post update to capital calculations and is not a subject of this thesis. Approaches from Extreme Value Theory (EVT) have gained popularity in the area of operational risk in recent years, with its focus on the behaviour of processes at extreme levels making it a natural candidate for operational risk modelling. However, the adoption of EVT in operational risk modelling has encountered several obstacles with the main one being the scarcity of data leading to substantial statistical uncertainty for both parameter and capital estimates. This Master thesis evaluates Bayesian Inference approaches to extreme value estimation and implements a method to reduce these uncertainties. The results indicate that the Bayesian Inference approaches gives a significant reduction of the statistical uncertainties compared to more traditional estimators and also performs well when applied on real-world data sets.

Sammanfattning

När Advanced Measurement Approach skall implementeras krävs det att banker använder fyra olika informationskällor för att bedöma sitt kapitalkrav för operationell risk. De tre kvantitativa källorna som används för att bygga den framtida förlustdistributionen är intern förlustdata, extern förlustdata och scenarioanalys. Den fjärde källan, affärsmiljö och interna kontrollfaktorer, behandlas som en ex-post-uppdatering till kapitalberäkningen och är inte ett föremål för denna uppsats. Extremvärdesmetoder har ökat i popularitet inom operationell risk de senaste åren där deras fokus på processers beteende på extremnivåer är väl lämpat för operationell riskmodellering. Likväl har införandet av extremvärdesmetoder i operationell riskmodellering stött på flera hinder varav bristen på lämplig data är den största. Denna brist leder till väsentlig statistisk osäkerhet för både parameter- och kapitalestimat. Detta examensarbete utvärderar Bayesianska Inferensmetoder för extremvärdesestimering och implementerar en metod för att reducera nämnda osäkerheter. Resultaten indikerar at de Bayesianska metoderna ger en signifikant reduktion av de statistiska osäkerheterna jämfört med mer traditionella metoder. Också när metoden används på verklig förlustdata uppnås låg osäkerhet.

Acknowledgements

I would like to thank my supervisor at Swedbank, Alexander Jöhnemark, for his availability, encouragement and very helpful feedback and discussion during my work with the thesis. Also, I want to thank the entire Group Operational Risk at Swedbank, on which behalf I conducted the thesis project, for facilitating all the practical aspects of the project and for making me feel so welcome and appreciated from day one. Lastly, my thanks goes to my supervising professor at KTH, Filip Lindskog, for valuable input during the whole thesis work.

Erik Dahlberg
Stockholm, June 1, 2015

Contents

Chapter 1

Introduction

Under Basel II banks can choose between three methods to calculate their operational risk capital requirement. The three methods are

- The Basic Indicator Approach
- The Standardized Approach
- The Advanced Measurement Approach (AMA)

of which the Advanced Measurement Approach is the most sophisticated method which allows the bank to develop an internal model to calculate their capital requirement subject to approval by the regulatory authority.

Under the AMA the modeller tries to build the future loss distribution using three main quantitative sources, namely internal loss data, external loss data and scenario analysis. Incorporation of these elements plays a crucial role for the capital requirement estimation in operational risk. The fourth data element, business environment and internal control factors, is treated as an ex-post update to capital calculations and is out-of-scope for this thesis.

Extreme value approaches have gained popularity in the area of operational risk in recent years, with its focus on the behaviour of processes at extreme levels making it a natural candidate for operational risk modelling. According to extreme value theory (EVT) a generalized Pareto distribution (GPD) is well suited for modelling extreme losses because it under assumptions represents the domain of attraction of independendent losses beyond a high-level treshold, called GPD-threshold. The adoption of EVT in operational risk modelling has encountered several obstacles with the main one being the scarcity of data and another being the identification of an appropriate GPD threshold. The combination of these two issues leads to substantial statistical uncertainty for both parameter and capital estimates. In order to combat the aforementioned data scarcity institutions use external data

and scenario analysis as complements to their internal loss data, with these two data sources leading to their own challenges.

In the last couple of years Bayesian Inference methods has gained interest in the field of operational risk modelling. Especially when calibrating the parameters of the loss distribution models used for calculating the capital requirement they seem to have some desirable properties. One publication aiming to be a reference in this subject is by Shevchenko [16], he describes the Loss Distribution Approach as well as some Bayesian Inference methods that could be applicable for operational risk. A paper by Ergashev et al. [7] more closely examines the use of Bayesian Inference methods and the challenges involved.

The aim of this thesis will be to investigate if the Bayesian Inference (BI) methods indeed outperforms the more traditional Maximum Likelihood estimator (MLE) when estimating the parameters of a GPD in small data sets. Also a study of a full loss distribution with a Lognormal body and GPD tail will be presented. In the end this model will be tested in a real world environment using actual loss data.

The analysis of this work shows that the BI method significantly outperforms MLE already when the data set is of the size 200 data points. It gives good results for the full loss distribution as well as reasonable results in the real world setting.

The background for the thesis, including it's real world context, is presented in Chapter 2. Chapter 3 describes some mathematical concepts used in the thesis. Chapter 4 presents the parameter estimation method using Bayesian Inference. Chapter 5 presents the obtained results and finally Chapter 6 concludes and provides a forward look.

Chapter 2

Background

2.1 Operational Risk

When banks allocate capital against capital losses it can be viewed as self-insurance. The three main categories that attract a capital charge in financial institutions are credit risk, market risk and operational risk. Of these, operational risk can be viewed as the newest one since it did not require any explicit capital allocation until recently. In the past operational risk was assumed to be implicitly covered by the capital allocation made for credit risk. The concept of operational risk is in general related to the way a firm (not just financial firms) operates rather than changes in the market or credit ratings.

According to Shevchenko [16], operational risk accounts for approximately 15-25 % of the total capital allocated in many banks and is the second largest risk after credit risk. Even though explicit capital allocation to operational risk is a relatively new matter for most banks the management of operational risk is not. It has always been important for the financial industry to prevent external fraud, internal fraud and processing errors, all of which are typical operational risk events. To give an illustration of operational risk processes the following example is presented:

Example 1: Consider a foreign exchange deal where the trader makes use of market inaccuracy by:

- buying USD 100 million for SEK 862 million (exchange rate: USD 1 = SEK 8.62)
- selling USD 100 million for SEK 862,087,221 (exchange rate: USD 1 = SEK 8.62087221)

hence making a profit of SEK 87,221. However, the banks back office makes a mistake and the settlement is delayed a few days which leads to penalties of SEK 100,000 to be paid by bank. In total, because of the operational error made, the bank records a loss on the deal of SEK 12,779.

This fictional example illustrates just one operational risk event that could lead to a small loss for the bank. However, the consequences of operational errors could be much more severe, as shown by the following real-world examples:

Example 2: ***The Barings Bank downfall*** From 1992 to 1994 trader Nick Leeson made unauthorized speculative trades in futures and options from the Barings Bank's Singapore office leading to losses that in the end of 1994 had reached £208 million. He was able to do so because he was in charge of both the trading and the back office settling in the office, roles usually divided between different people, enabling him to hide his losses in an error account. When he, in early 1995, tried to recuperate his losses by betting that the Japanese stock market would not change overnight disaster struck as the Kobe earthquake sent Asian markets plummeting. In the end losses approached an amount that was twice the Barings Bank's available trading capital and the bank was declared insolvent in February 1995. The inadequate separation of front and back office responsibilities that led to the downfall is an example of an operational error.

Example 3: ***Société Générale and Jérôme Kerviel*** Until early 2008 trader Jérôme Kerviel took unauthorized trading positions that eventually lost his employer, french bank Société Générale, €4.9 billion. Mr. Kerviel claims that high-ranking officers knew of his trades but ignored them as long as he was making money for the bank and Société Générale claims that he was acting entirely on his own and was able to do so because of his knowledge of the banks back-office systems. Whichever is right it must be seen as an operational risk error when the banks internal systems fails to detect trades larger than the bank's entire market capitalization. The banking supervising authority in France fined SocGen €4 million in 2008 for their laxity. [17]

Since the financial crisis, regulators in many countries have been coming down hard on banks that fail to control risks like these properly. Other examples are the fine of £30 million paid by Swiss bank UBS for their inadequate controls when Kweku Adoboli, a trader in their London office, lost them $2.3 billion as well as JPMorgan Chase being fined around $1 billion for failures related to losses of $6 billion made by trader Bruno Iksil.[17]

2.2 The Definition of Operational Risk

Before the Basel Committee on Banking Supervision (BCBS) in 2001 issued the proposal for what we today refer to as Basel II there was no widely accepted definition of operational risk. It was mostly seen as anything that was not credit or market risk. Now, in the Basel II framework [1, p. 144], the following explanation is used:

Definition: Operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk.

2.3 Quantifying Operational Risk under Basel II

As mentioned in the introduction Basel II allows for three different approaches to quantify the operational risk capital requirement:

- The Basic Indicator Approach
- The Standardized Approach
- The Advanced Measurement Approach (AMA)

Of which the AMA is the approach that allows the bank to develop an internal model for calculating the capital charge and hence is the sole approach considered in this thesis. Under AMA the bank need to satisfy several criteria before being granted approval by the authority, including:

- The model should include internal and external data as well as scenario analysis and factors reflecting the business environment and internal control systems
- The capital requirement should be calculated as the 99.9 % confidence level for a holding period of one year.
- Diversification benefits will be accepted if the dependence modelling is approved

To get approval the bank should demonstrate that the model accurately describes its operational risk exposure in all cells of the Basel II matrix, presented in tables tables 2.1, 2.2 & 2.3, where losses are divided into eight business lines and seven event types.

The requirements under Basel II for the AMA leads to several challenges when modelling operational risk. The use of three different quantitative data sources are meant to give the modeller as much data as possible to work with, however the data sources come with different characteristics and the implementation of all three into the same model can be challenging. Internal data describes the bank's own risk profile while the external data represents the risk profile of the industry as a whole. While the bank operates within this industry it is not given that it follows the industry risk profile, more often it will not. Scenario data on the other hand represents a forward view on possible extreme losses and as such must be treated with care. The fourth factor of business environment and internal control systems is treated as en ex-post update to the internal risk model and will not be considered in this thesis.

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

Table 2.1: Basel II business lines (BL) [1, p. 147]

j	Event Type, ET(j)
1	Internal fraud
2	External fraud
3	Employment Practices and Workplace Safety
4	Client, Products & Business Practices
5	Damage to Physical Assets
6	Business disruption and system failures
7	Execution, Delivery & Process Management

Table 2.2: Basel II Event types (ET) [1, p. 305]

	ET(1)	ET(2)	$\cdots$	ET(j)	$\cdots$	ET(7)
BL(1)						
BL(2)						
$\vdots$						
BL(i)						
$\vdots$						
BL(8)						

Table 2.3: Basel II matrix.

2.4 Extreme Value Theory in Operational Risk

A few operational risk events are rare but have a major impact on the bank, so called *low-frequency/high-severity* risks. It is recognized that these operational risks have

heavy tailed distributions and, due to its simple fitting procedure, the lognormal distribution is a popular choice for modeling the severity distribution. However, due to the high quantile level requirement for operational risk capital charge, accurate modelling of extremely high losses (the tail of the severity distribution) is critical and it is often useful to use other heavy tailed distributions for the tail of the severity distribution. This is where Extreme Value Theory (EVT) comes into the picture. The tail of the severity distribution is often modelled using a Generalized Pareto Distribution (GPD) with the tail limited from below by a so called GPD-treshold. This gives a body of the severity distribution limited from above by the same treshold.

In other words, using this approach the body of the distribution consists of small losses that occur frequently and the tail of the distribution consists of large losses that occur infrequently.

2.5 Current situation

The limited amount of data available to operational risk modellers has lead them into challenges when using EVT because of its need for large tail-event samples [7]. However, while data sufficiency is a major obstacle faced when trying to fit a GPD to data, it is not the only one. Fitting a GPD requires the identification of an appropriate GPD-treshold, a task that is crucial as a misidentified treshold can greatly impact the parameter estimates. Some of the most used methods for treshold estimation, such as the Hill estimator or fixing a percentage for tail data, does not always give clear indications that the treshold has been selected appropriately. Many times they cause the tail sample to be "polluted" with non-tail data points.

These issues in combination leads to large uncertainty when estimating both the parameters and the capital requirement. Especially uncertainty in the shape parameter of the GPD, crucial for shaping the tail of the distribution, leads to substantial uncertainty in the capital estimates as illustrated by the example on the next page:

Example 4: The following picture illustrates how uncertainty in the shape parameter estimate of a GPD leads to uncertainty in capital estimates. All numbers are estimated on simulated data sets of a GPD with true shape parameter equal to 0.9 and true capital equal to 86 MSEK.

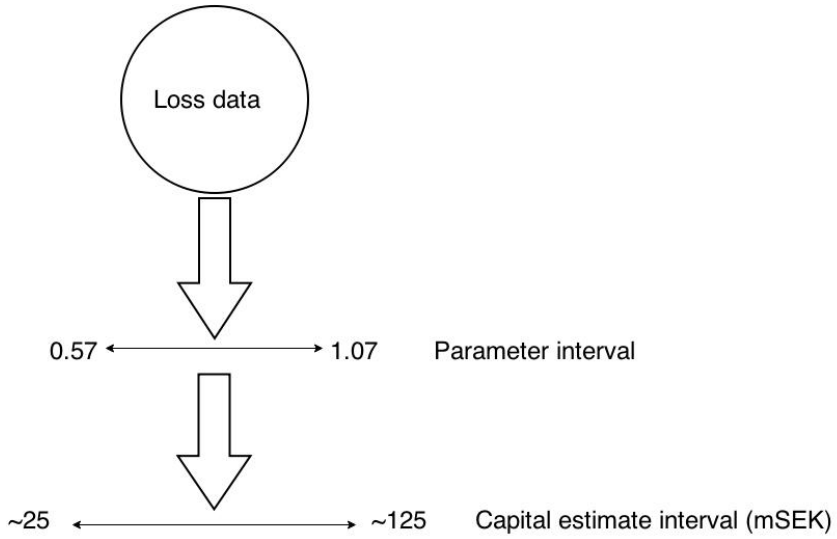


Figure 2.1: Illustration of the parameter uncertainty's effect on capital estimation.

As can be seen the estimated parameter interval leads to an interval for possible capital estimates of $\sim$ 100 MSEK, i.e. larger than the true value of the capital. This uncertainty is something the business wants to reduce to ensure that it properly insures itself against operational risks. Hence, reducing the uncertainty of parameter estimates will be the main focus of the thesis.

Chapter 3

Mathematical background

3.1 Loss Distribution Approach (LDA)

One of the most common ways to create a model that fulfills the requirements of the Advanced Measruement Approach (AMA) is the Loss Distribution Approach. With this modeling technique the modeler splits all operational losses into homogeneous segments, e.g. the Basel Matrix (see table 2.3), and for each segment a loss distribution is created. This distribution should represent the expectation of total losses that can occur with a one-year time horizon. Using the notation presented in [16] the LDA model can be written as:

$$Z_t = \sum_{j=1}^{J} Z_t^{(j)}; \quad Z_t^{(j)} = \sum_{i=1}^{N_t^{(j)}} X_i^{(j)}(t) \tag{3.1}$$

The following notations are present in the above equations:

- Z_t is the annual loss
- t = 1,2,... represents discrete time counted in annual units
- $Z_t^{(j)}$ is the annual loss in risk cell j which is modelled as a compound loss over one year with *frequency*, $N_t^{(j)}$, which is given by a e.g. Poisson process, and *severities*, $X_i^{(j)}(t), i = 1, ..., N_t^{(j)}$
- the most usual approach is to model the frequencies and severities by independent random variables

Under this model the capital the bank should hold is defined as the 0.999 Value at Risk (VaR) which is the quantile of the distribution for the next year annual loss Z_{T+1}:

$$\mathrm{VaR}_q[Z_{T+1}] = \inf\{z \in \mathbf{R} : \Pr[Z_{T+1} > z] \leq 1 - q\} \tag{3.2}$$

at the level $q = 0.999$.

The focus of this thesis is using the Bayesian Inference Method, described next, to estimate the probability distribution in one cell of the Basel Matrix. I.e. the focus of the thesis is not to create the full loss distribution of the bank, Z_t, but rather to determine if the BI Method gives a good estimation of the distribution in one cell, $Z_t^{(j)}$. If the method works well it could be applied to all cells and then the full loss distribution could be estimated from these cells.

3.2 Bayesian Inference Method

Bayesian inference is a statistical method of inference in which Bayes' theorem (first presented in *An Essay towards solving a Problem in the Doctrine of Chances* (1763) [2]) is used to update the probability estimate of a proposition as additional information becomes available. The initial degree of confidence is called the prior and the updated degree of confidence is called the posterior.

Let us consider a random vector of loss data $\mathbf{X} = (X_1, ..., X_n)$ who has a joint density for a given vector of parameters $\phi = (\phi_1, ..., \phi_K)$, which we denote $h(\mathbf{x}|\phi)$. In the Bayesian approach, both parameters and observations are considered random. Then their joint density is

$$h(\mathbf{x}, \phi) = h(\mathbf{x}|\phi)\pi(\phi) = \pi(\phi|\mathbf{x})h(\mathbf{x}) \tag{3.3}$$

where $\pi(\phi)$ is the probability density of the parameters, known as the prior density function. $\pi(\phi)$ will typically depend on a set of further parameters, known as *hyper-parameters*, which is omitted for simplicity of notation. These hyper-parameters are used in the densities used as prior distributions. $\pi(\phi|\mathbf{x})$ is the density of parameters given data $\mathbf{X}$, known as the posterior density, $h(\mathbf{x}, \phi)$ is the joint density of observed data and parameters and $h(\mathbf{x}|\phi)$ is the density of observations for given parameters. This is the same as a likelihood function if you consider it a function of ϕ, i.e. $l_{\mathbf{X}(\phi)} = h(\mathbf{x}|\phi)$. $h(\mathbf{x})$ is the marginal density of $\mathbf{X}$. If $\pi(\phi)$ is continuous then it can be written as

$$h(\mathbf{x}) = \int h(\mathbf{x}|\phi)\pi(\phi)d\phi \tag{3.4}$$

if $\pi(\phi)$ is discrete then the integration should be replaced by a corresponding summation.

Using equation (3.3), the Bayes' theorem, says that: *The posterior density can be calculated as:*

$$\pi(\phi|\mathbf{x}) = \frac{1}{h(\mathbf{x})}h(\mathbf{x}|\phi)\pi(\phi) \tag{3.5}$$

Here, $h(\mathbf{x})$ plays the role of normalisation constant and hence the posterior distribution can be viewed as a combination of prior knowledge (contained in $\pi(\phi)$) with information from the data (contained in $h(\mathbf{x}|\phi)$).
Given that $h(\mathbf{x})$ is a normalisation constant, we often write the posterior as

$$\pi(\phi|\mathbf{x}) \propto h(\mathbf{x}|\phi)\pi(\phi) \tag{3.6}$$

The Bayesian inference approach permits a reliable estimation of distributions' parameters even if the quantity of data, denoted n, is limited. When n becomes larger, the weight of the likelihood component increases such that the posterior distribution tends to the likelihood function if $n \to \infty$, as a consequence the parameters obtained from both approaches converge. A useful result of this is that the data selected to inform the likelihood component may lead the model and, as a consequence, the capital charge.

One of the papers considered for this master thesis [11] proposes a two step Bayesian Inference approach in order to obtain the parameters of the statistical distribution used to characterize the severity. Scenarios are used to build the prior distributions of the parameters ($\pi(\phi)$), which is refined using the external data as informant of the likelihood component. This results in an initial posterior function ($\pi(\phi|Y)$) which is then used as a prior distribution in the next step where the likelihood component is informed by internal data. This leads to a second posterior distribution which will allow for estimation of the parameters of the severity distribution used in building the loss distribution function in the LDA model.

3.3 Markov Chain Monte Carlo (MCMC)

When modeling the Loss Distribution the target densities will usually not be standard densities and hence Markov Chain Monte Carlo methods are useful for sampling the parameters of the severity distribution. Another paper considered for this project [7] suggests that Metropolis-Hastings (MH) algorithm is well suited for producing samples from a given target density. The target density of a parameter (or set of parameters) is the full conditional density of that parameter (set) conditioned on the rest of the parameters of the model. Bayes' theorem implies that this full conditional density is equivalent to the posterior density. By construction, each sample from the MH algorithm [1] constitutes a Markov chain of dependent draws. The algorithm is based on a proposal density that generates a proposal value and a probability of move used to determine whether the proposal value should be taken as the next draw from the target density. If the proposal value is rejected, the last draw of the chain is retained as the next draw. Here we present a short introduction to the two constituent parts of MCMC methods, namely Monte Carlo integration

[1] The MH algorithm used for this thesis as well as its implementation will be further explained in chapter 4.

and Markov Chains. For further information about MCMC methods the reader is referred to Gilks et al. *Markov Chain Monte Carlo in practice* [10].

3.3.1 Monte Carlo integration

Let X be a vector of k random variables, with distribution $\pi(.)$. The task of Bayesian Inference is to evaluate the expectation

$$E[f(X))] = \int f(x)\pi(x)dx \tag{3.7}$$

for some function, f.

Monte Carlo integration evaluates $E[f(X))]$ by drawing samples, $\{X_t, t = 1, ..., n\}$, from the distribution $\pi(.)$ and the using these samples to approximate

$$E[f(X))] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t) \tag{3.8}$$

which means that the population mean of the function, $f(X)$, is estimated by a sample mean. If the samples are independent, the law of large numbers will ensure that the approximation is accurate by increasing the sample size, n, sufficiently.

Drawing samples independently from $\pi(.)$ is usually not an easy task as this density seldom is a standard density. The independence is, however, not necessary. The X_t's can be generated by any process that samples throughout the support of π in correct proportions. A *Markov chain* having π as its stationary distribution is one possibility for doing this. This is what is called *Markov Chain Monte Carlo*, MCMC.

3.3.2 Markov chains

A Markov chain is a discrete time stochastic process $\{X_0, X_1, ..., X_{t+1}, ...\}$ where at each time, $t \geq 0$, the next state X_{t+1} only depends on the current state, X_t. I.e., given X_t, the next state X_{t+1} does not depend further on the history of the chain. Mathematically this property can be written as

$$P(X_{t+1} \in A|X_0, X_1, ..., X_t) = P(X_{t+1} \in A|X_t) \tag{3.9}$$

for any set A. This distribution, $P(.|.)$, is called the *transition kernel* of the chain. Under certain regularity conditions the Markov chain will gradually 'forget' its initial state and eventually converge to a unique *stationary* distribution, which does not depend on either time t or the starting point X_0.

In the notation of the Monte Carlo integration section we have that after a sufficiently long *burn-in period* of say m iterations, the points $\{X_t, t = m+1, ..., n\}$ will be dependent samples approximately coming from the wanted distribution π. Then

the output of the Markov chain can be used to estimate the expectation $E[f(X))]$, with X having π as its distribution. After discarding the burn-in samples we get the following estimator:

$$E[f(X)] \approx \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t) \tag{3.10}$$

Chapter 4

Methodology

This chapter starts with section 4.1 presenting two common examples of methods to achieve Markov Chain Monte Carlo sampling as well as motivating the choice of method for this thesis work. Section 4.2 describes in detail the implementation of this method and section 4.3 describes a two-step procedure used to incorporate both external and internal data in the model. Section 4.4 shortly explains peaks over treshold and specifies the full loss distribution evaluated in the thesis.

4.1 Examples of Markov Chain Monte Carlo methods

There exists a large amount of methods for achieving Markov Chain Monte Carlo (MCMC) sampling when trying to estimate parameters of a model, however the two most commonly used seems to be the Metropolis-Hastings algorithm and the Gibbs sampler. Both W.R. Gilks [10] and D. Gamerman [8], two books on the subject of MCMC simulation, mainly describes these two algorithms in their work. An introduction to both methods as well as arguments for and against them are presented below.

4.1.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is an (almost) universal algorithm that creates a Markov chain with a stationary distribution for the parameters given the data set, i.e. $\pi(\theta|\mathbf{x})$ (see equation (3.5)), when direct sampling is difficult. It was developed by Metropolis et al. [15], for use in mechanical physics, and then generalized by Hastings [12] for a more statistical setting.

Before introducing the MH algorithm in full let us look at the closely related Acceptance-Rejection Sampling (AR), a classical simulation technique that generates non-Markovian and (usually) independent samples.

Say that there exists an absolutely continuous *target density*, $\pi(x) = f(x)/K$, where

$f(x)$ is the unnormalized density and K is the (possibly unknown) normalizing constant. Now, samples should be generated from this density $\pi(x)$.

Let $q(x)$ be a density that can be simulated by some known method, and suppose there exits a constant c for which $f(x) \leq cq(x)$ for all x. Then, to obtain a random variate from $\pi(x)$:

Algorithm: A-R Sampling
1. Generate a candidate X from $q(\cdot)$ and a value u from the uniform distribution, $\mathcal{U}(0,1)$, on the interval $(0,1)$.
2. If $u \leq f(X)/cq(X)$:
- return X = y.
Else:
- go to 1.

Now let us turn our focus towards the MH algorithm. As in the A-R method, suppose there exists a density that can generate candidates. Since the MH algorithm wants to create a Markov chain, the density is allowed to depend upon the current state of the process. The *candidate-generating density* or *proposal density* is denoted $q(Y|X_t)$. This density can be interpreted as saying that when the process is at a state X_t, the density generates a value Y from $q(Y|X_t)$. In order for the chain to be reversible, it can be shown (see e.g. [4, p. 329]) that the probability of move must be set to

$$\alpha(X_t, Y) = \min\left(1, \frac{\pi(Y|\mathbf{x})q(X_t|Y)}{\pi(X_t|\mathbf{x})q(Y|X_t)}\right). \tag{4.1}$$

If the candidate is accepted, the next state is $X_{t+1} = Y$. Otherwise, the next state is $X_{t+1} = X_t$. Schematically, the MH algorithm looks like:

Algorithm: MH Algorithm
1. Initialize algorithm with arbitrary value X_0
2. For $i = 1, \cdots, N$ do:
- Generate Y from $q(\cdot\|X_i)$ and u from $\mathcal{U}(0,1)$.
If $u \leq \alpha(X_i, Y)$:
- set $X_{i+1} = Y$
Else:
- set $X_{i+1} = X_i$
3. Return the obtained values: $\{X_1, X_2, \cdots, X_N\}$

Some remarks need to be made about the MH algorithm:

- The MH algorithm is specified by the candidate-generating proposal density, however this can have any form and the stationary distribution of the Markov chain will still be $\pi(\theta|\mathbf{x})$. For a justification of this please refer to [10, p. 7].
- As the calculation of the acceptance probability $\alpha(X_t, Y)$ is a fraction of two posterior densities, the normalizing constant $h(\mathbf{x})$ of eq. (3.5) is not needed as it will cancel.
- If the candidate generating density is symmetric the probability of acceptance reduces to $\pi(Y|\mathbf{x})/\pi(X|\mathbf{x})$. Hence if $\pi(Y|\mathbf{x}) \geq \pi(X|\mathbf{x})$ the chain moves to Y; otherwise, it will move with probability $\pi(Y|\mathbf{x})/\pi(X|\mathbf{x})$. I.e., if the jump is "uphill", it is always accepted, if "downhill" it is accepted with a non-zero probability.

The selection of the proposal density is crucial for the success of the method. On one hand one needs a not-so-low acceptance rate and on the other hand one needs a good mixing of samples. This leads to a trade-off between selecting a proposal density that allows for small steps and high acceptance rates against a density that gives large steps and good mixing of samples but a lower acceptance rate.

It is also crucial that the proposal density is easy to sample from as the whole point of the algorithm is to switch from sampling from the difficult density $\pi(\theta|\mathbf{x})$ to many generations from $q(Y|X_t)$.

All in all the MH algorithm is by far the most general way of generating a Markov Chain that samples from a known distribution and is applicable to a wide range of problems. It was also found to be the algorithm of choice for this thesis and its implementation will be presented in section 4.2.

4.1.2 Gibbs sampling

Gibbs sampling is a Markov Chain Monte Carlo algorithm for generating random variables for a distribution indirectly, without having to calculate the density. The name originates from Gibbs random fields in image-processing starting with Geman and Geman [9] in 1984.

The Gibbs sampler requires the knowledge of the full conditional distributions of the parameters and can be thought of as a practical implementation of the fact that the knowledge of these distributions is sufficient to determine a joint distribution. Let us consider a random vector $\mathbf{X}$ that has a joint density $f(\mathbf{x})$ and denote the full conditional distributions by $f_i(x_i|\mathbf{x}_{-i})$. The sampler then does the following steps:

1. Initialize $x_2^{l=0}, \cdots, x_N^{l=0}$ with some arbitrary values

2. For $l = 1, \cdots, L$:
   1) simulate x_1^l from $f_1(x_1|x_2^{l-1}, \cdots, x_N^{l-1})$
   2) simulate x_2^l from $f_2(x_2|x_1^l, x_3^{l-1} \cdots, x_N^{l-1})$
   $\vdots$
   N) simulate x_N^l from $f_N(x_N|x_1^l, \cdots, x_{N-1}^l)$

3. Do an increment, $l = l + 1$, and return to step 2

Under some quite general conditions $f(\mathbf{x})$ is a stationary distribution of the chain generated by this algorithm; and the chain being ergodic with a limiting distribution $f(\mathbf{x})$, i.e. the distribution of x^l converges to $f(x)$ for large l.

One concern with the Gibbs sampler is that it may have slow convergence and limited possibilities to control the convergence rate. This may lead to unnecessary long computation time required to reach convergence, i.e. an inefficient algorithm. Also Gibbs sampling does not take into account the previous value of the component being updated and is therefore seen as somewhat restrictive. The requirement of knowledge of the full conditional distributions may also cause some problems as they are not always easily obtained.

In general, choosing between Metropolis-Hastings or Gibbs is not a question of which is better, instead it is more a question of which one is more to your liking. In fact, MH and Gibbs are often used in combination with each other. In this thesis MH was chosen as it seemed easier to implement and that the ability to control convergence was seen as important.

4.2 Implementation of Metropolis-Hastings algorithm

When describing the MH algorithm used for this thesis it is important to keep in mind the equation that expresses the posterior distribution (i.e. the target distribution of the sampling) as proportional to the prior distribution multiplied by a likelihood component:

$$\pi(\theta|\mathbf{x}) \propto h(\mathbf{x}|\theta)\pi(\theta) \tag{4.2}$$

where $\mathbf{x}$ corresponds to the data sample evaluated. This equation is included rather than the one with the normalization constant (eq. (3.5)) as the normalization constant cancels in the calculation of probability of move in the MH algorithm (see eq. (4.1)).

To explain the implementation of MH algorithm a model that was analysed in the thesis work is used as a base. In this model the random variables X follows a Generalized Pareto distribution with treshold parameter equal to zero, i.e. $X \sim$

$\mathcal{GPD}(\xi, \beta, \tau = 0)$, hence the parameters that will be estimated are the scale, ξ, and shape, β, parameters. The algorithm looks as follows:

Algorithm: MH for GP distributed data

1. Initialize $\beta_{(i=0)}$ and $\xi_{i=0}$ within the support of $\pi(\beta|\mathbf{x})$ and $\pi(\xi|\mathbf{x})$

2. For $i = 1, ..., N$ do:

- Set $\beta_{(i)} = \beta_{(i-1)}$
- Generate proposal β_* from $q(\beta_*|\beta_{(i)})$
- Accept proposal with probability

$$\alpha(\beta_{(i)}, \beta_*) = \min\left\{1, \frac{\pi(\beta_*|\mathbf{x}, \xi_{(i-1)})q(\beta_{(i)}|\beta_*)}{\pi(\beta_{(i)}|\mathbf{x}, \xi_{(i-1)})q(\beta_*|\beta_{(i)})}\right\} \tag{4.3}$$

- i.e. generate Z from uniform dist. $\mathcal{U}(0, 1)$ and set $\beta_{(i)} = \beta_*$ if $\alpha(\beta_{(i)}, \beta_*) > Z$
- Set $\xi_{(i)} = \xi_{(i-1)}$
- Generate proposal ξ_* from $q(\xi_*|\xi_{(i)})$
- Accept proposal with probability

$$\alpha(\xi_{(i)}, \xi_*) = \min\left\{1, \frac{\pi(\xi_*|\mathbf{x}, \beta_i)q(\xi_{(i)}|\xi_*)}{\pi(\xi_{(i)}|\mathbf{x}, \beta_i)q(\xi_*|\xi_{(i)})}\right\} \tag{4.4}$$

- i.e. generate Z from uniform dist. $\mathcal{U}(0, 1)$ and set $\xi_{(i)} = \xi_*$ if $\alpha(\xi_{(i)}, \xi_*) > Z$

3. Do an increment, $i = i + 1$, return to step 2

The crucial part of this algorithm is the choice of the *proposal* density, i.e. the density denoted by $q(\cdot|\cdot)$ in the algorithm. The density used in this work is usually referred to as a tailored proposal density, and is specific for each block of variables simulated. The next section explains it thoroughly:

4.2.1 Tailoring the proposal density

The idea of the tailored proposal density can be described like this: if you need to have a good acceptance rate as well as a good mixing of parameters in your sample (within the support of the distribution) your proposal density should be similar in shape and location to your target distribution. Chib and Greenberg [4] suggests this approach as one possibility for choosing a proposal generating density and

also states that this leads to an *independence chain* as the proposals are generated independently of the current location of the chain.

The most typical choice of a proposal density is a density that creates a so called *random walk chain*. In this case the candidate Y is drawn according to the process $Y = X + Z$ where Z is called the increment random variable and follows a distribution q_1. The name random walk chain comes from the fact that the candidate is equal to the current value plus some noise introduced by Z. To show how the tailored proposal density improves upon this consider figure 4.1:

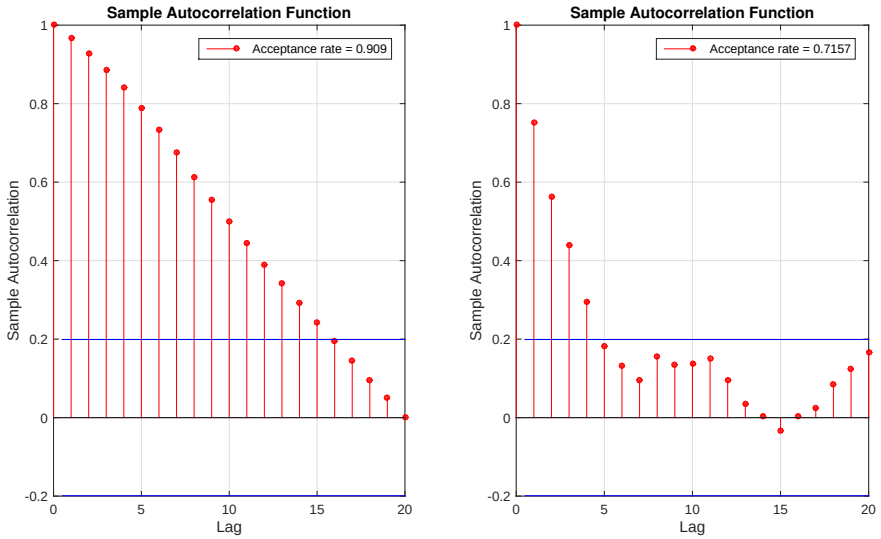


Figure 4.1: Autocorrelation functions of (left) a MH algorithm run with a *random walk chain* proposal density, (right) a MH algorithm run with a *tailored proposal density*. The acceptance rates of the random walk chain was 0.909 and of the tailored proposal density it was 0.7157.

Slowly decaying autocorrelation function indicates that the algorithm has a low mixing of parameters, as is evident in figure 4.1 the autocorrelation function of the tailored proposal density decays much faster than the autocorrelation function of the random walk chain. The acceptance rate of the tailored proposal density is still quite high, so the tailored proposal density gives a good mixing of parameters while at the same time maintaining a high acceptance rate making it more effective when trying to generate the true probability distribution of the parameters.

Consider the picture in figure 4.2. It shows the plotted likelihood function of the shape parameter, ξ, from a generated data set with true parameter value $\xi = 0.7$. The tailored proposal density wants to approximate this shape in order to be able to generate a good sample of possible parameters. There are numerous ways of

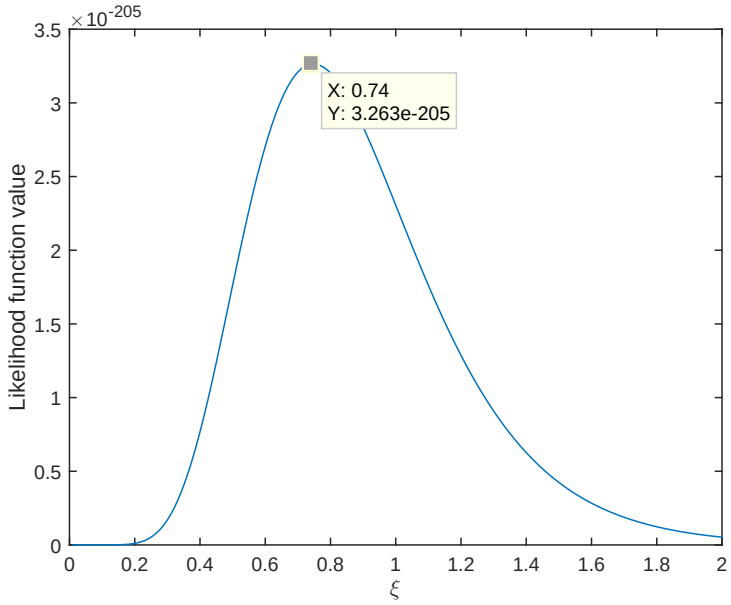


Figure 4.2: Likelihood plot of the shape parameter, ξ, of the GPD for a simulated data set with true value $\xi = 0.7$.

achieving this but to keep the sample generation as simple as possible a normal density is chosen as the proposal density.

The next issue is how to specify the parameters of the normal density. Since we want it to have the same mode as the target density the mode of the likelihood function is a natural candidate. This value is found by using a search heuristic called *simulated annealing*. The next parameter we should specify is the standard deviation of our proposal normal density, this is done by approximating the curvature of the likelihood function at the mode and then calculating the standard deviation from this. The next part explains both procedures in more detail.

Searching for the mode, simulated annealing

As previously mentioned, one challenge of this approach is the search for the mode. Although the target density of figure 4.2 only has one mode the possibility exists that, with complex target densities, we can have several local modes in the density function (see figure 4.3). Then many standard optimization strategies may get "caught" in these local modes which are not necessarily the global mode. To overcome this challenge we use an optimization strategy called *simulated annealing*.

Simulated annealing (SA) is an optimization heuristic with a probabilistic acceptance criteria that allows for moves towards both worse and better positions (com-

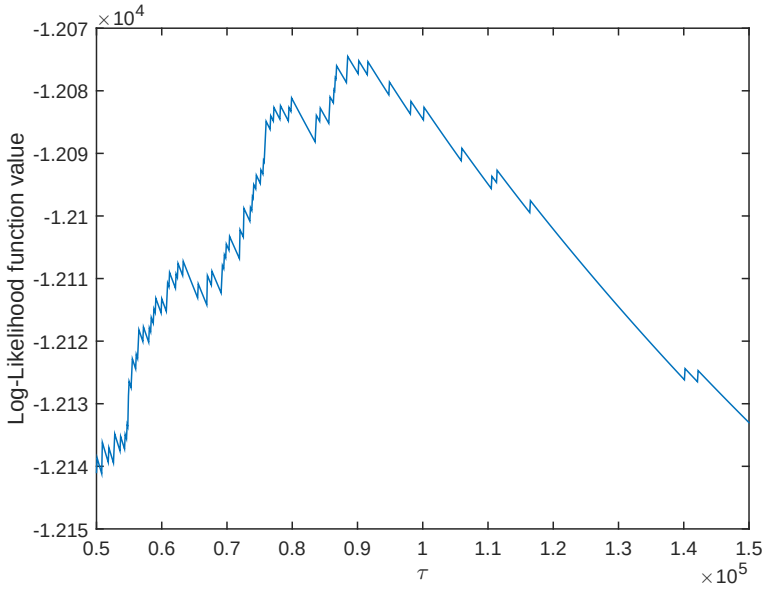


Figure 4.3: Log-Likelihood plot of the treshold parameter, τ, of the GPD for a simulated data set with true value $\tau = 10^5$. Note the many local optima of the function.

pared to what is the optimal position), which was first described for optimization problems by Kirkpatrick in 1983 [14]. The method iteratively suggests new, slight and random modifications to a current solution and hence moves gradually through the search space. In order to escape local optima, which is the crucial part of SA, the algorithm will accept modifications not only for the better, but also for the worse. The probabilistic acceptance criterion is used to decide whether or not to accept a worse move, but the probability of doing so declines over time according to a so-called cooling schedule, hence the method will eventually converge.

The simulated annealing algorithm used is inspired by [6] and described as follows:

Algorithm: Simulated annealing
1. Initialize parameters: - $\mathbf{\Theta}_{old} = \mathbf{\Theta}_0$, initial point - $t = t_0 > 0$, initial temperature - $g_{old} = g(\mathbf{\Theta}_{old})$, initial value of objective function 2. Main loop - Propose new move as: $\mathbf{\Theta}_{new} = \mathbf{\Theta}_{old} + c \times \mathbf{e} \times z$ - Compute new value of objective function: $g_{new} = g(\mathbf{\Theta}_{new})$ If $g_{old} - g_{new} > 0$, accept new value else accept move with probability $p = e^{(g_{old} - g_{new})/t}$ - Reduce the temperature according to $t = \nu t$ - Return to first point of main loop 3. When temperature reaches predetermined "floor", stop and return $\mathbf{\Theta}$ and g_{old}
Note: In the proposal of the new move c is a proportionality constant, $\mathbf{e}$ is a unit vector with a one in the position of the parameter for which the mode is being calculated and z is a $\mathcal{N}(0,1)$-distributed r.v.

Calculation of curvature and standard deviation

When the mode is found the standard deviation of the proposal normal density is estimated at the mode using the likelihood function. This is done by first calculating the approximative curvature of the likelihood function at the mode. This curvature is approximated by the negative Hessian at the mode, i.e.

$$\kappa = - \left. \frac{\partial^2 \mathcal{L}(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_1^2} \right|_{\theta_1 = \theta_1^*} \tag{4.5}$$

where $\mathbf{x}$ is the vector of data points, $\boldsymbol{\theta}$ is the vector of parameters and θ_1^* is the calculated mode of parameter θ_1.

When the curvature is found it is subsequently used to calculate the standard deviation of the normal density. For a plane curve given the curvature is given by

$$\kappa(x) = \frac{|f''(x)|}{[1 + (f'(x))^2]^{3/2}} \tag{4.6}$$

For the normal density with probability distribution function given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.7}$$

we have the relevant derivatives as

$$f'(x) = \frac{-(x-\mu)e^{\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma^3\sqrt{2\pi}} \tag{4.8}$$

$$f''(x) = \frac{\frac{(x-\mu)^2 e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma^4} - \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma^2}}{\sigma\sqrt{2\pi}} \tag{4.9}$$

However, evaluating these at the mode, i.e. setting $x = \mu$, we get

$$f'(x) = 0 \tag{4.10}$$

$$f''(x) = \frac{\frac{-1}{\sigma^2}}{\sigma\sqrt{2\pi}} = \frac{-1}{\sigma^3\sqrt{2\pi}} \tag{4.11}$$

Using these results, eq. (4.6) becomes

$$\kappa(\mu) = \frac{1}{\sigma^3\sqrt{2\pi}} \tag{4.12}$$

and solving for σ, we get the standard deviation used in the tailored proposal density:

$$\sigma = \left(\frac{1}{\kappa\sqrt{2\pi}}\right)^{\frac{1}{3}} \tag{4.13}$$

Role in sampling

The two steps described above are performed at each step of the algorithm, hence the proposal density moves through the search space together with the parameter sample. When convergence is reached, i.e. when the optimal mode is found, the proposal density will more or less fix to the same value, with only small changes from step to step. Then it will start sampling parameter values around this optimal mode and the algorithm will decide which ones to accept or not.

This means that the normal density calculated as above serves as both proposal density and as a search function in the algorithm, "helping" the algorithm find the best values together with the posterior target density, $\pi(\boldsymbol{\theta}|\mathbf{x})$. Together with this the shape of the proposal density ensures that a good mixing of possible parameter

values are obtained. Figure 4.4 illustrates how the proposal density approximates the target density and it is clear that after convergence is reached the proposal density will investigate a large part of the support of the target density, which is a desired property of the proposals.

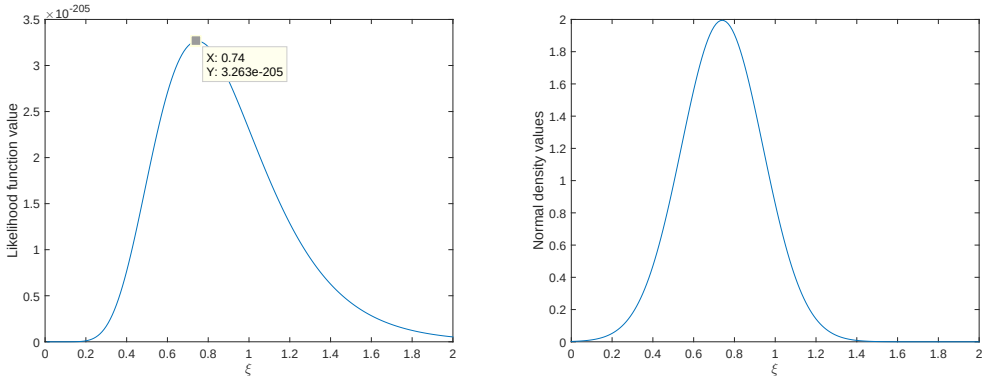


Figure 4.4: Left graph shows the plotted likelihoodfunction, right graph shows the normal proposal density used to sample values in Metropolis-Hastings algorithm. Note the similarity in shape and location of the two plots.

4.2.2 Analysing the output

When the algorithm is executed it will produce samples of the parameters we want to analyse. Before this analysis can start the part of the sample that has been produced before convergence was reached needs to be discarded. This part of the sample is usually referred to as the *burn-in period* and is illustrated in figure 4.5.

After the burn-in period is cut, several statistical measures is used to analyse the parameter sample. Mean, mode, standard deviation and other statistical measures such as upper and lower quantiles are calculated. These will then be compared to the performance of other parameter estimators, mainly against the Maximum Likelihood Estimator.

The upper and lower quantiles of the BI method are calculated empirically from the produced parameter sample in each case so that 95 % of the obtained values in the sample lies within these two quantiles.

The corresponding MLE values are obtained by first calculating the asymptotic variances of the maximum likelihood estimates, then taking the square root of these as standard deviations in a normal distribution and finding the lower and upper quantiles of this distribution corresponding to 95 % of the values being within these two quantiles. Hence, the two confidence intervals should be comparable.

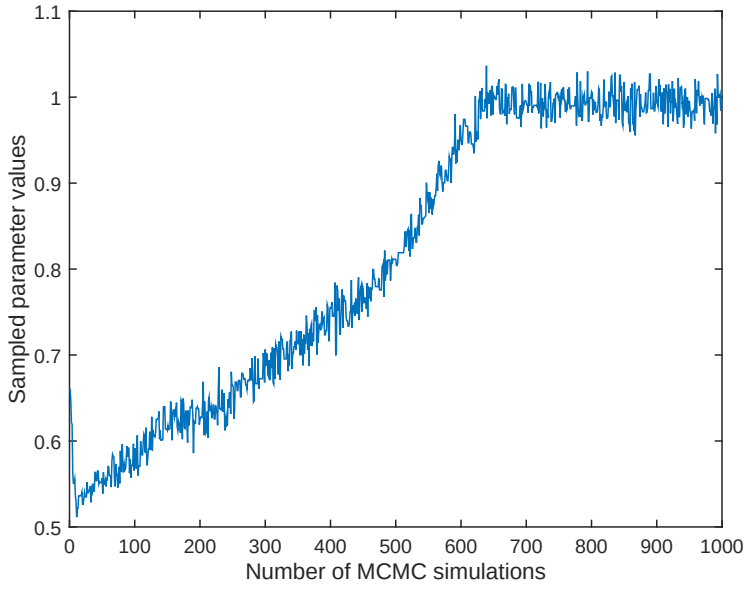


Figure 4.5: Example of burn-in period in Metropolis-Hastings sampling of parameter values. In this case the burn-in period would be set somewhere between 600 and 700 simulations and all values before this would be discarded.

The standard deviation of the MLE is calculated using the "68-95-99.7-rule", i.e. that the 95 % confidence interval roughly corresponds to two standard deviations on each side of the mean. Hence, the standard deviation of the MLE is calculated by solving for σ in equation (4.14).

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.95 \tag{4.14}$$

It could also be desirable to obtain an approximative distribution of the parameters. This is especially interesting when the sample is to be used as a prior distribution for another run of the MCMC algorithm and will be explained further in section 4.3.

Calculating the capital

In the end the number of true interest for the bank is the capital that needs to be allocated. Going back to section 2.3, it is stated that it should be calculated as the 99.9 % confidence level for a holding period of one year which translates to the one in a thousand year loss. The 99.9 % confidence level is calculated as the $\text{VaR}_{0.999}$, and to obtain the one year holding period an article by Böcker et al. [3] suggests the following analytical approximation

$$\mathrm{VaR}_\alpha(t) = F^{-1}\left(1 - \frac{1-\alpha}{E[N(t)]}\right), \quad \alpha \to 1 \tag{4.15}$$

where the expectation in the denominator of the fraction is the expected number of losses in time period t.

4.3 Two-step Bayesian Approach

The requirement to have both external and internal data in the model leads to challenges originating in the data sets' different characteristics. An approach designed to make the combination of these data more reliable is suggested by Hassani et al. [11] and is based on the construction of two successive posterior distribution functions. Using the Metropolis-Hastings algorithm described above the first posterior is obtained, which is then used as prior distribution in the second run of the algorithm to obtain the second posterior distribution.

Using the notation of section 3.2, with additions $\mathbf{y}$ as external data and $\mathbf{x}$ as internal data, the method uses two steps to produce the final distribution of parameters:

1. Take prior π_0 and let the likelihood component be informed by external data:

$$\pi_1(\phi|\mathbf{y}) \propto h(\mathbf{y}|\phi)\pi_0(\phi) \tag{4.16}$$

2. The posterior π_1 is then used as prior and the likelihood component is now informed by internal data:

$$\pi_2(\phi|\mathbf{x}) \propto h(\mathbf{x}|\phi)\pi_1(\phi) \tag{4.17}$$

The justification of this approach lies in the property that the Bayesian posterior distribution implies that the larger the quantity of data used, the larger the weight of the likelihood component. The consequence is, with N representing the number of data points in vector $\mathbf{x}$,

$$\pi(\phi|\mathbf{x}) \propto \pi(\phi)h(\mathbf{x}|\phi) \overset{N\to\infty}{\to} h(\mathbf{x}|\phi), \tag{4.18}$$

As a result the order of the Bayesian integration of the components (data) is significant. Due to this property, in the worst case, the second posterior distribution will be entirely driven by the internal data set. However, as the internal data sets usually are very small compared to the external data sets, a model that is somewhat driven by internal data will be obtained, but the external data will still have a significant impact in the results.

4.3.1 Identification of prior distributions

One of the crucial steps of the two-step Bayesian approach is the identification of prior distributions to use in the second step. Once the first step algorithm is run and the burn-in period is discarded what is left is the simulated parameter sample of the external data set, or the posterior distributions of the parameters. In the two-step approach they are used as prior distributions for a new run of the MH algorithm and hence we need to specify the distributions in terms of probability distribution functions to be able to quantify their impact on the probability of acceptance of new parameters in the second run.

Here an example of the analysis conducted when determining the prior distribution of a parameter for the second step of the approach is presented. Figure 4.6 presents a simulated parameter sample from the first run of the MH algorithm to which a probability distribution should be fitted.

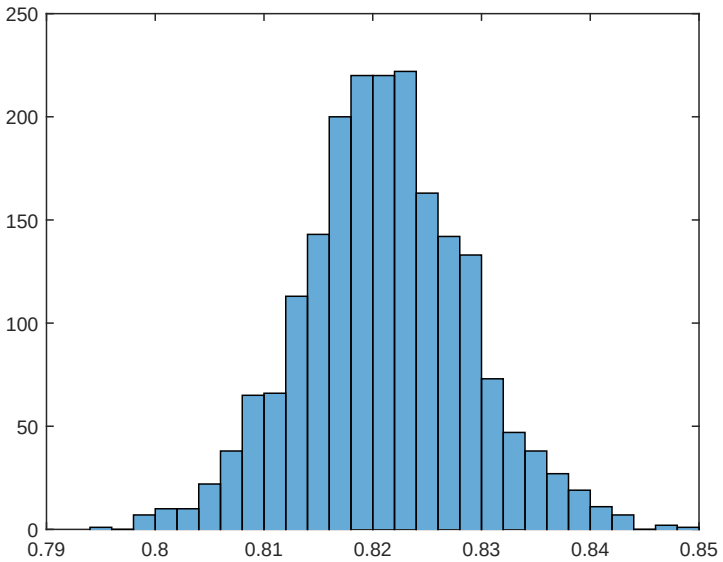


Figure 4.6: Example of parameter sample from one run of Metropolis-Hastings algorithm.

When trying to fit a distribution to a parameter sample some standard methods are used. Figure 4.7 shows the same parameter sample with the graph of a fitted normal density added, together with a QQ-plot of the same fitted density. The QQ plot indicates that the normal distribution slightly misses some of the tails of the sample, whereas this is not a bad fit it could be made better. One way of making the fit better is by using a kernel density estimation. Hassani et al. [11] suggests

using an Epanechnikov kernel for this purpose.

An Epanechnikov kernel is a kernel density estimator introduced by Epanechnikov [5] and is defined as

$$K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{|u|\leq 1} \tag{4.19}$$

The fitted Epanechnikov kernel together with the corresponding QQ plot is presented in figure 4.8.

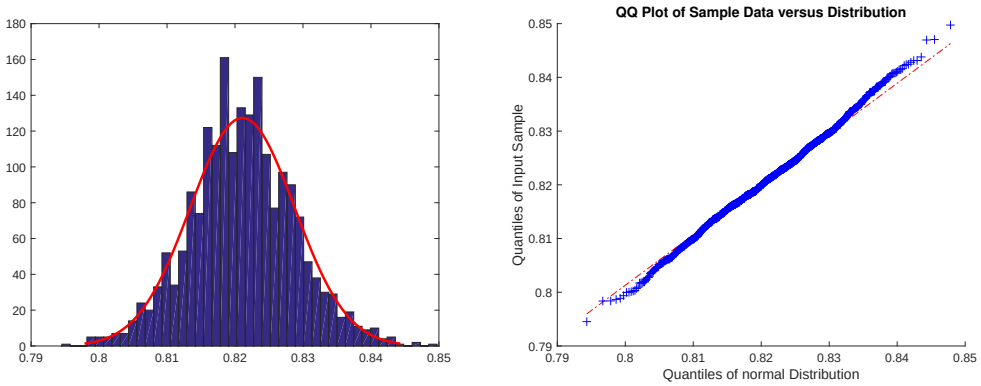


Figure 4.7: Left picture shows the example parameter sample with a fitted normal distribution, right picture shows a QQ plot of the fitted normal distribution versus the parameter sample.

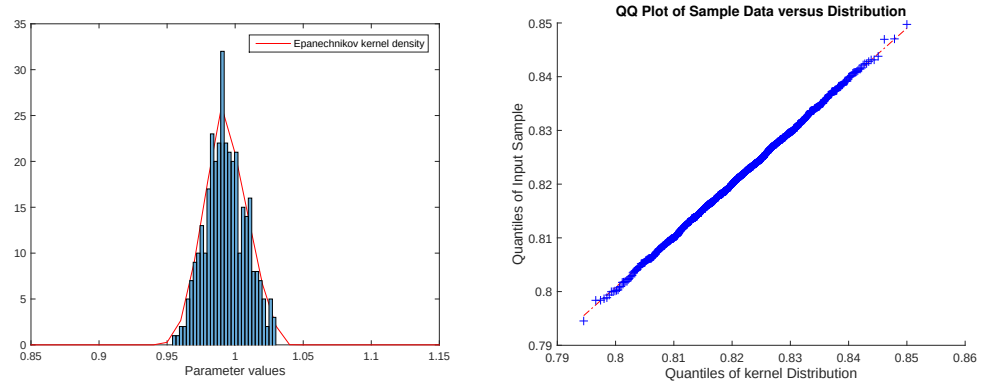


Figure 4.8: Left picture shows the example parameter sample with a fitted kernel density estimation, right picture shows a QQ plot of the fitted kernel density versus the parameter sample.

Hassani et al. [11] argues that the parametric solution may bias the construction of the densities as it requires fitting statistical to empirical distributions. They argue

that to stay as close as possible to the empirical distributions, and to the data, a non-parametric Kernel approach should be chosen. In addition, this study found that using the Kernel density estimation gives a more realistic parameter sample in the second run of the MH algorithm as illustrated by figure 4.9. The left sample in the first row is generated with a fitted normal distribution as prior, whereas the right sample in the first row was generated with an Epanechnikov kernel as prior. The sample on the second row is generated without any prior distribution for comparison. As can be seen the normal prior gives a second peak in the histogram concentrated around the mean of the fitted normal distribution. The Kernel prior instead gives a fatter tail on the side of the mean of the prior but a much smoother sample. This agrees with the intent to make the final distribution internal data driven, i.e. driven by the data added in the second step, but informed by the external data added in the first step.

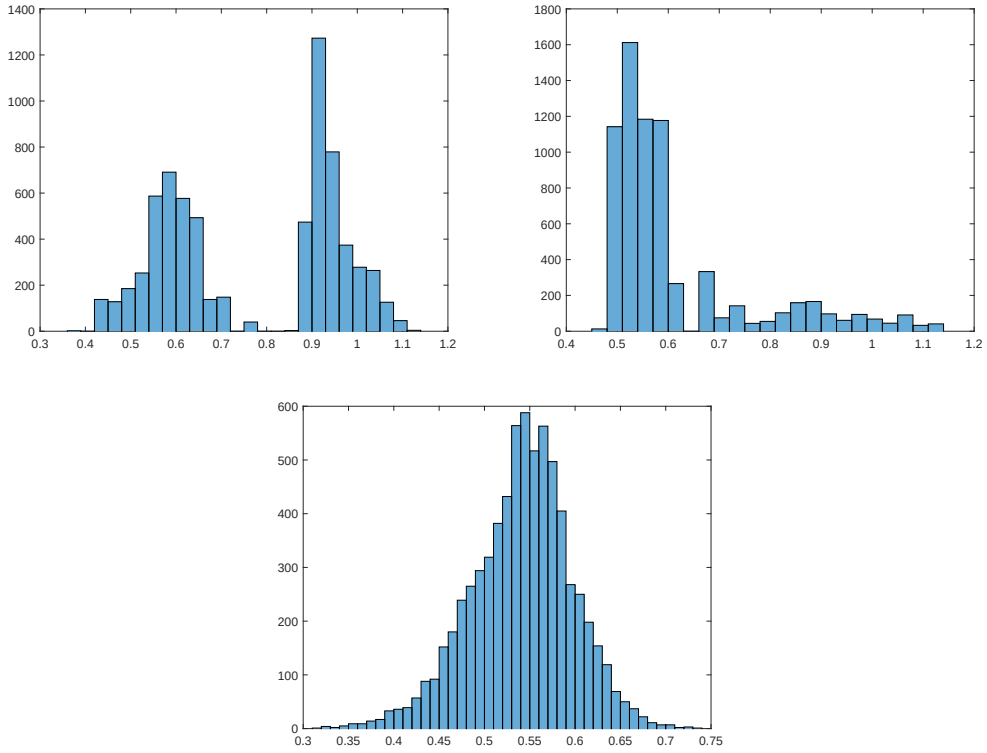


Figure 4.9: Comparison of resulting sample from second step of the two-step approach with different priors. Top left has a normal prior specified from the first step results, top right has a kernel density estimated and bottom has no prior distribution.

Hence, following this argumentation, the choice of prior for the second step of the

two-step approach is a Kernel density estimation with an Epanechnikov kernel.

4.4 Peaks over treshold

As mentioned in section 2.4, a few operational risk events are so called *low-frequency/ high-severity* risk. When modelling these the peaks over treshold method comes in useful. The method is a part of extreme value theory and lets the modeller extrapolate the tail of the distribution outside the range of the data sample, usually by appending another probability distribution above some high treshold, τ. A previous master thesis on the subject of operational risk [13] finds that the Generalized Pareto distribution gives the best fit of the data sample above a treshold. Hassani et al. [11] mentions that losses falling below the GPD treshold are commonly not modelled, but the inclusion of these losses in operational risk adds a positive value to the operational risk capital and therefore rigorous modelling should include these as well. They also argue that the choice of body distribution does not play a major role in the capital as long as the GPD treshold is accurate, hence, for its simplicity the Log-normal distribution is chosen as the body. The model will then be:

$$X \sim \begin{cases} \mathcal{LN}(\mu, \sigma) & \text{if } X \leq \tau \\ \mathcal{GPD}(\tau, \beta, \xi) & \text{if } X > \tau \end{cases} \tag{4.20}$$

and model like this will be estimated using the Bayesian Inference approach in this report.

Chapter 5

Results

5.1 Simulated data sets

The models evaluated in this part of the study were all tested on simulated data sets, i.e. the parameters of the model are fixed, simulate data samples from these models and try to recreate the parameter values using the methodology described in chapter 4.

5.1.1 GPD model

The first model to be evaluated was a Generalized Pareto Distribution with the following specifications:

$$X \sim \mathcal{GPD}(\xi = 0.7, \beta = 1 \times 10^6, \tau = 0) \tag{5.1}$$

I.e. the shape parameter, $\xi = 0.7$, the scale parameter, $\beta = 1 \times 10^6$ and the treshold parameter, τ, is zero and will not be estimated. Three data sets differing in size will be simulated to investigate the performance of the Bayesian Inference method relative to the Maximum Likelihood Estimator and try to determine when, in terms of number of data points, the BI method starts outperforming the MLE significantly.

The probability distribution function of the GPD, given parameters ξ, β and $\tau = 0$ is

$$f(x|\xi, \beta, \tau = 0) = \frac{1}{\beta}\left(1 + \xi\frac{x}{\beta}\right)^{-(1/\xi+1)} \tag{5.2}$$

hence, referring to eq. (4.2), the likelihood component for this setting is

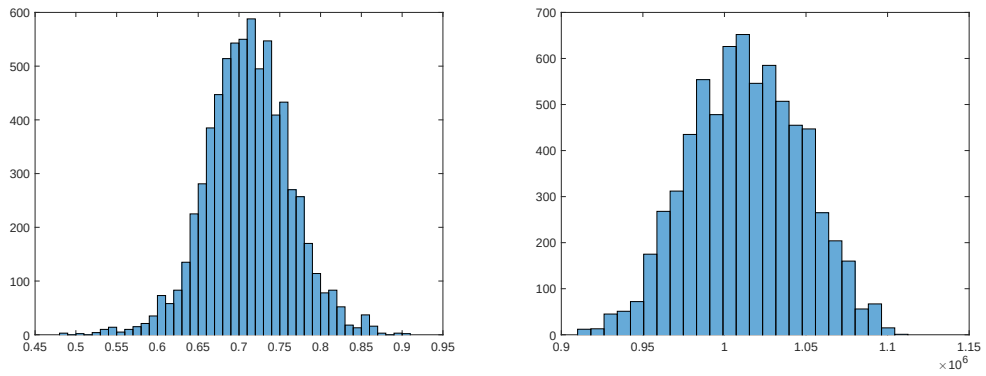
$$h(\mathbf{x}|\xi, \beta) = \prod_{i=1}^{N} \frac{1}{\beta}\left(1 + \xi\frac{x}{\beta}\right)^{-(1/\xi+1)} \tag{5.3}$$

where N is the number of simulated data points. The prior component of eq. (4.2) was set as wide uniform priors as the purpose of this part of the study is to compare BI with MLE. The effect of priors will be presented later.

The results will be presented with a table showing statistical measures of both MLE and BI and histograms of realized parameter samples to illustrate their distribution. In the end box plots for graphical comparison of all three data sets is shown.

Data set 1 - 1000 data points

Bayesian Inference					
Parameter	True Value	Mean	St.Dev.	Lower	Upper
ξ	0.7	0.712	0.0513	0.608	0.817
β	1×10^6	1.01×10^6	3.41×10^4	9.49×10^5	1.08×10^6
Maximum Likelihood Estimation					
	True Value	MLE	St.Dev	Lower	Upper
ξ	0.7	0.713	0.0530	0.607	0.819
β	1×10^6	1.01×10^6	5.30×10^4	9.04×10^5	1.14×10^6

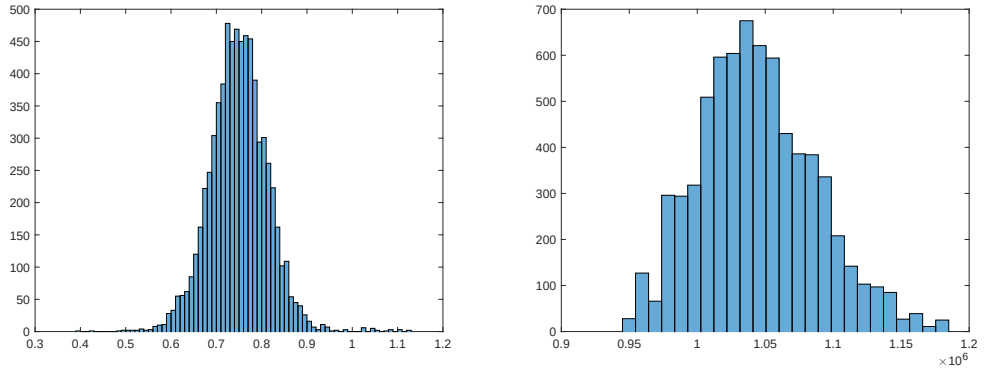


(a) ξ (b) β

Figure 5.1: Histograms of realized parameter values for ξ and β from the BI approach. Data set 1, 1000 data points

Data set 2 - 200 data points

Bayesian Inference					
Parameter	True Value	Mean	St.Dev.	Lower	Upper
ξ	0.7	0.749	0.0639	0.621	0.871
β	1×10^6	1.04×10^6	4.29×10^4	9.68×10^5	1.14×10^6
Maximum Likelihood Estimation					
	True Value	MLE	St.Dev	Lower	Upper
ξ	0.7	0.75	0.122	0.506	0.994
β	1×10^6	1.04×10^6	1.20×10^5	8.01×10^5	1.35×10^6

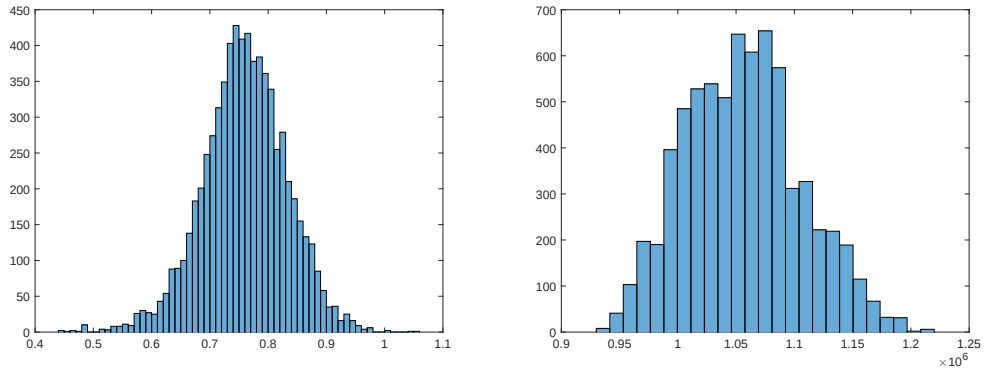


(a) ξ (b) β

Figure 5.2: Histograms of realized parameter values for ξ and β from the BI approach. Data set 2, 200 data points

Data set 3 - 50 data points

Bayesian Inference					
Parameter	True Value	Mean	St.Dev.	Lower	Upper
ξ	0.7	0.76	0.072	0.612	0.897
β	1×10^6	1.06×10^6	4.99×10^4	9.66×10^5	1.16×10^6§
Maximum Likelihood Estimation					
	True Value	MLE	St.Dev	Lower	Upper
ξ	0.7	0.761	0.2595	0.244	1.28
β	1×10^6	1.05×10^6	2.21×10^5	6.09×10^5	1.82×10^6



(a) ξ (b) β

Figure 5.3: Histograms of realized parameter values for ξ and β from the BI approach. Data set 3, 50 data points

Box plots of parameter quantiles

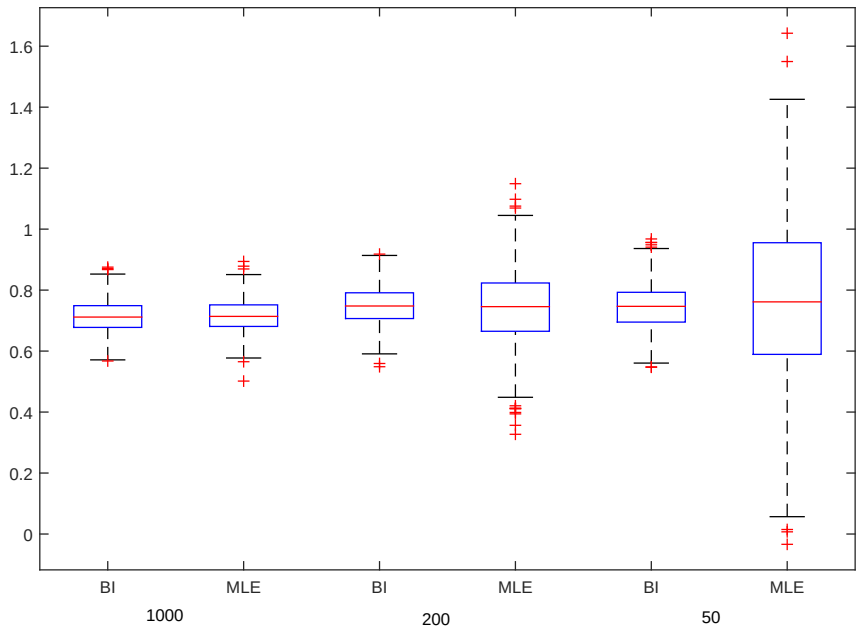


Figure 5.4: Box plot of ξ-quantiles from BI and MLE methods, grouped after size of data set.

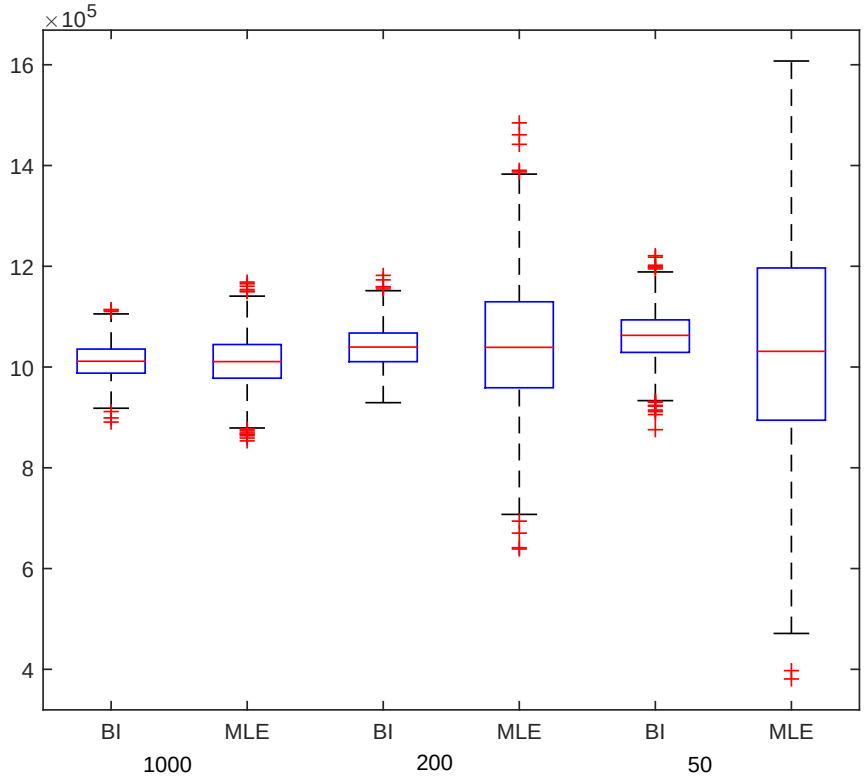


Figure 5.5: Box plot of β-quantiles from BI and MLE methods, grouped after size of data set.

As can be seen from both the tables of the statistical measures as well as from the box plots the BI method starts outperforming the MLE already at 200 data points. At 50 data points the BI method reduces the confidence interval by 72 % and 84 % for ξ and β respectively. It is also worth noting that the confidence intervals from the BI method stays slim as the data set size decreases, compared to those of the MLE which almost doubles from data set to data set. The parameter sample histograms indicate nice distributions, with especially the histograms of ξ indicating similarity to a normal distribution. This is as expected given the selection of a tailored normal distribution as the proposal density.

5.1.2 GPD with Informative Priors

The Advanced Measurement Approach requires the inclusion of scenario analysis in the Loss Distribution Approach and the most used method of incorporating this is to use it for specifying priors. The model is the same as in the previous two studies, i.e. a GPD with parameters $\xi = 0.7$, $\beta = 10000$ and $\tau = 0$.

This study will just assume that the scenario analysis has been performed and that the resulting priors are known and focus on their impact on the parameter estimates.

The first prior to be imposed is a situation where the scenario analysis has given two lognormal distributions as priors for the two parameters with the following specifications:

$$\pi(\beta) \sim \mathcal{LN}(13, 1), \quad \pi(\xi) \sim \mathcal{LN}(0, 0.5) \tag{5.4}$$

With these priors the mean and mode for each parameter are:

Variable	Mean	Mode
β	7.2942×10^5	1.6275×10^5
ξ	1.284	0.6065

That these values does not match the true parameter values are intentional so any effect from the priors will be possible to study.

The second prior to be studied is a beta distribution for the scale and a gamma distribution for the shape, with the following specifications:

$$\pi(\beta) \sim \Gamma(3, 2 \times 10^5), \quad \pi(\xi) \sim \text{Beta}(5, 2) \tag{5.5}$$

With these priors the mean and mode for each parameter are:

Variable	Mean	Mode
β	6×10^5	4×10^5
ξ	0.714	0.8

The final prior to be tested is one with a serious misspecification, with the purpose to study if the dataset will be able to correct for this or not. This time the prior for the shape is a beta distribution and the prior for scale is a log-normal distribution with the following specifications:

$$\pi(\beta) \sim \mathcal{LN}(12, 2), \quad \pi(\xi) \sim \text{Beta}(3, 4) \tag{5.6}$$

With these priors the mean and mode for each parameter are:

Variable	Mean	Mode
β	4.4241×10^5	2.2026×10^4
ξ	0.4286	0.4

The histograms and parameter statistics of three runs for each prior specification are presented.

Sample size	**Variable**	**Mean**	**St.Dev.**	**Lower**	**Upper**
1000	β	1.01×10^6	3.56×10^4	9.52×10^5	1.08×10^6
	ξ	0.714	0.0564	0.6	0.826
200	β	1.04×10^6	4×10^4	9.68×10^5	1.12×10^6
	ξ	0.752	0.0612	0.631	0.873
50	β	1.05×10^6	4.23×10^4	9.72×10^5	1.13×10^6
	ξ	0.671	0.0751	0.519	0.817

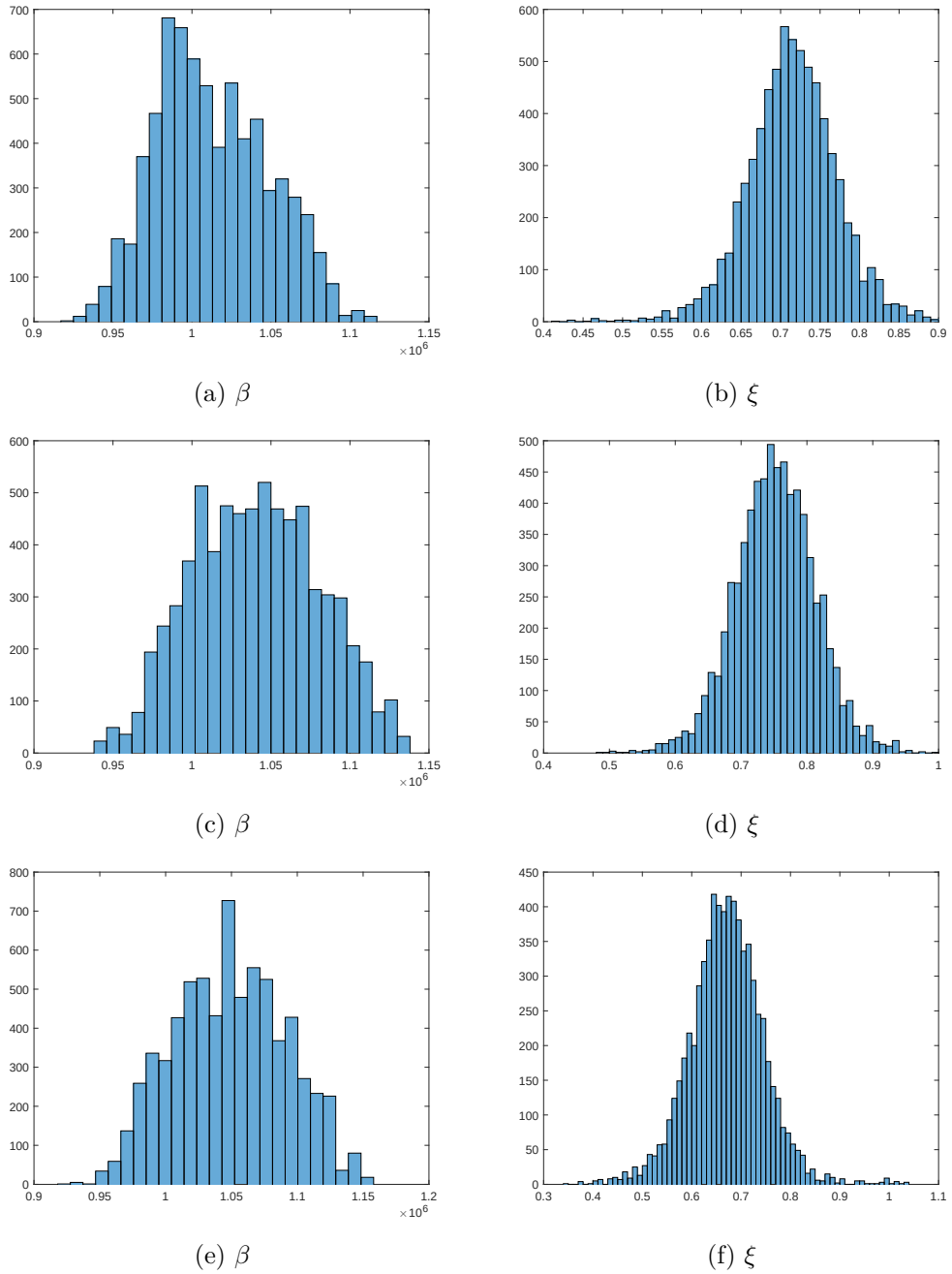
Table 5.1: Statistics for simulation with both priors lognormal.

Sample size	**Variable**	**Mean**	**St.Dev.**	**Lower**	**Upper**
1000	β	10^6	3.99×10^4	9.24×10^5	1.09×10^6
	ξ	0.723	0.0533	0.621	0.826
200	β	1.04×10^6	3.79×10^4	9.71×10^5	1.11×10^6
	ξ	0.751	0.0599	0.629	0.865
50	β	1.04×10^6	4.46×10^4	9.6×10^5	1.13×10^6
	ξ	0.681	0.0757	0.532	0.835

Table 5.2: Statistics for simulation with one gamma and one sigma prior.

Sample size	**Variable**	**Mean**	**St.Dev.**	**Lower**	**Upper**
1000	β	1.02×10^6	3.99×10^4	9.52×10^5	1.11×10^6
	ξ	0.7	0.056	0.584	0.802
200	β	1.06×10^6	4.87×10^4	9.64×10^5	1.16×10^6
	ξ	0.728	0.0602	0.6	0.845
50	β	1.05×10^6	4.57×10^4	9.52×10^5	1.14×10^6
	ξ	0.66	0.0702	0.516	0.795

Table 5.3: Statistics for simulation with misspecified priors.

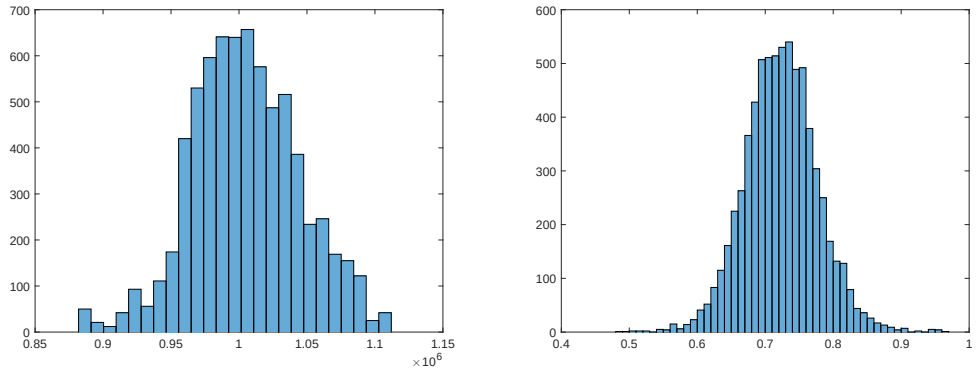


(a) β (b) ξ

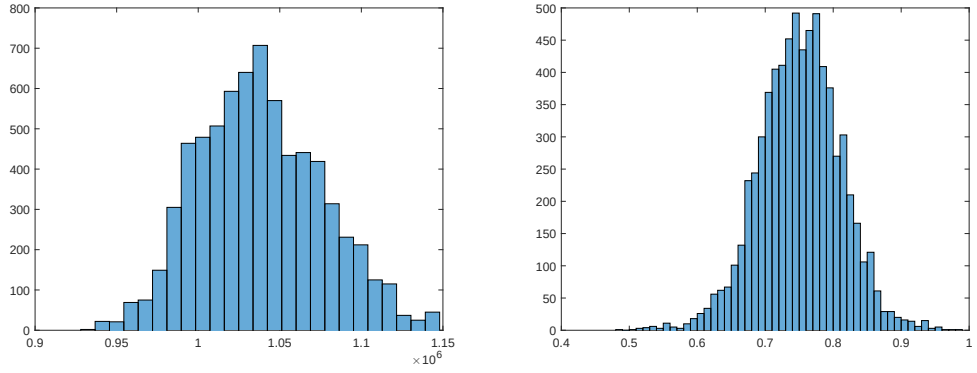
(c) β (d) ξ

(e) β (f) ξ

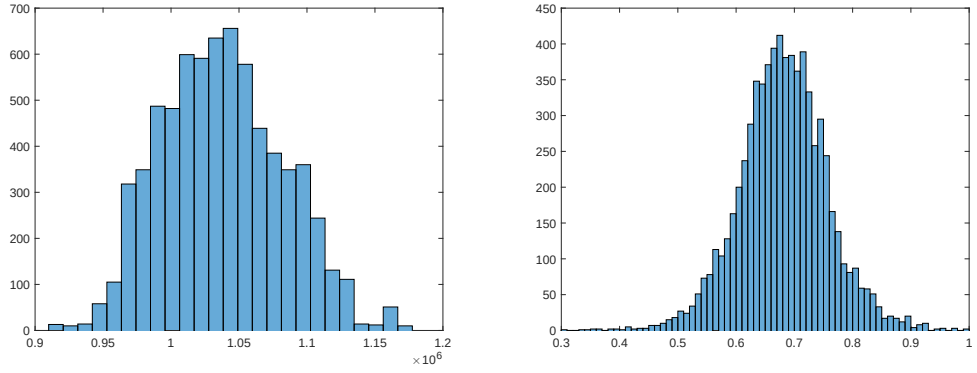
Figure 5.6: Results from simulation with both priors lognormal.The top row shows histograms of realized parameter values from simulation with sample size 1000. The middle row shows plotted histograms of realized parameter values from simulation with sample size 200. The bottom row shows plotted histograms of realized parameter values from simulation with sample size 50.



(a) β (b) ξ

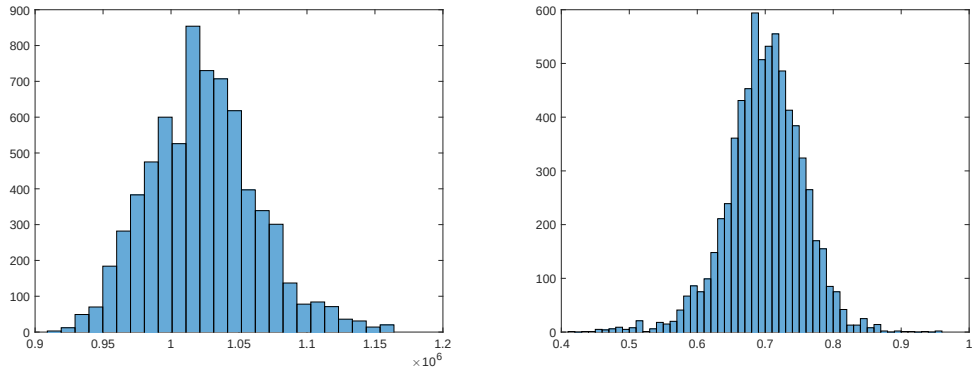


(c) β (d) ξ

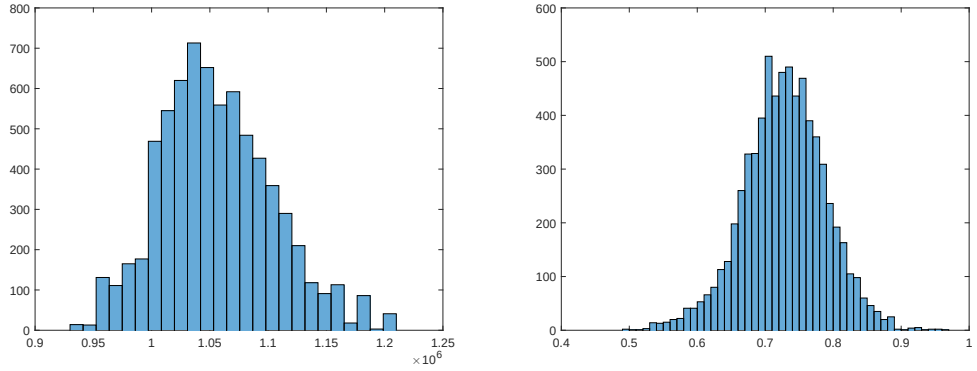


(e) β (f) ξ

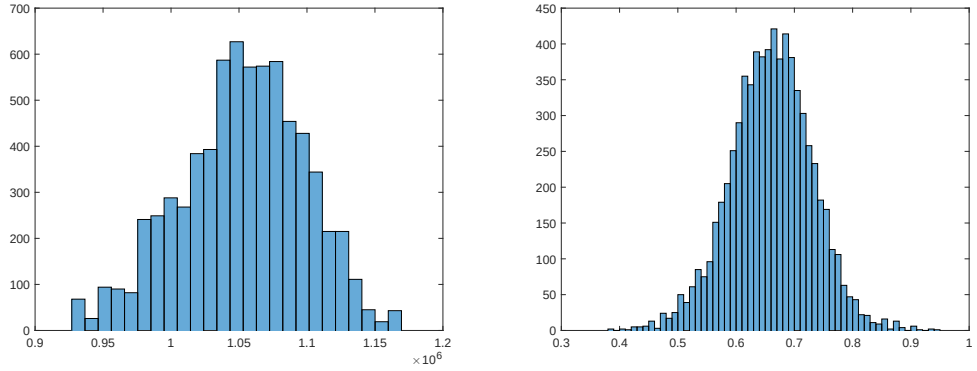
Figure 5.7: Results from simulation with one gamma and one beta prior. The top row shows histograms of realized parameter values from simulation with sample size 1000. The middle row shows plotted histograms of realized parameter values from simulation with sample size 200. The bottom row shows plotted histograms of realized parameter values from simulation with sample size 50.



(a) β (b) ξ



(c) β (d) ξ



(e) β (f) ξ

Figure 5.8: Results from simulation with misspecified priors. The top row shows histograms of realized parameter values from simulation with sample size 1000. The middle row shows plotted histograms of realized parameter values from simulation with sample size 200. The bottom row shows plotted histograms of realized parameter values from simulation with sample size 50.

The first remark to be made is about increased acceptance rates, which was evident during the run of this part of the study. It seems that with these prior distributions the algorithm will accept more generated values which was to be expected as it assigns greater probability to some different parts of the support of the target distribution.

Furthermore, the only really visible effect of the specification of the priors seems to be some slightly wider confidence intervals most clearly seen in the smaller sample size. If a specific comment should be made it would be that the misspecified priors seem to move the sample a little bit to the left in the histograms. Also, the histograms seems to have some skewness, but nothing of great significance.

What these results seem to indicate is that it is the data sample that will drive the parameter estimates when doing Bayesian Inference, even with a misspecified prior the data sample and the MCMC-iteration seem to be able to compensate and sample around the correct values. This is an effect that is desirable as scenario analysis still have some faults in the way it is collected (an issue that is not adressed by this thesis). Hence, these results are satisfactory when moving forward with the thesis work.

5.1.3 Two-step GPD model

This part of the thesis will investigate the "Two-step Bayesian Approach" suggested by [11] (and explained in section 4.3) by creating two separate distributions, one will represent external data and the other will represent internal data. The models are specified as:

$$\begin{aligned} X_{external} &\sim \mathcal{GPD}(\tau = 0, \beta = 1 \times 10^6, \xi = 0.6) \\ X_{internal} &\sim \mathcal{GPD}(\tau = 0, \beta = 1 \times 10^5, \xi = 0.8) \end{aligned} \tag{5.7}$$

To put some context to the matter we look at the 99,9 % VaR for these loss distributions, which is the base for the capital estimates. Table 5.4 shows the true and simulated capital estimate for the two models of (5.7) and in the last row the estimated capital for a situation where you pool the data from both models in the same sample and run the Bayesian Inference framework to estimate the parameters.

Model	**True $\mathbf{VaR_{0.999}}$**	**Estimated $\mathbf{VaR_{0.999}}$**
External	1.0349×10^8	0.84955×10^8
Internal	3.1274×10^7	3.6477×10^7
Combined		1.1362×10^8

Table 5.4: True and estimated $\mathrm{VaR}_{0.999}$ for different models.

As can be seen the estimated capital for the combined data model is above both the external and internal data models capital estimates which seems unlikely.

To make this study as close to the real world situation as possible several data sets will be generated, all from the distributions of eq. (5.7), with number of data points indicated in table 5.5.

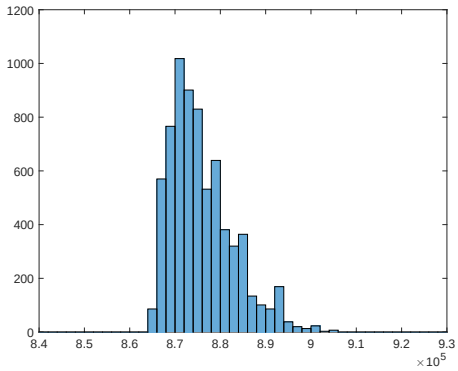
Data sets	
# external data points	# internal data points
10000	200
8000	100
6000	50

Table 5.5: Data sets for testing the Two-step Bayesian approach. Please note that all numbers have been chosen to represent real world data amount in the different Basel risk cells.

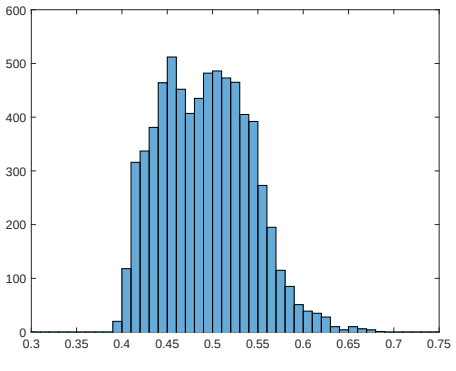
After the first run with the external data model the priors were specified according to the procedure described in section 4.3 using kernels and the second run was performed using data generated by the internal data model. Histograms of resulting parameter samples as well as parameter statistics and the final estimated $\text{VaR}_{0.999}$'s are presented.

Data set size (external,internal)	**Parameter**	**Mean**	**Std**	**Lower**	**Upper**
(10000,200)	β	8.76×10^5	7×10^3	8.67×10^5	8.94×10^5
	ξ	0.492	0.0495	0.411	0.593
(8000,100)	β	8.6×10^5	7.7×10^3	8.49×10^5	8.77×10^5
	ξ	0.472	0.0646	0.344	0.589
(6000,50)	β	9.02×10^5	8.19×10^3	8.91×10^5	9.21×10^5
	ξ	0.494	0.0535	0.39	0.601

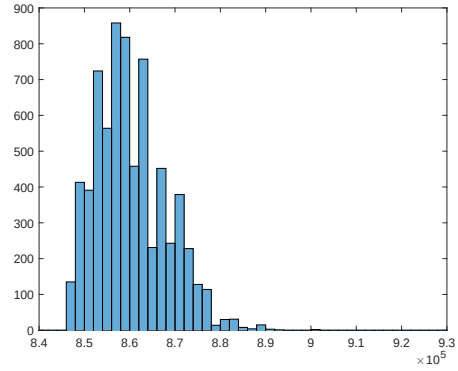
Table 5.6: Parameter statistics from three different sizes of data sets for the second step of the Bayesian Inference approach.



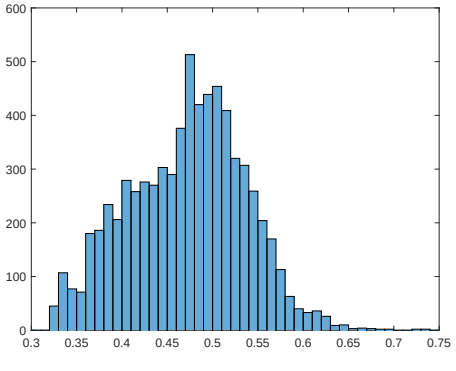
(a) β, Largest data set, (10000,200)



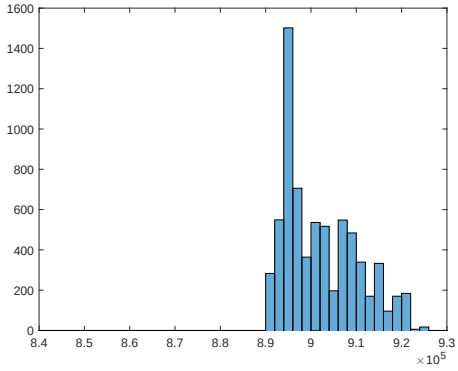
(b) ξ, Largest data set, (10000,200)



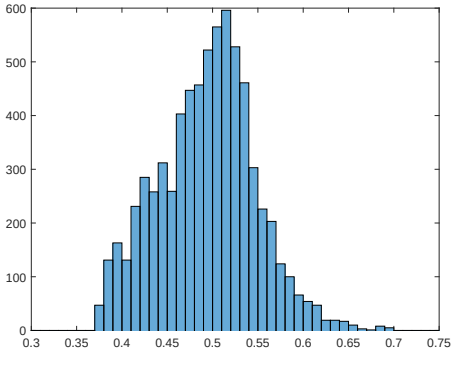
(c) β, Middle data set, (8000,100)



(d) ξ, Middle data set, (8000,100)



(e) β, Smallest data set, (6000,50)



(f) ξ, Middle data set, (6000,50)

Figure 5.9: Histograms of parameter samples from the 2nd run of the Bayesian Approach with posteriors from 1st run as priors.

Data set size (external,internal)	**Estimated $\mathbf{VaR_{0.999}}$**
(10000,200)	5.1496×10^7
(8000,100)	4.6535×10^7
(6000,50)	5.3570×10^7

Table 5.7: Capital estimates based on parameter values from second run of Bayesian Inference approach.

As can be seen from table 5.7 the final capital estimates, or $\text{VaR}_{0.999}$ all lies in between the capital estimates of internal data only and external data only from table 5.4. This is in line with how the models were specified in eq. (5.7) and with the real world expectations. The parameter samples show that the scale parameter, β, shifts a little bit from data set to set but not very significant with the statistics table indicating very similar estimates. For the crucial shape parameter, ξ, however, all estimates stays very close to each other from sample to sample. Also, the standard deviations and upper and lower quantiles seems stable as the data sets get smaller.

Another interesting feature of the results is that the shape parameter distributions indicate that the two-step model lowers this parameter significantly compared to the specified parameter of both the external and internal model. Instead the scale parameter lies in between but closer to the external value. As the final capital estimates lies in between both models' true capital values, and knowing that the shape parameter is crucial for the extreme value behaviour of the distribution, it seems that since the method finds a quite high scale parameter value it lowers the shape parameter to give a sensible final capital estimate.

All in all, the two-step method gives reasonable results with the capital estimates within the expected range as well as stable parameter estimates.

5.1.4 Log-normal body with GPD tail

One approach to create a full distribution for one cell of the Basel matrix, table 2.3, is to use a log-normal distribution for the body and then attach a GPD tail. The Bayesian Inference Approach to determining the parameters of this approach will be investigated here. The model is specified as follows:

$$X \sim \begin{cases} \mathcal{LN}(\mu = 9, \sigma = 2) & \text{if } 0 < X \leq 1 \times 10^5 \\ \mathcal{GPD}(\tau = 1 \times 10^5, \beta = 1 \times 10^6, \xi = 0.9) & \text{if } X > 1 \times 10^5 \end{cases} \tag{5.8}$$

This time, all parameters will be estimated and comparison will be made to the MLE. The probability distribution function of this model is:

$$\begin{aligned} f(x|\mu, \sigma, \tau, \beta, \xi) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) I_{\{0<x\leq\tau\}} \\ + \frac{1 - F_{\mathcal{LN}}(\tau|\mu, \sigma)}{\beta} \left(1 + \xi \frac{x - \tau}{\beta}\right)^{-(1/\xi+1)} I_{\{x>\tau\}} \end{aligned} \tag{5.9}$$

where $F_{\mathcal{LN}}(\tau|\mu, \sigma)$ is the cumulative distribution function of the Log-normal distribution evaluated at the treshold, τ. Therefore, referring to equation (4.2), the likelihood component for this setting is:

$$\begin{aligned} h(\mathbf{x}|\mu, \sigma, \tau, \beta, \xi) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma X_i} \exp\left(-\frac{(\log X_i - \mu)^2}{2\sigma^2}\right) I_{\{0<X_i\leq\tau\}} \\ + \frac{1 - F_{\mathcal{LN}}(\tau|\mu, \sigma)}{\beta} \left(1 + \xi \frac{X_i - \tau}{\beta}\right)^{-(1/\xi+1)} I_{\{X_i>\tau\}} \end{aligned} \tag{5.10}$$

with N being the number of data points. The prior component was again set as wide uniform priors. The results are presented with graphs illustrating the sample paths, statistics of parameter values with comparison to MLE as well as the histograms to illustrate the distributions of parameters. 1800 data points were generated from the body of the distribution and 200 data points from the GPD tail.

It should be noted that the treshold parameter is not possible to estimate using Maximum Likelihood Estimation, instead modellers use other methods such as the Hill estimator or some rule of thumb like taking 10 % of losses as the tail. Hence, the table reporting the results only have NA's in the row of treshold in the Maximum Likelihood part.

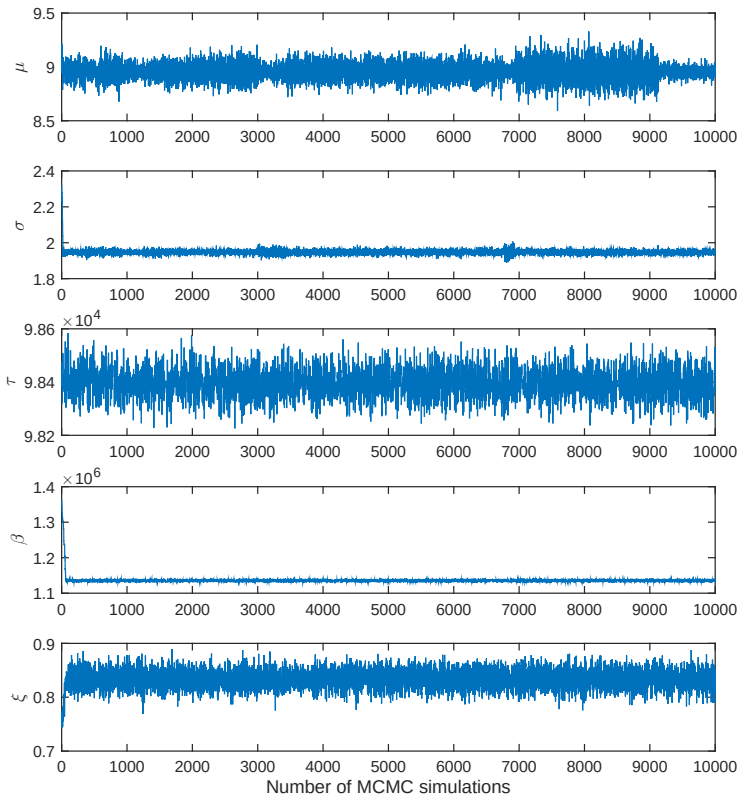
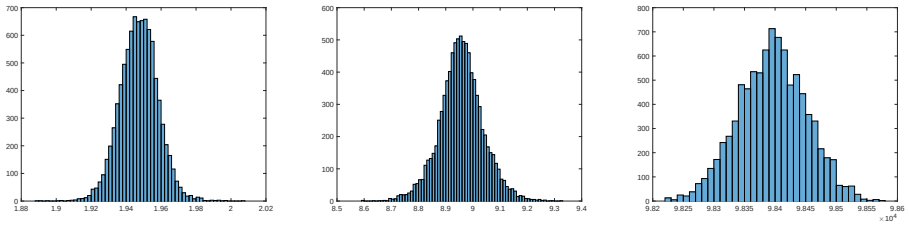


Figure 5.10: Sample paths for all five estimated parameters of Log-normal-GPD model.

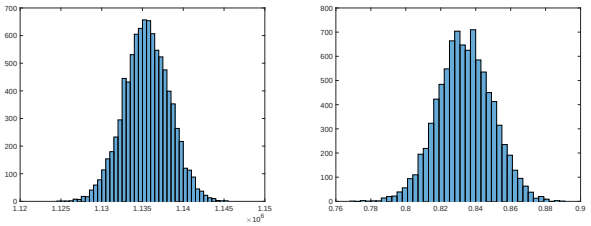
From figure 5.10 the sample paths indicate nice parameter generations around the value the method finds to be the most probable, an observation that is confirmed by the histograms in figure 5.11. All histograms show near normal behaviour, as expected when not imposing any priors. The statistics of table 5.8 indicate that the BI approach performs significantly better than MLE when regarding the standard deviations and confidence intervals, for all parameters except the mean, μ, of the Log-normal body. However, the difference is very small and the BI approach can be said to be performing well also for this parameter.

Parameter statistics				
	Bayesian Inference			
Parameter	Mean	Std	Lower	Upper
μ	8.9573	0.0807	8.7923	9.1223
σ	1.9473	0.0109	1.9261	1.9683
τ	9.8396×10^4	0.0055×10^4	9.8286×10^4	9.8596×10^4
β	1.1355×10^6	0.0028×10^6	1.1311×10^6	1.14×10^6
ξ	0.8339	0.0156	0.8032	0.8644
	Maximum Likelihood			
Parameter	MLE	Std	Lower	Upper
μ	9.1222	0.0506	9.0210	9.2234
σ	2.3076	0.0369	2.2382	2.3814
τ	NA	NA	NA	NA
β	1.3813×10^6	0.1869×10^6	1.0871×10^6	1.7550×10^6
ξ	0.7161	0.1095	0.4971	0.9351

Table 5.8: Table of parameter statistics of the Log-normal GPD model from Bayesian Inference and Maximum Likelihood Estimation.



(a) σ (b) μ (c) τ



(d) β (e) ξ

Figure 5.11: Histograms of parameter samples from the Log-normal-GPD model produced using the BI approach.

5.2 Real world data set

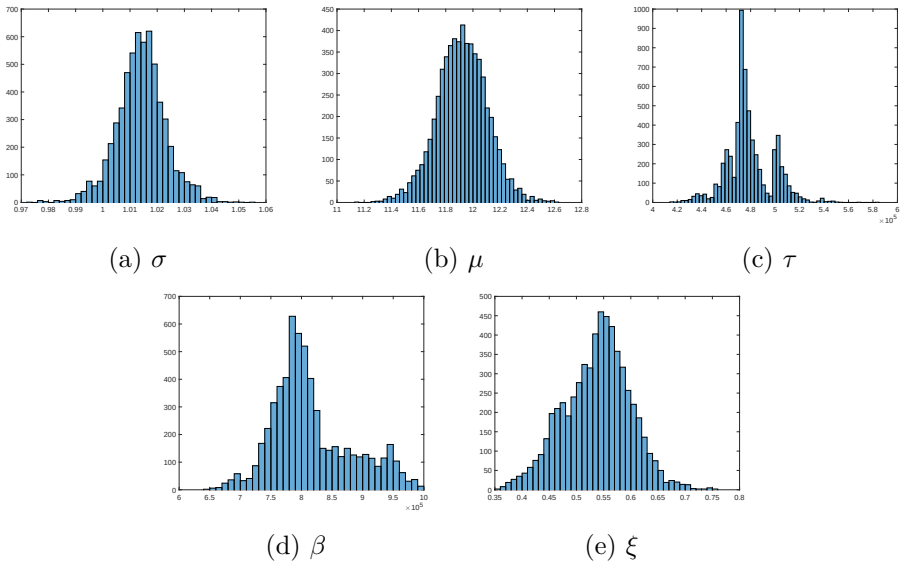
Disclaimer: none of the presented parameter estimates, capital estimates or loss distributions represents the true risk profile or capital allocation of Swedbank AB (publ). The loss data of the bank has been transformed before use, as cleared by Swedbank, and the results should therefore be viewed as illustrative only.

Following the satisfactory results of the BI method when tested on the model in section 5.1.4 this model will now be tested on a data set from the real world. The previous master thesis on the subject of operational risk [13] performs some exploratory data analysis concluding that operational risk loss data includes both high-frequency/low-severity data as well as the aforementioned low-frequency/high severity data. It also raises the possibility that the tail and body data does not originate from the same distribution, hence the Log-normal-GPD model will be tested here.

In line with purpose of this thesis the model will be tested on one cell of the Basel Matrix and the two-step approach described in section 4.3 will be used. This means that firstly, the external data points of the cell will be informing the likelihood component and the posterior from this run will then be used to inform the prior component in the next step when the internal data is introduced into the model. The selected cell of the Basel matrix contained 5252 external and 81 internal data points. The results presented will be the parameter statistics after the second run, as well as the histograms indicating the parameter distributions. To see if the model gives reasonable results the one in a thousand years loss will be presented as well, together with its interval based on setting the shape parameter, ξ, to its upper and lower limit values.

Parameter statistics				
Bayesian Inference				
Parameter	Mean	Std	Lower	Upper
μ	11.9195	0.1858	11.5405	12.2865
σ	1.0141	0.0090	0.9951	1.0331
τ	4.7886×10^5	0.1860×10^5	4.4023×10^5	5.1691×10^5
β	8.1566×10^5	0.6523×10^5	7.0378×10^5	9.5964×10^5
ξ	0.5379	0.0608	0.4136	0.6506

Table 5.9: Parameter statistics from the two-step BI approach on real data sets.



(a) σ (b) μ (c) τ

(d) β (e) ξ

Figure 5.12: Histograms of parameter samples from the Log-normal-GPD model produced using the two-step BI approach informed by real data sets.

To start with the standard deviations and confidence intervals in table 5.9 should be noted. For the tail parameters, and especially for ξ, both st.dev. and confidence intervals are very small, particularly compared to the Maximum Likelihood estimator that did not even converge in this data set. Compared to the previous two-step study of simulated data sets the interval and standard deviation for β increases quite a bit and also compared to the simulated study of the Log-normal-GPD which did not include the two-step approach. This is mainly due to that the internal data set is very small compared to that of the previous Log-normal-GPD study which contained 200 tail data points and the data set of the previous two-step analysis which contained 50 data points of which all were tail points. In this real-world setting the method found an optimal treshold giving ten internal data points in the tail. With this in mind the method performs reasonably well.

It is also interesting to see the actual capital number for this cell. Using eq. (4.15) the capital number was calculated to be ≈ 116 million SEK. Assessing the validity of this number becomes more of a guessing game but given the largest loss in this Basel cell in the last five years was ≈ 4.5 million SEK it does not seem entirely unreasonable.

A table giving the different parameters impact on the capital number is also presented. The table is created by letting all parameters assume their mean value from the resulting sample of the two-step method and then setting one parameter at a time to their confidence intervals' upper and lower points.

Capital estimate intervals		
Capital number with all parameters at mean value: ≈ 116 mSEK		
Parameter	Lower	Upper
μ	81.82	157.88
σ	114.79	118.67
τ	124.21	109.48
β	106.49	129.85
ξ	65.97	210.78

Table 5.10: Table showing the effect of changing each parameter to their upper and lower confidence interval values while keeping all other parameters at mean. The parameter that is changed is in the leftmost column. All numbers are million SEK.

Not surprisingly it is the shape parameter, ξ, of the GPD-tail that gives the largest interval in capital estimate. More surprisingly perhaps, is the fact that the second largest interval comes from the location parameter, μ, in the Log-normal body. To explain this we look at the explicit version of eq. (4.15) used here:

$$\text{VaR}_{0.999}(1y) = \frac{\beta}{\xi}\left[\left(\frac{1-\kappa}{1-F_{\mathcal{LN}}(\tau|\mu,\sigma)}\right)^{-\xi} - 1\right] + \tau \tag{5.11}$$

where

$$\kappa = 1 - \frac{1-0.999}{E[N(1y)]} \tag{5.12}$$

Figure 5.13 shows that as the location parameter increases the CDF-value at the estimated treshold goes down. This leads to a decrease in the value of the fraction inside the brackets of equation (5.11), and since it is taken to a negative power it means that the value inside the brackets will increase. Hence the estimated $VaR_{0.999}(1y)$ will increase. The quite significant increase can be attributed to the rapid decrease of the CDF-value between the lower and upper confidence interval limits for the location parameter. Intuitively this can be understood as the point where the GPD-tail is attached getting a higher probability when the location of the Log-normal distribution increases. Hence, the tail gets a bigger impact on the capital numbers as μ increases, everything else equal.

One should also observe that as the treshold increases the estimated capital decreases, as evident from table 5.10. This can be explained by roughly the same mechanism, as the treshold goes up the survival function value at the treshold decreases and the tail will get lower impact on the capital estimate, everything else equal.

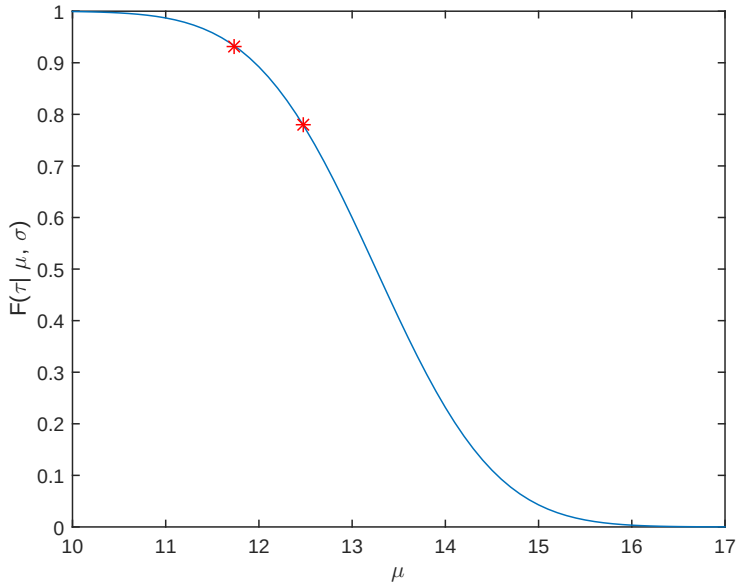


Figure 5.13: Plotted cumulative distribution function of a log-normal distribution at the estimated treshold for different values of the location μ. The confidence interval for the estimated location parameter is marked by red dots.

Finally, to demonstrate the performance of the algorithm in the real-world environment, a comparison to the parameter interval presented in figure 2.1 is made. In that example the shape parameter interval was ≈ 0.5, using the estimated mean of the shape parameter in this real-world study and taking the same interval as upper and lower confidence limits the following interval for the capital estimate was obtained:

Capital estimate interval	
Lower	Upper
49.10 mSEK	1031.90 mSEK

Hence, comparing this interval with the interval for ξ in table 5.10 we see a 72 % reduction in confidence interval for the capital estimate. So, even though the interval from the Bayesian Approach is still quite large it is a significant improvement from the current situation.

Chapter 6

Conclusion

In terms of the objective of this Master thesis, to reduce the statistical uncertainty about parameter and capital estimate, the results indicate good performance of the implemented method. When evaluating the first model, a simulated GPD-model, the BI approach showed significant decrease in parameter uncertainty when compared to MLE. This was evident in data sets as a large 200 data points, and even more visible as the data set size was decreased. The results indicated that BI managed to maintain slim confidence interval much longer than MLE when the data set size was decreased.

Real-world scenario analysis was not incorporated into the model, instead the simulated GPD-model was tested when imposing so called informative priors to give an idea of what impact the scenario analysis could have on the results. The outcome indicated some increased acceptance rate for the parameter samples but otherwise the only visible effect of the priors, even when intentionally misspecified, were some slightly wider confidence intervals. This indicates that it is the data sample that will drive the results of the BI approach and that the scenario analysis will be of informative nature only.

To simplify the inclusion of both external and internal data points in the results a so called "Two-step Bayesian Approach" was tested, using two models, one for external data and another for internal data. Also in this approach the outcome was slim confidence intervals for the parameters as well as reasonable capital estimates that was found to be between the capital estimates calculated separately for the external and internal data. One point of concern was that the final estimate for the shape parameter was found to be below that of both the external and internal data models, a suggestion for further work on this subject would be to investigate why this occurs and if it is an indication that the method is performing badly. This thesis relied on the credibility of the final capital estimates when assessing the reliability of the results, but further studies into this issue could be performed.

To create a full loss distribution for the whole loss space, a model using the Log-normal distribution as body and the GPD as tail was simulated and the BI approach implemented. Again, the outcome indicated a significant reduction of uncertainty compared to MLE except for the location parameter of the Log-normal distribution.

Finally, the BI approach was tested on a real-world data set using the Log-normal-GPD model and the "Two-step Bayesian approach". The outcome was that compared to the simulated data sets the method performed similarly, with some increases in confidence intervals for especially the treshold and scale parameters of the GPD-tail. However, when studying the impact of these confidence intervals on the capital estimate the thesis found that these parameters does not have a significant impact compared to e.g. the shape parameter and the results must be viewed as satisfactory. For the shape parameter the impact of the confidence interval on the final capital estimate was reduced by 72% compared to the interval presented in an initial example of the thesis.

The performance of the Bayesian Inference approach can be attributed to a couple of things. Firstly, the Bayesian Approach does not get stuck in flat density regions, instead it continues to move through the parameter space eventually converging when it finds the optimal value. The simulated annealing part of the tailoring of the proposal density is crucial in this task as it allows the Bayesian Approach to escape any local optima that may be found on its way to the global optimum point. Secondly, the tailored proposal density itself is very important for the performance of the approach as it allows for both a more efficient search for the optimal point as well as good sampling of parameters. When correctly specified, the tailored proposal density will firstly allow for quick convergence towards the optimum, and secondly give an accurate sampling around the optimum enabling the identification of the distribution of the parameters. The fact that Bayesian Inference should produce the probability distribution of each parameter, rather than their most likely point estimate, makes the ability to efficiently sample from a large part of the support of these distributions very important. Therefore, the tailored proposal density is helpful as it in itself is an approximation of the underlying parameter distribution and allows for this efficient sampling.

When it comes to the problem of using both external and internal data into the future loss distribution, Bayesian Inference has the advantage of allowing for incorporation of prior information about the parameters. This is utilized by the two-step approach, letting external data go first and then updating this result by internal data. This leads to the step of obtaining the correct distributions from the first step with external data very crucial for the results of the second step. The use of a kernel density estimator in this step allows the method to stay as close to the empirical distributions generated as possible, avoiding possible biases from fitting statistical to empirical distributions. Hence, the method ensures that the prior distributions then updated by internal data are correct and the final results could be considered

reliable.

To summarize, the Bayesian Inference approach shows significant improvements when trying to estimate the parameters of a loss distribution model in operational risk compared to more traditional estimators. The method implemented gives significant reductions of uncertainty for both parameter and capital estimates in all tests performed. The Master thesis hence shows that it is possible to achieve higher levels of certainty when trying to calculate the capital requirement of a financial institution by using the Bayesian Inference Approach.

Looking forward, apart from the already mentioned issue of the two-step approach, an area of further interest is when the loss distribution of each cell of the Basel Matrix is aggregated to build the total loss distribution. When doing this, benefits from diversification are allowed if the dependence modelling is done in a correct way. Hence, looking at the dependence modelling of the different Basel cells would be an interesting topic for further studies.

Bibliography

[1] Bank for International Settlements, *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel, June 2006, Available at www.bis.org

[2] Bayes, T., Price, R., *An Essay towards solving a Problem in the Doctrine of Chances*. Philosophical Transactions of the Royal Society of London 53, 370-418, 1763

[3] Böcker, K., Klüppelberg, C., *Operational VaR: a closed-form approximation*. RISK, 90-93, 2005

[4] Chib, S., Greenberg, E., *Understanding the Metropolis-Hastings algorithm*. The American Statistician, 49, 327-335, 1995

[5] Epanechnikov, V.A., *Non-Parametric Estimation of a Multivariate Probability Density*. Theory Probab. Appl., 14(1), 153-158, 1969

[6] Ergashev, B., *Should risk managers rely on maximum likelihood estimation method while quantifying operational risk?*. Journal of Operational Risk, 3, 2, 63-86, 2008

[7] Ergashev, B., Mittnik S., Sekeris, E., *A Bayesian Approach to Extreme Value Estimation in Operational Risk Modeling*. Working Paper Number 10, 2013, Center for Quantitative Risk Analysis (CEQURA), University of Munich

[8] Gamerman, D., Lopes, H.F, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd edition, Chapman & Hall, London, 2006

[9] Geman, S., Geman, D., *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741, 1984

[10] Gilks, W.R., Richardson, S., Spiegelhalter, D.J., *Markov Chain Monte Carlo in practice*. Chapman & Hall, London, 1996

[11] Hassani, B.K, Renaudin, A., *The Cascade Bayesian Approach for a controlled integration of internal data, external data and scenarios*. Documents de travail du Centre d'Economie de la Sorbonne 2013.09 - ISSN: 1955-611X

[12] Hastings, W.K., *Monte Carlo Sampling Methods Using Markov Chains and Their Applications* Biometrika, 57, 97-109, 1970

[13] Jöhnemark, A., *Modeling Operational Risk*. Master Thesis, KTH Royal Institute of Technology, 2012

[14] Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. *Optimization by Simulated Annealing*. Science, New Series, Vol. 220, No. 4598, p. 671-680, 1983

[15] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller. A.H., Teller,E. *Equations of State Calculations by Fast Computing Machines* Journal of Chemical Physics, 21, 1087-1092, 1953

[16] Shevchenko, P., *Modelling Operational Risk Using Bayesian Inference*. Springer-Verlag, Berlin Heidelberg, 2011.

[17] The Economist, *Société Générale's rogue trader St Jérôme*, Paris, May 24th 2014, [Internet] [cited April 27th 2015], Available at http://www.economist.com/news/finance-and-economics/21602732-you-cant-keep-good-crook-down-st-j-r-me