UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2018:31

An Application of A Bayesian Extreme Value Mixture Model to Reinsurance Excess-of-Loss Pricing and Extreme Value Threshold Selection

Asmir Prepic

Examensarbete i matematik, 15 hp
Handledare: Jesper Rydén
Examinator: Veronica Crispin Quinonez
Juni 2018

Department of Mathematics
Uppsala University

An Application of A Bayesian Extreme Value Mixture Model to Reinsurance Excess-of-Loss Pricing and Extreme Value Threshold Selection

Asmir Prepic

25 June 2018

Abstract

This thesis fits a flexible extreme value mixture model consisting of a non-parametric bulk part and a parametric tail part to Dansih fire insurance data using Bayesian methods and a estimation with a MCMC algorithm. The posterior distributions of the mixture model parameters are estimated together with the threshold which divides the data into the bulk part and the tail part. A further application on Excess-of-Loss reinsurance contract price estimation is done. The model gives a large variance non-symmetric distribution to all the parameters making the model difficult to use in pricing of reinsurance contract based on the results.

Contents

1 Introduction

One important application of extreme value analysis is the one applied by insurance companies to administrate and estimate cost of financial risk. Insurance seeks to share the risk among a collective as a protection against unforeseeable events with a negative economic cash flow to a individual who is a member of the collective. In other words, insurance is a method of sharing risk between individuals as members of the collective. Central to the definition of insurance is the word "unforeseeable" meaning that insurance covers events which are not known and quite often has small probability. By applying probability theory and statistics to insurance problems, companies can obtain control over stochastic events such that they can be predictable on average and with a estimated error. Insurance companies apply extreme value methods to analyze insurance claims (costs of the insured event) to the insured collective and plenty of time is spent by risk analysts trying to estimate the claims payouts from the insurance company. In particular insurance claims which have high costs are rare and are difficult to anticipate.

Difficulties arise in the analysis from the fact that low probability events such as high cost insurance claims happen seldom and information from the events and when they occur might be restricted. A common method for identifying rare events from a distribution as claim costs is the Peaks over Threshold which allows for finding the threshold, above which events are classified as extreme, by visual inspection. This method might require special knowledge about the events under study and does not either quantify the uncertainty of the threshold. After threshold selection, a Generalized Pareto Distribution is fitted to the observations above the threshold which might be problematic if the number of excesses is small. It is important to choose the threshold carefully for insurance companies who seek to purchase reinsurance contracts to cover extreme losses.

There are other methods developed for threshold selection in various forms which do rely on graphical diagnostic to set the threshold for the extreme events from a distribution. Some modern methods include the use of mixture models where extreme and non-extreme events are considered from two different distributions with and without applications of Bayesian methods for estimation. Behrens, Lopes, and Gamerman suggest such a method in [1] where the distributions below and above the threshold are considered parametric and where the threshold is treated as a parameter to be estimated. Other versions of such approaches are done by Oxley, Reale, Scarrott, and Zhao in [11] and Vrac and Naveau in [10] where they attempt on methodologies of automatizing of threshold selection and quantification of risk using parametric methods.

This thesis attempts to apply a method which estimates a non-parametric distribution for the density below a threshold and a Generalized Pareto Distribution (GPD) for the density above a threshold. This methodology requires application of sampling with MCMC and specification of prior distributions for the parameters.[3] An application is done on insurance claims data.

The thesis is organized as follows. The presentation of the theory applied is done in Section 2 starting with a short introduction to insurance mathematics, Bayesian statistics and followed by the specification of estimation procedure as well as the MCMC algorithm. Section 3 presents the data on which the application is attempted. Section 4 presents the results of the estimation.

2 Theory

The main theory is from Mikosh [9] and from Rydén and Rychlik[13]
The following section will present the theory used in this thesis starting with insurance and re-insurance moving to extreme value theory and application for insurance problems.

2.1 Applications of Probability in Insurance

The foundations of modern Risk theory was set in the early 20th century by the Swedish actuary Filip Lundberg in [4]. Lundberg focused on insurance claims arrivals and insurance claims sizes in order to find how much premium needs to be charged for the insurance company not to go bankrupt. The strength of his approach was that he was able to find efficient methods for modeling portfolios which had similar risk profile. This meant for example isolating a portfolio of only fire insurances for family homes and modeling their claims arrival process and distribution of claim payment sizes.

Central to Lundbergs ideas was three assumptions which are still used in applications in industries today often with case to case modifications. The three assumptions are as follows:

- Claims happen at random times T_i where they are ordered as $0 \leq T_1 \leq T_2 \leq T_3 \leq ...$
- When a claim arrives at time T_i it is attached with a claim payment (size or severity) X_i. The sequence of claim sizes $\{X_i\}$ is i.i.d. of non-negative random variables.
- $\{X_i\}$ and $\{T_i\}$ are independent.

The intuition for the assumptions are straightforward as there is only one assumption, i.i.d. claim sizes which indicate a portfolio of similar risks. By applying the assumptions a more straightforward and simplified modelling is obtained suggesting that the insurance companies should split their exposure into homogeneous portfolios in order to be able to monitor the risk closely. Independence between claim sizes and claim arrivals simplifies calculation but also means that it is unlikely that a period of many claim occurrences will at the same time be a period of larger claim sizes. These assumptions do not have to be met.

In order to understand re-insurance and insurance policies, an explanation of a general collective risk model will be presented in the following section. Using the previously stated assumptions, the model for processes of the number of claims is defined as a counting process

$$N = N(t)_{t \geq 0} \tag{1}$$

on $[0, \infty)$ where $N(t)$ is the number of claims occurred by time t.

Having specified the counting process $N(t)$, and claim sizes X_i, the total claim cost process is defined as

$$S(t) = \sum_{i=1}^{N(t)} X_i. \tag{2}$$

By controlling the process in (2), the insurance company can run the insurance business without going bankrupt. To reach the goal models, statistics and asymptotic behaviour of (2) need to be investigated. This thesis will focus on estimating the distribution of X_i in (2) as the scope will quickly become very large if the counting process is also addressed.

The insurance company meets the obligations of the insurance contracts by using (2) to set the necessary premium to cover the losses. Equation (2) is used for the purpose of estimating the premiums. The premium income needs to be set at a level which covers total losses of (2) which can be denoted S. The level of premium income over time is a deterministic function $p(t)$ since it is already set by the insurance company what to charge. Since $S(t)$ is random, one approach to setting the premium is to set it at $E[S(t)]$ with the reasoning is that the outcome of the total claim amounts will be centered around some central statistic. If $p(t) < E[S(t)]$, the insurance company will then have a loss, on average, or have a profit, on average, if $p(t) > E[S(t)]$. In order for the insurance company to perform better than average, it is common that the premium is set such that

$$p(t) = (1 + \rho)E[S(t)] \tag{3}$$

where ρ is referred to as a loading. This can be an arbitrary percentage or for example one standard deviation of $S(t)$. Because of the properties of the claim sizes and claim arrival times the expected loss is

$$E[S(t)] = \lambda t E[X_1], \tag{4}$$

if the counting process $N(t)$ is a homogeneous Poisson process. X_1 is generally a positive random variable such as exponential, gamma, log-normal or truncated normal.

2.2 Re-insurance and Claim Sizes

Some insurance portfolios have too volatile $S(t)$ (total loss processes) to be handled properly. In order to meet this volatility, insurance companies have mutual agreements to share the risk and the premium of the portfolio.[9] Sharing of risk is not much different from a community founding its own insurance company to share the risks among the insured who are also the owners. Such companies are usually referred to as mutual and have a long history in Europe. There are many different types of re-insurance agreements who all are designed to share the risk between insurance companies in their own way and many of the usual types of agreements aim to modify the total loss process $S(t)$ in some kind of way. Mikosh in [9] describes the three common treaties.

2.2.1 Proportional Reinsurance

Proportional loss reinsurance where a fraction ρ of $S(t)$ is shared between two or more companies. With proportional loss reinsurance the company will pay and receive

$$(1-\rho)S(t) \tag{5}$$

of the total claim cost and

$$(1-\rho)p(t) \tag{6}$$

of the premium.

2.2.2 Stop- Loss Reinsurance

Stop-loss reinsurance where the insurer covers the losses of the entire portfolio above a certain limit. This means that the insurer will not face the total $S(t)$ but rather $\min(S(t), K)$ where K is the agreed upon total limit. On the other hand the re-insurer will have the total cost function

$$R(t)_{SL} = \max(S(t) - K, 0) \tag{7}$$

2.2.3 Excess-of-Loss Reinsurance

Excess-of-loss reinsurance where the insurance company pays for all individual insurance claims up to a threshold r and the re-insurer pays the amount above K for any of the claims that reach the claims level. This gives the re-insurer total claim amount process of

$$R(t)_{XL} = \sum_{i=1}^{N(t)} \max(X_i - r, 0). \tag{8}$$

giving the loss function from the insurance company's perspective

$$S(t)_{insurance} = \sum_{i=1}^{N(t)} \min(X_i, r). \tag{9}$$

The scope of this paper is to choose the optimal value of r based on the distribution of X_i. The goal is to find $E[R(t)_{XL}] =$ premium for reinsurance. For the purpose of finding X_i, extreme value theory is used.

2.3 Extreme Value Theory

Extreme Value Theory can be described in different ways. In the Wikipedia article extreme value theory is described as "*... a branch of statistics dealing with large deviations from the median of probability distributions..*" Similarly extreme value theory is used to determine and analyze probabilities that more extreme events than already occurred will be observed.[2]

Let $X_1.X_2, ..., X_n$ be independent identically distributed random variables with distribution $F(x)$. The inference is generally focused around the maximum

$$M_n = \max(X_1, X_2, ..., X_n) \tag{10}$$

of the sequence. The distribution of (10) is easily derived by applying the rules for independent and identically distributed random variables as

$$\begin{aligned} F_{M_n}(x) &= P(X_1 < x, X_2 < x, ..., X_n < x) \\ &= P(X_1 < x)P(X_2 < x)...P(X_n < x) = F(x)^n. \end{aligned} \tag{11}$$

The distribution function of (10) is not completely trivial. A problem arises when n $\to \infty$. By the definition of (11) the distribution will be degenerate since $F_{M_m}(x) \to 0$ as n $\to \infty$. By introducing sequences of normalizing constants a_n and b_n and adjusting the distribution in (10) such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(a_n x + b_n)]^n \to G(x) \tag{12}$$

is not degenerate.[13] If (12) holds then $G(x)$ is said to belong to a Generalized Extreme Value Distribution (GEV) on the form

$$G(x|u,\sigma,\xi) = \begin{cases} \exp\left(- \left[1 - \xi \frac{(x-u)}{\sigma}\right]_+^{1/\xi} \right) & \xi \neq 0 \\ \exp\left(- \exp\left(- \frac{x-u}{\sigma}\right)\right) & \xi = 0, \end{cases} \tag{13}$$

where σ is a scale parameter, u is a location parameter and ξ is a shape parameter and where $x_+ = \max(x, 0)$.[13] More specifically, the Generalized Pareto Distribution (GPD), defined as

$$G(x|u,\sigma,\xi) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left[- \left(\frac{x-u}{\sigma}\right)\right]_+ & \xi = 0, \end{cases} \tag{14}$$

where $x > u$, $x_+ = \max(x, 0)$, $\xi > 0$ and $\sigma > 0$ is used in this thesis. The meaning is that the excess distribution above some threshold u will be approximated by (14). The aim of this thesis is mainly finding u and the parameters of GPD.

2.4 Peaks Over Threshold and Threshold Selection

Peaks over threshold, POT, refers to the study of equation (14) and determining how many observations exceed the threshold u. Expressing (14) in terms in distribution functions such that

$$U(x) = P(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}. \qquad 0 \leq x < x - u \tag{15}$$

From the theory of extreme value distributions, (15) will follow a GPD as specified in (14). The GPD of (14) is also valid for any threshold $v > u$ if the distribution above u also follows a GPD.[2] By assuming that the exceedenses above u are independent the number of exceedenses can be approximated by a Poisson distribution. This is used in reinsurance pricing by modelling the number of excedanses as Poisson random variables having GPD claim sizes.[9]

There are numerous ways of choosing the threshold as well as quantifying the uncertainty of u. Choosing a threshold that is too high leaves fewer observations for estimation of the parameters. On the other hand if it is too low, it might not be a GPD distribution. Threshold selection is therefore a choice between bias and variance.[2]

Techniques of visual inspection in some form has been common in determining the threshold u. One common method which has been used for years is the "Mean Excess Plot" which aims to give a visual description of the GPD behaviour for different values of u. This is based on the fact that the mean of a GPD distributed variable X is

$$E(X) = \frac{\sigma}{1 - \xi} \tag{16}$$

and by introducing an excess u

$$E(X - u > | X > u) = \frac{\sigma_u}{1 - \xi}. \tag{17}$$

The mean of a GPD should theoretically have linear property which means by introducing a high threshold $v > u$ should yield

$$E(X - v | X > v) = \frac{\sigma_u + \xi v}{1 - \xi}, \tag{18}$$

which is a linear transformation of (16).[2] Empirically the mean excess function is estimated as

$$\hat{e}(u) = \frac{\sum_{i=1}^{n} (X_i - u) 1_{[X_i > u]}}{\sum_{i=1}^{n} 1_{[X_i > u]}}. \tag{19}$$

.

There are many different approaches for choosing u all with different characteristics.[7] The choice of methodology can be dependent on the situation and the goal for the analysis. To estimate the whole distribution of the claim sizes X_i and not just the tail, a approach could be to split the distributions into a "bulk" part and a "tail" part. These type of distributions are generally refereed to as "mixed type distributions" and are usually difficult or impossible to solve for analytically.[7] Therefore in many of the cases in "mixture model" Bayesian and other techniques are used to fit the "bulk" and "tail" distributions as well as the threshold which can be treated like a parameter in a Bayesian framework.

The threshold u will be estimated using Bayesian technique in this thesis and the "bulk" part of the distribution will be set as a non-parametric Kernel density as in.[3] The next section describes Bayesian estimation techniques for distribution parameters followed by the theory of the application in this thesis.

2.5 Bayesian Statistics

The Bayesian approach to statistical inference is built from the concepts of Bayes formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{20}$$

where $P(A|B)$ is the probability of event A given that event B has happened and $P(A \cap B)$ is the probability that both events A and B happen and $P(B) > 0$.[13] It is important to note that if there is a partition $A_1, A_2, ..., A_k$ of a set S then it holds that the event B can be written

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i) \tag{21}$$

which is also known as the law of total probability. Equation (21) is sometimes expressed in the equation (20) as a normalizing constant $c = P(B)$ by rewriting (20) such that

$$P(A|B) = \frac{P(B|A)P(A)}{c}. \tag{22}$$

Expression (22) describes the problem in such way that first $P(B|A)P(A)$ is set and then normalized to be a valid probability. The goal is often to find the numerator of (22) in different applications of estimation procedures.

Bayesian estimation of the parameters utilizes the characteristics of Bayes formula to estimate some parameter of interest. Suppose some parameter θ is unknown and an estimation is required from some random sample of $(X_1, X_2, ..., X_n)$. θ is then, in a classical approach, estimated by using the information in the random sample. After the estimation is done, θ is considered fixed in the classical approach. This could be for example any of the parameters of a probability distribution. In the Bayesian approach on the other hand, θ is considered as a random variable, with a probability distribution for which some prior belief is set to. This means that before observing any data, there is an idea of in what range and with which weights θ is likely to be in. A sample is then obtained from a population and the prior belief about θ is updated with the information from the sample. The updated distribution about of θ becomes then the posterior distribution.

If the prior distribution of θ is denoted as $\pi(\theta)$ and the distribution of the sample $f(x|\theta)$, and the marginal of the sample

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta, \tag{23}$$

then the posterior of θ is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}. \tag{24}$$

The $m(x)$ part of equation (24) can be seen as a normalizing constant for the distribution $\pi(\theta|x)$ which will not be of central importance in this application. The posterior distribution $\pi(\theta|x)$ is conditional on the observed sample of x. By obtaining the posterior, inference can be drawn about the parameter value θ. For example, point estimation, intervals or simulations from the distribution. It is on these ideas that the distribution of the insurance losses will be estimated as well as the threshold for large claims.

2.6 Estimation of the Claims Distribution and Threshold

Since the aim of this thesis is mainly threshold selection and estimation of the tail distribution, the approach will be to set the distribution below the tail as a non-parametric kernel density. In addition, the distribution below the tail will be simultaneously estimated as the threshold and tail distribution. The reason for such approaches is the increased complexity of parametric models when implementing simultaneous estimation.[3] The model in this thesis will be a mixture model as explained in this section.

Suppose a sample $X_1, X_2, ..., X_n$ of independent identically distributed observations with distribution function $F(x)$. The goal is to estimate the distribution function $F(x)$. The approach in this thesis is to split

$$F(x|\lambda, u, \lambda_u, \xi, (X_1, X_2, ..., X_n)) = \begin{cases} (1-\phi_u)\frac{H(x)|\lambda,(X_1,X_2,...,X_n))}{H(u|\lambda,(X_1,X_2,...,X_n))} & x \leq u \\ (1-\phi_u) + \phi_u G(x|u, \sigma_u, \xi) & x > u \end{cases} \tag{25}$$

such that $H(x|\lambda, (X_1, X_2, ..., X_n))$ is a non-parametric distribution function depending only on the parameter λ, $G(x|u, \sigma_u, \xi)$ is a Generalized Pareto Distribution for the distribution above the threshold u and ϕ_u is the probability that a observation will be above the threshold u which is estimated as the proportion of the observations above the threshold u. The $\phi_u G(x|u, \sigma_u, \xi)$ is then a unconditional GPD function as opposed to $G(x|u, \sigma_u, \xi)$ which is not scaled with ϕ_u. The problem of this thesis is to estimate the parameters $\lambda, u, \sigma_u, \xi$ and ϕ_u and the approach will be a Markov Chain Monte Carlo algorithm.[3]

The kernel distribution $H(x)$ is as mentioned a non-parametric distribution function. The estimation of such a distribution function is specified directly from the information drawn from the sample $X_1, X_2, ..., X_n$ without estimations of any distribution. The procedure aims to capture the "true" distribution and density as generated by the data. A function

$$f(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) \tag{26}$$

defines the kernel density and is dependent on the sample and that the parameter $\lambda > 0$. K in (26) is usually a probability density chosen such that it is centered around zero. This naturally allows for a variety of choices for density functions that satisfies the conditions of (26). For example K can be a uniform or triangular distribution. In this thesis K will be a zero mean normal probability density function. For the reader's understanding, one can think of the kernel density as a histogram with finer and finer bins resulting in a smooth function. Central to the Kernel density estimation is the choice of the parameter λ which is called the bandwidth parameter.[3]. There is plenty of textbooks and literature dealing with the choice of λ but this is treated as a parameter which needs to be estimated and thus a likelihood and a prior distribution will be specified.

For parameter estimation in this thesis a likelihood function needs to be specified. The sample $(X_1, X_2, ..., X_n)$ can be denoted as $\mathbf{X}$ to simplify the notation in the thesis. If θ denotes the parameter vector $\theta = (\lambda, u, \mu, \sigma, \xi)$ then the likelihood function can be expressed as

$$L(\theta|\mathbf{X}) = L(\lambda, u|\mathbf{X})L(u, \mu, \sigma, \xi|\mathbf{X}). \tag{27}$$

The first likelihood of (27) is the likelihood for the bulk of distribution below the threshold u. Since the bulk of the distribution will be estimated with a kernel density, the likelihood only depends on λ and u with λ being the bandwidth parameter. The second part of the likelihood in (27) is the likelihood for the Generalized Pareto Distribution which also depends on u among other parameters. Since both of the likelihood functions depend on the same parameter, MCMC estimation will be necessary.[3]

The kernel function will be rewritten for easier notation in the following section. Recall that the kernel density is

$$f(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right). \tag{28}$$

By rewriting the $K()$ function such that

$$K_\lambda(x) = \frac{K\left(\frac{x}{\lambda}\right)}{\lambda} \tag{29}$$

giving the function

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(x - x_i) \tag{30}$$

and as before K is a standard normal density function.[6] The likelihood is needed since one of the parameters to be estimated is λ and is specified as

$$L(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{n}\sum_{j=1}^{n} K_\lambda(x_i - x_j) \tag{31}$$

for $j \neq i$.[3] The likelihood function is further adjusted to avoid degeneracy in the way that it is normalized.[3] The likelihood function has to further be adjusted such that it is only considering the bulk of the distribution. I.e. the part of the density which is below the threshold u. If the set for the observations below the threshold u is denoted $A = \{i : x_i \leq u\}$, the resulting likelihood function becomes

$$L(\lambda, u|\mathbf{X}) = \left\{ \frac{(1-\phi_u)}{\frac{1}{n}\sum_{k=1}^{n}\Phi(\frac{u-x_k}{\lambda})} \right\}^{|A|} \prod_{A} \frac{1}{(n-1)} \sum_{j=1}^{n} K_\lambda(x_i - x_j) \tag{32}$$

for $j \neq i$. Some definitions need to be made for equation (32). Φ is the standard normal distribution function and ϕ is the proportion of observations which are above the threshold u and A denotes the set of the observations below u. The $K_\lambda()$ is the normal probability density function.[3]

The second part of the likelihood which needs to be specified is the second part from (27). I.e.

$$L(u, \mu, \sigma, \xi|\mathbf{X}) \tag{33}$$

is a function of the threshold u the GPD parameters μ, σ and ξ conditioned on the observed sample $\mathbf{X} = (X_1, X_2, ..., X_n)$. Since the GPD density function has a different form depending if $\xi = 0$ or not the likelihood is

$$L_{pp}(u, \mu, \sigma, \xi|\mathbf{X}) = \begin{cases} exp\left\{ -n_b\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \prod_B \frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \\ \\ exp\left[-n_b exp\left(\frac{u-\mu}{\sigma}\right)\right] \prod_B \frac{1}{\sigma} exp\left(\frac{x_i-u}{\sigma}\right), \end{cases} \tag{34}$$

where the first part of (34) is $for \quad \xi \neq 0$ and the second part is $for \quad \xi = 0$. There are a few notes for the likelihood function in (34).[3] First of all, the set which it is applied for is the tail of the distribution. Since the bulk of the distribution was denoted $A = \{i : x_i \leq u\}$, the tail can be denoted $B = \{i : x_i > u\}$ and is the observations above the threshold u.[3] n_b in (34) is defined as the number of exceedenses above the threshold u. n_b does not have to be defined as such but it is defined so in the likelihood in (34). The rest of the parameters are as in earlier notation the parameters for a GPD.

2.7 Bayesian Estimation Specification

This section is using the theory from Section 2 to define the necessary parts for the Bayesian Estimation of the parameters. As mentioned, the estimation technique in this thesis will be Bayesian. Meaning that a specification of the prior distributions of all the parameters needs to be done before starting the estimation algorithm. The notation used will be the same as in.[3]

2.7.1 Priors for Tail Distribution Parameters

The definition requires specification of the parameters to estimate. Since one part of the distribution will be labeled as "bulk" and the other as "tail". There is a prior distribution for all the parameters involved which can be written as $\theta = (\lambda, u, \mu, \sigma, \xi)$.[3] The important assumption in this application is that the distributions for λ, u and (μ, σ, ξ) are independent, meaning that the multiplication rule of independence holds and the full prior for the parameter distributions is expressed as

$$\pi(\lambda, u, \mu, \sigma, \xi) = \pi(\lambda)\pi(u)\pi(\mu, \sigma, \xi). \tag{35}$$

[3] The prior distributions for all the parameters are explained in the following section.

There are a few methodologies to specify the priors for the GPD or the "tail" distribution. These include expert opinions on some quantiles or other information. Central for the specification of priors is to incorporate some sort of belief about the distribution of the parameters. There might for example be an idea of what the tail distribution looks like. This might expressed through the parameters that it is very unlikely (but not impossible or impossible if that is the prior information). This type of prior information is sometimes difficult to extract by simply analyzing the data. There are caps and floors and other type of claim specific information in insurance which claim handlers or other staff might know. The choice of the prior for the GPD tail distribution parameters in this thesis is a diffuse prior meaning that as little as possible prior information is incorporated in the prior distribution of the GPD parameters. The choice is a multivariate normal distribution with a uninformative arbitrary (large) variance and uncorrelated marginal distributions and arbitrary mean. The multivariate normal distribution with three variates is written as

$$f(x_1, x_2, x_3) = \frac{exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{(2\pi)^3 |\mathbf{\Sigma}|}} \tag{36}$$

where in the case of this thesis $x_1 = \mu, x_2 = \sigma, x_3 = \xi$, $\boldsymbol{\mu} =$ the vector with the means of (μ, σ, ξ). $\mathbf{\Sigma}$ in (36) is the covariance matrix for the joint distribution of (μ, σ, ξ) with diagonal entries being the variances of (μ, σ, ξ) as (V_μ, V_σ, V_ξ). The off-diagonal elements of $\mathbf{\Sigma}$ are set to 0 since the parameters are assumed to be uncorrelated.

2.7.2 Prior for Threshold u

There are plenty of ways to define the prior for the threshold u.[1] A discrete, uniform or any type of other prior can be defined. As with the other priors, the information from the prior distribution is set by how the parameters are set. In this thesis we would like priors without to big of a variance such that the prior distribution is somewhat centered around the mean. There are not any specific reasons for why the distribution is somewhat centered around the mean. It is just an arbitrary choice. The distribution for the threshold u applied in this thesis however is a normal distribution with mean u_u and variance σ_u.[1] The prior distribution applied is

$$\pi(u|\mu_u, \sigma_u, e_1) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \frac{exp(-0.5[(u-\mu_u)/\sigma_u]^2}{\phi[-(e_1-\mu_u)/\sigma_u]}, \tag{37}$$

where μ_u and σ_u are the mean and variance of the threshold u which is usually set at some high quantile.[3] The $\phi[-(e_1-\mu_u)/\sigma_u$ is a truncation factor which puts a restriction on the probability mass of (37). This means that the truncation sets the probability mass to 0 below the truncation point e_1. In the example of the normal distribution, $e_1 = 0$ means that the distribution has only positive mass on $x \in \mathbb{R}^+$, i.e. the positive real numbers. The choice of $e1$ can be anything to restrict the values of the prior for u. In the application of this thesis $e_1 = 0$ to not add to much information to the distribution.

2.7.3 Prior for Bulk Parameter λ

It is difficult to intuitively set a prior distribution for the bandwidth parameter λ. λ specifies in a way the "smoothness" of the non-parametric density.[6] This makes intuition and prior information difficult to include without touching the data at first. In this thesis, the prior for λ is set as a prior for the precision $1/\lambda^2$

$$\pi(\lambda|d_1, d_2) = \frac{1}{d_2^{d_1}\Gamma(d_1)} \left(\frac{1}{\lambda^2}^{d_1-1}\right) exp\left(-\frac{1}{\lambda^2 d_2}\right), \tag{38}$$

which is an inverse gamma distribution with parameters d_1 and d_2.[3] The choice of the inverse gamma for this application is the same as by the authors in [3] as there is not more information about this prior distribution. The authors in [3] argue that this prior is suitable for the application.

2.8 Estimation of Posterior

The goal of the thesis is to find a suitable threshold and estimate the posterior distribution of the parameters. By Using the theory in Section 2.5 a posterior distribution can be found for all the parameters involved in the likelihood function. In some general cases, the posterior can be found with an analytic solution based on the likelihood and the prior.[12]. Other applications where the posterior does not have a analytic solution or has a very difficult form, other estimation techniques need to be applied. One of these estimation techniques are Markov Chain Monte Carlo or MCMC for short. MCMC allows sampling from posterior distribution as well as finding statistics from posterior distribution. MCMC will not be treated in detail in this thesis. The reader can find information written by Dellaportas and Roberts in [5]. Only the algorithm will be outlined which is based on the prior parameters specified earlier in the thesis.

The goal is to simulate the posterior for $(\lambda, u, \mu, \sigma, \xi)$ with a Metropolis-Hastings algorithm using the following steps:

Step 0: set starting value $(\lambda^{(0)}, u^{(0)}, \mu^{(0)}, \sigma^{(0)}, \xi^{(0)})$

set iteration step $j \geq 1$

Step 1: Given $\xi^{(j-1)}$, generate $\xi^* \sim N(\xi^{(j-1)}, V_\xi)$, calculate

$$\alpha_\xi = \min\left\{ \frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^*|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j-1)}|\mathbf{X})}, 1 \right\} \tag{39}$$

and with probability α_ξ accept ξ^* and set $\xi^{(j)} = \xi^*$ else reject ξ^* and set $\xi^{(j)} = \xi^{(j-1)}$.

Step 2: Given $\sigma^{(j-1)}$, generate $\sigma^* \sim LN(\log(\sigma^{(j-1)}), V_\sigma)$, calculate

$$\alpha_\sigma = \min\left\{ \frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^*, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j)}|\mathbf{X})} \frac{LN(\sigma^{(j-1)}|\log(\sigma^{(*)}), V_\sigma)}{LN(\sigma^{(*)}|\log(\sigma^{(j-1)}), V_\sigma)}, 1 \right\} \tag{40}$$

and with probability α_σ accept σ^* and set $\sigma^{(j)} = \sigma^*$ else reject σ^* and set $\sigma^{(j)} = \sigma^{(j-1)}$.

Step 3: Given $\mu^{(j-1)}$, generate $\mu^* \sim N(\mu^{(j-1)}, V_\mu)$, calculate

$$\alpha_\mu = \min\left\{ \frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(*)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1 \right\} \tag{41}$$

and with probability α_μ accept μ^* and set $\mu^{(j)} = \mu^*$ else reject μ^* and set $\mu^{(j)} = \mu^{(j-1)}$.

Step 4: Given $u^{(j-1)}$, generate $u^* \sim N(u^{(j-1)}, V_u)\mathbb{I}_{(\min(x_1,\dots,x_n),\max(x_1,\dots,x_n))}$, calculate

$$\alpha_u = \min\left\{\frac{\pi(\lambda^{(j-1)}, u^{(*)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \frac{(\Phi((M-u^*)/\sqrt{V_u}) - (\Phi((m-u^*)/\sqrt{V_u})}{(\Phi((M-u^{(j-1)})/\sqrt{V_u}) - (\Phi((m-u^{(j-1)})/\sqrt{V_u})}, 1\right\} \tag{42}$$

and with probability α_u accept u^* and set $u^{(j)} = u^*$ else reject u^* and set $u^{(j)} = u^{(j-1)}$.

Step 5: Given $\lambda^{(j-1)}$, generate $\lambda^* \sim LN(\log(\lambda^{(j-1)}), V_\lambda)$, calculate

$$\alpha_\sigma = \min\left\{\frac{\pi(\lambda^{(*)}, u^{(j)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \frac{LN(\lambda^{(j-1)}|\log(\lambda^{(*)}), V_\lambda)}{LN(\lambda^{(*)}|\log(\lambda^{(j-1)}), V_\lambda)}, 1\right\} \tag{43}$$

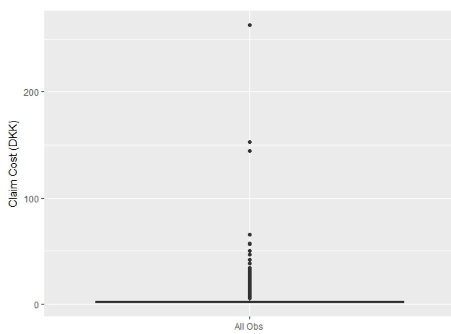
and with probability α_λ accept λ^* and set $\lambda^{(j)} = \lambda^*$ else reject λ^* and set $\lambda^{(j)} = \lambda^{(j-1)}$.

set j = 2 and return to step 1.[3] The algorithm in this thesis is set to generate 20000 observations of the parameters with the first 5000 being the burn in period.

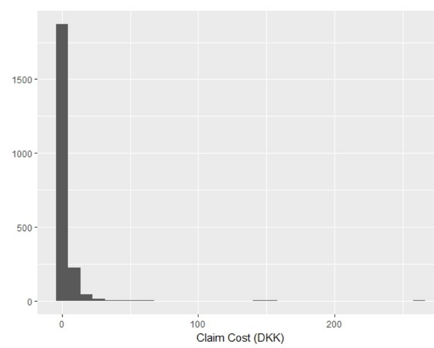
3 Data

The data used in this thesis is the "Danish Fire Claims" data set provided by most statistical packages for different software. Data consists of 2167 observations with a date stamp and the size of the claim between 3rd January 1980 and 31 December 1990 given by Mette Rytgaard at Copenhagen Reinsurance. The data are adjusted for inflation and are expressed in 1985 level Millions DKK. The whole set is used in this thesis without any checks for auto correlation or other time series related structure in the data.

Figure 1 illustrates the distribution of the Danish fire insurance claim data without any transformations. The figure clearly suggests that there is presence of heavy tails. Especially the box plot illustrates this fact. Figure 2 shows the QQplot for the claim cost and a exponential distribution together with a plot of the whole data set. Figure 2 also quite clearly suggests that there is a presence of a heavy tailed distribution.



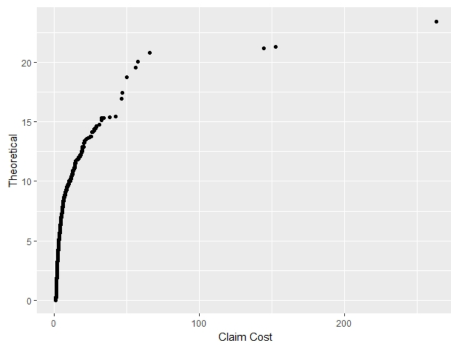
(a) Boxplot of Claim Cost



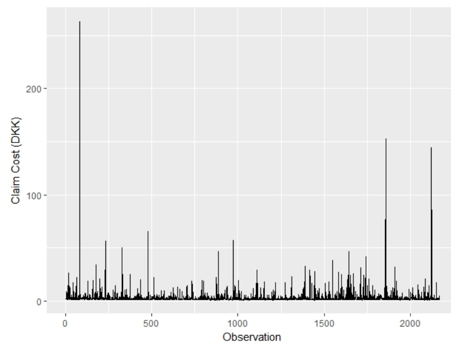
(b) Histogram of Claim Cost

Figure 1: Distribution of Danish Fire Insurance Claim Data

Figure 3 illustrates the sample densities both non-transformed and on the scale of the natural logarithm. The densities are non-parametric and although it is difficult to see, they suggest that the tails are "heavier" in both distributions. There are a few basic statistics presented in Table 1 which illustrate how the Claim Costs are distributed. The conclusion from this information is that there is likelihood for heavy tails and fitting some sort of distribution which allow for heavy tails is appropriate.

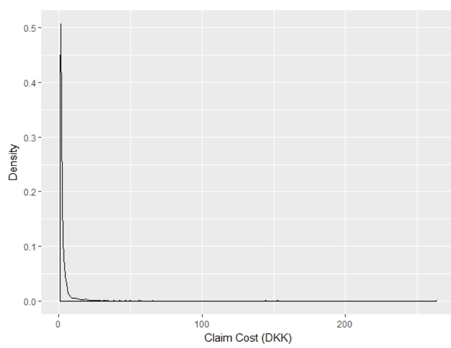


(a) QQ-plot

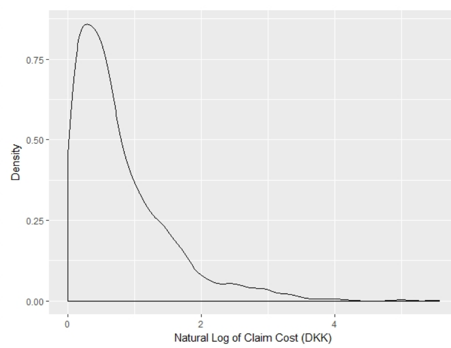


(b) Claim Cost Danish Fire Insurance

Figure 2: QQplot and Plot of Danish Fire Insurance Claim Data (10^6 DKK).



(a) Density Claim Cost



(b) Density Log of Claim Cost

Figure 3: Density Plots of Danish Fire Insurance Claim Data (10^6 DKK).

Statistic	Value
Mean	3.385
St.Dev	8.507
90% Quantile	5.542
95% Quantile	9.973
99% Quantile	26.043

Table 1: Statistics of Danish Fire Insurance Claims

4 Results

The estimates of the parameters and the results from running the algorithm in Section 2.8 are presented in this section. The algorithm was run in R 20000 times where the first 5000 runs were considered a burn in period. In addition to the algorithm, a mean excess plot is provided based on Section 2.4.

4.1 Mean Excess Plot

A Mean Excess Plots assists in evaluating whether the GPD distribution is a reasonable assumption for the data set or not.[8] Generally, one is looking for where and for which choices of the threshold u the Mean Excess Plot is linear. Linearity in the plot suggests that there is a good fit to a Pareto Distribution. Although Figure 4 is somewhat difficult to read, there is a suggestion of linearity in the Mean Excess function for the thresholds up to threshold value of 20. Note that the bulk of the distribution is below the very high threshold values.

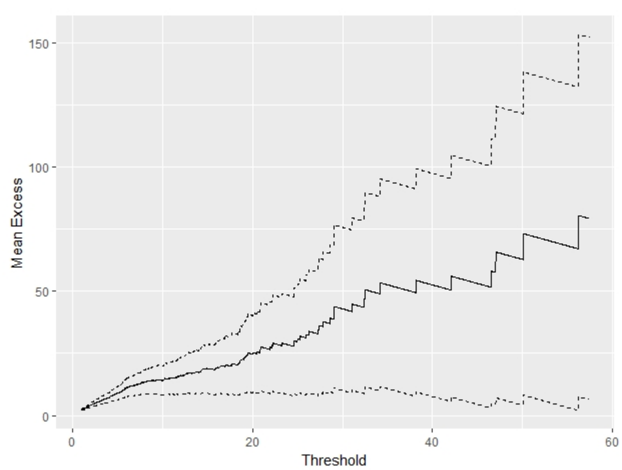


Figure 4: Mean Excess Plot Dansih Fire Insurance Claims

4.2 Distribution and Statistics for the Parameters

The priors for the parameters are presented in Table 2. The choice for the priors are arbitrary such that they have large variance but can satisfy the necessary conditions for the parameters. For example σ in a GPD needs to be positive hence a positive value is chosen with marginally normal distribution with unlikely negative observations but large enough variance to allow for the data to influence the posterior distribution. The priors were also chosen such that the GPD with the prior parameters do not have astronomically large values. Other priors can of course be chosen given that there is more information about the data. This is discussed in the discussion Section (5).

After running the MCMC algorithm the posterior distribution $\pi(\lambda, u, \mu, \sigma, \xi|\mathbf{x})$ is obtained. Figure 5 contains all the plots for the posterior distributions and Table 3 the mean and the standard deviation of the parameters. The densities in Figure 5 do not show any immediate suggestions for empirical parameters as the probability masses vary in a non predictive pattern. The distributions for σ and λ suggest a central tendency but with very large deviations making inference from the plots difficult. For the parameters u, μ and ξ the variance is quite large and probability mass not symmetric.

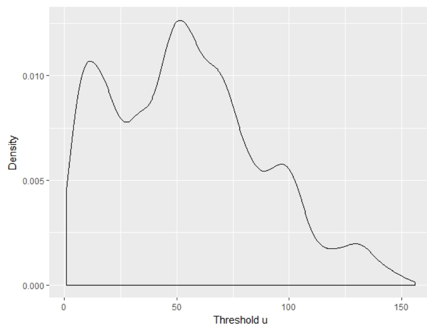
Table 2 presents the mean and standard deviation for the posterior parameter estimates. The main parameter of interest in this thesis is the threshold parameter which is set quite high considering that the 99% quantile is 26.043. The other parameters have, as the density plots suggested, large standard deviation which implies that there are difficulties in the definition of the parameters.

Prior Parameter	Parameters	Parameter Value
u	μ_u, σ_u	$\mu_u = 35, \sigma_u = 50$
λ	$d1, d2$	$d1 = 2, d2 = 2$
μ	$\mu_\mu,\ \sigma_\mu$	$\mu_\mu = 100,\ \sigma_\mu = 40$
σ	$\mu_\sigma,\ \sigma_\sigma$	$\mu_\sigma = 10,\ \sigma_\sigma = 2$
ξ	$\mu_\xi,\ \sigma_\xi$	$\mu_\xi = 1,\ \sigma_\xi = 5$

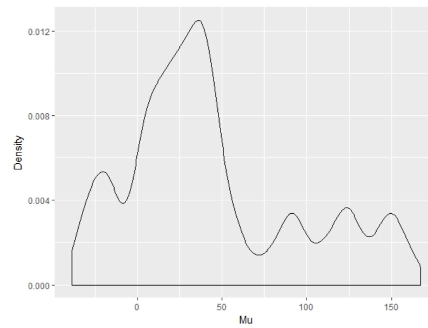
Table 2: Choice of Priors for the Parameters

Parameter	Mean	S.Dev
$\hat{u}$	54.635	33.526
$\hat{\lambda}$	$1.384 \cdot 10^{-6}$	$7.576 \cdot 10^{-6}$
$\hat{\mu}$	44.187	50.862
$\hat{\sigma}$	$1.000 \cdot 10^{66}$	$1.978 \cdot 10^{67}$
$\hat{\xi}$	20.470	20.768

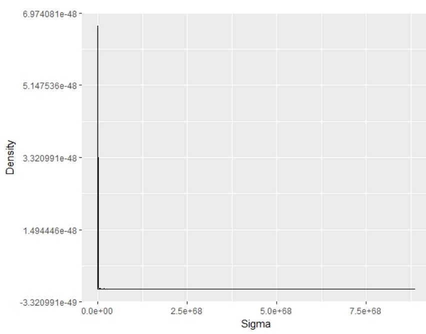
Table 3: Statistics of the Posterior Parameters



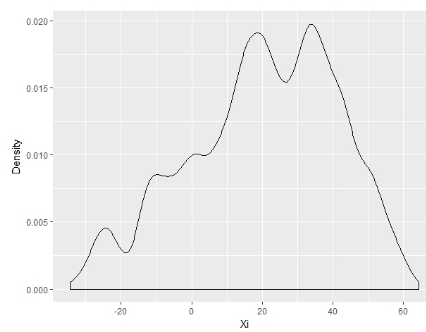
(a) Density for Threshold



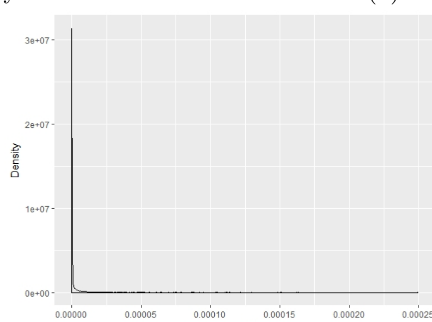
(b) Density for μ



(c) Density for σ



(d) Density for ξ



(e) Density for λ

Figure 5: The Posterior Distributions for Threshold, Kernel, and GPD Distribution Parameters

4.3 Application to Reinsurance Pricing

By applying the theory from Section 2.1 the price for a excess of loss reinsurance contract has a a straight forward calculation. There is however an important assumption that needs to be done. Using Equation (4) the price of a collective insurance portfolio is

$$E[S(t)] = \lambda t E[X_1], \tag{44}$$

so we need λ and $E[X_1]$. λ is the average number of claims for a time period. In this thesis we will use a year. The assumption required is that the claim arrivals are constant. This is not always the case but it is not central to this thesis hence the assumption of constant arrivals is set. This will suggest that approximately 0.3 claims above the threshold $u = 54.635$ will arrive every year. $E[X_1]$ is the expected value of the tail distribution with the parameters estimated after the MCMC. In this application $E[X_1]$ becomes very large and is estimated to

$$E[X_1] = 7.831 \cdot 10^{133} \tag{45}$$

giving the price of the Excess-of-Loss contract

$$E[S(t)_{XL}]) = 7.831 \cdot 10^{133} \cdot 0.3, \tag{46}$$

which is a impossibly high price to apply in practice. This pricing suggest that the GPD estimated heavy tails and a likelihood that the Claim Cost will be very high. The GPD on this form will not be applied in practice since it gives to large estimate of the average claim cost above the threshold to be realistic. Other considerations are taken into account such that it is impossible that a property can be worth so much. Further comments about this result are in the discussion section.

5 Discussion

This thesis aims to fit a distribution to an insurance claims data set, estimate a threshold for the tail distribution and use the information to estimate the price of a Excess of Loss insurance contract. The choice of a Bayesian analysis with a non-parametric distribution below the threshold and a GPD distribution in the tail is because because it refrains from visual methods for specifying the threshold. A MCMC algorithm is applied to estimate the posterior of all the parameters involved in the threshold analysis. These parameters were then used to estimate a price for a Excess of Loss reinsurance contract without upper limit.

The result of the thesis is that the posterior distributions are non-symmetric with large variance for the threshold and GPD parameters. This conclusion makes inference and interpretation difficult for the intended application. It is very difficult to estimate a reasonable price for the reinsurance contract with the estimated GPD parameters. Since the values of the parameters indicate a very long and fat tail, the price of the insurance contract becomes extremely large. Although the pricing will not be done this way in practice (one would apply maximum coverage and other distribution applications) it should be of use when analyzing how the distribution in the tail behaves and inference regarding whether or not a GPD for the tail is a good fit. Unfortunately the distribution in the tails gives a mean which approaches infinity and is useless for the estimation of the reinsurance price in this appliction. However, a different approach for the reinsurance price might be to use the threshold distribution to estimate a data-driven threshold and then apply some other distribution for the the part of the sample above the tail. One obvious approach is to use the empirical mean or the observations above the estimated threshold u from the MCMC algorthm.

There are some improvements that can be done in order to obtain a more reasonable result form the algorithm. Since the algorithm requires some prior information in terms of the prior distributions there is a possibility to tweak the distributions in order to specify the priors more correctly. For example, the GPD parameters were set as a uninformative trinomial normal distribution but could be really anything or expressed as quantiles from the sample. Another method could be to fix the prior parameters and estimate the rest of the using the same algorithm as outlined in Section 2. In addition to test different prior parameters and parameter distributions, the MCMC could be tweaked a little bit to specify the acceptance parameter differently as it decides if the proposal parameter should be accepted or not. Furthermore, the proposal distribution and variance could be experimented with in order to obtain reasonable results.

Expect obvious specification improvements in specification of the algorithm there can be done some changes to the data. It would for example be interesting to run the algorithm on the natural logarithm of the claim sizes to investigate if there is a presence of more smoothness and how that would impact the threshold and tail distribution. For now, the algorithm would be mainly of use in the estimation of the threshold density but not for the tail distribution.

References

[1] Lopes H.F. Gamerman D. Behrens, C.N. Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling 4* (3), 227–244, 2004.

[2] Bawa J. Trenner L. Coles, S. and P Dorazio. *An introduction to statistical modeling of extreme values.* Springer, 2001.

[3] Lee D. MacDonald A. Reale M. Russell G. Darlowb, B. and C.J. Scarrott. A flexible extreme value mixture model. *Computational Statistics and Data Analysis* 55 2137–2157, 2011.

[4] Lundberg F. *Approximerad framställning av sannolikhetsfunktionen. Återförsäkring av kollektivrisker.* Akad. Afhandling. Almqvist och Wiksell, Uppsala, 1909.

[5] Dellaportas P. Roberts G.O. An Introduction to MCMC. In: Møller J. (eds) *Spatial Statistics and Computational Methods. Lecture Notes in Statistics,* vol 173. Springer, New York, NY, 2003.

[6] Wand M.P. Jones, M.C. *Kernel Smoothing.* Chapman and Hall/CRC, 1995.

[7] A. MacDonald and C. Scarrot. A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *Revstat - Statistical Journal* 10(1):33-60, 2012.

[8] A. McNeil. Estimating the Tails of Loss Severity Distributions Using Extreme Value. *ASTIN Bulletin*, 27(1):117–137, 1997.

[9] T. Mikosh. *Non-Life Insurance Mathematics : An Introduction with the Poisson Process.* Springer, 2009.

[10] M. Naveau, P. Vrac. Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resourc. Res.*, 43(W07402), 2007.

[11] Reale M. Scarrott C.J. Zhao X. Oxley, L. Extreme value modelling for forecasting the market crisis. *Appl. Fin. Econ.*, 20, 63–72, 2010.

[12] Sheldon M. Ross. *Introduction to Probability Models.* Academic Press, 2010.

[13] J. Rychlik, I. Rydén. *Probability and risk analysis: An introduction for engineers.* Springer, 2006.