Stockholms universitet

Tail Estimation: a Comparative Simulation Study of Extreme Value Theory and Importance Sampling

Betty Frankl

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2024:8
Matematisk statistik
Maj 2024

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Bachelor Thesis **2024:8**
http://www.math.su.se

Tail Estimation: a Comparative Simulation Study of Extreme Value Theory and Importance Sampling

Betty Frankl*

May 2024

Abstract

Extreme value theory was developed specifically for analysing extreme events that failed to be analysed by conventional methods. It is difficult to find alternative methods that could be used in practice. However, in a simulation study where the underlying distribution is known, there are alternative methods that could be used. Importance sampling is a Monte Carlo variance reduction technique, that can be used to handle rare events. In this simulation study we will compare these two methods in order to highlight differences and gain deeper knowledge.

We focus on tail estimation and place particular emphasis on the extreme right tails. In these cases there are typically very few, or no, observations in the area of interest. The methodology we use from extreme value theory is the peak over threshold model and the generalized Pareto distribution (GPD) method. For importance sampling we use an extreme order biasing technique.

The comparison of the methods reveals key differences in their approaches and performance. Importance sampling proves both effectiveness, compared to Monte Carlo, and reliability, due to its ability to generate unbiased estimates. However, both importance sampling and Monte Carlo show their limitations for small sample sizes, often resulting in zero estimates. The GPD method exhibits variability in performance across distributions and sample sizes. The deviation from the true values are high in many cases. Importantly, though, the GPD method has the capacity to produce non-zero estimates which is significant for real-world applications where zero probability assumptions may not be valid.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: betty.frankl@gmail.com. Supervisor: Johannes Heiny Ola Hössjer.

Acknowledgement

First of all, I want to thank my beloved family for their unwavering support and patience throughout this journey. A heartfelt thank you to Pablo Frankl, whose presence and kindness have been invaluable.

I have been very lucky to have such incredible supervisors, Johannes Heiny and Ola Hössjer. Johannes, who helped me come up with such a great thesis idea, has also contributed with his expertise in extreme value theory, greatly benefiting the theoretical work. Ola has been a good sounding board where I have been able to discuss things, and he also provided me with invaluable feedback.

I should also mention ChatGPT, who helped me a lot with my English, generating plots, and handling tables, although our collaboration has sometimes been challenging.

Lastly, I want to thank my amazing friends and colleagues who have made this period incredibly fun, especially with all dancing nights and Friday swims in Brunnsviken. A special thank you to Lars Lidvall whose help and support have meant the world to me.

Contents

1 Introduction

Analysing extreme events holds significant importance across various fields, such as insurance and finance. In practice, these events are of great importance since they are hard to predict and can cause serious consequences. The challenge of analysing these events, though, is the limited data available for analysis, making conventional methods unsuitable. To address this issue, researchers have developed extreme value theory and specific methods to analyse these events. Within this theory there are different approaches, for example the peak over threshold (POT) model and block maxima model.

There are multiple ways to analyse and questions to address when dealing with extreme values. In this study, however, we will specifically concentrate on tail estimation, placing emphasis on the extreme right cases, meaning tails such that the tail probability is very small. Additionally, we examine distributions with different tail behaviours, or in different *maximum domains of attraction* (MDA), to investigate potential variations.

In practical applications, it is difficult to find alternatives to extreme value theory when estimating these extreme right tails. On the other hand, in a simulation study where we know the underlying distribution, there are some alternative methods that could be used. One of these is importance sampling, a Monte Carlo method that can be used to handle the challenges posed by rare events. The idea is to sample from another distribution and thereby reduce the variance of the regular Monte Carlo estimator. This enables us to achieve the same results as Monte Carlo with less data.

In this thesis we will compare extreme value theory, or more specifically the POT model and the generalised Pareto distribution (GPD) method, and important sampling for tail estimation. We will include Monte Carlo as well. It should be mentioned that the comparison of these methods is a bit tricky since they are based on different assumptions. Importance sampling requires knowledge about the underlying distribution since it is a simulation based technique used to estimate statistical properties of a particular distribution. In the GPD method we will not assume the underlying distribution. Despite this disparity, comparing these methods is valuable from a theoretical point of view, since this can offer crucial insights and deepen our understanding of both the GPD method and importance sampling.

The primary references for this thesis is Embrechts et al. [3] for extreme value theory, and Srinivasan [11] for importance sampling.

2 Extreme Value Theory

In this section, we will first introduce two central results; the Fisher-Tippet theorem and the Pickands–Balkema–de Haan theorem. If not explicitly stated otherwise, the reference for the first two subsections is Chapter 3.1-4 in [3], while Chapter 6.5 from the same source serves as the reference for the remaining parts of this section.

2.1 Fisher-Tippet Theorem

We begin by examining the sample maxima. Let $X, X_1, X_2, ...$ denote a sequence of independent and identically distributed non-degenerate random variables with distribution function F. Non-degenerate means that there is no value x_0 for which $\mathbb{P}(X = x_0) = 1$. The *sample maxima* are given by

$$M_n = \max(X_1, ..., X_n), \quad n \in \mathbb{N}.$$

Since the random variables are independent and identically distributed we can easily obtain the distribution function of M_n as follows

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, ..., X_n \leq x) = F^n(x). \tag{1}$$

As we are interested in extremes, we want to focus on the right tail of the distribution function (the left tail can be addressed analogously). The right endpoint of F is defined by

$$x_F = \sup(x \in \mathbb{R} : F(x) < 1).$$

If we now consider the behavior of M_n it can be shown that, for $x_F \leq \infty$, the following applies:

$$\mathbb{P}(M_n \leq x) = F^n(x) \to \begin{cases} 0 & \text{if } x < x_F \\ 1 & \text{if } x \geq x_F \end{cases} \qquad \text{as } n \to \infty.$$

Thus M_n converges in probability to x_F as $n \to \infty$. We know that the sequence M_n is non-decreasing, thus M_n converges almost surely to x_F under the same conditions.

However, this fact does not provide a lot of information. As Coles writes, knowledge about F is usually missing in practice and small discrepancies in an estimate of F can lead to substantial discrepancies for F^n [2, p. 45-46]. Even if we knew the exact distribution F, we could not draw any significant conclusions from this, since M_n converges to a degenerate distribution.

Hence we cannot say anything about quantiles, confidence intervals or other interesting statistical properties.

An alternative strategy is to accept that F is unknown and instead look for limit distributions for a transformation of M_n. This approach closely resembles the approach in the central limit theorem.

Theorem 2.1.1 (Central limit theorem)
Let $S_n = X_1 + X_2 + ... + X_n$ be a sum of independent and identically distributed random variables with finite expected value μ and variance σ. Then the following holds

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty,$$

where $\xrightarrow{d}$ refers to convergence in distribution [5, p. 162].

In a similar way we now want to examine the centred and normalised M_n. Let us consider probabilities of the form

$$\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) = \mathbb{P}\left(M_n \leq c_n x + d_n\right),$$

where $d_n \in \mathbb{R}$ are the centering constants and $c_n > 0$ are the normalising constants.

Theorem 2.1.2 (Fisher-Tippet theorem)
Let $\{X_i\}_{i=1}^n$ be a sequence of independent and identically distributed random variables, with distribution F. If there exist norming constants $c_n > 0$, $d_n \in \mathbb{R}$, and a non-degenerate distribution function H, such that

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H, \quad \text{as } n \to \infty.$$

Then it follows that H belongs to one of the following three distribution functions

$$\textit{Fréchet}: \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & x > 0 \end{cases} \qquad \alpha > 0.$$

$$\textit{Weibull}: \quad \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{\alpha}), & x \leq 0 \\ 1, & x > 0 \end{cases} \qquad \alpha > 0.$$

$$\textit{Gumbel}: \quad \Lambda(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}.$$

See [3, p. 121] for a statement of this result, and [9, p. 9-11] for a proof.

The three types of distributions, Φ_α, Ψ_α, and Λ, presented in Theorem 2.1.2, are referred to as extreme value distributions.

For a distribution F, if it satisfies the conditions outlined in Theorem 2.1.2, we say that F belongs to the maximum domain of attraction (MDA) of the corresponding extreme value distribution. We write this, in for example the Gumbel case, as $F \in \text{MDA}(\Lambda)$.

It turns out that it is convenient to use a one-parameter representation of the standard extreme value distributions.

Definition 2.1.1 (The generalised extreme value distribution (GEV)) *Let H_ξ be the distribution function defined by*

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-\frac{1}{\xi}}) & \text{if} \quad \xi \neq 0, \\ \exp(-e^{-x}) & \text{if} \quad \xi = 0, \end{cases}$$

where $1 + \xi x > 0$.

The parameter ξ is referred to as the shape parameter and it corresponds to α by

$$\begin{cases} \text{Fréchet}(\Phi_\alpha): & \xi = \alpha^{-1} > 0, \\ \text{Gumbel}(\Lambda): & \xi = 0, \\ \text{Weibull}(\Psi_\alpha): & \xi = -\alpha^{-1} < 0. \end{cases} \tag{2}$$

2.2 POT & Pickands–Balkema–de Haan Theorem

Before we can present our next central result, the Pickands–Balkema–de Haan theorem, we will introduce some necessary theoretical framework.

In extreme value theory there are two main models used to model extremes; the block maxima model and the POT model. In the block maxima model, the data is divided into blocks and the maximum within each block is selected for analysis. The POT model selects all values above a certain threshold. This model is often regarded as more efficient in utilizing data [4] and the associated theory is also very appealing when dealing with tail estimation. As a consequence, we will focus on the POT model in this thesis.

Letting $N_u = \sum_{i=1}^{n} \mathbb{1}(X_i \geq u)$, we denote the excesses over a threshold u by $Y_1, Y_2, ...Y_{N_u}$. Accordingly we have that $Y_i = X_i - u$ for all i such that $X_i > u$. In Figure 1 we can see an illustration of the POT model, which is based on excesses over the threshold u.

The Peak Over Threshold Model (POT)

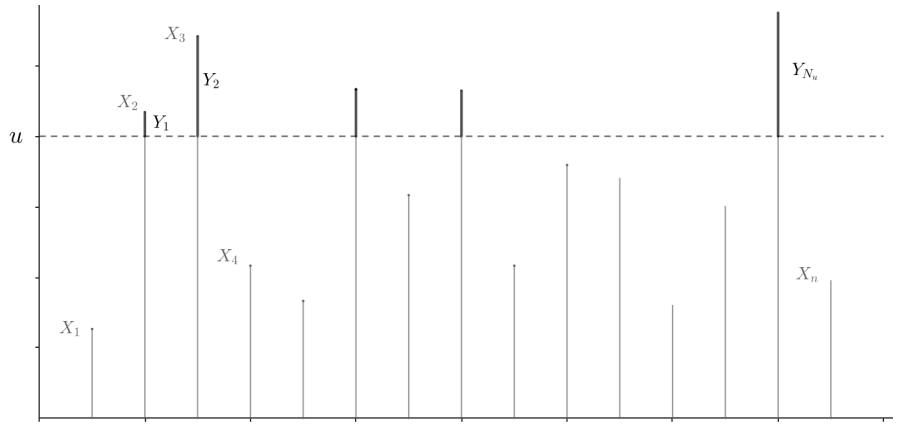

Figure 1: Independent and identically distributed variables $X_1, X_2, ..., X_n$ and excesses $Y_1, Y_2, ..., Y_{N_u}$ over the threshold u.

Definition 2.2.1 (Excess distribution function)
The excess distribution function F_u of X, or analogously F, given the threshold $u < x_F$ is given by

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u), \quad x \geq u.$$

Definition 2.2.2 (Generalised Pareto distribution (GPD))
The distribution function $G_{\xi,\beta}$ of the GPD, where $\xi \in \mathbb{R}$ and $\beta > 0$ is given by

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \frac{\xi x}{\beta})^{\frac{-1}{\xi}} & if \quad \xi \neq 0, \\ 1 - e^{\frac{-x}{\beta}} & if \quad \xi = 0, \end{cases} \qquad x \in D(\xi, \beta),$$

where

$$D(\xi, \beta) = \begin{cases} [0, \infty[& if \quad \xi \geq 0, \\ [0, \frac{-\beta}{\xi}] & if \quad \xi < 0. \end{cases}$$

The parameters ξ and β are known as the shape and scale parameters, respectively [8, p. 275].

Now that we have all the necessary tools at our disposal, we are ready to introduce the next theorem.

Theorem 2.2.1 (Pickands–Balkema–de Haan theorem)
Let $\xi \in \mathbb{R}$. We can find a positive function β such that

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

if and only if $F \in MDA(H_\xi)$.

See [8, p. 277] for a statement of this result, and the proof can be found in [7].

What Theorem 2.2.1 essentially tells us is that we can approximate F_u with the GPD. However, it is important to note that this approximation requires a sufficiently large threshold u, and the estimation of β depends on the choice of u. Therefore, careful consideration is needed when selecting the threshold.

2.3 Estimation Excess Over Threshold

First, let us denote $\overline{F}(x) = 1 - F(x)$, which is the right tail of F. We want to estimate $\bar{F}(x) = \delta$, where δ is a small positive number, for example 0.0001.

Theorem 2.2.1 tells us how we can approximate the exceedance over some high threshold u. Now we will show how to use this to find the tails of interest. From Definition 2.2.1 we get that, for $y > 0$

$$\begin{aligned} F_u(y) &= \mathbb{P}(X \leq y + u | X > u) \\ &= \frac{\mathbb{P}(X \leq y + u, X > u)}{\mathbb{P}(X > u)} \\ &= \frac{F(y+u) - F(u)}{1 - F(u)}. \end{aligned} \tag{3}$$

We can see this illustrated in Figure 2.

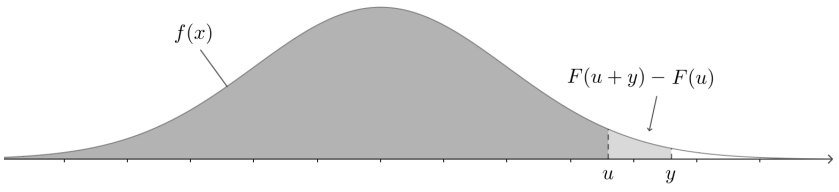

Figure 2: An example of a density function $f(x)$ and its relation between the threshold u, excess y and the corresponding distribution function $F(x)$.

Our main focus will now be on $\overline{F}(y+u)$. From (3) we get that

$$F_u(y) = \frac{\overline{F}(u) - \overline{F}(y+u)}{\overline{F}(u)} \iff \overline{F}(y+u) = \overline{F}(u)\overline{F}_u(y).$$

Now we have a nice expression for $\overline{F}(y+u)$ from which we will be able to proceed. For the remainder of this section, we employ the same method as

described in [3, p. 352-358]. Following this approach, we proceed by estimate $\overline{F}(u)$ and $\overline{F}_u(y)$ separately to achieve an approximation of $\overline{F}(y+u)$.

It comes natural to use the empirical distribution function to estimate $\overline{F}(u)$. Hence, we get that

$$\widehat{\overline{F}(u)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq u) = \frac{N_u}{n}.$$

Theorem 2.2.1 tells us that $F_u(y) \approx G_{\xi,\beta(u)}(y)$ which is equivalent to $\overline{F}_u(y) \approx \overline{G}_{\xi,\beta(u)}(y)$. We can approximate $\overline{G}_{\xi,\beta(u)}$ by estimating the parameters ξ and β and insert them into $\overline{G}$. In section 2.5 we will describe this in more detail.

Assuming that we have obtained the estimates $\hat{\xi}$ and $\hat{\beta}$, we get the approximation

$$\widehat{\overline{F}_u(y)} = \overline{G}_{\hat{\xi},\hat{\beta}}(y).$$

Recalling the relationship $\overline{G}_{\xi,\beta}(x) = 1 - G_{\xi,\beta}(x)$ we finally get that

$$\widehat{\overline{F}(y+u)} = \begin{cases} \frac{N_u}{n}\left(1 + \frac{\hat{\xi}y}{\hat{\beta}}\right)^{\frac{-1}{\hat{\xi}}}, & \hat{\xi} \neq 0, \\ \frac{N_u}{n} e^{\frac{-y}{\hat{\beta}}}, & \hat{\xi} = 0. \end{cases}$$

2.4 Threshold Selection

The selection of u involves a trade-off between high variance and high bias of the GPD parameter estimates $\hat{\xi}$ and $\hat{\beta}$. A low threshold value tends to result in high bias, whereas a high threshold value leads to high variance. While theoretically, achieving an optimal bias-variance trade-off is possible, Embrechts et al. [3, p. 355-356] argue that, in practice, we need a different approach. They recommend a graphical approach, involving the study of plots of the empirical mean excess function and parameter estimates for various threshold values. Coles [2] also recommend a similar method.

Definition 2.4.1 (Mean excess function)
The mean excess function $e(u)$ of X, or analogously F, given the threshold $u < x_F$ is given by

$$e(u) = E(X - u | X > u).$$

For a random variable X with a GPD and parameters $\xi < 1$ and $\beta > 0$ the following holds for $u < x_F$

$$e(u) = \frac{\xi u + \beta}{1 - \xi} = \frac{\xi}{1-\xi}u + \frac{\beta}{1-\xi}, \quad \xi u + \beta > 0.$$

Thus the mean excess function is linear in this case.

Consider $X, X_1, ..., X_n$ with any distribution that belongs to some MDA with $\xi < 1$. Note that $X - u = Y$ for $X > u$. Given the sample $Y_1, Y_2, ..., Y_{N_u}$ the empirical mean excess function is given by

$$e_n(u) = \frac{1}{N_u} \sum_{i=1}^{N_u} Y_i, \quad u > 0.$$

With Theorem 2.2.1 in mind, the graphical approach to selecting u will involve choosing a sufficiently large value for u such that $e_n(x)$ exhibits approximately linear behavior for $x \geq u$. However, this is not an easy task, and one should not expect to find a unique choice of u. Hence, this method should be combined with other graphical techniques. We will see these techniques demonstrated in the simulation study (Section 4).

2.5 GPD Parameter Estimation

There are multiple ways to estimate the parameters. Unfortunately, there is no single method to obtain a reliable result. As Embrechts et al. [3] write, it is necessary to consider multiple approaches. There are several simple methods that can be applied. The most common one is probably the maximum likelihood estimator (MLE), which is mostly considered reliable. It is important to note, though, that this method is only suitable for $\xi > -0.5$. Two other common methods are probability weighted moments (PWM) and method of moments. However, these methods are only suitable for $\xi \geq 0$, which means that they are not appropriate to use for a distribution in MDA(Ψ). For cases where $\xi < 0$, alternative methods like generalized PWM exist, but they are more difficult to implement [1].

The choice of method for estimation is indeed important for the final result. However, due to the scope of this thesis, which focuses on comparing two different methods for high quantile estimation across distributions within each MDA, we will limit ourselves to one estimation method. The most neutral choice here appears to be the MLE, given its widespread use and relatively reliable performance across all three types of MDAs.

The following result on the MLE is taken from [2, p. 80-81]. We obtain the MLEs $\hat{\xi}$ and $\hat{\beta}$ of the GPD by maximizing the log-likelihood function with respect to ξ and β, given the sample $\boldsymbol{y} = \{y_1, y_2, \ldots, y_{N_u}\}$. For $\xi \neq 0$ the log-likelihood function is given by

$$\ell(\xi, \beta; \boldsymbol{y}) = -N_u \ln(\beta) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \ln\left(1 + \frac{\xi}{\beta} y_i\right)$$

assuming that $1 + \frac{\xi}{\beta} y_i > 0$ for all $i \in \{1, 2, ..., N_u\}$. If $\xi = 0$ we obtain the following log-likelihood function

$$\ell(\beta; \boldsymbol{y}) = -N_u \ln(\beta) - \frac{1}{\beta} \sum_{i=1}^{N_u} y_i.$$

Unfortunately it will not be possible to maximize these log-likelihood functions analytically, but as we will see later in the simulation study (Section 4), we can do this numerically using R.

Another advantage of using the MLE is its asymptotic normality. This enables us to construct confidence intervals, which we will use for the plots in the simulation study (Section 4).

3 Importance Sampling

Importance sampling is a Monte Carlo based simulation method that hastens the occurrences of rare events and thereby improves the Monte Carlo method in these cases. Srinivasan [11] demonstrates how both Monte Carlo and importance sampling can be used for tail estimation. The results in this section are mainly taken from this reference. We start by investigating how the Monte Carlo method can be used for high quantile estimation. After that, we will try to improve the method by importance sampling. Note that we use the same notation as previously and all random variables $X, X_1, X_2, ...$ remain independent and identically distributed.

3.1 Monte Carlo Tail Estimation

Our goal is to estimate $\overline{F}(t) = 1 - F(t) = \mathbb{P}(X \geq t) = \delta$ where t is the threshold value such that the event $x > t$ is rare. Srinivasan [11, p. 1-2] describes how the the Monte Carlo method can be used in these cases.

The Monte Carlo method is based on conducting multiple independent identically distributed Bernoulli trials. In the tail estimation case, the Bernoulli variable corresponds to the indicator function $\mathbb{1}(X \geq t)$, which returns 1 if the the observation is above t and 0 otherwise. Say that we use n trials, this gives us a sequence $\{X_i\}_{i=1}^n$ where each one of the random variables has the same success probability δ. Let N_t denote the number of n trials that lie above t. Then N_t has a binomial distribution with the same success probability δ. The corresponding density is given by

$$\mathbb{P}(N_t = k) = \binom{n}{k} \delta^k (1 - \delta)^{n-k}, \quad k \in \{0, 1, ..., n\}. \tag{4}$$

The MLE $\hat{\delta}_{MC}$ is obtained by maximizing the probability function (4) with respect to δ. We find the maximum by determining the value of δ for which

the first derivative of the log-likelihood is zero, indicating a critical point. Additionally, to ensure that this critical point corresponds to a maximum, we verify that the second derivative is negative. This gives us the following result:

$$\hat{\delta}_{MC} = \frac{N_t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq t).$$

This estimate is unbiased, meaning that its expected value is equal to the true value of the parameter. This can be shown by standard calculations for expected values. Using the variance formula for a binomial distribution and standard calculations for variances we get that

$$\text{Var}(\hat{\delta}_{MC}) = \text{Var}\left(\frac{N_t}{n}\right) = \frac{1}{n}(\delta - \delta^2). \tag{5}$$

When δ is close to zero we get that

$$\text{Var}(\hat{\delta}_{MC}) \approx \frac{\delta}{n}.$$

When δ is small there is a low probability we will have observations above t. If our sample does not provide any observations in this specific area, $\hat{\delta}_{MC}$ will end up being zero, which does not tell us much. If we happen to obtain a non-zero estimate it might still not be very informative due to a large variance of $\hat{\delta}_{MC}$. In practice we will need a very large sample size n to achieve low variance with the Monte Carlo method (see [11, p. 2] for more details).

3.2 Importance Sampling Tail Estimation

Importance sampling serves as an alternative approach to the regular Monte Carlo method. Importance sampling offers a potential solution to the issue of high variance of the estimate $\hat{\delta}_{MC}$. Srinivasan [11, p. 2-3] describes the method in the tail estimation case.

The basic idea of importance sampling is to use a *biasing density* $f_*(x)$ that allows rare events to occur more frequently. This enables us to improve our estimate $\hat{\delta}_{MC}$ by lowering its variance. Note that the distribution of X is referred to as the *target distribution*. Using the previous notation, we know that δ is equal to the expected value of the Bernoulli variable. Thus the

following holds

$$\begin{aligned}\delta &= E\left(\mathbb{1}(X \geq t)\middle| X \sim f(x)\right) \\ &= \int \mathbb{1}(x \geq t) f(x)\, dx \\ &= \int \mathbb{1}(x \geq t) \frac{f(x)}{f_*(x)} f_*(x)\, dx \\ &= E\left(\mathbb{1}(X \geq t) \frac{f(X)}{f_*(X)} \middle| X \sim f_*(x)\right). \end{aligned} \tag{6}$$

From now on, to simplify the notation, let us denote:

$$E_*(X) = E(X|X \sim f_*(x)), \quad E(X) = E(X|X \sim f(x))$$

and corresponding notation for the variances. Also, let

$$W(x) = \frac{f(x)}{f_*(x)}.$$

The new expected value (6) motivates to use the Monte Carlo method and estimate δ in the following way

$$\hat{\delta}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \geq t) W(X_i),$$

where $X_1, \ldots, X_n$ are independent and identically distributed random variables with density f_*. As in the regular Monte Carlo case, this estimate is also unbiased (as shown in (6)). In order to derive the variance of $\hat{\delta}_{IS}$ we first use the fact that all random variables X_i are independent and identically distributed and then the property $\mathrm{Var}(X) = E(X^2) - E^2(X)$. This gives us that

$$\begin{aligned}\mathrm{Var}_*(\hat{\delta}_{IS}) &= \mathrm{Var}_* \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \geq t) W(X_i) \right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_* \left(\mathbb{1}(X \geq t) W(X) \right) \\ &= \frac{1}{n} \mathrm{Var}_* \left(\mathbb{1}(X \geq t) W(X) \right) \\ &= \frac{1}{n} \left(E_* \left(\mathbb{1}^2(X \geq t) W^2(X) \right) - E_* \left(\mathbb{1}(X \geq t) W(X) \right)^2 \right). \end{aligned} \tag{7}$$

The importance sampling problem now comes down to finding a biasing density such that $\mathrm{Var}(\hat{\delta}_{IS})$ is lower than $\mathrm{Var}(\hat{\delta}_{MC})$.

3.3 Optimal Biasing

Srinivasan [11, p. 4-7] shows how the the optimal biasing density is derived. By Jensens Inequality we have that

$$E_* \left(\mathbb{1}^2(X \geq t) W^2(X) \right) \geq E_* \left(\mathbb{1}(X \geq t) W(X) \right)^2 . \tag{8}$$

Since our aim is to lower the variance in (7) we want to reduce the left-hand side in (8) as much ass possible. If we find a biasing density f_* such that the equality in (8) holds, the variance will become zero. Thus, such a biasing density must be optimal. Since

$$E_* \left(\mathbb{1}(X \geq t) W(X) \right)^2 = E \left(\mathbb{1}(X \geq t) \right)^2 = \delta^2$$

the equality in (8) holds if and only if

$$\mathbb{1}(X \geq t) W(X) = \delta, \quad \text{for} \quad X \sim f_*(X).$$

From the definition of W we get that the optimal biasing density is given by

$$f_*^{opt}(x) = \frac{1}{\delta} \mathbb{1}(x \geq t) f(x). \tag{9}$$

This can be verified by simply replacing $f_*(X)$ in $W(X)$ by $f_*^{opt}(X)$ in (8).

The optimal biasing density concentrates all its probability mass above t, yet remains proportional to the target density f by the factor δ. However, if we knew δ, there would be no point of estimating it in the first place. Hence we will assume that we do not know this proportionality constant. Hence we need to figure out another biasing density that still reduces the variance in (7).

3.4 Biasing Methods

There are numerous methods for selecting biasing density. Srinivasan [11, Chapter 4] summarizes a few of the most commonly used ones, demonstrates how these can be improved and proposes an alternative which employs extreme order statistics. The author shows how adjusting the methods appropriately can increase the number of observations in the tail, and thereby substantially reduce the variance of $\hat{\delta}_{IS}$.

The author also introduces the concept of blind biasing, which involves biasing without knowledge of the target distribution f. In theory, this would be the ideal method for making a fair comparison to the extreme value approach. However, blind biasing remains a subject of ongoing study, and

current research does not provide evidence of variance improvement. Therefore, we will proceed with biasing methods that utilize knowledge of the target distribution.

In order to asses the performance of a method, the *simulation gain*, denoted Γ, is used. It is defined as the ratio between the number of observations, in the Monte Carlo case, required to achieve a certain level of variance, and the corresponding number in the Importance Sampling case. In the context of tail estimation the simulation gain [11, p. 8] can be expressed as

$$\Gamma = \frac{\delta(1-\delta)}{E_*\left(\mathbb{1}^2(X \geq t)W^2(X)\right) - \delta^2}. \tag{10}$$

Note that this ratio also corresponds to the ratio between the variance of $\hat{\delta}_{MC}$ and the variance of $\hat{\delta}_{IS}$.

The simulation gain is not a perfect measure of the biasing density performance. Firstly, since we do not know δ we cannot calculate its actual value. Instead we will have to estimate it by first estimating the variances. Secondly, we cannot solely rely on the simulation gain to completely assess the performance of the various biasing methods, as their effectiveness may vary across different distributions and thresholds.

When selecting a biasing method for this thesis, I have examined simulation gain in both specific examples and in more general cases for different methods. Unfortunately, there is limited literature available on the comparison of these methods and their performance. Therefore, I have drawn conclusions from [11, Chapter 4] by taking part of the authors ideas, taken optimal biasing into consideration and analysing the simulation gain to evaluate the different methods.

The choice of biasing density is crucial for the performance of importance sampling. The optimal approach would naturally involve assessing several different methods. However, this strategy would be time-consuming, especially if we were to further refine the methods to maximize their performance. Therefore, our aim is to select one method that remains sufficiently effective.

Although some of the methods may seem very promising, they might be very complicated to implement. Thus, to make a trade-off between complexity and potential for variance improvement, we conclude that the proposed biasing method based on extreme order statistics, is the most suitable choice for this study. In Figure 3, we demonstrate and compare an extreme order biasing method with optimal biasing.

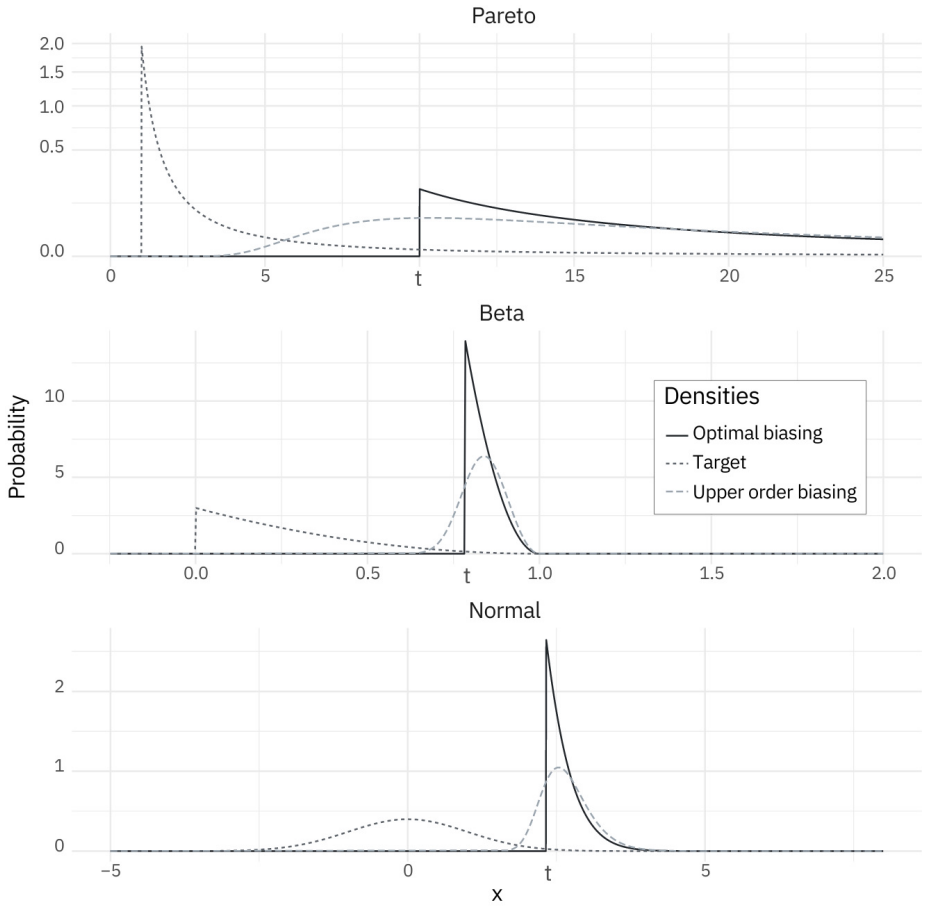

Figure 3: Different biasing methods for Pareto distribution with location parameter 1 and shape parameter 2, standard normal distribution and beta distribution with parameters 1 and 3. These are the three distributions we will use in the simulation study, for more details, see Section 4.1. The threshold t is chosen such that $\delta = F(t) = 10^{-4}$ for each distribution. Note that the y-axis in the Pareto plot is scaled using a square root transformation to improve readability. Note that upper order density biasing is one of the two useful extreme order biasing methods (see Section 3.5).

3.5 Extreme Order Density Biasing

There are essentially two useful biasing methods considering extreme order densities; the 1st upper order density, $\max(X_1, ..., X_n)$ or the 1st lower order density, $\min(X_1, ..., X_n)$. Since variance reduction in theses cases does not depend on the target distribution we can easily compare the different methods. The comparison clearly shows that the upper order density is

the superior choice (for details, see [11, p. 43]). Hence, we now proceed to describe the upper order biasing method.

Since the performance of the upper order biasing method is independent of the target distribution, given a threshold value t, the result will not depend on the tail behavior, or which MDA it belongs to. Instead, the performance depends on the number of observations used. As we shall see there is an optimal number of observations n_{opt}, for which the variance of $\hat{\delta}_{IS}$ is reduced maximally compared to regular Monte Carlo. The variance improvement of this method will also subside as $n \to \infty$. Note that n_{opt} increases as $t \to x_F$, but so does the maximum simulation gain. The results in this section can be found in [11, p. 38-46]. In Figure 4 we see how the simulation gain depends on n for a selection of δ values.

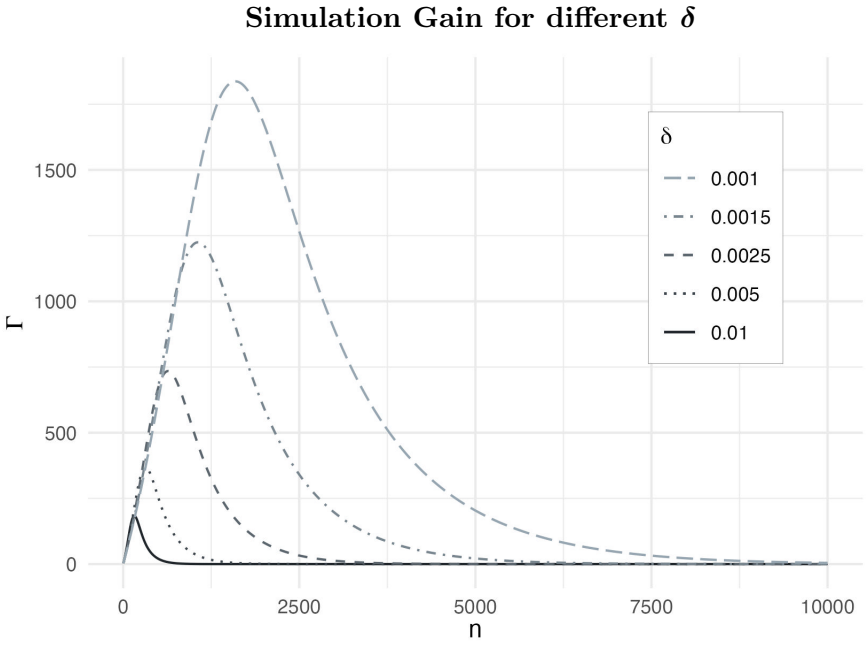

Figure 4: Simulation gain Γ of the upper order biasing method for different tail probabilities δ.

The upper order density of n random variables is derived the same way as M_n (1) in Section 2.1. Remember that all X_i are independent and identically distributed. We thus get that

$$F_*(x) = F^n(x) \Longrightarrow f_*(x) = nF^{n-1}(x)f(x),$$

where f_* is the derivative of F_*.

Maximizing the simulation gain (10) is equivalent to minimizing the quantity

$$I = E_* \left(\mathbb{1}^2(X \geq t) W^2(X) \right).$$

As we shall see, $I = I(n)$ is a function depending on the number of random variables n used, for a given δ. We can derive the expression for $I(n)$ in the following way

$$\begin{aligned} I(n) &= E_* \left(\mathbb{1}^2(X \geq t) W^2(X) \right) \\ &= \int_{-\infty}^{\infty} \mathbb{1}^2(x \geq t) W^2(x) f_*(x)\, dx \\ &= \int_t^{\infty} \frac{f^2(x)}{f_*(x)}\, dx = \frac{1}{n} \int_t^{\infty} f(x) F^{1-n}(x)\, dx \\ &= \begin{bmatrix} F = F(x) \\ dF = f(x)dx \\ t \to F(t) \\ \infty \to 1 \end{bmatrix} = \frac{1}{n} \int_{F(t)}^{1} F^{1-n}\, dF. \end{aligned} \tag{11}$$

From standard integral calculations we conclude that

$$I(n) = \frac{1}{n} \left[\frac{F^{2-n}}{2-n} \right]_{F(t)}^{1} = \frac{1}{n} \left(\frac{1}{2-n} - \frac{F^{2-n}(t)}{2-n} \right) = \frac{1 - (1-\delta)^{2-n}}{n(2-n)}, \tag{12}$$

where we use the equality $F(t) = 1 - \delta$ in the last step. Since the expression in (12) only depends on n and δ we draw the conclusion that the simulation gain (10) for the upper order biasing method is independent of the target distribution given a certain tail probability δ, or analogously, threshold t. Through simple, but tedious calculations, one can also show that $I(n)$ is strictly convex when $n > 0$ and thereby has a unique minimum.

We can also rewrite (12) to obtain an expression that does not depend on δ. By using the fact that $F(t) = 1 - \delta$ we get that

$$I(n) = \frac{F(t)^{2-n} - 1}{n(n-2)}. \tag{13}$$

In Table 1 we see a selection of n_{opt} values and the corresponding simulation gain for different tail probabilities. Note that t in these cases corresponds to the $1 - \delta$ quantile for the target distribution distribution.

As we see in Table 1, both the optimal number of observation and the maximal simulation gain increases as $\delta \to 0$.

Table 1: The maximal simulation gain for different values of δ; the probability of exceeding the threshold t.

δ	n_{opt}	Γ
0.01	159	183
10^{-4}	15 936	18 377
10^{-7}	15 936 242	18 377 668

4 Simulation Study

4.1 Selection of Distributions for Simulation Study

We limit ourselves to continuous distributions in order to avoid challenges involving discrete distributions. In many cases, the normalised sample maxima of discrete distributions does not converge to any extreme value distribution (see [3, Theorem 3.1.3]). Although it may be feasible to apply importance sampling in discrete cases, most of the theory considering tail estimation is based on a continuous distribution assumption.

In the importance sampling setting, assuming continuity, we are not restricted in the choice of distributions. Therefore, our distribution selection revolves around the limitations present in extreme value theory, or more specifically, the limitations regarding the shape parameter ξ. As mentioned in section 2.4 and 2.5, the parameter ξ for each distribution has to be in the interval $]-0.5, 1[$ in order for our methods to be justified.

Since we want to examine possible differences for different tail behaviours we choose one distribution from each MDA. In the Gumbel case we know that $\xi = 0$ for all distributions. Therefore we can select any distribution in this MDA (see [3, Table 3.4.4] for examples). The most natural choice here is the standard normal distribution, denoted $\mathrm{N}(0, 1)$. In the Fréchet and Weibull cases, we need to be more careful.

We will now derive the shape parameter ξ for the remaining distributions and conclude that these two are suitable for the comparison. The theoretical basis for this section is taken from [3, Section 3.3].

First, let us introduce an essential concept. We say that a function f is regularly varying with index $-\eta$ for some $\eta \geq 0$ if the following holds:

$$\lim_{x\to\infty} \frac{f(xt)}{f(x)} = t^{-\eta}, \quad t > 0. \tag{14}$$

We denote this $f \in \mathcal{R}_{-\eta}$.

In the case of the Fréchet MDA the following statement holds

$$F \in \mathrm{MDA}(\Phi_\alpha) \iff \overline{F} \in \mathcal{R}_{-\alpha}. \tag{15}$$

Recall from (2) in section 2.1 that $\xi = \alpha^{-1} > 0$ in the Fréchet case. Hence we want to find a distribution such that $\overline{F} \in \mathcal{R}_{-\alpha}$ where $\alpha > 1$ since this gives us $\xi < 1$. Consider the Pareto distribution with location parameter a and shape parameter b. Denote the corresponding distribution function F. Using (14) and (15) we get that, for $x > a$

$$\overline{F}(x) = \left(\frac{a}{x}\right)^b \in \mathcal{R}_{-b} \Longrightarrow \xi = \frac{1}{b}.$$

Hence, if we choose $a = 1$ and $b = 2$ the Pareto distribution meets the requirements for the simulation study. We denote this distribution as Par(location= 1, shape= 2).

For the Weibull MDA the following holds:

$$F \in \text{MDA}(\Psi_\alpha) \Longleftrightarrow x_F < \infty,\ \overline{F}(x_F - x^{-1}) \in \mathcal{R}_{-\alpha}.$$

We know that the beta distribution has bounded support with $x_F = 1$. The density function of a beta distribution with parameters a and b is given by

$$f(x) = nx^{a-1}(1-x)^{b-1}, \quad 0 < x < 1, \quad a, b > 0$$

where K is a constant. We note that

$$\begin{aligned}
\lim_{x\to\infty} \frac{f(1-(xt)^{-1})}{f(1-x^{-1})} &= \lim_{x\to\infty} \frac{K(1-(xt)^{-1})^{a-1}(xt)^{1-b}}{K(1-x^{-1})^{a-1}x^{1-b}} \\
&= \lim_{x\to\infty} \frac{(1-(xt))^{a-1}t^{1-b}}{(1-x^{-1})^{a-1}} \\
&= t^{1-b},
\end{aligned}$$

which implies that $f(1-x^{-1}) \in \mathcal{R}_{1-b}$.

By Karamatas theorem we get that

$$\begin{aligned}
\overline{F}(x_F - x^{-1}) &= \int_{1-x^{-1}}^{1} f(y)\, dy \\
&\sim x^{-1} f(1-x^{-1}),
\end{aligned}$$

thus

$$\lim_{x\to\infty} \frac{\overline{F}(1-(xt)^{-1})}{\overline{F}(1-x^{-1})} = \lim_{x\to\infty} \frac{(xt)^{-1}f(1-(xt)^{-1})}{x^{-1}f(1-x^{-1})} = t^{-1}t^{1-b} = t^{-b}$$

which implies that $\overline{F}(1-x^{-1}) \in \mathcal{R}_{-b}$. In the Weibull case we know from (2) that $\xi = -\alpha^{-1}$. Hence, if we choose a beta distribution with parameters $a = 1$ and $b = 3$ we have that $\xi = -\frac{1}{3}$ which is suitable for the simulation study. We denote this distribution as Beta$(1, 3)$.

4.2 The Inverse Transformation Method

To simulate random variables we mainly use built in functions in R. There is one case though, where we have to perform the simulations ourselves. When we use importance sampling we have to simulate observations from our biasing density.

The easiest and most straightforward method for continuous distributions is the inverse transformation method, which makes use of Proposition 4.2.1.

Proposition 4.2.1
Let $U \sim U(0,1)$. For any continuous distribution function F it holds that, if we define

$$X = F^{-1}(U),$$

then X has distribution F. [10, p. 683-685]

Hence, to simulate from the biasing density we can simply compute the inverse of F_*. In our case F_* is the upper order distribution function corresponding to each target distribution. By standard calculations we obtain the result

$$X_* = F^{-1}(U^{\frac{1}{n}}), \tag{16}$$

where n denotes the sample size and X_* denotes a random variable with the upper order biasing density. Note that F in (16) refers to the inverse of the target distribution function.

4.3 Selection of Sample Sizes and Tail Probabilities

In order to compare the GPD method and importance sampling, it is important to choose the sample sizes and tail probabilities δ with caution.

As a start, we aim to select a small δ, such that the Monte Carlo method is unlikely to provide reliable results. Then we want to choose various sample sizes such that improvement can be observed for all methods. We also wish to include at least one very small sample size value.

After some experimenting we decide to go with $\delta = 10^{-4}$ as a start. We conclude that 10, 100, 1000, and 100 000 are suitable choices to observe differences. However, we also want to include n_{opt}, which in this case is $n = 15\ 936$ (see Table 1).

To further investigate the extreme right tail cases, we also include $\delta = 10^{-7}$. To compare the two tail probabilities, we maintain the same sample sizes as for $\delta = 10^{-4}$. We also refrain from using the specific n_{opt} value for $\delta = 10^{-7}$ since this value is very large, which could lead to computational challenges. With the current sample sizes, we will still be able to observe improvement in importance sampling compared to Monte Carlo.

4.4 Plots for Threshold Selection

In the GPD approach, threshold selection u is crucial for the reliability of the estimate $\hat{\xi}$ (we will not address β since this is only a scale parameter). As discussed in section 2.4, to find a suitable threshold we visually examine two different plots for each distribution and sample size. The first plot shows the empirical mean excess function and the second one shows $\hat{\xi}$, both across various threshold values. Showing all of these plots would be excessive, hence we will illustrate the methods with a few examples from each distribution. The remaining plots will be attached to the appendix.

What we are looking for in the empirical mean excess plot is a threshold high enough, so that the points look approximately linear to the right of the threshold. As we shall see this is not a trivial task. The plots are often hard to asses and the there are multiple possible choices. Hence, to facilitate the threshold selection we will use the second plot as well.

What we are looking for in the $\hat{\xi}$ plot is a bias-variance trade-off. The estimate is asymptotically unbiased for $u \to x_F$, but the larger the threshold, the fewer observations to estimate $\widehat{\overline{F}}(u)$ and the GPD parameters from. Hence, we will try to find a threshold as high as possible without the $\hat{\xi}$ plot showing too much instability.

For the $\hat{\xi}$ plot we also construct confidence intervals to facilitate the assessment. To do this we use the fact that the MLE of ξ is asymptotically normal [1]. The following method on how to construct confidence intervals is taken from [6, p. 97-101].

Assuming the MLE $\hat{\xi}$ is approximately normal we can construct a 95 % Wald confidence interval in the following way

$$I = \left(\hat{\xi} \pm z_{0.975}\mathrm{se}(\hat{\xi})\right)$$

where se denotes the standard error of $\hat{\xi}$, whereas $z_{0.975}$ is the 0.975-quantile of a standard normal distribution.

It shall be mentioned that the approximate normality probably does not hold for our smallest sample size $n = 10$. Despite this fact, the confidence intervals are included in the plots for technical convenience. However, the reader should interpret these intervals cautiously and not place too much emphasis on them.

We start by examining the beta distribution which is in the Weibull MDA (Ψ), in which the shape parameter $\xi < 0$. In Figure 5 we can see a few illustrating examples of what one can encounter in this case since $\hat{\xi}$ often becomes negative aswell. If $\hat{\xi} < 0$ we have the restriction $0 < x - u < \frac{-\hat{\beta}}{\hat{\xi}}$, which limits the options for u. Since $\hat{\xi}$, $\hat{\beta}$ and u are connected it is hard to

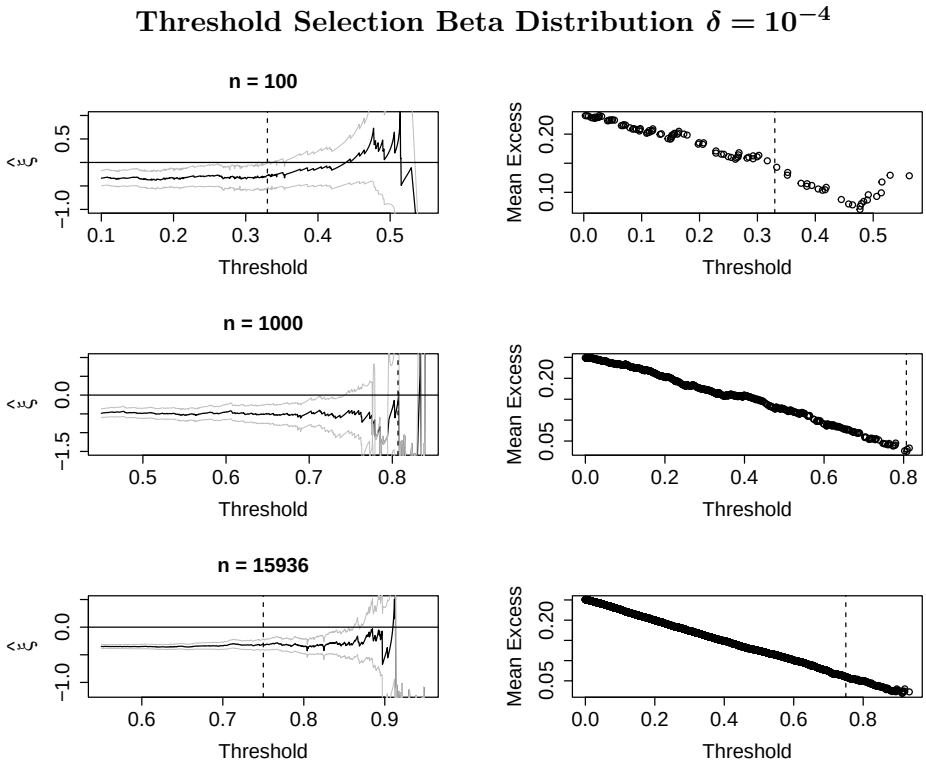

Figure 5: Threshold selection for Beta$(1, 3)$ and $\delta = 10^{-4}$. The chosen thresholds u, indicated by the dotted lines, are: 0.33, 0.8068136 and 0.75 aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

find a simple rule to get this condition met. In general, though, if it is not met, one should try to find a larger threshold.

For $n = 1000$ we can see one of these cases illustrated. In this case we have to choose a much higher u than we would have preferred, from visually examining the plots. Thus we have to choose a threshold resulting in a seemingly high variance for $\hat{\xi}$. Note that this restriction, in general, also requires a larger u for $x \to x_F$. Consequently we had to choose partly different thresholds for $\delta = 10^{-7}$ (see appendix).

For $n = 100$ we can see that the the empirical mean excess function seems to turn upwards for $u \approx 4.8$. According to Embrechts [3, p. 320] this could be a sign of clustering in the data. Hence, one should not place to much importance in this phenomenon as it likely does not accurately represent the overall shape of the theoretical mean excess function. For $n = 15936$

we can see an example where the threshold choice seems relatively clear and straightforward.

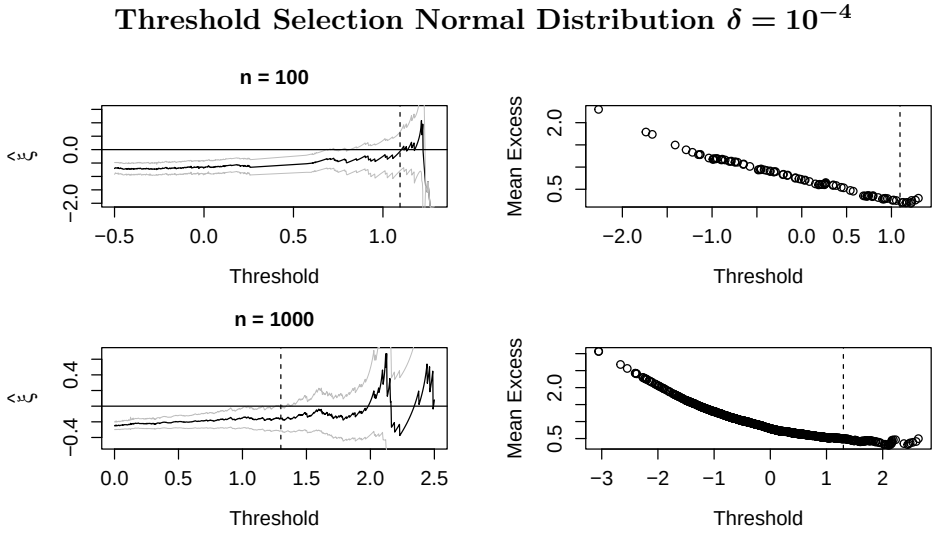

Figure 6: Threshold selection for $\mathrm{N}(0,1)$ and $\delta = 10^{-4}$. The chosen thresholds u, indicated by the dotted lines, are: 1.095477 and 1.3 aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

In the normal distribution case, which is in the Gumbel MDA (Λ), the shape parameter $\xi = 0$. But since we are using $\hat{\xi}$ which in almost all cases are non-zero we can also face the same challenges as if $\hat{\xi} < 0$.

In Figure 6 we can see this illustrated for $n = 100$, where a more natural choice of threshold would be slightly lower, by visually examining the plots. Note that, in this case we also needed to adjust some of the thresholds for $\delta = 10^{-7}$ (see appendix).

For $n = 1000$ we see an example where the choice is more intuitive. The empirical mean excess function seems approximately linear and the variance of $\hat{\xi}$ does not seem to be to high.

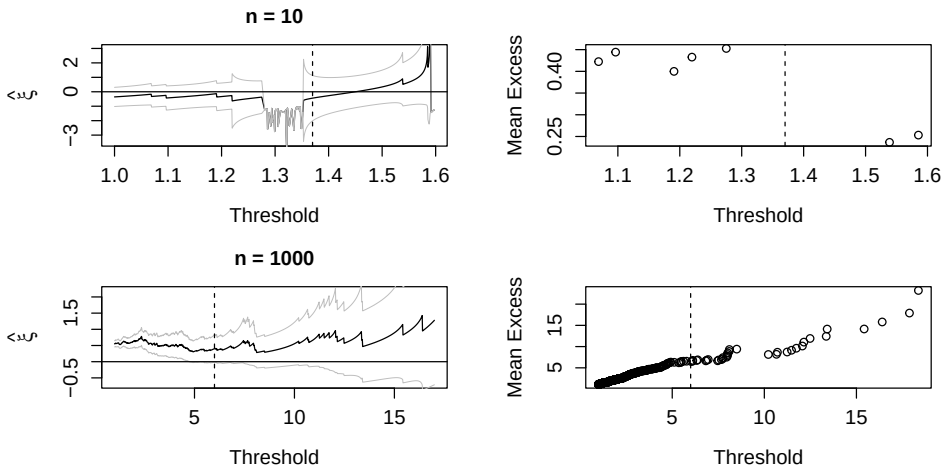

Figure 7: Threshold selection for Par(location $= 1$, shape $= 2$). The chosen thresholds u, indicated by the dotted lines, consistent across both values of δ, are: 1.37 and 6 aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

In the Pareto distribution case, which is in the Fréchet MDA (Φ), the shape parameter $\xi > 0$. Consequently the estimate $\hat{\xi}$ will typically also be positive. This is very helpful when selecting thresholds since our only restriction is that $x - u > 0$. Consequently, we will have more possible options choosing u.

In Figure 7 we can see this illustrated for $n = 1000$ where the choice of u seems quite intuitive. The empirical mean excess function seems approximately linear and we still have a reasonable amount of data to estimate $\widehat{\overline{F}}(u)$ and the GPD parameters. For $n = 10$ though, we are facing the recurring problem of selecting a threshold when dealing with very few observations. In this case one have to review all possible threshold options, which are very limited, and try to find one that seems reasonable.

5 Result

In Table 2 and 3 the values of the tail estimates are presented. To make the results more intuitive, we also present the results in Figure 8 and 9. Note the scaled y-axis which leads to small deviations being perceived as larger and conversely for large deviations. Consequently, it is important for the reader to be aware of this adjustment.

Table 2: Tail estimates for $\delta = 10^{-4}$

n	Beta			Normal			Pareto			δ
	MC	IS	GPD	MC	IS	GPD	MC	IS	GPD	
10	0	0	$6.4 \cdot 10^{-3}$	0	0	$2.3 \cdot 10^{-3}$	0	0	$4.8 \cdot 10^{-4}$	10^{-4}
100	0	$1.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-7}$	0	$1.0 \cdot 10^{-4}$	$1.7 \cdot 10^{-11}$	0	0	$7.3 \cdot 10^{-6}$	10^{-4}
1000	0	$1.1 \cdot 10^{-4}$	$9.6 \cdot 10^{-9}$	0	$1.0 \cdot 10^{-4}$	$8.4 \cdot 10^{-5}$	0	$1.1 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	10^{-4}
15936	$1.3 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$7.3 \cdot 10^{-5}$	$1.3 \cdot 10^{-4}$	$9.9 \cdot 10^{-5}$	$1.2 \cdot 10^{-4}$	$6.3 \cdot 10^{-5}$	$9.9 \cdot 10^{-5}$	$8.6 \cdot 10^{-5}$	10^{-4}
100000	$1.2 \cdot 10^{-4}$	$9.8 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$1.1 \cdot 10^{-4}$	$1.3 \cdot 10^{-4}$	$9.2 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	10^{-4}

In Table 2 we observe that Monte Carlo returns zero for small n values. Importance sampling produce accurate results for most n values. Note that for $n = 100\ 000$, in both the beta and Pareto case, importance sampling seems to lower its accuracy slightly. In Figure 8 we notice that the GPD method deviates a lot from the true value $\delta = 10^{-4}$, especially for small sample sizes in the beta and normal cases.

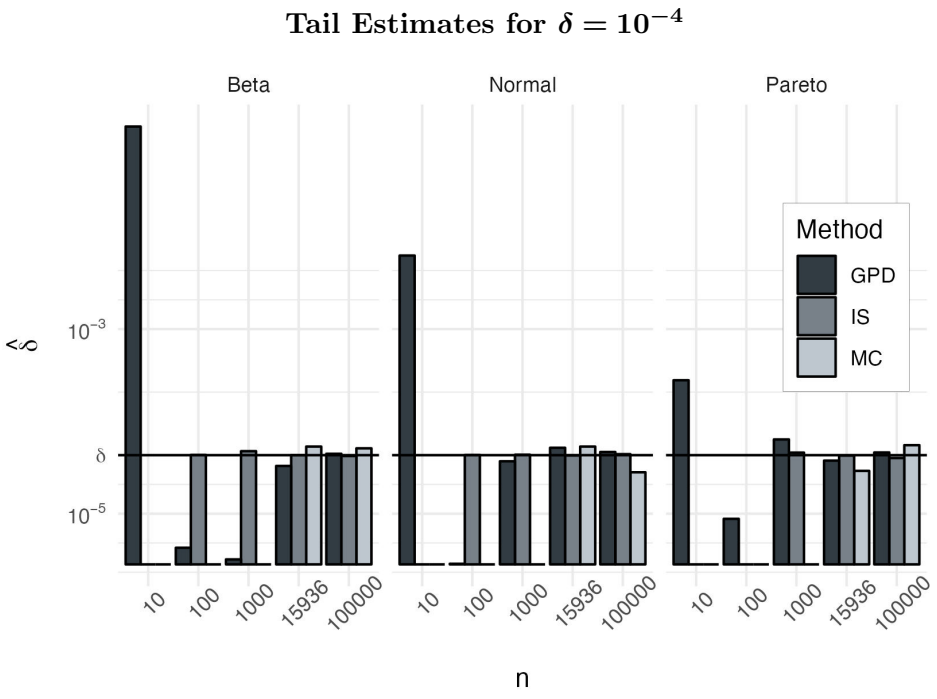

Figure 8: Tail estimates across the three methods compared to the true tail probability $\delta = 10^{-4}$. Please note that the y-axis is scaled using a square root transformation to improve readability.

Table 3: Tail estimates for $\delta = 10^{-7}$

n	Beta			Normal			Pareto			δ
	MC	IS	GPD	MC	IS	GPD	MC	IS	GPD	
10	0	0	$5.2 \cdot 10^{-3}$	0	0	$5.9 \cdot 10^{-4}$	0	0	$2.5 \cdot 10^{-6}$	10^{-7}
100	0	0	$7.4 \cdot 10^{-5}$	0	0	$8.9 \cdot 10^{-5}$	0	0	$7.9 \cdot 10^{-11}$	10^{-7}
1000	0	$1.0 \cdot 10^{-6}$	$2.2 \cdot 10^{-4}$	0	0	$2.9 \cdot 10^{-8}$	0	0	$7.7 \cdot 10^{-7}$	10^{-7}
15936	0	$6.7 \cdot 10^{-8}$	$8.0 \cdot 10^{-8}$	0	$1.2 \cdot 10^{-7}$	$1.5 \cdot 10^{-9}$	0	$1.1 \cdot 10^{-7}$	$4.9 \cdot 10^{-8}$	10^{-7}
100000	0	$1.0 \cdot 10^{-7}$	$2.2 \cdot 10^{-7}$	0	$1.0 \cdot 10^{-7}$	$3.7 \cdot 10^{-11}$	0	$1.0 \cdot 10^{-7}$	$5.7 \cdot 10^{-8}$	10^{-7}

In Table 3 we observe the changes for each method given an even smaller value of δ. Monte Carlo returns zero for all n values. Importance sampling also shows similar tendencies. In Figure 9 we notice that the GPD method returns high errors, especially for small sample sizes in the beta and normal cases. Note though, that the estimates are non-zero and in the Pareto case it does not deviate as much from $\delta = 10^{-7}$.

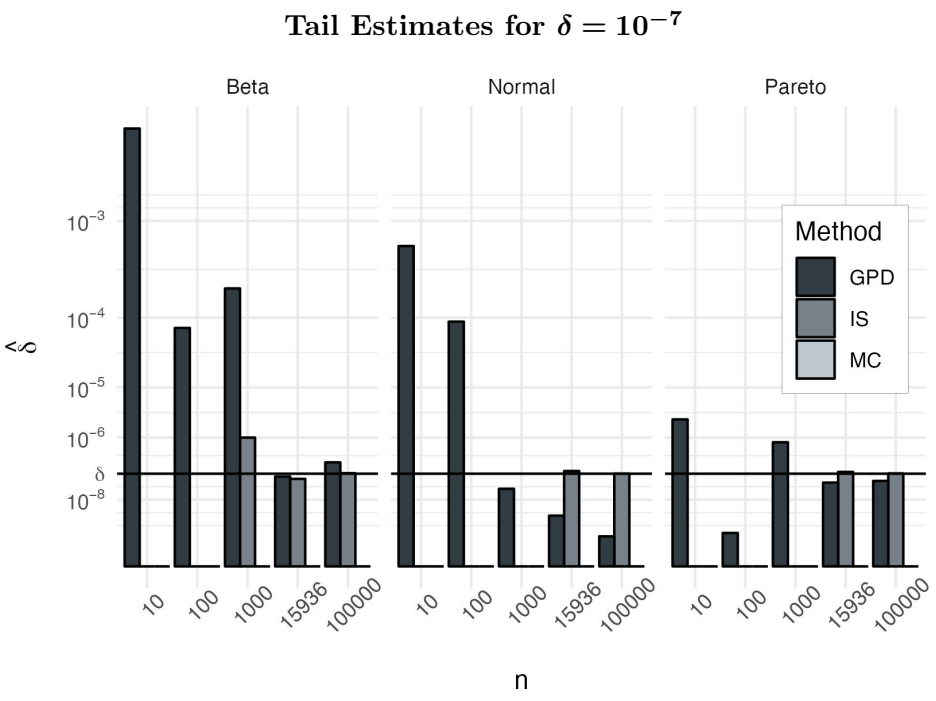

Figure 9: Tail estimates across the three methods compared to the true tail probability $\delta = 10^{-7}$. Please note that the y-axis is scaled using a square root transformation to improve readability.

6 Discussion

The simulation study shows that importance sampling with upper order biasing is a very convenient method that demonstrates reliable results since the estimator is unbiased. As expected it clearly requires fewer observations compared to Monte Carlo given the same accuracy. A problem though, is that for small sample sizes, both importance sampling and Monte Carlo returns zero. This is a consequence of having no observations in the area of interest. We also note that, in some cases, importance sampling seems to slightly lower its performance for very large sample sizes (see Table 2 in the Pareto and beta case). This could be due to numerical problems, since the biasing density involves division by a probability to the power of many thousands. However, note that the upper order biasing method is expected to result in equal performance compared to Monte Carlo as n becomes very large. As a consequence, this might not be a huge problem since Monte Carlo improves as n increases.

The GPD method manages to perform better than Monte Carlo in many cases and sometimes even importance sampling. However, the GPD method also seems to have some problems, especially concerning the variety of performance. The deviation from the true value δ is very high for small sample sizes, especially in the Weibull (beta distribution) and Gumbel (normal distribution) cases. This could be a consequence of the challenges posed in the threshold selection (see Section 4.4). In the Pareto case though, it actually seems to perform quite well, even compared to importance sampling. This becomes even more clear when lowering δ to 10^{-7} for small sample sizes. Here the GPD method really shows its strength.

Extreme value theory was developed to solve real world problems, were one does not know the underlying distribution. If the sample size is small and we consider some extreme right tail, it would be hard to achieve any information from a method that requires observations in the actual tail. With this in mind, despite large errors in some cases, the GPD method has an advantage since we do not have to rely on data in the tail region of interest. This way extreme value theory gives us the possibility to estimate very rare events of great importance, based on very limited data.

It is valuable that the GPD method can return non-zero estimates. The alternative would might be to guess or assume that the probability is zero, which we often know in advance, is not the case. However, there are of course downsides, and it only works under the assumption of independent and identical distributed random variables (which in many important applications do not apply).

A disadvantage of the GPD method used in this study is that we have to manually examine all the plots. This is time consuming and also leaves a

lot to decide for the person analysing the plots. This can also affect the reliability of the results since the true ξ values were known beforehand, leading to potential bias. There might be automated procedures though, which would be interesting to examine as well.

7 Conclusion

Despite the different prior knowledge required for the GPD method and importance sampling we are able to draw some important conclusions from this simulation study.

In the importance sampling case, upper order biasing is a very convenient and effective method. It is easy to use and the performance does not depend on the target distribution. It performs very well compared to Monte Carlo. Both methods return zero estimates though, if the sample size is too small. There could also potentially be numerical challenges for the upper order biasing method. However, this needs further investigation in order to draw any definite conclusions.

The performance of the GPD method varies widely and it shows high errors in many cases. We observe some differences between the MDAs by studying the different distributions. The challenges arising from $\xi \leq 0$ becomes clear when selecting thresholds. Since $\hat{\xi}$ is negative in most of these cases, the options for threshold selection is clearly limited, both in the Weibull (beta distribution) and the Gumbel (normal distribution) domains of attraction. This can potentially be an explanation to why the GPD method in these cases shows large errors. In the in Fréchet (Pareto distribution) case this method seems to perform quite well, even for very small sample sizes.

Another aspect that should be mentioned is that the GPD method almost never returns zero estimates. As we argue in the discussion (Section 6), this is an advantage considering real world problems where we know that the probability is not zero.

Since the GPD method we use in this thesis involves manually examination of plots, the results might be slightly biased due to prior knowledge about the shape parameter ξ.

References

[1] Z. Bermudez and S. Kotz, *Parameter estimation of the generalized pareto distribution—part i*, Journal of Statistical Planning and Inference, 140 (2010), pp. 1353–1373.

[2] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer, 2001.

[3] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events: for Insurance and Finance*, Springer, 1997.

[4] A. Ferreira and L. De Haan, *On the block maxima method in extreme value theory: Pwm estimators*, The Annals of statistics, (2015), pp. 276–298.

[5] A. Gut, *An Intermediate Course in Probability*, Springer, 2nd ed., 2009.

[6] L. Held and D. Sabanés Bové, *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*, Springer, 2 ed., 2020.

[7] M. R. Ledbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and series*, Springer, 1983.

[8] A. J. McNeil, R. F., and P. Embrechts, *Quantitative Risk Management*, Princeton University Press, 2005.

[9] S. I. Resnick, *Extreme Values, Regular Variation, and Point Processes*, Springer, 2008.

[10] S. M. Ross, *Introduction to Probability Models*, Elsevier, 12th ed., 2019.

[11] R. Srinivasan, *Importance sampling: Applications in communications and detection*, Springer, 2002.

Appendix

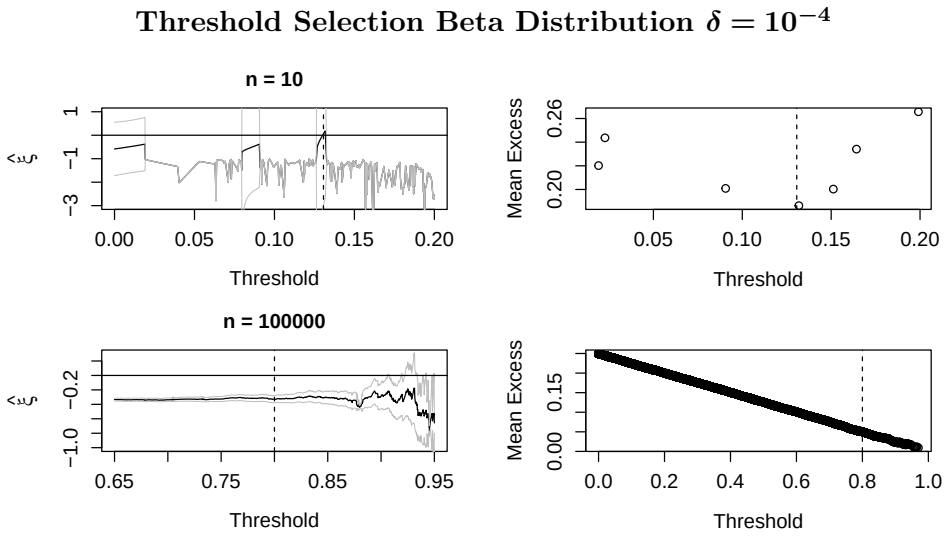

Figure 10: Threshold selection for Beta$(1, 3)$ and $\delta = 10^{-4}$. The chosen thresholds u, indicated by the dotted lines, are: 0.130653266 and 0.8, aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

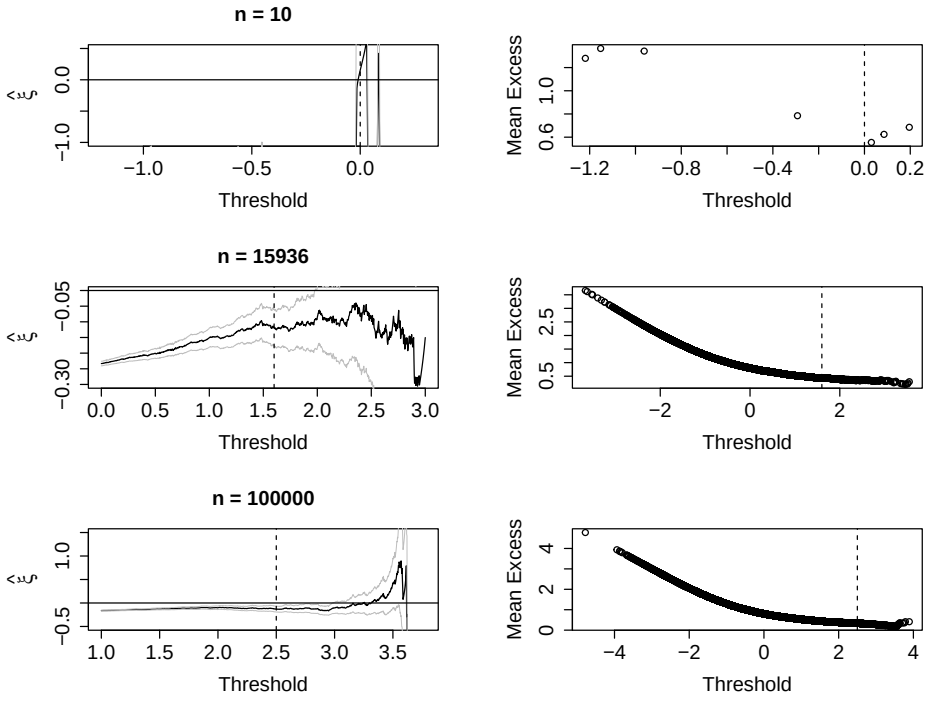

Figure 11: Threshold selection for Beta$(1, 3)$ and $\delta = 10^{-4}$. The chosen thresholds u, indicated by the dotted lines, are: 0, 1.6 and 2.5, aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

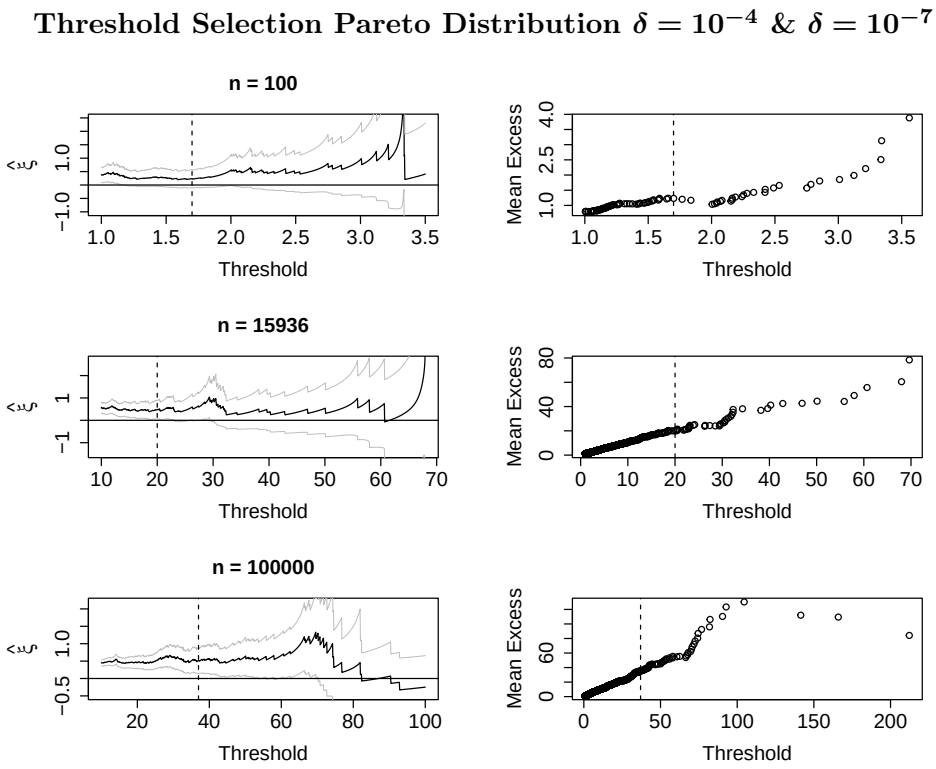

Figure 12: Threshold selection for Par(location $= 1$, shape $= 3$). The chosen thresholds u, indicated by the dotted lines, consistent across both δ are: 1.7, 20 and 37, aligned with increasing sample sizes n. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

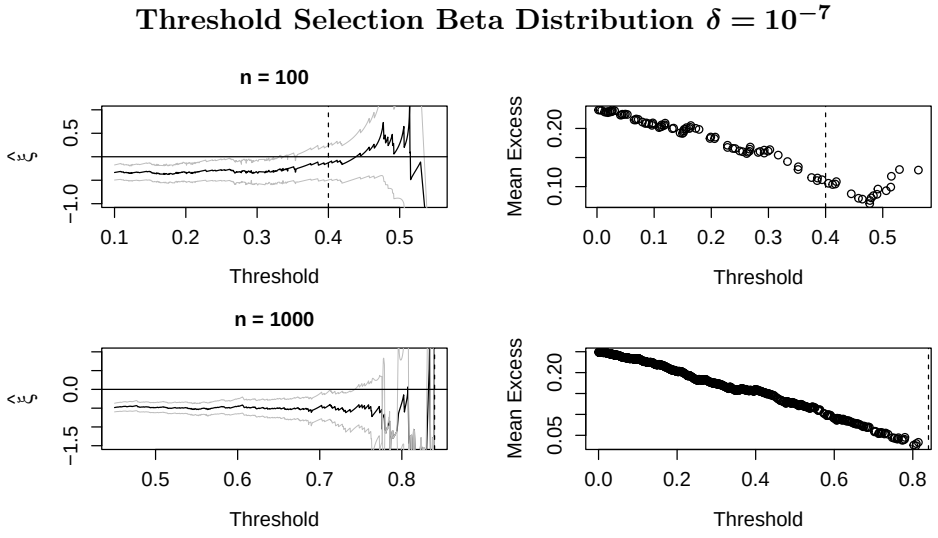

Figure 13: Threshold selection for Beta$(1, 3)$ and $\delta = 10^{-7}$. The chosen thresholds u, indicated by the dotted lines, are: 0.4 and 0.83988, aligned with increasing sample sizes n. Note that the selected thresholds for the remaining values of n is the same as for $\delta = 10^{-4}$. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.

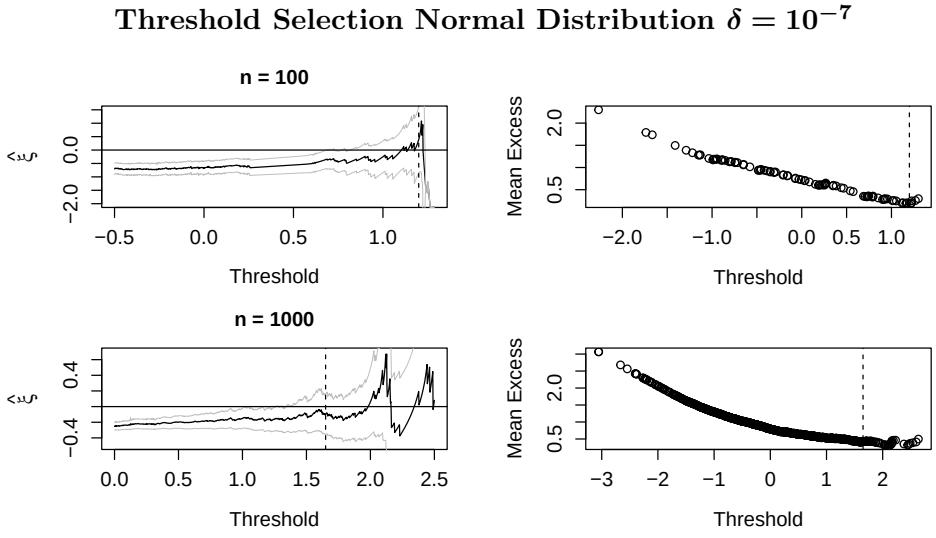

Figure 14: Threshold selection for $\mathrm{N}(0, 1)$ and $\delta = 10^{-7}$. The chosen thresholds u, indicated by the dotted lines, are: 1.2 and 1.65, aligned with increasing sample sizes n. Note that the selected thresholds for the remaining values of n is the same as for $\delta = 10^{-4}$. The plots to the left illustrate the shape parameter estimate $\hat{\xi}$ across various threshold values u. The plots to the right show the empirical mean excess function for different threshold values.