Univariate Extreme Value Modelling

Zhang Ruiyang

Contents

Preface

This notes aims to survey various extreme value models for a sequence of univariate and independent and identically distributed (i.i.d.) random variables. Statistical inference mostly concern with the main part (within 2 s.d., say) of distributions, and extreme events are often poorly represented by those models. However, under some contexts, it is the extreme event and not the non-extreme event that we are interested in. For example, in hydrology, we are interested in the extreme water levels which will case flooding. Therefore, new approaches and methods are needed for these types of statistical analyses.

We will start with an introduction of extreme value theory, and go over some of the basics of the theory. The extreme value modelling, at least in the univariate i.i.d. case, has three main groups of models, each uses different proportions of the data. In Chapter 2, we will go over the first group of models, using only block-wise data. In Chapter 3, we will discuss methods that use all data beyond a threshold. In Chapter 4, we will explain the mixture models, which use all the data. Two simulation studies will be carried out and analysed in Chapter 5.

This notes is prepared as a summary for the work I have done while doing the 6-week STOR-i internship at Lancaster University during the summer of 2022. The internship is funded by STOR-i and supervised by Conor Murphy and Lídia André.

Lancaster, UK
August 2022

Chapter 1

Introduction

Fort Collins is a city in northern Colorado, USA. It does not rain a lot - it gets 16 inches of rain on average per year, while the US average is 38 inches. However, during 27 - 28 July, 1997, Fort Collins was hit by a slow moving storm and it dumped 14.5 inches of rain in 31 hours, creating flash flooding and destroyed many parts of the city. This catastrophe caused 5 deaths, dozens injuries, and over $200 million in property damage.

For Fort Collins, such heavy precipitation is certainly rare and extreme. In this type of context, we would like to have a better understanding of those rare events and find a more suitable model for those in order to better prepare for future catastrophes. From a statistical perspective, the theory that deals with modelling of this kind is the extreme value theory, and we will be exploring some of it in this notes.

The following figure is a plot of daily maximum precipitation data at Fort Collins between 1900 and 1999.

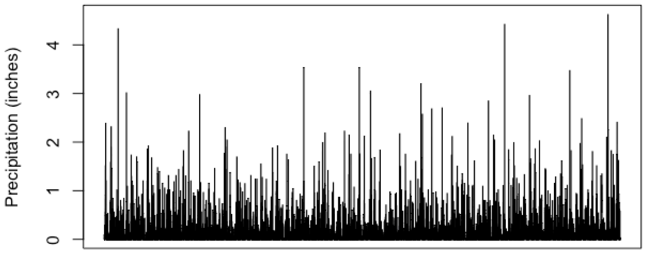


Figure 1.1: Fort Collins Daily Maximum Precipitation Data, 1900-1999

Now, we are interested in modelling the extreme events. In this case, the larger the value, the more extreme the event. One of the most direct thing we can do is to highlight the maximum precipitation for each year and model using those. In the following figure, we have included the annual maximum precipitation and coloured them in red.

This approach, by dividing the data into blocks with fixed size and picking the maximum from them, is known as the block-wise method, which will be discussed in detail in Chapter 2. The block maximums can be modelled using a generalised extreme value distribution, and we can do

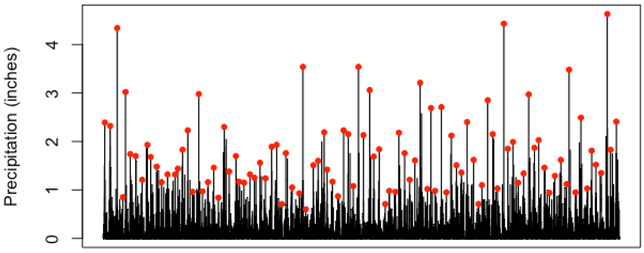


Figure 1.2: Fort Collins Daily Maximum Precipitation Data with Annual Maximum, 1900-1999

inference based on that - for example, to compute the 99% and 99.9% quantiles. One issue with this approach is the selection of block size. In this example, by taking the annual maximum, we use a block size of 365 (or 366 for a leap year), and this is rather intuitive. However, there are cases where this choice is not as intuitive, which will impose issues on the modelling. Another issue is that this approach is rather wasteful, as we have neglected the majority of data.

A different approach, building on the previous, is to select all data with values greater than a threshold. In the following figure, we have included horizontal threshold at 0.8, coloured in red.

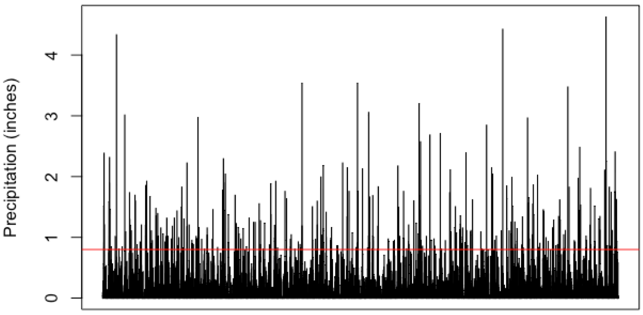


Figure 1.3: Fort Collins Daily Maximum Precipitation Data with Threshold, 1900-1999

This approach of selecting a threshold and picking all the 'peaks' over the threshold is known as the Peaks-Over-Thresholds method, and we will discuss it in detail in Chapter 3. The values above a well-selected threshold can be modelled by a generalised Pareto distribution, and inferences could be done based on that. However, this relies on a good threshold, and the quality of the modelling depends heavily on the quality of threshold, which is hard to select. Some methods of threshold selection are going to be mentioned in Chapter 3 as well.

Another approach one could take is to model with all the data using a mixture model. This approach will be described in Chapter 4. Most of the mixture models follow the same idea: the non-extreme events follow one distribution while the extreme events follow another distribution (usually generalised Pareto distribution), and by finding a good way to mix the two distributions will offer us a good model for the data. This approach is especially helpful when we are interested in doing inferences for both the extreme and the non-extreme events. As we are involved with more data and more complicated models, one natural issue is the increase in computational cost. Another point of concern is that if our only goal is to model the extreme, it might not be too useful to include the non-extreme data. The inclusion of the non-extreme distribution may not influence the extreme modelling too much, making efforts to implement a mixture model is relatively futile one.

The last chapter of this notes will include two simulation studies to investigate some of the methods mentioned in earlier chapters. Most of the content in this notes is based on Coles (2001) [3], and we will survey some of the more recent methods and citations will be included when necessary.

Chapter 2

Block-wise Models

In this chapter, we will be looking at the Extreme Value Theorem, which provides the theoretical foundation for using the Generalised Extreme Value distribution to model the block maximums. After that, we will discuss some of the inferences we can do with the Generalised Extreme Value distribution.

2.1 Extreme Value Theorem

Consider a sequence of univariate random variables $X_1, X_2, \cdots$ that are independently and identically distributed (i.i.d.), with a common CDF F. We can think of this sequence as a stream of independent samples from a particular distribution. One thing we can consider is how the average behaves as we get more and more samples, i.e. $\bar{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$. From central limit theorem, we know that $\sqrt{n}(\bar{X_n} - \mu) \to N(0, \sigma^2)$ in distribution, where μ and σ are the mean and standard deviation of the distribution for X_i. This result is key to the further development of Probability and Statistics, which is why it is called the 'central' limit theorem.

Another thing that was considered for the same stream of samples is the maximum of them, i.e. $M_n := \max(X_1, \cdots X_n)$. What distribution might this value follow asymptotically? This line of study begins the investigation of extreme value theory, but at the initial stage it was merely studied for mathematical curiosity.

Let us first play around with these concepts for a bit. For $M_n = z$, we need to have $X_1, \cdots, X_n \leq z$, which gives us the following equation

$$\mathbb{P}(M_n \leq z) = \mathbb{P}(X_1 \leq z, X_2 \leq z, \cdots, X_n \leq z) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq z) = F^n(z)$$

where final two equalities are due to i.i.d. of X_is.

From this equality, if we can obtain the CDF of the samples, we can obtain the distribution for M_n. However, in a lot of real-life applications, the CDF of samples are not obtainable - we simply do not know. We could use the empirical CDF of the samples instead, but the uncertainty would be huge if we start to compute F^n for some large n, which makes this approach not very fruitful.

Another consequence of the above equality is that, as $n \to \infty$, if z^+ is the upper end point of F, i.e. $z^+ := \inf_z F(z) = 1$, for any $z \leq z^+$ we would have $F^n(z) \to 0$, and $F^n(z^+) = 1$. The distribution of M_n would then degenerate to a point mass on z^+, which is not desirable. To remedy this issue, we find two sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ and study

$$M_n^* = \frac{M_n - b_n}{a_n}$$

instead. For appropriate choices of $\{a_n > 0\}$ and $\{b_n\}$, the degeneracy issue is avoided. The choice of these constants is a delicate matter, but we will assume they have already been found and carry on with the discussion of asymptotic properties of M_n^*. We will only show some examples of $\{a_n > 0\}$ and $\{b_n\}$ for some given F later. Interested readers could refer to Kotz and Nadarajah (2000) [9] for more details on this topic.

Before moving on to finally describing the result, one final remark is that we are only considering the maximum and not the minimum since these two are interchangeable. To illustrate this point, we have the following equality

$$\min(X_1, \cdots, X_n) = \max(-X_1, \cdots, -X_n).$$

Now, the theorem about the maximum of a stream of i.i.d. samples is known as the **Extreme Value Theorem**, sometimes it is also called the Fisher–Tippett–Gnedenko theorem. Here is its statement.

Theorem 1. (Extreme Value Theorem) If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] \to G(z)$$

as $n \to \infty$, where G is a non-degenerate (i.e. not point mass) distribution function, then G has the form

$$G_\xi(x) = \exp\Big(- (1 + \xi x)^{-1/\xi}\Big), \qquad 1 + \gamma x > 0$$

or just

$$G_0(x) = \exp(-\exp(-x)).$$

The proof of this theorem is omitted. It could be found at many places, e.g. Haan and Ferreira (2006) [7].

The function G (both G_ξ and G_0) is known as the **generalised extreme value distribution** (GEV). The value of ξ determines the decay of the tail of the distribution, and it is one of the most essential parameters that we would like to study for GEV. We call ξ the **extreme-value index** (EVI) of the distribution, and we will mention briefly how its value affects the distribution as well as ways to estimate it. This will happen in Section 2.3.

This chapter aims to discuss block-wise methods of extreme value modelling. To illustrate how the GEV mentioned just now is related to block-wise maximum, we can simply consider $X_1, \cdots, X_n$ as n data in one time block, and $M_n := \max(X_1, \cdots, X_n)$ as the block maximum. Depending on the data, the division could be by month or year, or some other suitably sized interval. We have already established that given suitable rescaling of M_n, the rescaled random variable M_n^* follows GEV, which allows us to make further inferences about extreme values of the original data.

For Theorem 1, we have assumed that for a large enough n, we can use GEV to model the distribution of maxima of long sequences. The normalising constants mentioned in the theorem assumption seem to be hard to find, but it could actually be resolved rather easily in practice. Assuming we have

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \to G(z)$$

as $n \to \infty$, we have, for large enough n,

$$\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] \approx G(z).$$

Equivalently, we have

$$\begin{aligned} \mathbb{P}\left[M_n \leq z\right] &\approx G[(z - b_n)/a_n] \\ &= G^*(z) \end{aligned}$$

where G^* is another GEV. So, it is irrelevant in practice that the parameters of G and G^* are different, since parameters of both need to estimated anyway.

This motivates the following approach of modelling extremes of i.i.d. data sequence. We will block the data into sequences of length n where n is large, and this generates a series of block maxima $M_{n,1}, M_{n,2}, \cdots, M_{n,m}$ which the GEV could be fitted. With this, estimates of extreme quantiles of the annual maximum distribution can then be obtained by the following equations

$$z_p = \begin{cases} b - \frac{a}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] & \text{for } \xi \neq 0 \\ b - a\log\{-\log(1-p)\} & \text{for } \xi = 0 \end{cases}$$

where $G(z_p) = 1 - p$. We normally call z_p as the **return level** associated with the **return period** $1/p$, since we expect the level z_p to be exceeded on an average of $1/p$ years. If we choose to block the data annually, z_p is exceeded by the annual maximum in any particular year with probability p.

2.2 Generalised Extreme Value Distribution

Generalised extreme value distribution unites three families of extreme value distributions - Gumbel, Fréchet, and Weibull. Which of the three families the GEV represents depends on the value of EVI ξ. When $\xi = 0$, we get Type I distribution, the Gumbel distribution. When $\xi > 0$, we get Type II distribution, the Fréchet distribution. When $\xi < 0$, we get Type III distribution, the Weibull distribution.

The Gumbel distribution has CDF

$$F(x) = \exp\{-\exp[-(x-b)/a]\}$$

where b is its location parameter and a is its scale parameter.

The Fréchet distribution has CDF

$$F(x) = \begin{cases} 0 & z \leq b \\ \exp\{-(\frac{z-b}{a})^{-\alpha}\} & z > b \end{cases}$$

where b is its location parameter, a is its scale parameter, and α is the shape parameter.

The Weibull distribution has CDF

$$F(x) = \begin{cases} \exp\{-[-(\frac{z-b}{a})^{\alpha}]\} & z < b \\ 1 & z \geq b \end{cases}$$

where b is its location parameter, a is its scale parameter, and α is the shape parameter.

In the following figure, we have plotted the pdf of these three distributions for some arbitrary parameters. We can see that Fréchet distribution has a lower end point at b, while the Weibull distribution has a upper end point at b. The Gumbel distribution has no end points.

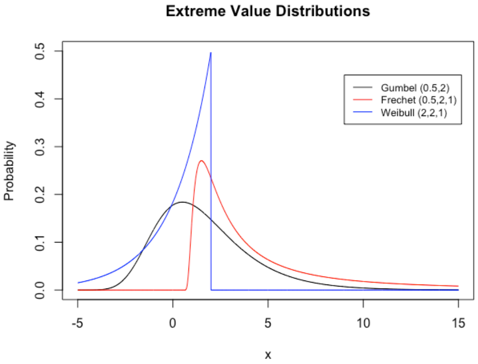


Figure 2.1: Gumbel, Fréchet, and Weibull Distributions

At this point, let us emphasise once again the effect of EVI ξ on the distribution. For $\xi < 0$, the distribution has an upper end point. Examples of such distributions include the uniform distribution and the Beta distribution. For $\xi = 0$, the distribution has a light tail and the tail decays exponentially. Such distributions include the normal distribution and the exponential distribution. For $\xi > 0$, the distribution has a heavy tail, and its tail decays polynomially like in the case of Pareto and Student t. In real-life applications, say environmental data and seismological data, we will almost always have data with positive EVI - which is why many methodologies in latter chapters assume $\xi > 0$.

Most of the times in science, a simpler and more general form is preferred. It is rather troublesome to have three separate distributions, especially when they serve similar modelling goals. This is why we would prefer the **generalised** extreme value distribution rather than these three extreme value distributions.

The generalised extreme value (GEV) distribution has CDF

$$\exp\Big\{-\Big[1+\xi\Big(\frac{z-b}{a}\Big)\Big]^{-1/\xi}\Big\}$$

on $\{z : 1+\xi(z-b)/a > 0\}$, where ξ is the extreme-value index (EVI), b is the location parameter, a is the scale parameter. When $\xi = 0$, the CDF becomes

$$\exp\Big\{-\exp\Big(\frac{z-b}{a}\Big)\Big\}.$$

The EVI also determines the tail behaviour of the GEV. For EVI $\xi < 0$, the distribution has an upper end point. For EVI $\xi = 0$, the tail is exponentially decaying. For EVI $\xi > 0$, the tail is polynomially decaying.

2.3 GEV Examples

In this section, we will study a few examples of how we can get M_n, the maximum of i.i.d. samples $X_1, X_2, \cdots, X_n$ following CDF F, to follow an GEV, by using well-selected constant sequences. The examples mentioned in this section is identical to the examples used in Coles (2001) [3].

2.3.1 Example 1

If $X_1, X_2, \cdots$ is a sequence of independent standard exponential Exp(1) variables with $F(x) = 1 - e^{-x}$ for $x > 0$. We let $a_n = 1$ and $b_n = \log n$, and we get

$$\begin{aligned}\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] &= F^n(z + \log n)\\ &= \Big[1 - \exp[-(z + \log n)]\Big]^n\\ &= \Big[1 - \exp(-z)/n)\Big]^n\\ &\to \exp(-e^{-z})\end{aligned}$$

as $n \to \infty$ for each fixed $z \in \mathbb{R}$. This means, the M_n in this case follows the Gumbel distribution with $b = 0$ and $a = 1$. The EVI ξ is 0.

2.3.2 Example 2

If $X_1, X_2, \cdots$ is a sequence of independent standard Fréchet variables with $F(x) = \exp(-1/x)$ for $x > 0$. We let $a_n = n$ and $b_n = 0$, and we get

$$\begin{aligned}\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] &= F^n(nz)\\ &= \Big(\exp[-1/(nz)]\Big)^n\\ &= \exp[-1/z]\end{aligned}$$

as $n \to \infty$ for each fixed $z > 0$. This means, the M_n in this case follows the Fréchet distribution with $b = 0$, $a = 1$, and $\alpha = 1$. The EVI ξ is 1.

2.3.3 Example 3

If $X_1, X_2, \cdots$ is a sequence of independent standard uniform variables Unif[0,1] with $F(x) = x$ for $x \in [0, 1]$. For fixed $z < 0$, suppose $n > -z$ and we let $a_n = 1/n$ and $b_n = 1$, and we get

$$\begin{aligned}\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] &= F^n(n^{-1}z + 1)\\ &= \Big(1 + \frac{z}{n}\Big)^n\\ &\to e^z\end{aligned}$$

as $n \to \infty$ for each fixed $z \in \mathbb{R}$. This means, the M_n in this case follows the Weibull distribution with $b = 0$, $a = 1$, and $\alpha = 1$. The EVI ξ is -1.

2.4 Inferences for GEV

After having a statistical model, the next goal will be to make some inferences about it. There are many objects we can infer, and many ways to do them. In this section, we will only be focusing on a small fraction of them. We will cover the maximum likelihood estimator (MLE) and inference for return levels.

2.4.1 Maximum Likelihood Estimator

Under the assumption that block maxima $Z_1, \cdots, Z_m$ are independent variables having the GEV distributions, the log-likelihood for the GEV parameters when $\xi \neq 0$ is given by

$$\ell(b, a, \xi) = -m \log a - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - b}{a} \right) \right] - \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - b}{a} \right) \right]^{-1/\xi}$$

provided that

$$1 + \xi \left(\frac{z_i - b}{a} \right)$$

for $i = 1, 2, \cdots, m$.

When the data inequality is violated, it implies that some of the data falls beyond an end-point of the GEV, in which case the likelihood is 0 and the log-likelihood is $-\infty$.

When $\xi = 0$, a separate treatment is needed and we will be using the Gumbel limit of the GEV. For this case, the log-likelihood becomes

$$\ell(b, a) = -m \log a - \sum_{i=1}^{m} \left(\frac{z_i - b}{a} \right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{z_i - b}{a} \right) \right\}.$$

No analytic solution is available for the maximum of the above likelihoods, but common optimisation techniques could be applied to approximate the MLEs, although additional care is needed to make sure the data inequality is not violated as well as the ξ value assumption.

A potential difficulty with using MLEs for GEV is that the regularity conditions that are required for the usual asymptotic properties associated with the MLE may not the satisfied due to the end points of the distribution. In Smith (1985) [12], this problem is studied in detail and the following results are obtained:

- when $\xi > -0.5$, MLEs are regular, and have the usual asymptotic properties
- when $-1 < \xi < -0.5$, MLEs are generally obtainable, but do not have the standard asymptotic properties
- when $\xi < -1$, MLEs are unlikely to be obtainable

However, the cases where $\xi \leq -0.5$ implies the distribution has a very short bounded upper tail, which is rarely encountered in applications.

2.4.2 Return Level Inference

If we have the MLEs, we can estimate the return levels by substitution. The maximum likelihood estimate of the $1/p$ return level z_p for $0 < p < 1$ is obtained by

$$\hat{z}_p = \begin{cases} \hat{b} - \frac{\hat{b}}{\hat{\xi}} \left[1 - y_p^{-\hat{\xi}} \right] & \text{for } \hat{\xi} \neq 0 \\ \hat{b} - \hat{a} \log y_p & \text{for } \xi = 0 \end{cases}$$

where $y_p = -\log(1-p)$.

The interpretation of return level inferences, especially for return levels corresponding to long return periods, need to be careful. First, the normal approximation of the distribution of the MLE may be poor. More fundamentally, estimates and their measure of precision are based on the assumption that the model is correct, which may not always be the case in applications. This means, measures of uncertainty on return levels should properly be regarded as lower bounds, as they could be much greater if uncertainty due to model correctness were taken into account.

Chapter 3

Peaks-Over-Threshold Models

In this chapter, we will start with motivating the Peaks-Over-Threshold approach for extreme value modelling, and describe the Generalised Pareto Distribution, which is the theoretically justified model for exceedance beyond a threshold. Next, the Hill estimator and its corresponding Hill plot will be studied. They serve as inferential tools for the shape parameter of the Generalised Pareto Distribution. Then, we will mention a few methods to select the thresholds. Two graphical methods - mean residual plot and parameter stability plot - are mentioned, and one numerical method - Conor's Method - is mentioned.

3.1 Generalised Pareto Distribution

In Chapter 2, we have been discussing block-wise approaches for extreme value modelling. Block-wise approaches tend not to be too data-efficient. For example, if we only consider annual maximum, we would have wasted a lot of data that we have collected, resulted in a less accurate model. Additionally, we may have more than one extreme events in one block, and selecting only the maximum of each block will disregard that. This issue could be partially remedied by selecting r largest order statistics, but that may result in the selection of less than ideal data points. Also, the model for r largest order statistics, although is somewhat similar to that of block-wise maximum, is rather cumbersome to write down and do inference on. It is thus quite clear that block-wise models are intuitive but may not be the best option in many cases. As an alternative, we may consider the peaks-over-threshold models which select all data with values beyond a particular threshold.

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables following the distribution function F. We will let u be our threshold, and consider all values beyond u as extreme. Since all the random variables are i.i.d., we will pick an arbitrary one and denote it as X to simplify the notation. With this, we can obtain the following equation:

$$\begin{aligned}
F(u+y) &= \mathbb{P}(X \leq u+y) = \mathbb{P}(X \leq u) + \mathbb{P}(u \leq X \leq u+y) \\
&= F(u) + \mathbb{P}(X \geq u, X \leq u+y) \\
&= F(u) + \mathbb{P}(X \leq u+y \; | X \geq u)\mathbb{P}(X \geq u) \\
&= F(u) + F_u(y)(1-F(u))
\end{aligned}$$

where $F_u(y) := \mathbb{P}(X \leq u+y \; | X \geq u)$ and $y > 0$. After some rearrangement of the above equation, we get the following statement:

$$1 - F_u(y) = \mathbb{P}(X \geq u + y \mid X \geq u) = \frac{1 - F(u+y)}{1 - F(u)}$$

for $y > 0$. The value $\mathbb{P}(X \geq u + y \mid X \geq u)$ describes the behaviour of threshold exceedances, which is the extreme events in this context. The above equality tells us that, if we know F, we could obtain the distribution function of the threshold exceedances, and that is the extreme value modelling that we are trying to achieve. Sadly, F is not always obtainable in real life applications, so we cannot get the extreme value modelling directly in this manner.

For $X_1, \cdots, X_n$ with distribution F and $M_n := \max(X_1, \cdots, X_n)$, we assume it satisfies the conditions of Theorem 1, i.e. we have sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq z\Big] \to G(z)$$

as $n \to \infty$, where G is a non-degenerate (i.e. not point mass) distribution function, then G has the form

$$G_\xi(x) = \exp\Big(- (1 + \xi x)^{-1/\xi}\Big), \qquad 1 + \xi x > 0$$

or just

$$G_0(x) = \exp(-\exp(-x)).$$

We will derive here, in a rather hand-wavy manner, the distribution for conditional exceedance F_u. For large enough n and $\xi \neq 0$, we have

$$F^n(z) \approx \exp\left\{ - \left[1 + \xi\left(\frac{z-b}{a}\right)\right]^{-1/\xi}\right\}$$

which is equivalent to

$$n \log F(z) \approx - \left[1 + \xi\left(\frac{z-b}{a}\right)\right]^{-1/\xi}$$

For large z, $F(z) \to 1$, so $\log F(z) \sim F(z) - 1$, and we have

$$1 - F(z) \sim \frac{1}{n}\left[1 + \xi\left(\frac{z-b}{a}\right)\right]^{-1/\xi}.$$

Evaluating the above equation at $z = u$ and $z = u + y$ respectively gives us

$$1 - F(u) \sim \frac{1}{n}\left[1 + \xi\left(\frac{u-b}{a}\right)\right]^{-1/\xi}$$

and

$$1 - F(u+y) \sim \frac{1}{n}\left[1 + \xi\left(\frac{u+y-b}{a}\right)\right]^{-1/\xi}$$

which lead to

$$\mathbb{P}(X \geq u + y \mid X \geq u) = \frac{1 - F(u+y)}{1 - F(u)} = \frac{\left[1 + \xi\left(\frac{u+y-b}{a}\right)\right]^{-1/\xi}}{\left[1 + \xi\left(\frac{u-b}{a}\right)\right]^{-1/\xi}} = \left[1 + \frac{\xi y}{a_u}\right]^{-1/\xi}$$

where $a_u = a + \xi(u - b)$. Thus, we get the distribution of F_u, which is

$$F_u(y) = 1 - \left[1 + \frac{\xi y}{a_u}\right]^{-1/\xi}$$

when $\xi \neq 0$ and $y > 0$. A similar computation can be adapted for the case where $\xi = 0$, which gives us the following distribution function:

$$F_u(y) = 1 - e^{-y/a_u}$$

where a_u is defined in the same way as the $\xi \neq 0$ case. This distribution is known as the **generalised Pareto distribution** (GPD).

The above derivation of GPD is lifted from Coles (2001) [3]. The rigorous version of the above statement is known as the Pickands–Balkema–De Haan theorem, and its proof can be found at many places, for example Leadbetter et al. (1983) [10].

To summarise, the GPD at threshold / location u has two parameters: location ξ and scale $a_u = a + \xi(u - b)$. For simplicity of notation, we will denote the two parameters simply as ξ and σ where σ is a_u. This means, the CDF of GPD becomes

$$F_u(y) = \begin{cases} 1 - \left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi} & \text{for } \xi \neq 0 \\ 1 - e^{-y/\sigma} & \text{for } \xi = 0, \end{cases}$$

where $y > 0$. The density function of GPD is plotted below.

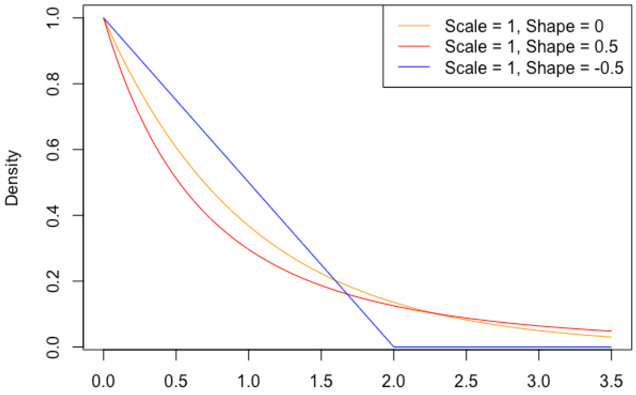


Figure 3.1: GPDs with Different Shape Parameter

A very nice property of GPD is its parameter stability. For a true threshold u of some samples, the exceedance of samples over u follow the GPD asymptotically. The property is that, if select the samples over a higher threshold v with $v > u$, the exceedance follow a GPD as well, with the same shape parameter and slightly different scale parameter. This property is going to be exploited later in this chapter when we discuss the diagnostics plots.

3.2 GPD Examples

In Chapter 2, we have mentioned three examples of GEV. In this chapter, we will do the same and illustrate the same three examples but with GPD instead of GEV. These examples are identical to those mentioned in Coles (2001) [3].

3.2.1 Example 1

For exponential model Exp(1) with CDF $F(x) = 1 - e^{-x}$ for $x > 0$, by direct calculation, we get

$$\frac{1 - F(u+y)}{1 - F(u)} = \frac{\exp(-u-y)}{\exp(-u)} = e^{-y}$$

for $y > 0$. This is GEV with $\xi = 0$ and $a_u = 1$, and it is an exact result for all thresholds $u > 0$.

3.2.2 Example 2

For the standard Freéchet model with CDF $F(x) = \exp(-1/x)$ for $x > 0$, we have

$$\frac{1 - F(u+y)}{1 - F(u)} = \frac{1 - \exp(-(u+y)^{-1})}{1 - \exp(-u^{-1})} \sim \left(1 + \frac{y}{u}\right)^{-1}$$

as $u \to \infty$ for all $y > 0$. This is the GPD with $\xi = 1$ and $a_u = u$.

3.2.3 Example 3

For the uniform distribution model Unif[0,1] with CDF $F(x) = x$ on $x \in [0, 1]$, we have

$$\frac{1 - F(u+y)}{1 - F(u)} = \frac{1 - (u+y)}{1-u} = 1 - \frac{y}{1-u}$$

for $y \in [0, 1-u]$. This is the GPD with $\xi = -1$ and $a_u = 1 - u$.

3.3 Hill Estimator

One of the key parameters of GPD is the shape parameter ξ. Its value influences significantly how the distribution will behave, and thus is rather important if we can find ways to estimate it. The Hill estimator, proposed in Hill (1975) [8], provides a theoretically justified method to estimate this parameter. In fact, the paper works with a more general class of distributions - heavy-tailed distribution F with its tail varying regularly with tail index $1/\xi$ - and GPD belongs to this class.

For a sequence of identically distribution (not necessarily independent) random variables $X_1, \cdots, X_n$, we will order them and denote them as $X_{(i)}$ with $X_{(1)}$ being the largest. Using this notation, the Hill estimator of ξ based on the k upper order statistics is defined by

$$\hat{\xi}_k = \frac{1}{k-1} \sum_{j=1}^{k-1} \log\left(\frac{X_{(j)}}{X_{(k)}}\right)$$

for $2 \leq k \leq n$.

In practice, the Hill estimator is used as follows. First, we will compute the Hill estimator at each possible k, then plot the points $\left(k, \hat{\xi}_k\right)$. We will then select the estimator from the set of $\hat{\xi}_k$ that are stable.

An example of Hill estimator plot is shown below, using the Fort Collins precipitation data.

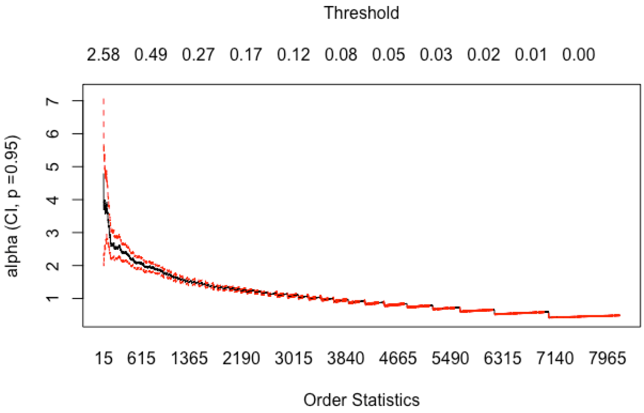


Figure 3.2: Hill Plot for Fort Collins Precipitation Data

As we can see, the Hill estimator plot decreases towards the end, and it is hard to find a region where the plot is stable. We may possibly guess a value between 0 and 1 is plausible, but there is no guarantee if we only look at this plot. The Hill plot will work sometimes, but when it does not it provide very limited information, thus making it not a desirable diagnostic tool.

3.4 Threshold Selection for GPD

For a sequence of i.i.d. data $X_1, \cdots, X_n$, we consider the set of data beyond the threshold u, $\{X_i : X_i > u\} = \{X'_1, X'_2, \cdots, X'_k\}$. For those, we will consider their exceedance over the threshold, and get $\{Y_1, Y_2, \cdots, Y_k\}$ with $Y_i = X'_i - u$. These Y_i follow GPD asymptotically due to the theoretical result mentioned earlier in this chapter, so we can try to fit a GPD using these Y_i.

This sounds great, but it requires a selection of good threshold u. If the threshold is too low, we may have bias as the fit of GPD might be poor. If the threshold is too high, we may have a good fit of GPD but there will be less data which cause an increase in variance. A trade-off between bias and variance thus exists for the problem of threshold selection.

There are many ways to select a threshold, and it is still a big topic of research in extreme value theory, especially when we try to implement the methods for data of varying structures and properties. Some of the many methods will be mentioned in this notes, but this is certainly not extensive.

3.4.1 Mean Residual Life Method

If we have a random variable Y that follows a GPD with parameters ξ and σ, its expectation is

$$\mathbb{E}(Y) = \frac{\sigma}{1-\xi}$$

when $\xi < 1$. The expectation will be ∞ if $\xi \geq 1$, so the following discussion assumes ξ to be less than 1.

For a sequence of i.i.d. data $X_1, X_2, \cdots, X_n$, we denote an arbitrary term of them simply as X. For its threshold u_0, we have the following equation

$$\mathbb{E}[X - u_0 | X > u_0] = \frac{\sigma_{u_0}}{1-\xi}$$

where σ_{u_0} is the scale parameter for threshold u_0. For a threshold u with $u > u_0$, the exceedance will still follow a GPD but just with different parameters. So, we get an equation that is similar to the one above:

$$\mathbb{E}[X - u | X > u] = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi(u - u_0)}{1-\xi}.$$

If we treat the left-hand side of the above equation as a function of u, we get a function $f(u)$ defined by

$$f(u) := \mathbb{E}[X - u | X > u] = \frac{\sigma_{u_0} + \xi u}{1-\xi} = \frac{\sigma_{u_0}}{1-\xi} + (u - u_0) \cdot \frac{\xi}{1-\xi}.$$

It is clear that, as u varies, $f(u)$ changes linearly with gradient $\xi/(1-\xi)$. This property motivates the **mean residual life** approach of threshold selection.

We can create a plot with points of the form

$$\left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right)$$

where n_u is the number of exceedance for threshold u and $x_{(i)}$ are the exceedance points. The second coordinate of each of the points above is an approximation of $f(u)$, and this quantity is called as the mean residual life, which is why the plot is known as the mean residual life plot. This means, the plot should be linear after a certain value of u, and that certain point will be the true threshold u_0 based on our previous derivations. This method was first proposed in Yang (1978) [15].

The idea behind this method sounds intuitive and easy, but it turns out to be hard to work with in many real life applications. The hard part is that it is rather unclear how one should choose the point u_0 for which the points afterwards are linear.

A mean residual life plot has been created for the Fort Collins precipitation data mentioned in Chapter 1. The grey lines are the 95% confidence interval based on the approximated normality of sample mean.

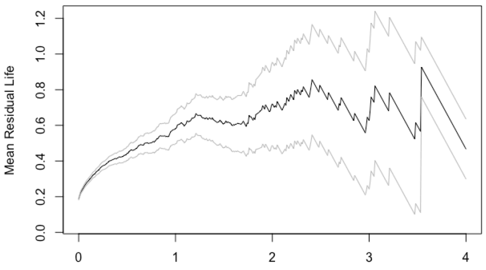


Figure 3.3: Mean Residual Life Plot of England Monthly Average Temperature Data

From the figure above, it is hard to locate u_0 as there are multiple plausible values if we only focus on the linearity criteria. Here, we can pick $u_0 = 1.2$, $u_0 = 2.4$, or $u_0 = 3.1$. All data after

1.2 seems to be linear, yet there is a small kink at 2.4 and another one at 3.1, so it might not be an ideal threshold. Also, threshold at 3.1 might be too high, and leave us with too little data. So, we might end up choosing 2.4 as our threshold.

The above example is already a relatively clear-cut one, and it could certainly get a lot worse. Also, the reasoning behind the choice of 2.4 is not the most convincing, and a lot of subjectivity are involved. Some more, when we have more than one variables of our data, we will need to apply this procedure of threshold selection to all variables, which will be very time consuming. Overall, the mean residual life plot is rather intuitive and visual, but requires a lot of ad hoc judgements and does not adapt well with increasing dimensionality of data.

3.4.2 Parameter Stability Plot

If we have found a good threshold, we can use MLE to estimate the parameters ξ and σ of GPD. Recall that the CDF of GPD is

$$F_u(y) = \begin{cases} 1 - \left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi} & \text{for } \xi \neq 0 \\ 1 - e^{-y/\sigma} & \text{for } \xi = 0. \end{cases}$$

Assume $y_1, \cdots, y_k$ are k excesses of a threshold u. If $\xi \neq 0$, the log-likelihood is

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi y_i/\sigma)$$

provided that $1 + \xi y_i/\sigma > 0$ for all i. If not, $\ell(\sigma, \xi) = -\infty$. When $\xi = 0$, the log-likelihood is of the form

$$\ell(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i.$$

The above results require a selection of threshold, yet the threshold is normally unknown. However, we may consider to exploit the above results to help us obtain a good threshold.

For true threshold u_0 of a sequence of data, data beyond u_0 can be modelled by a GPD. Additionally, data beyond u where $u > u_0$ follows a GPD as well. The shape parameters of these two GPDs are identical, but the other parameter σ will vary. We have, by definition,

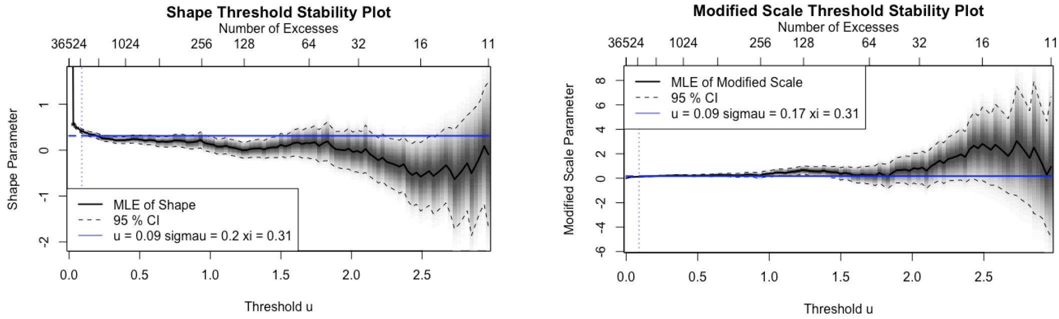
$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$

where σ_u is σ for GPD with threshold u, and similar for σ_{u_0}. This can be adjusted in order for it to be constant as well. We will declare a new parameter σ^* defined by

$$\sigma^* := \sigma_u - \xi u = \sigma_{u_0} - \xi u_0,$$

and this clearly is independent of the value of u. So, as a result, the parameters σ^* and ξ should be constant for any thresholds above the true threshold u_0, and we can use this property to select threshold. It is not hard to realise this approach is rather similar to that of mean residual plot. This method was first proposed in Davison and Smith (1990) [4].

An example of threshold stability plot is included below, using the same Fort Collins Precipitation data as before.



(a) Shape Parameter Stability Plot

(b) Modified Scale Parameter Stability Plot

Figure 3.4: Threshold Stability Plots of Fort Collins Precipitation Data

If we read directly off the graph, we can realise the estimated threshold in this case is 0.09 inch, which is very different from the result we obtain using mean residual life plot.

3.4.3 Conor's Method

The two methods shown above are both graphical, and have existed for a while. Since then, newer methods have appeared, such as Northrop and Coleman (2014) [11] and many others. Here, we will present a method[1], proposed by Conor Murphy[2] that is rather recent and we believe it performs better than many of the existing methods.

As described earlier this chapter, for a sequence of i.i.d. random samples, condition on the value of the random samples being greater than a threshold u, the exceedance of the random sample over the threshold u follows the GPD asymptotically. From this result, we can infer that for sufficiently large number of samples, the exceedance over the correct threshold can be fitted nicely by a GPD. This motivates a series of methods for threshold selection, by picking a few threshold options and test how well the exceedance fit the GPD in each case. There are many obstacles with this approach. Among them, one major obstacle is that a direct comparison of each threshold is not feasible due to the fact that each fitted model uses different data sets.

Another useful information that one could pick up from the asymptotic GPD result is that, although we have shown the convergence is true, we do not really know how fast the convergence is - the convergence rate varies with sample distribution. If the convergence is fast, a small sample size will be sufficient. However, if the convergence rate is low, we would require a large sample size, usually more than what we have, to achieve the desired convergence. This leads to issues for threshold selection, since we may never get a good enough fit of GPD regardless of which threshold we try. To tackle the general issue of potential slow convergence to GPD, a set of theory, known as the **penultimate extreme value theory**, has been developed. Early results of this theory can be found in Smith (1985) [12]. Roughly speaking, the idea of this theory is that since the full convergence cannot be achieved with the limited sample size, we will try to find a model that describes the distribution when we are close to full convergence yet not exactly converged to GPD. This new model will serve as a replacement of the GPD in the later modelling, and motivate modelling using a modified GPD with step-wise parameters instead of constant parameters in the usual cases. Threshold selection models that develop using this idea

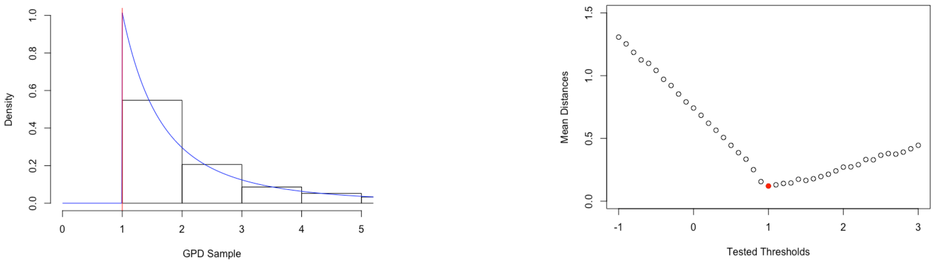
[1] As of now, it is unpublished.

[2] One of the two supervisors of my internship.

include Wadsworth and Tawn (2012) [14] and Northrop and Coleman (2014) [11] . The method we propose here, however, does not employ this line of thought.

The method we are presenting here follows the GPD fit approach. For each of the proposal thresholds, we will estimate the GPD parameters via MLE using the exceedance data. This is rather standard. Next, in order to find which proposed threshold to select, we will measure the deviance between the exceedance and GPD with estimated parameters using the average distance at many fixed equally gapped quantiles. This distance, adapted from the work of Varty et al (2021) [13], measures the closeness of GPD fit at various equally spaced quantile levels, which can be visually viewed as the average error of the QQ plot to the diagonal line of perfect fit. Loosely speaking, this distance, if we only use the given samples, will indicate the bias of the threshold well, but indicate the variance poorly. To remedy that, we will use bootstrapping to resample exceedance and measure that same distance. The final comparison between various proposal thresholds compares the mean of average calculated bootstrapped distances at each threshold level, and the threshold that minimises this quantity will be selected. Essentially, this method incorporates both the bias and variance of the GPD fit to select the threshold.

We have sampled GPD samples with shape = 0.5, scale = 1, and threshold = 0. The true threshold is thus 1. We apply this method to this sample. The proposal thresholds are from 0 quantile to 95 quantile with increment of 5 quantile. We do 100 bootstrapping for each threshold level, and the deviance distance is computed at 500 equally spaced quantile levels. The estimated threshold of our method is around 0, and in Figure 3.5a we have plotted a histogram of samples along with the fitted GPD curve (in blue) with estimated parameters on it. The red vertical line is the estimated threshold. In Figure 3.5b, we have plotted the mean distances at various proposed thresholds, with the minimum point highlighted in red. This value of this point gives us the threshold, as shown in Figure 3.5a.



(a) Conor's Method with GPD Samples

(b) Mean Distance Plot with Minimum in Red

Figure 3.5: Conor's Method Demonstration

Chapter 4

Mixture Models

In this chapter, we will study several mixture models to do univariate extreme value modelling. We will go over three methods - the standard extreme value mixture model, the dynamically weighted mixture model, and the hybrid Pareto distribution model.

Disclaimer: In each section, the notations are directly lifted from the papers each section is referring from, so they are not standardised.

4.1 Standard Extreme Value Mixture Model

The simplest extreme value mixture model was proposed by Behrens Lopes Gamerman (2004) [1], and similar models have been widely developed since then.

In the previous two chapters, we have considered extreme value modelling that use partial data. Here, we will consider the mixture models, which employ all the collected data. A mixture model will be a mix of two (or more) distributions, usually for different parts of the distribution. Here, we are mixing two distributions, one for the 'bulk' and one for the 'tail'. The bulk distribution tends to vary, while the tail distribution is almost always assumed to be following the GPD (or some form of it) as we have theoretical guarantees for the behaviour of exceedance over threshold. The mixing is generally a weighted average of the two distribution, and different choices of weights will result in different models.

If we denote the bulk distribution CDF to be H and the tail GPD to have CDF G , then the CDF of the mixture F is given by

$$F(x) = \begin{cases} H(x) & x \leq u \\ H(u) + (1 - H(u))G(x) & x > u, \end{cases}$$

where u is the threshold. The density function of the mixture is thus given by

$$f(x) = \begin{cases} h(x) & x \leq u \\ (1 - H(u))g(x) & x > u, \end{cases}$$

where h and g are densities of H and G respectively.

A problem with this model is that, although the CDF is continuous, the density function may contain a discontinuity at u. This could be remedied by imposing additional constraints on the density function.

Also, the reasoning behind using GPD to model the extreme events is that the the non-extreme events provide little information about the extreme behaviour. So, it is rather counter-intuitive to consider the bulk in the modelling. The inclusion of a bulk distribution may also create additional issue to the model, as we may have misspecified the bulk and sabotage the overall fit. Misspecification can be protected by some reformulation, as shown in Chapter 3 of Dey and Yan (2016) [5].

4.2 Dynamically Weighted Mixture Model

This section is based on Frigessi, Haug, and Rue (2002) [6].

Let $X_1, X_2, \cdots, X_n$ be **non-negative** i.i.d. random variables with common probability density function $l(x)$ given by (with parameters omitted in the formula)

$$l(x) = \frac{(1-p)f + pg}{Z}$$

where

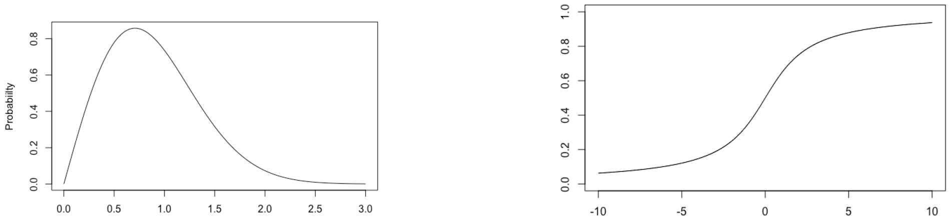
$$g(x;\xi,\sigma) = \frac{1}{\sigma}\left(1 + \frac{\xi x}{\sigma}\right)^{-1-1/\xi}$$

is the density of GPD, $p(x,\theta)$ is increasing in x, takes values in $(0,1]$, and satisfies $\lim_{x\to x_\infty} p = 1$. The function f is some density function of the bulk distribution, while Z is the normalising constant.

The bulk distribution is chosen to be a Weibull distribution with density

$$f(x;\beta,\lambda) = \beta\lambda^\beta x^{\beta-1}\exp[-(\lambda x)^\beta]$$

for positive β and λ. Figure 4.1a is a plot of this distribution - notice it has light tail.



(a) Weibull Distribution with shape = 2, scale = 1

(b) Weight Function with $\mu = 0$ and $\tau = 2$

Figure 4.1: (a) Weibull Distribution (b) Weight Function

For the weighting function p, two options are proposed in the paper. We will denote them as p_1 and p_2.

The weight p_1 is defined as follows:

$$p(x;\theta) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\tau}\right)$$

where $\theta = (\mu, \tau)$ and $\mu, \tau > 0$. The parameter μ indicates its location, while parameter $1/\tau$ indicates its steepness. Note that this function is the CDF of a Cauchy distribution. A plot of this function is shown in Figure 4.1b, with $\mu = 0$ and $\tau = 2$. As the value of x increases, the function will tend to 1, which means the mixture model will tend to be more like GPD g and less like the bulk f. The normalising constant Z in this case has an ugly formula and could only be computed numerically. The exact formula can be found in the paper [6].

The second weight function p_2 has a simpler definition. It is just a step function that take 0 before θ and 1 after θ. A formal definition is shown below.

$$p(x;\theta) = \begin{cases} 1, & x \geq \theta \\ 0, & x < \theta \end{cases}$$

This means, the mixture model will be following the bulk distribution f fully when $x < \theta$ and be following the tail distribution GPD g fully when $x \geq \theta$. This function is, in fact, p_1 with $\tau \to 0$. It is not hard to notice that this parameter θ is simply the threshold for the GPD. One thing that the readers should note is that using this weight function, the mixture model is generally going to have a discontinuity at $x = \theta$, which may not be an ideal property. Because of this, we will mostly use the model with weight function 1.

So, the model with f being the density of Weibull distribution and p being the weight function 1 mentioned earlier is commonly known as the **dynamically weighted mixture** model (DWM). The density of DWM is $l(x) = [(1-p)f + pg]/Z$, and this equation can be reformulated to make the model as a pure mixture model. We define the following terms

$$A := \int_0^\infty f(1-p)\, dx \qquad B := \int_0^\infty gp\, dx.$$

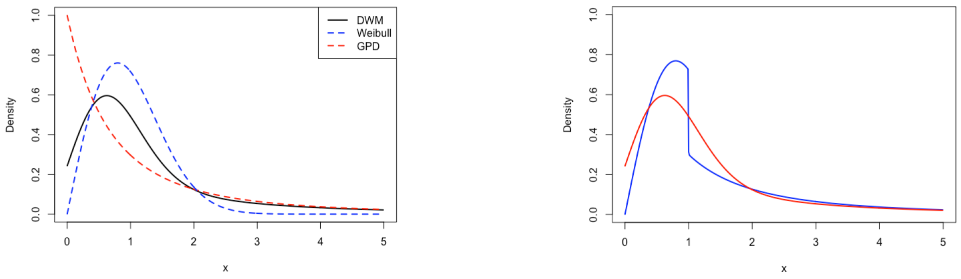
Using this, we have

$$\pi := \frac{B}{A+B} \qquad g_1(x) := f(1-p)/A \qquad g_2(x) := gp/B.$$

Compiling them together, we have the density of DWM as

$$l(x) = (1-\pi)g_1 + \pi g_2.$$

Figure 4.2a includes DWM, Weibull, and GPD. The are plotted together so the readers can see how the mixture is done. The exact parameters used in this plot can be found in the paper [6].



(a) Plot Example in Frigessi et al. (2002)

(b) Step-Wise Weight vs. Continuous Weight

Figure 4.2: DWM Plots

As mentioned earlier, we can choose the weight function to be step-wise (option 2) instead of continuous (option 1). We have also remarked that it is not ideal to use the step-wise weight as it creates a discontinuity. To graphically illustrate this point, we have plotted a DWM using the step-wise weight (in blue) and a DWM with continuous weight (in red) on the same graph, see Figure 4.2b. Do note that the parameters of both these two plots are identical except for the part that determines the weight function.

The following figure is a demonstration of fitting some Weibull distributed data using a DWM. The histogram is of the samples from Weibull with shape = 2 and scale = 1, and the black curve is the density function. The red curve in the figure is the density function with estimated parameters, which is quite close to the black curve - the fit is good!

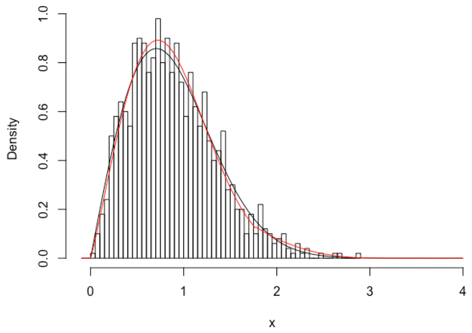


Figure 4.3: Fitting a Weibull with shape = 2, scale = 1

4.3 Hybrid Pareto Model

This section is based on Carreau and Bengio (2009) [2].

In this section, we will introduce a different mixture model that is designed for data that are asymmetric and have heavy tails. This model is known as the hybrid Pareto distribution (HPD). Essentially, this model is a mix of a Gaussian bulk with a GPD tail, under some smoothness constraints at the junction point of connection.

The two ingredients of this model, as said earlier, are the normal distribution and the GPD. The density of the normal is simply

$$f_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

where the subscript of f denotes the parameters. The density of the GPD is

$$g_{\xi,\beta}(y-\alpha) = \begin{cases} \frac{1}{\beta}\left(1+\frac{\xi}{\beta}(y-\alpha)\right)^{-1/\xi-1} & \xi \neq 0 \\ \frac{1}{\beta}\exp\left(-\frac{y-\alpha}{\beta}\right) & \xi = 0, \end{cases}$$

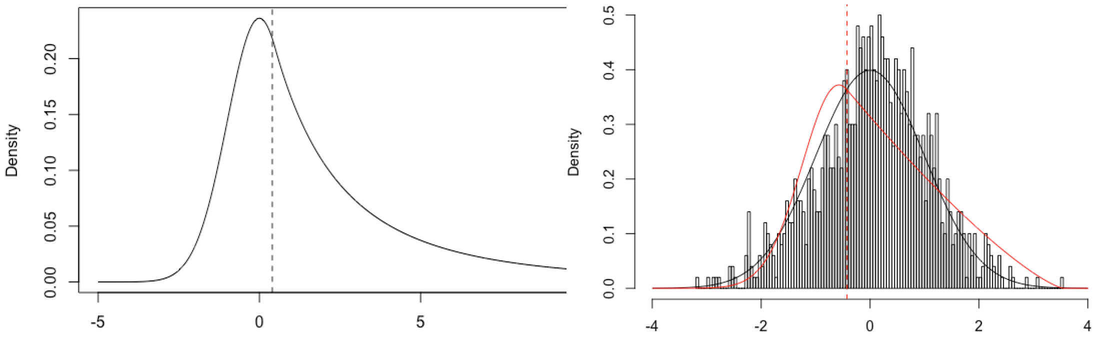
where $y \geq \alpha$ when $\xi \geq 0$ and $\alpha \leq y \leq \alpha - \beta/\xi$ when $\xi < 0$.

There are five parameters $\mu, \sigma, \xi, \alpha, \beta$, where α is the threshold for GPD while the rest are standard parameters for each respective distribution. We would like to have some form of continuity at α, which impose two constraints: (1) continuity at α, $f(\alpha) = g(0)$ (2) continuity of derivative at α, $f'(\alpha) = g'(0)$. This restrict some freedom of parameter choice, so we set ξ, μ, σ to be free while α, β to be determined on the choice for those three parameters.

The derivation of the density for HPD is omitted here. A detailed derivation can be found in the original paper by Carreau and Bengio (2009) [2]. We would only show the density function for HPD here, which is given by

$$h(y) = \begin{cases} \frac{1}{\gamma} f_{\mu,\sigma}(y) & y \leq \alpha \\ \frac{1}{\gamma} g_{\xi,\beta}(y-\alpha) & y > \alpha, \end{cases}$$

where γ is the normalising constant. The density function of HPD is plotted in Figure 4.4a, where the vertical dotted line is the threshold.



(a) Hybrid Pareto Distribution with $\mu = 0$, $\sigma = 1$, $\xi = 0.4$

(b) Fitting Normal Samples with Hybrid Pareto Distribution

Figure 4.4: Hybrid Pareto Model Plots

We tried to fit the HPD to a normal sample. Note that since the HPD is designed to better model data that are asymmetric and heavy-tailed, this fit is not going to be too good. In the Figure 4.4b, a histogram of data is shown, with the black curve representing the true density and the red curve representing the fitted model density. The red vertical line is the threshold of the fitted model.

Chapter 5

Simulation Study

In this chapter, we conducted two simulation studies with the threshold selection methods described in previous chapters. The first study investigates the rate of convergence to the GPD for different distributions. The second study compares the performance of Conor's Method, DWM, and HPD when they try to estimate the threshold for samples from different distributions.

Remark: Even though not explicitly studied, the computing time required for HPD is the shortest, and DWM is the longest.

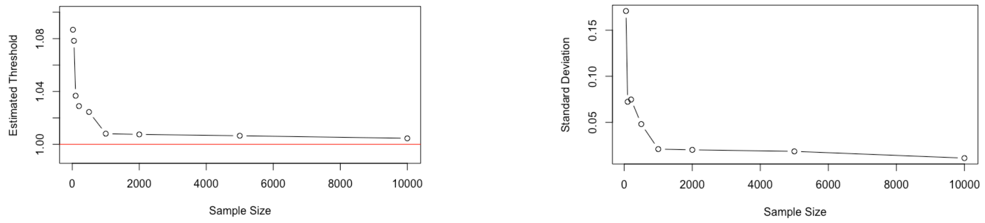
5.1 Convergence to GPD Simulation

In Chapter 3, we have introduced the GPD, as well as the theoretical justification of modelling exceedance over a threshold of a sequence of data using the GPD. The theorem shows the convergence to GPD, but the actual rate of convergence is not described in the theorem. Different distributions will converge at different rate, and this simulation study aims to investigate it.

Conor's Method, described in Section 3.4.3, is a good method to estimate the threshold. If we have a sufficiently large sample size, we will be expected to obtain a better threshold since the threshold selection method depended on how well the exceedance fit a GPD. Though not directly proved, it is still plausible to believe in the correlation between the trend of estimated threshold and the convergence to GPD.

The simulation experiment is designed as follows. For varying sample sizes (20, 50, 100, 200, 500, 1000, 2000, 5000, 10000), we will generate samples of that size following one of three distributions (GPD with positive shape, Standard Normal, Beta) and estimate the threshold using Conor's Method. The procedure of generating sample and estimating threshold is repeated for 100 times for each of the sample sizes. For GPD samples, we know the true threshold so we will compare our estimations with that. For the other two distributions, we do not know their true thresholds but can still observe the trend. For each distribution, we will show two plots. The first plot is the average threshold estimation for each of the sample sizes. The second plot is the standard deviation of the 100 estimations.

For GPD samples, as shown in Figure 5.1a and Figure 5.1b, the estimated threshold gets stabilised after sample size 1000. Not much improvements have been made afterwards even though we have obtained a lot more samples. Additionally, if we pay closer attention to the values of the estimated threshold, the estimation is still quite good with small sample size. For example, at size 50, the estimation is 1.08, which is not too far from the truth. Also, we can see that the standard deviation decreases as sample size increases, indicating an decrease in the variance of the estimator.

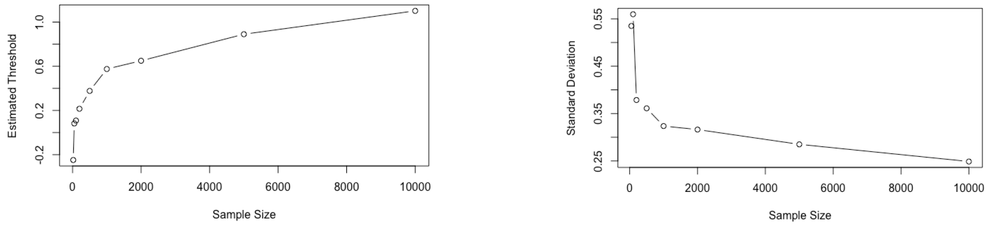


(a) Threshold Estimation

(b) Standard Deviation

Figure 5.1: GPD Samples with shape = 0.5, scale = 0.5, location = 1

For standard normal samples, as shown in Figure 5.2a and Figure 5.2b, the estimator is still not very stable after sample size 5000. The increment is decreasing, which hints the convergence. We can certainly infer that standard normal distribution takes a long while to converge. The standard deviation plot illustrates a general decreasing trend, as predicted.

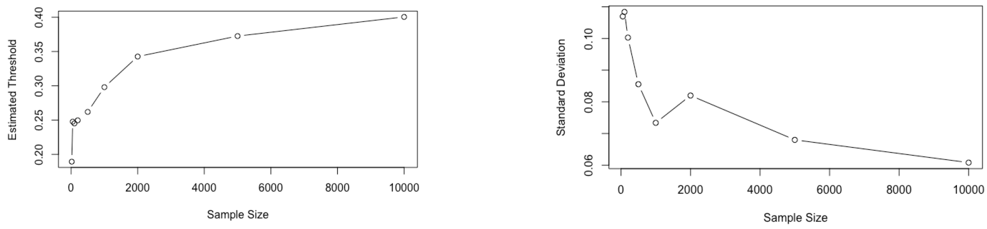


(a) Threshold Estimation

(b) Standard Deviation

Figure 5.2: Standard Normal Samples

For Beta(2,5) samples, as shown in Figure 5.3a and Figure 5.3b, the estimator is still not very stable after sample size 5000. The increment is decreasing, which hints the convergence. We can certainly infer that standard normal distribution takes a long while to converge. The standard deviation plot illustrates a general decreasing trend, as predicted.



(a) Threshold Estimation

(b) Standard Deviation

Figure 5.3: Beta(2,5) Samples

In Figure 5.4, we have combined Figure 5.2a and Figure 5.3a into the same graph. This allows us to make more direct comparison, and enable us to compare the scales. It is thus very clear to us the standard normal converges much slower than that of Beta(2,5).

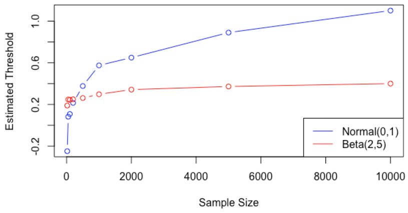


Figure 5.4: Comparison between Normal(0,1) and Beta(2,5)

5.2 Return Levels Estimation Comparison

In this section, we will study and compare the performance of different univariate extreme value models when the samples received follow different distributions.

The three models we will be studying here are Conor's Method (described in Section 3.4.3), DWM (described in Section 4.2) with continuous Cauchy weight function, and HPD (described in Section 4.3). These three models are selected since they are automated, and they are believed to be performing well in general. Graphical methods like mean residual life plot are not desirable if we want to run simulation studies, since they cannot be automated - a statistician needs to make a decision at each plot to determine the estimation. Also, the first method belong to the POT class of methods, while the latter two belong to the mixture class. Thus, we do have a good pool of methods to compare.

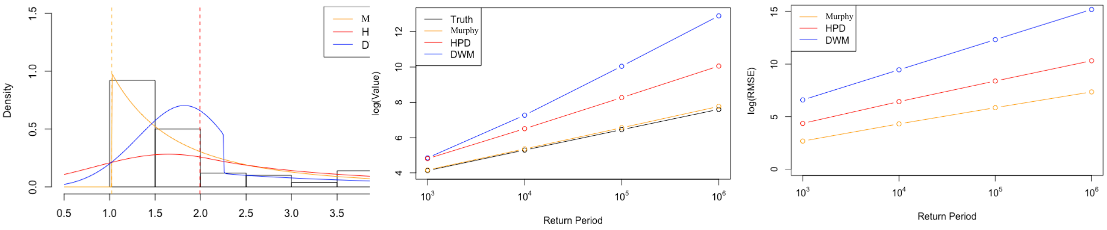
The underlying distributions of the samples will be GPD, Normal, and Beta. More specifically, the GPD will have scale = 1, shape = 0.5, the normal will be standard normal so mean = 0, variance = 1, and the Beta will have shape parameters 2 and 5. The reasoning behind this choice is that we are interested to see how the methods perform when the tail index (or EVI) ξ varies. The Normal distribution has an exponentially decaying heavy tail and $\xi = 0$. The Beta distribution has an upper end point and $\xi < 0$. The GPD distribution we sample from is set to have a positive ξ and thus a polynomially decaying light tail.

The simulation goes like the following. For each distribution and each model, we will generate a sample of size 1000, and fit the model using this sample. This process will be repeated for 100 times, and all the 100 sets of estimated parameters will be collected. The final estimated parameters for each model of each distribution will be the average of the 100 values. Using the final estimated parameters, we will then compute various return levels and create plots of the various densities for comparison.

5.2.1 GPD

One of the main quantity we would like to infer is the return levels, especially with a high return period. If we have 1000 samples, we would be interested in the 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6} return level. Since we will likely to have no samples at these levels, those return levels would be hard to estimate. So, we would like to investigate how well various models do regarding this task. Note that the return period is the reciprocal of the return level. From Figure 5.5b, we can notice the DWM does poorly, especially at higher return periods. Also, we can notice that Conor's Method provides the closest estimation at all return levels, and have the least RMSEs.

In Figure 5.5a, we have plotted the fitted curves together, along with a histogram of the samples as Figure 5.5b and a plot of RMSE as Figure 5.5c.

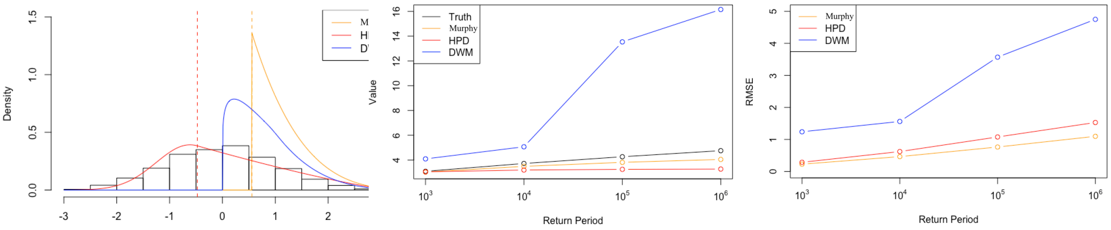


(a) Fitted Distributions (b) Estimated Return Periods (c) Root Mean Squared Error

Figure 5.5: GPD samples with scale = 1, shape = 0.5, location = 1

5.2.2 Standard Normal Distribution

For standard normal samples, we cannot compute its true threshold, so we are unable to compare how the threshold estimation of each model is. From Figure 5.6b, we can notice that the DWM is the furthest from the truth, with Conor's Method closest. In Figure 5.6a, we have plotted the fitted curves together, along with a histogram of the samples as Figure 5.6b and a RMSE plot as Figure 5.6c.

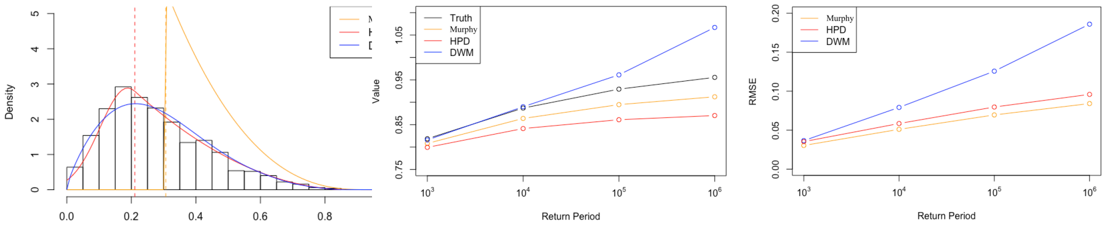


(a) Fitted Distributions (b) Estimated Return Periods (c) Root Mean Squared Error

Figure 5.6: Standard Normal Samples

5.2.3 Beta Distribution

For Beta samples, we cannot compute its true threshold, so we are unable to compare how the threshold estimation of each model is. From Figure 5.7b, we can notice that Conor's Method performs the best with closest return level estimations as well as least RMSEs, while the DWM overestimates and produce large RMSEs. In Figure 5.7a, we have plotted the fitted curves together, along with a histogram of the samples as Figure 5.7b and a RMSE plot as Figure 5.7c.



(a) Fitted Distributions

(b) Estimated Return Periods

(c) Root Mean Squared Error

Figure 5.7: Beta(2,5) Samples

Bibliography

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

[2] Carreau, J., Bengio, Y. (2008). A Hybrid Pareto Model for Asymmetric Fat-Tailed Data: the Univariate Case. *Extremes.* **12**(1), 53-76.

[3] Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values.* London: Springer London.

[4] Davison, A. C., Smith, R. L. (1990). Models for Exceedances Over High Thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, **52**(3), 393–425.

[5] Dey, D. K., Yan, J. (2016). *Extreme Value Modeling and Risk Analysis: Methods and Applications.* Chapman and Hall/CRC.

[6] Frigessi, A., Haug, O., Rue, H. (2002) A Dynamic Mixture Model for Unsupervised Tail Estimation without Threshold Selection. *Extremes.* **5**, 219–235 .

[7] Haan, L., Ferreira, A. (2006). *Extreme Value Theory: an Introduction.* New York London: Springer.

[8] Hill, B. M. (1975). A Simple General Approach to Inference about the Tail of a Distribution. *The Annals of Statistics*, **3**(5).

[9] Kotz, S., Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications.* Imperial College Press.

[10] Leadbetter, M. R., Lindgren, G., Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* New York, NY: Springer New York.

[11] Northrop, P.J., Coleman, C.L. (2014) Improved threshold diagnostic plots for extreme value analyses. *Extremes.* **17**, 289–303.

[12] Smith, R. L. (1985). Maximum Likelihood Estimation in a Class of Nonregular Cases. *Biometrika*, **72**(1), 67–90.

[13] Varty, Z., Tawn, J. A., Atkinson, P. M., Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv.* 2102.00884v1.

[14] Wadsworth, J.L., Tawn, J.A. (2012) Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B (Methodological).* **74**(3):543–567.

[15] Yang, G. L. (1978). Estimation of a Biometric Function. *The Annals of Statistics*, **6**(1).