Introduction to Extreme Value Analysis in R

Motivation and Block Maxima Approach

Lídia André[1] and Soraia Pereira[2]

[1]STOR-i CDT, Lancaster University, UK
[2]CEUAL, Faculdade de Ciências da Universidade de Lisboa, Portugal

23 March 2022

Why Extremes?

Extreme events, although rare, have a huge human impact

Various applications:

- financial sector - *e.g.* portfolio risk
- environmental research -*e.g.* catastrophes

Figure 1: Flood

Figure 2: Earthquake

Why Extremes?

Common statistical approaches are **not suitable** for modelling extreme events

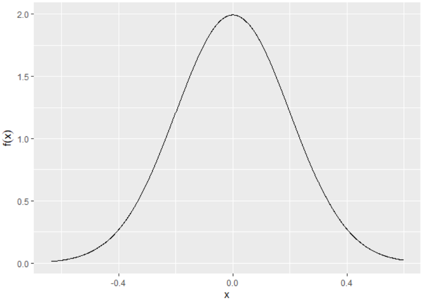

Figure 3: Normal Density

- the majority of points are concentrated towards the centre of the distribution
- estimation will be hard as observations in the tails are scarce

Why Extremes?

There are several methodologies to Extreme Value Theory (EVT)

The most common are:

- Block Maxima Approach - modelled with the Generalised Exreme Value Distribution
- Peaks Over Threshold Approach - modelled with the Generalised Pareto Distribution

Block Maxima Approach

The **aim** is to model and characterise the behaviour of the maximum *(minimum)* of a series of independent and identically distributed (i.i.d.) random variables, *i.e.,*

$$M_n = \max\{X_1, \ldots, X_n\}.$$

The modelling consists in

- splitting the data into m blocks of size n of sequences of observations
- a sequence of **block maxima** $M_{n,1}, \ldots, M_{n,m}$ is generated
  - usually the blocks correspond to a one-year period - **annual maxima**
- the choice of block size needs care
  - small n may lead to **biased** results
  - large n may lead to **higher** variance (m is smaller)

Generalised Extreme Value Distribution

Extremal Types Theorem

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left[\frac{M_n - b_n}{a_n} \leq z\right] \rightarrow G(z) \quad \text{as } n \rightarrow \infty, \tag{1}$$

where G is a non-degenerate distribution function, then G belongs to the Generalised Extreme Value (GEV) family of models

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2}$$

defined on the set $\left\{z : 1 + \xi\left(z - \mu\sigma\right) > 0\right\}$, where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are the location, scale and shape (or **tail index**) parameters, respectively.

The GEV distribution is often used to model **block maxima**

Generalised Extreme Value Distribution

There are 3 particular cases of the GEV distribution:

- when $\xi > 0$, we have the Fréchet distribution

$$G(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases}$$

- when $\xi < 0$, we have the Weibull distribution

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

- when $\xi \to 0$, we have the Gumbel distribution

$$G(z) = \exp\left\{-\exp\left\{-\frac{z-b}{a}\right\}\right\}, \quad z \in \mathbb{R};$$

Generalised Extreme Value Distribution

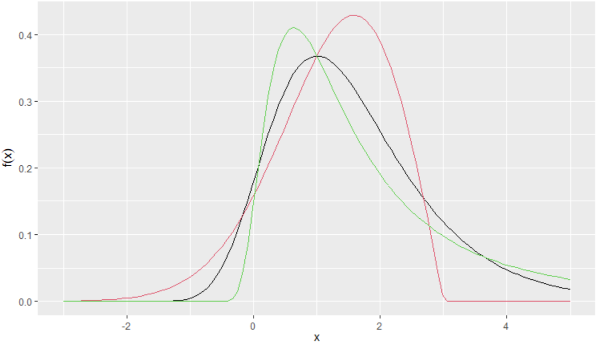

Figure 4: Fréchet (green), Weibull (red) and Gumbel (black) distributions

Return Levels

A useful, and often obtained, quantity in EVT is the return level. It is obtained by inverting (2)

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \left[-\log(1-p)\right]^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma\left[-\log(1-p)\right], & \xi = 0. \end{cases} \tag{3}$$

- z_p is exceeded by the annual maximum in any particular year with probability p - it represents the **return level** associated with the return period $\frac{1}{p}$

Return Levels

We can also obtain the **return level plot** - a plot of the level expected to be exceeded on average once in p years against the (logarithm of the) return period p

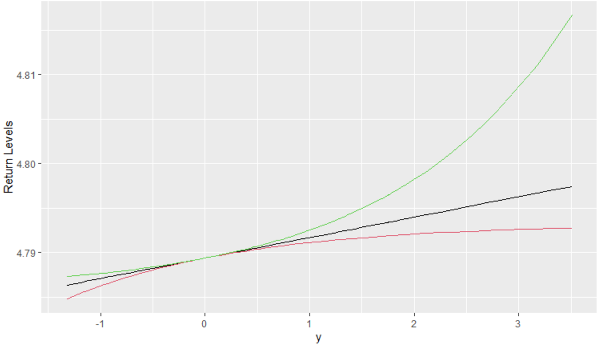

Figure 5: Return Levels for the Fréchet (green), Weibull (red) and Gumbel (black) distributions

Inference

Inference for Generalised Extreme Value Distributions is usually done by **Maximum Likelihood Estimation**

Once obtained the maximum likelihood estimators for the GEV parameters, we can sub them into the return level and obtained the **estimated return levels**

The r-largest Approach

We might wish to model the behaviour of the r**-largest order statistics** within a block instead of just the maximum

In this way we are able to estimate the GEV parameters with **more than** just the single largest observation within each block

Defining $M_n^{(i)}$ as the ith largest observation, we have the limiting joint distribution of

$$\left(\frac{M_n^{(1)} - b_n}{a_n}, \ldots, \frac{M_n^{(r)} - b_n}{a_n}\right)$$

for some choice of r.

The r-largest Approach

The likelihood to which we employ maximum likelihood estimation is then given by

$$L(\mu,\sigma,\xi) = \underbrace{\exp\left\{-\left(1+\xi\frac{M_n^{(r)}-\mu}{\sigma}\right)_+^{-1/\xi}\right\}}_{r\text{-largest observation}} \prod_{i=1}^{r} \underbrace{\frac{1}{\sigma}\left(1+\xi\frac{M_n^{(i)}-\mu}{\sigma}\right)_+^{-1/\xi-1}}_{i\text{th largest observation}} \tag{4}$$

Extentions

We have been assuming i.i.d. variables but in practice a lot of application are **non-identically** distributed. There are some possibilities to model **block maxima** and r **largest**

Possible Models:

- **Linear time trend in mean**

$$\mathrm{GEV}(\mu_0 + \mu_1 t, \sigma, \xi)$$

- **Linear time trend in scale**

$$\mathrm{GEV}(\mu, \exp\{\sigma_0 + \sigma_1 t\}, \xi)$$

References I

Coles, S. (2001).
An Introduction to Statistical Modeling of Extreme Values, volume 208 of *Springer Series in Statistics*.
Springer-Verlag, London, U.K.

Lee, C. (2021/2022).
Lecture Notes for Lancaster University MATH456/556.