Einführung in die Extremwertstatistik

Henning Rust
henning.rust@fu-berlin.de

Meteorologische Extremereignisse, SoSe 2023
Version from 24. Mai 2024, 09:07

Ronald A. Fisher as a child.
Contents

1 Introduction 3

2 Nature of the Problem 3
  2.1 Motivation 3
  2.2 An Overview 4

3 Some Basics of Extreme Value Theory 5
  3.1 Probabilistic Framework 6
  3.2 Fisher-Tippett Theorem 8
    3.2.1 Gumbel Distribution (Emil Julius Gumbel, 1891-1966) 9
    3.2.2 Fréchet Distribution (Maurice Fréchet, 1878-1973) 10
    3.2.3 Weibull Distribution (Waloddi Weibull, 1887-1979) 11
  3.3 Generalised Extreme Value Distribution 11
  3.4 Probability of Exceedances, Return-Periods and Return-Levels 13
    3.4.1 Probability of Exceedance 13
    3.4.2 Return-Level and Return-Period 14
    3.4.3 Why is the concept of return-periods dangerous? 14
  3.5 Working with Minima 14
  3.6 Relaxing assumptions 15

4 Statistical Modelling (of Extreme Values) 16
  4.1 Maximum-Likelihood Estimation 16
  4.2 Confidence Intervals 17
    4.2.1 Propagation of Errors (Delta Method) 17
    4.2.2 Profile Likelihood 18
  4.3 Model Diagnostics 18
    4.3.1 Quantile-Quantile-Plot (QQ-plot) 19
    4.3.2 Probability-probability plot (pp-plot) 20
  4.4 Relaxing assumptions 20

5 Non-Stationary Extreme Value Models 21
  5.1 General Idea 21
  5.2 Linear models 21
  5.3 Static Link-Functions 22
  5.4 Résumé 22
  5.5 Example: A Seasonal Extreme Value Model 22
    5.5.1 Amplitude and phase 23
    5.5.2 Annual exceedance probabilities and associated levels 23

6 Model Selection 23

7 What has not been covered 25
1 Introduction

Extreme value statistics (EVS) tries to answer the question:

| What size of events do we have to expect to occur at which frequency/probability? |

This question typically arises for natural catastrophes, such as earthquakes, tsunamis, forest fires, land slides, storms, flash floods, and so on. A further application is the design of infrastructure. Infrastructure should withstand events with a certain well defined (typically small) probability of occurrence. Examples are streets, bridges, dams, sewer systems, nuclear power plants, wind turbines and so on. Finally, also the financial market, particularly the insurance industry uses concepts of extreme value statistics to calculate their primes for insurance or re-insurance. There are many more examples which EVS provides a useful framework for.

There are also some examples, where EVS is not a useful concept for, see Fig. 1

Figure 1: Examples where EVS is not useful: extreme ironing (left), extreme piercing (right).

This text closely follows the book of Stuart [?][?], definitively worth reading.

2 Nature of the Problem

2.1 Motivation

Suppose, we are given a series of daily rainfall sums at the station Potsdam. This is a very nice example, since Potsdam has precipitation recordings for over 100 years. Figure 2.1 shows the time series; we can see from the plot that only a small fraction of the more than 36500 observations determine the distribution of extremes.

We are asked to estimate the probability of occurrence of extreme precipitation or, more colloquial, what is the size of the extreme precipitation event which occurs
on average once in a hundred or a thousand years. Frequently, that involves making statements about events of a size which has not been observed so far. What to do?

A histogram as in Fig. 2.1 should be a good way to go but the interesting part, the extremes between 60mm/day and about 100mm/day are represented only by very few events (see right plot in Fig. 2.1).

We could use the largest event in that series (August 8th, 1978) as an estimate for the 100-year event, probably not a bad choice. But what if somebody else had started observing 50-years earlier and stopped in 1960? Than we would not have observed this event. Maybe only smaller ones or several larger ones. If an estimate of the 100-yr event is made like this, there is obviously a strong dependence on the observation period.

Although we can never get rid of this dependence, we can reduce it by introducing extra information about the behaviour of extremes. This extra information is based on rigourous mathematical grounds and is the essence of extreme value theory, as introduced in the next sections.

2.2 An Overview

When dealing with extreme events, the notion of risk emerges. Risk is a function of the hazard probability (i.e. the probability of occurrence of an event), the exposure (i.e. what is actually exposed to this hazard) and the vulnerability (i.e. how easily is the exposed object/subjet damaged). Extreme value statistics can only inform about the hazard probability; the two other ingredients for risk are not covered here. Figure 2 embeds extreme value statistic into the context of hazard probability and the prerequisites for using and understanding the theory (basics of statistical modelling). It further shows the two main routes: the block-maxima and the peak-over-threshold approach to extreme value statistics.

Figure 2: Advance organizer of the course: embedding extreme value statistics into its aims (hazard probabilities) and prerequisites (basics of statistical modelling).

The goal is a probability distribution function for extremes which is the basis for the hazard probability.
The historical development is associated with some prominent names and can be roughly lined out as follows: Leonard Tippett (1902-1985) worked as researcher in the British cotton industry aiming to make cotton stronger. He found out that the strength of a thread depends on the strength of its weakest fibres. Together with Ronald Aymer Fisher (1890-1962), Tippett was able to prove three limiting theorems for the distribution of extremes given independent random variables (Fisher-Tippett Theorem). Emil Julius Gumbel (1891-1966) collected these results in a book *Statistics of Extremes*. Also the Gumbel-distribution appeared in this book. It was possible to generalize the results of Tippett to weakly dependent variables.

Figure 3: Sir Ronald Aymer Fisher, FRS (1890-1962), picture taken in 1913 (left) and Emil Julius Gumbel (1891-1966) (right).
A short list on the development of extreme values statistics:

**Basics** starting in the 1920ieth by Fisher and Leonard Tippett

**Augmented** in the 1940ieth by Gnedenko,

**Applied** in the 1950ieth by Gumbel and Jenkinson,

**Generalised** from the 1970ieth until today by Pickands, Balkema, de Haan, Leadbetter and many others.

Today, extreme value statistics is still a vivid field of research with an international biannual conference called EVA. Actual fields of research are, e.g., multivariate or spatial extreme value models.

### 3.1 Probabilistic Framework

**Random Variables** Given a record of observations \( x_i, i = 1, \ldots, \) e.g., daily temperatures, hourly precipitation sums, daily discharge, average or maximum wind speed, etc. We conceive these observations \( x_i \) as realisations of a series of random variables (RV) \( X_i, i = 1, \ldots \). Conceiving observations as realisations of random variables might seem awkward at first sight. However, think of a measurement with an expected value from your signal and a random measurement error. The measurement error is typically a part of the measurement process which you can not (or desire not to) fully describe with a model or physical knowledge and is considered as “random”. This does not mean that we do not know anything about its variation, we know at least some (statistical) properties, such as mean and/or variance, etc. As soon as a system exceeds a certain complexity, it is not possible to fully describe it; either because initial conditions cannot be known in advance or the system itself cannot be properly described (both true for, e.g., the atmosphere). Throwing a dice is already too complex to be described/controlled in deterministic way and it makes up the prototype-example for a random process.

**Assigning Probability to Random Variables** A continuous random variable \( X \) is a variable with all possible values being real numbers associated with a random phenomenon. More concrete, it is a mapping from an unpredictable process to the real numbers \( X : E \rightarrow \Omega \subset \mathbb{R} \). As a function, a random variable needs to be measurable and thus probabilities can be assigned to its outcomes. For a continuous random variable, this is typically done on the image space with a probability density function \( f_X(x) \geq 0 \) or a probability distribution function \( 0 \leq F_X(x) = \int_{-\infty}^{x} f_X(y) \, dy \leq 1 \), see Fig. 4 for an example of a Gaussian (or normal) random variable.

**Block Maxima Approach** Let \( X_i \) be a series of independent and identically distributed (iid) RV, all with the same PDF \( F(x) \). Consider the maxima \( M_n \) of a block (subset) of size \( n \) of this series

\[
M_n = \max\{X_1, X_2, \ldots, X_n\}. \tag{1}
\]
Figure 4: Left: probability density function for various Gaussian (normally) distributed random variables. Right: associated probability distribution function.

\( M_n \) are called the block-maxima of blocks with block size \( n \). The maxima \( M_n \) can be characterised by their probability distribution \( Pr\{M_n < x\} \) which assigns a probability to \( M_n \) of exceeding (or equivalently non-exceeding) a given value \( x \).

For applications, the aim is to specify the probability of \( M_n \) exceeding a given value.

Formally, the probability distribution of \( M_n \) can be characterised by the joint – and in the case of iid \( X_i \) also by the marginal – distribution of the \( X_i \),

\[
Pr\{M_n < x\} = Pr\{X_1 < x, X_2 < x, \ldots, X_n < x\} \tag{2}
\]

\[
= Pr\{X_1 < x\}Pr\{X_2 < x\} \ldots Pr\{X_n < x\}
\]

\[
= F(x)F(x) \ldots F(x)
\]

\[
= \{F(x)\}^n.
\]

The problem of knowing the probability distribution of the \( M_n \) would be solved if we only knew \( F(x) \). Then we could assign a probability of exceedance to every maximum \( M_n \) using only the probability distribution function \( F(x) \) of the underlying RV \( X \). Unfortunately, in a general setting, we do not have access to \( F(x) \). Thus the central problem of extreme value theory is

What is a suitable probability distribution function describing \( M_n \)?

For exactly this question in the specified univariate (iid) setting, extreme value theory (EVT) provides adequate models under quite mild assumptions, even better: it provides a family of parametric PDFs. The unknown parameters of the PDF can be estimated by classical parameter estimation strategies. The resulting PDF is then completely know and probabilities for exceedances of a given value can be specified.
3.2 Fisher-Tippett Theorem

Starting Point  Consider a series of RV $X_i$ and their block maxima

$$M_n = \max\{X_1, X_2, \ldots, X_n\}$$

for a block size $n$. If we gradually increase the block size $n$, the distribution of block maxima $M_n$ moves towards larger values and gets more peaked (smaller in width), see Fig. 5.

![Figure 5: Maxima $M_n$ for $n = 10$ (top) and $n = 100$ (bottom)](image)

Introducing Shift and Scaling  In a next step, we account for the shift and scaling using sequences $a_n > 0$ and $b_n$ to shift and scale the maxima as

$$M^*_n = \frac{M_n - b_n}{a_n}. \quad (3)$$

As a result, the distribution of the maxima does not change it’s location (due to $b_n$) and it’s width (due to $a_n$), see Fig. 6, but it approaches a particular fixed distribution (red line).

![Figure 6: Convergence of the maxima taken from a Gaussian random variable against a fixed distribution (red line).](image)

The Theorem  This astonishing behaviour is stated by the following Fisher-Tippett Theorem (FTT) or Three-Types or Extreme Value Theorem or Fisher-Tippett-Gnedenko-Theorem.

**Theorem 1** (Fisher-Tippett). If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$, such that

$$\Pr\left\{ \frac{M_n - b_n}{a_n} \leq x \right\} \rightarrow G(x), \quad \text{as} \quad n \rightarrow \infty$$

where $G$ is a non-degenerate distribution function, then $G$ belongs to one of the following families:

$$I : G(x) = \exp\left\{ - \exp\left[ - \left( \frac{x - b}{a} \right) \right] \right\} \quad -\infty < x < \infty;$$

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

$$III : G(x) = \begin{cases} \exp\left\{ - \left( \frac{x - b}{a} \right)^{a} \right\} & x < b, \\ 1 & x \geq b. \end{cases}$$
for parameters $a > 0$, $b$ and, in the case of families II and III, $\alpha > 0$.

This theorem basically states that the properly rescaled maxima $\frac{M_n - b_n}{a_n}$ converge in distribution to a variable having a distribution within one of these three families. The three are the extreme value distributions and are also known as Type I Gumbel, Type II Fréchet, and Type III Weibull. They all have a location and scale parameter $b$ and $a$ and the Types II and III have a shape parameter $\alpha$.

The remarkable point of this result is that independent of the original distribution $F$, the maxima $M_n$ converge towards one of these three distributions. No other distribution is possible.

For a given problem with given observations, this theorem dramatically reduces the space of an appropriate model; even more, it states that only these three parametric forms of the distribution are possible models for the distribution we were looking for in the very beginning. The one specifying the probability of exceedance for the maxima $M_n$.

In a sense, the Fisher-Tippett-Theorem is similar to the central limit theorem which states that averages of iid random variables are Gaussian (normally) distributed. In the case of the FTT, the operation is not average (which is a properly normalized sum) over a block of size $n$ but maximum out of a block of size $n$. The resulting distribution is not Gaussian but one of the three types mentioned above.

### 3.2.1 Gumbel Distribution (Emil Julius Gumbel, 1891-1966)

In the simplest case of $b = 0$ and $a = 1$, the Gumbel distribution reads

$$G(z) = \exp^{-\exp^{-z}}, \quad -\infty < z < \infty. \quad (4)$$

The density function for this case is

$$g(z) = \exp^{z+\exp^{-z}}, \quad -\infty < z < \infty. \quad (5)$$

For this standard Gumbel case, the distribution function and the density function are shown in Fig. 7.

**Standard Gumbel**

Figure 7: Distribution function (left) and density function (right) of the standard Gumbel distribution.

The notable characteristic of the Gumbel distribution is the exponentially decaying tail ($\sim \exp^{-z}$) in the density function towards large values $z$ (light tailed). The Gumbel distribution arises as the limiting distribution for the following starting distributions (e.g.)
- normal or Gaussian distribution,
- log-normal distribution,
- exponential distribution,
- Gamma distribution,
- Gumbel distribution.

### 3.2.2 Fréchet Distribution (Maurice Fréchet, 1878-1973)

The standard Fréchet distribution for \( b = 0 \) and \( a = 1 \) reads

\[
G(z) = \begin{cases} 
0 & z \leq 0 \\
\exp(-z^{-a}) & z > 0
\end{cases},
\]

with a shape parameter \( \alpha > 0 \). The density function is

\[
g(z) = \alpha z^{-1-\alpha} \exp^{-z^{-\alpha}}, \quad -\infty < z < \infty.
\]

Figure 8 gives the distribution and density functions

**Standard Fréchet**

![Figure 8: Distribution function (left) and density function (right) of the standard Fréchet distribution.](image)

Here, the tail behaviour (and that is the important difference between the tree types) is algebraic (or according to a power-law) \( \sim z^{-\alpha} \) (**heavy tailed**). That implies a slower decay in probability towards large values than the exponential decay of the Gumbel distribution, in other words, extreme events are more likely with respect to the non-extreme events in this distribution than it is the case for the Gumbel distribution. The Fréchet distribution arises as the limiting distribution for the following starting distributions (e.g.)

- Cauchy distribution,
- log-gamma distribution,
- Pareto distribution,
- Fréchet distribution.
3.2.3 Weibull Distribution (Waloddi Weibull, 1887-1979)

The standard Weibull distribution for $b = 0$ and $a = 1$ reads

$$G(z) = \begin{cases} 
\exp\left\{-(-z)^\alpha\right\} & z < 0 \\
1 & z \geq 0 
\end{cases}$$

again with a shape parameter $\alpha > 0$. This distribution is particularly remarkable as the upper end-point is fixed, cf. Fig. 9.

There is no slow or fast decay of probability towards large values. Beyond the upper end-point the probability for events is zero, the end-point is fixed (fixed end-point). This behaviour arises, e.g. for the

- uniform distribution,
- beta distribution,
- Weibull distribution.

Consider a simple example. The random variable $Y$ we started with is a dice with 6 sides. If we take maxima over a certain block size, these maxima will never exceed 6. The upper end-point is thus fixed.

The important distinction between the three types is their behaviour for large values: exponential decay for Gumbel, algebraic decay for Fréchet and a fixed end-point for the Weibull distribution. There are thus basically three things which could happen for large values of block-maxima: either the decay in probability is fast (exponential, Gumbel), slow (algebraic, Fréchet) or there is a fixed end-point (Weibull). Nothing else is possible. Convenient characteristics of these three distributions can be found on https://wikipedia.org.

As this is already a very remarkable result, comparable to the central limit theorem in classical statistics, it would be better if we had only one parametric form and not three.

3.3 Generalised Extreme Value Distribution

A more satisfying result for analysing extreme values arose if we could combine the three families of the FTT into one parametric form. This lead to the following variant of the FTT (Von Mises (1954) and Jenkinson (1955)):
Theorem 2. If there exist sequences of constants \( \{a_n > 0\} \) and \( \{b_n\} \), such that

\[
P_r \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \to G(z), \quad \text{as} \quad n \to \infty
\]

where \( G \) is a non-degenerate distribution function, then \( G \) is a member of the GEV family:

\[
G(z) = \begin{cases} 
\exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} & \text{if } \xi \neq 0 \\
\exp \left\{ \left( \exp \left( \frac{z - \mu}{\sigma} \right) \right) \right\} & \text{if } \xi = 0
\end{cases}
\]

(9)

defined on \( \{ z : 1 + \xi (z - \mu)/\sigma > 0 \} \), where \(-\infty < \mu < \infty\), \( \sigma > 0 \) and \(-\infty < \xi < \infty\).

This model for extremes has three parameters, a location parameter \( \mu \), a scale parameter \( \sigma \) and a shape parameter \( \xi \), cf. Figs. 10. The latter decides which FTT-family the GEV belongs to. For \( \xi > 0 \) the GEV is equivalent to the Type II or Fréchet distribution and \( \xi < 0 \) corresponds to the Type III or Weibull distribution. The Type I or Gumbel distribution arises for the shape parameter \( \xi = 0 \).

Generalised Extreme Value Distribution (GEV)

Figure 10: Generalised extreme value distribution with zero shape (Gumbel, black), positive shape (Fréchet, red) and negative shape (Weibull, green). Left: probability distribution function. Right: probability density function.

Generalised Extreme Value Distribution (GEV)

Figure 11: Parameters of the generalised extreme value distribution.

Interpretation We now interpret the latter variant of the FTT as follows: for block maxima \( M_n \) for large values of \( n \) (and not in the strict sense of \( n \to \infty \)) the GEV is very likely an adequate parametric model for describing the occurrence probabilities of extremes. Hence, the FTT suggest to use the GEV as a parametric model for block-maxima. We thus solved the problem we started out with: we searched for a PDF describing the probability of exceedance for the extremes (block maxima).

Normalisation Constants There is one more technical issue to cope with. We have to taken care about the sequences of normalizations constants \( b_n \) and \( a_n > 0 \) and, additionally, we have to estimate the location, scale and shape parameters \( \mu, \sigma \) and \( \xi \). These two problems, however, can be reduced to one. Consider the case

\[
P_r \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \approx G(z)
\]
for \( n \) being large enough, then
\[
Pr \{ M_n \leq z \} \approx G((z - b_n)/a_n) = G^*(z),
\]
and \( G^* \) is also a member of the GEV family. We can thus directly estimate the parameters of \( G^* \) as the distribution for the maxima \( M^*_n \). The distinction between \( G^* \) and \( G \) is irrelevant in practice.

### 3.4 Probability of Exceedances, Return-Periods and Return-Levels

Let’s suppose that we have managed to estimate the location, scale and shape parameters of a GEV from observations and we are now in the situation that we can describe the distribution of the block maxima with a parametric distribution (namely the GEV). What does that help for the original question: “What size of events do we have to expect to occur at which frequency?”

#### 3.4.1 Probability of Exceedance

Recall that the probability distribution function \( G(z) \) actually gives the probability for not exceeding a value of size \( z \)
\[
Pr\{Z \leq z\} = G(z).
\]
Thus the probability of exceeding a value of \( z \) is
\[
Pr\{Z > z\} = 1 - G(z),
\]
see also Fig. 12.

**Probability of exceedance**

Figure 12: Probability of exceedances and return levels exemplified with a Gumbel distribution and a 10% chance of exceedance. The red surface under the density function part is the probability of exceedance: \( Pr\{Z > z\} = 1 - G(z) = 1/10 \), the green surface is the probability of non-exceedance: \( Pr\{Z < z\} = G(z) = 1 - 1/10 \).

The expression \( 1 - G(z) \) now associates a probability of exceedance to every value \( z \), thus once we have \( G(z) \) (in a parametric form), we can obtain a probability of exceedance for every \( z \).

Typically one seeks for the level \( z \) to be exceeded with a given probability \( p \). This can be obtained from inverting the PDF, i.e. specifying the quantile function \( z_p = G^{-1}(1-p) \). As we are in possession of the parametric form of the GEV \( G(z) \) and have estimated the parameters, this inversion can be done analytically. The desired quantiles are then given by
\[
z_p = \begin{cases} 
\mu + \sigma \log(y_p) & \xi \to 0 \\
\mu + \frac{\sigma}{\xi} \left( (y_p)^{\xi} - 1 \right) & \xi \neq 0
\end{cases}
\]
with \( y_p = -1/\log(1 - p) \). This follows from inverting Eq. (10) and setting \( p = Pr\{Z > z\} \). \( z_p \) is the level of exceedance associated to the probability \( p \).
3.4.2 Return-Level and Return-Period

It is very unfortunate, but as it has turned out that many people have difficulties to develop an intuition for what probability means. As it is actually not so bad to express probabilities in terms of relative frequencies, or the so called empirical probability, it is bad – but very common in geosciences and engineering – to express the probability of exceedances in return-periods (units of time). Based on the probability \( p \) of exceeding the level \( z_p \), we can define a time \( T = 1/p \) such that \( z_T = z_p \) is exceeded on average once every \( T \) time units. If we consider annual maxima, i.e. a block-size of a year, \( z_T \) will be exceeded on average every \( T \) years (if \( G(z) \) does not change, i.e. is stationary). This gives rise to calling \( z_T \) the return-level associated to the return-period \( T \). The relation between the return-level \( z_T \) and the corresponding return-period \( T \) can be expressed in a return-level plot, Fig. 13. The scaling of the abscissa is such that the Gumbel case

![Figure 13: Return level plot for standard Gumbel, Fréchet and Weibull.](image)

shows up as straight lines in the return-level plot, the Fréchet case as a concave line and the Weibull case as convex line. This plot provides the solution to the question we entered this lecture with: “What size of events do we have to expect to occur at which frequency?”

3.4.3 Why is the concept of return-periods dangerous?

It is not necessary to express probabilities \( p \) in terms of return-periods \( T = 1/p \) with the unit of time. This originates very likely from interpreting probabilities as relative frequencies \( \frac{a}{n} \), i.e. number of events \( a \) over number of total trials \( n \). This is a very common idea in frequentist statistics and justified by the Central Limit Theorem [e.g., ?]. The dangerous step is mapping of the number of trials on the time axis we live on. This has two stupid implications:

1. Most people say “the level \( z_T \) is exceeded once every \( T \) years” and forget about “on average”. This raises wrong expectations and might lead to unfavourable decisions (“I rebuild my house, this will not come again in the next 99 years, they said”)

2. If we identify \( T \) with our time axis, this implies that we are locked into a stationary concept. This is definitively not useful in the frame of a changing climate.

A suggestion for a better wording is “\( z_p \) is exceeded on average once every \( 1/p \) times by the annual maximum value”. Although slightly more stringent, this subtle differences in words might not help to solve the communication misunderstandings mentioned above.

3.5 Working with Minima

For applications with extremely small values (minima), similar theorems hold, see [e.g., ?]
3.6 Relaxing assumptions

What happens to $M_n$ if $X_i, i = 1, \ldots$ are not independent and identically distributed?

**What if $X_i$ are not independent?** In most practical cases, the Fisher-Tippett Theorem still holds [?], maybe larger $n$ required.

**What if $X_i$ are not identically distributed?** In many situations, the Fisher-Tippett Theorem still holds, e.g. seasonal processes: here annual block sizes effectively reduce to periods where maxima occur. Those periods, where maxima never occur, do not count into the block size. To increase block sizes to larger blocks than annual, one could use biannual block sizes.
4 Statistical Modelling (of Extreme Values)

In the last sections, we stated the problem and motivated a parametric model for the description of block-maxima. What was left to do was the estimation of the parameters in the GEV, the location ($\mu$), scale ($\sigma$) and shape ($\xi$) parameters. This problem of parameter estimation from observations basically corresponds to a standard setting found in statistical modelling. There are various ways to treat this problem, e.g. with various types of moment-based estimator, but the most common and versatile way to address this problem is the maximum likelihood estimation.

4.1 Maximum-Likelihood Estimation

Suppose we have realisations $x_1, \ldots, x_N$ of a series of independent and identically distributed (iid) random variables $X_1, \ldots, X_N$ with probability density function $f(x; \theta_0)$. The parameters of the pdf are summarised in the unknown parameter vector $\theta_0$. The likelihood for a parameter vector $\theta$ given the observations is then defined as

$$L(\theta) = \prod_{i=1}^{N} f(x_i; \theta).$$

(12)

The factorisation is due to the assumption of independence. It is convenient to define the log-likelihood as

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{N} \log f(x_i; \theta).$$

(13)

The assumptions of independence and identically distribution can be relaxed with mild consequences. This is beyond the scope of this introduction. The book of ? touches slightly on this topic.

The maximum-likelihood estimator is defined as

$$\hat{\theta}_0 = \arg \max_{\theta} l(\theta),$$

(14)

it is thus the parameter vector $\theta$ which maximises the likelihood. For simple cases, such a maximisation can be done analytically by setting the derivative to zero. In general, however, one has to solve Eq. (14) numerically. For the maximum-likelihood estimator, one can show convergence towards the true value $\theta_0$ for $N \to \infty$.

**Theorem 3** (Asymptotic distribution of the MLE). *Let $x_1, \ldots, x_N$ be independent realisations from a distribution within the family $F$, and let $l(.)$ and $\theta_0$ denote respectively the log-likelihood and the maximum-likelihood estimator of the $d$-dimensional model parameter vector $\theta_0$. Then, under suitable regularity conditions, for large $N$,

$$\hat{\theta}_0 \sim \text{MVN}_d(\theta_0, I_{E}^{-1}(\theta_0)),$$

(15)

where

$$I_{E}(\theta) =
\begin{bmatrix}
  e_{1,1} & \cdots & \cdots & e_{1,d}(\theta) \\
  \vdots & \ddots & \ddots & \vdots \\
  \vdots & \ddots & \ddots & \vdots \\
  e_{d,1} & \cdots & \cdots & e_{1,d}(\theta)
\end{bmatrix}$$

(16)
with
\[
e_{i,j}(\theta) = E \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta) \right]. \tag{17}
\]

The matrix \( I_E(\theta) \) measures the expected curvature of the log-likelihood surface and is called the expected information matrix. This theorem is typically used to obtain approximate confidence intervals as shown in the following.

### 4.2 Confidence Intervals

Based on the previous theorem, we can obtain approximate confidence intervals for individual components of \( \theta_0 = (\theta_1, \ldots, \theta_d) \). If \( \psi_{i,j} \) is an arbitrary term of the inverse of \( I_E(\theta) \), then for large \( N \)
\[
\hat{\theta}_i \sim N(\theta_i, \psi_{i,i}). \tag{18}
\]

Thus, if we knew \( \psi_{i,i} \), an approximate \((1 - \alpha)\) confidence interval for \( \theta_i \) is
\[
\theta_i \pm z_{\alpha/2} \sqrt{\psi_{i,i}}, \tag{19}
\]
where \( z_{\alpha/2} \) is the \( \alpha/2 \) quantile of the standard normal distribution. Recall that for 95%-confidence intervals \( \alpha = 0.05 \), \( z_{0.05} \). As the expected information matrix \( I_E(\theta) \) is not known, a good approximation can be obtained by using the observed derivatives in Eq. (17). This way to obtain confidence intervals is based on asymptotic normality.

There are alternative methods to quantify the uncertainty of the maximum-likelihood estimator, such as the profile-likelihood method or bootstrapping. The profile-likelihood is described further below.

#### 4.2.1 Propagation of Errors (Delta Method)

As we typically need confidence intervals for the return-levels, we need a way to propagate the confidence intervals from the parameters \( \theta \) of the PDF \( G(x) \) to the return-levels \( x_T \). We thus need the two following theorems.

**Theorem 4.** If \( \hat{\theta}_0 \) is the maximum likelihood estimate (MLE) of \( \theta_0 \), and \( \phi = g(\theta) \) is a scalar function, then the MLE of \( \phi_0 \) is given by \( \hat{\phi}_0 = g(\hat{\theta}_0) \).

**Theorem 5** (Delta Method). Let \( \hat{\theta}_0 \) be the large-sample MLE of the \( d \)-dimensional parameter \( \theta_0 \) with approximate variance-covariance matrix \( V_\theta \). Then if \( \phi = g(\theta) \) is a scalar function, the MLE of \( \phi_0 = g(\theta_0) \) satisfies
\[
\hat{\phi}_0 \sim N(\phi_0, V_\phi), \tag{20}
\]
where
\[
V_\phi = \nabla \phi^T V_\theta \nabla \phi, \tag{21}
\]
with
\[
\nabla \phi = \left[ \frac{\partial \phi}{\partial \theta_1}, \ldots, \frac{\partial \phi}{\partial \theta_d} \right]^T \tag{22}
\]
evaluated at \( \theta_0 \).
Thus confidence intervals for $\hat{\phi}_0$ (e.g., the estimates for return levels $x_T$) can be derived in the same way as for $\hat{\theta}_0$, given in Eq. (19).

Other approaches, such as the profile-likelihood, are more appropriate for confidence intervals. This is beyond the scope of this text but can be found in the book of ?.

4.2.2 Profile Likelihood

An alternative to the Gaussian asymptotic confidence intervals for GEV parameters (Sec. 4.1 is the evaluation of the Profile Likelihood. This approach is typically more accurate than the asymptotic one but also somewhat more costly.

We consider a single entry $\theta_i$ of the parameter vector $\theta$. The parameter vector with all other elements $\theta_j, j \neq i$ but $\theta_i$ is denoted as $\theta_{-i}$, thus $\theta = (\theta_i, \theta_{-i})$. The log-likelihood function can then be written as $l(\theta) = l(\theta_i, \theta_{-i})$. We can now define the profile log-likelihood as

$$l_p(\theta_i) = \max_{\theta_{-i}} l(\theta_i, \theta_{-i}).$$ (23)

For a fixed value of $\theta_i$ we maximise the log-likelihood with respect to all other components. Geometrically, this is now the profile of the log-likelihood surface when looking from the $\theta_i$ axis.

This can be straightforwardly generalized to more than one component of $\theta$. Consider a $d$-dimensional parameter vector $\theta = (\theta^{(1)}, \theta^{(2)})$ with component $\theta^{(1)}$ containing $k$ components and $\theta^{(2)}$ containing the remaining $d - k$ components. The profile log-likelihood for $\theta^{(1)}$ is

$$l_p(\theta^{(1)}) = \max_{\theta^{(2)}} l(\theta^{(1)}, \theta^{(2)}).$$ (24)

One can now show that under suitable regularity conditions for large $n$

$$D_p(\theta^{(1)}) = 2\{l(\hat{\theta}_0) - l_p(\hat{\theta}^{(1)})\} \sim \chi^2_{k}$$ (25)

with $\hat{\theta}_0$ being the maximum likelihood estimator and $l(\hat{\theta}_0)$ the log-likelihood at its maximum and $l_p(\hat{\theta}^{(1)})$ the profile likelihood for $\theta^{(1)}$.

This can now be used to obtain approximate $1 - \alpha$ confidence intervals for a single component $\theta_i$, $C_\alpha = \{\theta_i : D_p(\theta_i) < c_\alpha\}$ with $c_\alpha$ being the $(1 - \alpha)$ quantile of the $\chi^2_{k}$ distribution. This approximation is usually more accurate than the Gaussian limit given above.

Confidence intervals with the profile likelihood can be also obtained for a quantile $z_p$ by reparametrizing the GEV. To this end, we replace the location parameter with

$$\mu = z_p + \frac{\sigma}{\xi} \left[1 - \{\log(1 - p)\}^{-\xi}\right].$$ (26)

4.3 Model Diagnostics

This is typically done with probability and quantile plots (pp-plots and qq-plots, respectively); also the return level plot or the density plot can be used for diagnostics, see Fig. 15. In the following the qq-plot and the pp-plot is briefly discussed, for the return-level plot, see Sec. 3.4.2.
Profile Likelihood

Figure 14: Profile Likelihood for location (top), scale (center) and shape (bottom) for the Potsdam annual maxima of daily maximum temperatures.

Model Diagnostics

Figure 15: Model diagnostics using a probability plot (top, left), a quantile plot (qq-plot, top right), a density plot (bottom, left) and a return level plot (bottom, right).

4.3.1 Quantile-Quantile-Plot (QQ-plot)

A quantile-quantile or QQ-plots is a visual model diagnostics tool and compares a set of observations to a given distribution function (model), or, alternatively, a second set of observations. It is a visual aid to see whether two distributions (an empirical and a model, or two empirical) are similar.

Consider the empirical cumulative distribution function (ecdf). The observations $x_j, j = 1, \ldots, N$ are sorted according to their size

$$x_{(1)} < x_{(2)} < \ldots < x_{(i)} < x_{(i+1)} < \ldots < x_{(N)}$$

and to every observation is associated a probability of non-exceedance according to the empirical cumulative distribution function

$$\text{ecdf}_1(x_{(i)}) = i/N = \hat{p}_i. \quad (28)$$

This is also a possible estimator for the distribution function as a whole. Other estimators with different properties do also exist, e.g.

$$\text{ecdf}_2(x_{(i)}) = (i - 0.5)/N = \hat{p}_i. \quad (29)$$

The second estimator leaves space at both ends of the distribution, for small and for large values, such that neither 0 or 1 can be attained, which makes sense in many cases.

Now use the quantile function $F^{-1}$ of a given distribution function $F$ and calculate for all $\hat{p}_i$ the associated quantile $q_i = F^{-1}(\hat{p}_i)$. $F^{-1}(\hat{p}_i)$ is thus the “theoretical quantile” associated with the estimated probability $\hat{p}_i$ of $x_i$. What basically happens here is that an non-parametric (empirical) operation (ecdf) is inverted with an inverse function based on a parametric model: $F^{-1}_{\text{model}}(F_{\text{empirical}}(x))$. If both fit perfectly together, this results in the identity $F^{-1}_{\text{model}}(F_{\text{empirical}}(x)) = I(x)$. Thus this plots highlights deviations between model and observation. If plotting the tupel

$$(F^{-1}(\hat{p}_i), x_i) \quad (30)$$

leads roughly to a straight line, Fig. 15, model and data fit together. To what extent a variation from the exact straight line is tolerable is typically assessed by bootstrapping. The R-library evd does that by default, the R-library ismev unfortunately not.
4.3.2 Probability-probability plot (pp-plot)

The pp-plot is used to compare two distribution functions by plotting them against each other for the same quantile, i.e. plotting the tuple \((F_{model}(x_i), F_{empirical}(x_i))\). Again, a straight line indicates a good fit, see Fig. 15.

4.4 Relaxing assumptions

For maximum-likelihood estimation of GEV parameters, we assume \(M_n\) to be independent and identically distributed.

**What if \(M_n\) are not independent?** The MLE theorem does not hold. This is not problematic in many cases as the point estimates remain valid, however, the inverse of the expected Fisher-Information matrix is not a valid estimate of the estimators uncertainty anymore. Confidence intervals need to be obtained in different ways. One could either include the dependence as a model in the estimation procedure (max-stable processes) or obtain estimates for confidence intervals using adequate bootstrap approaches.

**What if \(M_n\) are not identically distributed?** If the change in distribution can be explicitly included in the model, a non-stationary GEV can be estimated, see Sect.5.
5 Non-Stationary Extreme Value Models

5.1 General Idea

At the beginning, we assumed that the observations can be conceived as realisations of a sequence of random variables which were identically distributed. That means in particular that besides the random effect, no changes were possible. That is a very strict assumption and to be relaxed now. We now allow the block maxima $Z$ to be characterised by a GEV with a location parameter $\mu(t)$ varying in time, i.e.

$$Z_t \sim \text{GEV}(\mu(t), \sigma, \xi). \quad (31)$$

We do, however, have to make an assumption for the variation in time.

5.2 Linear models

The most straight-forward assumption (which can, in fact, be very general) is a linear model for $\mu(t)$. The simplest linear model has a constant term and one further covariate. This can be, e.g. time

$$\mu(t) = \beta_0 + \beta_1 t, \quad (32)$$

with $\beta_0$ being a constant offset and $\beta_1$ the rate of change in time, e.g. for annual maxima the annual rate of change, cf. Fig. 16.

Figure 16: GEV with a linear trend in time in the location parameter

A more complex linear model includes more than one factor of influence. Here, this could be a quadratic component in time, see Fig. 17.

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2, \quad (33)$$

Figure 17: GEV with a quadratic trend in time in the location parameter

Or a change-point model (Fig. 18),

$$\mu(t) = \begin{cases} 
\mu_0 & \text{for } t \leq t_0 \\
\mu_1 & \text{for } t > t_0 
\end{cases} \quad (34)$$

Figure 18: GEV with a change point in time in the location parameter

The dependence on time is very intuitive but if you think about it, a time dependence can frequently expressed indirectly, by a dependence on another variable, a so-called covariate (as it varies with (co) the variable in question). In atmospheric science, the El Niño phenomenon is a popular covariate, it can be included in a model in the following way

$$\mu(t) = \beta_0 + \beta_1 \text{SOI}(t). \quad (35)$$
5.3 Static Link-Functions

Besides these simple linear models, also constrains can be included easily. For example the scale parameter needs to be positive and a time dependent scale parameter can be written as

\[ \sigma(t) = \exp(\beta_0 + \beta_1 t). \]  

(36)

Here, the exponential function ensures that \( \sigma(t) > 0 \). Such functions can be included in the model as well. To maintain the linear structure of the predictor, functions \( g(\cdot) \) are formally used to transform the parameters, e.g.

\[ \ln(\sigma(t)) = \beta_0 + \beta_1 t \]

(37)

and these functions \( g(\cdot) \) are called link functions. The corresponding function \( g^{-1}(\cdot) \) (e.g. the \( g^{-1}(\cdot) = \exp(\cdot) \) in Eq. (36)) is called inverse link function.

5.4 Résumé

In principle, these linear models can be used to model a time dependence of the shape parameter as well. However, as this parameter is very difficult to estimate, it is usually not recommended to try doing so.

Different hypotheses for the linear models can be challenged with classical model selection approaches.

The whole concept presented in this chapter is similar to generalised linear models \([?]\), in fact, it is the concept of vector generalised linear models \([?]?)\). For other than linear models (namely additive model), be referred to [e.g., ??].

5.5 Example: A Seasonal Extreme Value Model

For daily precipitation sums in the mid-latitudes, it turns out that a monthly block-size is sufficient for a good approximation of the resulting maxima with a GEV. Hence, we can capture the seasonal cycle in the GEV characterising the monthly maxima using harmonic functions. This yields a set of linear models for the location \( \mu(t) \) (Eq. (38)) and scale parameter \( \sigma(t) \) (Eq. (39)), while the shape parameter \( \xi \) (Eq. (40)) is kept constant:

\[ \mu_i = \beta_{\mu,0} + \sum_{k=1}^{K} (\beta_{\mu,\sin,k} \sin(k \omega c_i) + \beta_{\mu,\cos,k} \cos(k \omega c_i)) \]  

(38)

with \( \omega = 2\pi / 365.25 \), \( c_i \) centre of \( i \)-th month,

\[ \sigma_i = \beta_{\sigma,0} + \sum_{l=1}^{L} (\beta_{\sigma,\sin,l} \sin(l \omega c_i) + \beta_{\sigma,\cos,l} \cos(l \omega c_i)) \]  

(39)

and

\[ \xi_i = \xi_0. \]  

(40)

More than the first order of the Fourier series can be used and the orders \( L \) and \( K \) can determined with model selection strategies (here cross-validation). For most stations in Germany, the first order in \( \mu_i \) and the second order for \( \sigma_i \) are sufficient. An example is given in Fig. 19.
Figure 19: Example for a seasonal extreme value model for monthly maxima of daily precipitation sums. The box plots represent the monthly maxima of the observations, the solid lines the 0.25, 0.5, 0.75 and 0.99 quantile of the seasonal GEV. The latter quantile is the 100-year return level.

5.5.1 Amplitude and phase

One can now talk about amplitudes of return levels, e.g. the amplitude of the 100-yr return level for 1200 station in Germany is depicted in Fig. 20. The analysis is based on a seasonal model with first order harmonic in the location parameter and second order in the scale parameter, the shape parameter is constant for all stations.

Amplitude and maximum of the 100-yr return-level

Figure 20: Amplitude (left) of the 100-yr return-level and month of maximum 100-yr return-level for 1200 stations in Germany.

5.5.2 Annual exceedance probabilities and associated levels

From the monthly resolved distribution functions, we can obtain levels \( z_p \) which are associated with an annual probability of exceedance \( p \). The annual non-exceedance probability is

\[
\Pr\{M_n \leq z\} = \Pi_{\text{dec} \text{mon}=\text{jan}}^\text{dec} G_{\text{mon}}(z) = 1 - p.
\]

(41)

Fixing now an annual exceedance probability \( p \), we can solve numerically for \( z_p \)

\[
\Pi_{\text{dec} \text{mon}=\text{jan}}^\text{dec} G_{\text{mon}}(z) = 1 - p.
\]

(42)

Other examples for seasonal extreme value models for precipitation are [e.g. ???????].

6 Model Selection

If we need to decide between various extreme value models, we have different options: a classical hypothesis testing setting based on the deviance (or likelihood-ratio) test, or, alternatively, an approach based on information theory, using information criteria, such as the Akaike Information Criterion (AIC), the Bayesian-Schwarz Information Criterion (BIC), or similar ones.

Here, we have a look at the deviance test. The deviance is defined as

\[
D = 2\{l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\}
\]

(43)
with $l_1(\mathcal{M}_1)$ being the maximized log-likelihood for a model $\mathcal{M}_1$ and $l_0(\mathcal{M}_0)$ the maximised log-likelihood for another model $\mathcal{M}_0$. The latter model must be included in the former one, one says $\mathcal{M}_0$ is nested\(^1\) into $\mathcal{M}_1$. What that means in particular can be seen by dividing the parameter vector $\theta$ of $\mathcal{M}_1$ into two subsets $\theta = (\theta^{(1)}, \theta^{(2)})$. Now Model $\mathcal{M}_0$ has basically the same structure as $\mathcal{M}_1$, however, the first $k$ components of $\theta$, namely $\theta^{(1)}$ are constrained to zero. For example, consider a linear trend in time in the location parameter

$$\mu = \mu_0 + \mu_1 t \quad (44)$$

and scale $\sigma$ and shape $\xi$ remain constant. Model $\mathcal{M}_1$ models this trend and thus has 4 parameters:

$$\theta = (\theta^{(1)}, \mu_0, \sigma, \xi) = (\theta^{(1)}, \theta^{(2)}) \quad (45)$$

Another model $\mathcal{M}_0$ might be also suitable, it does, however, not model the linear trend in time for the location parameter, thus $\mu$, $\sigma$, and $\xi$ are constants. Model $\mathcal{M}_0$ has 3 parameters which can be written in the same way as Eq. (46) setting $\theta^{(1)}$ to zero

$$\theta = (0, \theta^{(2)}) = (\theta^{(1)}, \theta^{(2)}). \quad (46)$$

We thus have two models

$$\mathcal{M}_1 : \theta = (\mu_1, \mu_0, \sigma, \xi)$$
$$\mathcal{M}_0 : \theta = (0, \mu_0, \sigma, \xi) \quad (47)$$

with $\mathcal{M}_0$ nested in $\mathcal{M}_1$. The dimension of the sub-component $\theta^{(1)}$ of $\theta$ is $k = 1$, the difference in the number of parameters of the model.

For a test, we set up the null-hypothesis $H_0$: “The smaller model, $\mathcal{M}_0$, is a valid simplification of the larger model, $\mathcal{M}_1$.” We want to test this hypothesis at a level of significance $\alpha$ using the deviance (Eq. (43)) as the test statistic.

**Theorem 6** (Deviance test). Suppose $\mathcal{M}_0$ with parameter $\theta^{(2)}$ is a sub-model of $\mathcal{M}_1$ with parameter $\theta = (\theta^{(1)}, \theta^{(2)})$ under the constraint that the $k$-dimensional sub-vector $\theta^{(1)} = 0$. Let $l_0(\mathcal{M}_0)$ and $l_1(\mathcal{M}_1)$ be the maximised values of the log-likelihood for models $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. A test for validity of $\mathcal{M}_0$ relative to $\mathcal{M}_1$ at the level of significance $\alpha$: reject $\mathcal{M}_0$ in favour of $\mathcal{M}_1$ if $D = 2\{l_1(\mathcal{M}_1) - l_0(\mathcal{M}_0)\} > c_\alpha$, where $c_\alpha$ is the $(1 - \alpha)$ quantile of the $\chi^2_k$ distribution.

---

\(^1\)This notion of nested is very different from the one used when a regional model is forced by a global climate model.
7 What has not been covered

- confidence intervals with likelihood profiling, see [?]
- $r$-largest order statistics, see [?]
- dependent data, see [?]
- point-process models, see [?]
- multivariate extremes
  - Copulas
  - max-stable processes

For further reading, I suggest to start with the book of Stuart Coles [?] and the references therein [e.g., ??]. Further interesting articles include [??????].