An introduction to extreme value theory

Thomas Mikosch

(University of Copenhagen)

(Wroclaw, 13 – 23.01.2014)
INTRODUCTION TO EXTREME VALUE THEORY

Fall

T. Mikosch
## Contents

1  What is risk and how does one model it? ................................................. 1

2  Heavy-tailed distributions  

   2.1 How can we detect the heaviness of a tail? .................................... 5  
   2.1.1 A check for the existence of moments .................................... 5  
   2.1.2 Quantiles and the QQ-plot ................................................... 7  
   2.2 The ruin probability of an insurance business ............................... 10  
   2.2.1 Standard models for the total claim amount of an insurance portfolio  
   2.2.2 The risk process and the probability of ruin ............................. 12  
   2.2.3 The ruin probability under a small claim condition .................... 14  
   2.2.4 Slowly and regularly varying functions and random variables ........ 18  
   2.2.5 Aggregation of regularly varying risks ................................... 19  
   2.2.6 Subexponential distributions ............................................... 21  
   2.2.7 The Cramér-Lundberg theorem for subexponential claims ............ 23  

3  Asymptotic theory for maxima .............................................................. 27  

   3.1 Limit probabilities for maxima .................................................. 27  
   3.2 Weak convergence of maxima under affine transformations .............. 32  
   3.2.1 Appendix: Convergence to types ........................................... 37  
   3.3 Maximum domains of attraction and norming constants ................... 39  
   3.3.1 The maximum domain of attraction of the Fréchet distribution  
   3.3.2 The maximum domain of attraction of the Weibull distribution  
   3.3.3 The maximum domain of attraction of the Gumbel distribution ....... 43  
   3.4 The generalized extreme value distribution and the generalized Pareto distribution ................................................................. 55  

4  Statistical analysis of extremes ............................................................ 64  

   4.1 Some examples ............................................................................. 64  
   4.2 Exploratory data analysis for extremes ....................................... 70  
   4.2.1 Quantile-quantile plots ....................................................... 70  
   4.2.2 The mean excess function ..................................................... 71  
   4.2.3 The return period ............................................................... 77  
   4.2.4 Records as an exploratory tool .............................................. 80  
   4.3 Parameter estimation for the generalized extreme value distribution  
   4.3.1 Maximum likelihood estimation ............................................. 83  
   4.3.2 Some theory about order statistics ....................................... 86  
   4.3.3 An extension to upper order statistics ................................... 93  
   4.3.4 Method of probability-weighted moments .............................. 93  
   4.3.5 Tail and quantile estimation, a first go .................................. 95  
   4.4 Estimating under maximum domain of attraction conditions .......... 96  
   4.4.1 Introduction ......................................................................... 96  
   4.4.2 Estimating the shape parameter $\xi$ ..................................... 98  
   4.4.3 Estimating the norming constants ....................................... 115  
   4.4.4 Tail and quantile estimation ............................................... 118  
   4.5 Fitting excesses over a threshold ................................................. 121  
   4.5.1 Fitting the GPD ................................................................. 121
4.5.2 An application to real data ........................................ 126

5 Dependence and extremes .............................................. 137
  5.1 The extremes of a stationary sequence ............................ 137
  5.2 The extremal index .................................................... 145
    5.2.1 Definition and elementary properties .......................... 145
    5.2.2 Interpretation and estimation of the extremal index .......... 148
    5.2.3 Estimating the extremal index from data ..................... 152
1 What is risk and how does one model it?

Men have always been aware of risks. However, only over the last 300 years the scientific understanding of risks has developed. The major reason for this more recent development was the foundation of that science which gives the firm basis for risk measurement: probability theory. Only since the notion of probability as a quantitative measure of uncertainty existed one could talk about risks in a scientific sense. Only as late as in the 18th century people became aware of the law of large numbers which tells us that averages of homogeneous data tend to fluctuate around a fixed number, the mean or expectation of the underlying distribution of the data, and the central limit theorem, another deep mathematical tool, describes how large the fluctuations of the data around the expectation can be. These two mathematical tools were discovered in order to understand gambling risks. They were far from “useful practical” applications. The insurance business and the risk management of banks in the 18th and 19th century were merely based on intuition and experience, not on quantitative measures of risks. This is not surprising since modern probability theory started developing only in the first half of the 20th century. In the 1930s Kolmogorov’s mathematical axioms summarized this development and turned probability theory from a collection of vague statements into hard core science. At the same time, due to efforts of outstanding people such as Ronald Fisher, statistics turned from a murky topic into an accepted part of human knowledge. Its basis is probability theory and its aim is to fit probabilistic models to real-life data.

Probabilistic and statistical tools for describing financial and insurance risks have been used only for some decades. In the first half of the 20th century actuaries did not know anything about probability theory. Only a few advanced people (among them were the Swedes Filip Lundberg and Harald Cramer) proposed probabilistic models to describe the total claim amount of an insurance business and to characterize the probability of ruin of an insurance portfolio. Their results laid the foundations to modern actuarial science which is based on probability theory and uses methods which have been developed in applied probability branches such as Markov chains, renewal or queuing theory. In finance, probabilistic models have been used rather late. Although the modeling of financial time series has some tradition in econometrics, only over the last 25 years probability theory has had a very successful period in finance. This is closely related to the enormous growth of the financial derivative market and the call for pricing of derivatives such as options, futures, etc. By now, stochastic calculus and stochastic control are basic building blocks for the pricing of derivatives and the hedging of portfolios. Over the last few years, however, it has been realized that it is very dangerous to believe in a mathematical formula for the price of a derivative. Such a formula is based on a certain mathematical model, i.e., on a mathematical belief, which in real life is not always appropriate, and therefore it may lead to large financial losses. Examples of such behavior have been observed several times in financial history, so on the remarkable Black Monday in October 1987 when the New York Stock Exchange crashed and billions of U.S. Dollars were lost within a couple of hours. A more recent story of failure is the Long Term Capital Management (LTCM) debacle. LTCM was a hedge fund founded by, among others, Scholes and Merton who, together with Black, had in 1973 given a theoretical justification for the celebrated Black-Scholes option pricing formula. The latter has been considered as the basis for modern derivative pricing, and Scholes and Merton were awarded the Nobel prize for economics in 1997. However, LTCM crashed quite badly in October 1998, despite (or because of) using this formula. The LTCM crash caused a 13.7% loss of the U.S. Dollar against the Japanese Yen within one week. Newsweek (19 October, 1998) asked: “The buck is bruised. So is another big hedge fund. What’s going on?” By now, the derivative business has exceeded an annual volume of more than $US 15 trillion per year. Thus, because of its enormous volume, it has become one of the biggest financial risks in modern
This danger was realized quite early. In 1992, the Basel Committee of the Bank for International Settlements (representing 27 European members plus the U.S., Canada, Japan, Australia, and South Africa) presented proposals to estimate market risk and to define the resulting capital requirements to be implemented in the banking sector. The European Union (EEC 93/6) approved a directive, effective January 1996, that mandates banks and investment firms to set capital aside to cover market risks. In the U.S., the Securities and Exchange Commission fulfills a similar regulatory function. Measuring and estimating risks in its various forms has become a major challenge to probability theory and statistics. Companies such as RiskMetrics have specialized in advising the financial industry how to measure and estimate risks, and government regulators control financial institutions as to whether they satisfy certain risk standards. Using historical data, the financial institutions have to show that they have kept enough reserves to cover possible losses above a certain threshold. This threshold is commonly called Value at Risk (VaR) and it is usually defined as the 5% or 1% quantile of the distribution underlying the returns in a financial portfolio; see Figure 1.1 for an illustration. The determination of that quantile is a delicate statistical problem, and we will learn how extreme value theory can help to solve it.

Classical financial time series analysis and insurance mathematics has been mostly dealing with “small risks”. The normal, exponential or gamma distributions for claims sizes are typical representatives of this kind of approach. For these distributions, an event outside a $3\sigma$ bound around the expectation is very unlikely; it is practically impossible. These distributions have been chosen because they are convenient from a mathematical point of view, i.e., the mathematics becomes easier. However, these distributions do not really fit real-life data and they underestimate the risk represented by the data. If there were any very large or very small values in a sample, they would be considered as atypical (outliers say) and removed from the sample or they had to be treated by separate means.

In contrast to the latter approach, we will mainly deal with “large” risks, i.e., those risks which are closely related to the very small values (such as the extreme negative financial returns = extreme losses) or the very large values in a sample (such as large claims in an insurance portfolio or large gains in a financial context). This means we will deal with the extremes in such a sample: including the maximum, the minimum, and the upper or lower order statistics of a sample. We will learn how to determine the distributions of those quantities, about their limiting behavior and how to approximate their probability distributions by statistical means.

The question as to what an “extremal event” or “large risk” means clearly depends on the point of view of the person who asks this question. However, whatever definition one takes, most will
agree that Table 1.2, taken from Sigma [2] contains extremal events in the context of insurance. When looked upon as single events, each of them exhibits some common features.

<table>
<thead>
<tr>
<th>Losses</th>
<th>Date</th>
<th>Event</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 000</td>
<td>08/24/92</td>
<td>Hurricane “Andrew”</td>
<td>USA</td>
</tr>
<tr>
<td>11 838</td>
<td>01/17/94</td>
<td>Northridge earthquake in California</td>
<td>USA</td>
</tr>
<tr>
<td>5 724</td>
<td>09/27/91</td>
<td>Tornado “Mireille”</td>
<td>Japan</td>
</tr>
<tr>
<td>4 931</td>
<td>01/25/90</td>
<td>Winterstorm “Daria”</td>
<td>Europe</td>
</tr>
<tr>
<td>4 749</td>
<td>09/15/89</td>
<td>Hurricane “Hugo”</td>
<td>P. Rico</td>
</tr>
<tr>
<td>4 528</td>
<td>10/17/89</td>
<td>Loma Prieta earthquake</td>
<td>USA</td>
</tr>
<tr>
<td>3 427</td>
<td>02/26/90</td>
<td>Winter storm “Vivian”</td>
<td>Europe</td>
</tr>
<tr>
<td>2 373</td>
<td>07/06/88</td>
<td>Explosion on “Piper Alpha” offshore oil rig</td>
<td>UK</td>
</tr>
<tr>
<td>2 282</td>
<td>01/17/95</td>
<td>Hanshin earthquake in Kobe</td>
<td>Japan</td>
</tr>
<tr>
<td>1 938</td>
<td>10/04/95</td>
<td>Hurricane “Opal”</td>
<td>USA</td>
</tr>
<tr>
<td>1 700</td>
<td>03/10/93</td>
<td>Blizzard over eastern coast</td>
<td>USA</td>
</tr>
<tr>
<td>1 600</td>
<td>09/11/92</td>
<td>Hurricane “Iniki”</td>
<td>USA</td>
</tr>
<tr>
<td>1 500</td>
<td>10/23/89</td>
<td>Explosion at Philips Petroleum</td>
<td>USA</td>
</tr>
<tr>
<td>1 453</td>
<td>09/03/79</td>
<td>Tornado “Frederic”</td>
<td>USA</td>
</tr>
<tr>
<td>1 422</td>
<td>09/18/74</td>
<td>Tornado “Fifi”</td>
<td>Honduras</td>
</tr>
<tr>
<td>1 320</td>
<td>09/12/88</td>
<td>Hurricane “Gilbert”</td>
<td>Jamaica</td>
</tr>
<tr>
<td>1 238</td>
<td>12/17/83</td>
<td>Snowstorms, frost</td>
<td>USA</td>
</tr>
<tr>
<td>1 236</td>
<td>10/20/91</td>
<td>Forest fire which spread to urban area</td>
<td>USA</td>
</tr>
<tr>
<td>1 224</td>
<td>04/02/74</td>
<td>Tornadoes in 14 states</td>
<td>USA</td>
</tr>
<tr>
<td>1 172</td>
<td>08/04/70</td>
<td>Tornado “Celina”</td>
<td>USA</td>
</tr>
<tr>
<td>1 168</td>
<td>04/25/73</td>
<td>Flooding caused by Mississippi in Midwest</td>
<td>USA</td>
</tr>
<tr>
<td>1 048</td>
<td>05/05/95</td>
<td>Wind, hail and floods</td>
<td>USA</td>
</tr>
<tr>
<td>1 005</td>
<td>01/02/76</td>
<td>Storms over northwestern Europe</td>
<td>Europe</td>
</tr>
<tr>
<td>950</td>
<td>08/17/83</td>
<td>Hurricane “Alicia”</td>
<td>USA</td>
</tr>
<tr>
<td>923</td>
<td>01/21/95</td>
<td>Storms and flooding in northern Europe</td>
<td>Europe</td>
</tr>
<tr>
<td>923</td>
<td>10/26/93</td>
<td>Forest fire which spread to urban area</td>
<td>USA</td>
</tr>
<tr>
<td>894</td>
<td>02/03/90</td>
<td>Tornado “Herta”</td>
<td>Europe</td>
</tr>
<tr>
<td>870</td>
<td>09/03/93</td>
<td>Typhoon “Yancy”</td>
<td>Japan</td>
</tr>
<tr>
<td>865</td>
<td>08/18/91</td>
<td>Hurricane “Bob”</td>
<td>USA</td>
</tr>
<tr>
<td>851</td>
<td>02/16/80</td>
<td>Floods in California and Arizona</td>
<td>USA</td>
</tr>
</tbody>
</table>

Table 1.2 The 30 most costly insurance losses 1970–1995. Losses are in million $US at 1992 prices. For a precise definition of the notion of catastrophic claim in this context see Sigma [2].

- Their (financial) impact on the (re)insurance industry is considerable. As stated in Sigma [2], at $US 150 billion, the total estimated losses in 1995 amounted to ten times the cost of insured losses – an exceptionally high amount, more than half of which was accounted for by the Kobe earthquake. Natural catastrophes alone caused insured losses of $US 12.4 billion, more than half of which were accounted for by four single disasters costing some billion dollars each; the Kobe earthquake, hurricane “Opal”, a hailstorm in Texas and winter storms combined with floods in Northern Europe. Natural catastrophes also claimed 20,000 of the 28,000 fatalities in the year of the report.
• They are difficult to predict a long time ahead. It should be noted that 28 of the insurance losses reported in Table 1.2 are due to natural events and only 2 are caused by man-made disasters.

• If looked at within the larger context of all insurance claims, they are rare events.

Extremal events in insurance and finance have (from a mathematical point of view) the advantage that they are mostly quantifiable in units of money. However most such events have a non-quantifiable component which more and more economists are trying to take into account. Going back to the data presented in Table 1.2, extremal events may clearly correspond to individual (or indeed grouped) claims which by far exceed the capacity of a single insurance company; the insurance world’s reaction to this problem is the creation of a reinsurance market. One does not however have to go to this grand scale. Even looking at standard claim data within a given company one is typically confronted with statements like “In this portfolio, 20% of the claims are responsible for more than 80% of the total portfolio claim amount”. This is an extremal event statement as we shall discuss more in detail later on.

Comments

Bernstein [1] gives a very nice historical account of “The Remarkable Story of Risk”. He shows how probability theory and statistics have enabled men to take uncertain events into financial and actuarial calculations and to hedge against major financial losses.

References


2 Heavy-tailed distributions

In this section we will consider some realistic distributions for modeling insurance claims and financial returns. These distributions are heavy-tailed compared to the normal or the exponential distributions. Recall that for any random variable $X$ with distribution function $F$,

$$F(x) = 1 - F(x) = P(X > x) \quad \text{and} \quad F(-x) = P(X \leq -x), \quad \text{for } x > 0,$$

are the right and left tail of the distribution of $X$. If

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-0.5y^2}}{\sqrt{2\pi}} \, dy, \quad x \in \mathbb{R},$$

denotes the distribution function of the standard normal distribution and

$$\lim_{x \to \infty} \frac{F(x)}{\Phi(x)} = \infty,$$

the right tail of $X$ is heavier than the right tail of the standard normal distribution. Similarly, if $Y$ has a standard exponential distribution with $P(Y > x) = e^{-x}$ for $x > 0$ and

$$\lim_{x \to \infty} \frac{F(x)}{P(Y > x)} = \infty,$$
the right tail of $X$ is heavier than the right tail of $Y$.

Now, we could proceed with infinitely many other distributions and compare their tails. However, it is more reasonable to search for a “natural” class of heavy-tailed distributions in the context of finance and insurance, and this class of functions has been identified as the class of subexponential distributions. In what follows, we want to give some motivation why these distributions are “natural heavy-tailed” ones. For this reason we make an excursion to risk theory, i.e., to that part of insurance mathematics that studies the event that an insurance business is subject to ruin or bankruptcy. This is a classical insurance problem and has been dealt with since the 1930s, in particular by Harald Cramér, one of the outstanding probabilists and statisticians of the 20th century, and his Stockholm school.

Before we come to the more theoretical part we look at some simple empirical tools for detecting how heavy the tails of a distribution are.

2.1 How can we detect the heaviness of a tail?

2.1.1 A check for the existence of moments

![Figure 2.1](image1)

**Figure 2.1** Left: Danish fire insurance data. Right: The plot of $T_1(p)$ for various values of $p$. The plot suggests that the 2nd moment of the data might be infinite.

![Figure 2.2](image2)

**Figure 2.2** Left: Log-returns of the S&P500 index. Right: The plot of $T_1(p)$ for various values of $p$. The plot suggests that the 4th moment of the data might be infinite.

One way to characterize a heavy-tailed distribution is to assume that for a random variable $X$ with this distribution

$$E|X|^p < \infty \quad \text{for all } p < p_0 \quad \text{and} \quad E|X|^p = \infty \quad \text{for } p > p_0.$$
In this generality, it is hard to use statistics for detecting the borderline value \( p_0 \). Indeed, the class of distributions with property (2.1) is too large do make any reasonable inference on it. However, we can use a heuristic tool which is based on the strong law of large numbers.

Recall that a sequence \( (X_i) \) of random variables is said to obey the strong law of large numbers if the relation

\[
X_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} S_n \xrightarrow{a.s.} c
\]

holds for some constant \( c \). If \( (X_i) \) is iid, Kolmogorov’s strong law of large numbers tells us that (2.2) holds if and only if the expectation of \( |X_1| \) is finite, i.e., \( E|X_1| < \infty \), and then \( c = EX_1 \). Thus if \( E|X_1| = \infty \) we may expect that a plot of the sample mean as a function of \( n \) behaves very irregularly.

A consequence of the strong law of large numbers is that

\[
\frac{1}{n} M_n = \frac{1}{n} \max_{i=1,...,n} X_i \xrightarrow{a.s.} 0,
\]

provided that \( E|X_1| < \infty \). Indeed, by the strong law of large numbers (2.2),

\[
n^{-1} |X_n| = n^{-1} |S_n - S_{n-1}| \xrightarrow{a.s.} |EX_1 - EX_1| = 0.
\]

Then

\[
n^{-1} |M_n| \leq n^{-1} \max_{i=1,...,n} X_i + \max_{i=n_0+1,...,n} i^{-1} |X_i|.
\]

By virtue of (2.4) the right-hand side can be made arbitrarily small as \( n \to \infty \) if we choose \( n_0 \) fixed but arbitrarily large. This implies (2.3).

For fixed \( p > 0 \) and a sample \( X_1, \ldots, X_n \) of iid random variables consider the quantities

\[
T_i(p) = \frac{\max_{j=1,...,i} |X_j|^p}{|X_1|^p + \cdots + |X_i|^p}, \quad i = 1, \ldots, n.
\]

Notice that by the strong law of large numbers and by (2.3),

\[
T_n(p) = \frac{n^{-1} \max_{j=1,...,n} |X_j|^p}{n^{-1} ||X_1|^p + \cdots + |X_n|^p} \xrightarrow{a.s.} 0, \quad \frac{0}{E|X_1|^p} = 0,
\]

provided that \( E|X_1|^p < \infty \). Moreover, one can show that \( T_n(p) \xrightarrow{a.s.} 0 \) if and only if \( E|X_1|^p < \infty \), and (2.5) remains valid for wide classes of dependent (stationary ergodic) sequences \( (X_i) \) which fact is crucial when one deals with data from finance.

Thus it is reasonable to consider a plot of \( T_i(p) \) against \( i \) and to check whether the so defined graph is close to zero for large \( i \). This method works sufficiently well, as the illustrations in Figures 2.1 and 2.2 show. But this graphical method cannot give more than an indication as to what the limiting value \( p_0 \) might be. As a matter of fact, one needs long samples in order to get any idea about \( p_0 \), and it is quite impossible to distinguish between a value \( p_0 \) and \( p_0 \pm \epsilon \) for small \( \epsilon \).

In any case: we see that financial data and insurance claim sizes are not normally or exponentially distributed and that their distributional tails are much heavier than those of these standard distributions.
2.1.2 Quantiles and the QQ-plot

Another simple means to detect the heaviness of a tail is a quantile-quantile plot, or for short QQ-plot. Quantiles correspond to the “inverse” of a distribution function, which is not always well-defined (distribution functions are not necessarily strictly increasing).

In the following definition we focus on a left-continuous version.

**Definition 2.4** (Generalised inverse of a monotone function)

Suppose $h$ is a non-decreasing function on $\mathbb{R}$. The generalised inverse of $h$ is defined as

$$ h^-(t) = \inf \{ x \in \mathbb{R} : h(x) \geq t \}. $$

(We use the convention that the infimum of an empty set is $\infty$.)

**Definition 2.5** (Quantile function)

The generalised inverse of the distribution function $F$

$$ F^-(t) = \inf \{ x \in \mathbb{R} : F(x) \geq t \}, \quad 0 < t < 1, $$

is called the quantile function of the distribution function $F$. The quantity $x_t = F^-(t)$ defines the $t$-quantile of $F$. $\square$

If $F$ is monotone increasing (such as the distribution function $\Phi$ of the standard normal distribution), we see that $F^- = F^{-1}$, i.e., the ordinary inverse of $F$. An illustration of the quantile function is given in Figure 2.3. Notice that intervals where $F$ is constant turn into jumps of $F^-$, and jumps of $F$ turn into intervals of constancy for $F^-$.  

In this way we can define the generalised inverse of the empirical distribution function $F_n$, of a sample $X_1, \ldots, X_n$, i.e.,

$$ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \quad x \in \mathbb{R}. $$

It is easy to verify that $F_n$ has all properties of a distribution function:

- $F_n(-\infty) = 0$, $F_n(\infty) = 1$.
- $F_n(x) \leq F_n(y)$ for $x \leq y$.

---

**Figure 2.3** An “interesting” distribution function $F$, its quantile function $F^-(t)$ (left) and the corresponding function $F^+(1 - x^{-1})$ (right).
• $F_n$ is right-continuous.

If we order the values of the sample as follows:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)},$$

the resulting ordered sample consists of the order statistics of the sample. For example, $X_{(k)}$ is the $k$th order statistic of the sample. Clearly, $X_{(1)}$ is the minimum and $X_{(n)}$ the maximum of the sample. In general, various order statistics have the same value: if some of the $X_i$’s coincide. If this happens one says that the sample has ties. If the $X_i$’s are iid with a density the sample cannot have ties with probability 1. Indeed, in this case the event \( \{X_i = x, X_j = x\} \) has probability 0 for any $i$ and $j$.

Since the empirical distribution function of a sample is a distribution function, one can calculate its quantile function $F_{n}^{-}$ which we call the empirical quantile function. Now assume that the sample has no ties. It is not difficult to see that

$$F_n(X_{(k)}) = k/n, \quad k = 1, \ldots, n,$$

i.e., $F_n$ jumps by $1/n$ at every value $X_{(k)}$ and is constant in $[X_{(k)}, X_{(k+1)})$. This means that the empirical quantile function $F_{n}^{-}$ jumps at the values $k/n$ by $X_{(k)} - X_{(k-1)}$ and remains constant in $((k-1)/n, k/n]$:

$$F_{n}^{-}(t) = \begin{cases} X_{(1)} & t \in (0, 1/n], \\ X_{(k)} & t \in ((k-1)/n, k/n], \quad k = 2, \ldots, n - 1, \\ X_{(n)} & t \in ((n - 1)/n, 1). \end{cases}$$

A fundamental result of probability theory, the Glivenko-Cantelli lemma tells us the following: if $X_1, X_2, \ldots$ is an iid sequence with distribution function $F$, then

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0,$$

implying that $F_n(x) \approx F(x)$ uniformly for all $x$, and one can show under additional conditions that this implies that $F_{n}^{-}(t) \approx F^{-}(t)$ uniformly for $t$-compact sets of $(0, 1)$.

This observation is the basic idea for the QQ-plot: if $X_1, \ldots, X_n$ were a sample with known distribution function $F$, we would expect that $F_{n}^{-}(t)$ is close to $F^{-}(t)$ for all $t \in (0, 1)$, provided $n$ is large. Thus, if we plotted $F_{n}^{-}(t)$ against $F^{-}(t)$ for $t \in (0, 1)$ we should roughly see a straight line.

It is common to plot the graph

$$\left\{ \left( X_{(k)}, F^{-}\left( \frac{k}{n+1}\right) \right) \right\}, \quad k = 1, \ldots, n$$

for a given distribution function $F$. (Modifications of the plotting positions have been used as well; see pp. 292-293 in Embrechts et al. [2] ) Chambers [1] gives the following properties of a QQ-plot:

1. **Comparison of distributions.** If the data were generated from a random sample of the reference distribution, the plot should look roughly linear. This remains true if the data come from a linear transformation of the distribution.
2. Outliers. If one or a few of the data values are contaminated by gross error or for any reason are markedly different in value from the remaining values, the latter being more or less distributed like the reference distribution, the outlying points may be easily identified on the plot.

3. Location and scale. Because a change of one of the distributions by a linear transformation simply transforms the plot by the same transformation, one may estimate graphically (through the intercept and slope) location and scale parameters for a sample of data, on the assumption that the data come from the reference distribution.

4. Shape. Some difference in distributional shape may be deduced from the plot. For example if the reference distribution has heavier tails (tends to have more large values) the plot will curve down at the left and/or up at the right.

For an illustration of (a) and (d) see Figure 2.6. For an illustration of (d) in a two-sided case see Figure 2.6(d). For an illustration in the case of the Danish fire insurance data and the S&P 500 data, see Figure 2.7.

![QQ-plots](image)

**Figure 2.6** QQ-plot of exponentially (a), uniformly (b), lognormally (c) distributed simulated data versus the exponential distribution. In (d) a QQ-plot of $t_4$-distributed data versus the standard normal distribution is given.

References


2.2 The ruin probability of an insurance business

2.2.1 Standard models for the total claim amount of an insurance portfolio

Before we determine the event of ruin we want to consider some models for the evolution of an insurance portfolio. What we can imagine is that we consider a portfolio of car insurance policies where the cars are roughly of the same value (price) and claims occurring in this portfolio cannot be fundamentally different from each other. We call this a \textit{homogeneous} portfolio.

- \textit{Claim sizes} are modeled by iid non-negative random variables $X_1, X_2, \ldots$. 

- Claims arrive at times $T_n = Y_1 + \cdots + Y_n$. The $Y_i$’s are supposed to be iid and positive, $Y_n = T_n - T_{n-1}$ is called \textit{$n$th inter-arrival time}. 

- The claim sizes are independent of their arrival times, i.e., $(X_n)$ and $(T_n)$, hence $(X_n)$ and $(Y_n)$, are mutually independent sequences.

- Until time $t$,

\begin{equation}
N(t) = \# \{ n : T_n \leq t \},
\end{equation}

claims have occurred. The integer-valued random variable $N(t)$ is the \textit{claim number}.

- The \textit{total claim amount} in the portfolio at time $t$ is given by

\[ S(t) = \sum_{i=1}^{N(t)} X_i. \]

- Premiums are paid continuously and linearly at rate $c > 0$, i.e., until time $t$ an amount of $ct$ (Dollars, say) has been accumulated.

- At time zero, the insurance business starts with the \textit{initial capital} $u > 0$ which is usually assumed to be “large”.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure2.png}
\caption{Left: QQ-plot of the Danish fire insurance data (see Figure 2.1) against the exponential distribution with the same expectation. The graph is curved down at the right end indicating that the right tail of the distribution of the data is significantly heavier than the exponential. Right: QQ-plot for the S&P500 data (see Figure 2.2) against the normal distribution with the same mean and variance. The graph is curved down at the right end and curved up at the left end indicating that the tails of the distribution of the data are much heavier than those of the normal distribution.}
\end{figure}
If the $Y_n$’s are iid exponential $\text{Exp}(\lambda)$, i.e.,
\[ P(Y_n > x) = e^{-\lambda x}, \quad x > 0, \]
then it can be shown that $(N(t), t \geq 0)$ constitutes a homogeneous Poisson process with parameter $\lambda$, i.e.,
- the process starts at zero: $N(0) = 0$ a.s.
- for every fixed $t$, $N(t)$ has a Poisson $\text{Poi}(\lambda t)$ distribution:
\[ P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, \ldots ; \]
- it has independent increments, i.e., if $0 = t_0 < t_1 < \cdots < t_n$, then the random variables
\[ N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1}) \]
are independent,
- it has stationary increments, i.e., for $s < t$
\[ N(t) - N(s) \overset{d}{=} N(t + h) - N(s + h) \quad \text{for all } h > 0.\]

See Figure 2.8 for an illustration of the sample paths of a Poisson process with parameter $\lambda = 1$.

In particular, for a homogeneous Poisson process,
\[ N(t) = N(t) - N(0) \overset{d}{=} N(t + h) - N(h), \quad h > 0.\]
Thus the increments $N(t) - N(s)$ for $s < t$ are $\text{Poi}((t - s)\lambda)$ distributed, where the process gained its name from.

As a matter of fact, every process $(N(t), t \geq 0)$ with the above properties has representation as a counting process (2.6) with iid $\text{Exp}(\lambda)$ inter-arrival times, and so we have learned about two possible definitions of a Poisson process. Since the random arrival times of claims can be interpreted in a broader context as renewal times (for example as the times when defect light bulbs have to be replaced or broken parts of a machine have to be changed), $(T_n)$ is often referred to as renewal process and the corresponding process $(N(t), t \geq 0)$ as renewal counting process.

For historical reasons, the above model of an insurance business is called the Cramér-Lundberg model if $(N(t), t \geq 0)$ is a homogeneous Poisson process. (Filip Lundberg in his 1900 PhD thesis proposed this model for modeling the insurance business; Harald Cramér made the model mathematically sound and used it extensively in his mathematical derivations.) The total claim amount process $(S(t), t \geq 0)$ is referred to as compound Poisson process. If $(N(t), t \geq 0)$ is a renewal counting process the model for the evolution of the insurance portfolio is called the renewal model. In other words, if the claim number process of an insurance portfolio is Poisson, the renewal model is called Cramér-Lundberg.

### 2.2.2 The risk process and the probability of ruin

The risk process corresponding to the renewal model is given by

$$U(t) = u + c t - S(t), \quad t \geq 0.$$  

An illustration of a path of such a process can be found in Figure 2.9. The event

$$\{U(t) < 0 \text{ for some } t \geq 0\}$$

is called ruin. If this happens, the total claim amount exceeds the initial capital plus the accumulated premium, and the owner of the portfolio is in principle bankrupt. A major question which has troubled mathematicians such as Harald Cramér is whether it is possible to determine the

![Figure 2.9 One realization of the risk process $(U(t))$.](image)
The probability of ruin, i.e.,
\[
\psi(u) = P\left(U(t) < 0 \text{ for some } t \geq 0\right)
\]
\[
= P\left(\inf_{t \geq 0} U(t) < 0\right)
\]
\[
= P\left(\inf_{t \geq 0} (S(t) - c\, t) > u\right)
\]
A glance at Figure 2.9 convinces one that ruin can occur only at the times when claims arrive, i.e., at times \(T_n\). Hence
\[
\psi(u) = P\left(\sup_{n \geq 0} (S(T_n) - c\, T_n) > u\right)
\]
\[
= P\left(\sup_{n \geq 0} \sum_{i=1}^{n} (X_i - c\, Y_i) > u\right),
\]
where we used that \(S(T_n) = \sum_{i=1}^{n} X_i\). Writing \(Z_i = X_i - c\, Y_i\), we arrive at
\[
\psi(u) = P\left(\max_{n \geq 0} \sum_{i=1}^{n} Z_i > u\right).
\]
The strong law of large numbers tells us that as \(n \to \infty\),
\[
\frac{1}{n} \sum_{i=1}^{n} Z_i \overset{a.s.}{\to} E(Z_i) = E(X_1) - c\, E(Y_1).
\]
Thus, if \(E(Z_i) > 0\), for large \(n\),
\[
\sum_{i=1}^{n} Z_i \approx n E(Z_i),
\]
which implies that \(\psi(u) = 1\) for any choice of \(u\), i.e., ruin is unavoidable with probability 1. Thus it is reasonable to assume that \(E(Z_i) < 0\) or, equivalently,
\[(2.7)\]
\[c E(Y_1) > E(X_1).\]
The latter is the net profit condition. It simply means that, on average, there must flow more premium than claims into the insurance business. Only in this case, the business is in a stable situation.
In the Cramér-Lundberg model, \(Y_i \sim Exp(\lambda)\), hence \(EY_i = 1/\lambda\), and therefore (2.7) turns into
\[(2.8)\]
\[c > \lambda E(X_1).\]
In what follows we will always assume that the net profit condition holds.
2.2.3 The ruin probability under a small claim condition

In the 1930s Harald Cramér used some deep mathematical theory (the theory of large deviations) to show that the probability of ruin $\psi(u)$ is negligible if the initial capital is "large" and the claims are "small". To make this statement precise one has to say what "large" and "small" means. Cramér defined in the Cramér-Lundberg model the following small claim condition: the moment generating function of the claim sizes exists in some neighborhood of the origin. This means that

$$m_X(h) = E e^{hx} < \infty \quad \text{for } h < h_0, \text{some } h_0 > 0. \tag{2.9}$$

(For negative $h$ the function is obviously finite since $e^{hx} \leq 1$.) As a matter of fact, Markov's inequality tells us that for $x > 0$ and $h > 0$,

$$P(X_1 > x) = P\left(e^{hx} > e^{hx}ight) \leq E \left(\frac{e^{hx}}{e^{hx} I_{[h,\infty)} \left(\frac{e^{hx}}{e^{hx}}\right)}\right) \leq \frac{E e^{hx}}{e^{hx}}.$$

Here and in what follows, $I_A$ denotes the indicator function of the set $A$,

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise}. \end{cases}$$

Hence, there exists a finite constant $K > 0$ such that

$$e^{hx} P(X_1 > x) \leq K,$$

and it is not difficult to see that

$$e^{hx} P(X_1 > x) \to 0 \quad \text{as } x \to \infty \text{ for } h' < h.$$

This condition means that claims exceeding a large threshold $x$ are exponentially small, i.e., they can in principle be neglected since they occur with a tiny probability. Most familiar distributions of classical statistics satisfy the small claim condition (2.9); see Table 2.10.

Under a small claim condition, Cramér showed the following result.

**Theorem 2.11** (Cramér-Lundberg theorem)

Consider the Cramér-Lundberg model including the net profit condition (2.8) and assume that the iid claim sizes $X_i$ are positive with probability 1. Assume there exists a $\nu > 0$ such that

$$m_X(\nu) = E e^{\nu X_1} = 1 + \frac{c\nu}{\lambda}. \tag{2.10}$$

Then the following relations hold.

1. For all $u \geq 0$,

$$\psi(u) \leq e^{-\nu u}.$$

2. If, moreover, $E(X_1 e^{\nu X_1}) < \infty$, and

$$\rho := \frac{c}{\lambda E(X_1)} - 1,$$

then

$$\lim_{u \to \infty} e^{\nu u} \psi(u) = \left[\frac{\nu}{\rho \lambda} \int_0^\infty x e^{\nu x} F(x) dx\right]^{-1} < \infty.$$
<table>
<thead>
<tr>
<th>Name</th>
<th>Tail $\bar{F}$ or density $f$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\bar{F}(x) = e^{-\lambda x}$</td>
<td>$\lambda &gt; 0$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$</td>
<td>$\alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\bar{F}(x) = e^{-cx^\tau}$</td>
<td>$c &gt; 0, \tau \geq 1$</td>
</tr>
<tr>
<td>Truncated normal</td>
<td>$f(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}$</td>
<td>—</td>
</tr>
<tr>
<td>Any distribution with bounded support</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.10 Claim size distribution functions: “small claims”. All distribution functions have support $(0, \infty)$.

3. In the case that the claims sizes are exponentially distributed, i.e., $\bar{F}(x) = e^{-x/\mu}$ for $x > 0$, one has an exact expression for $\psi(u)$:

$$
\psi(u) = \frac{1}{1 + \rho} \exp \left\{ -\frac{\rho}{\mu(1 + \rho)} u \right\}, \quad u \geq 0.
$$

For a proof of this result we refer to Embrechts et al. [1], Section 1.2, Feller [2], Sections VI.5, XI.7a, or Grandell [3].

The fundamental, so-called *Cramér-Lundberg condition* (2.10), can also be written as

$$
\int_0^\infty e^{-ux} \bar{F}(x) dx = \frac{c}{\lambda}.
$$

It follows immediately from the definition of Laplace–Stieltjes transform that, whenever $\nu$ in (2.10) exists, it is uniquely determined; see also Grandell [3], p. 58. The constant $\nu$ is also referred to as the *Lundberg exponent* or *adjustment coefficient* of the underlying risk process.

Since the definition of the Lundberg exponent requires the existence of the moment generating function $m_X(h)$ in some neighborhood of the origin we see that the Cramér-Lundberg ruin estimate is applicable only under a small claim condition in the sense we discussed above.

An intuitive interpretation of the Cramér-Lundberg ruin estimate is the following. If the claim sizes are “small” and the initial capital $u$ is “very large” the probability of ruin, i.e., bankruptcy of the insurance business, is exponentially small, i.e., very unlikely to happen at all. Some sophisticated analysis of the ruin probability shows that in the case of small claims ruin can occur only if a large number of claims builds up over a short time horizon. This changes when large claims are involved: ruin will happen out of the blue – due to one unusually large claim which determines the whole portfolio.

If one scans the literature with the following question in mind:

*Which distributions do actually fit claim size data?*
<table>
<thead>
<tr>
<th>Name</th>
<th>Tail $\tilde{F}$ or density $f$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>$f(x) = \frac{1}{\sqrt{2\pi} \sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$</td>
<td>$\mu \in \mathbb{R}$, $\sigma &gt; 0$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$\tilde{F}(x) = \left(\frac{\kappa}{\kappa + x}\right)^\alpha$</td>
<td>$\alpha, \kappa &gt; 0$</td>
</tr>
<tr>
<td>Burr</td>
<td>$\tilde{F}(x) = \left(\frac{\kappa}{\kappa + x^\tau}\right)^\alpha$</td>
<td>$\alpha, \kappa, \tau &gt; 0$</td>
</tr>
<tr>
<td>Benktander-type-I</td>
<td>$\tilde{F}(x) = (1 + 2(\beta/\alpha) \ln x) e^{-\beta(\ln x)^2 - (\alpha + 1) \ln x}$</td>
<td>$\alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td>Benktander-type-II</td>
<td>$\tilde{F}(x) = e^{\alpha/\beta x - (1 - \beta) e^{-\alpha x^\beta}/\beta}$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$0 &lt; \beta &lt; 1$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\tilde{F}(x) = e^{-\alpha x^\tau}$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$0 &lt; \tau &lt; 1$</td>
</tr>
<tr>
<td>Loggamma</td>
<td>$f(x) = \frac{\alpha^\beta}{\Gamma(\beta)} (\ln x)^{\beta-1} x^{-\alpha-1}$</td>
<td>$\alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td>Truncated</td>
<td>$\tilde{F}(x) = P(</td>
<td>X</td>
</tr>
<tr>
<td>$\alpha$-stable</td>
<td>where $X$ is $\alpha$-stable</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.12** Claim size distributions: “large claims”. All distribution functions have support $(0, \infty)$ except for the Benktander cases and the loggamma with $(1, \infty)$. 

16
then most often one will find one of the distributions listed in Table 2.12. All the distributions in Table 2.10 allow for the construction of the Lundberg exponent. For the ones in Table 2.12 however, this exponent does not exist. For that reason we have labelled the two tables with “small claims”, respectively “large claims”. A more precise discussion of these distributions follows below.

A detailed study of the properties of the distributions listed in Table 2.12 with special emphasis on insurance is to be found in Hogg and Klugman [4].

For the sake of argument, assume that we have a portfolio following the Cramér–Lundberg model for which individual claim sizes can be modeled by a Pareto distribution

$$F(x) = (1 + x)^{-\alpha}, \quad x \geq 0, \quad \alpha > 1.$$ 

It then follows that $EX_1 = \int_0^\infty (1 + x)^{-\alpha} dx = (\alpha - 1)^{-1}$ and the net profit condition amounts to $\rho = c(\alpha - 1)/\lambda - 1 > 0$. Question:

*Can we work out the exponential Cramér–Lundberg estimate in this case, for a given premium rate $c$ satisfying the above condition?*

The answer to this question is *no*.

Indeed, in this case, for every $\nu > 0$

$$\int_0^\infty e^{\nu x} (1 + x)^{-\alpha} dx = \infty,$$

i.e., there is *no* exponential Cramér–Lundberg estimate in this case.

However, it turns out that most individual claim size data are modeled by such distributions; see for instance Hogg and Klugman [4] and Ramla-Hansen [5, 6] for very convincing empirical evidence on this. We have seen in Section 2.1 that insurance data do not satisfy a small claim condition, and we will give further evidence for this fact in the sequel. So clearly, classical risk theory has to be adjusted to take this observation into account. In the next section we discuss the class of subexponential distributions which will be the candidates for loss distributions in the heavy-tailed case. A detailed discussion of the theory of subexponential distributions is rather technical, so we content ourselves with an overview of that part of the theory which is most easily applicable within risk theory in particular and insurance and finance in general. We will present the large-claims equivalent of the Cramér–Lundberg estimate below.

Before we can do that we make an excursion to a class of functions which turns out to be useful for modeling heavy tails: the regularly varying functions.

References


2.2.4 Slowly and regularly varying functions and random variables

Although the distributions in Table 2.12 may look very different some of them have very similar tails. Those include Pareto, Burr, stable and loggamma distributions. In particular, their right tails can be written in the form

\[ \overline{F}(x) = 1 - F(x) = P(X > x) = \frac{L(x)}{x^\alpha}, \quad x > 0, \]

for some constant \( \alpha > 0 \) and a positive function \( L(x) \). This function has the interesting property that

\[ \lim_{x \to \infty} \frac{L(cx)}{L(x)} = 1 \quad \text{for all } c > 0. \]

A function with this property is called slowly varying (at infinity). Examples of such functions are constants, logarithms, positive powers of logarithms, iterated logarithms.

Every slowly varying function has representation for \( x \geq x_0 \), some \( x_0 > 0 \),

\[ L(x) = c_0(x) \exp \left\{ \int_{x_0}^{x} \frac{\varepsilon(t)}{t} dt \right\}, \tag{2.11} \]

where \( \varepsilon(t) \to 0 \) as \( t \to \infty \) and \( c_0(t) \) is a positive function satisfying \( c_0(t) \to c_0 \) for some positive constant \( c_0 \). From the latter property one can show that for every \( \delta > 0 \),

\[ \lim_{x \to \infty} \frac{L(x)}{x^\delta} = 0 \quad \text{and} \quad \lim_{x \to \infty} x^\delta L(x) = \infty, \]

i.e., it is “small” compared to any power function.

If \( L(x) \) is slowly varying and \( \delta \in \mathbb{R} \), the function

\[ f(x) = x^\delta L(x), \quad x > 0, \]

is said to be regularly varying with index \( \delta \). Moreover, if a non-negative-valued random variable \( X \) has tail \( P(X > x) = L(x)x^{-\alpha} \) we say that \( X \) is regularly varying with index \( \alpha \).

An alternative way to define regular variation with index \( \delta \) is to require

\[ \lim_{x \to \infty} \frac{f(cx)}{f(x)} = c^\delta, \quad \text{for all } c > 0. \tag{2.12} \]

Using the representation (2.11) it can be shown that (2.12) even holds uniformly for \( c \)-compact sets of \((0, \infty)\), i.e., for any compact (=closed and bounded) set \( K \subset (0, \infty) \),

\[ \lim_{x \to \infty} \sup_{c \in K} \left| \frac{f(cx)}{f(x)} - c^\delta \right| = 0. \tag{2.13} \]

Regular variation is one possible way to describe “small” deviations from exact power law behavior. It is hard to believe that social or natural phenomena can be described by exact power law behavior. It is, however, known that various phenomena, such as Zipf’s law, fractal dimensions, the probability of exceedances of high thresholds by certain iid data, can be well described by “almost power law functions” in a suitable sense. Regular variation is a reasonable concept in this
context. It has been carefully studied for many years; see Bingham et al. [1] for an encyclopaedic treatment.

One of the surprising properties of regular variation is that one can integrate regularly varying functions in an asymptotic sense as if they were power laws. To be more precise, consider an exact power function \( f(x) = x^\delta \) for some \( \delta \in \mathbb{R} \). Then for \( x \geq x_0 > 0 \),

\[
\int_{x_0}^x f(y) \, dy = \begin{cases} 
\frac{x^{\delta+1}}{\delta + 1} - \frac{x_0^{\delta+1}}{\delta + 1} & \text{for } \delta \neq -1, \\
\log x - \log x_0 & \text{for } \delta = -1.
\end{cases}
\]

Analogously, for \( x > 0 \),

\[
\int_x^\infty f(y) \, dy = \begin{cases} 
\frac{x^{\delta+1}}{\delta + 1} & \text{for } \delta < -1, \\
\infty & \text{for } \delta \geq -1.
\end{cases}
\]

The corresponding relations for regularly varying functions with index \( \delta \), i.e., \( f(x) = x^\delta L(x) \) for a slowly varying \( L(x) \), are then given by

\[
\begin{align*}
\int_{x_0}^x f(y) \, dy & \sim -\frac{x f(x)}{\delta + 1} & \text{as } x \to \infty \text{, for } \delta > -1, \\
& \text{is slowly varying for } \delta = -1. \\
\int_x^\infty f(y) \, dy & \sim \frac{x f(x)}{\delta + 1} & \text{as } x \to \infty \text{, for } \delta < -1, \\
& \infty \text{ for } \delta > -1.
\end{align*}
\]

Here and in what follows, for any positive functions \( h(x) \) and \( g(x) \),

\[
(2.16) \quad h(x) \sim g(x) \quad \text{as } x \to \infty \text{ means that } \lim_{x \to \infty} \frac{h(x)}{g(x)} = 1.
\]

Relations (2.14) and (2.15) are referred to as Karamata’s theorem named after a Croatian mathematician who lived in the first half of the 20th century and laid the foundations to this theory. His motivation was by no means a practical one; only later it turned out that regular variation is a useful concept, as we will see soon. Karamata’s theorem can be proved by using representation (2.11).

References


### 2.2.5 Aggregation of regularly varying risks

The insurance business is about the aggregation of claims with size \( X_i \) in portfolios and about determining suitable premiums for them. The latter problem can be solved only if one has an impression of the size of the aggregated claims. In other words, we want to find theoretical means in order to describe the distribution of the total claim amount \( S(t) = \sum_{i=1}^{N(t)} X_i \). This task, however, is a delicate one, and so we attempt an easier one first: determine the tail of the distribution of \( n \) aggregated independent risks, i.e., of the random variable

\[
S_n = X_1 + \cdots + X_n.
\]
where the $i$th risk is described by the non-negative random variable $X_i$.

The following simple result describes the distribution of $S_n$ under the assumption that each $X_i$ is regularly varying with index $\alpha > 0$, i.e., there exist a slowly varying function $L_i(x)$ such that

$$
(2.17) \quad \mathcal{F}_i(x) = P(X_i > x) = \frac{L_i(x)}{x^\alpha}, \quad x > 0.
$$

We start with two risks $X_1$ and $X_2$.

**Lemma 2.13** Assume that $X_1$ and $X_2$ are independent non-negative random variables with regularly varying tails of index $\alpha > 0$, i.e., (2.17) holds. Then $X_1 + X_2$ is regularly varying with the same index. More precisely,

$$
P(X_1 + X_2 > x) \sim P(X_1 > x) + P(X_2 > x) = x^{-\alpha} (L_1(x) + L_2(x)), \quad x \to \infty.
$$

In what follows, we use the symbols $o(1)$ and $O(1)$ for asymptotic relations. Here

$$
h(x) = o(g(x)) \quad \text{as} \quad x \to x_0 \quad \text{means that} \quad \lim_{x \to x_0} \frac{|h(x)|}{g(x)} = 0.
$$

Analogously,

$$
h(x) = O(g(x)) \quad \text{as} \quad x \to x_0 \quad \text{means that} \quad \limsup_{x \to x_0} \frac{|h(x)|}{g(x)} < \infty.
$$

**Proof.** Write $G(x) = P(X_1 + X_2 \leq x)$ for the distribution function of $X_1 + X_2$. Using $\{X_1 + X_2 > x\} \supset \{X_1 > x\} \cup \{X_2 > x\}$, one easily checks that

$$
\overline{G}(x) \geq (\mathcal{F}_1(x) + \mathcal{F}_2(x))(1 - o(1)).
$$

If $0 < \delta < 1/2$, then from

$$
\{X_1 + X_2 > x\} \subset \{X_1 > (1 - \delta)x\} \cup \{X_2 > (1 - \delta)x\} \cup \{X_1 > \delta x, X_2 > \delta x\},
$$

it follows that

$$
\overline{G}(x) \leq \mathcal{F}_1((1 - \delta)x) + \mathcal{F}_2((1 - \delta)x) + \mathcal{F}_1(\delta x) \mathcal{F}_2(\delta x)
= \left(\mathcal{F}_1((1 - \delta)x) + \mathcal{F}_2((1 - \delta)x)\right)(1 + o(1)).
$$

Hence

$$
1 \leq \liminf_{x \to \infty} \frac{\overline{G}(x)}{\mathcal{F}_1(x) + \mathcal{F}_2(x)} \leq \limsup_{x \to \infty} \frac{\overline{G}(x)}{\mathcal{F}_1(x) + \mathcal{F}_2(x)} \leq (1 - \delta)^{-\alpha},
$$

which proves the result upon letting $\delta \downarrow 0$. \qed

An important corollary obtained via induction on $n$ is the following:

**Corollary 2.14** Assume that $X_1, \ldots, X_n$ are $n$ iid non-negative risks, each of them being regularly varying with index $\alpha > 0$. Then

$$
P(S_n > x) \sim n \ P(X_1 > x), \quad x \to \infty.
$$

\[20\]
Suppose now that \( X_1, \ldots, X_n \) are iid with distribution function \( F \) as in the above corollary. Denote the partial sum of \( X_1, \ldots, X_n \) by \( S_n = X_1 + \cdots + X_n \) and their maximum by \( M_n = \max(X_1, \ldots, X_n) \). Then for all \( n \geq 2 \),

\[
P(M_n > x) = \frac{1}{n}
\]

\[
= \frac{1}{n} \sum_{k=0}^{n-1} F^k(x) 
\]

\[
\sim \frac{1}{n} \int F(x) \, dx, \quad x \to \infty.
\]

Therefore, with the above notation, Corollary 2.14 can be reformulated as: if \( X_i \) is regularly varying with index \( \alpha > 0 \) then

\[
P(S_n > x) \sim P(M_n > x), \quad x \to \infty.
\]

This implies that for distributions with regularly varying tails, the tail of the distribution of the sum \( S_n \) is mainly determined by the tail of the distribution of the maximum \( M_n \). This is exactly one of the intuitive notions of heavy-tailed distribution or large claim. Hence, stated in a somewhat vague way:

> Under the assumption of regular variation, the tail of the maximum determines the tail of the sum.

This is somewhat similar to the observation we made in Section 2.1.1. However, the crucial difference to the latter is that we compared \( M_n \) and \( S_n \) directly, i.e., we considered the random ratio \( M_n/S_n \) and studied its properties. Here we study the deterministic ratio of tail probabilities \( P(M_n > x)/P(S_n > x) \).

### 2.2.6 Subexponential distributions

We learnt in the previous section that for iid regularly varying random variables \( X_1, X_2, \ldots \) with positive index \( \alpha \) the tail of the sum \( S_n = X_1 + \cdots + X_n \) is essentially determined by the tail of the maximum \( M_n = \max_{i=1, \ldots, n} X_i \). To be precise, we found out that \( P(S_n > x) \sim P(M_n > x) \) as \( x \to \infty \) for every \( n = 1, 2, \ldots \). The latter relation can be taken as a natural definition for “heavy-tailedness” of a distribution:

**Definition 2.15** (Subexponential distribution)
The positive random variable \( X \) and its distribution are said to be subexponential if for a sequence \((X_i)\) of iid random variables with the same distribution as \( X \) the following relation holds:

\[
(2.18) \quad \text{For all } n \geq 2: \quad P(S_n > x) \sim P(M_n > x), \quad x \to \infty.
\]

The class of all subexponential distributions is denoted by \( S \).

One can show that the defining property (2.18) holds for all \( n \geq 2 \) if it holds for some \( n \geq 2 \); see Section 1.3.2 in [2] for details.

As we have learnt in Section 2.2.5, \( P(M_n > x) \sim nF(x) \) as \( x \to \infty \), and therefore the defining property (2.18) can also be formulated as

For all \( n \geq 2 \):

\[
\lim_{x \to \infty} \frac{P(S_n > x)}{F(x)} = n.
\]

In what follows we want to consider some elementary properties of subexponential distributions. For complete proofs we refer to Section 1.3.2 in [2].
Lemma 2.16 (Some basic properties of subexponential distributions)

1. If \( F \in \mathcal{S} \), then for any \( y > 0 \),

\[
\lim_{x \to \infty} \frac{F(x - y)}{F(x)} = 1.
\]

The convergence is uniform for \( y \)-compact sets.

2. If (2.19) holds then, for all \( \varepsilon > 0 \),

\[
e^{\varepsilon x} F(x) \to \infty, \quad x \to \infty.
\]

3. If \( F \in \mathcal{S} \) then, given \( \varepsilon > 0 \), there exists a finite constant \( K \) so that for all \( n \geq 2 \),

\[
\frac{P(S_n > x)}{F(x)} \leq K(1 + \varepsilon)^n, \quad x \geq 0.
\]

For the proof of (3), see Lemma 1.3.5 in [2].

Proof of (1). For \( x \geq y > 0 \),

\[
\frac{F^{2y}(x)}{F(x)} = 1 + \int_0^y \frac{F(x - t)}{F(x)} dF(t) + \int_y^\infty \frac{F(x - t)}{F(x)} dF(t)
\]

\[
\geq 1 + F(y) + \frac{F(x - y)}{F(x)} (F(x) - F(y)).
\]

Thus, for \( x \) large enough so that \( F(x) - F(y) \neq 0 \),

\[
1 \leq \frac{F(x - y)}{F(x)} \leq \left( \frac{F^{2y}(x)}{F(x)} - 1 - F(y) \right) (F(x) - F(y))^{-1}.
\]

In the latter estimate, the right-hand side tends to 1 as \( x \to \infty \). This proves (2.19).

The property (2.19) means that \( \overline{F}(\ln x) \) is slowly varying so that uniform convergence follows from the uniform convergence theorem for slowly varying functions; see Section 2.2.4.

Proof of (2). By virtue of (1), the function \( \overline{F}(\ln x) \) is slowly varying. But then the conclusion that \( x^\varepsilon \overline{F}(\ln x) \to \infty \) as \( x \to \infty \) follows immediately from the representation theorem for slowly varying functions; see Section 2.2.4.

Lemma 2.16(2) justifies the name “subexponential” for \( F \in \mathcal{S} \); indeed \( \overline{F}(x) \) decays to 0 slower than any exponential \( e^{-\varepsilon x} \) for \( \varepsilon > 0 \). Furthermore, since for any \( \varepsilon > 0 \):

\[
E e^{\varepsilon X} \geq E (e^{\varepsilon X} I_{[y, \infty)}) \geq e^{\varepsilon y} \overline{F}(y), \quad y \geq 0,
\]

it follows from Lemma 2.16(2) that for \( F \in \mathcal{S} \), \( E e^{\varepsilon X} = \infty \) for all \( \varepsilon > 0 \). Therefore the moment generating function of a subexponential distribution is infinite for all positive \( \varepsilon \). This result was first proved by Chistyakov [1], Theorem 2.

As follows from the proof of Lemma 2.16(2) the latter property holds true for the larger class of distributions satisfying (2.19). The latter relation can be taken as another definition of heavy-tailed...
distribution. It means that the tails $P(X > x)$ and $P(X + y > x)$ are not significantly different for any fixed $y$. In particular, it says that for any $y > 0$ as $x \to \infty$,
\[
\frac{P(X > x + y)}{P(X > x)} = \frac{P(X > x + y, X > x)}{P(X > x)} = P(X > x + y \mid X > x) \to 1.
\]
Thus, once $X$ exceeded a high threshold $x$ it is very likely to exceed an even higher threshold $x + y$. This changes completely when we look, for example, at an exponential or a normal random variable.

Property (2.19) helps one to exclude certain distributions from the class $S$. However, it is in general difficult to determine whether a given distribution is subexponential.

**Example 2.17** (Examples of subexponential distributions)
The large claim distributions in Table 2.12 are subexponential. The small claim distributions in Table 2.10 are not subexponential. However, the tail of a subexponential distribution can be very close to an exponential distribution. Not only the heavy-tailed Weibull distributions with tail
\[
F(x) = e^{-e^{-x}}, \quad x \geq 0, \quad \text{for some } \tau \in (0, 1),
\]
but also the distributions with tail
\[
F(x) \sim e^{-x^{\beta}}, \quad x \geq 0, \quad \text{for some } \beta > 0,
\]
are subexponential. We refer to Section 1.4.1 in [2] for details. \hfill \square

### 2.2.7 The Cramér-Lundberg theorem for subexponential claims

Recall the Cramér-Lundberg model for an insurance business from Section 2.2.1. In Section 2.2.3 we approximated the probability of ruin in a homogeneous portfolio, consisting of iid claims with distribution function $F$. If the latter has a finite moment generating function in some neighborhood of zero, the Cramér-Lundberg Theorem 2.11 tells us that the ruin probability is exponentially small for a large initial capital: $\psi(u) \sim e^{-\nu u}$ as $u \to \infty$ for some positive constant $\nu$, the adjustment coefficient. The latter was defined via the moment generating function.

For large claims one cannot use the idea of Cramér’s proof. However, one can prove the following result (see for example Grandell [4], pp. 216-217):

\begin{equation}
\psi(u) = \frac{\rho}{1 + \rho} \sum_{n=0}^{\infty} (1 + \rho)^{-n} P(S_n^s > u).
\end{equation}

The constant
\[
\rho = \frac{c}{\lambda E(X_1)} - 1,
\]
also played a prominent role in the Cramér-Lundberg theorem, where $c > 0$ is the premium rate, $\lambda$ the intensity of the underlying Poisson process and $EX_1$ is the expected claim size. The quantity $\rho$ is positive by virtue of the net profit condition; cf. (2.8). Moreover,
\[
S_0^s = 0, \quad S_n^s = X_1^s + \cdots + X_n^s, \quad n \geq 1,
\]
is a random walk of iid positive random variables $X_i^s$ with distribution given by
\[
F_I(x) = \frac{1}{EX_1} \int_0^x F(y) \, dy, \quad x \geq 0.
\]
Clearly, for this definition $EX_1$ must be finite. For obvious reasons, $F_I$ is called the integrated tail distribution of $F$. A fact from measure theory tells us that

$$EX_1 = \int_0^\infty \mathcal{F}(x) \, dx,$$

and therefore $F_I$ defines a distribution, indeed.

Relation (2.21) has a nice probabilistic interpretation. Write $p = (1+\rho)^{-1}$. Then we know that the corresponding geometric series converges:

$$\sum_{n=0}^{\infty} p^n = \frac{1}{1-p} = \frac{1+\rho}{\rho}.$$ 

Hence,

$$P(Y = n) = (1-p) \, p^{n-1} = \rho \, (1+\rho)^{-n}, \quad n = 1, 2, \ldots ,$$

defines a geometrically distributed random variable $Y$ with values $n = 1, 2, \ldots$. Notice that {\{Y = n\}} is the event that in $n$ independent success-failure trials (success probability $(1-p)$, failure probability $p$) success occurs for the first time.

Assume that $Y$ and $(X_i^*)$ are independent. The random variable

$$S_Y^* = \sum_{i=1}^{Y-1} X_i^*$$

is said to have a compound geometric distribution. Using the standard rules for conditional probabilities, we see that the following holds:

$$P(S_Y^* \leq u) = \sum_{n=0}^{\infty} P(Y = n+1) \, P(S_Y^* \leq u \mid Y = n+1)$$

$$= \sum_{n=0}^{\infty} P(Y = n+1) \, P(S_n^* \leq u \mid Y = n+1)$$

$$(2.22) \quad = \sum_{n=0}^{\infty} P(Y = n+1) \, P(S_n^* \leq u),$$

where we used that $S_n^*$ and $Y$ are independent. Thus we arrived at

$$P(S_Y^* \leq u) = \frac{\rho}{1+\rho} \sum_{n=0}^{\infty} (1+\rho)^{-n} \, P(S_n^* \leq u)$$

$$= 1 - \psi(u).$$

Hence the non-ruin probability $1 - \psi(u)$ has interpretation as a compound geometric probability with compounding distribution $F_I$.

A glance at the relation (2.21) suggests the following idea. If $X_i^*$ with distribution function $F_I \in \mathcal{S}$ we know by definition of the subexponential distributions that as $u \to \infty$,

$$P(S_n^* > u) \sim n \, P(X_i^* > u) = n \, \mathcal{F}_I(u).$$

24
Hence one might hope that, in an asymptotic sense, one can replace in (2.21) the quantity \(P(S_n^0 > u)\) by \(nF_I(u)\). In general this would not be allowed since one has to interchange the limit as \(u \to \infty\) with the infinite series \(\sum_{n=0}^{\infty} \). However, the property (2.20) of a subexponential distribution ensures that this is possible. Hence we conclude from (2.20) that the following result holds for large claims.

\[
\psi(u) \sim \frac{\rho}{1 + \rho} \sum_{n=1}^{\infty} (1 + \rho)^{-n} n \frac{1}{n} \frac{\rho^{-1}}{F_I(u)}.
\]

Here we used the fact that

\[
\frac{d}{dp} \left[ \sum_{n=0}^{\infty} (1 + \rho)^{-n} \right] = \frac{d}{dp} \frac{1 + \rho}{\rho} = -\frac{1}{\rho^2}
\]

\[
= -\sum_{n=1}^{\infty} n (1 + \rho)^{-n-1} = -\frac{1}{1 + \rho} \sum_{n=1}^{\infty} n (1 + \rho)^{-n}.
\]

**Theorem 2.18** (The Cramér–Lundberg theorem for large claims)

*Consider the Cramér-Lundberg model with net profit condition \(\rho > 0\) and \(F_I \in \mathcal{S}\), then*

(2.23)

\[
\psi(u) \sim \rho^{-1} F_I(u), \quad u \to \infty.
\]

It was shown by Embrechts and Veraverbeke [3] that \(F_I \in \mathcal{S}\) is also necessary for the result (2.23).

It is worthwhile to compare the ruin probabilities under a small and a large claim condition. From the Cramér-Lundberg Theorem 2.11 we know that, under a small claim regime, for some positive \(\nu\),

\[
\psi(u) \sim e^{-\nu u} \quad \text{as } u \to \infty.
\]

In contrast to the latter, the Embrechts and Veraverbeke theorem gives another benchmark result under a large claim condition, i.e., subexponentiality of the integrated claim size distribution function \(F_I\):

\[
\psi(u) \sim \frac{1}{\rho} F_I(u).
\]

**Example 2.19** (Pareto claims)

Assume that the claim sizes are Pareto distributed, i.e.,

\[
F(x) = (1 + x)^{-\alpha} \quad \text{for some } \alpha > 1.
\]

The assumption \(\alpha > 1\) ensures that \(EX_1 < \infty\). Straightforward calculation gives

\[
F_I(x) = \frac{1}{EX_1} \int_x^{\infty} F(y) \, dy = \frac{1}{EX_1} \int_x^{\infty} (1 + y)^{-\alpha} \, dy
\]

\[
= (1 + x)^{-\alpha+1}.
\]

Since \(F_I\) is regularly varying, it is subexponential and therefore (2.23) yields

\[
\psi(u) \sim \rho^{-1} (1 + u)^{-\alpha+1} \quad \text{as } u \to \infty.
\]

This means that the probability of ruin \(\psi(u)\) is non-negligible even for large values of \(u\). □

25
Using Karamata’s theorem (see p. 19), it is not difficult to see that regular variation of $\overline{F}$ with index $-\alpha < -1$ implies regular variation of $\overline{F_I}$ with index $-\alpha + 1$. Hence the same argument as for the Pareto distribution yields that the ruin bound (2.23) holds.

Theorem 2.18 shows that, whenever subexponential interigated tail distributions $F_I$ occur, ruin is much more more likely and therefore a portfolio with large claims is much more dangerous than a portfolio with small claims. It is in general difficult to check whether $F_I \in S$; see Section 1.4.2 in [2]. There it is shown that Theorem 2.18 is applicable to the large claims distributions of Table 2.12.

The Embrechts and Veraverbeke result is only a qualitative one; it can be shown that the asymptotic approximation in (2.23) can be very bad for suitably chosen distributions $F$. However, almost all claim size distributions fitted to real-life data in actuarial history have been heavy-tailed (subexponential), and therefore this result shows that the extremes of the sample of the claims play a prominent role for determining the ruin probability. Since in a subexponential portfolio one large claim (such as one big hurricane or earthquake; see Table 1.2 for various prominent examples) can essentially determine the total claim amount, caution has to be taken. In practice the danger of very large claims is usually taken care of by reinsurance products. The basic idea of reinsurance is that primary insurers share part of their premiums with other insurance companies with the aim to cover large losses exceeding a sufficiently high threshold. Over the last few decades it has become common to share the risk (and the profit) between insurance companies by buying certain “layers” of risk from various companies. This has the advantage that, when a large claim occurs, the primary insurer is not responsible for the whole claim size, but only for a part of it. Moreover, various insurance companies have specialized in the largest claims which exceed the highest insurable threshold. In this case, extreme value theory immediately enters in order to determine the distribution of the high level excesses and to calculate premiums.

Another consequence of the techniques leading to the ruin formula (2.23) is that one can determine the tail of the total claim amount of subexponential claim sizes. Indeed, recall that

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

is compound Poisson, i.e., $N(t)$ is Pois($\lambda t$) and independent of $(X_i)$. Then the same calculations as for the geometric compound case (see (2.22)) yield that

$$P(S(t) > u) = \sum_{n=0}^{\infty} P(N(t) = n) \, P(S_n > u)$$

$$= \sum_{n=0}^{\infty} e^{-\lambda} \left( \frac{(\lambda t)^n}{n!} \right) P(S_n > u).$$

Using the subexponential relation $P(S_n > u) \sim n \overline{F}(u)$ and $EN(t) = \lambda t$, we arrive at

$$P(S(t) > u) \sim \sum_{n=0}^{\infty} e^{-\lambda} \left( \frac{(\lambda t)^n}{n!} \right) n \, \overline{F}(u)$$

$$= EN(t) \, \overline{F}(u) = \lambda \, t \, \overline{F}(u), \quad u \to \infty. \quad (2.24)$$

This complements the results for the sums $S_n$ with deterministic index. Moreover, it gives a more realistic view at the very large total claim amounts than the central limit theorem does. Recall that

$$ES(t) = \lambda \, EX_1 \, t \quad \text{and} \quad \text{var}(S(t)) = \lambda \, t \, [\text{var}(X_1) + (EX_1)^2].$$

Then by the central limit theorem for random sums, for $u$ fixed and $t \to \infty$,

$$P(S(t) > ES(t) + u \sqrt{\text{var}(S(t))}) \to \overline{F}(u), \quad (2.25)$$
where $\Phi(u) = \int_{-\infty}^{u} \varphi(x) \, dx$ and $\varphi(x) = \exp\{ -0.5 x^2 / 2 \} / \sqrt{2\pi}$. The standard normal tail $\overline{\Phi}(u)$ decays to zero faster than exponentially in $u$. Indeed, the so-called Mill’s ratio (which is obtained by an application of l'Hôpital’s rule to $\overline{\Phi}(u)/(u^{-1} \varphi(u))$) tells us that

$$(2.26) \quad \overline{\Phi}(u) \sim u^{-1} \varphi(u) \quad \text{as } u \to \infty.$$ 

Relation (2.24) gives one a more realistic picture of the far-out tail $P(S(t) > u)$ than the central limit theorem (2.25); the latter approximation is poor in the tails.

References


3 Asymptotic theory for maxima

In this section we start the investigation of the extremes in a sample. In what follows, $X, X_1, X_2, \ldots$ is a sequence of iid non-degenerate random variables with common distribution function $F$. We closely follow Chapter 3 in Embrechts et al. [4].

3.1 Limit probabilities for maxima

We want to investigate the fluctuations of the sample maxima

$$M_1 = X_1, \quad M_n = \max_{i=1,\ldots,n} X_i, \quad n \geq 2.$$ 

Corresponding results for minima can easily be obtained from those for maxima by using the identity

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

Later we will continue with the analysis of the upper order statistics of the sample $X_1, \ldots, X_n$.

There is of course no difficulty in writing down the exact distribution function of the maximum $M_n$:

$$P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = F^n(x), \quad x \in \mathbb{R}, \quad n \geq 1.$$ 

Extremes happen “near” the upper end of the support of the distribution (i.e., that set where the random variables $X_i$ assume their values), hence, intuitively, the asymptotic behavior of $M_n$ must be related to the distribution function $F$ in its right tail near the right endpoint which we denote by

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

We immediately obtain, for all $x < x_F$,

$$P(M_n \leq x) = F^n(x) \to 0, \quad n \to \infty,$$ 

27
and, in the case $x_F < \infty$, we have for $x \geq x_F$ that

$$P(M_n \leq x) = F^n(x) = 1.$$  

Thus $M_n \xrightarrow{P} x_F$ as $n \to \infty$, where $x_F \leq \infty$. (Here and in what follows, $\xrightarrow{P}$ denotes convergence in probability.) Since the sequence $(M_n)$ is non-decreasing in $n$, it converges a.s., and hence we conclude that

$$(3.1) \quad M_n \xrightarrow{a.s.} x_F, \quad n \to \infty.$$  

This fact does not provide a lot of information. More insight into the order of magnitude of maxima is given by weak convergence results for centered and normalized maxima. This is one of the main topics in classical extreme value theory. For instance, the fundamental Fisher-Tippett theorem given below has the following content: if there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that

$$(3.2) \quad c_n^{-1} (M_n - d_n) \xrightarrow{d} H, \quad n \to \infty,$$  

for some non-degenerate distribution $H$, then $H$ must be of the type of one of the three so-called standard extreme value distributions. Here and in what follows, $\xrightarrow{d}$ denotes convergence in distribution. This means that

$$P(c_n^{-1} (M_n - d_n) \leq x) \to H(x) \quad \text{as } n \to \infty$$  

for the distribution function $H$ at all continuity points $x$ of $H$. In what follows, we slightly abuse notation by denoting the distribution $H$, its distribution function and a random variable with distribution $H$ by the same symbol.

Consequently, one has to consider probabilities of the form

$$P \left( c_n^{-1} (M_n - d_n) \leq x \right),$$  

which can be rewritten as

$$(3.3) \quad P(M_n \leq u_n),$$  

where $u_n = u_n(x) = c_n x + d_n$. We first investigate (3.3) for general sequences $(u_n)$, and afterwards come back to affine transformations as in (3.2). We ask:

**Which conditions on $F$ ensure that the limit of $P(M_n \leq u_n)$ for $n \to \infty$ exists for appropriate constants $u_n$?**

It turns out that one needs certain continuity conditions on $F$ at its right endpoint. This rules out many important distributions. For instance, if $F$ has a Poisson distribution, then $P(M_n \leq u_n)$ never has a limit in $(0, 1)$, whatever the sequence $(u_n)$. This implies that the normalized maxima of iid Poisson distributed random variables do not have a non-degenerate limit distribution. This remark might be slightly disappointing, but it shows the crucial difference between sums and maxima. In the former case, the central limit theorem yields the normal distribution as limit under the very general moment condition $EX_1^2 < \infty$. If $EX_1^2 = \infty$ the relatively small class of infinite variance $\alpha$-stable limit distributions for normalized and centered sums enters. Only in that very heavy-tailed case conditions on the tails of the distribution of $X_1$ guarantee the existence of a limit distribution. Thus, in contrast to sums, we always need rather delicate conditions on the tail $F$ to ensure that $P(M_n \leq u_n)$ converges to a non-trivial limit, i.e., a number in $(0, 1)$.  

In what follows we answer the question above. We commence with an elementary result which is crucial for the understanding of the weak limit theory of sample maxima. It will become a standard tool throughout.
Proposition 3.1 (Poisson approximation)

For given $\tau \in [0, \infty]$ and a sequence $(u_n)$ of real numbers the following are equivalent

\begin{align}
(3.4) \quad n \overline{F}(u_n) & \to \tau, \\
(3.5) \quad P(M_n \leq u_n) & \to e^{-\tau}.
\end{align}

Proof. Consider first $0 \leq \tau < \infty$. If (3.4) holds, then

$$P(M_n \leq u_n) = F^n(u_n) = (1 - \overline{F}(u_n))^n = \left(1 - \frac{\tau}{n} + o\left(\frac{1}{n}\right)\right)^n,$$

which implies (3.5). Conversely, if (3.5) holds, then $\overline{F}(u_n) \to 0$. (Otherwise, $\overline{F}(u_n)$ would be bounded away from 0 for some subsequence $(n_k)$. Then $P(M_{n_k} \leq u_{n_k}) = (1 - \overline{F}(u_{n_k}))^{n_k}$ would imply $P(M_{n_k} \leq u_{n_k}) \to 0$.) Taking logarithms in (3.5) we have

$$-n \ln (1 - \overline{F}(u_n)) \to \tau.$$

Since $-\ln(1 - x) \sim x$ for $x \to 0$ this implies that $n\overline{F}(u_n) = \tau + o(1)$, giving (3.4).

If $\tau = \infty$ and (3.4) holds, but (3.5) does not, there must be a subsequence $(n_k)$ such that $P(M_{n_k} \leq u_{n_k}) \to \exp\{-\tau'\}$ as $k \to \infty$ for some $\tau' < \infty$. But then (3.5) implies (3.4), so that $n_k\overline{F}(u_{n_k}) \to \tau' < \infty$, contradicting (3.4) with $\tau = \infty$. Similarly, (3.5) implies (3.4) for $\tau = \infty$. □

Clearly, Poisson’s limit theorem is the key behind the above proof. Indeed, assume for simplicity $0 < \tau < \infty$ and define $B_n = \sum_{i=1}^{n} I_{\{X_i > u_n\}}$. This quantity has a binomial distribution with parameters $(n, \overline{F}(u_n))$. An application of Poisson’s limit theorem yields $B_n \xrightarrow{d} \text{Poisson}(\tau)$ if and only if $EB_n = n\overline{F}(u_n) \to \tau$ which is nothing but (3.4). Also notice that $P(M_n \leq u_n) = P(B_n = 0) \to \exp\{-\tau\}$. This explains why (3.5) is sometimes referred to as Poisson approximation to the probability $P(M_n \leq u_n)$.

Evidently, if there exists a sequence $(u_n^{(\tau)})$ satisfying (3.4) for some fixed $\tau > 0$, then we can find such a sequence for any $\tau > 0$. For instance, if $(u_n^{(1)})$ satisfies (3.4) with $\tau = 1$, $u_n^{(\tau)} = u_n^{(1)}$ obeys (3.4) for an arbitrary $\tau > 0$.

By (3.1), $(M_n)$ converges a.s. to the right endpoint $x_F$ of the distribution function $F$, hence

$$P(M_n \leq x) \to \begin{cases} 0 & \text{if } x < x_F, \\
1 & \text{if } x > x_F. \end{cases}$$

The following result extends this kind of 0-1 behavior.

Corollary 3.2 Suppose that $x_F < \infty$ and

$$\overline{F}(x_F^-) = F(x_F) - F(x_F^-) > 0.$$

Then for every sequence $(u_n)$ such that

$$P(M_n \leq u_n) \to \rho,$$

either $\rho = 0$ or $\rho = 1$. 

29
Proof. Since $0 \leq \rho \leq 1$, we may write $\rho = \exp\{-\tau\}$ with $0 \leq \tau \leq \infty$. By Proposition 3.1 we have $n\overline{F}(u_n) \to \tau$ as $n \to \infty$. If $u_n < x_F$ for infinitely many $n$ we have $\overline{F}(u_n) \geq \overline{F}(x_F- \to 0$ for those $n$ and hence $\tau = \infty$. The other possibility is that $u_n \geq x_F$ for all sufficiently large $n$, giving $n\overline{F}(u_n) = 0$, and hence $\tau = 0$. Thus $\tau = \infty$ or 0, giving $\rho = 0$ or 1.

This result shows in particular that for a distribution function with a jump at its finite right endpoint no non-degenerate limit distribution for $M_n$ exists, whatever the normalization.

A similar result is true for certain distributions with infinite right endpoint as we see from the following characterization, given in Leadbetter et al. [8], Theorem 1.7.13.

**Theorem 3.3** Let $F$ be a distribution function with right endpoint $x_F \leq \infty$ and let $\tau \in (0, \infty)$. There exists a sequence $(u_n)$ satisfying $n\overline{F}(u_n) \to \tau$ if and only if

$$(3.6) \quad \lim_{x \uparrow x_F} \frac{\overline{F}(x)}{\overline{F}(x-)} = 1 \quad \text{and} \quad F(x_F-) = 1.$$  

The result applies in particular to discrete distributions with infinite right endpoint. If the jump heights of the distribution function do not decay sufficiently fast, then a non-degenerate limit distribution for maxima does not exist. For instance, if $X$ is integer-valued and $x_F = \infty$, then (3.6) translates into $\overline{F}(n)/\overline{F}(n - 1) \to 1$ as $n \to \infty$.

These considerations show that some intricate asymptotic behavior of $(M_n)$ exists. The discreteness of a distribution can prevent the maxima from converging and instead forces “oscillatory” behavior. Nonetheless, in this situation it is often possible to find a sequence $(c_n)$ of integers such that $(M_n - c_n)$ is tight; i.e., every subsequence of $(M_n - c_n)$ contains a weakly convergent subsequence. This is true for the examples to follow; see Aldous [1], Section C2, Leadbetter et al. [8], Section 1.7.

**Example 3.4** (Poisson distribution)

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0, \quad \lambda > 0.$$  

Then

$$\overline{F}(k) = 1 - \frac{F(k) - F(k - 1)}{\overline{F}(k - 1)}$$

$$= 1 - \frac{\lambda^k}{k!} \left( \sum_{r=k}^{\infty} \frac{\lambda^r}{r!} \right)^{-1}$$

$$= 1 - \left( 1 + \sum_{r=k+1}^{\infty} \frac{k!}{r!} \lambda^{r-k} \right)^{-1}.$$  

The latter sum can be estimated as

$$\sum_{s=1}^{\infty} \frac{\lambda^s}{(k+1)(k+2) \cdots (k+s)} \leq \sum_{s=1}^{\infty} \left( \frac{\lambda}{k} \right)^s = \frac{\lambda/k}{1 - \lambda/k}, \quad k > \lambda,$$

which tends to 0 as $k \to \infty$, so that $\overline{F}(k)/\overline{F}(k - 1) \to 0$. Theorem 3.3 shows that no non-degenerate limit distribution for maxima exists and, furthermore, that no limit of the form $P(M_n \leq u_n) \to \rho \in (0, 1)$ exists, whatever the sequence of constants $(u_n)$.
Example 3.5 (Geometric distribution)

\[ P(X = k) = p(1 - p)^k, \quad k \in \mathbb{N}_0, \quad 0 < p < 1. \]

In this case, since \( F(k) = (1 - p)^{k+1} \), we have that

\[ \frac{F(k)}{F(k - 1)} = 1 - p. \]

Hence again no limit \( P(M_n \leq u_n) \to \rho \) exists except for \( \rho = 0 \) or 1.

Example 3.6 (Negative binomial distribution)

\[ P(X = k) = \binom{v + k - 1}{k} p^v (1 - p)^k, \quad k \in \mathbb{N}_0, \quad 0 < p < 1, v > 0. \]

For \( v \in \mathbb{N} \) the negative binomial distribution generalizes the geometric distribution in the following sense: the geometric distribution models the waiting time for the first success in a sequence of independent trials, whereas the negative binomial distribution models the waiting time for the \( v \)th success.

Using Stirling’s formula, we obtain

\[ \frac{F(k)}{F(k - 1)} = 1 - p \in (0, 1); \]

i.e., no limit \( P(M_n \leq u_n) \to \rho \) exists except for \( \rho = 0 \) or 1. \( \square \)

Comments

Extreme value theory is a classical topic in probability theory and mathematical statistics. Its origins go back to Fisher and Tippett [6]. Since then a large number of books and articles on extreme value theory has appeared. The interested reader may, for instance, consult the following textbooks: Aldous [1], Beirlant, Teugels and Vynckier [2], Coles [3], Embrechts et al. [4], Falk et al. [5], Gumbel [7], Leadbetter et. al. [8], Reiss [9] and Resnick [10].

References


3.2 Weak convergence of maxima under affine transformations

Now we come back to the characterization of the possible limit laws for the maxima $M_n$ of the iid sequence $(X_n)$ under positive affine transformations.

In this section we answer the question:

What are the possible (non-degenerate) limit laws for the maxima $M_n$
when properly normalized and centered?

This question turns out to be closely related to the following:

Which distributions satisfy for all $n \geq 2$ the identity in law

$$
\max (X_1, \ldots, X_n) \overset{d}{=} c_n X + d_n
$$

for appropriate constants $c_n > 0$ and $d_n \in \mathbb{R}$?

The question is, in other words, which classes of distribution functions $F$ are closed (up to affine transformations) for maxima.

**Definition 3.7** (Max-stable distribution)

A non-degenerate random variable $X$ (the corresponding distribution or distribution function) is called max-stable if it satisfies (3.7) for iid $X, X_1, \ldots, X_n$, appropriate constants $c_n > 0, d_n \in \mathbb{R}$ and every $n \geq 2$.

From now on we refer to the centering constants $d_n$ and the normalizing constants $c_n$ jointly as norming constants.

Assume for the moment that $(X_n)$ is a sequence of iid max-stable random variables. Then (3.7) may be rewritten as follows

$$
c_n^{-1} (M_n - d_n) \overset{d}{=} X.
$$

We conclude that every max-stable distribution is a limit distribution for maxima of iid random variables. Moreover, max-stable distributions are the only limit laws for normalized maxima.

**Theorem 3.8** (Limit property of max-stable laws)

The class of max-stable distributions coincides with the class of all possible (non-degenerate) limit laws for (properly normalized and centered) maxima of iid random variables.

**Proof.** It remains to prove that the limit distribution of affinely transformed maxima is max-stable. Assume that for appropriate norming constants,

$$
\lim_{n \to \infty} F^n (c_n x + d_n) = H(x), \quad x \in \mathbb{R},
$$

32
for some non-degenerate distribution function $H$. We anticipate here (and indeed state precisely in Theorem 3.9) that the possible limit distribution functions $H$ are continuous functions on the whole real line.

Then for every $k \in \mathbb{N}$

$$
\lim_{n \to \infty} F^{n_k} (c_n x + d_n) = \left( \lim_{n \to \infty} F^n (c_n x + d_n) \right)^k = H^k (x), \quad x \in \mathbb{R}.
$$

Furthermore,

$$
\lim_{n \to \infty} F^{n_k} (c_n x + d_n) = H(x), \quad x \in \mathbb{R}.
$$

By the convergence to types theorem (see Theorem 3.17 below) there exist constants $\tilde{c}_k > 0$ and $\tilde{d}_k \in \mathbb{R}$ such that

$$
\lim_{n \to \infty} \frac{c_n}{c_n} = \tilde{c}_k \quad \text{and} \quad \lim_{n \to \infty} \frac{d_n - d_n}{c_n} = \tilde{d}_k,
$$

and for iid random variables $Y_1, \ldots, Y_k$ with distribution function $H$,

$$
\max (Y_1, \ldots, Y_k) \overset{d}{=} \tilde{c}_k Y_1 + \tilde{d}_k. \quad \Box
$$

The following result is the basis of classical extreme value theory.

**Theorem 3.9** (Fisher-Tippett theorem, limit laws for maxima)

Let $(X_n)$ be a sequence of iid random variables. If there exist norming constants $c_n > 0$, $d_n \in \mathbb{R}$ and some non-degenerate distribution function $H$ such that

$$
c_n^{-1} (M_n - d_n) \overset{d}{\to} H,
$$

then $H$ belongs to the type of one of the following three distribution functions:

---

**Figure 3.10** Densities of the standard extreme value distributions. We chose $\alpha = 1$ for the Fréchet and the Weibull distribution.
Fréchet: \[ \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\{-x^{-\alpha}\}, & x > 0 \end{cases} \quad \alpha > 0. \]

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

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

The proof is based on certain functional equations whose solution yields the above limit distributions as the only max-stable ones. See Section 3.2 in [1] for more details.

The limit law in (3.9) is unique only up to affine transformations. If the limit appears as \( H(cx + d) \), i.e.,
\[
\lim_{n \to \infty} P \left( c_n^{-1} (M_n - d_n) \leq x \right) = H(cx + d),
\]
then \( H(x) \) is also a limit under a simple change of norming constants:
\[
\lim_{n \to \infty} P \left( \tilde{c}_n^{-1} (M_n - \tilde{d}_n) \leq x \right) = H(x)
\]
with \( \tilde{c}_n = c_n/c \) and \( \tilde{d}_n = d_n - dc_n/c \). The convergence to types theorem (see Theorem 3.2.1 below) shows precisely how affine transformations, weak convergence and types are related.

In Tables 2.10 and 2.12 we defined a Weibull distribution function for \( c, \alpha > 0 \). For \( c = 1 \) it is given by
\[
F_\alpha(x) = 1 - e^{-x^\alpha}, \quad x \geq 0,
\]
which is the distribution function of a positive random variable. The Weibull distribution \( \Psi_\alpha \), as a limit distribution for maxima, is concentrated on \( (-\infty, 0) \):
\[
\Psi_\alpha(x) = 1 - F_\alpha(-x), \quad x < 0.
\]

In the context of extreme value theory we follow the convention and refer to \( \Psi_\alpha \) as the Weibull distribution. We hope to avoid any confusion by a clear distinction between the two distributions whose extremal behaviour is completely different.

Though, for modeling purposes, the types of \( \Lambda \), \( \Phi_\alpha \) and \( \Psi_\alpha \) are very different, from a mathematical point of view they are closely linked. Indeed, one immediately verifies the following properties. Suppose \( X > 0 \), then
\[
X \text{ has distribution function } \Phi_\alpha \text{ if and only if } \quad \ln X^\alpha \text{ has distribution function } \Lambda \quad \text{if and only if}
\]
\[
-X^{-1} \text{ has distribution function } \Psi_\alpha.
\]

These relationships will appear again and again in various disguises in the sequel.

**Definition 3.12** (Extreme value distribution and extremal random variable)
The distribution functions \( \Phi_\alpha \), \( \Psi_\alpha \) and \( \Lambda \) as presented in Theorem 3.9 are called standard extreme value distributions, the corresponding random variables standard extremal random variables. Distribution functions of the types of \( \Phi_\alpha \), \( \Psi_\alpha \) and \( \Lambda \) are extreme value distributions; the corresponding random variables extremal random variables.

\[\square\]
Figure 3.11 Evolution of the maxima $M_n$ of standard exponential (top) and Cauchy (bottom) samples. A sample path of $(M_n)$ has a jump whenever $X_n > M_{n-1}$ (we say that $M_n$ is a record). The graph seems to suggest that there occur more records for the exponential than for the Cauchy random variables. However, the distribution of the number of record times is approximately the same in both cases as can be shown by theoretical means. The qualitative differences in the two graphs are due to a few large jumps for Cauchy distributed variables. Compared with those the smaller jumps are so tiny that they “disappear” from the computer graph; notice the difference between the vertical scales.

By Theorem 3.8, the extreme value distributions are precisely the max-stable distributions. Hence if $X$ is an extremal random variable it satisfies (3.8). In particular, the three cases in Theorem 3.9 correspond to

- **Fréchet**: $M_n \overset{d}{=} n^{1/\alpha} X$
- **Weibull**: $M_n \overset{d}{=} n^{-1/\alpha} X$
- **Gumbel**: $M_n \overset{d}{=} X + \ln n$.

**Example 3.13** (Maxima of exponential random variables)

See also Figures 3.11 and 3.15. Let $(X_i)$ be a sequence of iid standard exponential random variables. Then

$$P(M_n - \ln n \leq x) = (P(X \leq x + \ln n))^n = (1 - e^{-x})^n \to \exp\{\exp(-e^{-x})\} = \Lambda(x), \quad x \in \mathbb{R}.$$ 

For comparison recall that for iid Gumbel random variables $X_i$,

$$P(M_n - \ln n \leq x) = \Lambda(x), \quad x \in \mathbb{R}. \qed$$

35
Figure 3.15 The distribution function \( P(M_n - \ln n \leq x) \) for \( n \) iid standard exponential random variables and the Gumbel distribution function (top). In the bottom figure the relative error \( P(M_n - \ln n > x)/\bar{F}(x) - 1 \) of this approximation is illustrated.

Example 3.14 (Maxima of Cauchy random variables)
See also Figures 3.11 and 3.16. Let \((X_i)\) be a sequence of iid standard Cauchy random variables. The standard Cauchy distribution has density

\[
f(x) = \left(\pi \left(1 + x^2\right)\right)^{-1}, \quad x \in \mathbb{R}.
\]

By l'Hospital’s rule we obtain

\[
\lim_{x \to \infty} \frac{\bar{F}(x)}{(\pi x)^{-1}} = \lim_{x \to \infty} \frac{f(x)}{\pi \left(1 + x^2\right)} = \lim_{x \to \infty} \frac{\pi x^2}{\pi \left(1 + x^2\right)} = 1,
\]
Figure 3.16 The distribution function $P(\pi n^{-1} M_n \leq x)$ of $n$ iid standard Cauchy random variables and the Fréchet distribution function $\Phi_1$ (top). In the bottom figure the relative error $P(\pi n^{-1} M_n > x)/\Phi_1(x) - 1$ of this approximation is illustrated.

giving $\overline{F}(x) \sim (\pi x)^{-1}$. This implies

$$P \left( M_n \leq \frac{n x}{\pi} \right) = \left( 1 - \frac{nx}{\pi} \right)^n$$

$$= \left( 1 - \frac{1}{n x} + o(1/n) \right)^n$$

$$\rightarrow \exp \{-x^{-1}\} = \Phi_1(x), \quad x > 0. \quad \Box$$

### 3.2.1 Appendix: Convergence to types

We say that the random variables $X$ and $Y$ (and the corresponding distributions and distribution functions) belong to the same type or are of the same type if there exist constants $a \in \mathbb{R}$ and $b > 0$
such that

\[ X \overset{d}{=} bY + a. \]

The following result is a particularly important tool for weak convergence. It tells us that the limit law of a sequence of random variables is uniquely determined up to changes of location and scale. For a proof we refer to Resnick [5].

**Theorem 3.17** (Convergence to types theorem)

Let \( A, B, A_1, A_2, \ldots \) be random variables and \( b_n > 0, \beta_n > 0 \) and \( a_n, \alpha_n \in \mathbb{R} \) be constants. Suppose that

\[ b_n^{-1}(A_n - a_n) \overset{d}{\to} A. \]

Then the relation

\[ (3.10) \quad \beta_n^{-1}(A_n - a_n) \overset{d}{\to} B \]

holds if and only if

\[ (3.11) \quad \lim_{n \to \infty} b_n/\beta_n = b \in [0, \infty), \quad \lim_{n \to \infty} (a_n - \alpha_n)/\beta_n = a \in \mathbb{R}. \]

If (3.10) holds then \( B \overset{d}{=} bA + a \) and \( a, b \) are the unique constants for which this holds.

When (3.10) holds, \( A \) is non-degenerate if and only if \( b > 0 \), and then \( A \) and \( B \) belong to the same type. \( \square \)

It is immediate from (3.11) that the constants \( a_n \) and \( b_n \) are uniquely determined only up to the asymptotic relation (3.11).

**Comments**

Theorem 3.9 marked the beginning of extreme value theory as one of the central topics in probability theory and statistics. The limit laws for maxima were derived by Fisher and Tippett [2]. A first rigorous proof is due to Gnedenko [3]. De Haan [4] subsequently applied regular variation as an analytical tool. His work has been of great importance for the development of modern extreme value theory.

**References**


3.3 Maximum domains of attraction and norming constants

In the preceding section we identified the extreme value distributions as the limit laws for normalized maxima of iid random variables; see Theorem 3.9. This section is devoted to the question:

Given an extreme value distribution \( H \), what conditions on the distribution function \( F \) imply that the normalized maxima \( M_n \) converge weakly to \( H \)?

Closely related to this question is the following:

How may we choose the norming constants \( c_n > 0 \) and \( d_n \in \mathbb{R} \) such that

\[
(3.12) \quad c_n^{-1} (M_n - d_n) \xrightarrow{d} H.
\]

Can it happen that different norming constants imply convergence to different limit laws?

The last question can be answered immediately: the convergence to types theorem (see Theorem 3.17) ensures that the limit law is uniquely determined up to affine transformations.

**Definition 3.18** (Maximum domain of attraction)

We say that the random variable \( X \) (the distribution function \( F \) of \( X \), the distribution of \( X \)) belongs to the maximum domain of attraction of the extreme value distribution \( H \) if there exist constants \( c_n > 0 \), \( d_n \in \mathbb{R} \) such that (3.12) holds. We write \( X \in \text{MDA}(H) \) (\( F \in \text{MDA}(H) \)).

Notice that the extreme value distribution functions are continuous on \( \mathbb{R} \), hence \( c_n^{-1} (M_n - d_n) \xrightarrow{d} H \) is equivalent to

\[
\lim_{n \to \infty} P (M_n \leq c_n x + d_n) = \lim_{n \to \infty} F^n (c_n x + d_n) = H(x), \quad x \in \mathbb{R}.
\]

The following result is an immediate consequence of Proposition 3.1 and will be used throughout the following sections.

**Proposition 3.19** (Characterization of MDA(\( H \))

The distribution function \( F \) belongs to the maximum domain of attraction of the extreme value distribution \( H \) with norming constants \( c_n > 0 \), \( d_n \in \mathbb{R} \) if and only if

\[
\lim_{n \to \infty} nF(c_n x + d_n) = -\ln H(x), \quad x \in \mathbb{R}.
\]

When \( H(x) = 0 \) the limit is interpreted as \( \infty \).

For every standard extreme value distribution we characterize its maximum domain of attraction. Using the concept of regular variation this is not too difficult for the Fréchet distribution \( \Phi_\alpha \) and the Weibull distribution \( \Phi_\alpha \); see Sections 3.3.1 and 3.3.2. The maximum domain of attraction of the Gumbel distribution \( \Lambda \) is not so easily characterized; it consists of distributions whose right tail decreases to zero faster than any power function. This will be made precise in Section 3.3.3. If \( F \) has a density, simple sufficient conditions for \( F \) to be in the maximum domain of attraction of some extreme value distribution are due to von Mises. We present them below for practical (and historical) reasons.

The following concept defines an equivalence relation on the set of all distribution functions.
Definition 3.20 (Tail-equivalence)
Two distribution functions $F$ and $G$ are called tail-equivalent if they have the same right endpoint, i.e., if $x_F = x_G$, and
\[
\lim_{x \uparrow x_F} \frac{F(x)}{G(x)} = c
\]
for some constant $0 < c < \infty$.

We show that every maximum domain of attraction is closed with respect to tail-equivalence, i.e., for tail-equivalent $F$ and $G$, $F, G \in \text{MDA}(H)$ if and only if $G \in \text{MDA}(H)$. Moreover, for any two tail-equivalent distributions one can take the same norming constants. This will prove to be of great help for calculating norming constants which, in general, can become a rather tedious procedure.

Theorem 3.8 identifies the max-stable distributions as limit laws for affinely transformed maxima of iid random variables. The sample maximum $M_n$ is the empirical version of the $(1 - n^{-1})$-quantile of the underlying distribution function $F$. Therefore the latter is an appropriate centering constant.

3.3.1 The maximum domain of attraction of the Fréchet distribution $\Phi_\alpha(x) = e^{-x^{-\alpha}}$

In this section we characterize the maximum domain of attraction of $\Phi_\alpha$ for $\alpha > 0$. By Taylor expansion,
\[
1 - \Phi_\alpha(x) = 1 - \exp \left\{ -x^{-\alpha} \right\} \sim x^{-\alpha}, \quad x \to \infty,
\]
hence the tail of $\Phi_\alpha$ decreases like a power law. We ask:

How far away can we move from a power tail and still remain in MDA($\Phi_\alpha$)?

We show that the maximum domain of attraction of $\Phi_\alpha$ consists of distribution functions $F$ whose right tail is regularly varying with index $-\alpha$. For $F \in \text{MDA}(\Phi_\alpha)$ the constants $d_n$ can be chosen as $0$ (centering is not necessary) and the $c_n$ by means of the quantile function, more precisely by
\[
c_n = F^c((1 - n^{-1}) \quad = \quad \inf \left\{ x \in \mathbb{R} : F(x) \geq 1 - n^{-1} \right\}
\]
(3.13)
\[
= \inf \left\{ x \in \mathbb{R} : \left( \frac{1}{F} \right)(x) \geq n \right\}
= \left( \frac{1}{F} \right)^{-c}(n).
\]

Theorem 3.21 (Maximum domain of attraction of $\Phi_\alpha$)
The distribution function $F$ belongs to the maximum domain of attraction of $\Phi_\alpha$, $\alpha > 0$, if and only if $F(x) = x^{-\alpha}L(x)$ for some slowly varying function $L$.

If $F \in \text{MDA}(\Phi_\alpha)$, then
\[
c_n^{-1} M_n \xrightarrow{d} \Phi_\alpha,
\]
where the norming constants $c_n$ can be chosen according to (3.13).

Notice that this result implies in particular that every $F \in \text{MDA}(\Phi_\alpha)$ has an infinite right endpoint $x_F = \infty$. Furthermore, the norming constants $c_n$ form a regularly varying sequence, more precisely, $c_n = n^{1/\alpha}L_1(n)$ for some slowly varying function $L_1$.

Proof. Let $F$ be regularly varying with index $-\alpha < 0$. By the choice of $c_n$ and regular variation,
\[
\bar{F}(c_n) \sim n^{-1}, \quad n \to \infty,
\]
(3.15)
and hence $F(c_n) \to 0$ giving $c_n \to \infty$. For $x > 0$,

$$nF(c_n, x) \sim \frac{F(c_n^x)}{F(c_n)} \to x^{-\alpha}, \quad n \to \infty.$$  

For $x < 0$, immediately $F^n(c_n, x) \leq F^n(0) \to 0$, since regular variation requires $F(0) < 1$. By Proposition 3.19, $F \in \text{MDA}(\Phi_\alpha)$.

We only sketch the converse proof. Assume that $\lim_{n \to \infty} F^n(c_n, x + d_n) = \Phi_\alpha(x)$ for all $x > 0$ and appropriate $c_n > 0$, $d_n \in \mathbb{R}$. This leads to

$$\lim_{n \to \infty} F^n(c_{n+1}, x + d_{n+1}) = \Phi_{1/s}(x) = \Phi_\alpha(s^{-1/\alpha} x), \quad s > 0, \ x > 0.$$  

By the convergence to types theorem (Theorem 3.17)

$$c_{n+1}/c_n \to s^{1/\alpha} \quad \text{and} \quad (d_{n+1} - d_n)/c_n \to 0.$$  

We can extend $(c_n)$ to a positive function $c(x)$ by setting $c(x) = c_{[x]}$. Then $c(x)$ is a regularly varying function with positive index, in particular $c_n \to \infty$; for details see Section 1.9 in Bingham et al. [1]. Assume first that $d_n = 0$, then $nF(c_n, x) \to x^{-\alpha}$, in particular

$$\frac{nF(c_n, x)}{nF(c_n)} = \frac{F(c_n, x)}{F(c_n)} \to x^{-\alpha}.$$  

It remains to replace $c_n$ by an arbitrary function $g(y) \to \infty$ as $y \to \infty$, i.e.,

$$\frac{F(g(y), x)}{F(g(y))} \to x^{-\alpha}.$$  

This follows from regular variation of $c(x)$; see Proposition A3.8(a) in [2]. The case $d_n \neq 0$ is more involved, indeed one has to show that $d_{n+1}/c_n \to 0$. If the latter holds one can repeat the above argument by replacing $d_n$ by 0. For details on this, see Bingham et al. [1], Theorem 8.13.2.

We have found the answer to the above question:

$$F \in \text{MDA}(\Phi_\alpha) \iff \frac{F}{F(x)} \quad \text{is regularly varying with index } -\alpha.$$  

Thus we have a simple characterization of MDA$(\Phi_\alpha)$. Notice that this class of distribution functions contains “very heavy-tailed distributions”. Indeed, regular variation of $F$ with index $-\alpha$ implies that

$$E(X^+)^{\delta} = \int_0^\infty P(X^+ > x) \, dx = \int_0^\infty F(x^{1/\delta}) \, dx \quad \left\{ \begin{array}{ll} = \infty & \delta > \alpha, \\ < \infty & \delta < \alpha. \end{array} \right.$$  

Thus they may be appropriate distributions for modeling large insurance claims and large fluctuations of prices, log-returns, etc.

Von Mises found some easily verifiable conditions on the density of a distribution for it to belong to some maximum domain of attraction ([2], Corollary 3.3.8).

**Corollary 3.22** (Von Mises condition)

*Let $F$ be a distribution function with density $f$ satisfying*

$$(3.16) \quad \lim_{x \to \infty} \frac{x f(x)}{F(x)} = \alpha > 0,$$

*then $F \in \text{MDA}(\Phi_\alpha)$.*
The class of distribution functions \( F \) with regularly varying tail \( \overline{F} \) is obviously closed with respect to tail-equivalence (Definition 3.20). The following result gives us some insight into the structure of \( \text{MDA}(\Phi_\alpha) \). Besides this theoretical aspect, it will turn out to be a useful tool for calculating norming constants.

**Proposition 3.23** (Closure property of \( \text{MDA}(\Phi_\alpha) \))

Let \( F \) and \( G \) be distribution functions and assume that \( F \in \text{MDA}(\Phi_\alpha) \) with norming constants \( c_n > 0 \), i.e.,

\[
\lim_{n \to \infty} F^n(c_n x) = \Phi_\alpha(x), \quad x > 0.
\]

Then

\[
\lim_{n \to \infty} G^n(c_n x) = \Phi_\alpha(cx), \quad x > 0,
\]

for some \( c > 0 \) if and only if \( F \) and \( G \) are tail-equivalent with

\[
\lim_{x \to \infty} \overline{F}(x)/\overline{G}(x) = e^\alpha.
\]

**Proof of the sufficiency.** For the necessity part see Resnick [5], Proposition 1.19. Suppose that \( \overline{F}(x) \sim q \overline{G}(x) \) as \( x \to \infty \) for some \( q > 0 \). By Proposition 3.19 the limit relation (3.17) is equivalent to

\[
\lim_{n \to \infty} n \overline{F}(c_n x) = x^{-\alpha}
\]

for all \( x > 0 \). For such \( x \), \( c_n x \to \infty \) as \( n \to \infty \) and hence, by tail-equivalence,

\[
n \overline{G}(c_n x) \sim n q^{-1} \overline{F}(c_n x) \to q^{-1} x^{-\alpha},
\]

i.e., again by Proposition 3.19,

\[
\lim_{n \to \infty} G^n(c_n x) = \exp \left\{ - \left( q^{1/\alpha} x \right)^{-\alpha} \right\} = \Phi_\alpha \left( q^{1/\alpha} x \right).
\]

Now set \( c = q^{1/\alpha} \).

By Theorem 3.21, \( F \in \text{MDA}(\Phi_\alpha) \) if and only if \( \overline{F} \) is regularly varying with index \( -\alpha < 0 \). The representation theorem for regularly varying functions (see Section 2.2.4) implies that every \( F \in \text{MDA}(\Phi_\alpha) \) is tail-equivalent to an distribution function with density satisfying (3.16). We can summarize this as follows:

\[
\text{MDA}(\Phi_\alpha) \text{ consists of distribution functions satisfying the von Mises condition (3.16) and their tail-equivalent distribution functions.}
\]

We conclude this section with some examples.

**Example 3.24** (Pareto-like distributions)

- Pareto
- Cauchy
- Burr
- Stable with exponent \( \alpha < 2 \).

The respective densities or distribution functions are given in Table 2.12. All these distributions are Pareto-like in the sense that their right tails are of the form

\[
\overline{F}(x) \sim K x^{-\alpha}, \quad x \to \infty,
\]

42
for some $K$, $\alpha > 0$. Obviously $F$ is regularly varying with some index $-\alpha < 0$ which implies that $F \in \text{MDA}(\Phi_\alpha)$ and as norming constants we can choose $c_n = (Kn)^{1/\alpha}$; see Theorem 3.21. Then

$$(Kn)^{-1/\alpha} M_n \xrightarrow{d} \Phi_\alpha.$$ 

The Cauchy distribution was treated in detail in Example 3.14.

**Example 3.25** (Loggamma distribution)

The loggamma distribution has tail

$$P(x) \sim \frac{\alpha^{\beta-1}}{\Gamma(\beta)} (\ln x)^{\beta-1} x^{-\alpha}, \quad x \to \infty, \quad \alpha, \beta > 0.$$ 

Hence $F \in \text{MDA}(\Phi_\alpha)$. According to Proposition 3.23 we choose $c_n$ by means of the tail-equivalent right-hand side of (3.18). On applying (3.13) and taking logarithms we find we have to solve

$$(3.19) \quad \alpha \ln c_n - (\beta - 1) \ln(\ln c_n - \ln(\alpha^{\beta-1}/\Gamma(\beta))) = \ln n.$$ 

The solution satisfies

$$\ln c_n = \alpha^{-1} (\ln n + \ln r_n),$$

where $\ln r_n = o(\ln n)$ as $n \to \infty$. We substitute this into equation (3.19) and obtain

$$\ln r_n = (\beta - 1) \ln(\alpha^{-1} \ln n (1 + o(1))) + \ln \left(\alpha^{\beta-1}/\Gamma(\beta)\right).$$

This gives the norming constants

$$c_n \sim \left((\Gamma(\beta))^{-1} (\ln n)^{\beta-1} n\right)^{1/\alpha}.$$ 

Hence

$$\left((\Gamma(\beta))^{-1} (\ln n)^{\beta-1} n\right)^{-1/\alpha} M_n \xrightarrow{d} \Phi_\alpha.$$ 

### 3.3.2 The maximum domain of attraction of the Weibull distribution $\Psi_\alpha(x) = e^{-(x^{-\alpha})}$

In this section we characterize the maximum domain of attraction of $\Psi_\alpha$ for $\alpha > 0$. An important, though not at all obvious fact is that all distribution functions $F$ in $\text{MDA}(\Psi_\alpha)$ have finite right endpoint $x_F$. As was already indicated on p. 34, $\Psi_\alpha$ and $\Phi_\alpha$ are closely related, indeed

$$\Psi_\alpha(-x^{-1}) = \Phi_\alpha(x), \quad x > 0.$$ 

Therefore we may expect that also $\text{MDA}(\Psi_\alpha)$ and $\text{MDA}(\Phi_\alpha)$ will be closely related. The following theorem confirms this.

**Theorem 3.26** (Maximum domain of attraction of $\Psi_\alpha$)

*The distribution function $F$ belongs to the maximum domain of attraction of $\Psi_\alpha$, $\alpha > 0$, if and only if $x_F < \infty$ and $P(x_F - x^{-1}) = x^{\alpha - 1} L(x)$ for some slowly varying function $L$. If $F \in \text{MDA}(\Psi_\alpha)$, then

$$(3.20) \quad c_n^{-1} (M_n - x_F) \xrightarrow{d} \Psi_\alpha,$$

where the norming constants $c_n$ can be chosen as $c_n = x_F - F^{-1}(1 - n^{-1})$ and $d_n = x_F$. 

43
We omit the proof. Hence,

\[
F \in \text{MDA}(\Psi_\alpha) \quad \text{if and only if } x_F < \infty \text{ and } \\
\overline{F}(x_F - x^{-1}) \text{ is regularly varying with index } -\alpha < 0.
\]

Thus MDA(\Psi_\alpha) consists of distribution functions \( F \) with support bounded to the right. They may not be the best choice for modeling extremal events in insurance and finance, precisely because \( x_F < \infty \). Though clearly in all circumstances in practice there is a (perhaps ridiculously high) upper limit, we may not want to incorporate this extra parameter \( x_F \) in our model. Often distributions with \( x_F = \infty \) should be preferred since they allow for arbitrarily large values in a sample. Such distributions typically belong to MDA(\Psi_\alpha) or MDA(\Lambda).

In the previous section we found it convenient to characterize membership in MDA(\Psi_\alpha) via the density of a distribution function; see Corollary 3.22.

**Corollary 3.27** (Von Mises condition)

*Let \( F \) be a distribution function with density \( f \) which is positive on some finite interval \((z, x_F)\). If*

\[
\lim_{x \uparrow x_F} \frac{(x_F - x) f(x)}{\overline{F}(x)} = \alpha > 0,
\]

*then \( F \in \text{MDA}(\Psi_\alpha) \). \qed

**Proposition 3.28** (Closure property of MDA(\Psi_\alpha))

*Let \( F \) and \( G \) be distribution functions with right endpoints \( x_F = x_G < \infty \) and assume that \( F \in \text{MDA}(\Psi_\alpha) \) with norming constants \( c_n > 0 \); i.e.,

\[
\lim_{n \to \infty} F^n(c_n x + x_F) = \Psi_\alpha(x), \quad x < 0.
\]

*Then*

\[
\lim_{n \to \infty} G^n(c_n x + x_F) = \Psi_\alpha(c x), \quad x < 0,
\]

*for some \( c > 0 \) if and only if \( F \) and \( G \) are tail-equivalent with*

\[
\lim_{x \uparrow x_F} \frac{\overline{F}(x)}{\overline{G}(x)} = c^{-\alpha}.
\]

\qed

The representation theorem for regularly varying functions implies that every \( F \in \text{MDA}(\Psi_\alpha) \) is tail-equivalent to a distribution function with density satisfying (3.21). We summarize this as follows:

MDA(\Psi_\alpha) consists of distribution functions satisfying the von Mises condition (3.21) and their tail-equivalent distribution functions.

We conclude this section with some examples of prominent MDA(\Psi_\alpha)-members.

**Example 3.29** (Uniform distribution on \((0, 1)\))

Obviously, \( x_F = 1 \) and \( \overline{F}(1 - x^{-1}) = x^{-1} \) is regularly varying with index \(-\alpha = -1\). Then by Theorem 3.26 we obtain \( F \in \text{MDA}(\Psi_1) \). Since \( \overline{F}(1 - n^{-1}) = n^{-1} \), we choose \( c_n = n^{-1} \). This implies in particular

\[
n(M_n - 1) \overset{d}{\to} \Psi_1.
\]

\qed
Example 3.30 (Power law behavior at the finite right endpoint)
Let $F$ be a distribution function with finite right endpoint $x_F$ and distribution tail

$$F(x) = K (x_F - x)^\alpha, \quad x_F - K^{-1/\alpha} \leq x \leq x_F, \quad K, \alpha > 0.$$ 

By Theorem 3.26 this ensures that $F \in \text{MDA} (\Psi_\alpha)$. The norming constants $c_n$ can be chosen such that $F(x_F - c_n) = n^{-1}$, i.e., $c_n = (n K)^{-1/\alpha}$ and, in particular,

$$(nK)^{1/\alpha} \left( M_n - x_F \right) \rightarrow \Psi_\alpha. \quad \square$$

Example 3.31 (Beta distribution)
The beta distribution has density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1, \quad a, b > 0.$$ 

Notice that $f(1 - x^{-1})$ is regularly varying with index $-(b - 1)$ and hence, by Karamata’s theorem (see (2.14) and (2.15) on p. 19),

$$F(1 - x^{-1}) = \int_{1-x^{-1}}^{1} f(y) dy = \int_{x}^{\infty} f(1 - y^{-1})y^{-2} dy \sim x^{-1}f(1 - x^{-1}).$$

Hence $F(1 - x^{-1})$ is regularly varying with index $-b$ and

$$F(x) \sim \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)} (1 - x)^{b}, \quad x \uparrow 1.$$ 

Thus the beta distribution function is tail-equivalent to a distribution function with power law behavior at $x_F = 1$. By Proposition 3.28 the norming constants can be determined by this power law tail which fits into the framework of Example 3.30 above. \quad \square

3.3.3 The maximum domain of attraction of the Gumbel distribution $\Lambda(x) = e^{-e^{-x}}$

Von Mises functions

The maximum domain of attraction of the Gumbel distribution $\Lambda$ covers a wide range of distribution functions $F$. Although there is no direct link with regular variation as in the maximum domains of attraction of the Fréchet and Weibull distribution, we will find extensions of regular variation which allow for a complete characterization of MDA($\Lambda$).

A Taylor expansion argument yields

$$1 - \Lambda(x) \sim e^{-x}, \quad x \rightarrow \infty,$$

hence $\Lambda(x)$ decreases to zero at an exponential rate. Again the following question naturally arises:

How far away can we move from an exponential tail and still remain in MDA($\Lambda$)?

We will see in the present and the next section that MDA($\Lambda$) contains distribution functions with very different tails, ranging from moderately heavy (such as the lognormal distribution) to light (such as the normal distribution). Also both cases $x_F < \infty$ and $x_F = \infty$ are possible. Before we give a general answer to the above question, we restrict ourselves to some $F \in \text{MDA}(\Lambda)$ with a density which have a simple representation, proposed by von Mises. These distributions provide an important building block of this maximum domain of attraction, and therefore we study them in detail. We will see later (Theorem 3.40 and Remark 4 after Theorem 3.40) that one only has to consider a slight modification of the von Mises functions in order to characterize MDA($\Lambda$) completely.
**Definition 3.32** (Von Mises function) Let $F$ be a distribution function with right endpoint $x_F \leq \infty$. Suppose there exists some $z < x_F$ such that $F$ has representation

\[
\bar{F}(x) = c \exp \left\{ -\int_z^x \frac{1}{a(t)} \, dt \right\}, \quad z < x < x_F, \tag{3.22}
\]

where $c$ is some positive constant, $a(\cdot)$ is a positive function with density $a'$ and $\lim_{x \uparrow x_F} a'(x) = 0$. Then $F$ is called a von Mises function, the function $a(\cdot)$ the auxiliary function of $F$. \hfill \Box

**Remark.** 1) Relation (3.22) should be compared with the Karamata representation (2.11) of a regularly varying function. Substituting into (3.22) the function $a(x) = x/\delta(x)$ such that $\delta(x) \to \alpha \in [0, \infty)$ as $x \to \infty$, (3.22) becomes a regularly varying tail with index $-\alpha$. We will see later (see Remark 2 below) that the auxiliary function of a von Mises function with $x_F = \infty$ satisfies $a(x)/x \to 0$. It immediately follows that $\bar{F}(x)$ decreases to zero much faster than any power law $x^{-\alpha}$.

We give some examples of von Mises functions.

**Example 3.33** (Exponential distribution)

\[
\bar{F}(x) = e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0.
\]

$F$ is a von Mises function with auxiliary function $a(x) = \lambda^{-1}$. \hfill \Box

**Example 3.34** (Weibull distribution)

\[
\bar{F}(x) = \exp \{-c x^\tau\}, \quad x \geq 0, \quad c, \tau > 0.
\]

$F$ is a von Mises function with auxiliary function

\[
a(x) = c^{-1} \tau^{-1} x^{1-\tau}, \quad x > 0.\]

\hfill \Box

**Example 3.35** (Erlang distribution)

\[
\bar{F}(x) = e^{-\lambda x} \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}, \quad x \geq 0, \quad \lambda > 0, n \in \mathbb{N}.
\]

$F$ is a von Mises function with auxiliary function

\[
a(x) = \lambda^{-1} x^{n-1}, \quad x > 0.
\]

Notice that $F$ is the $\Gamma(n, \lambda)$ distribution function. \hfill \Box

**Example 3.36** (Exponential behavior at the finite right endpoint) Let $F$ be a distribution function with finite right endpoint $x_F$ and distribution tail

\[
\bar{F}(x) = K \exp \left\{ -\frac{\alpha}{x_F - x} \right\}, \quad x < x_F, \quad \alpha, K > 0.
\]

46
$F$ is a von Mises function with auxiliary function

$$a(x) = \frac{(x_F - x)^2}{\alpha}, \quad x < x_F.$$ 

For $x_F = 1$, $\alpha = 1$ and $K = e$ we obtain for example

$$\overline{F}(x) = \exp \left\{ - \frac{x}{1 - x} \right\}, \quad 0 \leq x < 1. \quad \square$$

**Example 3.37** (Differentiability at the right endpoint)

Let $F$ be a distribution function with right endpoint $x_F < \infty$ and assume there exists some $z < x_F$ such that $F$ is twice differentiable on $(z, x_F)$ with positive density $f = F'$ and $F''(x) < 0$ for $z < x < x_F$. Then it is not difficult to see that $F$ is a von Mises function with auxiliary function $a = \overline{F}/f$ if and only if

$$\lim_{x \uparrow x_F} \frac{F'(x) F''(x)}{f^2(x)} = -1. \tag{3.23}$$

Indeed, let $z < x < x_F$ and set $Q(x) = -\ln \overline{F}(x)$ and $a(x) = 1/Q'(x) = \overline{F}(x)/f(x) > 0$. Hence $F$ has representation (3.22). Furthermore,

$$a'(x) = -\frac{\overline{F}(x) F''(x)}{f^2(x)} - 1$$

and (3.23) is equivalent to $a'(x) \to 0$ as $x \uparrow x_F$.

Condition (3.23) applies to many distributions of interest, including the normal distribution; see Example 3.43. \square

In Remark 1 above we gained some indication that regular variation does not seem to be the right tool for describing von Mises functions. The tail function $\overline{F}$ is said to be *rapidly varying* if

$$\lim_{x \to \infty} \frac{\overline{F}(xt)}{\overline{F}(x)} = \begin{cases} 0 & \text{if } t > 1, \\ \infty & \text{if } 0 < t < 1. \end{cases}$$

Some of the important results for regularly varying functions can be extended to rapidly varying functions in a natural way; see Theorem A3.12 in [2].

**Proposition 3.38** (Properties of von Mises functions)

Every von Mises function $F$ has a positive density $f$ on $(z, x_F)$. The auxiliary function can be chosen as $a(x) = \overline{F}(x)/f(x)$. Moreover, the following properties hold.

1. If $x_F = \infty$, then $\overline{F}$ is rapidly varying and

$$\lim_{x \to \infty} \frac{x f(x)}{\overline{F}(x)} = \infty. \quad \tag{3.24}$$

2. If $x_F < \infty$, then $\overline{F}(x_F - x^{-1})$ is rapidly varying and

$$\lim_{x \uparrow x_F} \frac{(x_F - x)f(x)}{\overline{F}(x)} = \infty. \quad \tag{3.25}$$

47
Remarks. 2) It follows from (3.24) that \( \lim_{x \to \infty} x^{-1} a(x) = 0 \), and from (3.25) that \( a(x) = o(x_F - x) = o(1) \) as \( x \uparrow x_F \).

3) Note that \( a^{-1}(x) = f(x)/\overline{F}(x) \) is the hazard rate of \( F \).

For the proof, see p. 141 in [2].

Now we can show that von Mises functions belong to the maximum domain of attraction of the Gumbel distribution. Moreover, the specific form of \( \overline{F} \) allows one to calculate the norming constants \( c_n \) from the auxiliary function.

**Proposition 3.39** (Von Mises functions and MDA(\( \Lambda \)))

Suppose the distribution function \( F \) is a von Mises function. Then \( F \in \text{MDA}(\Lambda) \). A possible choice of norming constants is

\[
(3.26) \quad d_n = F^{-} (1 - n^{-1}) \quad \text{and} \quad c_n = a(d_n),
\]

where \( a \) is the auxiliary function of \( F \).

**Proof.** Representation (3.22) implies for \( t \in \mathbb{R} \) and \( x \) sufficiently close to \( x_F \) that

\[
\frac{\overline{F}(x + t a(x))}{\overline{F}(x)} = \exp \left\{ - \int_x^{x+ta(x)} \frac{1}{a(u)} \, du \right\}.
\]

We set \( v = (u - x)/a(x) \) and obtain

\[
(3.27) \quad \frac{\overline{F}(x + t a(x))}{\overline{F}(x)} = \exp \left\{ - \int_0^t \frac{a(x)}{a(x + va(x))} \, dv \right\}.
\]

We show that the integrand converges locally uniformly to 1. For given \( \varepsilon > 0 \) and \( x \geq x_0(\varepsilon) \),

\[
|a(x + va(x)) - a(x)| = \left| \int_x^{x+va(x)} a'(s) \, ds \right| = \varepsilon |v| a(x) \leq \varepsilon |t| a(x),
\]

where we used \( a'(x) \to 0 \) as \( x \uparrow x_F \). This implies for \( x \geq x_0(\varepsilon) \) that

\[
\left| \frac{a(x + va(x))}{a(x)} - 1 \right| \leq \varepsilon |t|.
\]

The right-hand side can be made arbitrarily small, hence

\[
(3.28) \quad \lim_{x \uparrow x_F} \frac{a(x)}{a(x + va(x))} = 1,
\]

uniformly on bounded \( v \)-intervals. This together with (3.27) yields

\[
(3.29) \quad \lim_{x \uparrow x_F} \frac{\overline{F}(x + t a(x))}{\overline{F}(x)} = e^{-t}
\]

uniformly on bounded \( t \)-intervals. Now choose the norming constants \( d_n = F^{-} (1 - n^{-1}) \) and \( c_n = a(d_n) \). Then (3.29) implies

\[
\lim_{n \to \infty} n \overline{F}(d_n + tc_n) = e^{-t} = - \ln \Lambda(t), \quad t \in \mathbb{R}.
\]

An application of Proposition 3.19 shows that \( F \in \text{MDA}(\Lambda) \). \( \square \)
Characterizations of MDA(Λ)

Von Mises functions do not completely characterize the maximum domain of attraction of Λ. However, a slight modification of the defining relation (3.22) of a von Mises function yields a complete characterization of MDA(Λ).

For a proof of the following result we refer to Resnick [5], Corollary 1.7 and Proposition 1.9.

**Theorem 3.40** (Characterization I of MDA(Λ))
The distribution function \( F \) with right endpoint \( x_F \leq \infty \) belongs to the maximum domain of attraction of Λ if and only if there exists some \( z < x_F \) such that \( F \) has representation

\[
\bar{F}(x) = c(x) \exp \left\{ - \int_z^x \frac{g(t)}{a(t)} \, dt \right\}, \quad z < x < x_F,
\]

where \( c \) and \( g \) are measurable functions satisfying \( c(x) \to c > 0 \), \( g(x) \to 1 \) as \( x \uparrow x_F \), and \( a(x) \) is a positive function with density \( a'(x) \) having \( \lim_{x \uparrow x_F} a'(x) = 0 \).

For \( F \) with representation (3.30) we can choose

\[
d_n = F^+(1 - n^{-1}) \quad \text{and} \quad c_n = a(d_n)
\]
as normalizing constants.

A possible choice for the function \( a \) is

\[
a(x) = \int_x^{x_F} \frac{\bar{F}(t)}{\bar{F}(x)} \, dt, \quad x < x_F,
\]

Motivated by von Mises functions, we call the function \( a \) in (3.30) an auxiliary function for \( F \).

**Remarks.** 4) Representation (3.30) is not unique, there being some trade-off possible between the functions \( c \) and \( g \). The following representation can be employed alternatively; see Resnick [5], Proposition 1.4:

\[
\bar{F}(x) = c(x) \exp \left\{ - \int_z^x \frac{1}{a(t)} \, dt \right\}, \quad z < x < x_F,
\]

for functions \( c \) and \( a \) with properties as in Theorem 3.40.

5) For a random variable \( X \) the function \( a(x) \) as defined in (3.31) is nothing but the mean excess function

\[
a(x) = E(X - x \mid X > x), \quad x < x_F;
\]

see also Section 3.4 for a discussion on the use of this function. Later the mean excess function will turn out to be an important tool for statistical fitting of extremal event data.

Another characterization of MDA(Λ) is suggested in the proof of Proposition 3.39. There one shows that every von Mises function satisfies the following relation for some positive function \( \tilde{a} \):

\[
\lim_{x \uparrow x_F} \frac{\bar{F}(x + t\tilde{a}(x))}{\bar{F}(x)} = e^{-t}, \quad t \in \mathbb{R}.
\]
**Theorem 3.41** (Characterization II of MDA(Λ))

The distribution function \( F \) belongs to the maximum domain of attraction of \( Λ \) if and only if there exists some positive function \( a \) such that (3.33) holds. A possible choice is \( a = a \) as given in (3.31).

The proof of this result is for instance to be found in de Haan [4], Theorem 2.5.1.

Now recall the notion of tail-equivalence (Definition 3.20). Similarly to the maximum domains of attraction of the Weibull and Fréchet distribution, tail-equivalence is an auxiliary tool to decide whether a particular distribution belongs to the maximum domain of attraction of \( Λ \) and to calculate the norming constants. In MDA(Λ) it is even more important because of the large variety of tails \( F \).

**Proposition 3.42** (Closure property of MDA(Λ) under tail-equivalence)

Let \( F \) and \( G \) be distribution functions with the same right endpoint \( x_F = x_G \) and assume that \( F \in \text{MDA}(Λ) \) with norming constants \( c_n > 0 \) and \( d_n \in \mathbb{R} \); i.e.,

\[
\lim_{n \to \infty} F^n(c_n x + d_n) = \Lambda(x), \quad x \in \mathbb{R}.
\]

Then

\[
\lim_{n \to \infty} G^n(c_n x + d_n) = \Lambda(x + b), \quad x \in \mathbb{R},
\]

if and only if \( F \) and \( G \) are tail-equivalent with

\[
\lim_{x \to x_F} \frac{F(x)}{G(x)} = e^b.
\]

**Proof of the sufficiency.** For a proof of the necessity see Resnick [5], Proposition 1.19. Suppose that \( F(x) \sim c G(x) \) as \( x \uparrow x_F \) for some \( c > 0 \). By Proposition 3.19 the limit relation (3.34) is equivalent to

\[
\lim_{n \to \infty} n F^n(c_n x + d_n) = e^{-x}, \quad x \in \mathbb{R}.
\]

For such \( x \), \( c_n x + d_n \to x_F \) and hence, by tail-equivalence,

\[
nc_n x + d_n \sim c e^{-1} F(c_n x + d_n) \to e^{-1} e^{-x}, \quad x \in \mathbb{R}.
\]

Therefore by Proposition 3.19,

\[
\lim_{n \to \infty} G^n(c_n x + d_n) = \exp \left\{ -e^{-\left( x + \ln c \right)} \right\} = \Lambda(x + \ln c), \quad x \in \mathbb{R}.
\]

Now set \( \ln c = b \)

The results of this section yield a further complete characterization of MDA(Λ).

**MDA(Λ)** consists of von Mises functions

and their tail-equivalent distribution functions.

This statement and the examples discussed throughout this section show that MDA(Λ) consists of a large variety of distributions whose tails can be very different. Tails may range from moderately heavy (lognormal, heavy-tailed Weibull) to very light (exponential, distribution functions with support bounded to the right). Because of this, MDA(Λ) is perhaps the most interesting among all maximum domains of attraction. As a natural consequence of the variety of tails in MDA(Λ), the norming constants also vary considerably. Whereas in MDA(Φα) and MDA(Ψα) the norming constants are calculated by straightforward application of regular variation theory, more advanced results are needed for MDA(Λ). A complete theory has been developed by de Haan involving certain subclasses of rapidly and slowly varying functions; see de Haan [4] or Bingham et al. [1], Chapter 3. Various examples below will illustrate the usefulness of results like Proposition 3.42.
Figure 3.44 Distribution functions of the normalized maxima of $n$ standard normal random variables and the Gumbel distribution function (top). In the bottom figure the relative error of this approximation for the tail is illustrated. The rate of convergence appears to be very slow.

Example 3.43 (Normal distribution)
See also Figure 3.44. Denote by $\Phi$ the distribution function and by $\varphi$ the density of the standard normal distribution. We first show that $\Phi$ is a von Mises function and check condition (3.23). An application of l'Hospital's rule to $\Phi'(x)/(x^{-1} \varphi(x))$ yields Mill's ratio, $\Phi'(x) \sim \varphi(x)/x$. Furthermore $\varphi'(x) = -x \varphi(x) < 0$ and

$$
\lim_{x \to \infty} \frac{\Phi(x) \varphi'(x)}{\varphi^2(x)} = -1.
$$

Thus $\Phi \in \text{MDA}(A)$ by Example 3.37 and Proposition 3.39. We now calculate the norming constants. Use Mill's ratio again:

$$
(3.35) \quad \overline{\Phi}(x) \sim \frac{\varphi(x)}{x} = \frac{1}{\sqrt{2\pi} \, x} \, e^{-x^2/2}, \quad x \to \infty,
$$

and interpret the right-hand side as the tail of some distribution function $G$. Then by Propo-
sition 3.42, $\Phi$ and $G$ have the same norming constants $c_n$ and $d_n$. According to (3.26), $d_n = G^\ast (1 - n^{-1})$. Hence look for a solution of $-\ln G (d_n) = \ln n$; i.e.,

\[
\frac{1}{2} d_n^2 + \ln d_n + \frac{1}{2} \ln 2\pi = \ln n.
\]

Then a Taylor expansion in (3.36) yields

\[
d_n = (2 \ln n)^{1/2} - \frac{\ln \ln n + \ln 4\pi}{2(2 \ln n)^{1/2}} + o \left( \frac{1}{\ln n} \right)
\]
as a possible choice for $d_n$. Since we can take $a(x) = \frac{\Phi(x)}{\varphi(x)}$ we have that $a(x) \sim x^{-1}$ and therefore

\[
c_n = a(d_n) \sim (2 \ln n)^{-1/2}.
\]

As the $c_n$ are unique up to asymptotic equivalence, we choose

\[
c_n = (2 \ln n)^{-1/2}.
\]

We conclude that

\[
\sqrt{2 \ln n} \left( M_n - \sqrt{2 \ln n} + \frac{\ln \ln n + \ln 4\pi}{2(2 \ln n)^{1/2}} \right) \xrightarrow{d} \Lambda.
\]

Note that $c_n \to 0$, i.e., the distribution of $M_n$ becomes less spread around $d_n$ as $n$ increases. \qed

Similarly, it can be proved that the gamma distribution also belongs to MDA ($\Lambda$).

Another useful trick to calculate the norming constants is via monotone transformations. If $g$ is an increasing function and $\hat{X} = g(x)$, then obviously

\[
\hat{M}_n = \max \left( \hat{X}_1, \ldots, \hat{X}_n \right) = g(M_n).
\]

If $X \in$ MDA ($\Lambda$) with

\[
\lim_{n \to \infty} P \left( M_n \leq c_n x + d_n \right) = \Lambda(x), \quad x \in \mathbb{R},
\]

then

\[
\lim_{n \to \infty} P \left( \hat{M}_n \leq g(c_n x + d_n) \right) = \Lambda(x), \quad x \in \mathbb{R}.
\]

In some cases, $g$ may be expanded in a Taylor series about $d_n$ and just linear terms suffice to give the limit law for $\hat{M}_n$, with changed constants $\hat{c}_n = c_n g'(d_n)$ and $\hat{d}_n = g(d_n)$. We apply this method to the lognormal distribution.

**Example 3.45** (Lognormal distribution)

Let $X$ be a standard normal random variable and $g(x) = e^{\mu + \sigma x}$, $\mu \in \mathbb{R}$, $\sigma > 0$. Then

\[
\hat{X} = g(X) = e^{\mu + \sigma X}
\]
defines a lognormal random variable. Since $X \in$ MDA ($\Lambda$) we obtain

\[
\lim_{n \to \infty} P \left( \hat{M}_n \leq e^{\mu + \sigma(c_n x + d_n)} \right) = \Lambda(x), \quad x \in \mathbb{R},
\]

where $c_n$, $d_n$ are the norming constants of the standard normal distribution as calculated in Example 3.43. This implies

\[
\lim_{n \to \infty} P \left( e^{-\mu - \sigma d_n} \hat{M}_n \leq 1 + \sigma c_n x + o(c_n) \right) = \Lambda(x), \quad x \in \mathbb{R}.
\]
Since \( c_n \to 0 \) it follows that
\[
e^{-\mu-\sigma d_n} \frac{\sigma e^{-\mu-\sigma d_n}}{\sigma c_n} \left( \tilde{M}_n - e^{\mu+\sigma d_n} \right) d \to \Lambda,
\]
so that \( \tilde{X} \in \text{MDA}(\Lambda) \) with norming constants
\[
\tilde{c}_n = \sigma e^{\mu+\sigma d_n}, \quad \tilde{d}_n = e^{\mu+\sigma d_n}.
\]

\[\square\]

**Further properties of distributions in MDA(\( \Lambda \))**

In the remainder of this section we collect some further useful facts about distributions in MDA(\( \Lambda \)).

**Proposition 3.46** (Existence of moments)

Assume that the random variable \( X \) has distribution function \( F \in \text{MDA}(\Lambda) \) with infinite right endpoint. Then \( F \) is rapidly varying. In particular, the moments \( E(X^+)^\alpha \) are finite for every \( \alpha > 0 \), where \( X^+ = \max(0, X) \).

In Section 3.3.2 we showed that the maximum domains of attraction of \( \Psi_\alpha \) and \( \Phi_\alpha \) are linked in a natural way. Now we show that MDA(\( \Phi_\alpha \)) can be embedded in MDA(\( \Lambda \)).

**Example 3.47** (Embedding MDA(\( \Phi_\alpha \)) in MDA(\( \Lambda \)))

Let \( X \) have distribution function \( F \in \text{MDA}(\Phi_\alpha) \) with norming constants \( c_n \). Define
\[
X^* = \ln(1 \vee X)
\]
with distribution function \( F^* \). By Proposition 3.19 and Theorem 3.21, \( F \in \text{MDA}(\Phi_\alpha) \) if and only if
\[
\lim_{n \to \infty} n \overline{F}(c_n x) = \lim_{n \to \infty} \frac{\overline{F}(c_n, x)}{\overline{F}(c_n)} = x^{-\alpha}, \quad x > 0.
\]

This implies that
\[
\lim_{n \to \infty} \frac{F^*(\alpha^{-1} x + \ln c_n)}{F^*(\ln c_n)} = \lim_{n \to \infty} \frac{F(c_n \exp \{ \alpha^{-1} x \})}{F(c_n)} = e^{-x}, \quad x \in \mathbb{R}.
\]

Hence \( F^* \in \text{MDA}(\Lambda) \) with norming constants \( c_n^* = \alpha^{-1} \) and \( d_n^* = \ln c_n \). As auxiliary function one can take
\[
a^*(x) = \int_x^\infty \frac{F^*(y)}{F^*(x)} \, dy
\]

\[\square\]

**Example 3.48** (Subexponential distributions and MDA(\( \Lambda \)))

Goldie and Resnick [3] characterize the distribution functions \( F \) that are both subexponential (we write \( F \in \mathcal{S} \)) and in MDA(\( \Lambda \)). Starting from the representation (3.30) for \( F \in \text{MDA}(\Lambda) \), they give necessary and sufficient conditions for \( F \in \mathcal{S} \). In particular, \( \lim_{x \to \infty} a(x) = \infty \) is necessary but not sufficient for \( F \in \text{MDA}(\Lambda) \cap \mathcal{S} \). A simple sufficient condition for \( F \in \text{MDA}(\Lambda) \cap \mathcal{S} \) is that \( a \) is eventually non-decreasing and that there exists some \( t > 1 \) such that
\[
\liminf_{x \to \infty} \frac{a(tx)}{a(x)} > 1.
\]

This condition is easily checked for the following distributions which are all von Mises functions and hence in MDA(\( \Lambda \)):
- **Benktander-type-I**

\[
\mathcal{F}(x) = (1 + 2(\beta/\alpha)\ln x)\exp\{-((\beta(\ln x)^2 + (\alpha + 1)\ln x)\}, \quad x \geq 1, \quad \alpha, \beta > 0.
\]

Here one can choose

\[
a(x) = \frac{x}{\alpha + 2\beta \ln x}, \quad x \geq 1.
\]

- **Benktander-type-II**

\[
\mathcal{F}(x) = e^{\alpha/\beta} x^{-(1-\beta)} \exp\left\{-\frac{\alpha}{\beta} x^\beta\right\}, \quad x \geq 1, \quad \alpha > 0, \ 0 < \beta < 1,
\]

with auxiliary function

\[
a(x) = \frac{x^{1-\beta}}{\alpha}, \quad x \geq 1.
\]

- **Weibull**

\[
\mathcal{F}(x) = e^{-c x^\tau}, \quad x \geq 0, \quad 0 < \tau < 1, \quad c > 0,
\]

with auxiliary function

\[
a(x) = e^{-c^{-1} \tau^{-1} x^{1-\tau}}, \quad x \geq 0.
\]

- **Lognormal**

with auxiliary function

\[
a(x) = \frac{\Phi(\sigma^{-1}(\ln x - \mu))\sigma x}{\varphi(\sigma^{-1}(\ln x - \mu))} \sim \frac{\sigma^2 x}{\ln x - \mu}, \quad x \to \infty.
\]

The critical cases occur when \( F \) is in the tail close to an exponential distribution. For example, let

\[
\mathcal{F}(x) \sim \exp\{-x(\ln x)^\alpha\}, \quad x \to \infty.
\]

For \( \alpha < 0 \) we have \( F \in \text{MDA}(\Lambda) \cap \mathcal{S} \) in view of Theorem 2.7 in Goldie and Resnick [3], whereas for \( \alpha \geq 0, \ F \in \text{MDA}(\Lambda) \) but \( F \notin \mathcal{S} \).

**Comments**

Various estimation methods will depend on an application of the Fisher-Tippett theorem and related results. The quality of those approximations will be crucial. Figure 3.15 suggests a fast rate of convergence in the case of the exponential distribution: already for \( n = 5 \) the distribution of the normalized maximum is quite close to \( \Lambda \), while for \( n = 50 \) they are almost indistinguishable. In contrast to this rapid rate of convergence, the distribution of the normalized maximum of a sample of normal random variables converges extremely slowly to its limit distribution \( \Lambda \); see Figure 3.44. This slow rate of convergence also depends on the particular choice of \( c_n \) and \( d_n \).

**References**


3.4 The generalized extreme value distribution and the generalized Pareto distribution

In Section 3.2 we have shown that the standard extreme value distributions, together with their types, provide the only non-degenerate limit laws for affinely transformed maxima of iid random variables. The three standard extreme value distributions can be represented in a compact one-parameter family \((H_\xi)_{\xi \in \mathbb{R}}\) of distributions containing the standard extreme value distributions, namely

\[
H_\xi = \begin{cases} 
\Phi_{1/\xi} & \text{if } \xi > 0, \\
\Lambda & \text{if } \xi = 0, \\
\Psi_{-1/\xi} & \text{if } \xi < 0.
\end{cases}
\]

The distribution \(H_\xi\) is referred to as the generalized extreme value distribution with parameter \(\xi\). Here \(\xi\) and \(\alpha\) are related:

- \(\xi = \alpha^{-1} > 0\) corresponds to the Fréchet distribution \(\Phi_\alpha\),
- \(\xi = 0\) corresponds to the Gumbel distribution \(\Lambda\),
- \(\xi = -\alpha^{-1} < 0\) corresponds to the Weibull distribution \(\Psi_\alpha\).

The following choice is by now widely accepted as the standard representation.

**Definition 3.49** (Jenkinson-von Mises representation of the extreme value distributions: the generalized extreme value distribution (GEV))

Define the distribution \(H_\xi\) by

\[
H_\xi(x) = \begin{cases} 
\exp \left\{ - (1 + \xi x)^{-1/\xi} \right\} & \text{if } \xi \neq 0, \\
\exp \{-\exp\{-x\}\} & \text{if } \xi = 0, 
\end{cases}
\]

where \(1 + \xi x > 0\).

Hence the support of \(H_\xi\) corresponds to

- \(x > -\xi^{-1}\) for \(\xi > 0\),
- \(x < -\xi^{-1}\) for \(\xi < 0\),
- \(x \in \mathbb{R}\) for \(\xi = 0\).

\(H_\xi\) is called a standard generalized extreme value distribution (GEV). One can introduce the related location-scale family \(H_{\xi, \mu, \psi}\) by replacing the argument \(x\) above by \((x - \mu)/\psi\) for \(\mu \in \mathbb{R}, \psi > 0\). The support has to be adjusted accordingly. We also refer to \(H_{\xi, \mu, \psi}\) as GEV. □
We consider the distribution $H_\xi$ as the limit of $H_\xi$ for $\xi \to 0$. With this interpretation
\[ H_\xi(x) = \exp \left\{ - (1 + \xi x)^{-1/\xi} \right\}, \quad 1 + \xi x > 0, \]
serves as a representation for all $\xi \in \mathbb{R}$. The densities of the standard GEV for $\xi = -1, 0, 1$ are shown in Figure 3.10.

The GEV provides a convenient unifying representation of the three extreme value types Gumbel, Fréchet and Weibull. Its introduction is mainly motivated by statistical applications as will be seen later. GEVs fitting will turn out to be one of the fundamental concepts. The following theorem is one of the basic results in extreme value theory. In a concise analytical way, it gives the essential information collected in the previous section on maximum domains of attraction. Moreover, it constitutes the basis for numerous statistical techniques to be discussed later. First recall the notion of the quantile function $F^{\xi-}$ of a distribution function $F$ and define
\[ U(t) = F^{\xi-}(1 - t^{-1}), \quad t > 0. \]

**Theorem 3.50** (Characterisation of MDA ($H_\xi$))

For $\xi \in \mathbb{R}$ the following assertions are equivalent:

1. $F \in \text{MDA} (H_\xi)$.

2. There exists a positive, measurable function $a(\cdot)$ such that for $1 + \xi x > 0$,
\[
\lim_{u \uparrow x} \frac{F(u + xa(u))}{F(u)} = \begin{cases} 
(1 + \xi x)^{-1/\xi} & \text{if } \xi \neq 0, \\
\exp(-x) & \text{if } \xi = 0.
\end{cases}
\]

3. For $x, y > 0$, $y \neq 1$,
\[
\lim_{s \to \infty} \frac{U(sx) - U(s)}{U(sy) - U(s)} = \begin{cases} 
x^\xi - 1 & \text{if } \xi \neq 0, \\
\frac{\ln x}{\ln y} & \text{if } \xi = 0.
\end{cases}
\]

**Sketch of the proof.** Below we give only the main ideas in order to show that the various conditions enter very naturally. Further details are to be found in the literature, for instance in de Haan [3].

(1)$\Leftrightarrow$(2) For $\xi = 0$ this is Theorem 3.41.

For $\xi > 0$ we have $H_\xi(x) = \Phi_{\alpha}(\alpha^{-1}(x + \alpha))$ for $\alpha = 1/\xi$. By Theorem 3.21, (1) is then equivalent to regular variation of $F$ with index $-\alpha$. By the representation (2.11) of slowly varying functions, for some $z > 0$,
\[
\Phi(x) = c(x) \exp \left\{ - \int_z^x \frac{1}{\alpha(t)} \, dt \right\}, \quad z < x < \infty,
\]
where $c(x) \to c > 0$ and $a(x)/x \to \alpha^{-1}$ as $x \to \infty$ locally uniformly. Hence
\[
\lim_{u \to \infty} \frac{\Phi(u + xa(u))}{\Phi(u)} = \left( 1 + \frac{x}{\alpha} \right)^{-\alpha},
\]
which is (3.39). If (2) holds, choose $d_n = (1/\Phi)^{-}(n) = U(n)$, then
\[
1/\Phi(d_n) \sim n,
\]
56
and with \( u = d_n \) in (3.39),
\[
(1 + \frac{x}{\alpha})^{-\alpha} = \lim_{n \to \infty} \frac{F(d_n + xa(d_n))}{F(d_n)} = \lim_{n \to \infty} \frac{nF(d_n + xa(d_n))}{F(d_n)} ,
\]
whence by Proposition 3.19, \( F \in \text{MDA}(H_\xi) \) for \( \xi = \alpha^{-1} \).

The case \( \xi < 0 \) can be treated similarly.

(2)\(\Rightarrow\)(3) We restrict ourselves to the case \( \xi \neq 0 \), the proof for \( \xi = 0 \) being analogous. For simplicity, we assume that \( F \) is continuous and increasing on \( (-\infty, x_F) \). Set \( s = 1/F(u) \), then (3.39) translates into
\[
A_s(x) = (sF(U(s) + xa(U(s))))^{-1} \to (1 + \xi x)^{1/\xi} , \quad s \to \infty .
\]
Now for every \( s > 0 \), \( A_s(x) \) is decreasing and for \( s \to \infty \) converges to a continuous function. Then also \( A_s^+(t) \) converges pointwise to the inverse of the corresponding limit function, i.e.,
\[
\lim_{s \to \infty} \frac{U(st) - U(s)}{a(U(s))} = \frac{t^{\xi} - 1}{\xi} .
\]

Now (3.40) is obtained by using (3.41) for \( t = x \) and \( t = y \) and taking the quotient. The proof of the converse can be given along the same lines. \( \square \)

**Remarks.** 1) Condition (3.39) has an interesting probabilistic interpretation. Indeed, let \( X \) be a random variable with distribution function \( F \in \text{MDA}(H_\xi) \), then (3.39) reformulates as
\[
\lim_{u \to x_F} P \left( \frac{X - u}{a(u)} > x \mid X > u \right) = \begin{cases} (1 + \xi x)^{-1/\xi} & \text{if } \xi \neq 0 , \\ e^{-x} & \text{if } \xi = 0 . \end{cases}
\]
Hence (3.42) gives a distributional approximation for the scaled excesses over the (high) threshold \( u \). The appropriate scaling factor is \( a(u) \). This interpretation will turn out to be crucial in many statistical applications.

2) A reformulation of relation (3.40) immediately leads to an estimation procedure for quantiles outside the range of the data.

In Remark 1 above we used the notion of excess. The following definition makes this precise.

**Definition 3.51** (Excess distribution function, mean excess function)

Let \( X \) be a random variable with distribution function \( F \) and right endpoint \( x_F \). For a fixed \( u < x_F \),
\[
F_u(x) = P(X - u \leq x \mid X > u) , \quad x \geq 0 ,
\]
is the excess distribution function of the random variable \( X \) (of the distribution function \( F \)) over the threshold \( u \). The function
\[
e(u) = E(X - u \mid X > u)
\]
is called the mean excess function of \( X \). \( \square \)

Excesses over thresholds play a fundamental role in many fields. Different names arise from specific applications. For instance, \( F_u \) is known as the excess-life or residual lifetime distribution function in reliability theory and medical statistics. In an insurance context, \( F_u \) is usually referred to as the excess-of-loss distribution function. A detailed discussion of some basic properties and statistical applications of the mean excess function and the excess distribution function are given later.
<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean Excess Function</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto</td>
<td>$\frac{\kappa + u}{\alpha - 1}$, $\alpha &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>Burr</td>
<td>$\frac{u}{\alpha \tau - 1} \left(1 + o(1)\right)$, $\alpha \tau &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>Loggamma</td>
<td>$\frac{u}{\alpha - 1} \left(1 + o(1)\right)$, $\alpha &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\frac{\sigma^2 u}{\ln u - \mu} \left(1 + o(1)\right)$</td>
<td></td>
</tr>
<tr>
<td>Benktander-type-I</td>
<td>$\frac{u}{\alpha + 2\beta \ln u}$</td>
<td></td>
</tr>
<tr>
<td>Benktander-type-II</td>
<td>$\frac{u^{1-\beta}}{\alpha}$</td>
<td></td>
</tr>
<tr>
<td>Weibull</td>
<td>$\frac{u^{1-\gamma}}{c \tau} \left(1 + o(1)\right)$</td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>$\beta^{-1} \left(1 + \frac{\alpha - 1}{\beta u} + o\left(\frac{1}{u}\right)\right)$</td>
<td></td>
</tr>
<tr>
<td>Truncated normal</td>
<td>$u^{-1} \left(1 + o(1)\right)$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.52 Mean excess functions for some standard distributions. The parametrisation is taken from Tables 2.10 and 2.12. The asymptotic relations are to be understood for $u \to \infty$.

**Example 3.53** (Calculation of the mean excess function)
Using the definition of $e(u)$ and partial integration, the following formulae are easily checked. They are useful for calculating the mean excess function in special cases. Suppose for ease of presentation that $X$ is a positive random variable with distribution function $F$ and finite expectation; trivial changes allow for support $(x_0, \infty)$ for some $x_0 > 0$. Then

$$e(u) = \int_u^{x_F} (x - u) dF(x) / F(u)$$

(3.44)

$$= \frac{1}{F(u)} \int_u^{x_F} F(x) dx, \quad 0 < u < x_F.$$  

Whenever $F$ is continuous,

(3.45)

$$F(x) = \frac{e(0)}{e(x)} \exp \left\{ - \int_0^x \frac{1}{e(u)} du \right\}, \quad x > 0.$$  

It immediately follows from (3.45) that a continuous distribution function is uniquely determined by its mean excess function. If, as in many cases of practical interest, $F$ is regularly varying with
index \(-\alpha\) for some \(\alpha > 1\), then an immediate application of Karamata’s theorem (see (2.14) and (2.15)) yields \(c(u) \sim u/(\alpha - 1)\) as \(u \to \infty\). In Table 3.52 the mean excess functions of some standard distributions are summarized.

The appearance of the right-hand side limit in (3.42) motivates the following definition.

**Definition 3.54** (The generalized Pareto distribution (GPD))

Define the distribution function \(G_\xi\) by

\[
G_\xi(x) = \begin{cases} 
1 - (1 + \xi x)^{-1/\xi} & \text{if } \xi \neq 0, \\
1 - e^{-x} & \text{if } \xi = 0,
\end{cases}
\]

where

\[
x \geq 0 \quad \text{if } \xi \geq 0, \\
0 \leq x \leq -1/\xi \quad \text{if } \xi < 0.
\]

\(G_\xi\) is called a standard generalized Pareto distribution (GPD). One can introduce the related location-scale family \(G_{\xi,\nu,\beta}\) by replacing the argument \(x\) above by \((x - \nu)/\beta\) for \(\nu \in \mathbb{R}, \beta > 0\). The support has to be adjusted accordingly. We also refer to \(G_{\xi,\nu,\beta}\) as GPD.

As in the case of \(H_0\), \(G_0\) can also be interpreted as the limit of \(G_\xi\) as \(\xi \to 0\). The distribution function \(G_{\xi,0,\beta}\) plays an important role for fitting tails to high threshold excesses of data. To economize on notation, we will denote

\[
G_{\xi,\beta}(x) = 1 - \left(1 + \frac{\xi \alpha}{\beta} x\right)^{-1/\xi}, \quad x \in D(\xi, \beta),
\]

where

\[
x \in D(\xi, \beta) = \begin{cases} 
[0, \infty) & \text{if } \xi \geq 0, \\
[0, -\beta/\xi] & \text{if } \xi < 0.
\end{cases}
\]

Whenever we say that \(X\) has a GPD with parameters \(\xi\) and \(\beta\), it is understood that \(X\) has distribution function \(G_{\xi,\beta}\).

Time to summarize:

<table>
<thead>
<tr>
<th>The GEV</th>
<th>(H_\xi, \xi \in \mathbb{R}), describes the limit distributions of normalized maxima.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The GPD</td>
<td>(G_{\xi,\beta}, \xi \in \mathbb{R}, \beta &gt; 0), appears as the limit distribution of scaled excesses over high thresholds.</td>
</tr>
</tbody>
</table>

GPD fitting is one of the most useful concepts in the statistics of extremal events. Here we collect some basic probabilistic properties of the GPD.

**Theorem 3.58** (Properties of GPD)
Figure 3.55 Densities of the GPD for different parameters $\xi$ and $\beta = 1$.

Figure 3.56 Densities of the GPD for $\xi < 0$ and $\beta = 1$. Recall that they have compact support $[0, -1/\xi]$. 
1. Suppose $X$ has a GPD with parameters $\xi$ and $\beta$. Then $EX < \infty$ if and only if $\xi < 1$. In the latter case

$$E \left( 1 + \frac{\xi}{\beta} X \right)^{-r} = \frac{1}{1 + \xi r}, \quad r > -1/\xi,$$

$$E \left( \ln \left( 1 + \frac{\xi}{\beta} X \right) \right)^k = \xi^k k!, \quad k \in \mathbb{N},$$

$$EX \left( G_{\xi,\beta}(X) \right)^r = \frac{\beta}{(r + 1 - \xi)(r + 1)}, \quad (r + 1)/|\xi| > 0.$$

If $\xi < 1/r$ with $r \in \mathbb{N}$, then

$$EX^r = \frac{\beta^r \Gamma(\xi^{-1} - r)}{\xi^{r+1} \Gamma(1 + \xi^{-1})} r!.$$

2. For every $\xi \in \mathbb{R}$, $F \in \text{MDA}(H_\xi)$ if and only if

$$\lim_{u \uparrow \tau} \sup_{0 < x \leq u - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0$$

for some positive function $\beta$.

3. Suppose $x_i \in D(\xi,\beta), i = 1, 2$, then

$$\frac{G_{\xi,\beta}(x_1 + x_2)}{G_{\xi,\beta}(x_1)} = \frac{G_{\xi,\beta + \xi x_1}(x_2)}{G_{\xi,\beta}(x_1)}.$$

4. Assume that $N$ is Poisson($\lambda$), independent of the iid sequence $(X_n)$ with a GPD with parameters $\xi$ and $\beta$. Write $M_N = \max(X_1, \ldots, X_N)$. Then

$$P(M_N \leq x) = \exp \left\{ -\lambda \left( 1 + \frac{x}{\beta} \right)^{-1/\xi} \right\} = H_{\xi,\mu,\psi}(x),$$

Figure 3.57 GPD for different parameters $\xi$ and $\beta = 1$. 

1. Suppose $X$ has a GPD with parameters $\xi$ and $\beta$. Then $EX < \infty$ if and only if $\xi < 1$. In the latter case

$$E \left( 1 + \frac{\xi}{\beta} X \right)^{-r} = \frac{1}{1 + \xi r}, \quad r > -1/\xi,$$

$$E \left( \ln \left( 1 + \frac{\xi}{\beta} X \right) \right)^k = \xi^k k!, \quad k \in \mathbb{N},$$

$$EX \left( G_{\xi,\beta}(X) \right)^r = \frac{\beta}{(r + 1 - \xi)(r + 1)}, \quad (r + 1)/|\xi| > 0.$$

If $\xi < 1/r$ with $r \in \mathbb{N}$, then

$$EX^r = \frac{\beta^r \Gamma(\xi^{-1} - r)}{\xi^{r+1} \Gamma(1 + \xi^{-1})} r!.$$

2. For every $\xi \in \mathbb{R}$, $F \in \text{MDA}(H_\xi)$ if and only if

$$\lim_{u \uparrow \tau} \sup_{0 < x \leq u - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0$$

for some positive function $\beta$.

3. Suppose $x_i \in D(\xi,\beta), i = 1, 2$, then

$$\frac{G_{\xi,\beta}(x_1 + x_2)}{G_{\xi,\beta}(x_1)} = \frac{G_{\xi,\beta + \xi x_1}(x_2)}{G_{\xi,\beta}(x_1)}.$$

4. Assume that $N$ is Poisson($\lambda$), independent of the iid sequence $(X_n)$ with a GPD with parameters $\xi$ and $\beta$. Write $M_N = \max(X_1, \ldots, X_N)$. Then

$$P(M_N \leq x) = \exp \left\{ -\lambda \left( 1 + \frac{x}{\beta} \right)^{-1/\xi} \right\} = H_{\xi,\mu,\psi}(x),$$

61
where \( \mu = \beta \xi^{-1} (\lambda^\xi - 1) \) and \( \psi = \beta \lambda^\xi \).

5. Suppose \( X \) has GPD with parameters \( \xi < 1 \) and \( \beta \). Then for \( u < x_F \),

\[
e(u) = E(X - u \mid X > u) = \frac{\beta + \xi u}{1 - \xi}, \quad \beta + u\xi > 0.
\]

**Proof.** (1) and (3) follow by direct verification.

(2) In Theorem 3.50 (see Remark 1) we have already proved that \( F \in \text{MDA}(H_\xi) \) if and only if

\[
\lim_{u \uparrow x_F} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0
\]

where \( \beta(u) = a(u) \). Because the GPD is continuous, the uniformity of the convergence follows

(4) One immediately obtains

\[
P(M_N \leq x) = \sum_{n=0}^{\infty} e^{-\lambda n} \frac{\lambda^n}{n!} G_{\xi,\beta}^n(x)
= \exp\{-\lambda G_{\xi,\beta}(x)\}
= \exp\left\{-\lambda \left(1 + \frac{x}{\beta(\lambda^\xi - 1)}\right)^{-1/\xi}\right\}
= \exp\left\{-\left(1 + \frac{x\xi(\lambda^\xi - 1)\beta}{\beta \lambda^\xi}\right)^{-1/\xi}\right\}, \quad \xi \neq 0.
\]

The case \( \xi = 0 \) reduces to

\[
P(M_N \leq x) = \exp\left\{-e^{-\left(x - \ln \lambda / \beta\right)}\right\}.
\]

(5) This result immediately follows from the representation (3.44). \( \square \)

**Remarks.** 4) Theorem 3.58 summarizes various properties which are essential for the special role of the GPD in the statistical analysis of extremes.

5) The property (3) above is often reformulated as follows: the class of GPDs is closed with respect to changes of the threshold. Indeed the left-hand side in (3.49) is the conditional probability that, given our underlying random variable exceeds \( x_1 \), it also exceeds the threshold \( x_1 + x_2 \). The right-hand side states that this probability is again of the generalized Pareto type.

6) Property (2) above suggests a GPD as appropriate approximation of the excess distribution function \( F_u \) for large \( u \). This result goes back to Pickands [6] and is often formulated as follows. For some function \( \beta \) to be estimated from the data,

\[
\overline{F}_u(x) = P(X - u > x \mid X > u) \approx \overline{G}_{\xi,\beta(u)}(x), \quad x > 0.
\]

Alternatively one considers for \( x > u \),

\[
P(X > x \mid X > u) = \overline{G}_{\xi(u),\beta(u)}(x).
\]

In both cases \( u \) has to be taken sufficiently large. Together (2) and (5) give us a nice graphical technique for choosing the threshold \( u \) so high that an approximation of the excess distribution

62
function $F_u$ by a GPD is justified: given an iid sample $X_1, \ldots, X_n$, construct the empirical mean excess function $e_n(u)$ as sample version of the mean excess function $e(u)$. From (5) we have that the mean excess function of a GPD is linear, hence check for a $u$-region where the graph of $e_n(u)$ becomes roughly linear. For such $u$ an approximation of $F_u$ by a GPD seems reasonable. Later we will use this approach for fitting excesses over thresholds.

7) From Proposition 3.1 (see also the succeeding Remark) we have learnt that the number of the exceedances of a high threshold is roughly Poisson. From Remark 6 we conclude that an approximation of the excess distribution function $F_u$ by a GPD may be justified. Moreover, it can be shown that the number of exceedances and the excesses are independent in an asymptotic sense; see Leadbetter [5].

8) Property (4) now says that in a model, where the number of exceedances is exactly Poisson and the excess distribution function is an exact GDP, the maximum of these excesses has an exact GEV.

The above remarks suggest the following approximate model for the exceedance times and the excesses of an iid sample:

- The number of exceedances of a high threshold follows a Poisson process.
- Excesses over high thresholds can be modeled by a GPD.
- An appropriate value of the high threshold can be found by plotting the empirical mean excess function.
- The distribution of the maximum of a Poisson number of iid excesses over a high threshold is a GEV.

In interpreting the above summary, do look at the precise formulation of the underlying theorems. If at this point you want to see some of the above in action; see for instance Smith [8], p. 461.

**Comments**

In this section we summarized some of the probabilistic properties of the GEV and the GPD. They are crucial for the statistical analysis of extremal events. The GEV will be used for statistical inference of data which occur as iid maxima of certain time series, for instance annual maxima of rainfall, windspeed, etc. Theorem 3.50 opens the way to tail and and high quantile estimation. Part (2) of this theorem leads immediately to the definition of the GPD, which goes back to Pickands [6]. An approximation of the excess distribution function by the GPD has been suggested by hydrologists under the name peaks over threshold (POT) method which will be considered later. Weak limit theory for excess distribution functions originates from Balkema and de Haan [1]. Richard Smith and his collaborators have further developed the theory and applications of the GPD in various fields. Basic properties of the GPD can for instance be found in Smith [7]. Detailed discussions on the use of the mean excess function in insurance are to be found in Beirlant et al. [2] and Hogg and Klugman [4].
References


4 Statistical analysis of extremes

4.1 Some examples

In the previous sections we have introduced a multitude of probabilistic models in order to describe, in a mathematically sound way, extremal events in the one-dimensional case. The real world however often informs us about such events through *statistical data*: major insurance claims, flood levels of rivers, large decreases (or indeed increases) of stock market values over a certain period of time, extreme levels of environmental indicators such as ozone or carbon monoxide, wind-speed values at a certain site, wave heights during a storm or maximal and minimal performance values of a portfolio. All these, and indeed many more examples, have in common that they concern questions about extreme values of some underlying set of data. At this point it would be utterly foolish (and indeed very wrong) to say that all such problems can be cast into one or the other probabilistic model treated so far: this is definitely not the case! Applied mathematical (including statistical) modeling is all about trying to offer the applied researcher (the finance expert, the insurer, the environmentalist, the biologist, the hydrologist, the risk manager, ...) the necessary set of tools in order to deduce scientifically sound conclusions from data. It is however also very much about reporting correctly: the data have to be presented in a clear and objective way, precise questions have to be formulated, model-based answers given, always stressing the underlying assumptions. The whole process constitutes an art: statistical theory plays only a relatively small, though crucial role here.

The previous sections have given us a whole battery of techniques with which to formulate in a mathematically precise way the basic questions underlying extreme value theory. This section aims at going one step further: based on data, we shall present statistical tools allowing us to link questions asked in practice to a particular (though often non-unique) probabilistic model. Our treatment as regards these statistical tools will definitely not be complete, though we hope it will be representative of current statistical methodology in this fast-expanding area. The reader will meet data, basic descriptive methods, and techniques from mathematical statistics concerning estimation and testing in extreme value models.

After the mathematical theory of maxima and heavy-tailed distributions presented in the previous sections, we now turn to the crucial question:

64
How do extreme values manifest themselves in real data?

A full answer to this question would not only take most of the present section, one could write whole volumes on it. Let us start by seeing how in practice extremes in data manifest themselves. We do this through a series of partly hypothetical, partly real examples. At a later stage in the section, we will come back to some of the examples for a more detailed analysis.

Example 4.1 (River Nidd data)
A standard data-set in extreme value theory concerns flows of the river Nidd in Yorkshire, England; the source of the data is the Flood Studies Report NERC [8]. We are grateful to Richard Smith for having provided us with a copy of the data. The basic set contains 154 observations on flow data above 65 CUMECS over the 35-year period 1934-1970. A crude de-clustering technique was used by the hydrologists to prepare these data. Though the full set contains a series of values for each year, for a first analysis only the annual maxima are considered. In this way, intra-year dependencies are avoided and a valid assumption may be to suppose that the data \( x_1, \ldots, x_{35} \) are realizations from a sequence \( X_1, \ldots, X_{35} \) of iid random variables all with common extreme value distribution \( H \) say. Suppose we want to answer questions like:

- What is the probability that the maximum flow for the next year will exceed a level \( x \)?
- What is the probability that the maximum flow for the next year exceeds all previous levels?
- What is the expected length of time (in years say) before the occurrence of a specific high quantity of flow?

Clearly, a crucial step forward in answering these questions would be our gaining knowledge of the distribution \( H \). The theory of Section 3 gives us relevant parametric models for \( H \); see the Fisher-Tippett theorem (Theorem 3.9) where the extreme value distributions enter. Standard statistical tools such as maximum likelihood estimation (MLE) are available.

Example 4.3 (Insurance claims)
Suppose our data consist of fire insurance claims \( x_1, \ldots, x_n \) over a specified period of time in a well-defined portfolio, as for instance presented in Figure 4.4. Depending on the type of fire causing the specific claims, a condition of the type “\( x_1, \ldots, x_n \) come from an iid sample \( X_1, \ldots, X_n \) with distribution function \( F \)” may or may not be justified. Suppose for the sake of argument that the underlying portfolio is such that the above assumption can be made. Questions we want to answer (or tasks we want to perform) could be:

- Calculate next year’s premium volume needed in order to cover, with sufficiently high probability, future losses in this portfolio.
- What is the probable-maximum-loss of this portfolio if the latter is defined as a high (for instance the 0.999-) quantile of the distribution function \( F \)?
- Given that we want to write an excess-of-loss cover with priority \( a_k \) (also referred to as attachment point) resulting in a one-in-\( k \)-year event, how do we calculate \( a_k \)? The latter means that we want to calculate \( a_k \) so that the probability of exceeding \( a_k \) equals 1/\( k \).

Again, as in the previous example, we are faced with a standard statistical fitting problem. The main difference is that in this case we do not immediately have a specific parametric model (such as the extreme value distributions in Example 4.1) in mind. We first have to learn about the data:
Is $F$ light- or heavy-tailed?

What are its further shape properties: skewed, flat, unimodal, ...?

In the heavy-tailed case fitting by a subexponential distribution (see Section 2.2.6) might be called for. The method of exceedances to be explained soon will be relevant.

\[ \square \]

**Example 4.5** (ECOMOR reinsurance)

The ECOMOR reinsurance contract stands for “Le Traité d’Excédent du Coût Moyen Relatif” and was introduced by the French actuary Thépaut as a novel contract aiming to enlarge the reinsurer’s flexibility in constructing customised products. Suppose over a period $[0, t]$, the claims $x_1, \ldots, x_{n(t)}$ are received by the primary insurer. The ECOMOR contract binds the reinsurer to cover (for a specific premium) the excesses above the $k$th largest claim. This leads us to a model where $X_1, \ldots, X_{N(t)}$ are (conditionally) iid with specific model assumptions on the underlying distribution function $F$ of $X_i$ and on the counting process $(N(t))$. It is perhaps worthwhile to stress that, though innovative in nature, ECOMOR never was a success.

\[ \square \]

**Example 4.6** (Value-at-Risk)

Suppose a financial portfolio consists of a number of underlying assets (bonds, stocks, derivatives, ...), all having individual (though correlated) values at any time $t$. Every asset has its specific Profit–Loss (P&L) distribution, which can be represented as a probability distribution governing the (random) changes of value. Through the estimation of portfolio covariances, the portfolio manager then estimates the overall portfolio P&L distribution. Management and regulators may now be interested in setting “minimal requirements” or, for the sake of argument, a maximal limit on the potential losses. A possible quantity is the so-called Value-at-Risk (VaR) measure briefly treated in the discussion of Figure 1.1. There the VaR is defined as the 5% quantile of the P&L distribution. The following questions are relevant.

- Estimate the VaR for a given portfolio.
Figure 4.4 4580 claims from a fire insurance portfolio. The values are multiples of 1000 SFr. The corresponding histogram of the claims ≤ 5000 SFr (left) and of the remaining claims exceeding 5000 SFr (right). The data are very skewed to the right. The x-axis of the histogram on the rhs reaches up to 250 due to a very large claim around 225; see also the top figure.

- Estimate the probability that, given we exceed the VaR, we exceed it by a certain amount. This corresponds to the calculation of the so-called shortfall distribution.

The first question concerns quantile estimation for an estimated distribution function, in many cases outside the range of our data. The second question obviously concerns the estimation of the excess distribution function as defined in Section 3.4 (modulo a change of sign: we are talking about losses!). The theory presented in the latter section advocates the use of the generalized Pareto distribution as a natural parametric model in this case.

Example 4.8 (Fighting the arch-enemy with mathematics)
The above heading is the actual title of an interesting paper by de Haan [5] on the famous Dutch dyke project following the disastrous flooding of parts of the Dutch provinces of Holland and Zeeland on February 1, 1953, killing over 1800 people. In it, de Haan gives an account of the theoretical and applied work done in connection with the problem of how to determine a safe height for the sea dykes in the Netherlands. More than with any other event, the resulting work by Dutch mathematicians under van Dantzig gave the statistical methodology of extremal events a decisive push. The statistical analyses also made a considerable contribution to the final decision making about the dyke heights. The problem faced was the following: given a small number p (in the range of $10^{-4}$ to $10^{-3}$), determine the height of the sea dykes such that the probability that there is a flood in a given year equals p. Again, we are confronted with a quantile estimation
problem. From the data available, it was clear that one needed estimates well outside the range of the data. The seawater level in the Netherlands is typically measured in (N.A.P. + $x$) m (N.A.P. = Normal Amsterdam Peil, the Dutch reference level corresponding to mean sea level). The 1953 flood was caused by a (N.A.P. + 3.85) m surge, whereas historical accounts estimate a (N.A.P. + 4) m for the 1570 flood, the worst recorded. The van Dantzig report estimated the $(1 - 10^{-4})$–quantile as (N.A.P. + 5.14) m for the annual maximum. That is, the one–in–ten–thousand–year surge height is estimated as (N.A.P. + 5.14) m. We urge all interested in extreme value statistics to read de Haan [5].

Many more examples with an increasing degree of complexity could have been given including:

- non–stationarity (seasonality, trends),
- sparse data,
- multivariate observations,
- infinite-dimensional data (for instance continuously monitored processes).

The literature cited throughout contains a multitude of examples. Besides the work mentioned already by Smith on the river Nidd and de Haan’s paper on the dyke project, we call the following papers to the reader’s attention:

- Rootzén and Tajvidi [11] where a careful analysis of Swedish wind storm losses (i.e. insurance data) is given. Besides the use of standard methodology (fitting of generalised extreme value and Pareto distributions), problems concerning trend analysis enter, together with a covariate analysis looking at the potential influence from numerous environmental factors.

- Resnick [10] considers heavy tail modelling in a huge data set ($n \approx 50000$) in the field of the teletraffic industry. Besides giving a very readable and thought provoking review of some of the classical methods, extremes in time series models are specifically addressed.

- Smith [12] applies extreme value theory to the study of ozone in Houston, Texas. A key question concerns the detection of a possible trend in ground–level ozone. Such a study is particularly interesting as air–quality standards are often formulated in terms of the highest level of permitted emissions.
The above papers are not only written by masters at their trade (de Haan, Resnick, Rootzén, Smith), they also cover a variety of fields (hydrology, insurance, electrical engineering, environmental research).

Within the context of finance, numerous papers analysing specific data are being published; see Figure 4.7 for a typical example of financial return data. A paper which uses up-to-date statistical methodology on extremes is for instance Danelson and de Vries [2] where models for high frequency foreign exchange recordings are treated. See also Müller et al. [7] for more background on the data. Interesting case studies are also to be found in Barnett and Turkman [1], Falk, Hüsler and Reiss [3], and Longin [6]. The latter paper analyses US stock market data.

We hope that the examples above have singled out a series of problems. We now want to present their statistical solutions. There is no way in which we can achieve completeness concerning the statistical models now understood: the definitive book on this still awaits the writing.

The following sections should offer the reader both hands-on experience of some basic methods, as well as a survival kit to get him/her safely through the “jungle of papers on extreme value statistics”. The outcome should be a better understanding of those basic methods, together with a clear(er) overview of where the field is heading to. This chapter should also be a guide on where to look for further help on specific problems at hand.

Of the more modern textbooks containing a fair amount of statistical techniques we would like to single out Falk et al. [3] and Reiss [9]. The latter book also contains a large amount of historical notes. *It always pays to go back to the early papers and books written by the old masters*, and the annotated references in Reiss [9] could be your guide. *However, whatever you decide to read, don’t miss out on Gumbel [4]*!

References


4.2 Exploratory data analysis for extremes

One of the reasons why Gumbel’s book [3] is such a feast to read is its inclusion of roughly 100 graphs and 50 tables. The author very much stresses the importance of looking at data before engaging in a detailed statistical analysis. In our age of nearly unlimited computing power this graphical data exploration is becoming increasingly important. The reader interested in some recent developments in this area may for instance consult Chambers et al. [1], Cleveland [2] or Tufts [4]. In the sections to follow we discuss some of the more useful graphical methods.

References


4.2.1 Quantile-quantile plots

We learned in Section 2.1.2 about QQ-plots. They can be used as an exploratory tool to discover how heavy the tail of the distribution of the data is. It is not a precise statistical tool and only gives some indication about the fatness of the tail. So far we have considered only location-scale

![QQ-plot graphs](a) Gumbel distributed simulated data versus Gumbel distribution. GEV distributed data with parameters (b): $\xi = 0.3$, (c): $\xi = -0.3$, (d): $\xi = 0.7$, versus Gumbel. The values $\xi = 0.7$ and $\xi = 0.3$ are chosen so that $\alpha = 1/\xi$ either belongs to the range $(1,2)$ (typically encountered for insurance data) or $(3,4)$ (corresponding to many examples in finance).
families. In the case of the generalized extreme value distribution (GEV), see Definition 3.49,
\[
H_{\xi,\mu,\psi}(x) = \exp \left\{ - \left( 1 + \frac{x - \mu}{\psi} \right)^{-1/\xi} \right\}, \quad 1 + \xi (x - \mu) / \psi > 0,
\]
besides the location and scale parameters $\mu \in \mathbb{R}, \psi > 0$, a shape parameter $\xi \in \mathbb{R}$ enters, making immediate interpretation of a QQ-plot more delicate. Recall that $\Phi_\alpha, \Psi_\alpha$, and $\Lambda$ denote the standard extreme value distributions Fréchet, Weibull and Gumbel; see Definition 3.12. A preferred method for testing graphically whether our sample comes from $H_{\xi,\mu,\psi}$ would be to first obtain an estimate $\hat{\xi}$ for $\xi$ either by guessing or by one of the methods given below, and consequently work out a QQ-plot using $H_{\hat{\xi},0,1}$ where again $\mu$ and $\psi$ may be estimated either by visual inspection or through linear regression. These preliminary estimates are often used as starting values in numerical iteration procedures.

4.2.2 The mean excess function

Another useful graphical tool, in particular for discrimination in the tails, is the mean excess function. Note that we have already introduced this function in the context of the GEV; see Definition 3.51. We recall it here for convenience.

**Definition 4.10** (Mean excess function)
Let $X$ be a random variable with right endpoint $x_F$; then
\[
e(u) = E(X - u \mid X > u), \quad 0 \leq u < x_F,
\]
is called the mean excess function of $X$. □

The quantity $e(u)$ is often referred to as the mean excess over the threshold value $u$. This interpretation will be crucial in the sequel. In an insurance context, $e(u)$ can be interpreted as the expected claim size in the unlimited layer, over priority $u$. Here $e(u)$ is also called the mean excess loss function. In a reliability or medical context, $e(u)$ is referred to as the mean residual life function. In a financial risk management context, switching from the right tail to the left tail, $e(u)$ is referred to as the shortfall. A summary of the most important mean excess functions is to be found in Table 3.52.

In Example 3.53 we already noted that any continuous distribution function $F$ is uniquely determined by its mean excess function; see (3.44) and (3.45) for the relevant formulae linking $F$ to $e$ and vice versa.

**Example 4.12** (Some elementary properties of the mean excess function)
If $X$ is $Exp(\lambda)$ distributed, then $e(u) = \lambda^{-1}$ for all $u > 0$. Now assume that $X$ is a random variable with support unbounded to the right and distribution function $F$. If for all $y \in \mathbb{R}$,
\[
\lim_{x \to \infty} \frac{F(x - y)}{F(x)} = e^{\gamma y},
\]

71
Weibull: \( \tau < 1 \) or lognormal

or lognormal

Gamma: \( \alpha > 1 \)

Exponential

\[ e(u) = \int_u^{\infty} F(y) \, dy / F(u) \]

and apply Karamata’s theorem (see (2.14) and (2.15)) to \( F \circ \ln \). Notice that for \( F \in S \) (the class of subexponential distributions) (4.3) is satisfied with \( \gamma = 0 \) so that in this heavy-tailed case, \( e(u) \) tends to \( \infty \) as \( u \to \infty \). On the other hand, superexponential functions of the type \( F(x) \sim \exp\{-x^\alpha\} \), \( \alpha > 1 \), satisfy the limit relation (4.3) with \( \gamma = \infty \) so that the mean excess function tends to 0.

**Example 4.13** Recall that for \( X \) generalised Pareto the mean excess function is linear; see Theorem 3.58(e). The mean excess function of a heavy-tailed distribution function, for large values of the argument, typically appears to be between a constant function (for \( \text{Exp}(\lambda) \)) and a straight line with positive slope (for the Pareto case).

It was noticed by Benktander that interesting claim size distributions correspond to mean excess functions of the form

\[
e(u) = \begin{cases} 
  u^{1-\beta} / \alpha, & \alpha > 0, 0 \leq \beta < 1, \\
  u / (\alpha + 2\beta \ln u), & \alpha, \beta > 0.
\end{cases}
\]

Note that \( e(u) \) increases but the rate of increase decreases with \( u \). As a consequence, Benktander introduced the two families of distributions which, within the insurance world, now bear his name. The Benktander-type-I and -type-II classes are defined in Table 2.12.

A graphical test for tail behaviour can now be based on the empirical mean excess function \( e_n(u) \). Suppose that \( X_1, \ldots, X_n \) are iid with distribution function \( F \) and let \( F_n \) denote the empirical
Figure 4.14 The empirical mean excess function $e_n(u)$ of simulated data ($n = 1000$) compared with the corresponding theoretical mean excess function $e(u)$ (dashed line): standard exponential (top), lognormal (middle) with $\ln X \sim N(0, 4)$, Pareto (bottom) with tail index 1.7.

distribution function and $\Delta_n(u) = \{i : i = 1, \ldots, n, X_i > u\}$, then

$$
e_n(u) = \frac{1}{F_n(u)} \int_u^\infty F_n(y) \, dy = \frac{1}{\text{card} \Delta_n(u)} \sum_{i \in \Delta_n(u)} (X_i - u), \quad u \geq 0,
$$

(4.4)

with the convention that $0/0 = 0$. A mean excess plot (ME-plot) then consists of the graph

$$
\left\{(X_{(k)}, e_n(X_{(k)})) : k = 1, \ldots, n\right\}.
$$

The statistical properties of $e_n(u)$ can again be derived by using the relevant empirical process theory. For our purposes, the ME-plot is used only as a graphical method, mainly for distinguishing between light- and heavy-tailed models; see Figure 4.14 for some simulated examples. Indeed caution is called for when interpreting such plots. Due to the sparseness of the data available for

73
calculating $e_n(u)$ for large $u$-values, the resulting plots are very sensitive to changes in the data towards the end of the range; see for instance Figure 4.15. For this reason, more robust versions like median excess plots and related procedures have been suggested; see for instance Beirlant, Teugels and Vynckier [1] or Rootzén and Tajvidi [3]. For a critical assessment concerning the use of mean excess functions in insurance see Rytgaard [4].

**Example 4.16** (Exploratory data analysis for some examples from insurance and finance)

In Figures 4.17–4.19 we have graphically summarised some properties of three real data-sets. Two come from insurance, one from finance. The data underlying Figure 4.18 correspond to Danish fire insurance claims in millions of Danish Kroner (1985 prices). The data were communicated to us by Mette Rytgaard and correspond to the period 1980–1990, inclusively. There is a total of $n = 2493$ observations.

The second insurance data, presented in Figure 4.19, correspond to a portfolio of industrial fire data ($n = 8043$) reported over a two year period. This data-set is definitely considered by the portfolio manager as "dangerous", i.e., large claim considerations do enter substantially in the final premium calculation.

<table>
<thead>
<tr>
<th>Data</th>
<th>Danish</th>
<th>Industrial</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>2493</td>
<td>8043</td>
</tr>
<tr>
<td>min</td>
<td>0.3134</td>
<td>0.003</td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.157</td>
<td>0.587</td>
</tr>
<tr>
<td>median</td>
<td>1.634</td>
<td>1.526</td>
</tr>
<tr>
<td>mean</td>
<td>3.063</td>
<td>14.65</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>2.645</td>
<td>4.488</td>
</tr>
<tr>
<td>max</td>
<td>263.3</td>
<td>13520</td>
</tr>
<tr>
<td>$\hat{a}_{0.99}$</td>
<td>24.61378</td>
<td>184.0009</td>
</tr>
</tbody>
</table>

**Figure 4.15** The mean excess function of the Pareto distribution $F(x) = x^{-1.7}$, $x \geq 1$, together with 20 empirical mean excess functions $e_n(u)$ each based on simulated data ($n = 1000$) from the above distribution. Note the very unstable behaviour, especially towards the higher values of $u$. This is typical and makes the precise interpretation of $e_n(u)$ difficult; see also Figure 4.14.
Figure 4.17 Exploratory data analysis of BMW share prices. Top: the 500 largest values from the upper and lower tails. Middle: the corresponding log–histograms. Bottom: the ME–plot. See Example 4.16 for some comments.

Table 4.20 Basic statistics for the Danish and the industrial fire data; \(\hat{x}_{0.99}\) stands for the empirical 99%-quantile.

A first glance at the figures and Table 4.20 for both data-sets immediately reveals heavy-tailedness and skewedness to the right. The corresponding mean excess functions are close to a straight line which indicates that the underlying distributions may be modeled by Pareto-like distribution functions. The QQ-plots against the standard exponential quantiles also clearly show tails much heavier than exponential ones.

Whereas the insurance data may be supposed to represent iid observations, this is definitely not the case for the BMW daily log-return data underlying Figure 4.17. For the full data-set see Figure 4.7. The period covered is January 23, 1973 - July 12, 1996, resulting in \(n = 6,146\) observations on the log-returns. Nevertheless, we may certainly assume stationarity of the underlying times series so that many limit results (such as the SLLN) remain valid under general conditions. This would allow us to interpret the graphs of Figure 4.17 in a way similar to the iid case, i.e., we will assume
that the empirical plots (histogram, empirical mean excess function, QQ-plot) are close to their theoretical counterparts. Note that we contrast these tools for the positive daily log-returns and the absolute values of the negative ones. The log-histograms again show skewness to the right and heavy-tailedness. It is interesting to observe that the left and right tail of the distribution of the log-returns are different. Indeed, both the histograms and the ME-plots (mind the different slopes) indicate that the left tail of the distribution is heavier than the right one.

In Figure 4.17 we have singled out the 500 largest positive (left) and negative (right) log-returns over the above period. In Table 4.21 we have summarized some basic statistics for the three resulting data-sets: BMW—all, BMW-upper and BMW-lower. The nomenclature should be obvious.

We would like to stress that it is our aim to fit tail-probabilities (i.e., probabilities of extreme returns). Hence it is natural for such a fitting to disregard the “small” returns. The choice of 500 at this point is rather arbitrary; we will come back to this issue and indeed a more detailed analysis later.

Comments

The importance of the mean excess function (or plot) as a diagnostic tool for insurance data is nicely demonstrated in Hogg and Klugman [2]; see also Beirlant et al. [1] and the references therein.
Figure 4.19 Exploratory data analysis of insurance claims caused by industrial fire: the data (top left), the histogram of the log-transformed data (top right), the ME-plot (bottom left) and a QQ-plot against standard exponential quantiles (bottom right). See Example 4.16 for some comments.

References


4.2.3 The return period

In this section we are interested in answering the question:

What is the mean waiting time between specific extremal events?

This question is usually made precise in the following way. Let \((X_i)\) be a sequence of iid random variables with continuous distribution function \(F\) and \(u\) a given threshold. We consider the sequence \((I_{(X_i > u)})\) of iid Bernoulli random variables with success probability \(p = F(u)\). Consequently, the time of the first success

\[
L(u) = \min \{ i \geq 1 : X_i > u \}
\]
\[ P(L(u) = k) = (1 - p)^{k-1}p, \quad k = 1, 2, \ldots. \]

Notice that the random variables
\[ L_1(u) = L(u), \quad L_{n+1}(u) = \min \{ i > L_n(u) : X_i > u \} - L_n(u), \quad n \geq 1, \]
describe the time periods between two consecutive exceedances of \( u \) by \( (X_n) \). The \textit{return period of the events} \( \{ X_i > u \} \) is then defined as \( EL(u) = p^{-1} = (F(u))^{-1} \), which increases to \( \infty \) as \( u \to \infty \). For ease of notation we take distribution functions with unbounded support above. All relevant questions concerning the return period can now be answered straightforwardly through the corresponding properties of the geometric distribution. Below we give some examples.

Define
\[ r_k = P(L(u) \leq k) = p \sum_{i=1}^{k} (1 - p)^{i-1} = 1 - (1 - p)^k, \quad k \in \mathbb{N}. \]

Hence \( r_k \) is the probability that there will be at least one exceedance of \( u \) before time \( k \) (or within \( k \) observations). This gives a 1–1 relationship between \( r_k \) and the return period \( p^{-1} \).

One is often interested in the probability that there will be an exceedance of \( u \) \textit{before} the return period. Hence this probability becomes
\[ P(L(u) \leq EL(u)) = P(L(u) \leq \lfloor 1/p \rfloor) = 1 - (1 - p)^{\lfloor 1/p \rfloor}, \]
where \( \lfloor x \rfloor \) denotes the integer part of \( x \). For high thresholds \( u \), i.e., for \( u \uparrow \infty \) and consequently \( p \downarrow 0 \), we obtain
\[
\lim_{u \uparrow \infty} P(L(u) \leq EL(u)) = \lim_{p \downarrow 0} \left( 1 - (1 - p)^{\lfloor 1/p \rfloor} \right) = 1 - e^{-1} = 0.63212.
\]
This shows that for high thresholds the mean of \( L(u) \) (the return period) is larger than its median.

**Example 4.22** (Return period, \( t \)-year event)
Within an insurance context, a structure is to be insured on the basis that it will last at least 50 years with no more than 10% risk of failure. What does this information imply for the return period? Using the language above, the engineering requirement translates into
\[ P(L(u) \leq 50) \leq 0.1. \]
Here we tacitly assumed that a structure failure for each year \( i \) can be modelled through the event \( \{ X_i > u \} \), where \( X_i \) is a structure-dependent critical component, say. We assume the iid property of the \( X_i \). The above condition, solved for \( P(L(u) \leq 50) = 1 - (1 - p)^{50} = 0.1 \), now immediately implies that \( p = 0.002105 \), i.e., \( EL(u) = 475 \). In insurance language one speaks in this case about a 475-year event.

The important next question concerns the implication of a \( t \)-year event requirement on the underlying threshold value. By definition this means that for the corresponding threshold \( u_t \),

\[ t = EL(u_t) = \frac{1}{F(u_t)}, \]

hence

\[ u_t = F^{-1}(1 - t^{-1}). \]

In the present example, \( u_{475} = F^{-1}(0.9978) \). This leads us once more to the crucial problem of high quantile estimation. \( \square \)

**Example 4.23** (Continuation of Example 4.8)

In the case of the Dutch dyke example, recall that, assuming stationarity among the annual maxima of sea levels, the last comparable flood before 1953 took place in November 1570, so that in the above language one would speak about a 382-year event. The 1953 level hence corresponds roughly to the \((1 - 1/382)\)-quantile of the distribution of the annual maximum. The subsequent government requirements demanded dykes to be built corresponding to a 1000-to-10000-year event! \( \square \)

The above examples clearly stress the need for a solution to the following problems:

- Find reliable estimators for high quantiles from iid data.
- Because of increasing safety requirements, implying \( t \)-year events with increasingly higher \( t \), the iid assumption may not always be tenable. Moreover, most data in practice will exhibit dependence and/or non-stationarity. Find therefore quantile estimation procedures for non-iid data.

**Comments**

Return periods and \( t \)-year events have a long history in hydrology; see for instance Castillo [2] and Rosbjerg [4]. For relevant statistical techniques coming more from a reliability context, see Crowder et al. [3]; methods more related to medical statistics are to be found in Andersen et al. [1].

**References**


4.2.4 Records as an exploratory tool

Suppose that the random variables \(X_i\) are iid with distribution function \(F\). A record \(X_n\) occurs if \(X_n > M_{n-1} = \max(X_1, \ldots, X_{n-1})\). By definition we take \(X_1\) as a record. Record times \(L_n\) are the random times at which the process \((M_n)\) jumps. Define the record counting process as

\[
N_1 = 1, \quad N_n = 1 + \sum_{k=2}^{n} I_{\{X_k > M_{k-1}\}}, \quad n \geq 2.
\]

The following result (on the mean \(EN_n\)) may be surprising.

**Lemma 4.24** (Moments of \(N_n\))

Suppose \((X_i)\) are iid with continuous distribution function \(F\) and \((N_n)\) defined as above. Then

\[
EN_n = \sum_{k=1}^{n} \frac{1}{k} \quad \text{and} \quad \text{var}(N_n) = \sum_{k=1}^{n} \left( \frac{1}{k} - \frac{1}{k^2} \right).
\]

**Proof.** From the definition of \(N_n\) we obtain

\[
EN_n = 1 + \sum_{k=2}^{n} P(X_k > M_{k-1}) = 1 + \sum_{k=2}^{n} \int_{-\infty}^{+\infty} P(X_k > u) dP(M_{k-1} \leq u).
\]

Now use \(P(M_{k-1} \leq u) = F^{k-1}(u)\) and change the variable \(F(u)\) to \(x\), say (here the continuity of \(F\) comes in):

\[
EN_n = 1 + \sum_{k=2}^{n} \int_{0}^{1} (1 - x) d(x^{k-1}) = \sum_{k=1}^{n} k^{-1}.
\]

This yields the result for \(EN_n\). A similar argument works for \(\text{var}(N_n)\). \(\square\)

Notice that \(EN_n\) and \(\text{var}(N_n)\) are both of the order \(\ln n\) as \(n \to \infty\). More precisely, \(EN_n - \ln n \to \gamma\), where \(\gamma = 0.5772\ldots\) denotes Euler’s constant. As a consequence:

*the number of records of iid data grows very slowly!*

Before reading further, guess the answer to the next question:

*How many records do we expect in 100, 1000 or 10,000 iid observations?*

Table 4.25 contains the somewhat surprising answer.

**Example 4.27** (Records in real data)

In Figure 4.26 the total amount of sunshine hours in Vancouver during the month of July from 1909 until 1973 is given. The data are taken from Glick [2]. There are 6 records in these \(n = 64\) observations, namely for \(i = 1, 2, 6, 10, 23, 53\). Clearly one would need a much larger \(n\) in order to test confidently the iid hypothesis for the underlying data \(X_1, \ldots, X_{64}\) on the basis of the record values. If the data were iid, then we would obtain \(EN_{64} = 4.74\). The observed value of 6 agrees rather well. On the basis of these observations we have no reason to doubt the iid hypothesis. \(\square\)
\[
\begin{array}{|c|c|c|c|c|}
\hline
n = 10^k, \ k = & E N_n & \ln n & \ln n + \gamma & D_n \\
\hline
1 & 2.9 & 2.3 & 2.9 & 1.2 \\
2 & 5.2 & 4.6 & 5.2 & 1.9 \\
3 & 7.5 & 7.0 & 7.5 & 2.4 \\
4 & 9.8 & 9.2 & 9.8 & 2.8 \\
5 & 12.1 & 11.5 & 12.1 & 3.2 \\
6 & 14.4 & 13.8 & 14.4 & 3.6 \\
7 & 16.7 & 16.1 & 16.7 & 3.9 \\
8 & 19.0 & 18.4 & 19.0 & 4.2 \\
9 & 21.3 & 20.7 & 21.3 & 4.4 \\
\hline
\end{array}
\]

Table 4.25 Expected number of records \( E N_n \) in an iid sequence \( (X_n) \), together with the asymptotic approximations \( \ln n \), \( \ln n + \gamma \), and standard deviation \( D_n = \sqrt{\text{var}(N_n)} \), based on Lemma 4.24.

Notice that the above results on the expectation and variance of the number of records are independent of the distribution function \( F ! \). Continuity of \( F \) is the only assumption required. Some further intuition about the structure of records and record times is provided by the following considerations.

A record happens when there is a jump in the sequence \( (M_n) \) of partial maxima. The random times when the jumps happen (the record times) are denoted by \( L_1 < L_2 < \cdots \). Thus the sequence of records is given by \( (X_{L_n}) \). Now assume that \( F \) is continuous. Then the function \( R(x) = -\ln F(x) \) is continuous and non-decreasing and hence its generalized inverse (see p. 7)

\[
R^-(t) = \inf \{x : R(x) \geq t\}, \quad t > 0,
\]

is increasing. Direct calculation yields

\[
X_1 \overset{d}{=} R^-(E_1),
\]

where \( E_1 \) is a standard exponential random variable. Indeed,

\[
P (R^-(E_1) \leq x) = P (E_1 \leq R(x)) = 1 - e^{-R(x)} = F(x).
\]

Hence the sequences \( (M_n) \) and

\[
\left( \bigvee_{i=1}^{n} R^-(E_i) \right) = \left( R^-( \bigvee_{i=1}^{n} E_i ) \right)
\]
have the same distribution, where \((E_i)\) is an iid standard exponential sequence. Moreover, denoting by \((\tilde{L}_n) (= (L_n))\) the record times of the sequence \((R^\leftarrow (E_i))\), we have for the sequences of records that
\[
(X_{L_n}) \overset{d}{=} \left( R^\leftarrow \left( E^\leftarrow_{L_n} \right) \right).
\]

An immediate consequence of these observations is that the counting processes
\[
\tilde{N}(t) = \# \{ i \geq 1 : L_i \leq t \} \quad \text{and} \quad \tilde{N}(t) = \# \{ i \geq 1 : \tilde{L}_i \leq t \}, \quad t \geq 0,
\]
have the same distribution. In other words:

*The distribution of the number of records of an iid sequence with a continuous distribution function \(F\) in a given time interval does not depend on \(F\).*

If \(F\) is standard exponential then the records \((R^\leftarrow (E^\leftarrow_{L_n})) = (E^\leftarrow_{L_n})\) are the points of a homogeneous Poisson process on \([0, \infty)\) with intensity 1, i.e., the process
\[
N(t) = \# \{ i \geq 1 : E^\leftarrow_{L_i} \leq t \}, \quad t \geq 0,
\]
is a homogeneous Poisson process in the sense of Section 2.2.1. This follows from the observation that \((E^\leftarrow_{L_n})\) is Markov with transition probabilities \(\pi(x, (y, \infty)) = \exp\{-y - x\}\); see Resnick [5], Proposition 4.1.

It is not difficult to see that for a general iid sequence \((X_n)\) the process
\[
N_R(t) = \# \{ i \geq 1 : R^\leftarrow (E^\leftarrow_{L_i}) \leq t \}
\]
\[= \# \{ i \geq 1 : E^\leftarrow_{L_i} \leq R(t) \}
\]
\[= N(R(t)), \quad t \geq 0,
\]
constitutes an inhomogeneous Poisson process with mean measure of the interval \((a, b)\) given by
\[
\mu_R(a, b) = E(N_R(a, b)) = E[N_R(b) - N_R(a)] = E[N(R(b)) - N(R(b))] = R(b) - R(a).
\]

The latter relation only makes sense for \((a, b) \subset (x^*_F, x^*_F)\), where \(x^*_F\) and \(x^*_F\) are the left and right endpoints of the distribution \(F\). The inhomogeneity of the Poisson process \(N_R\) refers to the fact that \(N_R(a_i, b_i)\) are independent Poisson(\(\mu_R(a_i, b_i)\)) random variables provided \((a_i, b_i) \subset (x^*_F, x^*_F)\) are disjoint intervals.

We summarize as follows.

**Theorem 4.28** (Point process description of records)

*Let \(F\) be a continuous distribution function with left endpoint \(x^*_F\) and right endpoint \(x^*_F\). Then the records \((X_{L_n})\) of the iid sequence \((X_n)\) are the points of a Poisson process on \((x^*_F, x^*_F)\) with mean measure \(\mu_R\) given by
\[
\mu_R(a, b) = R(b) - R(a), \quad x^*_F < a \leq b < x^*_F, \quad \text{where} \quad R(x) = -\ln F(x).
\]

In particular, if \(F\) is standard exponential then \(R(t) = t\) and \((X_{L_n})\) constitute the points of a homogeneous Poisson process on \([0, \infty)\) with intensity 1.*
Comments

Since the fundamental paper by Foster and Stuart [1] numerous papers have been published on records; see for instance Pfeifer [4], Kapitel 4, Resnick [5], Chapter 4, and the references cited therein; see also Goldie and Resnick [3]. We find Glick [2] a very entertaining introduction. Smith [6] gives more information on statistical inference for records, especially in the non-iid case.

The exploratory techniques introduced so far all started from an iid assumption on the underlying data. Their interpretation becomes hazardous when applied in the non-iid case, as for instance to data exhibiting a trend. Various statistical de-trending techniques exist within the realm of regression theory and time series analysis. These may range from fitting of a deterministic trend to the data, averaging, differencing, ... 

References


4.3 Parameter estimation for the generalized extreme value distribution

Recall from (4.1) the generalized extreme value distribution (GEV)

\[ H_{\xi,\mu,\psi}(x) = \exp \left\{ - \left( 1 + \xi \frac{x - \mu}{\psi} \right)^{-1/\xi} \right\}, \quad 1 + \xi \frac{x - \mu}{\psi} > 0. \]

As usual the case \( \xi = 0 \) corresponds to the Gumbel distribution

\[ H_{0,\mu,\psi}(x) = \exp \left\{ - e^{-(x-\mu)/\psi} \right\}, \quad x \in \mathbb{R}. \]

The parameter \( (\xi, \mu, \psi) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \) consists of a shape parameter \( \xi \), location parameter \( \mu \) and scale parameter \( \psi \). For notational convenience, we shall either write \( H_\xi \) or \( H_\theta \) depending on the case in hand. In Theorem 3.50 we saw that \( H_\xi \) arises as the limit distribution of normalized maxima of iid random variables. Standard statistical methodology from parametric estimation theory is available if our data consist of a sample

\[ X_1, \ldots, X_n \text{ iid from } H_\theta. \]

We mention here that the assumption of \( X_i \) having an exact extreme value distribution \( H_\xi \) is perhaps not the most realistic one. In the next section we turn to a more tenable assumption that the \( X_i \) are approximately \( H_\xi \) distributed. The “approximately” will be interpreted as “belonging to the maximum domain of attraction of”.

83
Fitting of annual maxima

As already discussed in Example 4.1, data of the above type may become available when the $X_i$ can be interpreted as maxima over disjoint time periods of length $s$, say. In hydrology, which is the cradle of many of the ideas for statistics of extremal events, this period mostly consists of one year; see for instance the river Nidd data in Figure 4.2. The 1-year period is chosen in order to compensate for intra-year seasonalities. Therefore the original data may look like

$$
\mathbf{x}^{(i)} = \left( X_1^{(i)}, \ldots, X_s^{(i)} \right),
\mathbf{x}^{(2)} = \left( X_1^{(2)}, \ldots, X_s^{(2)} \right),
\ldots,
\mathbf{x}^{(n)} = \left( X_1^{(n)}, \ldots, X_s^{(n)} \right),
$$

where the vectors $(\mathbf{x}^{(i)})$ are assumed to be iid, but within each vector $\mathbf{x}^{(i)}$ the various components may (and mostly will) be dependent. The time length $s$ is chosen so that the above conditions are likely to be satisfied. The basic iid sample from $H_0$ on which statistical inference is to be performed then consists of

$$(4.8) \quad X_i = \max(X_1^{(i)}, \ldots, X_s^{(i)}), \quad i = 1, \ldots, n. \quad \quad \text{fitting of annual maxima.}$$

Below we discuss some of the main techniques for estimating $\theta$ in the exact model (4.7).

4.3.1 Maximum likelihood estimation

The set-up (4.7) corresponds to the standard parametric case of statistical inference and hence in principle can be solved by maximum likelihood methodology. Suppose that $H_0$ has density $h_\theta$. Then the likelihood function based on the data $\mathbf{x} = (X_1, \ldots, X_n)$ is given by

$$
L(\theta; \mathbf{x}) = \prod_{i=1}^n h_\theta(X_i) I_{\{1+\xi(X_i-\mu)/\psi > 0\}}.
$$

Denote by $\ell(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x})$ the log-likelihood function. The maximum likelihood estimator (MLE) for $\theta$ then equals

$$
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ell(\theta; \mathbf{x}),
$$

i.e., $\hat{\theta}_n = \hat{\theta}_n (X_1, \ldots, X_n)$ maximizes $\ell(\theta; \mathbf{x})$ over an appropriate parameter space $\Theta$. In the case of $H_{0,\mu,\psi}$ this gives us

$$
\ell((0, \mu, \psi); \mathbf{x}) = -n \ln \psi - \sum_{i=1}^n \exp \left\{ -\frac{X_i - \mu}{\psi} \right\} - \sum_{i=1}^n \frac{X_i - \mu}{\psi}.
$$
Differentiating the latter function with respect to $\mu$ and $\psi$ yields the likelihood equations in the Gumbel case:

\[
0 = n - \sum_{i=1}^{n} \exp \left\{ -\frac{X_i - \mu}{\psi} \right\},
\]

\[
0 = n + \sum_{i=1}^{n} \frac{X_i - \mu}{\psi} \left( \exp \left\{ -\frac{X_i - \mu}{\psi} \right\} - 1 \right).
\]

Clearly no explicit solution exists to these equations. The situation for $H_\xi$ when $\xi \neq 0$ is even more complicated, so that numerical procedures are called for. Jenkinson [3] and Prescott and Waklen [6, 7] suggest variants of the Newton–Raphson scheme. With the existence of the Fortran algorithm published in Hosking [2] and its supplement in Macleod [5], the numerical calculation of the MLE $\theta_n$ for general $H_0$ poses no serious problem in principle.

Notice that we said in principle. Indeed in the so-called regular cases maximum likelihood estimation offers a technique yielding efficient, consistent and asymptotically normal estimators. See for instance Cox and Hinkley [1] and Lehmann [4] for a general discussion on maximum likelihood estimation. Relevant for applications in extreme value theory, typical non-regular cases may occur whenever the support of the underlying distribution depends on the unknown parameters. Therefore, although we have reliable numerical procedures for finding the MLE $\hat{\theta}_n$, we are less certain about its properties, especially in the small sample case. For a discussion on this point see Smith [8]. In the latter paper it is shown that the classical (good) properties of the MLE hold whenever $\xi > -1/2$; this is not the case for $\xi \leq -1/2$.

As most distributions encountered in insurance and finance have support unbounded to the right (this is possible only for $\xi \geq 0$), the MLE technique offers a useful and reliable procedure in those fields.

At this point we would like to quantify a bit more the often encountered statement that for applications in insurance (and finance for that matter) the case $\xi \geq 0$ is most important. Clearly, all financial data must be bounded to the right; an obvious (though somewhat silly) bound is total wealth. The main point however is that in most data there does not seem to be clustering towards a well-defined upper limit but more a steady increase over time of the underlying maxima. The latter would then, for iid data, much more naturally be modelled within $\xi \geq 0$. A typical example is to be found in the Danish fire insurance data of Figure 4.18.

An example where a natural upper limit may exist is given in Figure 4.29. The data underlying this example correspond to a portfolio of water–damage insurance. In contrast to the industrial fire data of Figure 4.19, in this case the portfolio manager realises that large claims only play a minor role. Though the data again show an increasing ME–plot, for values above 5000, the mean excess losses are growing much slower than to be expected from a really heavy-tailed model, unbounded to the right. The ME–plot for these data should be compared with those for the Danish fire data (Figure 4.18) and the industrial fire data (Figure 4.19).

References


Figure 4.29 Exploratory data analysis of insurance claims caused by water: the data (top, left), the histogram of the log-transformed data (top, right), the ME-plot (bottom). Notice the kink in the ME-plot in the range (5,000, 6,000) reflecting the fact that the data seem to cluster towards some specific upper value.


4.3.2 Some theory about order statistics

So far we have based maximum likelihood estimation only on $n$ iid observations of maxima which we have assumed to follow exactly a GEV $H_\theta$. By appropriately defining the underlying time periods, we design independence into the model; see (4.8). Suppose now that, rather than just having the largest observation available, we possess the $k$ largest of each period (year, say). It is then natural to base the estimation on a finite number of upper order statistics instead of the maxima only.

For this reason we make a short excursion to order statistics and study some of their elementary properties.
Let $X, X_1, X_2, \ldots$ denote a sequence of iid non-degenerate random variables with common distribution function $F$. In this section we summarize some important properties of the upper order statistics of a finite sample $X_1, \ldots, X_n$. To this end we define the ordered sample

$$X_{(1)} \leq \cdots \leq X_{(n)}.$$ 

Hence $X_{(1)} = \min(X_1, \ldots, X_n)$ and $X_{(n)} = M_n = \max(X_1, \ldots, X_n)$. The random variable $X_{(n-k+1)}$ is called the $k$th upper order statistic. The notation for order statistics varies; some authors denote by $X_{(n)}$ the minimum and by $X_{(1)}$ the maximum of a sample. This leads to different representations of quantities involving order statistics.

The relationship between the order statistics and the empirical distribution function of a sample is immediate: for $x \in \mathbb{R}$ we introduce the empirical distribution function or sample distribution function

$$F_n(x) = \frac{1}{n} \text{card} \{i : 1 \leq i \leq n, X_i \leq x\} = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}, \quad x \in \mathbb{R},$$

where $I_A$ denotes the indicator function of the set $A$. Now

$$X_{(n-k+1)} \leq x \quad \text{if and only if} \quad \sum_{i=1}^{n} I_{\{X_i > x\}} < k,$$

which implies that

$$P \left( X_{(n-k+1)} \leq x \right) = P \left( F_n(x) > 1 - \frac{k}{n} \right).$$

Upper order statistics estimate tails and quantiles, and also exceed probabilities over certain thresholds. Recall the definition of the quantile function of the distribution function $F$

$$F^\leftarrow(t) = \inf \{x \in \mathbb{R} : F(x) \geq t\}, \quad 0 < t < 1.$$ 

87
For a sample $X_1, \ldots, X_n$ we denote the empirical quantile function by $F_n^{\leftarrow}$. If $F$ is a continuous function, then ties in the sample occur only with probability 0 and may thus be neglected, i.e., we may assume that $X_{(1)} < \cdots < X_{(n)}$. In this case $F_n^{\leftarrow}$ is a simple function of the order statistics, namely

\begin{equation}
F_n^{\leftarrow}(t) = X_{(n-k+1)} \quad \text{for } 1 - \frac{k}{n} < t \leq 1 - \frac{k-1}{n},
\end{equation}

for $k = 1, \ldots, n$.

Next we calculate the distribution function $F_{(n-k+1)}$ of the $k$th upper order statistic explicitly.

**Proposition 4.31** (Distribution function of the $k$th upper order statistic)

For $k = 1, \ldots, n$ let $F_{(n-k+1)}$ denote the distribution function of $X_{(n-k+1)}$. Then

1. $F_{(n-k+1)}(x) = \sum_{r=0}^{k-1} \binom{n}{r} F^r(x) F^{n-r}(x)$.

2. If $F$ is continuous, then

$$F_{(n-k+1)}(x) = \int_{-\infty}^{x} f_{(n-k+1)}(z) dF(z),$$

where

$$f_{(n-k+1)}(x) = \frac{n!}{(k-1)! (n-k)!} F^{n-k}(x) F^{k-1}(x);$$

i.e., $f_{(n-k+1)}$ is a density of $F_{(n-k+1)}$ with respect to $F$.

**Proof.** (1) For $n \in \mathbb{N}$ define

$$B_n = \sum_{i=1}^{n} I_{[X_i > x]}.$$

Then $B_n$ is a sum of $n$ iid Bernoulli variables with success probability

$$EI_{[X_i > x]} = P(X_1 > x) = \mathcal{F}(x).$$

Hence $B_n$ is a binomial random variable with parameters $n$ and $\mathcal{F}(x)$. An application of (4.9) gives for $x \in \mathbb{R}$

$$F_{(n-k+1)}(x) = P(B_n < k) = \sum_{r=0}^{k-1} P(B_n = r) = \sum_{r=0}^{k-1} \binom{n}{r} \mathcal{F}^r(x) \mathcal{F}^{n-r}(x).$$

(2) Using the continuity of $F$, we calculate

\begin{align*}
\frac{n!}{(k-1)! (n-k)!} \int_{-\infty}^{x} F^{n-k}(z) \mathcal{F}^{k-1}(z) dF(z) &= \frac{n!}{(k-1)! (n-k)!} \int_{-\infty}^{1} (1-t)^{n-k} t^{k-1} dt \\
&= \sum_{r=0}^{k-1} \binom{n}{r} \mathcal{F}^r(x) \mathcal{F}^{n-r}(x) = F_{(n-k+1)}(x).
\end{align*}
The latter follows from a representation of the incomplete beta function; it can be proved by multiple partial integration. See also Abramowitz and Stegun [1], formula 6.6.4.

Similar arguments lead to the joint distribution of a finite number of different order statistics. If for instance $F$ is absolutely continuous with density $f$, then the joint density of $(X_1, \ldots, X_n)$ is

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$ 

Since the $n$ values of $(X_1, \ldots, X_n)$ can be rearranged in $n!$ ways (by absolute continuity there are a.s. no ties), every specific ordered collection $(X_{(k)})_{k=1, \ldots, n}$ could have come from $n!$ different samples. This heuristic argument can be made precise; see for instance Reiss [2], Theorem 1.4.1, or alternatively use the transformation theorem for densities. The joint density of the ordered sample becomes:

$$f_{X_{(1)}, \ldots, X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i), \quad x_1 < \cdots < x_n.$$ 

(4.11)

The following result on marginal densities is an immediate consequence of equation (4.11).

**Theorem 4.32** (Joint density of $k$ upper order statistics)

*If $F$ is absolutely continuous with density $f$, then*

$$f_{X_{(n-k+1)}, \ldots, X_{(n)}}(x_1, \ldots, x_k) = \frac{n!}{(n-k)!} F^{n-k}(x_1) \prod_{i=1}^{k} f(x_i), \quad x_1 < \cdots < x_k.$$ 

Further quantities which arise in a natural way are the *spacings*, i.e., the differences between successive order statistics. They are for instance the building blocks of Hill’s estimator of the index of regular variation which we will study soon.

**Definition 4.33** (Spacings of a sample)

*For a sample $X_1, \ldots, X_n$ the spacings are defined by*

$$X_{(n-k+1)} - X_{(n-k)}, \quad k = 1, \ldots, n-1.$$ 

*For random variables with finite left (right) endpoint $\tilde{x}_F$ ($x_F$) we define the $n$th (0th) spacing as $X_{(1)} - X_{(0)} = X_{(1)} - \tilde{x}_F$ ($X_{(n+1)} - X_{(n)} = x_F - X_{(n)}$).*

**Example 4.34** (Order statistics and spacings of exponential random variables)

Let $(E_n)$ denote a sequence of iid standard exponential random variables. An immediate consequence of (4.11) is the joint density of an ordered exponential sample $(E_{(1)}, \ldots, E_{(n)})$:

$$f_{E_{(1)}, \ldots, E_{(n)}}(x_1, \ldots, x_n) = n! \exp \left\{- \sum_{i=1}^{n} x_i \right\}, \quad 0 < x_1 < \cdots < x_n.$$ 

From this we derive the joint distribution of exponential spacings by an application of the transformation theorem for densities. Define the transformation

$$T(x_1, \ldots, x_n) = (n x_1, (n-1)(x_2 - x_1), \ldots, x_n - x_{n-1}), \quad 0 < x_1 < \cdots < x_n.$$ 

89
Then \( \det(\partial T(\mathbf{x})/\partial \mathbf{x}) = n! \) and

\[
T^{-1}(x_1, \ldots, x_n) = \left(\frac{x_1}{n}, \sum_{j=1}^{2} \frac{x_j}{n-j+1}, \sum_{j=1}^{3} \frac{x_j}{n-j+1}, \ldots, \sum_{j=1}^{n} \frac{x_j}{n-j+1}\right), \quad x_1, x_2, \ldots, x_n > 0.
\]

Then the density \( g \) of \( (nE_1, (n-1)(E_2 - E_1), \ldots, E_n - E_{n-1}) \) is of the form

\[
g(x_1, \ldots, x_n) = \frac{1}{n!} f_{E_1, \ldots, E_n} \left( \frac{x_1}{n}, \sum_{j=1}^{2} \frac{x_j}{n-j+1}, \sum_{j=1}^{3} \frac{x_j}{n-j+1}, \ldots, \sum_{j=1}^{n} \frac{x_j}{n-j+1} \right)
= \exp \left\{ -\sum_{i=1}^{n} \sum_{j=1}^{i} \frac{x_j}{n-j+1} \right\}
= \exp \left\{ -\sum_{i=1}^{n} x_i \right\}.
\]

This gives for \( i = 1, \ldots, n \) that the random variables \( i(E_{n-i+1} - E_{n-i}) \) have joint density

\[
g(x_1, \ldots, x_n) = \exp \left\{ -\sum_{i=1}^{n} x_i \right\}, \quad x_1, \ldots, x_n > 0.
\]

This implies that the spacings

\[
E_{(n)} - E_{(n-1)}, E_{(n-1)} - E_{(n-2)}, \ldots, E_{(1)}
\]

are independent and exponentially distributed, and \( E_{(n-k+1)} - E_{(n-k)} \) has mean \( 1/k \) for \( k = 1, \ldots, n \), where we recall that \( E_{(0)} = 0 \).

\(\square\)

**Example 4.35** (Order statistics property of the Poisson process)

Let \( N = (N(t))_{t \geq 0} \) be a homogeneous Poisson process with intensity \( \lambda > 0 \); for a definition see p. 11. Then the arrival times \( T_i \) of \( N \) in \( (0, t] \), conditionally on \( \{N(t) = n\} \), have the same distribution as the order statistics of a uniform sample on \( (0, t) \) of size \( n \); i.e.,

\[
P(T_1, T_2, \ldots, T_{N(t)}) \in A | N(t) = n) = P(U_{(1)}, \ldots, U_{(n)}) \in A)
\]

for all Borel sets \( A \) in \( \mathbb{R}_+ \). This property is called the *order statistics property* of the Poisson process. It gives an intuitive description of the distribution of the arrival times of a Poisson process.

For a proof we assume that \( 0 < t_1 < \cdots < t_n < t \) and \( h_1, \ldots, h_n \) are all positive, but small enough such that the intervals \( J_i = (t_i, t_i + h_i], i = 1, \ldots, n \), are disjoint. Then

\[
P(T_i \in J_i, \ldots, T_n \in J_n | N(t) = n)
= P(T_i \in J_i, \ldots, T_n \in J_n, N(t) = n) / P(N(t) = n).
\]

Writing \( N(J_i) = N(t_i + h_i) - N(t_i), i = 1, \ldots, n \), and using the independence and stationarity of the increments of the Poisson process we obtain for the numerator that
Figure 4.36 Five realisations of the arrival times of a Poisson process $N$ with intensity 1, conditionally on $\{N(12) = 10\}$. They illustrate the order statistics property (Example 4.35).

\[
P(N(t_1) = 0, N(J_1) = 1, N(t_2) = N(t_1 + h_1) = 0, \ldots, N(t_n) - N(t_{n-1} + h_{n-1}) = 0, N(J_n) = 1, N(t) - N(t_n + h_n) = 0)
\]

\[
= P(N(t_1) = 0)P(N(J_1) = 1)P(N(t_2) - N(t_1 + h_1) = 0) \times \\
\cdots \times P(N(J_n) = 1)P(N(t) - N(t_n + h_n) = 0)
\]

\[
= e^{-\lambda t_1} e^{-\lambda_1 \lambda h_1} e^{-\lambda (t_2 - (t_1 + h_1))} \times \\
\cdots \times e^{-\lambda h_n - (t_n - h_{n-1})} e^{-\lambda h_n \lambda h_n} e^{-\lambda (t - (t_n + h_n))}
\]

\[
= \lambda^n e^{-\lambda t} \prod_{i=1}^{n} h_i.
\]

This implies

\[
P(T_1 \in J_1, \ldots, T_n \in J_n \mid N(t) = n) = \frac{n!}{t^n} \prod_{i=1}^{n} h_i.
\]

The conditional densities are obtained by dividing both sides by $\prod_{i=1}^{n} h_i$ and taking the limit for $\max_{1 \leq i \leq n} h_i \to 0$, yielding

\[
f_{T_1, \ldots, T_n \mid N(t)}(t_1, \ldots, t_n \mid n) = \frac{n!}{t^n}, \quad 0 < t_1 < \cdots < t_n < t.
\]

It follows from (4.11) that (4.12) is the density of the order statistics of $n$ iid uniform random variables on $(0, t)$. \qed
The following concept is called quantile transformation. It is extremely useful since it often reduces a problem concerning order statistics to one concerning the corresponding order statistics from a uniform sample. The proof follows immediately from the definition of the uniform distribution.

**Lemma 4.37 (Quantile transformation)**

Let \(X_1, \ldots, X_n\) be iid with distribution function \(F\). Furthermore, let \(U_1, \ldots, U_n\) be iid random variables uniformly distributed on \((0, 1)\) and denote by \(U_{(1)} < \cdots < U_{(n)}\) the corresponding order statistics. Then the following results hold:

1. \(F^{-1}(U_1) \overset{d}{=} X_1\).

2. For every \(n \in \mathbb{N}\),
   \[
   (X_1, \ldots, X_n) \overset{d}{=} (F^{-1}(U_{(1)}), \ldots, F^{-1}(U_{(n)})).
   \]

3. The random variable \(F(X_1)\) has a uniform distribution on \((0, 1)\) if and only if \(F\) is a continuous function. \(\square\)

**Example 4.38 (Simulation of upper order statistics)**

The quantile transformation above links the uniform distribution to some general distribution \(F\). An immediate application of this result is random number generation. For instance, exponential random numbers can be obtained from uniform random numbers by the transformation \(E_i = -\ln(1 - U_i)\). Simulation studies are widely used in an increasing number of applications. A simple algorithm for simulating order statistics of exponentials can be based on Example 4.34, which says that

\[
(E_{(n-i+1)} - E_{(n-i)})_{i=1,\ldots,n} \overset{d}{=} (\ln E_{(i-1)})_{i=1,\ldots,n},
\]

with \(E_{(0)} = 0\). This implies for the order statistics of an exponential sample that

\[
(E_{(n-i+1)})_{i=1,\ldots,n} \overset{d}{=} \left(\sum_{j=i}^{n} j^{-1} E_j \right)_{i=1,\ldots,n}.
\]

Order statistics and spacings of iid random variables \(U_i\) uniformly distributed on \((0, 1)\) and standard exponential random variables \(E_i\) are linked by the following representations; see e.g. Reiss [2], Theorem 1.6.7 and Corollary 1.6.9. We write \(\Gamma_n = E_1 + \cdots + E_n\), then

\[
(U_{(n)}, U_{(n-1)}, \ldots, U_{(1)}) \overset{d}{=} \left(\frac{\Gamma_n}{\Gamma_{n+1}}, \frac{\Gamma_{n-1}}{\Gamma_{n+1}}, \ldots, \frac{\Gamma_1}{\Gamma_{n+1}}\right),
\]

and

\[
(1 - U_{(n)}, U_{(n)} - U_{(n-1)}, \ldots, U_{(1)}) \overset{d}{=} \left(\frac{E_{n+1}}{\Gamma_{n+1}}, \ldots, \frac{E_1}{\Gamma_{n+1}}\right).
\]

The four distributional identities above provide simple methods for generating upper order statistics or spacings of the exponential or uniform distribution. \(\square\)

**References**


4.3.3 An extension to upper order statistics

So far, our data has consisted of $n$ iid observations of maxima which we have assumed to follow exactly a GEV $H_\theta$. By appropriately defining the underlying time periods, we design independence into the model; see (4.8). Suppose now that, rather than just having the largest observation available, we possess the $k$ largest of each period (year, say). In the notation of (4.8) this would amount to data

$$X^{(i)}_{(s-k+1)} \leq \cdots \leq X^{(i)}_{(s)} = X_i, \quad i = 1, \ldots, n.$$ 

Maximum likelihood theory based on these $k \times n$ observations would use the joint density of the independent vectors $(X^{(i)}_{(s-k+1)}, \ldots, X^{(i)}_{(s)})$, $i = 1, \ldots, n$. Only rarely in practical cases could we assume that for each $i$ the latter vectors are derived from iid data. If that were the case then maximum likelihood estimation should be based on the joint density of $k$ upper order statistics from a GEV as discussed in Theorem 4.32:

$$\frac{s!}{(s-k)!} H_\theta^{s-k} (x_1) \prod_{\ell=1}^{k} h_\theta (x_\ell), \quad x_1 < \cdots < x_k,$$

where, depending on $\theta$, the $x$-values satisfy the relevant domain restrictions. The standard error of the MLEs for $\mu$ and $\psi$ can already be reduced considerably if $k = 2$, i.e., we take the two largest observations into account. For a brief discussion on this method see Smith [2], Section 4.18, and Smith [1], where also further references and examples are to be found. The case $n = 1$, i.e., only one year of observations say, and $k > 1$ was first discussed in Weissman [3].

A final statement concerning maximum likelihood methodology, again taken from Smith [2], is worth stressing:

> The big advantage of maximum likelihood procedures is that they can be generalized, with very little change in the basic methodology, to much more complicated models in which trends or other effects may be present.

If the above quote has made you curious, do read Smith [2].

References


4.3.4 Method of probability-weighted moments

Among all the ad-hoc methods used in parameter estimation, the method of moments has attracted a lot of interest. In full generality it consists of equating model-moments based on $H_\theta$ to the corresponding empirical moments based on the data. Their general properties are notoriously unreliable on account of the poor sampling properties of second- and higher-order sample moments, a statement taken from Smith [2], p. 447. The class of probability-weighted moment estimators stands out as more promising. This method goes back to Hosking, Wallis and Wood [1]. Define

$$w_r (\theta) = E (X H_\theta (X)) , \quad r \in \mathbb{N}_0,$$

(4.13)
where $H_\theta$ is the GEV and $X$ has distribution function $H_\theta$ with parameter $\theta = (\xi, \mu, \psi)$. Recall that for $\xi \geq 1$, $H_\theta$ is regularly varying with index $1/\xi$. Hence $w_0$ is infinite. Therefore we restrict ourselves to the case $\xi < 1$. Define the empirical analogue to (4.13),

$$\hat{w}_r(\theta) = \int_{-\infty}^{+\infty} x H_\theta^r(x) dF_n(x), \quad r \in \mathbb{N}_0,$$

where $F_n$ is the empirical distribution function corresponding to the data $X_1, \ldots, X_n$. In order to estimate $\theta$ we solve the equations

$$w_r(\theta) = \hat{w}_r(\theta), \quad r = 0, 1, 2.$$

We immediately obtain

(4.14) $$\hat{w}_r(\theta) = \frac{1}{n} \sum_{j=1}^{n} X_{(j)} H_\theta^r(X_{(j)}), \quad r = 0, 1, 2.$$

Recall the quantile transformation from Lemma 4.37(b):

$$(H_\theta(X_{(1)}), \ldots, H_\theta(X_{(n)})) \overset{d}{=} (U_{(1)}, \ldots, U_{(n)}),$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ are the order statistics of an iid sequence $U_1, \ldots, U_n$ uniformly distributed on $(0, 1)$. With this interpretation, (4.14) can be written as

(4.15) $$\hat{w}_r(\theta) = \frac{1}{n} \sum_{j=1}^{n} X_{(j)} U_{(j)}^r, \quad r = 0, 1, 2.$$

Clearly, for $r = 0$, the right-hand side becomes $\overline{X}_n$, the sample mean. In order to calculate $w_r(\theta)$ for general $r$, observe that

$$w_r(\theta) = \int_{-\infty}^{+\infty} x H_\theta^r(x) dH_\theta(x) = \int_0^1 H_\theta^{-r}(y) y^r dy,$$

where for $0 < y < 1$,

$$H_\theta^{-r}(y) = \begin{cases} \mu - \frac{\psi}{\xi} (1 - (\ln y)^{-\xi}) & \text{if } \xi \neq 0, \\ \mu - \psi \ln(-\ln y) & \text{if } \xi = 0. \end{cases}$$

This yields for $\xi < 1$ and $\xi \neq 0$, after some calculation,

(4.16) $$w_r(\theta) = \frac{1}{r+1} \left\{ \mu - \frac{\psi}{\xi} \left(1 - \Gamma(1 - \xi)(1 + r\xi) \right) \right\},$$

where $\Gamma$ denotes the Gamma function $\Gamma(t) = \int_0^\infty e^{-u}u^{t-1} du$, $t > 0$. A combination of (4.15) and (4.16) gives us a probability-weighted moment estimator $\hat{\theta}_n^{(1)}$. Further estimators can be obtained by replacing $U_r^{(j)}$ in (4.15) by some statistic. Examples are:

- $\hat{\theta}_n^{(2)}$, where $U_{(j)}$ is replaced by any plotting position $p_{j,n}$, e.g. $p_{j,n} = j/n$.

- $\hat{\theta}_n^{(3)}$, where $U_r^{(j)}$ is replaced by $E U_r^{(j)}$.  

94
From (4.16), we immediately obtain

\[ w_0(\theta) = \mu - \frac{\psi}{\xi}(1 - \Gamma(1 - \xi)), \]

\[ 2w_1(\theta) - w_0(\theta) = \frac{\psi}{\xi} \Gamma(1 - \xi) \left( 2^\xi - 1 \right), \]

\[ 3w_2(\theta) - w_0(\theta) = \frac{\psi}{\xi} \Gamma(1 - \xi) \left( 3^\xi - 1 \right), \]

and hence

\[ \frac{3w_2(\theta) - w_0(\theta)}{2w_1(\theta) - w_0(\theta)} = \frac{3^\xi - 1}{2^\xi - 1}. \]

Applying any of the estimators above to the last equation yields an estimator \( \hat{\xi} \) of \( \xi \). Given \( \hat{\xi} \), the parameters \( \mu \) and \( \psi \) are then estimated by

\[ \hat{\psi} = \frac{(2\hat{w}_1 - \hat{w}_0) \hat{\xi}}{\Gamma(1 - \hat{\xi})(2^\hat{\xi} - 1)}, \]

\[ \hat{\mu} = \hat{w}_0 + \frac{\hat{\psi}}{\xi}(1 - \Gamma(1 - \hat{\xi})), \]

where \( \hat{w}_0, \hat{w}_1, \hat{w}_2 \) are any of the empirical probability-weighted moments discussed above. The case \( \xi = 0 \) can of course also be covered by this method.

For a discussion on the behaviour of these estimators see Hosking et al. [1]. Smith [2] summarises as follows.

The method is simple to apply and performs well in simulation studies. However, until there is some convincing theoretical explanation of its properties, it is unlikely to be universally accepted. There is also the disadvantage that, at present at least, it does not extend to more complicated situations such as regression models based on extreme value distributions.

References


4.3.5 Tail and quantile estimation, a first go

Let us return to the basic set-up of (4.7) and (4.8), i.e., we have an iid sample \( X_1, \ldots, X_n \) from \( H_\theta \). In this situation, a quantile estimator can be readily obtained. Indeed, by the methods discussed in the previous sections, we obtain an estimate \( \hat{\theta} \) of \( \theta \). Given any \( p \in (0, 1) \), the \( p \)-quantile \( x_p \) is defined via \( x_p = H_{\hat{\theta}}^\leq(p) \); see Definition 2.5. A natural estimator for \( x_p \), based on \( X_1, \ldots, X_n \), then becomes

\[ \hat{x}_p = H_{\hat{\theta}}^\leq(p). \]
By the definition of $H_\theta$ this leads to
\[
\hat{x}_p = \hat{\mu} - \frac{\hat{\psi}}{\xi} \left( 1 - \left( -\ln p \right)^{-1/\xi} \right).
\]
The corresponding tail estimate for $\overline{\mathbb{P}}_\theta(x)$, for $x$ in the appropriate domain, corresponds to
\[
\overline{\mathbb{P}}_{\hat{\theta}}(x) = 1 - \exp \left\{ - \left( 1 + \xi \frac{x - \hat{\mu}}{\hat{\psi}} \right)^{-1/\xi} \right\},
\]
where $\hat{\theta} = (\hat{\xi}, \hat{\mu}, \hat{\psi})$ is either estimated by the MLE or by a probability-weighted moment estimator.

Comments
A recommendable account of estimation methods for the GEV, including a detailed discussion of the pros and cons of the different methods, is Buishand [1]. Hosking [3] discusses the problem of hypothesis testing within GEV.

If the extreme value distribution is known to be Fréchet, Gumbel or Weibull, the above methods can be adapted to the specific distribution function under consideration. This may simplify the estimation problem in the case of $\xi \geq 0$ (Fréchet, Gumbel), but not for the Weibull distribution. The latter is due to non-regularity problems of the MLE as explained in Section 4.3.1. The vast amount of papers written on estimation for the three-parameter Weibull reflects this situation; see Embrechts et al. [2] for references. To indicate the sort of problems that may occur, we refer to Smith [4] who studies the Pareto-like probability densities
\[
f(x; K, \alpha) \sim c\alpha(K - x)^{\alpha - 1}, \quad x \uparrow K,
\]
where $K$ and $\alpha$ are unknown parameters.

References

4.4 Estimating under maximum domain of attraction conditions

4.4.1 Introduction
Relaxing condition (4.7), we assume in this section that for some $\xi \in \mathbb{R}$,
\[
X_1, \ldots, X_n \text{ are iid from } F \in MDA (H_\xi).
\]
By Proposition 3.19, $F \in MDA (H_\xi)$ is equivalent to
\[
\lim_{n \to \infty} n F(c_n x + d_n) = -\ln H_\xi (x)
\]
for appropriate norming sequences \((c_n)\) and \((d_n)\), and \(x\) belongs to a suitable domain depending on the sign of \(\xi\). Let us from the start be very clear about the fundamental difference between (4.7) and (4.17). Consider for illustrative purposes only the standard Fréchet case \(\xi = 1/\alpha > 0\). Now (4.7) means that our sample \(X_1, \ldots, X_n\) exactly follows a Fréchet distribution, i.e.,

\[
\overline{F}(x) = 1 - \exp\left\{-x^{-\alpha}\right\}, \quad x > 0.
\]

On the other hand, by virtue of Theorem 3.21 assumption (4.17) reduces in the Fréchet case to

\[
\overline{F}(x) = x^{-\alpha} L(x), \quad x > 0,
\]

for some slowly varying function \(L\). Clearly, in this case the estimation of the tail \(\overline{F}(x)\) is much more involved due to the non-parametric character of \(L\). In various applications, one would mainly (in some cases, solely) be interested in \(\alpha\). So (4.7) amounts to full parametric assumptions, whereas (4.17) is essentially semi-parametric in nature: there is a parametric part \(\alpha\) and a non-parametric part \(L\). Because of this difference, (4.17) is much more generally considered as inference for heavy-tailed distributions as opposed to inference for the GEV in (4.7).

A handwaving consequence of (4.18) is that for large \(u = c_n x + d_n\),

\[
n\overline{F}(u) \approx \left(1 + \xi \frac{u - d_n}{c_n}\right)^{-1/\xi},
\]

so that a tail-estimator could take on the form

\[
\left(\overline{F}(u)\right)^\hat{} = \frac{1}{n} \left(\frac{1}{1 + \hat{\xi} \frac{u - \hat{d}_n}{\hat{c}_n}}\right)^{-1/\hat{\xi}},
\]

for appropriate estimators \(\hat{\xi}\), \(\hat{c}_n\) and \(\hat{d}_n\). As (4.17) is essentially a tail-property, estimation of \(\xi\) may be based on \(k\) upper order statistics \(X_{(n-k+1)} \leq \cdots \leq X_{(n)}\). A whole battery of classical approaches has exploited this natural idea; see Section 4.4.2. The following mathematical conditions are usually imposed:

\[
\begin{align*}
(\text{a}) \quad & k(n) \to \infty \quad \text{use a sufficiently large number of order statistics,} \\
(\text{b}) \quad & \frac{n}{k(n)} \to \infty \quad \text{as we are interested in a tail property, we should also make sure to concentrate only on the upper order statistics. Let the tails speak for themselves.}
\end{align*}
\]

When working out the details later, we will be able to see where exactly the properties on \((k(n))\) enter. Indeed it is precisely this (for the moment perhaps redundant) degree of freedom \(k\) which will allow us to obtain the necessary statistical properties like consistency and asymptotic normality for our estimators.

From (4.19) we would in principle be in the position to estimate the quantile \(x_p = F^{\leftarrow}(p)\), for fixed \(p \in (0, 1)\), as follows

\[
\hat{x}_p = \hat{d}_n + \frac{\hat{c}_n}{\hat{\xi}} \left(n (1 - p) - \hat{\xi} - 1\right).
\]

97
Typically, we will be interested in estimating high $p$-quantiles outside the sample $X_1, \ldots, X_n$. This means that $p = p_n$ is chosen in such a way that $p > 1 - 1/n$, hence the empirical distribution function satisfies $F_n(p) = 0$ and does not yield any information about such quantiles. In order to get good estimators for $\xi$, $c_n$ and $d_n$ in (4.21) a subsequent trick is needed. Assume for notational convenience that $n/k \in \mathbb{N}$. A standard approach now consists of passing to a subsequence $(n/k)$ with $k = k(n)$ satisfying (4.20). The quantile $x_p$ is then estimated by

\begin{equation}
\hat{x}_p = \hat{a}_{n/k} + \hat{c}_{n/k} \left( \frac{n}{k} \left( 1 - p_n \right) \right)^{-\hat{\xi}} - 1 \right) .
\end{equation}

Why does this work? One reason behind this construction is that we need to estimate at two levels. First, we have to find a reliable estimate for $\xi$: this task will be worked out in Section 4.4.2. Condition (4.20) will appear very naturally. Second, we need to estimate the norming constants $c_n$ and $d_n$ which themselves are defined via quantiles of $F$. For instance, in the Fréchet case we know that $c_n = F^{-1}(1 - n^{-1})$; see Theorem 3.21. Hence estimating $c_n$ is equivalent to the problem of estimating $x_p$ at the boundary of our data range. By going to the subsequence $(n/k)$, we move away from the critical boundary value $1 - n^{-1}$ to the safer $1 - (n/k)^{-1}$. Estimating $c_{n/k}$ is so reduced to estimating quantiles within the range of our data. Similar arguments hold for $d_{n/k}$, and indeed for the Gumbel and Weibull case. We may therefore hope that the construction in (4.22) leads to a good estimator for $x_p$. The above discussion is only heuristic, a detailed statistical analysis shows that this approach can be made to work.

In the context of statistics of extremal events it may also be of interest to estimate the following quantity which is closely related to the quantiles $x_p$:

$$x_{p,r} = F^{-1}(p^{1/r}) , \quad r \in \mathbb{N} .$$

Notice that $x_p = x_{p,1}$. The interpretation of $x_{p,r}$ is obvious from

$$p = F^r(x_{p,r}) = P(\max \{ X_{n+1}, \ldots, X_{n+r} \} \leq x_{p,r}) ,$$

so $x_{p,r}$ is that level which, with a given probability $p$, will not be exceeded by the next $r$ observations $X_{n+1}, \ldots, X_{n+r}$. As an estimate we then obtain from (4.22)

$$\hat{x}_{p,r} = \hat{a}_{n/k} + \hat{c}_{n/k} \left( \frac{n}{k} \left( 1 - p^{1/r} \right) \right)^{-\hat{\xi}} - 1 \right) .$$

In what follows we will concentrate only on estimation of $x_p$; from the definition of $x_{p,r}$ it is clear how one has to proceed for general $r$. From the above heuristics we obtain a programme for the remainder of this section:

1. Find appropriate estimators for the shape parameter $\xi$ of the GEV.
2. Find appropriate estimators for the norming constants $c_n$ and $d_n$.
3. Show that the estimators proposed above yield reasonable approximations to the distribution tail in its far end and to high quantiles.
4. Determine the statistical properties of these estimators.

### 4.4.2 Estimating the shape parameter $\xi$

In this section we study different estimators of the shape parameter $\xi$ for $F \in \text{MDA}(H_\xi)$. We also give some of their statistical properties.
Method 1: Pickands’s estimator for $\xi \in \mathbb{R}$

The basic idea behind this estimator consists of finding a condition equivalent to $F \in \text{MDA}(H_{\xi})$ which involves the parameter $\xi$ in an easy way. The key to Pickands’s estimator and its various generalizations is Theorem 3.50, where it was shown that for $F \in \text{MDA}(H_{\xi})$, $U(t) = F^{\ast}((1 - t^{-1})$ satisfies

$$\lim_{t \to \infty} \frac{U(2t) - U(t)}{U(t) - U(t/2)} = 2^\xi.$$

Furthermore, this limit relation holds locally uniformly. Hence whenever the limit $\lim_{t \to \infty} c(t) = 2$ exists for a positive function $c$,

$$(4.23) \quad \lim_{t \to \infty} \frac{U(c(t)t) - U(t)}{U(t) - U(t/c(t))} = 2^\xi.$$

The basic idea now consists of constructing an empirical estimator using (4.23). To that effect, let

$$V(1) \leq \cdots \leq V(n)$$

be the order statistics from an iid sample $V_1, \ldots, V_n$ with common Pareto distribution function $F_Y(x) = 1 - x^{-1}$, $x \geq 1$. It follows in the same way as for the quantile transformation, see Lemma 4.37(b), that

$$(X(k))_{k=1,\ldots,n} \overset{d}{=} (U(V(k)))_{k=1,\ldots,n},$$

where $X_1, \ldots, X_n$ are iid with distribution function $F$. Notice that $V(n_{1-k+1})$ is the empirical $(1 - k/n)$-quantile of $F_Y$. Making use of the quantile transformation, it is possible to show that

$$\frac{k}{n} V(n_{k+1}) \overset{P}{\to} 1, \quad n \to \infty,$$

whenever $k = k(n) \to \infty$ and $k/n \to 0$. In particular,

$$V(n_{k+1}) \overset{P}{\to} \infty \quad \text{and} \quad \frac{V(n_{k+1})}{V(n_{k+1})} \overset{P}{\to} \frac{1}{2}, \quad n \to \infty.$$

Combining this with (4.23) and using a subsequence argument yields

$$\frac{U\left(V(n_{k+1})\right) - U\left(V(n_{k+1})\right)}{U\left(V(n_{k+1})\right) - U\left(V(n_{k+1})\right)} \overset{P}{\to} 2^\xi, \quad n \to \infty.$$

Motivated by the discussion above and by (4.23), we now define the **Pickands estimator**

$$\hat{\xi}^{(P)}_{n-k+1} = \frac{1}{\ln 2} \frac{X(n_{k+1}) - X(n_{k+1})}{X(n_{k+1}) - X(n_{k+1})}. \tag{4.24}$$

This estimator turns out to be weakly consistent provided $k \to \infty$, $k/n \to 0$:

$$\hat{\xi}^{(P)}_{n-k+1} \overset{P}{\to} \xi, \quad n \to \infty.$$

This was already observed by Pickands [3]. A full analysis on $\hat{\xi}^{(P)}_{n-k+1}$ is to be found in Dekkers and de Haan [1] from which the following result is taken.
Figure 4.40 Pickands–plot for the water–damage claim data; see Figure 4.29. The estimate of $\xi$ appears to be close to 0. The upper and lower lines constitute asymptotic 95% confidence bands.

**Theorem 4.39** (Properties of the Pickands estimator)
Suppose $(X_n)$ is an iid sequence with distribution function $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$. Let $\hat{\xi}^{(P)} = \hat{\xi}^{(P)}_{(n-k+1)}$ be the Pickands estimator (4.24).

1. (Weak consistency) If $k \to \infty$, $k/n \to 0$ for $n \to \infty$, then
   \[ \hat{\xi}^{(P)} \xrightarrow{P} \xi, \quad n \to \infty. \]

2. (Strong consistency) If $k/n \to 0$, $k/\ln \ln n \to \infty$ for $n \to \infty$, then
   \[ \hat{\xi}^{(P)} \xrightarrow{a.s.} \xi, \quad n \to \infty. \]

3. (Asymptotic normality) Under further conditions on $k$ and $F$ (see Dekkers and de Haan [1], p. 1799),
   \[ \sqrt{k} (\hat{\xi} - \xi) \xrightarrow{d} N(0, v(\xi)), \quad n \to \infty, \]
   where
   \[ v(\xi) = \frac{\xi^2 (2\xi+1)}{(2(2\xi-1)\ln 2)^2}. \]

**Remarks.** 1) This theorem forms the core of a whole series of results obtained in Dekkers and de Haan [1]; on it one can base quantile and tail estimators and (asymptotic) confidence interval constructions. The quoted paper [1] also contains various simulated and real life examples in order to see the theory in action. We strongly advice the reader to go through it, perhaps avoiding upon first reading the (rather extensive) technical details. The main idea behind the above construction goes back to Pickands [3]. A nice summary is to be found in de Haan [2] from which the derivation above is taken.

2) In the spirit of Section 4.4.1 notice that the calculation of Pickands’s estimator (4.24) involves
a sequence of upper order statistics increasing with \( n \). Consequently, one mostly includes a so-called Pickands–plot in the analysis, i.e.,

\[
\left\{ (k, \hat{\xi}_{n-k+1}^{(P)}): k = 1, \ldots, n \right\},
\]

in order to allow for a choice depending on \( k \). The interpretation of such plots, i.e., the optimal choice of \( k \), is a delicate point for which no uniformly best solution exists. It is intuitively clear that one should choose \( \hat{\xi}_{k,n}^{(P)} \) from such a \( k \)-region where the plot is roughly horizontal. We shall come back to this point later; see the Summary at the end of this section.

References


Method 2: Hill’s estimator for \( \xi = \alpha^{-1} > 0 \)

Suppose \( X_1, \ldots, X_n \) are iid with distribution function \( F \in \text{MDA}(\Phi_\alpha) \), \( \alpha > 0 \), thus \( F(x) = x^{-\alpha} L(x) \), \( x > 0 \), for a slowly varying function \( L \); see Theorem 3.21. Distributions with such tails form the prime examples for modelling heavy-tailed phenomena. For many applications the knowledge of the index \( \alpha \) of regular variation is of major importance. If for instance \( \alpha < 2 \) then \( E X_1^2 = \infty \). This case is often observed in the modelling of insurance data.

Empirical studies on the tails of daily log-returns in finance have indicated that one frequently encounters values \( \alpha \) between 3 and 4; see for instance Guillaume et al. [1], Longin [2] and Loretan and Phillips [3]. Information of the latter type implies that, whereas covariances of such data would be well defined, the construction of confidence intervals for the sample autocovariances and autocorrelations on the basis of asymptotic (central limit) theory may be questionable as typically a finite fourth moment condition is asked for.

The Hill estimator of \( \alpha \) essentially takes on the following form:

\[
(4.25) \quad \hat{\alpha}^{(H)} = \hat{\alpha}^{(H)}(k) = \left( \frac{1}{k} \sum_{j=1}^{k} \ln X_{(n-j+1)} - \ln X_{(n-k+1)} \right)^{-1},
\]

where \( k = k(n) \to \infty \) in an appropriate way, so that as in the case of Pickands’s estimator, an increasing sequence of upper order statistics is used. One of the interesting facts concerning (4.25) is that various asymptotically equivalent versions of \( \hat{\alpha}^{(H)} \) can be derived through essentially different methods, showing that the Hill estimator is very natural. Below we discuss some derivations.

References


The MLE approach

(Hill [1]). Assume for the moment that \( X \) is a random variable with distribution function \( F \) so that for \( \alpha > 0 \)
\[
P(X > x) = F(x) = x^{-\alpha}, \quad x \geq 1.
\]
Then it immediately follows that \( Y = \ln X \) has distribution function
\[
P(Y > y) = e^{-\alpha y}, \quad y \geq 0,
\]
i.e., \( Y \) is \( \text{Exp}(\alpha) \) and hence the MLE of \( \alpha \) is given by
\[
\hat{\alpha}_n = Y_n^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_j \right)^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_{(n-j+1)} \right)^{-1}.
\]

A trivial generalisation concerns
\[
(4.26) \quad F(x) = Cx^{-\alpha}, \quad x \geq u > 0,
\]
with \( u \) known. If we interpret (4.26) as fully specified, i.e., \( C = u^\alpha \), then we immediately obtain as MLE of \( \alpha \)
\[
(4.27) \quad \hat{\alpha}_n = \left( \frac{1}{n} \sum_{j=1}^{n} \ln \left( \frac{X_{(n-j+1)}}{u} \right) \right)^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_{(n-j+1)} - \ln u \right)^{-1}.
\]

Now we often do not have the precise parametric information of these examples, but in the spirit of MDA (\( \Phi_n \)) we assume that \( F \) behaves like a Pareto distribution function above a certain known threshold \( u \) say. Let

\[
(4.28) \quad K = \text{card } \{ i : X_{(i)} > u, \ i = 1, \ldots, n \}.
\]

Conditionally on the event \( \{ K = k \} \), maximum likelihood estimation of \( \alpha \) and \( C \) in (4.26) reduces to maximizing the joint density of \( (X_{(n-k+1)}, \ldots, X_{(n)}) \). From Theorem 4.32 we deduce
\[
\begin{align*}
\int_{X_{(n-k+1)}, \ldots, X_{(n)}} (x_1, \ldots, x_k) \\
= \frac{n!}{(n-k)!} \left( 1 - Cx_1^{-\alpha} \right)^{n-k} \left( Cx_1^{-\alpha} \right)^k \prod_{i=1}^{k} x_i^{-(\alpha+1)}, \quad u < x_1 < \cdots < x_k.
\end{align*}
\]

A straightforward calculation yields the conditional MLEs
\[
\begin{align*}
\hat{\alpha}_{(k)} &= \left( \frac{1}{k} \sum_{j=1}^{k} \ln \left( \frac{X_{(n-j+1)}}{X_{(n-k+1)}} \right) \right)^{-1} \\
&= \left( \frac{1}{k} \sum_{j=1}^{k} \ln X_{(n-j+1)} - \ln X_{(n-k+1)} \right)^{-1} \\
\hat{C}_{(k)} &= \frac{k}{n} X_{(n-k+1)}^{\alpha_{(k)}}.
\end{align*}
\]

102
Figure 4.41 Tail and quantile estimation based on a Hill–fit; see (4.29) and (4.30). The data are the Danish insurance claims from Example 4.16. Top, left: the Hill–plot for $\xi = 1/\alpha$ as a function of $k$ upper order statistics (lower horizontal axis) and of the threshold $u$ (upper horizontal axis), i.e., there are $k$ exceedances of the threshold $u$. Top, right: the fit of the shifted excess distribution function $F_u(x - u), x \geq u$, on log-scale. Middle: tail–fit of $F(x + u), x \geq 0$. Bottom: estimation of the 0.99–quantile as a function of the $k$ upper order statistics and of the corresponding threshold value $u$. The estimation of the tail is based on $k = 109 \ |u = 10|$ and $\alpha = \xi^{-1} = 1.618$. Compare also with the GPD–fit in Figure 4.60.
Figure 4.42 Tail and quantile estimation based on a Hill-fit; see \((4.29)\) and \((4.30)\). The data are the industrial fire claims from Example 4.16. Top, left: the Hill-plot for \(\xi = 1/\alpha\) as a function of \(k\) upper order statistics (lower horizontal axis) and of the threshold \(u\) (upper horizontal axis), i.e., there are \(k\) exceedances of the threshold \(u\). Top, right: the fit of the shifted excess distribution function \(F_u(x-u), x \geq u\), on log-scale. Middle: tail-fit of \(\overline{F}(x+u), x \geq 0\). Bottom: estimation of the 0.99-quantile as a function of the \(k\) upper order statistics and of the corresponding threshold value \(u\). The estimation of the tail is based on \(k = 149\) \((u = 100)\) and \(\alpha = \xi^{-1} = 1.058\). Compare also with the GPD-fit in Figure 4.60.
So Hill’s estimator has the same form as the MLE in the exact model underlying (4.27) but now having the deterministic \( u \) replaced by the random threshold \( X_{(n-k+1)} \), where \( k \) is defined through (4.28). We also immediately obtain an estimate for the tail \( \bar{F}(x) \)
\[
(4.29) \quad \hat{F}(x) = \frac{k}{n} \left( \frac{x}{\bar{X}_{(n-k+1)}} \right)^{-\hat{\alpha}(k)}
\]
and for the \( p \)-quantile
\[
(4.30) \quad \hat{x}_p = \left( \frac{n}{k} (1-p) \right)^{-1/\hat{\alpha}(k)} \bar{X}_{(n-k+1)}.
\]
From (4.29) we obtain an estimator of the excess distribution function \( F_u(x-u), x \geq u \), by using \( F_u(x-u) = 1 - \bar{F}(x)/\bar{F}(u) \). Examples, based on these estimators, are to be found in Figures 4.41 and 4.42 for the Danish, respectively industrial, fire insurance data. We will come back to these data more in detail in Section 4.5.2.

**Example 4.43** (The Hill estimator at work)
In Figures 4.41 and 4.42 we have applied the above methods to the Danish fire insurance data (Figure 4.18) and the industrial fire insurance data (Figure 4.19); for a preliminary analysis see Example 4.16. For the Danish data we have chosen as an initial threshold \( u = 10 \) (\( k = 109 \)). The corresponding Hill estimate has a value \( \hat{\xi} = 0.618 \). When changed to \( u = 18 \) (\( k = 47 \)), we obtain \( \hat{\xi} = 0.497 \). The Hill--plot shows a fairly stable behaviour in the range \((0.5, 0.7)\). As in most applications, the quantities of main interest are the high quantiles. We therefore turn immediately to Figure 4.41 (bottom), where \( \hat{x}_{0.99} \) is plotted across all relevant \( u-- \) (equivalently, \( k-- \)) values. For \( k \) in the region \((45, 110)\) the quantile--Hill--plot shows a remarkably stable behaviour around the value 24.7. This agrees perfectly with the empirical estimate of 24.6 for \( x_{0.99} \); see Table 4.20. This should be contrasted with the situation in Figure 4.42 for the industrial fire data. For the latter data, the estimate for \( \xi \) ranges from 0.945 for \( u = 100 \) (\( k = 149 \)) over 0.745 for \( u = 300 \) (\( k = 49 \)) to 0.799 for \( u = 500 \) (\( k = 29 \)). All estimates clearly correspond to infinite variance models! An estimate for \( x_{0.99} \) in the range \((180, 200)\) emerges, again in agreement with the empirical value of 184. We would like to stress at this point that the above discussion represents only the beginning of a detailed analysis. The further discussions have to be conducted together with the actuary responsible for the underlying data. \( \square \)

**References**

**The regular variation approach**
(de Haan [1]). This approach is in the same spirit as the construction of Pickands’s estimator, i.e., base the inference on a suitable reformulation of \( F \in \text{MDA}(\Phi_\alpha) \). Indeed \( F \in \text{MDA}(\Phi_\alpha) \) if and only if
\[
\lim_{t \to \infty} \frac{F(tx)}{F(t)} = x^{-\alpha}, \quad x > 0.
\]
Using partial integration, we obtain
\[
\int_t^\infty (\ln x - \ln t) dF(x) = \int_t^\infty \frac{F(x)}{x} \, dx,
\]
105
so that by Karamata’s theorem (see (2.14) and (2.15))

\[(4.31)\]
\[
\frac{1}{F(t)} \int_t^\infty \left(\ln x - \ln t\right) dF(x) \to \frac{1}{\alpha}, \quad t \to \infty.
\]

How do we find an estimator from this result? Two choices have to be made:

1. replace \( F \) by an estimator, the obvious candidate here is the empirical distribution function

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[X_i \leq x]} = \frac{1}{n} \sum_{i=1}^n I_{[X_{(i-n+1)} \leq x]},
\]

2. replace \( t \) by an appropriate high, data dependent level (recall \( t \to \infty \)); we take \( t = X_{(n-k+1)} \) for some \( k = k(n) \).

The choice of \( t \) is motivated by the fact that \( X_{(n-k+1)} \xrightarrow{a.s.} \infty \) provided \( k = k(n) \to \infty \) and \( k/n \to 0 \). From (4.31) the following estimator results

\[
\frac{1}{F_n(X_{(n-k+1)})} \int_{X_{(n-k+1)}}^\infty \left(\ln x - \ln X_{(n-k+1)}\right) dF_n(x)
\]

\[
\quad = \frac{1}{k-1} \sum_{j=1}^{k-1} \ln X_{(n-j+1)} - \ln X_{(n-k+1)}
\]

which, modulo the factor \( k-1 \), is again of the form \( (\hat{\alpha}_{(H)}^{(H)})^{-1} \) in (4.25). Notice that the change from \( k \) to \( k-1 \) is asymptotically negligible.

References


The mean excess function approach

This is essentially a reformulation of the approach above; we prefer to list it separately because of its methodological merit. Suppose \( X \) is a random variable with distribution function \( F \in \text{MDA}(\Phi_{\alpha}) \), \( \alpha > 0 \), and for notational convenience assume that \( X > 1 \) a.s. One can now rewrite (4.31) as follows (see also Example 4.12)

\[
E(\ln X - \ln t \mid \ln X > \ln t) \to \frac{1}{\alpha}, \quad t \to \infty.
\]

So denoting \( u = \ln t \) and \( e^*(u) \) the mean excess function of \( \ln X \) (see Definition 4.10) we obtain

\[
e^*(u) \to \frac{1}{\alpha}, \quad u \to \infty.
\]

Hill’s estimator can then be interpreted as the empirical mean excess function of \( \ln X \) calculated at the threshold \( u = \ln X_{(n-k+1)} \), i.e., \( e^*_n(\ln X_{(n-k+1)}) \).

We summarise as follows.
Suppose $X_1,\ldots,X_n$ are iid with distribution function $F \in \text{MDA} (\Phi_\alpha)$, $\alpha > 0$, then a natural estimator for $\alpha$ is provided by Hill’s estimator

$$
\hat{\alpha}^{(H)} = \left( \frac{1}{k} \sum_{j=1}^{k} \ln X_{(n-j+1)} - \ln X_{(n-k+1)} \right)^{-1},
$$

where $k = k(n)$ satisfies (4.20).

Below we summarize the main properties of the Hill estimator. Before looking at the theorem, you may want to refresh your memory on the meaning of linear processes and weakly dependent strictly stationary processes.

**Theorem 4.44 (Properties of the Hill estimator)**

Suppose $(X_n)$ is strictly stationary with marginal distribution $F$ satisfying for some $\alpha > 0$ and $L$ slowly varying,

$$
\overline{F}(x) = P(X > x) = x^{-\alpha} L(x), \quad x > 0.
$$

Let $\hat{\alpha}^{(H)} = \hat{\alpha}^{(H)}(k)$ be the Hill estimator (4.32).

1. (Weak consistency) Assume that one of the following conditions is satisfied:
   - $(X_n)$ is iid (Mason [3]),
   - $(X_n)$ is weakly dependent (Rootzén, Leadbetter and de Haan [5], Hsing [2]),
   - $(X_n)$ is a linear process (Resnick and Stāricā [4]).

   If $k \to \infty$, $k/n \to 0$ for $n \to \infty$, then
   $$
   \hat{\alpha}^{(H)} \overset{P}{\to} \alpha.
   $$

2. (Strong consistency) (Deheuvels, Häusler and Mason [1]) If $k/n \to 0$, $k/\ln \ln n \to \infty$ for $n \to \infty$ and $(X_n)$ is an iid sequence, then
   $$
   \hat{\alpha}^{(H)} \overset{\text{a.s.}}{\to} \alpha.
   $$

3. (Asymptotic normality) If further conditions on $k$ and $F$ are satisfied (see for instance the Comments below) and $(X_n)$ is an iid sequence, then
   $$
   \sqrt{k} \left( \hat{\alpha}^{(H)} - \alpha \right) \overset{d}{\to} N(0, \alpha^2).
   $$

**Remarks.** 3) Theorem 4.44 should be viewed as a counterpart to Theorem 4.39 on the Pickands estimator. Because of the importance of $\hat{\alpha}^{(H)}$, we prefer to formulate Theorem 4.44 in its present form for sequences $(X_n)$ more general than iid.

4) Do not interpret this theorem as saying that the Hill estimator is always fine. The theorem only says that rather generally the standard statistical properties hold. One still needs a crucial set of conditions on $\overline{F}$ and $k(n)$. In particular, second-order regular variation assumptions on $\overline{F}$ have to be imposed to derive the asymptotic normality of $\hat{\alpha}^{(H)}$. The same applies to the case of the Pickands estimator. Notice that these conditions are not verifiable in practice.
5) Depending on the slowly varying function $L$ in the tail of $F$, it can be shown that the rate of convergence of Hill’s estimator can be arbitrarily slow.

6) As in the case of the Pickands estimator, an analysis based on the Hill estimator is usually summarised graphically. The Hill-plot

$$\{(k, \hat{\alpha}^{(H)}(k)) : k = 2, \ldots, n\},$$

is instrumental in finding the optimal $k$.

7) The asymptotic variance of $\hat{\alpha}^{(H)}$ depends on the unknown parameter $\alpha$ so that in order to calculate the asymptotic confidence intervals an appropriate estimator of $\alpha$, typically $\hat{\alpha}^{(H)}$, has to be inserted.

8) The Hill estimator is very sensitive with respect to dependence in the data; see for instance Figure 4.48 in the case of an autoregressive process. For ARMA and weakly dependent processes special techniques have been developed, for instance by first fitting an ARMA model to the data and then applying the Hill estimator to the residuals. See for instance the references mentioned under part (a) of Theorem 4.44.

□

References


Figure 4.46 Pickands-, Hill- and DEdH-plots with asymptotic 95% confidence bands for 2000 absolute values of iid standard Cauchy random variables. The tail of the latter is Pareto-like with index $\xi = 1$. Recall that, for given $k$, the DEdH and the Hill estimator use the $k$ upper order statistics of the sample, whereas the Pickands estimator uses $4k$ of them. In the case of the Pickands estimator one clearly sees the trade-off between variance and bias; see also the discussion in the Comments.


Method 3: The Dekkers–Einmahl–de Haan Estimator for $\xi \in \mathbb{R}$

A disadvantage of Hill’s estimator is that it is essentially designed for $F \in \text{MDA}(H_\xi)$, $\xi > 0$. We have already stressed before that this class of models suffices for many applications in the realm of finance and insurance. In Dekkers, Einmahl und de Haan [1], Hill’s estimator is extended to cover the whole class $H_\xi$, $\xi \in \mathbb{R}$. In Theorem 3.26 we saw that for $F \in \text{MDA}(H_\xi)$, $\xi < 0$, the right endpoint $x_F$ of $F$ is finite. We can and do assume that $x_F > 0$, if necessary we shift the domain of the distribution function. In Section 3.3 we found that the maximum domain of attraction conditions for $H_\xi$ all involve some kind of regular variation. As for deriving the Pickands and Hill estimator, one can reformulate regular variation conditions to find estimators for any $\xi \in \mathbb{R}$. Dekkers et al. [1] come up with the following proposal:

\begin{equation}
\hat{\xi} = 1 + H_n^{(1)} + \frac{1}{2} \left( \frac{(H_n^{(1)})^2}{H_n^{(2)}} - 1 \right)^{-1},
\end{equation}
Figure 4.47 A “Hill horror plot”: the Hill estimator $\hat{\alpha}_{(k)}^{-1}$ from n iid realizations with distribution tail $\mathcal{F}_1(x) = 1/x$ (top line) and $\mathcal{F}_2(x) = 1/(x\ln x)$ (bottom line). The solid line corresponds to $\alpha = 1$. The performance of the Hill estimate for $\mathcal{F}_2$ is very poor. The value $k$ is the number of upper order statistics used for the construction of the Hill estimator.

where

$$H_n^{(1)} = \frac{1}{k} \sum_{j=1}^{k} (\ln X_{(n-j+1)} - \ln X_{(n-k)})$$

is the reciprocal of Hill’s estimator (modulo an unimportant change from $k$ to $k+1$) and

$$H_n^{(2)} = \frac{1}{k} \sum_{j=1}^{k} (\ln X_{(n-j+1)} - \ln X_{(n-k)})^2.$$  

Because $H_n^{(1)}$ and $H_n^{(2)}$ can be interpreted as empirical moments, $\hat{\xi}$ is also referred to as a moment estimator of $\xi$. To make sense, in all the estimators discussed so far we could (and actually should) replace $\ln x$ by $\ln(1 \vee x)$. In practice, this should not pose problems because we assume $k/n \to 0$. Hence the relation $X_{(n-k+1)} \overset{a.s.}{\to} x_F > 0$ holds.

At this point we *pause for a moment* and see where we are. First of all

*Do we have all relevant approaches for estimating the shape parameter $\xi$?*

Although various estimators have been presented, we have to answer this question by *no!* The above derivations were all motivated by analytical results on regular variation. An alternative approach is via point processes; one programme runs under the heading
Figure 4.48 A comparative study of the Hill–plots for 1000 iid simulated data from an AR(1) process \( X_t = \phi X_{t-1} + Z_t, \phi \in \{0.9, 0.5, 0.2\} \). The noise sequence \( (Z_t) \) comes from a symmetric distribution with exact Pareto tail \( P(Z > x) = 0.5x^{-10}, x \geq 1 \). One can show that, \( P(X > x) \sim cx^{-10} \). The solid line corresponds to the Hill estimator of the \( X_t \) as a function of the \( k \) upper order statistics. The dotted line corresponds to the Hill estimator of the residuals \( \hat{Z}_t = X_t - \hat{\phi}X_{t-1} \), where \( \hat{\phi} \) is the Yule-Walker estimator of \( \phi \). Obviously, the Hill estimator of the residuals yields much more accurate values. These figures indicate that the Hill estimator for correlated data has to be used with extreme care. Even for \( \phi = 0.2 \) the Hill estimator of the \( X_t \) cannot be considered as a satisfactory tool for estimating the index of regular variation. The corresponding theory for the Hill estimator of linear processes can be found in Resnick and Stărică [4].

Point process of exceedances,

or, as the hydrologists call it,

POT: the Peaks Over Threshold method.

Because of its fundamental importance we decided to spend a whole section on this method; see Section 4.5.

References

Comments

In the previous sections we discussed some of the main issues underlying the statistical estimation of
the shape parameter $\xi$. This general area is rapidly expanding so that at any moment of time is immediately outdated. The references cited are therefore not exhaustive and reflect our personal interest. The fact that a particular paper does not appear in the list of references does not mean that it is considered less important.

The Hill estimator: the bias-variance trade–off

Theorem 4.44 for iid data asserts that whenever $F(x) = x^{-\alpha}L(x)$, $\alpha > 0$, then the Hill estimator $\hat{\alpha}(H) = \hat{\alpha}_{[k]}(H)$ satisfies

$$\sqrt{k} \left( \hat{\alpha}(H) - \alpha \right) \xrightarrow{d} N \left( 0, \alpha^2 \right),$$

where $k = k(n) \to \infty$ at an appropriate rate. However, in the formulation of Theorem 4.44 we have not told you the whole story: depending on the precise choice of $k$ and on the slowly varying function $L$, there is an important trade–off between bias and variance possible. It all comes down to second-order behaviour of $L$, i.e., asymptotic behaviour beyond the defining property $L(tx) = L(x)$, $x \to \infty$. Typically, for increasing $k$ the asymptotic variance $\alpha^2/k$ of $\hat{\alpha}(H)$ decreases: so let us take $k$ as large as possible. Unfortunately,

when doing so, a bias may enter!

In our discussion below we follow de Haan and Peng [2]. Similar results are to be found in various other papers; see Embrechts et al. [1], p. 341.

The second-order property needed beyond $F(x) = x^{-\alpha}L(x)$ is that

$$\lim_{x \to \infty} \frac{F(tx)/F(x) - t^{-\alpha}}{a(x)} = t^{-\alpha} \frac{t^\rho - 1}{\rho}, \quad t > 0,$$

exists, where $a(x)$ is a measurable function of constant sign. The right-hand side of (4.34) is to be interpreted as $t^{-\alpha} \text{int}$ if $\rho = 0$. The constant $\rho \leq 0$ is the second-order parameter governing the rate of convergence of $F(tx)/F(x)$ to $t^{-\alpha}$. It naturally follows that $|a(x)|$ is regularly varying with index $\rho$. In terms of $U(t) = F^+(1 - t^{-1})$, (4.34) is equivalent to

$$\lim_{x \to \infty} \frac{U(tx)/U(x) - t^{1/\alpha}}{A(x)} = t^{1/\alpha} \frac{t^{\rho/\alpha} - 1}{\rho/\alpha},$$

where $A(x) = \alpha^{-2}a(U(x))$.

The following result is proved as Theorem 1 in de Haan and Peng [2].

**Theorem 4.49** (The bias–variance trade–off for the Hill estimator)

Suppose (4.35), or equivalently (4.34), holds and $k = k(n) \to \infty$, $k/n \to 0$ as $n \to \infty$. If

$$\lim_{n \to \infty} \sqrt{k} A \left( \frac{n^\rho}{A} \right) = \lambda \in \mathbb{R},$$

then as $n \to \infty$

$$\sqrt{k} \left( \hat{\alpha}(H) - \alpha \right) \xrightarrow{d} N \left( \frac{\alpha^3 \lambda}{\rho - \alpha}, \alpha^2 \right).$$

$\square$

112
Example 4.50 (The choice of the value $k$)

Consider the special case

$$F(x) = cx^{-\alpha}(1 + x^{-\beta})$$

for positive constants $c$, $\alpha$ and $\beta$. We can choose

$$a(x) = \beta x^{-\beta},$$

giving $\rho = -\beta$ in (4.34). Since $U(t) = (ct)^{1/\alpha}(1 + o(1))$, we obtain

$$A(x) = \frac{\beta}{\alpha^2}(cx)^{-\beta/\alpha}(1 + o(1)).$$

Then (4.36) yields $k$ such that

$$k \sim Cn^{(2\beta)/(2\beta + \alpha)}, \quad k \to \infty,$$

where $C$ is a constant, depending on $\alpha$, $\beta$, $c$ and $\lambda$. Moreover, $\lambda = 0$ if and only if $C = 0$, hence

$k = o(n^{(2\beta)/(2\beta + \alpha)})$.

From (4.36) it follows that for $k$ tending to infinity sufficiently slowly, i.e., taking only a moderate number of order statistics into account for the construction of the Hill estimator, $\lambda = 0$ will follow. In this case $\hat{\lambda}^{(H)}$ is an asymptotically unbiased estimator for $\alpha$, as announced in Theorem 4.44. The asymptotic mean squared error equals

$$\frac{1}{k} \left( \alpha^2 + \frac{\alpha^6\lambda^2}{(\rho - \alpha)^2} \right).$$

Theorem 4.49 also explains the typical behaviour of the Hill-plot showing large variations for small $k$ versus small variations (leading to a biased estimate) for large $k$. As always, the trick consists of finding the optimal $k$!

Results such as Theorem 4.49 are useful mainly from a methodological point of view. Condition (4.34) is rarely verifiable in practice. We shall come back to this point later; see the Summary at the end of this section.

References


A Comparison of Different Estimators of the Shape Parameter

At this point, we should pose the question

Which estimators of the shape parameter $\xi$ should one use?
Figure 4.51 (Warning) A comparative study of Hill-plots \( (20 \leq k \leq 1900) \) for 1900 iid simulated data from the distributions: standard exponential (top), heavy-tailed Weibull with shape parameter \( \alpha = 0.5 \) (middle), standard lognormal (bottom). The Hill estimator does not estimate anything reasonable in these cases. A (too) quick glance at these plots could give you an estimate of 3 for the exponential distribution. This should be a warning to everybody using Hill- and related plots! They must be treated with extreme care. One definitely has to contrast such estimates with the exploratory data analysis techniques from Section 4.2.
Figure 4.52 1864 daily log-returns (closing data) of the German stock index DAX (September 20, 1988 – August 24, 1995) (top) and the corresponding Hill–plot of the absolute values (bottom). It gives relatively stable estimates around the value 2.8 in the region $100 \leq k \leq 300$. This is much a wider region than in Figures 4.51. This Hill–plot is also qualitatively different from the exact Pareto case; see Figure 4.45. The deviations can be due to the complicated dependence structure of financial times series.

As so often in statistics, there is no clear-cut answer. It all depends on the possible values of $\xi$ and, as we have seen, on the precise properties of the underlying distribution function $F$. Some general statements can however be made. For $\xi = \alpha^{-1} > 0$ and distribution functions satisfying (4.34), de Haan and Peng proved results similar to Theorem 4.49 for the Pickands estimator (4.24) and the Dekkers–Einmahl–de Haan (DEdH) estimator (4.33). It turns out that in the case $\rho = 0$ the Hill estimator has minimum mean squared error. The asymptotic relative efficiencies for these estimators critically depend on the interplay between $\rho$ and $\alpha$. Both the Pickands and the DEdH estimator work for general $\xi \in \mathbb{R}$. For $\xi > -2$ the DEdH estimator has lower variance than Pickands’s. Moreover Pickands’s estimator is difficult to use since it is rather unstable; see Figures 4.45 and 4.46. There exist various papers combining higher–order expansions of $F$ together with resampling methods.

Several dozens of papers have been devoted to the problem of estimating the extremal index or the index of regular variation. Some of the relevant references can be found in Embrechts et al. pp. 344-345.

4.4.3 Estimating the norming constants

In the previous section we obtained estimators for the shape parameter $\xi$ given iid data $X_1, \ldots, X_n$ with distribution function $F \in \text{MDA}(H_\xi)$. Recall that the latter condition is equivalent to

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H_\xi$$

for appropriate norming constants $c_n > 0$ and $d_n \in \mathbb{R}$. We also know that this relation holds if and only if

$$nF(c_n x + d_n) \rightarrow -\ln H_\xi(x), \quad n \rightarrow \infty, \quad x \in \mathbb{R}.$$
As we have already seen in Section 4.4.1, norming constants enter in quantile and tail estimation; see (4.19). Below we discuss one method how norming constants can be estimated. Towards the end of this section we give some further references to other methods. In Section 3.3 we gave analytic formulae linking the norming sequences \( (c_n) \) and \( (d_n) \) with the tail \( \overline{F} \). For instance, in the Gumbel case \( \xi = 0 \) with right endpoint \( x_F = \infty \) the following formulae were derived in Theorem 3.41

\[
(4.37) \quad c_n = a \left( d_n \right), \quad d_n = F^\leftarrow \left( 1 - n^{-1} \right),
\]

where \( a(\cdot) \) stands for the auxiliary function which can be taken in the form

\[
a(x) = \int_x^\infty \frac{F(y)}{F(x)} \, dy.
\]

Notice the problem: on the one hand, we need the norming constants \( c_n \) and \( d_n \) in order to obtain quantile and tail estimates. On the other hand, (4.37) defines them as functions of just that tail, so it seems that

\[
\text{this surely is a race we cannot win!}
\]

Though this is partly true let us see how far we can get. We will try to convince you that the appropriate reformulation of the above sentence is:

\[
\text{this is a race which will be difficult to win!}
\]

Consider the more general set-up \( F \in \text{MDA}(H_\xi), \xi \geq 0 \), including the for our purposes most important cases of the Fréchet and the Gumbel distribution. In Example 3.47 we showed how one can unify these two maximum domains of attraction by the logarithmic transformation

\[
x^* = \ln(1 \vee x), \quad x \in \mathbb{R}.
\]

Together with Theorem 3.40 the following useful result can be obtained.

**Lemma 4.53** (Embedding MDA \((H_\xi), \xi \geq 0, \) in MDA \((\Lambda)\))

Let \( X_1, \ldots, X_n \) be iid with distribution function \( F \in \text{MDA}(H_\xi), \xi \geq 0 \), with \( x_F = \infty \) and norming constants \( c_n > 0 \) and \( d_n \in \mathbb{R} \). Then \( X^*_1, \ldots, X^*_n \) are iid with distribution function \( F^* \in \text{MDA}(\Lambda) \) and auxiliary function

\[
a^*(t) = \int_t^\infty \frac{F^*(y)}{F^*(t)} \, dy.
\]

The norming constants can be chosen as

\[
d^*_n = (F^*)^\leftarrow (1 - n^{-1}),
\]

\[
c^*_n = a^*(d^*_n) = \int_{d^*_n}^\infty \frac{F^*(y)}{F^*(d^*_n)} \, dy \sim n \int_{d^*_n}^\infty F^*(y) \, dy.
\]

\( \square \)

In Section 4.4.1 we tried to convince you that our estimators have to be based on the \( k \) largest order statistics \( X_{(n-k+1)}, \ldots, X_{(n)} \), where \( k = k(n) \to \infty \). From the above lemma we obtain estimators
if we replace \( F^* \) by the empirical distribution function \( F_n^* \):

\[
\hat{d}_{n/k}^* = X_{(n-k)} = \ln(1 \vee X_{(n-k)})
\]

\[
\hat{c}_{n/k}^* = \frac{n}{k} \int_{\hat{d}_{n/k}^*}^{\infty} F_n^*(y) \, dy
\]

\[
= \frac{n}{k} \int_{\ln X(n)}^{\ln X(n-k)} F_n^*(y) \, dy
\]

\[
= \frac{1}{k} \sum_{j=1}^{k} \ln X_{(n-j+1)} - \ln X_{(n-k)}.
\] (4.38)

The latter is a version of the Hill estimator. The change from \( k \) to \( k+1 \) in (4.38) is asymptotically unimportant.

Next we make the transformation back from \( F^* \) to \( F \) via

\[
\frac{n}{k} P \left( X^* > c_{n/k}^* x + d_{n/k}^* \right) = \frac{n}{k} P \left( X > \exp \left\{ c_{n/k}^* x + d_{n/k}^* \right\} \right), \quad x > 0.
\]

Finally we use that \( F^* \in \text{MDA}(\Lambda) \), hence the left-hand side converges to \( e^{-x} \) as \( n \to \infty \), provided that \( n/k \to \infty \). We thus obtain the tail estimator

\[
\hat{F}(x) = \frac{k}{n} \left( \exp \left\{ -\hat{d}_{n/k}^* + \ln x \right\} \right)^{-1/\hat{c}_{n/k}^*}
\]

\[
= \frac{k}{n} \left( \frac{x}{X_{(n-k)}} \right)^{-1/\hat{c}_{n/k}^*}.
\]

This tail estimator was already obtained by Hill; for the exact model (4.26); see (4.30). As a quantile estimator we obtain

\[
\hat{x}_p = \left( \frac{n}{k} (1-p) \right)^{-\hat{c}_{n,k}} X_{(n-k)}.
\]

Time to summarise all of this:

Let \( \hat{\xi}^{(H)} \) denote the Hill estimator for \( \xi \), i.e.,

\[
\hat{\xi}^{(H)} = \frac{1}{k} \sum_{j=1}^{k} \ln X_{(n-j+1)} - \ln X_{(n-k+1)}.
\]
Let $X_1, \ldots, X_n$ be a sample from $F \in \text{MDA}(H_\xi)$, $\xi \geq 0$, and $k = k(n) \to \infty$ such that $k/n \to 0$. Then for $x$ large enough, a tail estimator for $F(x)$ becomes

$$(F(x))^\hat{\xi} = \frac{k}{\overline{X(n-k)}} \left( \frac{x}{X(n-k)} \right)^{-1/\overline{\xi}_n}.$$ 

The quantile $x_p$ so that $F(x_p) = p \in (0,1)$ can be estimated by

$$\hat{x}_p = \left( \frac{n}{k} (1-p) \right)^{-\overline{\xi}_n} X_{(n-k)}.$$ 

Comments

It should be clear from the above that similar quantile estimation methods can be worked out using alternative parameter estimators as discussed in the previous sections.

4.4.4 Tail and quantile estimation

As before, assume that we consider a sample $X_1, \ldots, X_n$ of iid random variables with distribution function $F \in \text{MDA}(H_\xi)$ for some $\xi \in \mathbb{R}$. Let $0 < p < 1$ and $x_p$ denote the corresponding $p$-quantile.

The whole point behind the domain of attraction condition $F \in \text{MDA}(H_\xi)$ is to be able to estimate quantiles outside the range of the data, i.e., $p > 1 - 1/n$. The latter is of course equivalent to finding estimators for the tail $\overline{F}(x)$ with $x$ large. In Sections 4.3.5 and 4.4.3 we have already discussed some possibilities. Indeed, whenever we have estimators for the shape parameter $\xi$ and the normalizing constants $c_n$ and $d_n$, natural estimators of $x_p$ and $\overline{F}(x)$ can immediately be derived from the defining property of $F \in \text{MDA}(H_\xi)$. We want to discuss some of them more in detail and point at their most important properties and caveats. From the start we would like to stress that estimation outside the range of the data can be made only if extra model assumptions are imposed. There is no magical technique which yields reliable results for free. One could formulate this as in finance:

*There is no free lunch when it comes to high quantile estimation!*

In our discussion below, we closely follow the paper by Dekkers and de Haan [1]. The main results are formulated in terms of conditions on $U(t) = F^\leftarrow(1 - t^{-1})$ so that $x_p = U(1/(1-p))$. Denoting $U_n(t) = F_n^\leftarrow(1 - t^{-1})$, where $F_n^\leftarrow$ is the empirical quantile function,

$$U_n \left( \frac{n}{k-1} \right) = F_n^\leftarrow \left( 1 - \frac{k-1}{n} \right) = X_{(n-k+1)}, \quad k = 1, \ldots, n.$$ 

Hence $X_{(n-k+1)}$ appears as a natural estimator of the $(1-(k-1)/n)$-quantile. The range $[X_{(1)}, X_{(n)}]$ of the data allows one to make a within-sample estimation up to the $(1 - n^{-1})$-quantile. For high quantile estimation the following situations are of main interest:

1. high quantiles within the sample: $p = p_n \uparrow 1$, $n(1-p_n) \to c, c \in (1, \infty]$,

2. high quantiles outside the sample: $p = p_n \uparrow 1$, $n(1-p_n) \to c, 0 \leq c < 1$.

Case (a) for $c = \infty$ is addressed by the following result which is Theorem 3.1 in Dekkers and de Haan [1]. It basically tells us that we can just use the empirical quantile function for estimating $x_p$.

118
Theorem 4.54 (Estimating high quantiles I)
Suppose $X_1, \ldots, X_n$ is an iid sample from $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, and $F$ has a positive density $f$. Assume that the density $U'$ is regularly varying with index $\xi - 1$. Write $p = p_n$, and $k = k(n) = \lfloor n(1 - p_n) \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of $x$. If the conditions
\[ p_n \to 1 \quad \text{and} \quad n(1 - p_n) \to \infty \]
hold then
\[
\sqrt{2k} \frac{X_{(n-k+1)} - x_p}{X_{(n-k+1)} - X_{(n-2k+1)}} \overset{d}{\to} N \left( 0, 2^{2\xi+1} \frac{\xi^2}{(2\xi - 1)^2} \right).
\]
\[ \Box \]

Remark. 1) The condition "$U'$ is regularly varying with index $\xi - 1" can be reformulated in terms of $F$. For instance for $\xi > 0$, the condition becomes "$f$ is regularly varying with index $1 - 1/\xi". \]

In Theorem 3.50 we characterised $F \in \text{MDA}(H_\xi)$ through the asymptotic behaviour of $U$:
\[
\lim_{t \to \infty} \frac{U(tx) - U(t)}{U(ty) - U(t)} = \frac{\xi^y - 1}{y^\xi - 1}, \quad x, y > 0, \quad y \neq 1.
\]

For $\xi = 0$ the latter limit has to be interpreted as $\ln x / \ln y$. We can rewrite the above as follows
\[
(4.39) \quad U(tx) = \frac{x^\xi - 1}{1 - y^\xi} (U(t) - U(ty))(1 + o(1)) + U(t).
\]

Using this relation, a heuristic argument suggests an estimator for the quantiles $x_p$ outside the range of the data. Indeed, replace $U$ by $U_n$ in (4.39) and put $y = 1/2$, $x = (k-1)/(n(1-p))$ and $t = n/(k-1)$. Substitute $\xi$ by an appropriate estimator $\hat{\xi}$. Doing so, and neglecting $o(1)$-terms one finds the following estimator of $x_p$:
\[
(4.40) \quad \hat{x}_p = \frac{k/(n(1-p))}{1 - 2^{-\hat{\xi}}} (X_{(n-k+1)} - X_{(n-2k+1)}) + X_{(n-k+1)}.
\]

The following result is Theorem 3.3 in Dekkers and de Haan [1].

Theorem 4.55 (Estimating high quantiles II)
Suppose $X_1, \ldots, X_n$ is an iid sample from $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, and assume that $\lim_{n \to \infty} n(1-p) = c$ for some $c > 0$. Let $\hat{x}_p$ be defined by (4.40) with $\hat{\xi}$ the Pickands estimator (4.24). Then for every fixed $k > c$,
\[
\frac{\hat{x}_p - x_p}{X_{(n-k+1)} - X_{(n-2k+1)}} \overset{d}{\to} Y,
\]
where
\[
(4.41) \quad Y = \frac{(k/c)^{\hat{\xi}} - 2^{-\hat{\xi}}}{1 - 2^{-\hat{\xi}}} + \frac{1 - (Q_k/c)^{\hat{\xi}}}{\exp\{\xi H_k\} - 1}.
\]

The random variables $H_k$ and $Q_k$ are independent, $Q_k$ has a gamma distribution with parameter $2k + 1$ and
\[
H_k = \sum_{j=k+1}^{2k} \frac{E_j}{j}
\]
for iid standard exponential random variables $E_1, E_2, \ldots$. \[ \Box \]
Remarks. 2) The case $0 < c < 1$ of Theorem 4.55 corresponds to extrapolation outside the range of the data. For the extreme case $c = 0$, a relevant result is to be found in de Haan [2], Theorem 5.1. Most of these results depend on highly technical conditions on the asymptotic behaviour of $\hat{F}$. There is a strong need for comparative numerical studies on these high quantile estimators.

3) Approximations to the distribution function of $Y$ in (4.41) can be worked out explicitly.

4) As for the situation of Theorem 4.54, no results seem to exist concerning the optimal choice of $k$. For the consistency of the Pickands estimator $\hat{\xi}$, which is part of the estimator $\hat{\xi}_p$, one actually needs $k = k(n) \to \infty$; see Theorem 4.39.

5) In the case $\xi < 0$ similar results can be obtained for the estimation of the right endpoint $x_F$ of $F$. We refer the interested reader to Dekkers and de Haan [1] for further details and some examples.

References


Summary

Throughout Section 4.4, we have discovered various estimators for the important shape parameter $\xi$ of distribution functions in the maximum domain of attraction of the GEV. From these and further estimators, either for the location and scale parameters and/or norming constants, estimators for the tail $\hat{F}$ and high quantiles resulted. The properties of these estimators crucially depend on the higher-order behaviour of the underlying distribution tail $\hat{F}$. The latter is unfortunately not verifiable in practice.

On various occasions we hinted at the fact that the determination of the number $k$ of upper order statistics finally used remains a delicate point in the whole set-up. Various papers exist which offer a semi-automatic or automatic, so-called "optimal", choice of $k$. See for instance Beirlant et al. [1] for a regression based procedure with various examples to insurance data, and Danielson and de Vries [2] for an alternative method motivated by examples in finance. We personally prefer a rather pragmatic approach realising that, whatever method one chooses, the "Hill horror plot" (Figure 4.47) would fool most, if not all. It also serves to show how delicate a tail analysis in practice really is. On the other hand, in the "nice case" of exact Pareto behaviour, all methods work well; see Figures 4.45.

Our experience in analysing data, especially in (re)insurance, shows that in practice one is often faced with data which are clearly heavy-tailed and for which "exact" Pareto behaviour of $\hat{F}(x)$ sets in for relatively low values of $x$; see for instance Figure 4.19. The latter is not so obvious in the world of finance. This is mainly due to more complicated dependence structures in most of the finance data; compare for instance Figures 4.65 and 4.66. A "nice" example from the realm of finance was discussed in Figure 4.52. The conclusion "the data are heavy-tailed" invariably has to be backed up with information from the user who provided the data in the first place! Furthermore, any analysis performed has to be supported by exploratory data analysis techniques as outlined in Section 4.2. Otherwise, situations as explained in Figure 4.51 may occur.

It is our experience that in many cases one obtains a Hill- (or related) plot which tends to have a fairly noticeable horizontal stretch across different (often lower) $k$-values. A choice of $k$
in such a region is to be preferred. Though the above may sound vague, we suggest the user of extremal event techniques to experiment on both simulated as well as real data in order to get a feeling for what is going on. A final piece of advice along this route: never go for one estimate only. Calculate and plot always estimates of the relevant quantities (a quantile say) across a wide range of $k$-values; for examples see Figures 4.4.1 and 4.4.2. In the next section we shall come back to this point, replacing $k$ by a threshold $u$. We already warn the reader beforehand: the approach offered in Section 4.5 suffers from the same problems as those discussed in this section.

Comments

So far, we only gave a rather brief discussion on the statistical estimation of parameters, tails and quantiles in the heavy-tailed case. This area is still under intensive investigation so that at present no complete picture can be given. Besides the availability of a whole series of mathematical results a lot of insight is obtained through simulation and real life examples. In the next section some further techniques and indeed practical examples will be discussed. The interested reader is strongly advised to consult the papers referred to so far. An interesting discussion on the main issues is de Haan [4], where also applications to currency exchange rates, life span estimation and sea level data are given. In Einmahl [3] a critical discussion concerning the exact meaning of extrapolating outside the data is given. He stresses the usefulness of the empirical distribution function as an estimator.

References


4.5 Fitting excesses over a threshold

4.5.1 Fitting the GPD

Methodology introduced so far was obtained either on the assumption that the data come from a GEV (see Section 4.3) or belong to its maximum domain of attraction (see Section 4.4). We based statistical estimation of the relevant parameters on maximum likelihood, the method of probability-weighted moments or some appropriate condition of regular variation type. In Section 4.4 we laid the foundation to an alternative approach based on exceedances of high thresholds. The key idea of this approach is explained below.

Suppose

\[ X, X_1, \ldots, X_n \text{ are iid with distribution function } F \in \text{MDA}(H_\xi) \text{ for some } \xi \in \mathbb{R}. \]

First, choose a high threshold $u$ and denote by

\[ N_u = \text{card} \{ i : i = 1, \ldots, n, \ X_i > u \} \]
Figure 4.56 Data $X_1, \ldots, X_{13}$ and the corresponding excesses $Y_1, \ldots, Y_{N_u}$ over the threshold $u$. 

the number of exceedances of $u$ by $X_1, \ldots, X_n$. We denote the corresponding excesses by $Y_1, \ldots, Y_{N_u}$; see Figure 4.56. The excess distribution function of $X$ is given by

$$F_u(y) = P(X - u \leq y \mid X > u) = P(Y \leq y \mid X > u), \quad y \geq 0;$$

see Definition 3.51. The latter relation can also be written as

$$(4.42) \quad \overline{F}(u + y) = \overline{F}(u) \overline{F}_u(y).$$

Now recall the definition of the generalised Pareto distribution (GPD) from Definition 3.54: a GPD $G_{\xi, \beta}$ with parameters $\xi \in \mathbb{R}$ and $\beta > 0$ has distribution tail

$$\overline{G}_{\xi, \beta}(x) = \begin{cases} 
(1 + \frac{x}{\beta})^{-1/\xi} & \text{if } \xi \neq 0, \\
\frac{1}{e^{-x/\beta}} & \text{if } \xi = 0,
\end{cases} \quad x \in D(\xi, \beta),$$

where

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

Theorem 3.58(b) gives a limit result for $\overline{F}_u(y)$, namely

$$\lim_{u \uparrow x \in \mathbb{R}^+} \sup_{0 < x < x \in \mathbb{R} - u} \left| \overline{F}_u(x) - \overline{G}_{\xi, \beta(u)}(x) \right| = 0,$$

for an appropriate positive function $\beta$. Based on this result, for $u$ large, the following approximation suggests itself:

$$(4.43) \quad \overline{F}_u(y) \approx \overline{G}_{\xi, \beta(u)}(y).$$

122
It is important to note that $\beta$ is a function of the threshold $u$. In practice, $u$ will have to be taken sufficiently large. Given such a $u$, $\xi$ and $\beta = \beta(u)$ are estimated from the excess data, so that the resulting estimates depend on $u$; see our discussion below.

Relation (4.42) then suggests a method for estimating the far end tail of $F$ by estimating $\widetilde{F}_u(y)$ and $\widetilde{F}(u)$ separately. A natural estimator for $\widetilde{F}(u)$ is given by the empirical distribution function

$$
(\widetilde{F}(u)) = \frac{1}{n} \sum_{i=1}^{n} I(X_i > u) = \frac{N_u}{n}.
$$

On the other hand, the generalised Pareto approximation (4.43) (remember that $u$ is large!) motivates an estimator of the form

$$
(\widetilde{F}_u(y)) = G_{\hat{\xi}, \hat{\beta}}(y)
$$

for appropriate $\hat{\xi} = \hat{\xi}_n$ and $\hat{\beta} = \hat{\beta}_n$.

A resulting estimator for the tail $\widetilde{F}(u + y)$ for $y > 0$ then takes on the form

$$
(\widetilde{F}(u + y)) = \frac{N_u}{n} \left( 1 + \frac{y}{\hat{\beta}} \right)^{-1/\hat{\xi}}.
$$

In the Fréchet and Gumbel case ($\xi \geq 0$), the domain restriction in (4.45) is $y \geq 0$, clearly stressing that we estimate $\widetilde{F}$ in the upper tail. An estimator of the quantile $x_p$ results immediately:

$$
\hat{x}_p = u + \frac{\hat{\beta}}{\hat{\xi}} \left( \left( \frac{n}{N_u} (1 - p) \right)^{-\hat{\xi}} - 1 \right).
$$

Furthermore, for $\hat{\xi} < 0$ an estimator of the right endpoint $x_F$ of $F$ is given by

$$
\hat{x}_F = u - \frac{\hat{\beta}}{\hat{\xi}}.
$$

The latter is obtained by putting $\hat{x}_F = \hat{x}_1$ (i.e. $p = 1$) in (4.46). In Section 4.4.1 we said that the method of exceedances belongs to the realm of point process theory. From (4.45) and (4.46) this is clear: statistical properties of the resulting estimators crucially depend on the distributional properties of the point process of exceedances ($N_u$).

The above method is intuitively appealing. It goes back to hydrologists. Over the last 25 years they have developed this estimation procedure under the acronym of Peaks Over Threshold (POT) method. In order to work out the relevant estimators the following input is needed:

- reliable models for the point process of exceedances,
- a sufficiently high threshold $u$,
- estimators $\hat{\xi}$ and $\hat{\beta}$,
- and, if necessary, an estimator $\hat{\nu}$ for location.

If one wants to choose an optimal threshold $u$ one faces similar problems as for the choice of the number $k$ of upper order statistics for the Hill estimator. A value of $u$ too high results in too few exceedances and consequently high variance estimators. For $u$ too small estimators become biased.
Theoretically, it is possible to choose \( u \) asymptotically optimal by a quantification of a bias versus variance trade-off, very much in the same spirit as discussed in Theorem 4.49. In reality however, the same problems as already encountered for other tail estimators before do occur. We refer to the examples in Section 4.5.2 for illustrations on this. 

One method which is of immediate use in practice is based on the linearity of the mean excess function \( e(u) \) for the GPD. From Theorem 3.58(c) we know that for a random variable \( X \) with distribution function \( G_{\xi, \beta}, \)

\[
e(u) = E(X - u \mid X > u) = \frac{\beta + \xi u}{1 - \xi}, \quad u \in D(\xi, \beta), \xi < 1,
\]
hence \( e(u) \) is linear. Recall from (4.4) that the empirical mean excess function of a given sample \( X_1, \ldots, X_n \) is defined by

\[
e_n(u) = \frac{1}{N_u} \sum_{i \in \Delta_n(u)} (X_i - u), \quad u > 0,
\]
where as before \( N_u = \text{card} \{ i : i = 1, \ldots, n, X_i > u \} = \text{card} \Delta_n(u) \). The remark above now suggests a graphical approach for choosing \( u \):

choose \( u > 0 \) such that \( e_n(x) \) is approximately linear for \( x \geq u \).

The key difficulty of course lies in the interpretation of approximately. Only practice can tell! One often observes a change in the slope of \( e_n(u) \) for some value of \( u \). Referring to some examples on sulphate and nitrate level in an acid rain study Smith [3], p. 460, says the following:

*The general form of these mean excess plots is not atypical of real data, especially the change of slope near 100 in both plots. Smith [2] observed similar behaviour in data on extreme insurance claims, and Davison and Smith [1] used a similar plot to identify a change in the distribution of the threshold form of the River Nidd data. Such plots therefore appear to be an extremely useful diagnostic in this form of analysis.*

The reader should never expect a unique choice of \( u \) to appear. We recommend using plots, to reinforce judgement and common sense and compare resulting estimates across a variety of \( u \)-values. In applications we often prefer plots indicating the threshold value \( u \), as well as the number of exceedances used for the estimation, on the horizontal axes; the estimated value of the parameter or the quantile, say, is plotted on the vertical one. The latter is illustrated in Section 4.5.2. As can be seen from the examples, and indeed can be proved, all these plots exhibit the same behaviour as the Hill- and Pickands-plots before: high variability for \( u \) large (few observations) versus bias for \( u \) small (many observations, but at the same time the approximation (4.43) may not be applicable).

Concerning estimators for \( \xi \) and \( \beta \), various methods similar to those discussed in Section 4.4.2 exist.

**References**


**Figure 4.57** A “horror plot” for the MLE of the shape parameter $\xi$ of a GPD. The simulated data come from a distribution function given by $F(x) = 1/(x \ln x)$. Left: sample size 1000. Right: sample size 10000. The upper horizontal axis indicates the threshold value $u$, the lower one the corresponding number $k$ of exceedances of $u$. As for Hill estimation (see Figure 4.47) MLE also becomes questionable for such perturbed Pareto tails.

**Maximum likelihood estimation**

The following results are to be found in Smith [1].

Recall that our original data $x = (X_1, \ldots, X_n)$ are iid with common distribution function $F$. Assume $F$ is GPD with parameters $\xi$ and $\beta$, thus the density $f$ becomes

$$f(x) = \frac{1}{\beta} \left( 1 + \xi \frac{x}{\beta} \right)^{-\frac{1}{\xi} - 1}, \quad x \in D(\xi, \beta).$$

The log-likelihood function equals

$$\ell((\xi, \beta); x) = -n \ln \beta - \left( \frac{1}{\xi} + 1 \right) \sum_{i=1}^{n} \ln \left( 1 + \frac{\xi}{\beta} X_i \right).$$

Notice that the arguments of the above function have to satisfy the domain restriction $X_i \in D(\xi, \beta)$. For notational convenience, we have dropped that part from the likelihood function. Recall that $D(\xi, \beta) = [0, \infty)$ for $\xi \geq 0$. Now likelihood equations can be derived and solved numerically yielding the MLE $\hat{\xi}_n, \hat{\beta}_n$. This method works fine if $\xi > -1/2$, in the latter case one can show that

$$n^{1/2} \left( \hat{\xi}_n - \xi, \frac{\hat{\beta}_n}{\beta} - 1 \right) \overset{d}{\to} N(0, M^{-1}), \quad n \to \infty,$$

where

$$M^{-1} = \begin{pmatrix}
1 + \xi & -1 \\
-1 & 2
\end{pmatrix},$$

and $N(\mu, \Sigma)$ stands for the bivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$. The usual MLE properties like consistency and asymptotic efficiency hold.

Because of (4.43), it is more realistic to assume a GPD for the excesses $Y_1, \ldots, Y_N$, where $N = N_u$ is independent of the $Y_i$. The resulting conditional likelihood equations can be solved best
via a reparametrisation $(\xi, \beta) \to (\xi, \tau)$, where $\tau = -\xi / \beta$. This leads to the solution

$$\hat{\xi} = \bar{\xi}(\tau) = N^{-1} \sum_{i=1}^{N} \ln \left( 1 - \tau Y_i \right),$$

where $\tau$ satisfies

$$h(\tau) = \frac{1}{\tau} + \frac{1}{N} \left( \frac{1}{\bar{\xi}(\tau)} + 1 \right) \sum_{i=1}^{N} \frac{Y_i}{1 - \tau Y_i} = 0.$$  

The function $h(\tau)$, defined for $\tau \in (-\infty, \max(Y_1, \ldots, Y_N))$, is continuous at 0. Letting $u = u_n \to \infty$, Smith [1] derives various limit results for the distribution of $(\hat{\xi}_N, \hat{\beta}_N)$. As in the case of the Hill estimator (see Theorem 4.49), an asymptotic bias may enter. The latter again crucially depends on a second-order condition for $\chi$.

References


Method of probability–weighted moments

Similarly to our discussion in Section 4.3.4, Hosking and Wallis [1] also worked out a probability–weighted moment approach for the GPD. This is based on the quantities (see Theorem 3.58(a))

$$w_r = EZ \left( \overline{G}_{\xi, \beta}(Z) \right)^r = \frac{\beta}{(r + 1)(r + 1 - \xi)}, \quad r = 0, 1,$$

where $Z$ has GPD $G_{\xi, \beta}$. We immediately obtain

$$\beta = \frac{2w_0w_1}{w_0 - 2w_1} \quad \text{and} \quad \xi = 2 - \frac{w_0}{w_0 - 2w_1}.$$

If we now replace $w_0$ and $w_1$ by empirical moment estimators, one obtains the probability–weighted moment estimators $\hat{\beta}$ and $\hat{\xi}$. Hosking and Wallis [1] give formulae for the approximate standard errors of these estimators. They compare their approach to the MLE approach and come to the conclusion that in the case $\xi \geq 0$ the method of probability–weighted moments offers a viable alternative. However, as we have already stressed in the case of a GEV, maximum likelihood methodology allows us to fit much more general models including time dependence of the parameters and the influence of explanatory variables.

References


4.5.2 An application to real data

In the above discussion we have outlined the basic principles behind the GPD fitting programme. Turning to the practical applications, two main issues need to be addressed:

1. fit the conditional distribution function $F_0(x)$ for an appropriate range of $x$– (and indeed $u$–) values;
2. fit the unconditional distribution function $F(x)$, again for appropriate $x$-values.

Though formulae (4.44) and (4.45) in principle solve the problem, in practice care has to be taken about the precise range of the data available and/or the interval over which we want to fit. In our examples, we have used a set-up which is motivated mainly by insurance applications.

Take for instance the Danish fire insurance data. Looking at the ME-plot in Figure 4.18 we see that the data are clearly heavy-tailed. In order to estimate the shape parameter $\xi$ a choice of the threshold $u$ (equivalently, of the number $k$ of exceedances) has to be made. In the light of the above discussion concerning the use of ME-plots at this stage, we suggest a first choice of $u = 10$ resulting in 109 exceedances. This means we choose $u$ from a region above which the ME-plot is roughly linear. An alternative choice would perhaps be in the range $u \approx 18$. Figure 4.60 (top, left) gives the resulting estimates of $\xi$ as a function of $u$ (upper horizontal axis) and of the number $k$ of exceedances of $u$ (lower horizontal axis): the resulting plot is relatively stable with estimated values mainly in the range (0.4, 0.6). Compare this plot with the Hill-plot in Figure 4.41. For $u = 10$ maximum likelihood estimates $\hat{\xi} = 0.497$ (s.e. = 0.143) and $\hat{\beta} = 6.98$ result. A change to $u = 18$ yields $\hat{\xi} = 0.735$ (s.e. = 0.253) based on $k = 47$ exceedances.

From these estimates, using (4.44), an estimate for the (conditional) excess distribution function $F_u(x)$ can be plotted. Following standard practice in reinsurance, in Figure 4.60 (top, right) we plot the shifted distribution function $F_u(x-u), x \geq u$. In the language of reinsurance the latter procedure estimates the probability that a claim lies in a given interval, given that the claim has indeed pierced the level $u = 10$.

Though the above estimation (once $u = 10$ is chosen) only uses the 109 largest claims, a crucial question still concerns where the layer ($u = 10$ and above) is to be positioned in the total portfolio; i.e. we also want to estimate the tail of the unconditional distribution function $F$ which yields information on the frequency with which a given high level $u$ is pierced. At this point we need the full data-set and turn to formula (4.45). A straightforward calculation allows us to express $\hat{F}(z)$ as a three-parameter GPD:

\begin{equation}
\hat{F}(z) = 1 - \left(1 + \frac{z - u - \hat{\beta}}{\hat{\beta}}\right)^{-1/\hat{\xi}}, \quad z \geq u, \tag{4.47}
\end{equation}

where

$$\hat{\beta} = \frac{\hat{\beta}}{\hat{\xi}} \left(\frac{N_u}{n} \hat{\xi} - 1\right) \quad \text{and} \quad \hat{\beta} = \hat{\beta} \left(\frac{N_u}{n}\right).$$

We would like to stress that the above distribution function is designed only to fit the data well above the threshold $u$. Below $u$, where the data are typically abundant, various standard techniques can be used; for instance the empirical distribution function. By combining both, GPD above $u$ and empirical below $u$, a good overall fit can be obtained. There are of course various possibilities to fine-tune such a construction.

Finally, using the above fit to $F(z)$, we can give estimates for the $p$-quantiles, $p \geq F(u)$. In Figure 4.60 (bottom) we have summarised the 0.99-quantile estimates obtained by the above method across a wide range of $u$-values (upper horizontal axis), i.e. for each $u$-value a new model was fitted and $x_{0.99} \hat{u}$ estimated. Alternatively, the number of exceedances of $u$ is indicated on the lower horizontal axis. For these data a rather stable picture emerges. A value in the range (25, 26) follows. Confidence intervals can be calculated. The software needed to do these, and further analyses are discussed in the Notes and Comments below.

Figure 4.61 for the industrial fire data (see Figure 4.19) and Figure 4.58 for the BMW share prices (see Figure 4.17) can be interpreted in a similar way.
Figure 4.58 The positive BMW log-returns from Figure 4.17. Top, left: MLE of \( \xi \) as a function of \( u \) and \( k \) with asymptotic 95\% confidence band. Top, right: GPD-fit to \( F_u(x-u), x \geq u \), on log-scale. Middle: GPD tail-fit for \( \bar{F}(x+u), x \geq 0 \). Bottom: estimates of the 95\%-quantile as a function of the threshold \( u \) (upper horizontal axis) and of the corresponding number \( k \) of the upper order statistics (lower horizontal axis). A GPD with parameters \( \xi = 0.24, \beta = 0.013 \) is fitted, corresponding to \( u = 0.0355 \) and \( k = 100 \); i.e. the distribution almost has an infinite 4th moment.
Figure 4.59 The absolute values of the negative BMW log--returns from Figure 4.17. Top, left: MLE of $\xi$ as a function of $u$ and $k$ with asymptotic 95% confidence band. Top, right: GPD--fit to $F_u(x - u)$, $x \geq u$, on log--scale. Middle: GPD tail--fit for $F(x + u)$, $x \geq 0$. Bottom: estimates of the 95%--quantile as a function of the threshold $u$ (upper horizontal axis) and of the corresponding number $k$ of the upper order statistics (lower horizontal axis). A GPD with parameters $\xi = 0.343$, $\beta = 0.011$ is fitted, corresponding to $u = 0.0345$ and $k = 100$; i.e. the distribution has an infinite 3rd moment. As mentioned in the discussion of Figure 4.17, the left tail of the distribution appears to be heavier than the right one.
Figure 4.60 The Danish fire insurance data; see Figure 4.18. Top, left: MLE for the shape parameter $\xi$ of the GPD. The upper horizontal axis indicates the threshold $u$, the lower one the number $k$ of exceedances/upper order statistics involved in the estimation. Top, right: fit of the shifted excess distribution function $F_u(x-u)$, $x \geq u$, on log-scale. Middle: GPD tail-fit for $\bar{F}(x+u)$, $x \geq 0$. Bottom: estimates of the 0.99-quantile as a function of $u$ (upper horizontal axis) and $k$ (lower horizontal axis). A GPD with parameters $\xi = 0.497$ and $\beta = 6.98$ is fitted, corresponding to $k = 109$ exceedances of $u = 10$. Compare also with Figure 4.41 for the corresponding Hill-fit.
Figure 4.61 The industrial fire insurance data; see Figure 4.19. Top, left: MLE for the shape parameter $\xi$ of the GPD. The upper horizontal axis indicates the threshold $u$, the lower one the number $k$ of exceedances/upper order statistics involved in the estimation. Top, right: fit of the shifted excess distribution function $F_u(x-u)$, $x \geq u$, on log-scale. Middle: GPD tail-fit for $\overline{F}(x+u)$, $x \geq 0$. Bottom: estimates of the 0.99-quantile as a function of $u$ (upper horizontal axis) and $k$ (lower horizontal axis). A GPD with parameters $\xi = 0.747$ and $\beta = 48.1$ is fitted, corresponding to $k = 149$ exceedances of $u = 100$. Compare also with Figure 4.42 for the corresponding Hill-fit.
Mission improbable: how to predict the unpredictable

On studying the above data analyses, the reader may have wondered why we restricted our plots to $x_{0.99}$ for the Danish and industrial insurance data, and $x_{0.95}$ for the BMW data. In answering this question, we restrict attention to the insurance data. At various stages throughout the text we hinted at the fact that extreme value theory (EVT) offers methodology allowing for extrapolation outside the range of the available data. The reason why we are very reluctant to produce plots for high quantiles like 0.999 or even 0.9999, is that we feel that such estimates are to be treated with extreme care. We cite here a statement of Richard Smith: 

"There is always going to be an element of doubt, as one is extrapolating into areas one doesn’t know about. But what EVT is doing is making the best use of whatever data you have about extreme phenomena." Both fire insurance data-sets have information on extremes, and indeed EVT has produced models which make best use of whatever data we had at our disposal. Using these models, estimates for the $p$-quantiles $x_p$ for every $p \in (0, 1)$ can be given. The statistical reliability of these estimates becomes, as we have seen, very difficult to judge in general. Though we can work out approximate confidence intervals for these estimators, such constructions strongly rely on mathematical assumptions which are unverifiable in practice.

In Figures 4.62 and 4.63 we have reproduced the GPD estimates for $x_{0.999}$ and $x_{0.9999}$ for both the Danish and the industrial fire data. These plots should be interpreted with the above quote from Smith in mind. For instance, for the Danish fire insurance data we see that the estimate of about 25 for $x_{0.99}$ jumps at 90 for $x_{0.999}$ and at around 300 for $x_{0.9999}$. Likewise for the industrial fire, we get an increase from around 190 for $x_{0.99}$ to about 1 400 for $x_{0.999}$ and 10 000 for $x_{0.9999}$. These model-based estimates could form the basis for a detailed discussion with the actuary/underwriter/broker/client responsible for these data. One can use them to calculate so-called technical premiums, which are to be interpreted as those premiums which we as statisticians believe to most honestly reflect the information available from the data. Clearly many other factors have to enter at this stage of the discussion. We already stressed before that in dealing with high layers/extremes one should always consider total exposure as an alternative. Economic considerations, management strategy, market forces will enter so that by using all these inputs we are able to come up with a premium acceptable both for the insurer as well as the insured. Finally, once the EVT model-machinery (GPD for instance) is put into place, it offers an ideal platform for simulation experiments and stress-scenarios. For instance, questions about the influence of single or few observations and model-robustness can be analysed in a straightforward way. Though we have restricted ourselves to a more detailed discussion for the examples from insurance, similar remarks apply to financial or indeed any other kind of data where extremal events play an important role.

Comments

The POT method has been used by hydrologists for more than 25 years. It has also been suggested for dealing with large claims in insurance; see for instance Reiss [8] and Teugels [14]. It may be viewed as an alternative approach to the more classical GEV fitting.

In the present section, we gave a brief heuristic introduction to the POT. The practical use of the GPD in extreme value modelling is best to be learnt from the fundamental papers by Smith [11], Davison [1], Davison and Smith [2], North [6] and the references therein. Falk [3] uses the POT method for estimating $\xi$. Its theoretical foundation was already laid by Pickands [7] and developed further for instance by Smith [10] and Leadbetter [5].

The POT model is usually formulated as follows:

1. the excesses of an iid (or stationary) sequence over a high threshold $u$ occur at the times of
Figure 4.62  GPD-model based estimates of the 0.999-quantile (top) and 0.9999-quantile (bottom) for the Danish fire insurance data. WARNING: for the interpretation of these plots, read "Mission improbable" on p. 132.

Figure 4.63  GPD-model based estimates of the 0.999-quantile (top) and 0.9999-quantile (bottom) for the industrial fire data. WARNING: for the interpretation of these plots, read "Mission improbable" on p. 132.
a Poisson processes;

2. the corresponding excesses over \( u \) are independent and have a GPD;

3. excesses and exceedance times are independent of each other.

Here one basically looks at a space–time problem: excess sizes and exceedance times. Therefore it is natural to model this problem in a two–dimensional point process or in a marked point process setting: see Falk et al. [4], Section 2.3 and Section 10.3, and Leadbetter [5] for the necessary theoretical background. There also the stationary non–iid case is treated. Using these tools one can justify the above assumptions on the excesses and exceedance times in an asymptotic sense.

The POT method allows for fitting GPD models with time–dependent parameters \( \xi(t) \), \( \beta(t) \) and \( \nu(t) \), in particular one may include non–stationarity effects (trends, seasonality) into the model; see for instance Smith [11]. These are further attractive aspects of GPD fitting.

**Example 4.64** (Diagnostic tools for checking the assumptions of the POT method)

In Figures 4.65 and 4.66 we consider some diagnostic tools [suggested by Smith and Shively [12]] for checking the Poisson process assumption for the exceedance times in the POT model. Figure 4.65 (top) shows the excesses over \( u = 10 \) by the Danish fire insurance claims; see Figure 4.18. The left middle figure is a plot of the first sample autocorrelations of the excesses, indicating that the data are independent. In the right middle figure the corresponding inter–arrival times of the exceedances appear. If these times came from a homogeneous Poisson process they should be iid exponential; see p. 11. The (smoothed) curve in the figure indicates possible deviations from the iid assumptions; it is basically a smoothed mean value of the data and estimates the reciprocal of the intensity of the Poisson process. The curve is almost a straight line, parallel to the horizontal axis. In the left bottom figure a QQ–plot of the inter–arrival times versus exponential quantiles is given. The exponential fit is quite convincing. The sample autocorrelations of the inter–arrival times (bottom, right) indicate that they are indeed uncorrelated.

The picture changes for the absolute values of the negative log–returns of the BMW share prices (see Figure 4.17). In Figure 4.66 the excesses over \( u = 0.0344 \) are given. Their sample autocorrelations are close to zero (middle, left). However, the inter–arrival times of the exceedance times (middle, right) have a tendency to form clusters of large/small values. This does not seem to influence the constancy of their smoothed mean value curve. The corresponding QQ–plot (bottom, left) against exponential quantiles shows a clear deviation from a straight line. The sample autocorrelation at the first lag indicates dependence of inter–arrival times (bottom, right). A similar picture emerges for the positive log–returns.

Concerning software: most of the analyses done in this and other sections may be performed in any statistical software environment; pick your favourite. We have mainly used **S–Plus**. An introduction to the latter package is for instance to be found in Spector [13]. Venables and Ripley [15] give a nice introduction to modern applied statistics with S–Plus. The programs used for the analyses of the present chapter were written by Alexander McNeil and can be obtained from http://www.math.ethz.ch/~mcneil/software.html. Further S–Plus programs providing confidence intervals for parameters in the GPD have been made available by Nader Tajvidi under http://www.maths.lth.se/matstat/staff/nader. Various customised packages for extreme value fitting exist. Examples are **XTREMES**, which comes as part of Falk et al. [4]. In the context of risk management, **RiskMetrics** [9] forms an interesting software environment in which various of the techniques discussed so far, especially concerning quantile (VaR) estimation, are to be found.
Figure 4.65 Top: the excesses over $u = 10$ of the Danish fire insurance data; see Figure 4.18. Middle, left: the sample autocorrelations of the excesses. Middle, right: the inter-arrival times of the exceedances and smoothed mean values curve. Bottom, left: QQ-plot of the the inter-arrival times against exponential quantiles. Bottom, right: sample autocorrelations of these times. See Example 4.64 for further comments.
Figure 4.66 Top: the excesses over $\mu = 0.0344$ of the absolute values of the negative BMW log-returns; see Figure 4.17. Middle, left: the sample autocorrelations of the excesses. Middle, right: the inter-arrival times of the exceedances and smoothed mean values curve. Bottom, left: QQ-plot of the inter-arrival times against exponential quantiles. Bottom, right: sample autocorrelations of these times. See Example 4.64 for further comments.
5 Dependence and extremes

5.1 The extremes of a stationary sequence

One of the natural generalisations of an iid sequence is a strictly stationary process: we say that
the sequence of random variables \((X_n)\) is *strictly stationary* if its finite-dimensional distributions
are invariant under shifts of time, i.e.

\[
(X_{t_1}, \ldots, X_{t_m}) \overset{d}{=} (X_{t_1+h}, \ldots, X_{t_m+h})
\]

for any choice of indices \(t_1 < \cdots < t_m\) and integers \(h\). It is common to define \((X_n)\) with index set \(\mathbb{Z}\).

We can think of \((X_n)\) as a time series of observations at discrete equidistant instants of time where
the distribution of a block \((X_t, X_{t+1}, \ldots, X_{t+h})\) of length \(h\) is the same for all integers \(t\).

References


    2013–2035.


    Letters 12, 357–362.

    649–655.

    119–131.

    18, 21–25.


    (Ed.) Statistical Extremes and Applications. Proceedings of the International Conference at

    New York.

5 Dependence and extremes

5.1 The extremes of a stationary sequence

One of the natural generalisations of an iid sequence is a strictly stationary process: we say that
the sequence of random variables \((X_n)\) is *strictly stationary* if its finite-dimensional distributions
are invariant under shifts of time, i.e.

\[
(X_{t_1}, \ldots, X_{t_m}) \overset{d}{=} (X_{t_1+h}, \ldots, X_{t_m+h})
\]

for any choice of indices \(t_1 < \cdots < t_m\) and integers \(h\). It is common to define \((X_n)\) with index set \(\mathbb{Z}\).

We can think of \((X_n)\) as a time series of observations at discrete equidistant instants of time where
the distribution of a block \((X_t, X_{t+1}, \ldots, X_{t+h})\) of length \(h\) is the same for all integers \(t\).
For simplicity we use throughout the notion of a “stationary” sequence for a “strictly stationary” one. A strictly stationary sequence is naturally also stationary in the wide sense or second order stationary provided the second moment of \( X = X_0 \) is finite, i.e. \( EX_n = EX \) for all \( n \) and \( \text{cov}(X_n, X_m) = \text{cov}(X_0, X_{i_{n-m}}) \) for all \( n \) and \( m \).

It is impossible to build up a general extreme value theory for the class of all stationary sequences. Indeed, one has to specify the dependence structure of \((X_n)\). For example, assume \( X_n = X \) for all \( n \). This relation defines a stationary sequence and

\[
P(M_n \leq x) = P(X \leq x) = F(x), \quad x \in \mathbb{R}.
\]

Thus the distribution of the sample maxima can be any distribution \( F \). This is not a reasonable basis for a general theory.

The other extreme of a stationary sequence occurs when the \( X_n \) are mutually independent, i.e. \((X_n)\) is an iid sequence. In particular, we know that there exist only three types of different limit laws: the Fréchet distribution \( \Phi_\alpha \), the Weibull distribution \( \Psi_\alpha \) and the Gumbel distribution \( \Lambda \) (Fisher–Tippett Theorem 3.2.7). The distributions of the type of \( \Phi_\alpha \), \( \Psi_\alpha \), \( \Lambda \) are called extreme value distributions. In this section we give conditions on the stationary sequence \((X_n)\) which ensure that its sample maxima \((M_n)\) and the corresponding maxima \((\bar{M}_n)\) of an iid sequence \((\bar{X}_n)\) with common distribution function \( F(x) = P(\bar{X} \leq x) \) exhibit a similar limit behaviour. We call \((\bar{X}_n)\) an iid sequence associated with \((X_n)\) or simply an associated iid sequence. As before we write \( F \in \text{MDA}(H) \) for any of the extreme value distributions \( H \) if there exist constants \( c_n > 0 \) and \( d_n \in \mathbb{R} \) such that \( c_n^{-1}(\bar{M}_n - d_n) \overset{d}{\to} H \). For the derivation of the limit probability of \( P(\bar{M}_n \leq u_n) \) for a sequence of thresholds \((u_n)\) we made heavy use of the following factorisation property:

\[
(5.1) \quad P(\bar{M}_n \leq u_n) = P^n(\bar{X} \leq u_n) = \exp\left\{ n \ln \left( 1 - P(\bar{X} > u_n) \right) \right\} \approx \exp\left\{ -n \mathcal{F}(u_n) \right\}.
\]

In particular, we concluded in Proposition 3.1 that, for any \( \tau \in [0, \infty] \), \( P(\bar{M}_n \leq u_n) \to \exp\{-\tau\} \) if and only if \( n \mathcal{F}(u_n) \to \tau \in [0, \infty] \). It is clear that we cannot directly apply (5.1) to maxima of a dependent stationary sequence. However, to overcome this problem we assume that there is a specific type of asymptotic independence:

**Condition D\((u_n)\):** For any integers \( p, q \) and \( n \)

\[
1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_q \leq n
\]

such that \( j_1 - i_p \geq l \) we have

\[
\left| P\left( \max_{i \in A_1} X_i \leq u_n \right) - P\left( \max X_i \leq u_n \right) P\left( \max X_i \leq u_n \right) \right| \leq \alpha_{n,l},
\]

where \( A_1 = \{ i_1, \ldots, i_p \} \), \( A_2 = \{ j_1, \ldots, j_q \} \) and \( \alpha_{n,l} \to 0 \) as \( n \to \infty \) for some sequence \( l = l_n = o(n) \).

This condition as well as \( D'(u_n) \) below and their modifications have been intensively applied to stationary sequences in the monograph by Leadbetter, Lindgren and Rootzén [2]. Condition \( D(u_n) \)
is a distributional mixing condition, weaker than most of the classical forms of dependence restrictions. A discussion of the role of $D(u_n)$ as a specific mixing condition can be found in Leadbetter et al. [2]. Condition $D(u_n)$ implies, for example, that

\begin{equation}
(5.2) \quad P(M_n \leq u_n) = P^k(M_{[n/k]} \leq u_n) + o(1)
\end{equation}

for constant or slowly increasing $k$. This relation already indicates that the limit behaviour of $(M_n)$ and its associated sequence $(\tilde{M}_n)$ must be closely related. The following result (Theorem 3.3.3 in Leadbetter et al. [2]) even shows that the classes of possible limit laws for the normalised and centred sequences $(M_n)$ and $(\tilde{M}_n)$ coincide.

**Theorem 5.1** (Limit laws for maxima of a stationary sequence)

Suppose $c_n^{-1}(M_n - d_n) \xrightarrow{d} G$ for some distribution $G$ and appropriate constants $c_n > 0, d_n \in \mathbb{R}$. If the condition $D(c_n x + d_n)$ holds for all real $x$, then $G$ is an extreme value distribution.

**Proof.** Recall from Theorems 3.8, 3.9 and from Definition 3.12 that $G$ is an extreme value distribution if and only if $G$ is max–stable. By (5.2),

\[ P(M_{nk} \leq c_n x + d_n) = P^k(M_n \leq c_n x + d_n) + o(1) \rightarrow G^k(x) \]

for every integer $k \geq 1$, and every continuity point $x$ of $G$. On the other hand,

\[ P(M_{nk} \leq c_{nk} x + d_{nk}) \rightarrow G(x) . \]

Now we may proceed as in the proof of Theorem 3.8 to conclude that $G$ is max–stable. \hfill \Box

**Remark.** 1) Theorem 5.1 does not mean that the relations $c_n^{-1}(M_n - d_n) \xrightarrow{d} G$ and $c_n^{-1}(\tilde{M}_n - d_n) \xrightarrow{d} H$ hold with $G = H$. We will see later that $G$ is often of the form $H^\theta$ for some $\theta \in [0, 1]$ (see for instance Example 5.2); $\theta$ is then called extremal index. \hfill \Box

Thus max–stability of the limit distribution is necessary under the conditions $D(c_n x + d_n), x \in \mathbb{R}$. Next we want to find sufficient conditions for convergence of the probabilities $P(M_n \leq u_n)$ for a given threshold sequence $(u_n)$ satisfying

\begin{equation}
(5.3) \quad n \overline{F}(u_n) \rightarrow \tau
\end{equation}

for some $\tau \in [0, \infty)$. From Proposition 3.1 we know that (5.3) and $P(\tilde{M}_n \leq u_n) \rightarrow \exp\{-\tau\}$ are equivalent. But may we replace $(\tilde{M}_n)$ by $(M_n)$ under $D(u_n)$? The answer is, unfortunately, NO. All one can derive is

\[ \liminf_{n \rightarrow \infty} P(M_n \leq u_n) \geq e^{-\tau} ; \]

see the proof of Proposition 5.3 below.

**Example 5.2** (See also Figure 5.5.) Assume that $(Y_n)$ is a sequence of iid random variables with distribution function $\sqrt{F}$ for some distribution function $F$. Define the sequence $(X_n)$ by

\[ X_n = \max(Y_n, Y_{n+1}) , \quad n \in \mathbb{N} . \]

Then $(X_n)$ is a stationary sequence and $X_n$ has distribution function $F$ for all $n \geq 1$. From this construction it is clear that maxima of $(X_n)$ appear as pairs at consecutive indices.
Now assume that for $\tau \in (0, \infty)$ the sequence $u_n$ satisfies $u_n \uparrow x_F$ ($x_F$ is the right endpoint of $F$) and (5.3). Then $F(u_n) \to 1$ and

$$nP(Y_1 > u_n) = n \left(1 - \sqrt{F(u_n)}\right) = \frac{nF(u_n)}{1 + \sqrt{F(u_n)}} \to \frac{\tau}{2}.$$ 

Hence, by Proposition 3.1,

$$P(M_n \leq u_n) = P(\max(Y_1, \ldots, Y_n, Y_{n+1}) \leq u_n) = P(\max(Y_1, \ldots, Y_n) \leq u_n)F(u_n) \to e^{-\tau/2}.$$ 

Condition $D(u_n)$ is naturally satisfied: if $A_1$ and $A_2$ are chosen as in $D(u_n)$ and $l \geq 2$, then we can take $a_{n,l} = 0$. \hfill \Box

This example supports the introduction of a second technical condition.

**Condition** $D'(u_n)$: The relation

$$\lim_{k \to \infty} \lim_{n \to \infty} \sup_{j=2}^{[n/k]} P(X_1 > u_n, X_j > u_n) = 0.$$ 

**Remark.** 2) $D'(u_n)$ is an “anti-clustering condition” on the stationary sequence $(X_n)$. Indeed, notice that $D'(u_n)$ implies

$$E\sum_{1 \leq i < j \leq [n/k]} I_{\{X_i > u_n, X_j > u_n\}} \leq [n/k] \sum_{j=2}^{[n/k]} EI_{\{X_1 > u_n, X_j > u_n\}} \to 0,$$

so that, in the mean, joint exceedances of $u_n$ by pairs $(X_i, X_j)$ become very unlikely for large $n$. \hfill \Box

Now we have introduced the conditions which are needed to formulate the following analogue of Proposition 3.1; see Theorem 3.4.1 in Leadbetter et al. [2]:

**Proposition 5.3** (Limit probabilities for sample maxima)

Assume that the stationary sequence $(X_n)$ and the threshold sequence $(u_n)$ satisfy $D(u_n), D'(u_n)$. Suppose $\tau \in [0, \infty)$. Then condition (5.3) holds if and only if

$$\lim_{n \to \infty} P(M_n \leq u_n) = e^{-\tau}.$$ 

**Proof.** We restrict ourselves to the sufficiency part in order to illustrate the use of the conditions $D(u_n)$ and $D'(u_n)$. The necessity follows by similar arguments.

We have, for any $l \geq 1$,

$$\sum_{i=1}^{l} P(X_i > u_n) - \sum_{1 \leq i < j \leq l} P(X_i > u_n, X_j > u_n) \leq P(M_l > u_n) \leq \sum_{i=1}^{l} P(X_i > u_n).$$

140
Exploiting the stationarity of \( (X_n) \) we see that
\[
\sum_{i=1}^{l} P(X_i > u_n) = l \overline{F}(u_n),
\]
\[
\sum_{1 \leq i < j \leq l} P(X_i > u_n, X_j > u_n) \leq l \sum_{j=2}^{l} P(X_1 > u_n, X_j > u_n).
\]
Combining this and (5.5) for \( l = \lfloor n/k \rfloor \) ([x] denotes the integer part of x) and for a fixed k, we derive upper and lower estimates for \( P(M_{\lfloor n/k \rfloor} \leq u_n) \):
\[
1 - \frac{n}{k} \overline{F}(u_n) \leq P(M_{\lfloor n/k \rfloor} \leq u_n) \leq 1 - \frac{n}{k} \overline{F}(u_n) + \frac{n}{k} \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n).
\]
From (5.3) we immediately have
\[
\frac{n}{k} \overline{F}(u_n) \to \frac{\tau}{k}, \quad n \to \infty,
\]
and, by condition \( D'(u_n) \),
\[
\limsup_{n \to \infty} \frac{n}{k} \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n) = o(1/k), \quad k \to \infty.
\]
Thus we get the bounds
\[
1 - \frac{\tau}{k} \leq \liminf_{n \to \infty} P(M_{\lfloor n/k \rfloor} \leq u_n) \leq \limsup_{n \to \infty} P(M_{\lfloor n/k \rfloor} \leq u_n) \leq 1 - \frac{\tau}{k} + o(1/k).
\]
This and relation (5.2) imply that
\[
1 - \frac{\tau}{k} \leq \liminf_{n \to \infty} P(M_n \leq u_n)
\]
\[
\leq \limsup_{n \to \infty} P(M_n \leq u_n) \leq \left(1 - \frac{\tau}{k} + o(1/k)\right)^k.
\]
Letting \( k \to \infty \) we see that
\[
\lim_{n \to \infty} P(M_n \leq u_n) = e^{-\tau}.
\]
This concludes the proof. \( \square \)

**Example 5.4** (Continuation of Example 5.2)

We observed in Example 5.2 that condition (5.3) implies \( P(M_n \leq u_n) \to \exp\{ -\tau/2 \} \). We have already checked that \( D'(u_n) \) is satisfied. Thus \( D'(u_n) \) must go wrong. This can be easily seen: since \( X_1 \) and \( X_j \) are independent for \( j \geq 2 \) we conclude that
\[
n \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n)
\]
\[
= nP(X_1 > u_n, X_2 > u_n) + n(\lfloor n/k \rfloor - 2)P(X_1 > u_n) + n(\max(Y_1, Y_2, Y_3) > u_n) + \tau^2/k + o(1)
\]
\[
= n(1 - F^3(u_n)) + \tau^2/k + o(1), \quad n \to \infty.
\]
Figure 5.5 A realisation of the sequences \((Y_n)\) (top) and \((X_n)\) (bottom) with \(F\) standard exponential as discussed in Examples 5.2 and 5.4. Extremes appear in clusters of size 2.

We have
\[
n \left( 1 - F^3 (u_n) \right) = n \overline{F} (u_n) (1 + F^2 (u_n) + F(u_n)) \to 3\tau.
\]
Thus condition \(D'(u_n)\) cannot be satisfied. The reason for this is that maxima in \((X_n)\) appear in clusters of size 2. Notice that
\[
E \left( \sum_{i=1}^{n} I_{\{X_i > u_n, X_{i+1} > u_n\}} \right) = n \tau \left( X_1 > u_n, X_2 > u_n \right) \to 3\tau > 0,
\]
so that in the long run the expected number of joint exceedances of \(u_n\) by the pairs \((X_i, X_{i+1})\) stabilises around a positive number.

Proceeding precisely as in Section 3.3 we can now derive the limit distribution for the maxima \(M_n\):

**Theorem 5.6** (Limit distribution of maxima of a stationary sequence)

Let \((X_n)\) be a stationary sequence with common distribution function \(F \in \text{MDA}(H)\) for some extreme value distribution \(H\), i.e. there exist constants \(c_n > 0, d_n \in \mathbb{R}\) such that

\[
(5.6) \quad \lim_{n \to \infty} n \overline{F} (c_n x + d_n) = - \ln H(x), \quad x \in \mathbb{R}.
\]

Assume that for \(x \in \mathbb{R}\) the sequences \((u_n) = (c_n x + d_n)\) satisfy the conditions \(D(u_n)\) and \(D'(u_n)\). Then (5.6) is equivalent to each of the following relations:

\[
(5.7) \quad c_n^{-1} (M_n - d_n) \xrightarrow{d} H,
\]
\[
(5.8) \quad c_n^{-1} (\bar{M}_n - d_n) \xrightarrow{d} H.
\]

**Proof.** The equivalence of (5.6) and (5.8) is immediate from Proposition 3.19. The equivalence of (5.6) and (5.7) follows from Proposition 5.3. \(\square\)
From the discussion above we are not surprised about the same limit behaviour of the maxima of a stationary sequence and its associated iid sequence; the conditions $D(c_n x + d_n)$ and $D'(c_n x + d_n)$ force the sequence $(M_n)$ to behave very much like the maxima of an iid sequence. Notice that Theorem 5.6 also ensures that we can choose the sequences $(c_n)$ and $(d_n)$ in the same way as proposed in Section 3.3.

Thus the problem about the maxima of a stationary sequence has been reduced to a question about the extremes of iid random variables. However, now one has to verify the conditions $D(c_n x + d_n)$ and $D'(c_n x + d_n)$ which, in general, is tedious. Conditions $D(u_n)$ and $D'(u_n)$ have been discussed in detail in the monograph by Leadbetter et al. [2]. The case of a Gaussian stationary sequence is particularly nice: one can check $D(u_n)$ and $D'(u_n)$ via the asymptotic behaviour of the autocovariances

$$
\gamma(h) = \text{cov}(X_0, X_h), \quad h \geq 0.
$$

The basic idea is that the distributions of two Gaussian vectors are “close” to each other if their covariance matrices are “close”. Leadbetter et al. [2] make this concept precise by a so-called normal comparison lemma (their Theorem 4.2.1), a particular consequence of which is the estimate

$$
\left| P(X_{i_1} \leq u_n, \ldots, X_{i_k} \leq u_n) - \Phi^k(u_n) \right| 
\leq \text{const} \ n \sum_{h=1}^{k} |\gamma(h)| \exp \left( \frac{-u_n^2}{1 + |\gamma(h)|} \right)
$$

for $1 \leq i_1 < \cdots < i_k \leq n$. Here $(X_n)$ is stationary with marginal distribution function the standard normal $\Phi$, and it is assumed that $\sup_{h \geq 1} |\gamma(h)| < 1$. In particular,

$$
(5.9) \quad \left| P(M_n \leq u_n) - \Phi(u_n) \right| \leq \text{const} \ n \sum_{h=1}^{n} |\gamma(h)| \exp \left( \frac{-u_n^2}{1 + |\gamma(h)|} \right).
$$

Now it is not difficult to check conditions $D(u_n)$ and $D'(u_n)$. For details see Lemma 4.4.1 in Leadbetter et al. [2].

**Lemma 5.7** (Conditions for $D(u_n)$ and $D'(u_n)$ for a Gaussian stationary sequence)

Assume $(X_n)$ is stationary Gaussian and let $(u_n)$ be a sequence of real numbers.

1. Suppose the right-hand side in (5.9) tends to zero as $n \to \infty$ and $\sup_{h \geq 1} |\gamma(h)| < 1$. Then $D(u_n)$ holds.

2. If in addition $\limsup_{n \to \infty} n \Phi(u_n) < \infty$ then $D'(u_n)$ holds.

3. If $\gamma(n) \ln n \to 0$ and $\limsup_{n \to \infty} n \Phi(u_n) < \infty$ then both conditions $D(u_n)$ and $D'(u_n)$ are satisfied. $\square$

Now recall that the normal distribution $\Phi$ is in the maximum domain of attraction of the Gumbel law $\Lambda$; see Example 3.43. Then the following is a consequence of Lemma 5.7 and of Theorem 5.6. The constants $c_n$ and $d_n$ are chosen as in Example 3.43.

**Theorem 5.8** (Limit distribution of the maxima of a Gaussian stationary sequence)

Let $(X_n)$ be a stationary sequence with common standard normal distribution function $\Phi$. Suppose that

$$
\lim_{n \to \infty} \gamma(n) \ln n = 0.
$$

143
Then
\[
\sqrt{2 \ln n} \left( M_n - \sqrt{2 \ln n} + \frac{\ln \ln n + \ln 4\pi}{2(2 \ln n)^{1/2}} \right) \overset{d}{\to} \Lambda.
\]

The assumption $\gamma(n) \ln n \to 0$ is called Berman’s condition and is very weak. Thus Theorem 5.8 states that Gaussian stationary sequences have very much the same extremal behaviour as Gaussian iid sequences.

**Example 5.9** (Gaussian linear processes)

An important class of stationary sequences is that of the linear processes, which have an infinite moving average representation

\[(5.10)\]
\[X_n = \sum_{j=-\infty}^{\infty} \psi_j Z_{n-j}, \quad n \in \mathbb{Z},\]

where $(Z_n)_{n \in \mathbb{Z}}$ is an iid sequence and $\sum_{j} |\psi_j|^2 < \infty$. We also suppose that $EZ_1 = 0$ and $\sigma_Z^2 = \text{var}(Z_1) < \infty$. If $(Z_n)$ is Gaussian, so is $(X_n)$. Conversely, most interesting Gaussian stationary processes have representation (5.10); see Brockwell and Davis [1], Theorem 5.7.1, in particular, the popular (causal) ARMA processes. In that case the coefficients $\psi_j$ decrease to zero at an exponential rate. Hence the autocovariances of $(X_n)$, i.e.

\[\gamma(h) = E(X_0X_h) = \sigma_Z^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}, \quad h \geq 0,\]

decrease to zero exponentially as $h \to \infty$. Thus Theorem 5.8 is applicable to Gaussian ARMA processes.

Gaussian fractional ARIMA($p,d,q$) processes with $p,q \geq 1, d \in (0,1)$, enjoy a (causal) representation (5.10) with $\psi_j = j^{d-1}L(j)$ for a slowly varying function $L$; see Brockwell and Davis [1], Section 13.2. It is not difficult to see that the assumptions of Theorem 5.8 also hold in this case. Fractional ARIMA processes with $d \in (0,0.5)$ are a standard example of long memory processes where the sequence $\gamma(h)$ is not supposed to be absolutely summable. This shows that the restriction $\gamma(n) \ln n \to 0$ is indeed very weak in the Gaussian case.

In general, the extreme value behaviour of linear processes with subexponential noise $(Z_n)$ in MDA($\Lambda$) or MDA($\Phi_n$) varies a lot. One can show for subexponential $F$ that the limit distributions of $(M_n)$ are of the form $H^\theta$ for some $\theta \in (0,1]$ and an extreme value distribution $H$. This indicates the various forms of limit behaviour of maxima of linear processes, depending on their tail behaviour.

**Comments**

Extreme value theory for stationary sequences has been treated in detail in Leadbetter et al. [2]. There one can also find some remarks on the history of the use of conditions $D(u_n)$ and $D'(u_n)$.

A very recommendable review article is Leadbetter and Rootzén [3].

In summary, the conditions $D(u_n)$ and $D'(u_n)$ ensure that the extremes of the stationary sequence $(X_n)$ have the same qualitative behaviour as the extremes of an associated iid sequence. The main problem is to verify conditions $D(u_n)$ and $D'(u_n)$. For Gaussian $(X_n)$ this reduces to showing Berman’s condition, namely that $\gamma(n) = \text{cov}(X_0, X_n) = o(1/\ln n)$. It covers wide classes of Gaussian sequences, in particular ARMA and fractional ARIMA processes. We mention that Leadbetter et al. [2] also treated the cases $\gamma(n) \ln n \to c \in (0, \infty)$.
References


5.2 The extremal index

5.2.1 Definition and elementary properties

So far we presented a wealth of material on extremes. In most cases we restricted ourselves to iid observations. However, in reality extremal events often tend to occur in clusters caused by local dependence in the data. For instance, large claims in insurance are mainly due to hurricanes, storms, floods, earthquakes etc. Claims are then linked with these events and do not occur independently. The same can be observed with financial data such as exchange rates and asset prices. If one large value in such a time series occurs we can usually observe a cluster of large values over a short period afterwards.

The extremal index is a quantity which, in an intuitive way, allows one to characterise the relationship between the dependence structure of the data and their extremal behaviour. To understand this notion we first recall some of the examples of extremal behaviour for a strictly stationary sequence \((X_n)\) with marginal distribution \(F\). In this section we consider only this kind of model. As usual, \(M_n\) stands for the maximum of the sample \(X_1, \ldots, X_n\), \((X_n)\) is an associated iid sequence (i.e. with common distribution function \(F\)) and \((M_n)\) denotes the corresponding sequence of maxima.

**Example 5.10** Assume that the condition

\[
(5.11) \quad n F(u_n) \to \tau \in (0, \infty)
\]

holds for some non-decreasing sequence \((u_n)\).

(a) For an iid sequence \((X_n)\) we know from Proposition 3.1 that (5.11) is equivalent to the relation

\[
(5.12) \quad \lim_{n \to \infty} P(M_n \leq u_n) = e^{-\tau}.
\]

(b) Recall the conditions \(D(u_n)\) and \(D'(u_n)\) from Section 5.1. They ensure that the strictly stationary sequence \((X_n)\) has the same asymptotic extremal behaviour as an associated iid sequence. In particular, (5.11) implies (5.12) (see Proposition 5.3). Conditions \(D(u_n)\) and \(D'(u_n)\) are satisfied for large classes of Gaussian stationary sequences, including many Gaussian linear processes (for instance Gaussian ARMA and fractional ARIMA processes).

(c) Recall the situation of Example 5.2: starting with an iid sequence \((Y_n)\) with distribution function \(\sqrt{F}\), the strictly stationary sequence

\[
X_n = \max(Y_n, Y_{n+1}) , \quad n \in \mathbb{N},
\]

has distribution function \(F\) and

\[
M_n = \max(Y_1, \ldots, Y_{n+1}) , \quad n \in \mathbb{N}.
\]

145
If (5.11) is satisfied then
\[ \lim_{n \to \infty} P(M_n \leq u_n) = e^{-\tau/2}. \]

We know from Example 5.2 that condition \( D(u_n) \) is satisfied in this case, but \( D'(u_n) \) is not; see Example 5.4.

(d) Let \( (X_n) \) be a linear process
\[ X_n = \sum_{j=-\infty}^{\infty} \psi_j Z_{n-j}, \quad n \in \mathbb{Z}, \]
driven by iid noise \( (Z_t) \). Assume \( P(Z_t > x) = x^{-\alpha}L(x) \) as \( x \to \infty \) and that the tail balance condition
\[(5.13) \quad P(Z_t > x) \sim pP(\|Z_t\| > x) \quad \text{and} \quad P(Z_t \leq -x) \sim qP(\|Z_t\| > x)\]
holds for some \( p \in (0, 1), q = 1 - p. \) This implies that \( F \in \text{MDA}(\Phi_\alpha) \) for some \( \alpha > 0. \) Then we obtain from Corollary 5.5.3 in Embrechts et al. [1] that there exist constants \( c_n > 0 \) such that \( u_n = u_n(x) = c_n x \) satisfies (5.11) for \( \tau = \tau(x) = x^{-\alpha}, x > 0, \) and
\[
\lim_{n \to \infty} P(M_n \leq u_n(x)) = \Phi_\alpha(x), \quad x \in \mathbb{R}_+, \quad \text{and} \quad \lim_{n \to \infty} P(M_n \leq u_n(x)) = \Phi_\alpha^\theta(x), \quad x \in \mathbb{R}_+ .
\]
Here \( \Phi_\alpha \) denotes the standard Fréchet distribution and
\[
\theta = \left( \psi_+^\alpha p + \psi_-^\alpha q \right)/\|\psi\|_\alpha^n,
\]
where
\[
\psi_+ = \max_j (\psi_j \vee 0) \quad \text{and} \quad \psi_- = \max_j ((-\psi_j) \vee 0),
\]
\[
\|\psi\|_\alpha^n = \sum_{j=-\infty}^{\infty} |\psi_j|^\alpha (p I_{[\psi_j > 0]} + q I_{[\psi_j < 0]}).
\]

(e) Let \( (X_n) \) be a linear process driven by iid subexponential noise \( (Z_t) \) with \( F \in \text{MDA}(\Lambda) \), where \( \Lambda \) denotes the standard Gumbel distribution. We also assume the tail balance condition (5.13) and \( \max_j |\psi_j| = 1. \) Then we know from Corollary 5.5.12 in Embrechts et al. [1] that there exist constants \( c_n > 0 \) and \( d_n \in \mathbb{R} \) such that \( u_n = u_n(x) = c_n x + d_n \) satisfies (5.11) for \( \tau = \tau(x) = \exp\{t\}, x \in \mathbb{R}, \) and
\[
\lim_{n \to \infty} P(M_n \leq u_n(x)) = \Lambda(x), \quad x \in \mathbb{R}, \quad \text{and} \quad \lim_{n \to \infty} P(M_n \leq u_n(x)) = \Lambda^\theta(x), \quad x \in \mathbb{R},
\]
where
\[
\theta = (k_+^p + k_-^q)^{-1} ,
\]
and
\[
k_+ = \text{card}\{j : \psi_j = 1\} \quad \text{and} \quad k_- = \text{card}\{j : \psi_j = -1\}.
\]
\[
\Box
\]

146
The examples above follow similar patterns. Indeed, it is typical for stationary \((X_n)\) and \((u_n)\) satisfying (5.11) that \(P(M_n \leq u_n) \to \exp(-\theta \tau)\) for some \(\theta \in (0, 1]\), whereas \(P(M_n \leq u_n) \to \exp(-\tau)\). The latter fact indicates that exceedances of high threshold values \(u_n\) tend to occur in clusters for dependent data. This is something we might have expected when looking at real data-sets.

The above examples suggest the following definition which allows us to distinguish between the extremal behaviour of different dependence structures.

**Definition 5.11** (Extremal index)

Let \((X_n)\) be a strictly stationary sequence and \(\theta\) a non-negative number. Assume that for every \(\tau > 0\) there exists a sequence \((u_n)\) such that

\[
\lim_{n \to \infty} n \overline{F}(u_n) = \tau, \tag{5.14}
\]

\[
\lim_{n \to \infty} P(M_n \leq u_n) = e^{-\theta \tau}. \tag{5.15}
\]

Then \(\theta\) is called the extremal index of the sequence \((X_n)\). \(\square\)

**Remarks.** 1) The definition of the extremal index can be shown to be independent of the particular sequence \((u_n)\). More precisely, if \((X_n)\) has extremal index \(\theta > 0\) then, for any sequence of real numbers \((u_n)\) and \(\tau \in [0, \infty]\), the relations (5.14), (5.15) and \(P(M_n \leq u_n) \to \exp(-\tau)\) are equivalent; see Leadbetter [3]. A particular consequence is the following: if \(F \in \text{MDA}(H)\) for some extreme value distribution \(H\) then

\[
C^{-1}_n(M_n - d_n) \overset{d}{\to} H \quad \Leftrightarrow \quad C^{-1}_n(M_n - d_n - d) \overset{d}{\to} H^\theta
\]

for appropriate norming constants \(c_n > 0\) and \(d_n \in \mathbb{R}\).

2) Since an extreme value distribution \(H\) is max-stable (see Definition 3.7), \(H^\theta\) is of the same type as \(H\), i.e. there exist constants \(c > 0, d \in \mathbb{R}\) such that \(H^\theta(x) = H(cx + d)\). This also implies that the limits in (5.16) can be chosen to be identical after a simple change of the norming constants. \(\square\)

**Example 5.12** (Continuation of Example 5.10)

From the discussion in Example 5.10 it is immediate that the cases (a) and (b) (iid and weakly dependent stationary sequences) yield the extremal index \(\theta = 1\). In the case (c), \(\theta = 0.5\) (this type of example can naturally be extended for constructing stationary sequences with extremal index \(\theta = 1/k\) for any integer \(k \geq 1\)). The examples (d)-(e) (linear processes) show that we can get any number \(\theta \in (0, 1]\) as extremal index. \(\square\)

From these examples two natural questions arise:

*How can we interpret the extremal index \(\theta\)?*

and

*What is the range of the extremal index?*

Section 5.2.2 is devoted to the first problem. The second one has a simple solution:

\(\theta\) always belongs to the interval \([0, 1]\).
From Example 5.12 we already know that any number \( \theta \in (0, 1] \) can be an extremal index. The case \( \theta = 0 \) is somewhat pathological. We refer to Leadbetter, Lindgren and Rootzén [5] and Leadbetter [4] for some examples. The cases \( \theta > 0 \) are of particular practical interest. It remains to show that \( \theta > 1 \) is impossible, but this follows from the following easy argument:

\[
P\left(M_n \leq u_n \right) = 1 - P \left( \bigcup_{i=1}^{n} \{ X_i > u_n \} \right) \geq 1 - n \bar{F} \left( u_n \right).
\]

By definition of the extremal index, the left-hand side converges to \( e^{-\theta \tau} \) whereas the right-hand side has limit \( 1 - \tau \). Hence \( e^{-\theta \tau} \geq 1 - \tau \) for all \( \tau > 0 \) which is possible only if \( \theta \leq 1 \).

Next we ask:

**Does every strictly stationary sequence have an extremal index?**

Life would be easy if this were true. Indeed, extreme value theory for stationary sequences could then be derived from the corresponding results for iid sequences. The answer to the above question is (unfortunately) no.

**Example 5.13** Assume \( (X_n) \) is iid with \( F \in \text{MDA}(\Phi_n) \) and norming constants \( c_n > 0 \). Assume \( A \) is a positive random variable independent of \( (X_n) \). Then

\[
P \left( c_n^{-1} \max(A X_1, \ldots, A X_n) \leq x \right)
= P \left( c_n^{-1} M_n \leq A^{-1} x \right)
= E \exp \left\{ -x^{-\alpha} A^\alpha \right\}, \quad x > 0.
\]

It is worthwhile mentioning that, for large classes of stationary sequences \( (X_n) \), there exist real numbers \( 0 \leq \theta' \leq \theta' \leq 1 \) such that

\[
e^{-\theta' \tau} \leq \liminf_{n \to \infty} P \left( M_n \leq u_n \right) \leq \limsup_{n \to \infty} P \left( M_n \leq u_n \right) \leq e^{-\theta \tau}, \quad \tau > 0,
\]

for every sequence \( (u_n) \) satisfying (5.14). A proof of this result under condition \( D(u_n) \) is to be found in Leadbetter, Lindgren and Rootzén [5], Theorem 3.7.1.

### 5.2.2 Interpretation and estimation of the extremal index

We start with a somewhat simplistic example (taken from Weissman [14]) showing the relevance of the notion of the extremal index.

**Example 5.14** Assume a dyke has to be built at the seashore to protect against floods with 95% certainty for the next 100 years. Suppose it has been established that the 99.9 and 99.95 percentiles of the annual wave–height are 10 m and 11 m, respectively. If the annual maxima are believed to be iid, then the dyke should be 11 m high \((0.9995^{100} \approx 0.95)\). But if the annual maxima are stationary with extremal index \( \theta = 0.5 \), then a height of 10 m is sufficient \((0.999^{50} \approx 0.95)\). □

This example brings out already that estimation of the extremal index \( \theta \) must be a central issue in extreme value statistics for dependent data. Estimation of \( \theta \) will be based on a number of different probabilistic interpretations of the extremal index, leading to the construction of different estimators. Throughout we exclude the degenerate case \( \theta = 0 \).
A first (naive) approach to the estimation of $\theta$: the blocks method

Starting from the definition of the extremal index $\theta$, we have

$$P(M_n \leq u_n) \approx P^\theta(\tilde{M}_n \leq u_n) = F^\theta_n(u_n),$$

provided $nF(u_n) \to \tau > 0$. Hence

$$\lim_{n \to \infty} \frac{\ln P(M_n \leq u_n)}{n \ln F(u_n)} = \theta. \quad (5.17)$$

This simple limit relation suggests constructing naive estimators of $\theta$. Since we do not know $F(u_n)$ and $P(M_n \leq u_n)$, these quantities have to be replaced by estimators. An obvious candidate for estimating the tail $F(u_n)$ is its empirical version

$$\frac{N}{n} = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i > u_n]}.$$

This choice is motivated by the Glivenko–Cantelli theorem for stationary ergodic sequences $(X_n)$. To find an empirical estimator for $P(M_n \leq u_n)$ is not straightforward. Recall from Section 5.1 that condition $D(u_n)$ implies

$$P(M_n \leq u_n) \approx P^k(M_{[n/k]} \leq u_n) \quad (5.18)$$

for constant $k$ or slowly increasing $k = k(n)$. The approximation (5.18) forms the basis for the blocks method. For the sake of argument assume that $n = rk$ for integers $r = r(n) \to \infty$ and $k = k(n) \to \infty$. Otherwise, let $r = \lceil n/k \rceil$. This divides the sample $X_1, \ldots, X_n$ into $k$ blocks of size $r$:

$$X_1, \ldots, X_r; \ldots; X_{(k-1)r+1}, \ldots, X_{kr}. \quad (5.19)$$

For each block we calculate the maximum

$$M_r^{(i)} = \max \{ X_{(i-1)r+1}, \ldots, X_{ir} \}, \quad i = 1, \ldots, k.$$

Relation (5.18) then suggests the approximation

$$P(M_n \leq u_n) = P\left( \max_{1 \leq i \leq k} M_r^{(i)} \leq u_n \right) \approx P^k(M_r \leq u_n) \approx \left( \frac{1}{k} \sum_{i=1}^{k} I_{[M_r^{(i)} \leq u_n]} \right)^k = \left( 1 - \frac{K}{k} \right)^k.$$

A combination of these heuristic arguments with (5.17) leads to the following estimator of $\theta$:

$$\hat{\theta}^{(1)}_n = \frac{k}{n} \ln \left( 1 - \frac{K}{k} \right) = \frac{1}{r} \ln \left( 1 - \frac{K}{r(N/n)} \right). \quad (5.20)$$

Here $N$ is the number of exceedances of $u_n$ by $X_1, \ldots, X_n$ and $K$ is the number of blocks with one or more exceedances. A Taylor expansion argument yields a second estimator

$$\hat{\theta}^{(2)}_n = \frac{K}{N} = \frac{1}{r} \frac{K}{N/n} \approx \hat{\theta}^{(1)}_n. \quad (5.21)$$

149
The blocks method accounts for clustering in the data. If the event
\[
\left\{ M_r^{(i)} > u_n \right\} = \bigcup_{j=1}^{r} \{ X_{(i-1)r+j} > u_n \}
\]
happens one says that a cluster occurred in the \(i\)th block. These events characterise the extremal behaviour of \((X_n)\) if we assume that the size \(r(n)\) of the blocks increases slowly with \(n\). This gives us some feeling for the dependence structure in the sequence \((X_n)\). In this sense, the extremal index is a measure of the clustering tendency of high-threshold exceedances in a stationary sequence.

There has been plenty of hand-waving in the course of the derivation of the estimators \(\hat{\theta}_n^{(1)}\). Therefore the following questions naturally arise:

*What are the statistical properties of \(\hat{\theta}_n^{(1)}\) and \(\hat{\theta}_n^{(2)}\) as estimators of \(\theta\)?*

and

*Given that we have \(n\) observations, how do we choose the values of \(r\) (or \(k\)) and \(u_n\)?*

These are important questions. Partial answers are to be found in the literature; see the Comments below. A flavour of what one can expect as an answer is summarised in the following remark from Weissman and Cohen [15]:

*It turns out that it is not easy to obtain accurate estimates of \(\theta\).*

It seems that we are in a similar situation to that in Section 4.3, where we tried to estimate the index \(\xi\) of an extreme value distribution. This should not discourage us from considering some more estimators of \(\theta\), especially as we will meet alternative interpretations of the extremal index along the way.

**The extremal index as reciprocal of the mean cluster size**

This approach is based on results by Hsing, Hüsler and Leadbetter [2], Theorems 4.1 and 4.2. It uses point process theory which would enable us to define what a “cluster” of exceedances is. Since we do not have this tool, we rely on the intuitive meaning of such a cluster. Its size is an integer-valued random variable \(\xi\). Hsing et al. [2] show that \(\theta = (E\xi)^{-1}\), i.e. \(\theta\) can be interpreted as the reciprocal of the mean cluster size.

This interpretation of \(\theta\) suggests an estimator based on the blocks method:

\[
\hat{\theta}_n^{(2)} = \frac{\sum_{i=1}^{k} I_{\{M_r^{(i)} > u_n\}}}{\sum_{i=1}^{n} I_{\{X_i > u_n\}}} = \frac{K}{N},
\]

i.e. number \(K\) of clusters of exceedances divided by the total number \(N\) of exceedances. The same estimator has already been suggested as an approximation to \(\hat{\theta}_n^{(1)}\).

**The extremal index as conditional probability: the runs method**

O’Brien [9] proved, under a weak mixing condition, that the following limit relation holds:

\[
P(M_n \leq u_n) = (F(u_n))^{nP(M_2 \leq u_n | X_1 > u_n)} + o(1)
= \exp \left\{-nP \left( X_1 > u_n, M_2 \leq u_n \right) \right\} + o(1)
\]

150
provided \( n \mathcal{F}(u_n) \to \tau \). Here

\[
M_{2,s} = \max(X_2, \ldots, X_s),
\]

and \( s = s(n) \) satisfies \( s/n \to 0, \ s \to \infty \) and some more specific growth conditions. On the other hand, by definition of the extremal index,

\[
P(M_n \leq u_n) = \exp\{-\theta \tau\} + o(1).
\]

Hence, under the conditions above,

\[
\lim_{n \to \infty} \theta_n(s(n), u_n) = \lim_{n \to \infty} P(M_{2,s} \leq u_n \mid X_1 > u_n) = \theta.
\]

Thus \( \theta \) can be interpreted as a limiting conditional probability. The conditional probability \( \theta_n(s(n), u_n) \) is some measure of the clustering tendency of high threshold exceedances: \( M_{2,s} \) can be less than \( u_n \) only if \( X_1 \) is the last element in a cluster of values which exceed \( u_n \). If large values appear in clusters, there must be longer intervals between different clusters. As a consequence, \( P(M_n \leq u_n) \) will typically be larger than for independent observations.

O’Brien’s result has been used to construct an estimator of \( \theta \) based on runs:

\[
\hat{\theta}_n^{(3)} = \frac{\sum_{i=1}^{n-r} I_{A_{i,n}}}{\sum_{i=1}^{n} I_{\{X_i > u_n\}}} = \frac{\sum_{i=1}^{n-r} I_{A_{i,n}}}{N},
\]

where

\[
A_{i,n} = \{X_i > u_n, X_{i+1} \leq u_n, \ldots, X_{i+r} \leq u_n\}.
\]

This means we take any sequence of \( r = r(n) \) consecutive observations below the threshold as separating two clusters.

**Example 5.15** We re–consider Example 5.10(c); see also Example 5.2. We show that \( \theta = 1/2 \) can be calculated explicitly in the three different ways explained above. Assume that \( n \mathcal{F}(u_n) \to \tau > 0 \).

(a) The following is immediate from the definition of \( X_n \):

\[
\frac{\ln P(M_n \leq u_n)}{n \ln \mathcal{F}(u_n)} \to \frac{1}{2}.
\]

(b) High threshold exceedances of \( (X_n) \) typically appear in pairs. Hence \( E\xi = 2 \). Since \( \theta \) is the reciprocal of the mean cluster size \( E\xi, \theta = 2^{-1} \).

(c) Finally, consider the conditional probability

\[
P(X_2 \leq u_n, \ldots, X_s \leq u_n \mid X_1 > u_n)
= \frac{P(X_1 > u_n, X_2 \leq u_n, \ldots, X_s \leq u_n)}{P(X_1 > u_n)}
= \frac{F^{s/2}(u_n) - F^{(s+1)/2}(u_n)}{F(u_n)} = \frac{F^{s/2}(u_n)}{1 + F^{1/2}(u_n)}.
\]

The latter expression converges to \( 1/2 \) provided

\[
F^{s/2}(u_n) = \exp\{2^{-1} s \ln(1 - \mathcal{F}(u_n))\} \to 1.
\]

This is clearly satisfied if \( s = s(n) = o(n) \). \( \square \)
5.2.3 Estimating the extremal index from data

In this section we compare the performance of the estimators $\hat{\theta}_n^{(i)}$ of the extremal index $\theta$ both, for real and simulated data. Table 5.18 summarises the results for the exchange rate data presented in Figure 5.17. For data of this type one often claims that ARCH or GARCH models yield a reasonable fit. An ARCH(1) fit, based on maximum likelihood estimation, yields

\[ X_n = \sqrt{1.9 \cdot 10^{-5} + 0.5 X_{n-1}^2 \, Z_n}, \quad n \in \mathbb{N}, \]

for i.i.d $N(0,1)$ noise ($Z_n$). For the above model an ARCH(1) time series with the same length as for the exchange rate data was simulated. The estimators $\hat{\theta}^{(i)}$ are given in Table 5.19. From Table 8.4.23 in Embrechts et al. [1] we may read off the corresponding theoretical value $\hat{\theta} = 0.835$. This shows that $\theta$ is clearly underestimated. Also notice that the estimates strongly depend on the chosen threshold value $u$ and the size $r$.

In Figures 5.20–5.23 the number of exceedances of a given threshold $u$ in a cluster of observations is visualized for the above data. Both, the blocks and the runs method, are illustrated. For the former, $r$ denotes the block size as defined in (5.19). Every block is regarded as a cluster. For the same $r$ we define a cluster in the runs method as follows: it is a set of successive observations separated from the neighbouring sets by a least $r$ values below $u$. See Figure 5.16 for an illustration. The cluster size is then the number of exceedances of $u$ in the cluster.

Every figure consists of three pairs of graphs. For each pair the upper (lower) graph illustrates the blocks (runs) method for the same $u$ and $r$.
Figure 5.17 Log-returns of the exchange rate $\$US/\$UK, January 2, 1980–May 21, 1996 (top, left) and the corresponding sample autocorrelations of this time series (top, right), of its absolute values (bottom, left) and of its squares (bottom, right). The dotted lines indicate the 95% asymptotic confidence band for the sample autocorrelations of iid Gaussian res.

Comments

The concept of extremal index originates from Newell [8], Loynes [7] and O’Brien [10]. A firm definition was given by Leadbetter [3]. An overview of results concerning the extremal index is given in Smith and Weissman [13] and Weissman [14].

Weissman and Cohen [15] present various models where the extremal index can be calculated explicitly. Special methods have been developed for ARMA processes and Markov processes (see Leadbetter and Rootzén [6], Perfekt [11] and Rootzén [12]). For further references, see p. 425 in [1].

References


\[
\begin{array}{|c|c|c|c|c|c|c|c|c|}
\hline
u & r & \hat{\theta}^{(1)} & \hat{\theta}^{(2)} & \hat{\theta}^{(3)} & r & \hat{\theta}^{(1)} & \hat{\theta}^{(2)} & \hat{\theta}^{(3)} \\
\hline
0.015 & 100 & 0.524 & 0.347 & 0.147 & 50 & 0.613 & 0.480 & 0.213 \\
0.020 & 100 & 0.653 & 0.536 & 0.321 & 50 & 0.715 & 0.643 & 0.464 \\
0.025 & 100 & 0.689 & 0.636 & 0.545 & 50 & 0.758 & 0.727 & 0.636 \\
\hline
\end{array}
\]

Table 5.18 Estimators of $\theta$ for the exchange rate data of Figure 5.17 ($n = 4274$) for different thresholds $u$ and sizes $r$.

\[
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
u & r & \hat{\theta}^{(1)} & \hat{\theta}^{(2)} & \hat{\theta}^{(3)} & r & \hat{\theta}^{(1)} & \hat{\theta}^{(2)} & \hat{\theta}^{(3)} \\
\hline
0.015 & 100 & 0.625 & 0.403 & 0.149 & 50 & 0.820 & 0.612 & 0.403 \\
0.020 & 100 & 0.708 & 0.571 & 0.393 & 50 & 0.714 & 0.643 & 0.536 \\
0.025 & 100 & 0.632 & 0.583 & 0.583 & 50 & 0.694 & 0.667 & 0.583 \\
\hline
\end{array}
\]

Table 5.19 Estimators of $\theta$ for $n = 4274$ simulated data from the ARCH(1) model given in (5.24) for different thresholds $u$ and sizes $r$.


154
Figure 5.20 Clusters of exceedances by the blocks method (top figures) in comparison with the runs method (bottom figures) for the exchange rate data from Figure 5.17 (top). The chosen values are \( r = 100 \) and \( u = 0.015 \) (top two), \( u = 0.020 \) (middle two) and \( u = 0.025 \) (bottom two). These figures clearly indicate the dependence in the data.
\textbf{Figure 5.21} Continuation of Figure 5.20. The chosen values are $r = 50$ and $u = 0.015$ (top two), $u = 0.020$ (middle two) and $u = 0.025$ (bottom two).
Figure 5.22 Clusters of exceedances by the blocks method (top figures) in comparison with the runs method (bottom figures) for simulated ARCH(1) data (top) with parameters \( \hat{\lambda} = 0.5 \) and \( \hat{\beta} = 1.9 \cdot 10^{-5} \). The chosen values are \( r = 100 \) and \( u = 0.015 \) (top two), \( u = 0.020 \) (middle two) and \( u = 0.025 \) (bottom two).
Figure 5.23 Continuation of Figure 5.22. The chosen values are $r = 50$ and $u = 0.015$ (top two), $u = 0.020$ (middle two) and $u = 0.025$ (bottom two).