the **Renext** package
user guide

Yves Deville

October 3, 2013, Renext version 2.1.0

Contents

Abstract

The `Renext` package has been specified by IRSN. The main goal is to implement the statistical framework known as "méthode du renouvellement". This is similar to the Peaks Over Threshold (POT) method but the distribution of exceedances is not restricted to GPD. Data Over Threshold can be completed by historical data. Some utility functions of the package are devoted to event analysis or to graphical analysis.

Chapter 1

Introduction

This document was produced using **Renext 2.1.0**. Function calls may have changed in subsequent versions of the package.

Acknowledgments

We gratefully acknowledge the BEHRIG[1] members for their major contribution to designing, documenting and testing programs or datasets: Claire-Marie Duluc, Lise Bardet, Laurent Guimier and Vincent Rebour. We also gratefully acknowledge Yann Richet who encouraged this project from its beginning and provided assistance and many useful advices.

1.1 Goals

The **Renext** package has been specified and implemented by the french *Institut de Radioprotection et de Sûreté Nucléaire* (IRSN). The main goal is to implement in the R environment (R Development Core Team 2010) the statistical framework known within the community of french-speaking hydrologists as *Méthode du Renouvellement* and devoted to Extreme Values problems. This methodology appeared during the years 1980 and was well-accepted both by practitioners and researchers. Although the lack of freely available software may have limited its applicability, this method is still in use or referred to. The book in french by Miquel (1984) still provides an useful and frequently cited reference, while Parent and Bernier (2007) give a more recent presentation.

Although some connexions exist with the theory of Renewal Processes (Cox 1962), it must be said that the standard application of the "Renouvellement" relies on the much simpler Homogeneous Poisson Process (HPP) (Cox and Isham 1980), and is then similar to Peaks Over the Threshold (POT) method (Davison and Smith 1990). POT methods are widespread and are described e.g. in the book of Coles (2001) or that of Embrecht, Klüppelberg, and Mikosch (1996). There are several nice R packages devoted to POT or extreme values: `extRemes` (Gilleland, Katz, and Young 2004), `ismev` (Heffernan and Stephenson 2012), `evd` (Stephenson 2002), `POT` Ribatet (2009), `evdbayes` (Stephenson and Ribatet 2008). The package `nsRFA` (Viglione 2009) also contains useful functions for Extreme Values modelling.

Yet Another POT package?

- Contrary to most POT packages, the distribution of exceedances is not restricted to be in the Generalised Pareto Distributions (GPD) family and can be chosen within half a dozen of classical distributions including Weibull or gamma. Though theory says that GPD will be adequate for large enough thresholds, this is not a counter indication to the use of other distributions. Fitting e.g. Weibull or gamma exceedances might seem preferable to some practitioners and give good results for reasonably large return levels letting asymptotic theory do its job for very large return levels.

[1]IRSN *Bureau d'Expertise Hydrogéologique, Risques d'inondation et géotechnique.*

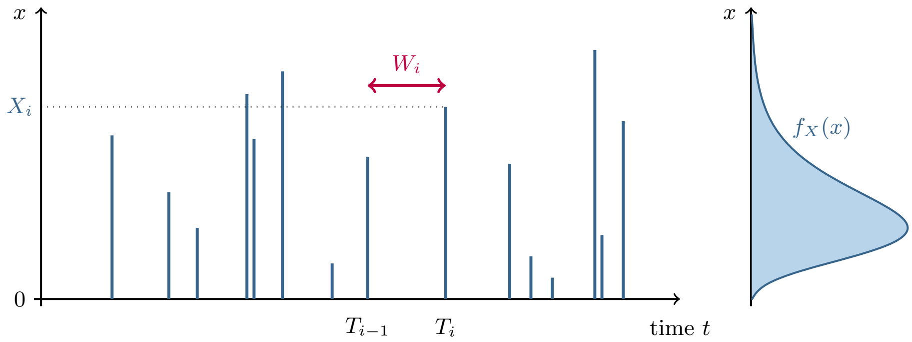


Figure 1.1: Events and levels. The random variable. $W_i = T_i - T_{i-1}$ can be called interevent.

- The package allows the use of *historical data* as explained in section 3.4. Such data can have considerable importance in practical contexts since fairly large periods can be concerned.

Unlike most R packages, **Renext** was not designed to implement innovative techniques arising from recent research in statistics but rather well accepted ones, as used by practitioners. The present document is not intended to be a manual of extreme values modelling but a presentation of the implemented tools with a limited statistical description of these.

The general framework for estimation is *Maximum Likelihood* (ML) and a black-box maximisation can be used with a quite arbitrary distribution of exceedances. For the sake of generality the inference mainly relies on the approximate *delta method*. The present version does not allow the use of covariables.

The package allows extrapolation to fairly large return periods (centuries). Needless to say, such extrapolations must be handled with great care.

1.2 Context and assumptions

1.2.1 Assumptions

The general context is the modelling of a *marked point process* (T_i, X_i). Events (e.g. floods) occur at successive random times T_i when a random variable "level" X_i is observed (e.g. flow). We assume that only *large* values of the level X are of interest. Thus even if the data are recorded on a regular basis (e.g. daily) the data can be soundly pruned to remove small or even moderately large values of X.

Under some general assumptions the instants T_i corresponding to large enough levels X_i should be well described by an *Homogeneous Poisson Process*. Recall that for HPP events the number N of events on a time interval of length w has a Poisson distribution with mean $\mu_N = \lambda \times w$. Moreover the numbers of T_i corresponding to disjoint intervals are independent. The parameter $\lambda > 0$ is called the *rate* and has the physical dimension of an inverse time: it will generally be given in inverse years or events by year. Another important property of the HPP is that the interevent random variables $W_i = T_i - T_{i-1}$ are independent with the same exponential distribution with mean $1/\lambda$.

Unless explicitly stated otherwise, we will make the following assumptions about the marked process

1. Events T_i occur according to a Homogeneous Poisson Process with rate λ.

2. Levels X_i form a sequence of independent identically distributed random variables with continuous distribution $F_X(x)$ and density $f_X(x)$.

3. The levels sequence and events sequence are independent.

The distribution $F_X(x)$ will be chosen within a parametric family and depends on a vector of parameters $\boldsymbol{\theta}_X$. This dependence can be enlightened using the notation $F_X(x; \boldsymbol{\theta}_X)$ when needed. The parameters of the whole model consist in λ and a vector $\boldsymbol{\theta}_X$.

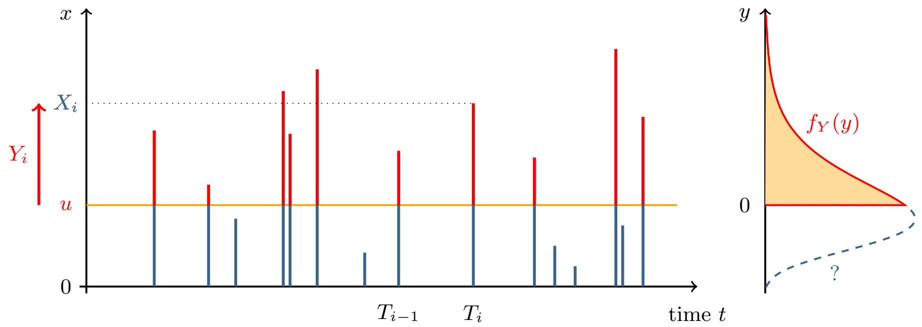


Figure 1.2: In POT only levels X_i with $X_i > u$ are modeled through exceedances $Y_i = X_i - u$. The lower part $x < u$ of the distribution $F_X(x)$ remains unknown.

1.2.2 Return periods

The *return period* of a given level x is the mean time between two events T_i with levels exceeding x, that is with $X_i > x$. Under the assumptions above, it is given by

$$T(x) = \frac{1}{\lambda\,[1 - F_X(x)]}. \tag{1.1}$$

Indeed the probability of $\{X_i > x\}$ is $1 - F_X(x)$ and the events with level exceeding x also form an HPP[2] (thinned HPP) with rate $\lambda\,[1 - F_X(x)]$. The mean interevent is the inverse rate.

Note that a complete knowledge of the distribution is not required since only large levels x are of interest.

1.2.3 Peaks Over Threshold (POT)

In the Peaks Over Threshold (POT) approach, only the upper part of the distribution $F_X(x)$ is modelled. More precisely, the interest is on the part $X > u$ where u is a *threshold*. The steps are

- Fix a suitable threshold u,
- Consider only the observations with level X_i greater than u i.e. with $X_i > u$,
- Estimate the rate of the events $X_i > u$ and fit a distribution to the exceedances $Y_i = X_i - u$.

The distribution of X conditional on $X > u$ is deduced from that of the exceedance Y by translation.

The threshold will often be chosen above the mode of X, leading to a decreasing density for the exceedance Y as suggested on figure 1.2. The distribution of Y typically has two parameters.

The determination of the threshold is a recognized difficulty in classical POT where only GPD exceedances are used. The situation is much more complex when non-GPD exceedances are used. The family of GPD distributions with a given shape parameter ξ can be said "stable for exceedances". With another threshold $v > u$ the estimation will use a smaller set of X_i but the underlying distribution of X conditional on $X > v$ is the same in the two cases. If a non-GPD distribution is used for the exceedances this is non-longer true. For instance if the exceedances over u are Weibull with shape $\alpha > 0$ and scale $\beta = 1$ i.e.

$$\Pr\{\,X > x \mid X > u\,\} = \exp\{-(x-u)^{\alpha}\} \qquad x > u$$

then the conditional distribution over a higher threshold $v > u$ is given by

$$\Pr\{\,X > x \mid X > v\,\} = \exp\{-\left[(x-u)^{\alpha} - (v-u)^{\alpha}\right]\} \qquad x > v > u$$

The distribution of the exceedance $X - v \mid X > v$ *is not* Weibull; it is a shifted version of the *Left Truncated Weibull* (LTW), see B.3.10.

[2]The is due to the independence of the two sequences X_i and T_i.

1.2.4 Link with other Extreme Values problems

Alternative approaches in Extreme Values modelling use time *blocks* of, say, one year and related by-block data. Popular examples are

- **block maxima**: for each block, only the maximal value is used in the analysis.
- r**-largest**: for each block the largest r observations (i.e. the r largest order statistics) are recorded. The number r may vary for different blocks.

Block maxima is obviously the special case $r = 1$ of the r-largest analysis, and using $r > 1$ largest observations when available leads to a better estimation. The r-largest analysis is described in chap. 3 of the book of Coles (2001). The distribution retained for the maxima or the r-largest is based on asymptotic considerations. Underlying the block data, one would generally find a continuous time process (e.g. temperature, sea surge), possibly observed at fixed times (e.g. high tide). The time-length of the blocks is generally chosen in order to reach a limit behaviour ignoring autocorrelation or seasonality in the continuous process.

Interestingly, the assumptions concerning the marked point process as stated before in 1.2.1 provide a framework to derive the distribution of the maxima or that of the r-largest observations over non-overlapping blocks. These distributions can be related to the distribution of the marks, see section 3.4 for the likelihood of a r-largest block, and appendix page 26 for a general study of the max. Maxima or r-largest observations can be viewed as *partial observations* of the marked process, or as the result of a *temporal aggregation* of this process. When the result of such an aggregation (i.e. maxima or r-largest) is known for one or several blocks with large durations, say decades or centuries, we may speak of *historical data*.

Although **Renext** primarily uses "OT data" as described above, it is possible to make use of supplementary data in a quite flexible fashion. Maxima and r-largest observations within block(s) can also be used, as well as the marks exceeding some known auxiliary threshold as sometimes called a *perception threshold*. A typical use of these possibilities is for historical data.

The notion of *return period* for the blocks framework differs from the one given above see discussion A.3 page 27. However, the difference between the two notions is confined to the small return periods context.

1.3 Data

1.3.1 Remarks

Model fitting functions in R usually have a formal argument specifying data with a *data.frame* object, the model being typically given by a *formula*. Due to the presence of heterogeneous types of data within a given "dataset", the arguments of **Renext** functions will take a slightly more complex form. For instance, it will generally be necessary to specify a duration or several block durations in complement to the vector of levels, to specify where missing periods (gaps) occurred, etc.

Some of the package functions require the use of `POSIX` objects representing date and time. R base package provides versatile functions to manage date/time or timestamps. See for instance the help of the `strptime` function.

As most R packages do, **Renext** comes with a few datasets taken from relevant literature or from real data examples. These datasets are usually given as lists objects with hopefully understandable element names.

1.3.2 OT data

The data used will mainly consist in recorded levels X_i or levels exceeding a reasonably low known threshold u_0. The POT modelling of such data will typically use a higher threshold $u > u_0$.

For instance the data `Brest` contain sea surge heights at high tide for the Brest gauging station. Only values exceeding $u_0 = 30\,\text{cm}$ are retained. More details about these data are provided in the package manual. The data are provided as a list with several parts.

```
> library(Renext)
> names(Brest)
```

```
[1] "info"      "describe"  "OTinfo"    "OTdata"    "OTmissing"
```

As their names may suggest the list elements contain Over Threshold (OT) data and information.

```
> head(Brest$OTdata, n = 4)
                 date  Surge comment
1 1846-01-13 23:59:39 35.989
2 1846-01-20 23:59:39 59.987
3 1846-01-23 23:59:39 45.986
4 1846-01-27 23:59:39 39.985
```

```
> str(Brest$OTinfo)
List of 4
 $ start      : POSIXt[1:1], format: "1845-12-31 23:59:39"
 $ end        : POSIXt[1:1], format: "2009-01-01"
 $ effDuration: num 148
 $ threshold  : num 30
```

The `OTdata` element is a data.frame indicating T_i (in time order) and the corresponding levels X_i. Note that the time part of the `POSIX` object may not be relevant. Here only the date part makes sense and the time part is by convention `"00:00"`. However on such a large period of time, it is affected by *leap seconds*, and `"00:00"` might appear as `"23:59"` the day before.

The `OTinfo` list mentions an *effective duration*. This is less than the time range which can be computed using the methods `range` and `diff` from the **base** package

```
> End <- Brest$OTinfo$end; Start <- Brest$OTinfo$start
> Dur <- as.numeric(difftime(End,  Start, units = "days"))/365.25
> Dur
[1] 162.9979
> Dur - as.numeric(Brest$OTinfo$effDuration)
[1] 15.37785
```

The difference – more than 15 years – is due to gaps or *missing periods*. The missing periods are described in the element `OTmissing`.

The `Brest` dataset has class `"Rendata"`. This is an S3 class defined in **Renext** to describe objects containing `OTdata` and possibly some extra information on missing periods or historical data. It has a `summary` method

```
> class(Brest)
[1] "Rendata"
> summary(Brest)
o Dataset Surge Heights at Brest (France)
   data 'Brest', variable 'Surge' (cm)

o OT data (main sample) from  1845-12-31  to  2009-01-01  (eff. dur. 147.62 years)

      n    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1289.00   30.02   33.65   38.31   41.76   46.58  143.90

o missing 'OT' periods, total 15.38 years

        n      Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
43.000000  0.002738  0.016430  0.038330  0.357600  0.086240  8.419000

o no 'MAX' historical data

o no 'OTS' historical data
```

The displayed information concerns the levels in the main OT sample and the possible gaps in this sample: number, duration (in years).

A `plot` method also exists

```
> plot(Brest)
```

which produces the plot on the left of figure 1.3.

1.3.3 Missing periods or gaps

A common problem in POT modelling is the existence of gaps within the observation period. These can result from many causes: damage or failure of the measurement system, human errors, strikes, wars, ...

Renext uses a natural description of the gaps within a dataset. They are stored as rows of a data.frame with two `POSIX` columns `start` and `end`

```
> head(Brest$OTmissing, n = 4)
                start                 end comment
1 1845-12-31 23:59:39 1846-01-03 23:59:39
2 1846-12-31 23:59:39 1847-01-20 23:59:39
3 1852-01-20 23:59:39 1852-02-07 23:59:39
4 1857-05-30 23:59:39 1859-11-23 23:59:39
```

Missing periods must be taken into account in the analysis. They should be displayed on timeplots showing events, since it is important to make a distinction between periods with no events and gaps, see figure 1.3. An important prerequisite to modelling is to ensure that the gaps occur independently from measured variables. For instance, storms can damage gauging systems for wind or sea level thus creating a non-independent (or endogenous) gap.

1.3.4 Historical data

As a possible complement to `OTdata`, we may have `MAXdata` that is: r-largest observations over one or several *blocks*. Such data require a complementary information: the block duration(s) which must be given in a chosen time unit.

The dataset `Garonne` is taken from Miquel (1984) where it is described. The data concern the french river *La Garonne* at the gauging station named *Le Mas d'Agenais* where many floods occurred during the past centuries. The data consist in both OT data and historical data. The variable is the river flow in m^3/s as estimated from the river level using a rating curve. The precision is limited and many ties are present among the flow values. The OT data contain flow values over the threshold $u = 2500\ \text{m}^3/\text{s}$. The historical data are simply the 12 largest flows for a period of about 143 years and will be used later.

```
> names(Garonne)
[1] "info"      "describe"  "OTinfo"    "OTdata"    "OTmissing" "MAXinfo"
[7] "MAXdata"
> Garonne$MAXinfo
                start        end duration
1 1769-12-31 23:59:39 1913-01-01   143.09
> head(Garonne$MAXdata, n = 4)
  block date Flow  comment
1     1 <NA> 7500 1 (1875)
2     1 <NA> 7400 2 (1770)
3     1 <NA> 7000 3 (1783)
4     1 <NA> 7000 4 (1855)
```

The `Garonne` dataset has class `"Rendata"`. The `plot` method for this class

```
> plot(Garonne)
```

produces a graphic displaying the historical period as on the right panel of figure 1.3. Note that the dates of the historical events are not known exactly and thus are as `NA` `POSIXct` objects. The historical levels are thus displayed as horizontal segments, while vertical segments would be used for known dates. The `plot` method for the class `Rendata` has a `show.hist` logical formal argument telling that historical periods should be shown (default value `TRUE`) or not.

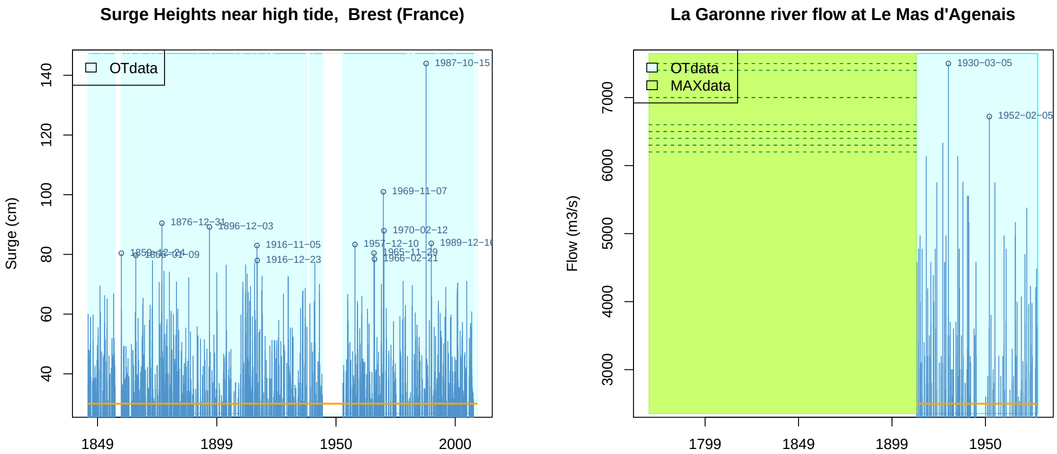


Figure 1.3: Graphics produced using the `plot` method of the `"Rendata"` class. On the left, the `Brest` object contains missing periods that are shown. On the right, the `Garonne` dataset contains information about an *historical period*, displayed as a green rectangle.

1.3.5 Aggregated data, counts

In some cases, the original data have been aggregated: the T_k are unknown and the X_k only have a block indication. For instance, we may know only the year for each event, or the year and the month. In a such scheme several events will fall in the same block. This situation is somewhat comparable to the r-largest context, but the data are here all the levels X_k over a known threshold and not only the largest levels. The difference is comparable to that between the two types of censoring (types I and II).

A difficulty with aggregated data concerns the treatment of missing information or missing data (gaps). There is usually no reason that missing periods should correspond to years and ignoring all blocks with a gap leads to a severe loss of information.

The use of aggregated data will be illustrated later in the discussion about `barplotRenouv`.

Chapter 2

Descriptive tools

2.1 Functional plots

2.1.1 Principles

Widespread graphical tools in statistics are *functional plots*, such as exponential plot, Weibull or Gumbel plot. In all cases, the plot is designed so that the theoretical distribution curve (exponential/Weibull/Gumbel) shows as a straight line. For instance the relations for distribution functions

$$\begin{aligned} -\log\left[1 - F_X(x)\right] &= (x-\mu)/\sigma \quad \text{(exponential)} \\ -\log\left[-\log F_X(x)\right] &= (x-\mu)/\sigma \quad \text{(Gumbel)} \end{aligned}$$

both show a linear relation between x and a transformed version $\phi(F)$ of $F_X(x)$, e.g. $\phi(F) = -\log\left[1-F\right]$ for the exponential case. The functional plots are obtained by plotting $[x,\, \phi(F)]$ still using the values of the probability F to display the unevenly spaced graduations on the y-axis. The Weibull plot is similar but also uses a (log) transformation of x.

With a sample X_i of size n one uses non-parametric estimates $\widetilde{F}_{[i]}$ of the values $F_X(X_{[i]})$ of the distribution function at the order statistics $X_{[i]}$. The n resulting points with ordinates $\widetilde{F}_{[i]}$ can be plotted with the transformed scale on the y-axis. Two classical options for the estimation and thus for the plotting positions are

$$\widetilde{F}_{[i]} \approx i/(n+1) \qquad \widetilde{F}_{[i]} \approx (i-0.3)/(n+0.4)$$

The first choice is motivated by the fact that $i/(n+1)$ is the expectation of $F_X(X_{[i]})$. The second option uses an approximation of the median.

As many other packages do, **Renext** provides exponential and Weibull plotting functions, namely `expplot` and `weibplot`

```
> expplot(x = Brest$OTdata$Surge, main = "expplot for \"Brest\"")
> weibplot(x = Brest$OTdata$Surge-30, main = "weibplot for \"Brest\" (surge - 30)")
```

producing the two plots on figure 2.1.

Note that the transformation $\phi(F)$ must not depend on unknown parameters. Therefore the Weibull plot produces a theoretical line only for the version with two parameters (shape and scale), and not for the three parameter one (with location).

2.1.2 Exponential vs Gumbel

While hydrologists often favour Gumbel plots, the exponential plot may also be used. The latter is better suited to the use of "OTdata" i.e. data where only values over a threshold u_0 are kept. Even if the original observations X_i are Gumbel, the conditional distribution $X_i \mid X_i > u_0$ will be close to an exponential for u_0 large enough, see B.1.3. This can be illustrated with a few simple R commands

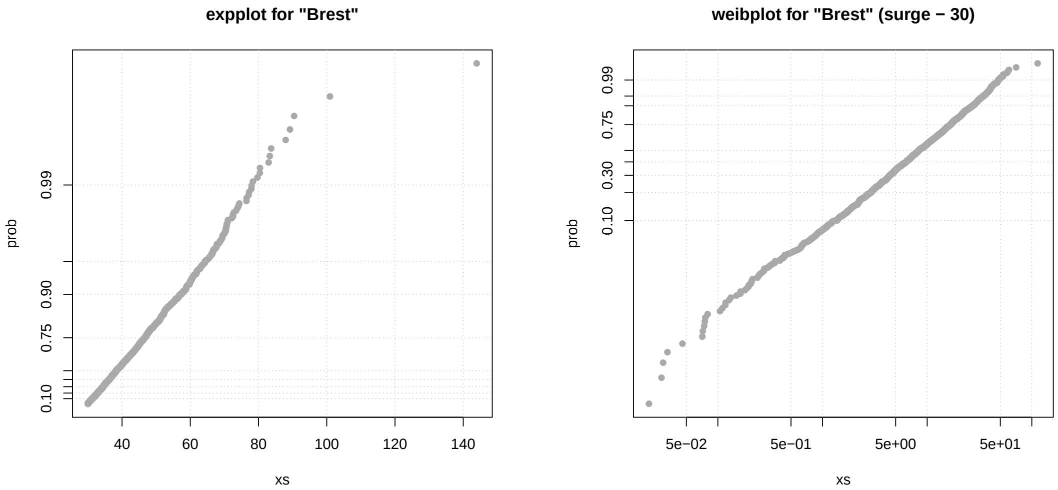


Figure 2.1: Exponential and Weibull plot for the Brest data. The variable `Surge` is used for the exponential plot. The threshold 30 cm is subtracted from `Surge` for the Weibull plot. The later uses a log-scale for `x`.

```
> library(evd); set.seed(136)
> X <- rgumbel(400); X <- X[X > 0.6]           ## X is truncated Gumbel
> n <- length(X);
> Z <- sort(X); F <- (1:n)/(n+1)               ## distribution function
> y.exp <- -log(1-F); y.gum <- -log(-log(F))
> plot(Z, y.exp, col = "red3", main = "exponential plot")

> plot(Z, y.gum, col = "SteelBlue3", main = "Gumbel plot")
```

The two plots are shown on figure 2.2. As a general fact the difference between exponential and Gumbel plots is restricted to the small values since the exponential and Gumbel distribution functions are close for large values.

2.2 Events and stationarity

Simple plots

The simplest plot for checking stationarity has points $[T_i, X_i]$ and can be obtained with R functions of the `graphics` package. The T_i and X_i will typically be available as two vectors of the same length or as two columns of a same data.frame object. For the example datasets of **Renext**, the T_i and X_i are given as two columns of the `OTdata` data frame

```
> plot(Flow ~ date, data = Garonne$OTdata, type = "h", main = "Flows > 2500 m3/s")
```

The graphics shows that several successive years had no exceedance over 2500 m^3/s during the second half of the 1940-1950 decade. This could lead to further investigations using the `subset` function

```
> subset(Garonne$OTdata, date >= as.POSIXct("1945-01-01") & date <= as.POSIXct("1950-01-01"))
         date Flow comment
96 1945-01-29 3200
```

The graphics can be enhanced using the `text` function in the `graphics` package to annotate special events or periods.

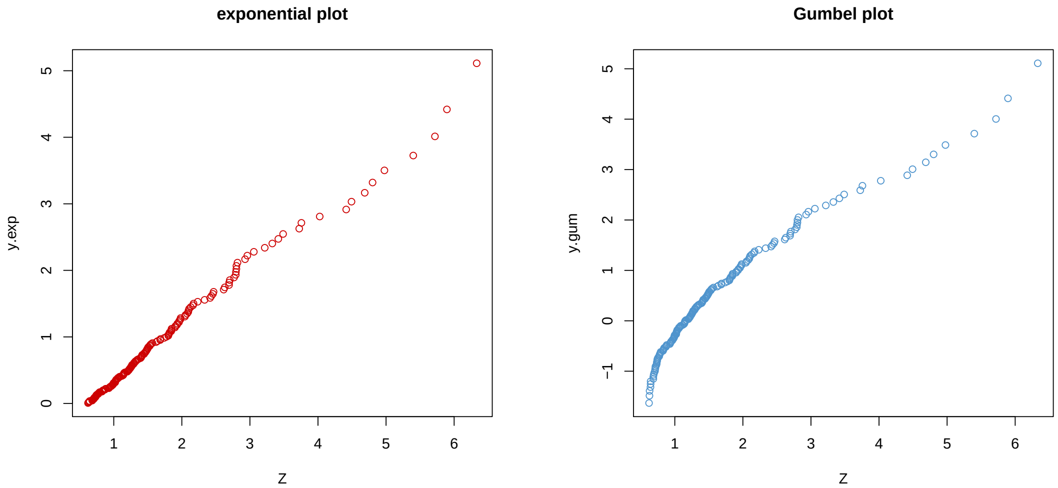


Figure 2.2: Truncated or ”thresholded” Gumbel random sample. Due to the truncation, the sample distribution is close to an exponential. The graduations for the y-axis are not in probability-scale.

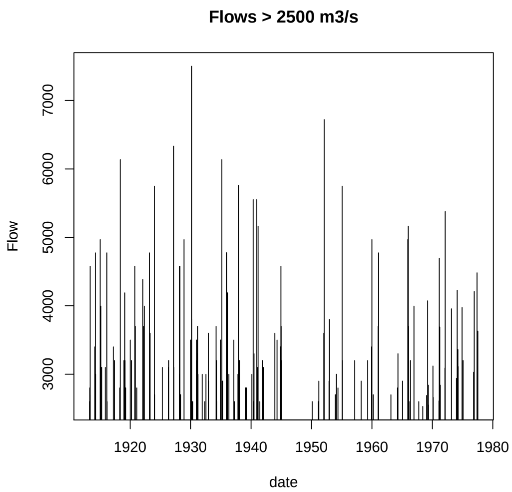


Figure 2.3: Simple plot of events for the `Garonne` data.

Uniformity

The `gof.date` function performs some tests to check the (conditional) uniformity of the events T_i as implied by the HPP assumption. It is based on the fact that for a given interval of time (s, t) the events T_i falling in the interval are jointly distributed as are the order statistics of a sample of the uniform distribution on (s, t). The sample size n is then random. Alternatively, the n events falling in an interval (T_k, T_{n+k+1}) also have this joint conditional distribution. In both cases a Kolmogorov-Smirnov (KS) test is well suited to check the uniformity.

The `gof.date` function mainly works with a `POSIX` object containing the events T_i as in

```
> gof.date(date = Garonne$OTdata$date)
```

which produces the plot on the left of figure 2.4. The empirical cumulative distribution function (ECDF) is compared to the uniform and the KS distance D_n is shown as a vertical segment. The displayed KS p-value tells that uniformity should be rejected at the significance level of 0.1%. Though less clearly than above, the plot points out that the years 1940-1950 had fewer events.

The `gof.date` function has optional args `start` and `end` to specify (and possibly restrict) the period on which the test is performed. By default these are taken as the first and last event in `date` and therefore only inner events are used in the ECDF.

Interevents

An important property of the HPP concerns the interevents $W_i = T_i - T_{i-1}$: the sequence W_i is independent and have exponential distribution with rate λ. Thus an exponentiality test might be performed to check the HPP assumption for observed data.

The `interevt` function computes the interevents W_i as numbers of days. The function returns a list with a `interevt` data.frame element containing the W_i in the `duration` column which can be used to check exponentiality. This can be done either with a plot - see figure 2.4 or with the test of exponentiality of the function `gofExp.test`

```
> ie <- interevt(date = Garonne$OTdata$date)
> names(ie)
[1] "interevt" "noskip"
> d <- ie$interevt$duration
> expplot(d, main = "Exponential plot for interevents")
> bt <- gofExp.test(d)
> bt
$statistic
[1] 193.9631

$df
[1] 149

$p.value
[1] 0.01557954

$method
[1] "Bartlett gof for exponential"
```

It seems unlikely to obtain a good exponential fit as far as events occurrence shows seasonality as is the case here. A seasonality can no longer result from another distribution of interevents – that is from a non-Poisson stationary renewal process. Increasing the threshold might improve the adequacy to the assumptions.

Missing periods or gaps

In practice the situation is somewhat more complex due to the possible existence of missing (or skipped) periods where no events have been recorded. Event rates should then be computed using *effective duration* that is: the total duration of measurement *ignoring missing periods*.

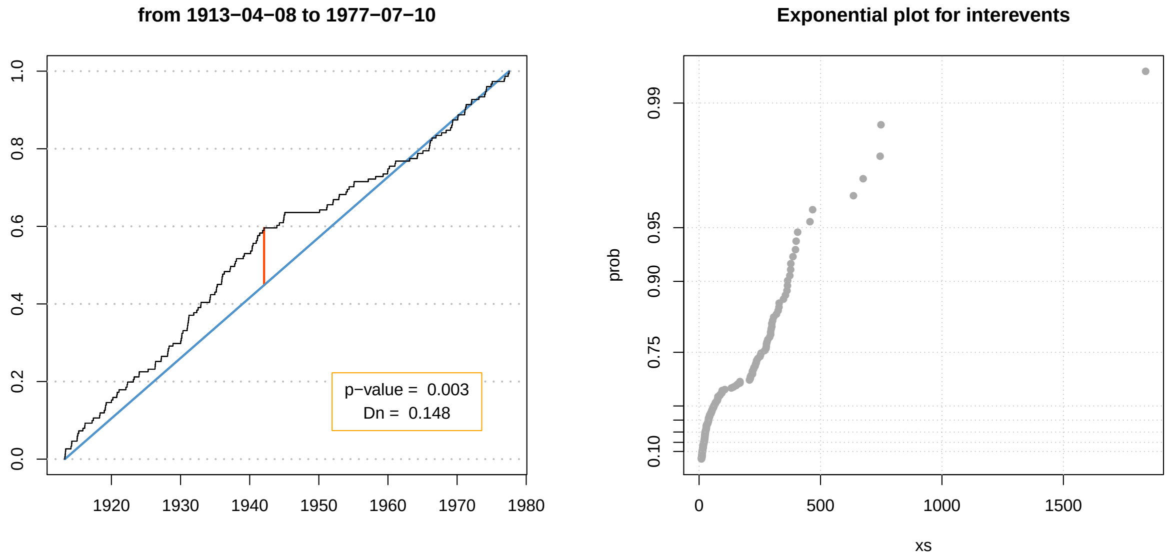


Figure 2.4: Analysis of the events for the `Garonne` data set (OTdata). Left panel: test for the uniformity of events with the KS distance shown as a vertical segment. Right panel : exponential plot for the interevents.

The functions `gof.date` and `interevt` can take this problem into consideration. The `gof.date` plot can display the missing periods or "gaps" provided that a suitable `skip` arg is given. For instance the following commands produce the plot on the left of figure 2.5

```
> gof.Brest  <- gof.date(date = Brest$OTdata$date, skip = Brest$OTmissing,
                         start = Brest$OTinfo$start, end = Brest$OTinfo$end)
> print(names(gof.Brest))
 [1] "effKS.statistic" "effKS.pvalue"    "KS.statistic"    "KS.pvalue"
 [5] "effnevt"         "nevt"            "rate"            "effrate"
 [9] "duration"        "effduration"     "noskip"
```

As their name may suggest, the returned list elements give the effective duration and the effective rate based on the true non-missing periods. The `noskip` element contains detailed information about each non-skipped period

```
> head(gof.Brest$noskip, n = 2)
         start          end duration nevt       rate        Dn         KS
1 1846-01-.... 1846-12-.... 0.991102   17 17.152624 0.2586935 0.17172882
2 1847-01-.... 1852-01-.... 4.999316   48  9.601314 0.2057777 0.02929104
```

For each period the rate has been computed as well as a KS test of uniformity. The power of the test is obviously limited for periods with few events.

The preceding call to `gof.date` corresponded to the default value of `plot.type` namely `"skip"` A drawback of the plot and KS test is that the comparison with the uniform is biased by the gaps. The KS distance D_n between the empirical and theoretical distributions can be amplified by the gaps when there are too few events or on the contrary be reduced by gaps when there are too much events. These two phenomena can be seen by comparing the two plots of figure 2.5 although the two KS statistics and p-value are here nearly identical. The right panel plot was produced using the non-default choice for the `plot.type` arg i.e. `plot.type= "omit"`, missing periods can be omitted on the plot and in the KS test computation.

```
> gof.Brest2  <- gof.date(date = Brest$OTdata$date,
                          skip = Brest$OTmissing, plot.type = "omit",
                          start = Brest$OTinfo$start, end = Brest$OTinfo$end)
```

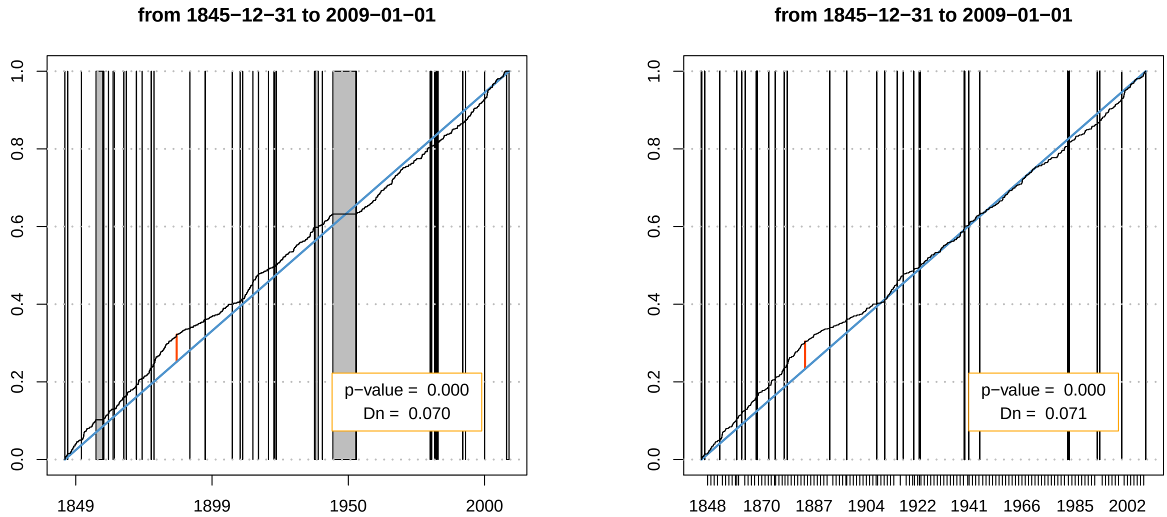


Figure 2.5: Using the `plot.type` arg of `gof.date` leads to the left panel (default value or `"skip"`), or the right one (value `"omit"`). Each missing period appears as a gray rectangle on the left graph and is flattened as a line on the right graph.

The time axis now has *unevenly* spaced ticks since it is obtained by concatenating the successive non-missing periods. More precisely, each retained time interval k begins at the first event T_{f_k} of a continuous observation period and ends at its last event T_{ℓ_k}. Each of the vertical lines shows an interval $(T_{\ell_k}, T_{f_{k+1}})$, which covers a missing period and is cut out as shown on figure 2.6. The displayed information on the right panel of figure 2.5 concerns `effKS.pvalue` and `effKS.statistic` of an "effective" KS test performed on non-missing periods. Provided that observation gaps occur independently from the events T_i, the interevents for couples of successive events falling in the same non-missing period can be used in a modified KS test. In the HPP case these interevents should be independent and identically distributed with exponential distribution thus concatenating them should produce an HPP hence an uniform conditional distribution of events.

For the `Brest` example, the test tells us that the uniformity of events should be rejected while the plot indicates that there were more events during the XIXth century than in during the XXth. Since large surges tend to occur more frequently in winter, further investigation of the gaps distribution would be useful.

2.2.1 Aggregated (counts) data

The `barplotRenouv` function draws a barplot for counts data and performs a few tests adapted to this context where events or interevents can no longer be used. The data used are n counts N_i for $i = 1$, 2, $\dots$, n. These counts must be on *disjoint intervals* or "blocks" with the *same duration*, e.g. one year. If events occur according to an HPP the N_i form a sample of a Poisson distribution. The barplot compares the empirical (or observed) frequencies to their theoretical counterparts i.e. the expectations. The theoretical distribution is estimated using the sample mean as Poisson parameter (Poisson mean).

The `Brest.years` object contains aggregated data for one-year blocks. Some blocks are incomplete and are listed in `Brest.years.missing` which can be used in `barplotRenvouv`

```
> data(Brest.years); data(Brest.years.missing)
> bp40 <- barplotRenouv(data = Brest.years, threshold = 40,
            na.block = Brest.years.missing, main = "threshold = 40 cm")
```

produces the graphic at the left of figure 2.7. Increasing the threshold

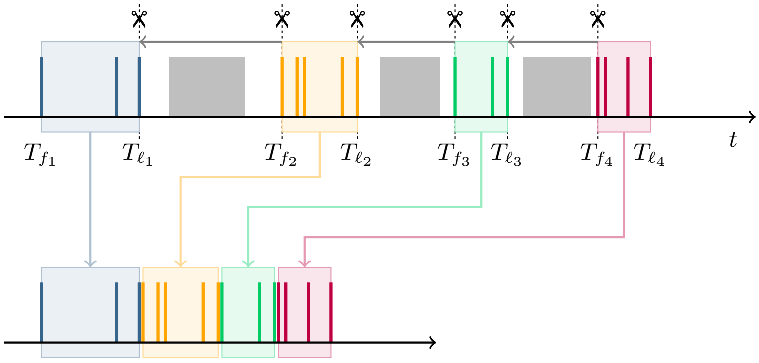


Figure 2.6: With `plot.type = "omit"`, the plot of `gof.date` only considers interevents for couples falling in the same non-missing period and concatenates them. The time interval $(T_{\ell_k}, T_{f_{k+1}})$ between the last event T_{ℓ_k} of the non-missing period k and the first event $T_{f_{k+1}}$ of the following non-missing period is "cut out". The two events T_{ℓ_k} and $T_{f_{k+1}}$ collapse into *one* event of the new Point Process. Note that a non-missing period with less than two events is cut out since it contains no valid interevent.

```
> bp50 <- barplotRenouv(data = Brest.years, threshold = 50,
                  na.block = Brest.years.missing, main = "threshold = 50 cm")
```

we get a barplot for the smaller sample at the right of figure 2.7. Note that the function guesses that the first column represents a block indication which may not be true with other data. Therefore the normal use would specify the `blockname` and `varname` formal arguments of `barplotRenouv`.

Great care is needed when the data contain missing periods since the number of events is then biased downward.

Goodness-of -fit

A popular test for Poisson counts is called *overdispersion test*. It is based on the fact that expectation and variance are equal in a Poisson distribution. The test statistic is

$$I = (n-1)\, S^2/\bar{N}$$

where $\bar{N}$ and S^2 are the sample mean and variance. Under the null hypothesis I is approximately distributed as $\chi^2(n-1)$. The statistic I tends to take large values when the observations N_i come from an overdispersed distribution such as the negative binomial. A one-sided test can therefore be used for a negative binomial alternative.

A Chi-square Goodness-of-fit test is also available to check the goodness-of-fit of the N_k to a Poisson distribution. In this test, the counts values N_k are summarized in a tabular format retaining m distinct values or group of adjacent values, together with the corresponding frequencies. The test statistic is

$$D^2 = \sum_{k=1}^{m} (O_k - E_k)^2 / E_k$$

where O_k and E_k are the observed and expected frequencies for the class k. For instance, the first class $k = 1$ can be $N = 0$ meaning that O_1 and E_1 are the number of intervals with no events recorded. Asymptotically (for large n)

$$D^2 \sim \chi^2(m-p-1)$$

where p is the number of parameters estimated from data, here $p = 1$ (for the mean of N). A one-sided test will reject the Poisson hypothesis when D^2 is too large[1].

A classical drawback of this test is that classes with a small expected count E_i should be grouped, in order to reach a minimal total of (say) 5.

```
> bp40$tests
```

[1]That is: $D^2 > \chi^2_\alpha$

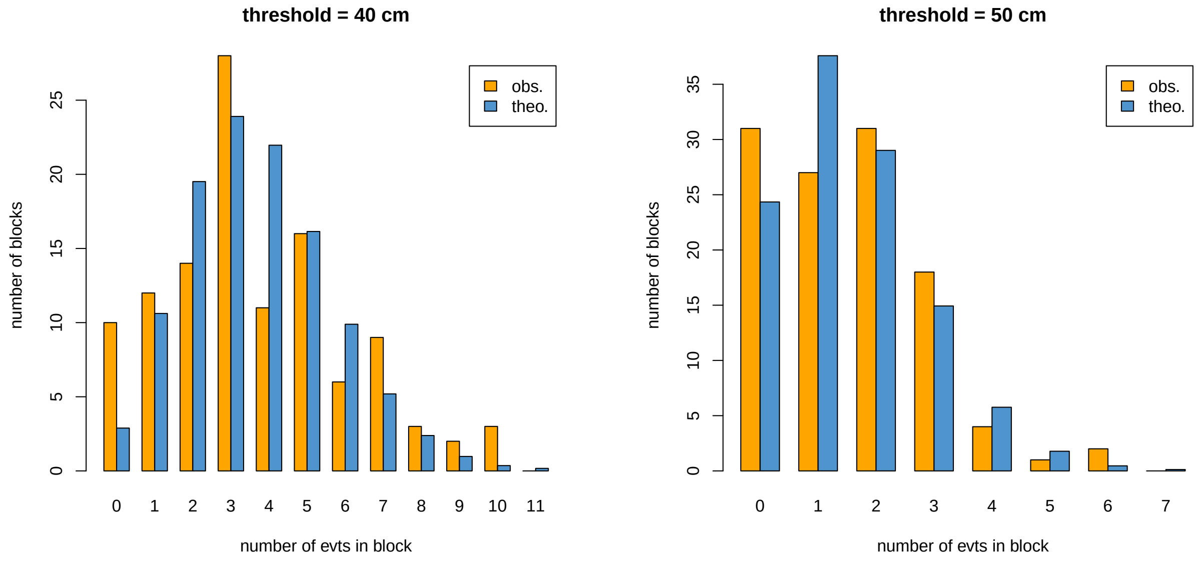


Figure 2.7: The two barplots produced with `barplotRenouv`. A bar height represents a number of blocks (here years) with the number of events given in abscissa.

```
     statistic  df      p.value
disp  181.4726 113 4.652672e-05
chi2   21.5105   5 6.485040e-04
> bp50$tests
      statistic  df   p.value
disp 131.022727 113 0.1181542
chi2   5.722912   3 0.1258975
```

For the dataset `Brest.years`, using a threshold of 50 cm leads to acceptable tests (at the 10% level), while 40 cm seems too small. For the chi-square test, more details (e.g. grouping) are available.

```
> bp50$freq
  obs.      theo. group
0   31 24.3452997     1
1   27 37.5857258     2
2   31 29.0135427     3
3   18 14.9309460     4
4    4  5.7628213     5
5    1  1.7793974     5
6    2  0.4578567     5
7    0  0.1244104     5
```

The values of N have been grouped in odrer to reach a minimal expected number of 5 for each group.

Note that for a fairly high threshold, the statistic N will generally take only the two values 0 and 1. Then the chi-square test which requires at least three classes will not be available.

Chapter 3

The `Renouv` function

3.1 Fitting POT for La Garonne

For the dataset `Garonne`, the OT data contain flow values over the threshold $u = 2500$ m^3/s. We can fit a POT model with any threshold $u \geqslant 2500$. As in Miquel (1984) we fit an exponential and a two parameters Weibull distribution using OT data only. The `Renouv` needs on input the *levels* given in a vector `x`, the *effective duration* `effDuration` – normally in years – and the *threshold*

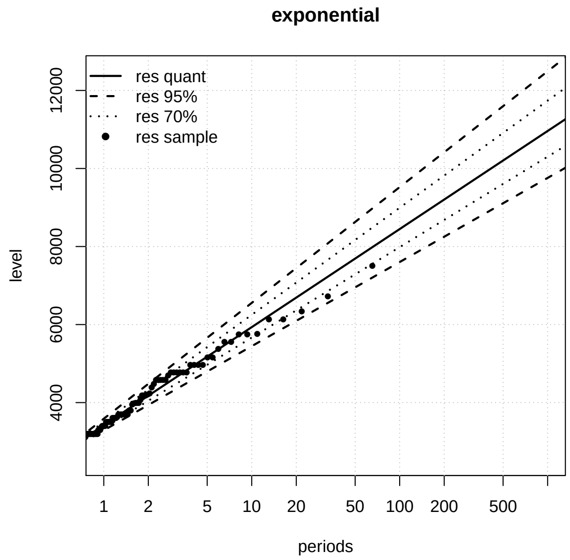
```
> fit.exp <- Renouv(x = Garonne$OTdata$Flow,
                    effDuration = 65, threshold = 2500,
                    distname.y = "exponential",
                    main = "exponential")
Special inference for the exponential case without history
> fit.exp$estimate
      lambda         rate
2.3230769231 0.0009160231
```

The result is mainly a list within which an `estimate` element gives the maximum likelihood estimates. The first element named `"lambda"` is the event rate expressed in *events by year*. The other elements are the ML estimates of the distribution for exceedances, with names corresponding to the probability functions – here one name `"rate"` for the exponential distribution parameter. Many other results are returned

```
> names(fit.exp)
 [1] "call"         "x.OT"         "y.OT"         "nb.OT"        "effDuration"
 [6] "threshold"    "distname.y"   "p.y"          "parnames.y"   "fixed.y"
[11] "trans.y"      "est.N"        "cov.N"        "est.y"        "cov.y"
[16] "corr.y"       "estimate"     "fixed"        "df"           "nobs"
[21] "p"            "opt"          "logLik"       "sigma"        "cov"
[26] "corr"         "history.MAX"  "history.OTS"  "funs"         "transFlag"
[31] "pct.conf"     "ret.lev"      "pred"         "infer.method" "KS.test"
[36] "expon.test"
> class(fit.exp)
[1] "Renouv"
```

The list is in fact an object with (S3) class `"Renouv"`. It is possible to display the results using the `summary` method, which would be invoked here simply by `summary(fit.exp)`. A few other S3 methods are available. For instance, `coef` extracts the estimated coefficients, and the ubiquitous `plot` method can be used to re-draw a return level plot from the fitted object. The `predict` can be used to compute return levels corresponding to given return periods.

The `distname.y` formal in `Renouv` is used to change the distribution for exceedances $Y_i = X_i - u$.



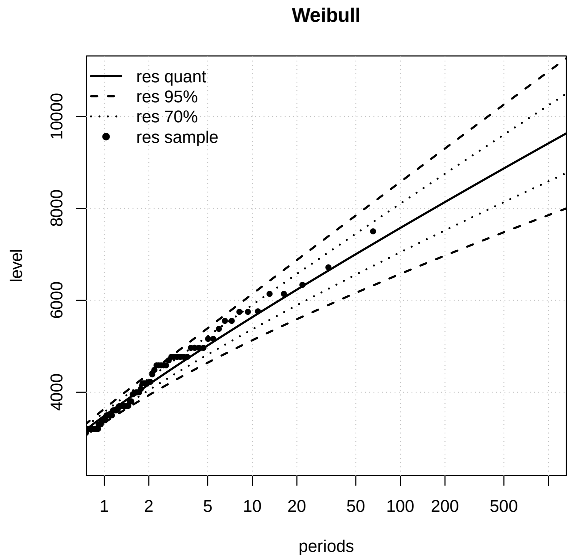


Figure 3.1: Return level plots for the example `Garonne` with two distributions for exceedances.

```
> fit.weibull <- Renouv(x = Garonne$OTdata$Flow,
                        effDuration = 65, threshold = 2500,
                        distname.y = "weibull",
                        main = "Weibull")
> fit.weibull$estimate
    lambda       shape       scale
  2.323077    1.139363 1145.889216
> fit.weibull$sigma
    lambda       shape       scale
0.18904932  0.07229351 86.17717237
```

The estimated parameters of the Weibull distribution and their standard deviation (list item `sigma`) show that the shape is close to 1.0, which corresponds to the exponential distribution. The two fits produced return level plots shown on figure 3.1.

3.2 Return level plot

3.2.1 Description

Renext uses a return level plot which may be qualified as *exponential*, and differs from the usual one which uses *Gumbel* scales. The main difference is that the exponential plot uses a log scale for return periods while the Gumbel plot uses a log-log scale. In both cases, the theoretical return level curve (exponential/Gumbel) shows as a straight line.

The difference between the two plots is restricted to the small levels/return periods, since the exponential and Gumbel distribution functions are close for large values. As it was advocated in the discussion about functional plots page 8, the exponential return level plot is better suited to the use of "OTdata" i.e. data where only values over a threshold u_0 are kept, even if the the original observations X_i are Gumbel see B.1.3.

Note that the return level plot is similar to the classical exponential plot, *but with the two axes x, y exchanged*. A concave (downward) RL plot indicates a distribution with a tail "lighter than the exponential" or even with finite end-point such as GPD with $\xi < 0$.

The displayed confidence limits are in all case pointwise and bilateral, and correspond to the confidence percents displayed which can be changed in the call. In most cases the confidence limits are approximate

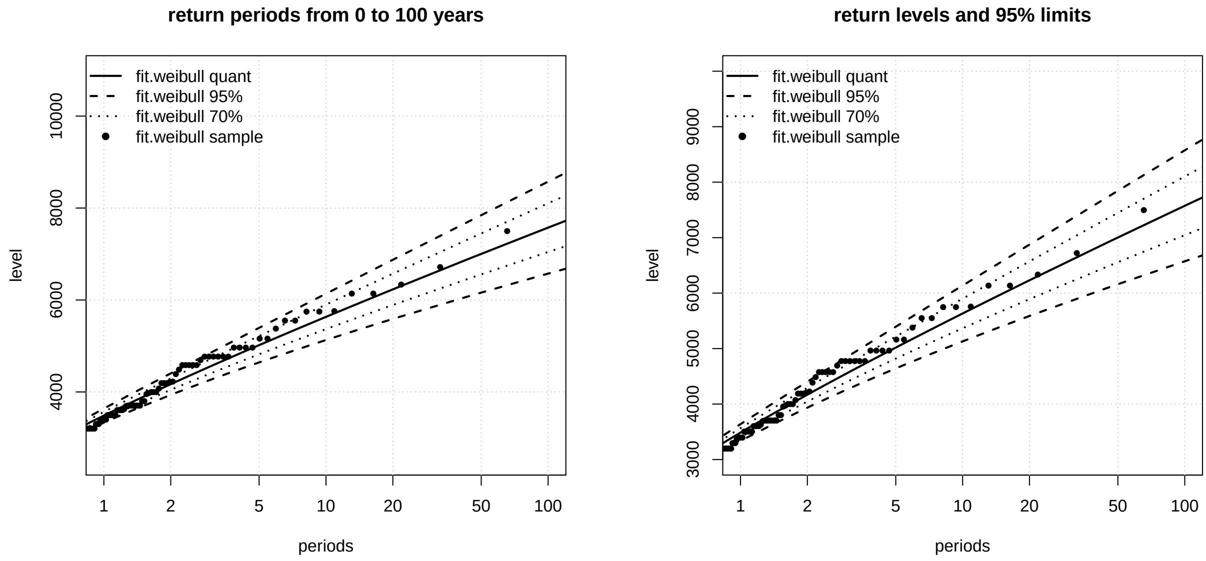


Figure 3.2: Changing the settings of the return level plot.

and obtained by using the *delta method* briefly described later. For some special cases with exponential distribution an exact inference is possible and used. The `infer.method` element in the list returned by `Renouv` provides information about this.

3.2.2 Plot method for `Renouv` objects

Once created with the `Renouv` function, an object of class `"Renouv"` can be used to (re)draw a return level plot and change some options. Useful changes concern the main title using the `main` argument, or axes labels `xlab`, `ylab`. Axis limits can also be set. For the return levels, this is done using the usual `ylim` argument. For the return periods, the limits are set using `Tlim` or `problim`. The first possibility works with a vector containing two return periods (in years); the second possibility requires a vector with two probabilities.

The two following code chunks produce the return level plots shown on figure 3.2. On left panel, we change the return periods axis limits.

```
> plot(fit.weibull,  Tlim = c(1, 100), main = "return periods from 0 to 100 years")
```

On the right panel we change both axes and the confidence level.

```
> plot(fit.weibull,
       Tlim = c(1, 100), ylim = c(3000, 10000),
       pct.conf = 95,
       main = "return levels and 95% limits")
```

Note that chosen percentage for the confidence limits `pct.conf = 95` must correspond to a value available in the object description. Otherwise, it is necessary to refit using `fitRenouv` with a suitable `pct.conf` argument.

3.3 Computational details

3.3.1 Maximum Likelihood theory

Estimation and inference in **Renext** mainly rely on the Maximum Likelihood (ML) theory. A relevant presentation can be found in Coles (2001, chap. 2) or in the *Further reading* references given there.

The standard application context of ML is when an ordinary sample i.e. n independent random variables X_i with the same distribution depending of an unknown vector $\boldsymbol{\theta}_X$ with density $f_X(x; \boldsymbol{\theta}_X)$. The likelihood function L is the joint density of the sample i.e.

$$L = \prod_{i=1}^{n} f_X(X_i;\, \boldsymbol{\theta}_X)$$

and the estimator $\widehat{\boldsymbol{\theta}}_X$ is the value of $\boldsymbol{\theta}_X$ maximising L. In some special cases the maximisation of L can have an explicit solution, but a numerical optimisation will generally be required. The ML theory warrants[1] the *asymptotic unbiasedness* and *asymptotic normality*: when n is large $\widehat{\boldsymbol{\theta}}_X$ has its expectation approximately equal to the true unknown $\boldsymbol{\theta}_X$, and it is approximately normally distributed.

The ML theory applies to more general situations where observations are no longer independent or can have different marginal distributions. This occurs when order statistics are used in the estimation as *historical data*.

The general principle of the `Renouv` function is to allow a large choice of distributions, yet trying to take advantage of the specific distribution/independence when possible. In most cases the maximisation of the likelihood is obtained using `optim` function of the `stats` package. When historical data are used they are considered as a complement to the ordinary data (exceedances) and two optimisations might be used.

3.3.2 Estimation and inference

The model uses a parameter vector $\boldsymbol{\theta} = \left[\lambda,\, \boldsymbol{\theta}'_X\right]'$ of length p formed with the HPP rate λ and the parameter vector $\boldsymbol{\theta}_X$ for the levels distribution.

When no historical data are used, the observed data consist in N events $[T_i,\, X_i]$ on a given period. Since events T_i and levels X_i are independent the likelihood is

$$L = \underbrace{\frac{(\lambda w)^N}{N!} e^{-\lambda w}}_{\text{events}} \times \underbrace{\prod_{i=1}^{N} f_X(X_i;\, \boldsymbol{\theta}_X)}_{\text{levels}}$$

where w is the time-length (i.e. the effective duration), and the log-likelihood is

$$\log L = N \log(\lambda w) - \lambda w - \log(N!) + \sum_{i=1}^{N} \log f_X(X_i;\, \boldsymbol{\theta}_X) \tag{3.1}$$

The ML estimation consists in two simple ML estimations: one for the events (rate estimation) and the other for levels. The ML estimate of the unknown rate λ is

$$\widehat{\lambda} = \frac{N}{w} = \frac{\text{number of events}}{\text{duration}}$$

Its variance is $\text{Var}[\widehat{\lambda}] = \lambda/w \approx \widehat{\lambda}/w$. Note that the number of events N is a *sufficient statistic* for λ: the events T_i are not used and the whole information they provide about λ is contained in N. The "X-part" of ML concerns an ordinary sample. The ML estimate $\widehat{\boldsymbol{\theta}}_X$ may be available in closed form in some cases (e.g. exponential).

When no historical data are used, it can be said that λ and $\boldsymbol{\theta}_X$ are orthogonal parameters. This is no longer true when historical data are used: the likelihood then takes a less favourable form (see below).

In a few cases with no historical data and favourable distribution (e.g. Weibull) it is possible to use the *expected* information matrix. But the general treatment in **Renext** is based on the *observed* information and the numerical derivatives. More precisely, the information matrix is obtained as the numerical hessian at convergence. The hessian can either be the element `hessian` returned by the `optim` function, or result form the use of the `hessian` function from `numDeriv` package: see the manual for more details.

[1] Under suitable *regularity conditions*.

3.3.3 Delta method

The *delta method* can be used to infer about a function[2] $\psi = \psi(\boldsymbol{\theta})$ of the parameter $\boldsymbol{\theta}$. For instance $\psi(\boldsymbol{\theta})$ can be the return period of a given level x (see 1.1). The transformed parameter estimate is $\widehat{\psi} = \psi(\widehat{\boldsymbol{\theta}})$. As a general result in the ML framework the transformed parameter estimate is asymptotically unbiased $\mathrm{E}[\widehat{\psi}] \approx \psi(\boldsymbol{\theta})$ and asymptotically normal with variance

$$\mathrm{Var}[\widehat{\psi}] \approx \boldsymbol{\delta}' \, \mathrm{Var}[\widehat{\boldsymbol{\theta}}] \, \boldsymbol{\delta}$$

where $\boldsymbol{\delta}$ is the gradient vector

$$\boldsymbol{\delta} = \frac{\partial \psi}{\partial \boldsymbol{\theta}} = \left[\frac{\partial \psi}{\partial \theta_1}, \, \frac{\partial \psi}{\partial \theta_2}, \, \dots, \, \frac{\partial \psi}{\partial \theta_p} \right]'$$

evaluated at $\widehat{\boldsymbol{\theta}}$, see Coles (2001, chap. 2).

Renext uses this approach with ψ taken as the level (or quantile) $x(T)$ corresponding to a given return period T. However the return level is related to a chosen probability of non-exceedance p (e.g. $p = 0.95$) which can be converted into a return period. Thus the relation is

$$T = \frac{1}{\lambda \times (1 - p)} \qquad F_X(x) = p$$

Since λ is unknown it is replaced by its ML estimation $\widehat{\lambda}$ and T is regarded as known. Thus the uncertainty about λ (usually small) is ignored in the relation between p and T. The gradient of the quantile function is computed numerically using a finite difference approximation.

3.3.4 Goodness-of-fit

As a general tool to assess the fit, the Kolmogorov-Smirnov (KS) test is computed in all cases.

The KS test normally requires a *completely specified* distribution for the null hypothesis while the *fitted* distribution is used here – thus generating a bias. In some special cases (normal, exponential) the bias could be corrected using an adaptation depending on the distribution as in Lilliefors test for the normal. However since the number of estimated parameters is small (usually 1 or 2 for the "exceedances part") the bias will be small provided that the number of exceedances is large enough, say 50 or more.

For some distributions such as exponential a specific test may be available. In the current version distribution-specific tests are limited to Bartlett's test of exponentiality.

Rounded measurements often lead to ties in the sample, which would without precaution generate a warning in the KS test. This can be avoided by "jitterizing" i.e. adding a small random noise to the observed values.

The graphical analysis of the fit using the return level plot is generally instructive. For exponential or Weibull exceedances, classical exponential or Weibull plot can also be drawn using the `expplot` and `weibplot` functions.

Note that when historical data are given, they are used during the estimation but not included in the empirical distribution in the KS test. In this case, the interpretation of the test needs further investigations.

3.4 Using historical data

3.4.1 Two types of historical data

Renext can use two kinds of historical data: *classical historical data* or "MAX" data, and *Over a Threshold Supplementary data*, or "OTS" data. In both cases, the data are structured in blocks and can be used only as complement to the main OT data which must continue to be provided.

MAX data complement the main OT data by r-largest blocks. Each block corresponds to a time interval of known duration w during which the r largest values are available. Blocks are assumed

[2] Smooth enough.

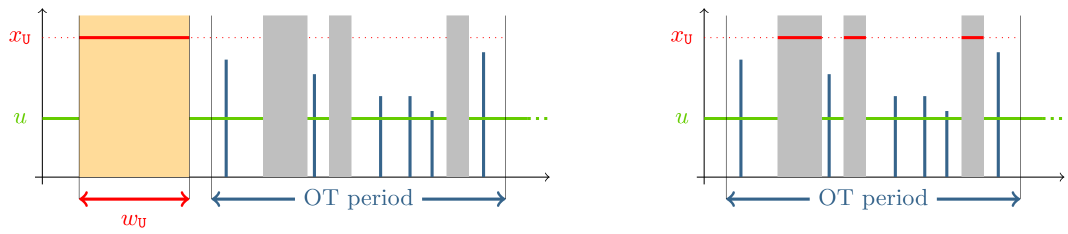


Figure 3.3: Unobserved level can provide information on an historical period (left) or on missing periods (right).

to be mutually disjoint and disjoint from the OT period. Neither the duration of blocks nor the number r of observations are assumed to be constant; hence each block b has a specified duration w_b and a number r_b of largest values.

OTS data complement the main OT data with other Over the Threshold data recorded on blocks with known duration and known exceedances. Again, blocks are assumed to be mutually disjoint and disjoint from the OT period and other historical blocks. For each such block b with known duration w_b, we must have a threshold u_b and all observations with levels exceeding u_b. The number r_b of such observations may be zero, in which case we may say that u_b is an *unobserved level*. The threshold u_b can not be smaller than the main threshold; it is sometimes called a *perception* threshold.

Unobserved levels (OTS data blocks with no observations) occur in some contexts where it is granted, or at least believed, that a given level say $x_{\mathtt{U}}$ was never exceeded during a period of time. For instance it can be granted that a river never flood over a given benchmark level during the last five centuries, or that the arch of a bridge was never reached since the construction. Such information has obviously a great potential impact on the estimation since it typically concerns very long periods, much longer than the observation period. If such an information exists, it can be used with the `Renouv` function. Note that the unobserved level can concern missing periods for OT data: although no data are available we may still know that no very high level occurred, see figure 3.3.

3.4.2 Likelihood

MAX data

Consider an historical "MAX" block of length $w_{\mathtt{H}}$. Let $Z_1 \geqslant Z_2 \geqslant \cdots \geqslant Z_r$ be the r largest observations. Their log-likelihood can be proved to be

$$\log L = r \log(\lambda w_{\mathtt{H}}) + \sum_{i=1}^{r} \log f_X(Z_i;\, \boldsymbol{\theta}_X) - \lambda w_{\mathtt{H}} \left[1 - F_X(Z_r;\, \boldsymbol{\theta}_X)\right] \tag{3.2}$$

When several blocks exist, they provide independent random vectors of observations with possibly different r and the log-likelihood is obtained by summing over blocks.

OTS data

The likelihood for an "OTS" block with threshold $u_{\mathtt{H}}$ is simpler to derive. According to the POT assumptions[3], the levels greater than $u_{\mathtt{H}}$ occur according to an HPP *thinning* the original HPP. This thinned process has rate $[1 - F_X(u_{\mathtt{H}})] \times \lambda$ because at each OT event, the level $u_{\mathtt{H}}$ can be exceeded with probability $1 - F_X(u_{\mathtt{H}})$. Let $w_{\mathtt{H}}$ be as before the block duration, and let $Z_1 \geqslant Z_2 \geqslant \cdots \geqslant Z_r$ be the r observations, with possibly $r = 0$. Up to an additive constant, the log-likelihood is

$$\log L = r \log(\lambda w_{\mathtt{H}}) + \sum_{i=1}^{r} \log f_X(Z_i;\, \boldsymbol{\theta}_X) - \lambda w_{\mathtt{H}} \left[1 - F_X(u_{\mathtt{H}};\, \boldsymbol{\theta}_X)\right] \tag{3.3}$$

This expression can be compared to (3.2). Replacing the block threshold $u_{\mathtt{H}}$ by the minimum observed value Z_r in the last formula leads to (3.2).

[3] See section 1.2.1 page 2.

When an OTS block contains no observation i.e. when $r = 0$, the log-likelihood (3.3) is simply

$$\log L = -\lambda w_{\mathtt{H}}\,[1 - F_X(u_{\mathtt{H}};\,\boldsymbol{\theta}_X)] \tag{3.4}$$

This is easily checked: on a period of length $w_{\mathtt{H}}$, the number of levels $> u_{\mathtt{H}}$ is Poisson with mean $\mu := [1 - F_X(u_{\mathtt{H}})] \times \lambda \times w_{\mathtt{H}}$. Hence the probability to observe no level $> u_{\mathtt{H}}$ is: $e^{-\mu}\mu^0/0! = e^{-\mu}$. Note that when only one OTS block is used with no observation and with $u_{\mathtt{H}}$ equal to the main threshold, the change in the log-likelihood is $-\lambda w_{\mathtt{H}}$ since then $F_X(u_{\mathtt{H}}) = 0$. This change is equivalent to that which would result from adding $w_{\mathtt{H}}$ to the effective duration for the main OT sample.

Remarks

Assume that we have only one historical block of type "MAX" and that it only contains the maximum Z_1 i.e. has $r = 1$. The contribution of the block to the log-likelihood (3.2) is

$$\log L = \log(\lambda w_{\mathtt{H}}) + \log f_X(Z_1;\,\boldsymbol{\theta}_X) - \lambda w_{\mathtt{H}}\,[1 - F_X(Z_1;\,\boldsymbol{\theta}_X)]$$

At the right hand side, the third term is identical to (3.4) with an unobserved level $u_{\mathtt{H}} = Z_1$ and a period length $w_{\mathtt{H}}$. The sum of the two first terms at right side is the extra contribution that would be added to the log-likelihood of the OT data if a new OT observation with level Z_1 had been done without changing the main OT period duration. Therefore, the same likelihood/results are obtained in the two following approaches

- Specify an historical MAX block of length $w_{\mathtt{H}}$ with $r = 1$ and level Z_1.
- Join the observed maximum Z_1 to the OT levels X_i, and specify that the level $u_{\mathtt{H}} := Z_1$ was never reached during a OTS block of length $w_{\mathtt{H}}$.

The second approach might seem natural to practitioners.

It can be shown that when historical data are used (MAX, OTS or both) the likelihood can be concentrated with respect to the rate λ, thus leading to the maximisation of function $\log L_{\mathrm{c}}(\boldsymbol{\theta}_X)$ depending on $\boldsymbol{\theta}_X$ only.

3.4.3 Example: using Garonne data

Specifying historical data

As said before, the `Garonne` dataset[4] contains historical data of type MAX, which can be used in the estimation. The data are described in the section 1.3.4 page 6. The historical data corresponds here to one block, and the following levels

```
> Garonne$MAXdata$Flow
 [1] 7500 7400 7000 7000 7000 6600 6500 6500 6400 6300 6300 6200
```

The duration is given in `Garonne$MAXinfo$duration` with value 143.09 years.

As a general rule, the historical data must be passed as a *list* of numeric vectors, each vector corresponding to one block. The (effective) durations are given as a numeric vector with the *same length as the list*. For the "MAX" case, the formal arguments to use are `MAX.data` (list) and `MAX.effDuration` (numeric vector).

Since the data corresponds here to one block, the list `MAX.data` contains only one vector and the vector `MAX.effDuration` is of length one. The two following fits produce the return level plots shown in figure 3.4.

```
> fit.exp.H <- Renouv(x = Garonne$OTdata$Flow,
                      effDuration = 65, threshold = 2500,
                      MAX.data=  list(Garonne$MAXdata$Flow),
                      MAX.effDuration = Garonne$MAXinfo$duration,
                      distname.y = "exponential",
                      main = "Garonne data, \"exponential\" with MAXdata")
```

[4] Provided as an object of class `"Rendata"`.

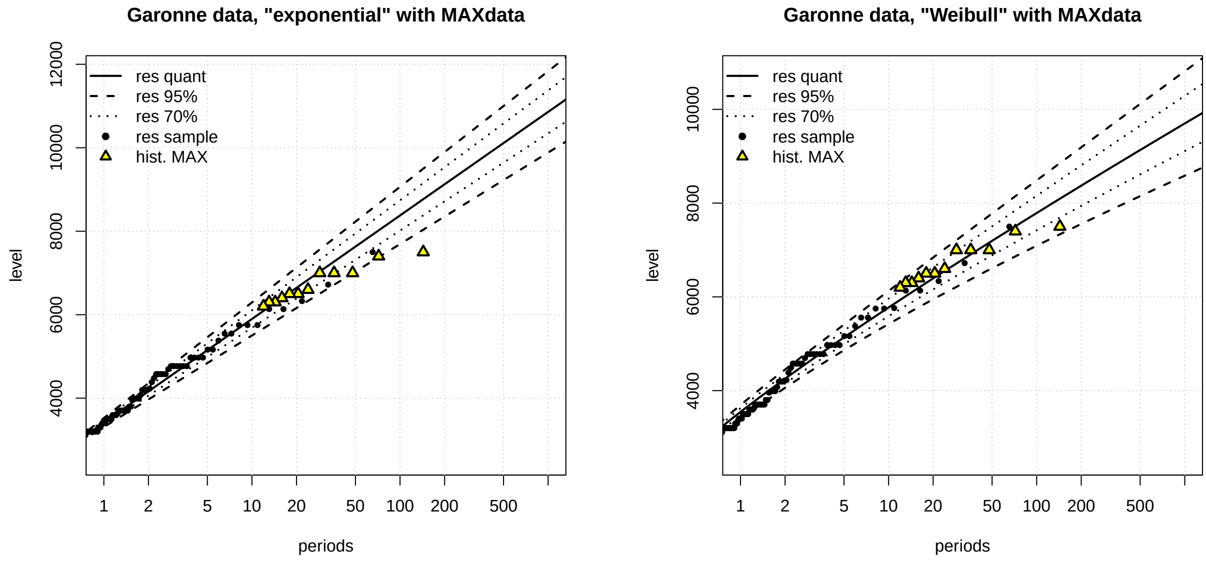


Figure 3.4: Return level plots for the example `Garonne` with two distributions for exceedances and historical data.

```
> fit.weib.H <- Renouv(x = Garonne$OTdata$Flow,
                       effDuration = 65, threshold = 2500,
                       MAX.data=  list(Garonne$MAXdata$Flow),
                       MAX.effDuration = Garonne$MAXinfo$duration,
                       distname.y = "weibull",
                       main = "Garonne data, \"Weibull\" with MAXdata")
```

The exponential fit is only slightly modified by the use of historical data. As said before, the parameter λ and $\boldsymbol{\theta}_X$ are no longer orthogonal when historical data are used

```
> fit.exp.H$corr
          lambda      rate
lambda 1.0000000 0.1705388
rate   0.1705388 1.0000000
```

Plotting positions

The historical data are displayed on the return level plot (see figure 3.4) as follows.

Consider a MAX block with r largest observations Z_k in decreasing order and with duration $w_{\mathtt{H}}$. Using the "non historical" data, we can give a prediction $\widetilde{N}_{\mathtt{H}}$ for the unknown number $N_{\mathtt{H}}$ of events on the historical period. A natural choice is $\widetilde{N}_{\mathtt{H}} = \widetilde{\lambda}\, w_{\mathtt{H}}$ where $\widetilde{\lambda}$ is the events rate on the OT period. Then the point Z_k will be associated to the probability of exceedance $1 - \widetilde{F} = k/(\widetilde{N}_{\mathtt{H}} + 1)$. For the largest value Z_1, we thus have $1 - \widetilde{F} = 1/(\widetilde{N}_{\mathtt{H}} + 1)$. When several historical blocks are available, the same principle can be used block by block.

For OTS data the principle is the same, except for an OTS block with no observation – that is, for the *unobserved level* case. Then the unobserved level (or the never exceeded threshold) $u_{\mathtt{H}}$ is shown as an horizontal segment with return periods ranging from 0 to $w_{\mathtt{H}}$.

Fitting from Rendata objects

Recall that a S3 class `"Rendata"` is defined in **Renext** in order to represent composite data with optional historical data. An object of class `"Rendata"` contains an OT sample, but also embeds useful pieces of information such as the effective duration for the OT sample or the variable name. It seems sensible to use these indications in a POT model by simultaneously passing them as formal arguments to the fitting function. For instance, when the OT sample of a "Rendata" object is used in a fit, the effective duration could consistently be taken from this object.

The Renouv can indeed be used by giving an x formal with class "Rendata" instead of a numeric vector.

```
> fitWithObj <- Renouv(x = Garonne)
```

Note that the threshold for the main OT is taken from the Rendata object, and will generally be too small for the POT modelling. It can be changed

```
> fitWithObj1 <- Renouv(x = Garonne, threshold = 3000)
```

Similarly, the effective duration of the object could be shortcut by giving a effDuration formal argument in the call. The distribution of the exceedances can be set in the usual way. In all cases, the summary method should be invoked on the fitted object.

Using "Rendata" objects passed as x formals can simplify the task of fitting many datasets files if these are read with the readXML function.

3.5 Fixing parameter values

3.5.1 Problem

In some situations one may want to fix one or several parameters in the distribution of exceedances and still perform a ML estimation for the remaining parameters. For instance, the shape of a Weibull distribution can be fixed while the scale is to be estimated. This can be viewed as a radical bayesian scheme with the fixed parameters receiving an 'ultra-informative' Dirac prior.

Renext supports fixed parameters, with some limitations. In the current version, the HPP rate parameter λ **can not be fixed**, and **at least one parameter must be estimated in the exceedance part**. Thus the full model must have at least two non-fixed parameters.

The specification of the fixed parameter is done using the fixed.par.y formal argument in Renouv. Its value must be a named list with names in the distribution parnames. As a general rule[5], the non-fixed (estimated) parameters must be given using the start.par.y arg with a similar list value.

3.5.2 Example

The fixed parameter option can work with or without historical data in the same manner.

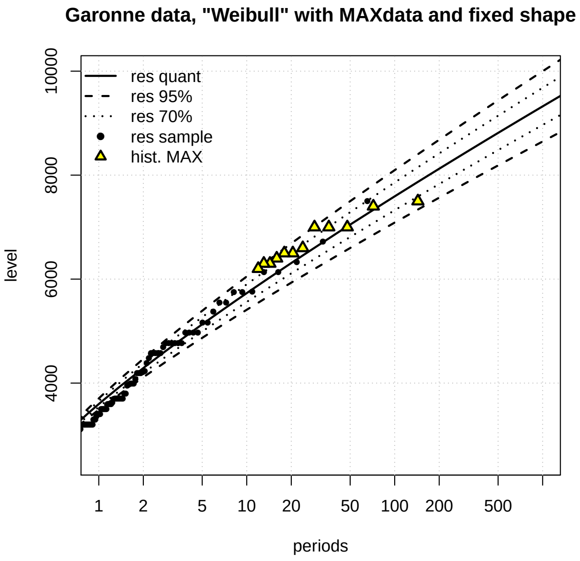
```
> fit.weib.fixed.H <-
    Renouv(x = Garonne$OTdata$Flow,
           effDuration = 65, threshold = 2500,
           MAX.data = list(Garonne$MAXdata$Flow),
           MAX.effDuration = Garonne$MAXinfo$duration,
           distname.y = "weibull",
           fixed.par.y = list(shape = 1.2),
           start.par.y = list(scale = 2000),
           trace = 0,
           main = "Garonne data, \"Weibull\" with MAXdata and fixed shape")
```

```
> fit.weib.fixed.H$estimate

     lambda       shape       scale
   2.381991    1.200000 1235.075301
```

With some distributions such as the SLTW some parameters *must* be fixed. Here the shift parameter delta is fixed to $\delta = 2000$ m^3/s meaning that we believe that exceedances over $u - \delta = 500$ are Weibull, even if we only know exceedances over the threshold $u = 2500$ m^3/s.

[5]In some special cases, this is unnecessary but harmless.



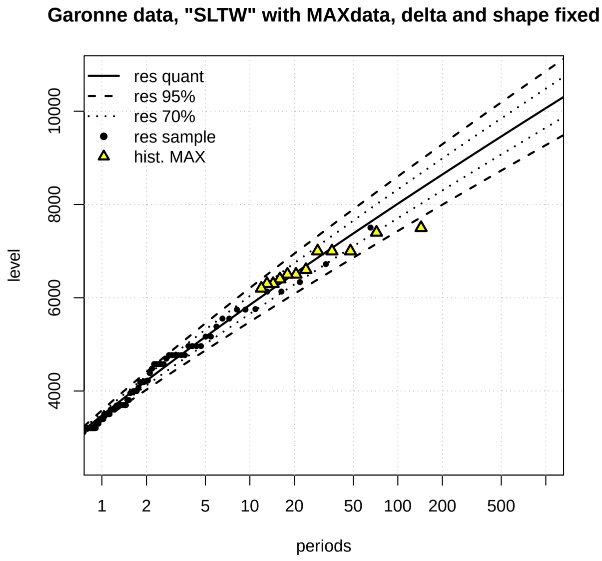


Figure 3.5: Return level plots for the example `Garonne` with two distributions with **fixed parameters** (and historical data).

```
> fit.SLTW.H <-
    Renouv(x = Garonne$OTdata$Flow,
           effDuration = 65, threshold = 2500,
           MAX.data = list(Garonne$MAXdata$Flow),
           MAX.effDuration = Garonne$MAXinfo$duration,
           distname.y = "SLTW",
           fixed.par.y = list(delta = 2000, shape = 1.2),
           start.par.y = list(scale = 2000),
           main = "Garonne data, \"SLTW\" with MAXdata, delta and shape fixed")
```

When some parameters are fixed the covariance contains structural zeros, and consequently the correlation matrix contains non-finite coefficients.

```
> fit.SLTW.H$cov
            lambda delta shape       scale
lambda   0.03493404     0     0   -2.282315
delta    0.00000000     0     0    0.000000
shape    0.00000000     0     0    0.000000
scale   -2.28231494     0     0 5674.167668
```

Appendix A

The “renouvellement” context

A.1 Marked point process

The *méthode du renouvellement* uses a quite general marked process $[T_i, X_i]$ for events and levels. As in 1.2.1 the two sequences ”events” and ”levels” are assumed to be independent, and the X_i are assumed to be independent and identically distributed with continuous distribution $F_X(x)$.

An alternative equivalent description of the events occurrence is through the associated *counting process* $N(t)$. This describes the joint distribution for the the numbers of events $N(t_k) - N(s_k)$ on an arbitrary collection of disjoint intervals (s_k, t_k). Although the most important and clearest context is the HPP, the theory can be extended to cover non-poissonian Lévy counting processes $N(t)$ e.g. Negative Binomial. However, the Negative Binomial Lévy Process implies the presence of multiple (simultaneous) events.

A.2 Some results

A.2.1 Compound maximum

Consider an infinite sequence of independent and identically distributed random variables X_k with continuous distribution $F_X(x)$. The maximum

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

has a distribution function given by $F_{M_n}(x) = F_X(x)^n$. Now let N be a random variable independent of the X_k sequence and taking non-negative integer values. The ”compound maximum”

$$M = \max(X_1, X_2, \ldots, X_N)$$

is a random variable with a mixed type distribution: it is continuous with a probability mass corresponding to the $N = 0$ case which can be considered as leading to the certain value $M = -\infty$. The distribution of M can be derived from that of X_k and N. Using $\Pr(M \leqslant x \mid N = n) = F_X(x)^n$ and the total probability formula we get

$$F_M(x) = \sum_{n=0}^{\infty} F_X(x)^n \Pr\{N = n\} = h_N[F_X(x)] \tag{A.1}$$

where $h_N(z) = \mathrm{E}\left(z^N\right)$ is the generating function of N.

When N has a Poisson distribution with mean $\mu_N = \lambda w$ the generating function is given by $h_N(z) = \exp\{-\mu_N\,[1 - z]\}$ and

$$F_M(x) = \exp\{-\lambda w\,[1 - F_X(x)]\} \tag{A.2}$$

When $F_X(x)$ is GPD it can be shown that M is[1] GEV see later.

[1] Up to its probability mass.

For large return levels x, we have $F_X(x) \approx 1$. The generating function $h_N(z)$ for $z = 1$ has a value $h_N(z) = 1$ and a first derivative $h'_N(z) = \mathrm{E}(N)$, leading to

$$1 - F_M(x) \approx \mathrm{E}(N)\,[1 - F_X(x)] \tag{A.3}$$

Equivalently

$$F_M(x) \approx F_X(x)^{\mathrm{E}(N)} \tag{A.4}$$

which tells that for large return levels, the distribution of M is approximately that of the maximum of $\mathrm{E}(N)$ independent X_k. Both formula (A.3) and (A.4) tell that the distribution of N only influences large return periods through its expectation. Consequently there is little point in choosing a non-Poisson distribution for N as far as the interest is focused on large return periods.

From formula (A.4) and the asymptotic behaviour of the maximum of n independent and identically distributed random variables (see B.1 later), it appears that when $\mathrm{E}(N)$ is large the distribution of M will generally be close to a suitably scaled GEV distribution.

A.2.2 Special cases

A case with special interest is when N is Poisson with mean $\mu_N = \lambda w$ and X has a Generalised Pareto Distribution (GPD). Then M follows[2] a Generalised Extreme Values (GEV) distribution.

Consider first the exponential case $F_X(x) = 1 - e^{-(x-\mu)/\sigma}$ for $x \geqslant \mu$. Then (A.2) writes as

$$F_M(x) = \exp\left\{-\lambda w\, e^{-(x-\mu)/\sigma}\right\}$$

which using simple algebra can be identified as the Gumbel distribution function with parameters $\mu^\star = \mu + \sigma\log(\lambda w)$ and $\sigma^\star = \sigma$.

In the general case where $F_X(x)$ corresponds to the GPD, $F_X(x) = 1 - [1 + \xi(x-\mu)/\sigma]_+^{-1/\xi}$ we have for $x \geqslant \mu$

$$F_M(x) = \exp\left\{-\lambda w\,[1 + \xi(x-\mu)/\sigma]_+^{-1/\xi}\right\}$$

which can be identified as `GEV`($\mu^\star$, $\sigma^\star$, ξ) with parameters $\mu^\star$ and $\sigma^\star$ depending on μ and σ.

Using this formalism we can derive the distribution of the maximum of the X_k on an arbitrary period of length w.

A.3 Return periods

In the general marked process context described above, the return period of a given level x can be defined using the thinned process $[T_i,\, X_i]$ of events with level exceeding x i.e. with $X_i > x$. The return period will be the expectation $T_X(x)$ of the interevent in the thinned process. In the rest of this section, we assume that events occur according to a HPP with rate $\lambda > 0$. Due to the independence of events and levels, the thinned event process also is an HPP with rate $\lambda(x) = \lambda[1 - F_X(x)]$. The return period is then given by

$$T_X(x) = \frac{1}{\lambda\,[1 - F_X(x)]}$$

Actually the interevent distribution is exponential with expectation $1/\lambda(x)$.

Still using the same probabilistic framework, we may consider the sequence of annual maxima or more generally the sequence M_n of maxima for successive non-overlapping time blocks with the same duration $w > 0$. The random variables M_n are independent with a common distribution $F_M(x)$ that can be determined as it was done in the last section. In this "block" context, the return period of a level x naturally expresses as a (non-necessarily integer) multiple of the block duration. Thus if $F_M(x) = 0.70$ i.e. if the level x if exceeded with 30% chance within a block, the return period is $1/0.3 \approx 3.33$ expressed in block duration unit. More generally, the *block* return period of the level x will be computed as

$$T_M(x) = \frac{w}{1 - F_M(x)} = \frac{\text{block duration}}{\text{prob. that } M \text{ exceeds } x} \tag{A.5}$$

[2] Up to its probability mass in $-\infty$.

A major difference between the two return periods $T_X(x)$ and $T_M(x)$ is that the level x can be exceeded several times within the same block, especially for small x. This difference may make ambiguous some statements about yearly return periods or yearly risks. For instance, the level x with a 100 years return period $T_X(x)$ is very likely to be exceeded twice or more within a given century[3].

Using the relation (A.2) between the distributions $F_X(x)$ and $F_M(x)$, the relation (A.5) becomes

$$T_M(x) = \frac{w}{1 - \exp\{-\lambda w\,[1 - F_X(x)]\}} \tag{A.6}$$

In practice, the interest will be focused on large levels x. In the expression at the denominator we may then use the approximation $1 - e^{-z} \approx z$ for small z, leading to $T_M(x) \approx T_X(x)$. Moreover the inequality $1 - e^{-z} \leqslant z$ for $z \geqslant 0$ shows that $T_M(x) \geqslant T_X(x)$ for all x. Using $1 - e^{-z} \approx z - z^2/2$, we even find a better approximation for moderately large levels x

$$T_M(x) \approx T_X(x) + \frac{w}{2}$$

The presence of the half-block length $w/2$ can be viewed as a rounding effect.

[3]Within a given century, the number $N(x)$ of events with levels $X_i > x$ is then Poisson with mean 1. Thus $\Pr\{N(x) = 0\} \approx 0.37$ and $\Pr\{N(x) > 1\} \approx 0.26$.

Appendix B

Distributions

B.1 Asymptotic theory and the GEV distribution

B.1.1 An important result

A central result of Extreme Values theory is the Fisher-Tippett-Gnedenko theorem below. The following conventions or definitions are used.

- Two probability distributions $F(x)$ and $G(x)$ are of same type when $G(x) = F(ax + b)$ for some constants $a > 0$ and b. All distributions of a given type are often written as $F_0([x - \mu]/\sigma)$ where $F_0(z)$ is a chosen member of the type, μ (location) and $\sigma > 0$ (shape) are parameters. The parameters μ and σ are not necessarily the mean nor the standard deviation.
- The notation z_+ is for the positive part of a number z, that is $z_+ = \max(z, 0)$.

Theorem (Fisher-Tippett-Gnedenko). *Let X_n be a sequence of independent and identically distributed random variables, and let $M_n = \max(X_1, X_2, \ldots, X_n)$. If there exists two sequences b_n and $a_n > 0$ such that $(M_n - b_n)/a_n$ has a non-degenerate limiting distribution $G(z)$, then that limiting distribution must be one of the following three types*

$$\begin{array}{ll} G(z) = \exp\{-e^{-z}\} & \textit{Gumbel or type I} \\ G(z) = \exp\{-z_+^{-\alpha}\} & \textit{Fréchet or type II} \\ G(z) = \exp\{-(-z)_+^{\alpha}\} & \textit{Weibull (reversed) or type III} \end{array}$$

where $\alpha > 0$ is a parameter for types II or III.

For each type, the distribution depends on μ and $\sigma > 0$ and possibly of $\alpha > 0$. E.g. the general Gumbel distribution is

$$G(x) = \exp\{-\exp[-(x - \mu)/\sigma]\}.$$

The third distribution corresponds to values $z \leqslant 0$ and is often called Weibull. This may create a confusion with the ordinary Weibull described later. A preferable appellation is *reversed Weibull.*

Each of the three possible limiting distributions is *max-stable* i.e. is closed for the maximum of independent and identically distributed random variables. For example if X_i are independent with the same Gumbel distribution, then their maximum M_n is also of Gumbel type.

The three possible limit distributions are fairly different. Some mathematical criteria allow to say whether a given distribution of X_k is in the *domain of attraction* of Gumbel, Fréchet or (reversed) Weibull. Some usual examples are found in the book of Kotz and Nadarajah (2005, chap. 1) and table B.1 gives the domains of attraction for the main distributions used in **Renext**. Broadly speaking, distributions with exponentially decaying upper tail (such as normal, exponential, gamma) fall in the domain of attraction of Gumbel. The Fréchet domain attracts heavy-tailed distributions (Pareto, Cauchy).

distribution of X_i	**limit of M_n**
exponential	Gumbel
Weibull	Gumbel
gamma	Gumbel
GPD $\xi = 0$ GPD $\xi > 0$ GPD $\xi < 0$	Gumbel Fréchet reversed Weibull
log-normal	Gumbel
finite mixture of exponentials	Gumbel
Pareto	Fréchet
Cauchy	Fréchet

Table B.1: Limit distribution for the maximum of a large number of independent levels X_i.

B.1.2 Generalised Extreme Values

The three types of the theorem above can be considered as special cases of the *Generalised Extreme Value* distribution depending of a shape parameter ξ

$$G(z) = \exp\left\{ -\left[1 + \xi\, z\right]_+^{-1/\xi} \right\}.$$

The sign of the shape parameter ξ is essential. When $\xi > 0$ we retrieve the Fréchet above up to a translation of z. For $\xi < 0$ we get the reversed Weibull up to a translation of z. When $\xi = 0$ the power $[1 + \xi\, z]^{-1/\xi}$ is to be replaced by its limit for $\xi \to 0$ which is e^{-z} and $G(z)$ is the Gumbel distribution function above.

Using a linear transform $z = (x - \mu)/\sigma$ with arbitrary μ and $\sigma > 0$ all distributions of the GEV type are obtained as

$$F(x) = \exp\left\{ -\left[1 + \xi\, \frac{x - \mu}{\sigma}\right]_+^{-1/\xi} \right\}. \tag{B.1}$$

This distribution is named GEV with scale parameter μ and shape parameter $\sigma > 0$, and it will be denoted as `GEV`(μ, σ, ξ). It is defined on the set of values x for which the bracketed expression within [ ] in (B.1) is non-negative that is

$\xi < 0$	$\xi = 0$	$\xi > 0$
$-\infty < x \leqslant \mu - \sigma/\xi$	$-\infty < x < +\infty$	$\mu - \sigma/\xi \leqslant x < +\infty$

Grouping the three distributions may be thought of as a purely formal trick. However, since the GEV distribution is regular at $\xi = 0$ we have a parametric family in the usual sense, with a parameter ξ. Thus it makes sense to estimate the parameter ξ without specifying its sign, or to give a confidence interval including the value $\xi = 0$. Note that the support of the distribution depends on the parameters and thus that Maximum Likelihood (ML) theory must be invoked with care.

B.1.3 Implication in POT

The Fisher-Tippett-Gnedenko theorem suggests that the GEV distribution should be systematically used to describe block maxima.

The implication in POT and the marked process context is less clear. When a large enough threshold u is chosen, the observations X_i exceeding u might be thought of as maxima of unobserved independent variables, suggesting the use of a three parameter GEV distribution with censoring $X_i > u$. Fortunately, the conditional GEV is approximately a Generalised Pareto Distribution (GPD) with only two parameters, thus the standard POT can be used, see B.3.2.

This justification is strengthened by the compound maximum results given in A.2 and the special cases A.2.2.

B.2 Probability distributions in POT

B.2.1 Levels vs exceedances

POT methods fit a distribution to the exceedances $Y_i = X_i - u$ over a fixed threshold u. The exceedances are positive by construction and might contain small values since the threshold will generally be taken greater than the mode of X.

In the rest of this section the letter X will be used for a level while Y is used for a positive exceedance random variable. The densities and distribution functions of X will be denoted as $f_X(x)$ and $F_X(x)$ while the Y subscript is used for Y. Thus

$$f_X(x) = f_Y(x-u), \qquad f_Y(y) = f_X(y+u).$$

For the distribution fitted in POT the threshold u *is not a parameter* to be estimated. Yet the probability functions for level X can have a location parameter. R functions used for Y can also have a location parameter with suitable default value for it.

B.2.2 Some indicators

The *coefficient of variation* CV of a positive random variable Y is the ratio of the standard deviation to the mean

$$\mathrm{CV} = \sqrt{\mathrm{Var}(Y)}/\mathrm{E}(Y). \tag{B.2}$$

Comparing this theoretical CV to its empirical equivalent $\widehat{\mathrm{CV}}$ is often instructive, keeping in mind that $\widehat{\mathrm{CV}}$ is subject to sampling flucuation. For an exponential distribution we have $\mathrm{CV} = 1$; a mixture of several exponentials corresponds to $\mathrm{CV} > 1$. When fitting distributions from the Pareto families, comparing $\widehat{\mathrm{CV}}$ to 1 will often be essential, see B.3.2 page 35 later.

B.2.3 Some useful probability functions

Several probability functions provide useful insights about the upper tail of a given distribution. Their name is related to *survival analysis* where the random variable of interest is the lifetime Y of a subject or item. The relation with POT is: increasing the POT threshold u is equivalent to selecting subjects still alive at "time" u.

The *survival function* value $S(y)$ is the probability $\Pr\{Y > y\} = 1 - F(y)$. The *hazard function* $h(v)$ is defined by

$$h(v)\,\mathrm{d}v = \Pr\left[v < Y \leqslant v + \mathrm{d}v \mid Y > v\right], \qquad v \geqslant 0$$

corresponding to the notion of instantaneous death rate. An usual equivalent definition is $h(v) = f(v)/S(v)$. In survival analysis, hazards are usually non-decreasing since a decreasing hazard would mean a "rejuvenation" effect. Yet in POT modelling, distributions often have decreasing hazards. A decreasing hazard implies the presence of a thick upper tail since rejuvenating subjects tend then to have a very long life.

The *mean residual life* MRL (or mean excess life) is defined as

$$\mathrm{MRL}(v) = \mathrm{E}\left(Y - v \mid Y > v\right), \qquad v \geqslant 0.$$

While a decreasing $\mathrm{MRL}(v)$ may seem natural, a distribution with long tail such as GPD can have an increasing mean residual life.

Another meaningful function is the *cumulative hazard* $H(y)$

$$H(y) = -\log S(y) = \int_0^y h(z)\,\mathrm{d}z, \qquad y \geqslant 0.$$

Increasing and decreasing hazards $h(y)$ are respectively equivalent to convex and concave cumulative hazards $H(y)$. When the distribution function $F(y)$ is plotted on an exponential plot, the ordinate used is in fact $H(x)$, see page 8. The concavity of the resulting curve is that of $H(y)$, and hence is related to the variation of $h(y)$. Distributions with increasing hazard $h(y)$ will give a convex (upward concave) curve on the exponential plot while a decreasing $h(y)$ leads to a concave (downward) one. The same effect is observed for the exponential return level plot but with axes exchanged hence with opposite concavity.

An alternative to the quantile function $q_X(p)$ of X is the following *return level function*. Consider an independent and identically distributed sequence X_i with survival $S_X(x)$; for a given $m > 1$ the value x_m that is exceeded on average once every m observations is given by the equation

$$S_X(x_m) = 1/m \qquad (m > 1) \tag{B.3}$$

and it can be called the return level with period m (or m-return level). This is an increasing function of m with limit for large m the upper end-point of the distribution of X. For many distributions the solution of (B.3) exist in closed form. In the POT context where levels X_i are observed on a rate of λ events by years, the value of m in (B.3) is to be divided by the rate λ to obtain the corresponding period T. Then x_m is the return level corresponding to period $T := m/\lambda$.

Since $1/m = S_X(x_m)$, we have $\log m = H_X(x_m)$. Thus plotting points $[\log m,\, x_m]$ i.e. points $[m,\, x_m]$ with a log scale for the first axis (return periods) is equivalent to plotting points $[x,\, H_X(x)]$, but with the two axes exchanged.

B.3 Distributions in Renext

B.3.1 Exponential

Definition

The exponential distribution has a survival function $S(y)$ and a density $f(y)$ given by

$$S(y) = e^{-\nu y}, \qquad f(y) = \nu\, e^{-\nu y}, \qquad y \geqslant 0 \tag{B.4}$$

where $\nu > 0$ is a parameter called *rate*.

Properties

The equation $S(y) = 1/m$ giving the "m years return level" has the explicit solution $y_m = \log(m)/\nu$.

The exponential distribution has constant hazard rate − a fact known as the "memorylessness property". It therefore also has a constant mean residual life.

The exponential is a special case of several families: Weibull (with shape $\alpha = 1$), GPD (with shape $\xi = 0$) and gamma (with shape $\alpha = 1$). For these three families, the shape parameter is in one-to-one relation with the coefficient of variation CV which can take values smaller or larger than 1. Within the three families, the exponential is characterized by $\mathrm{CV} = 1$.

The exponential distribution is closely related to Gumbel distribution. If Y is exponential then $V = -\log Y$ is Gumbel.

Estimation and inference

The exponential distribution has a well known ML inference from an ordinary sample Y_i of size n.

The ML estimator for ν is the inverse of the sample mean $\widehat{\nu} = 1/\bar{Y}$. Up to a scaling factor the exponential distribution is nothing but the $\chi^2(2)$ with two degrees of freedom. More precisely $2\nu\, Y_i \sim \chi^2(2)$. Multiplying the sum $\sum_i Y_i = n\,\bar{Y}$ by 2ν gives a "pivotal" quantity $V = 2\nu \times n\,\bar{Y}$ having a $\chi^2(2n)$ distribution. Since $V = 2n\,\nu/\bar{\nu}$, an exact confidence interval at the level $1-\alpha$ for ν is obtained as

$$\frac{\chi^2_{1-\alpha/2}}{2n} \times \widehat{\nu} \leqslant \nu \leqslant \frac{\chi^2_{\alpha/2}}{2n} \times \widehat{\nu}$$

where χ^2_α is the upper quantile for the $\chi^2(2n)$ distribution[1]. Exact confidence intervals are similarly derived for the distribution $F(y)$ with given y or for a m-return level y_m with m given.

[1] $\Pr\left\{\chi^2(2n) > \chi^2_\alpha\right\} = \alpha$

Goodness-of-fit

A specific goodness-of-fit test for the exponential distribution is sometimes called Bartlett's (or Moran's) test of exponentiality. The test statistic B_n involves the sample mean $\overline{Y}$ as well as the sample mean $\overline{\log Y}$ of the logged Y_i

$$B_n = b_n \times \left\{\log \bar{Y} - \overline{\log Y}\right\}, \qquad b_n = 2n \times \{1 + (n+1)/(6n)\}^{-1}.$$

Under the null hypothesis we have approximately $B_n \sim \chi^2(n-1)$ and a two-sided test is in order.

Remind that the goodness-of-fit can also be evaluated using a graphical analysis with an exponential plot.

Use in Renext

The exponential can be used in **Renext** under the two names `"exponential"` and `"exp"`. In both cases, the rate parameter ν of (B.4) is named `rate`. In the `Renouv` function, the choice of the distribution name among the two possible ones for the exponential has consequences.

- Using `distname.y = "exponential"` (which corresponds to the default value), the estimation and inference will be specific to the exponential. The test of exponentiality is computed and displayed by the `summary` method for the fitted object. When no historical data are used, the exact inference described above is used both for the parameter and the return levels.
- Using `distname.y = "exp"`, the distribution of the **stats** package is used in black-box mode, as it would be with any other available distribution. Thus the inference on the parameter and the return levels is based on the asympotic normality and the delta method.

The first possibility should obviously be preferred. In the second case, the likelihood is maximised numerically, and an initial value must be given using the `start.par.y` argument.

B.3.2 Generalised Pareto GPD

Definition

The Generalised Pareto Distribution (GPD) depends on three parameters μ (location), $\sigma > 0$ (scale) and ξ (shape). When $\xi \neq 0$, the survival function $S(y)$ and the density function $f(y)$ are given by

$$S(x) = \left[1 + \xi\,\frac{(x-\mu)}{\sigma}\right]_+^{-1/\xi} \qquad f(x) = \frac{1}{\sigma}\left[1 + \xi\,\frac{(x-\mu)}{\sigma}\right]_+^{-1/\xi-1} \qquad x \geqslant \mu \tag{B.5}$$

while the limit for $\xi \to 0$ is to be used for $\xi = 0$

$$S(x) = e^{-(x-\mu)/\sigma} \qquad f(x) = \frac{1}{\sigma}\,e^{-(x-\mu)/\sigma} \qquad x \geqslant \mu$$

which is a shifted exponential distribution with rate $1/\sigma$.

The distribution is defined for the values x with $x \geqslant \mu$ and $1 + \xi\,(x-\mu)/\sigma \geqslant 0$, that is

$\xi < 0$	$\xi = 0$	$\xi > 0$
$\mu \leqslant x \leqslant \mu - \sigma/\xi$	$\mu \leqslant x < +\infty$	$\mu \leqslant x < +\infty$

The value of the shape parameter ξ has a very strong influence, see figure B.1.

- When $\xi < 0$ the distribution has a finite upper end-point. As a special case, the uniform distribution is obtained with $\xi = -1$. The density function is decreasing for $-1 < \xi < 0$.
- When $\xi > 0$ the density is decreasing. The distribution tail thickens as ξ increases.

For most practical applications, the range of values for ξ is $(-0.5,\, 0.5)$.

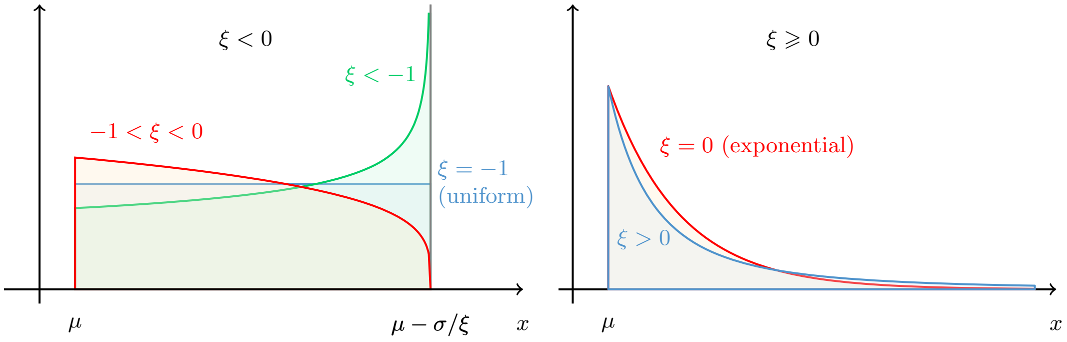


Figure B.1: GPD densities for $\xi < 0$ (left) and $\xi \geqslant 0$ (right). In the $\xi < 0$ case, the parameters are chosen in order to give the same support, i.e. μ and $-\sigma/\xi$ are kept constant.

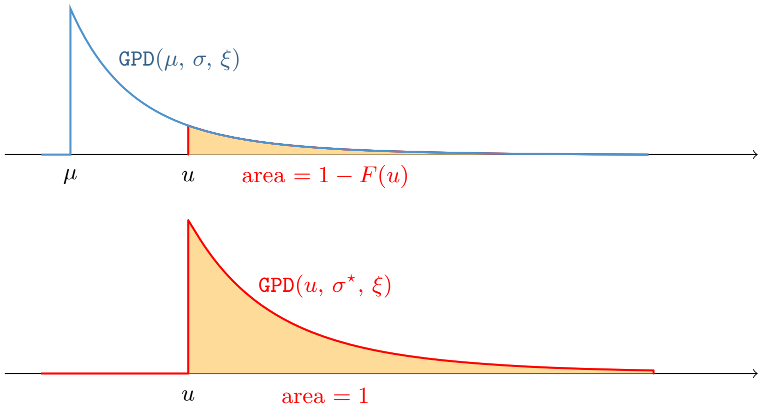


Figure B.2: "Stability for exceedances" of the GPD family.

Properties

The GPD has a finite expectation when $\xi < 1$ and a finite variance when $\xi < 1/2$ then given by

$$\mathrm{E}(X) = \mu + \frac{\sigma}{1-\xi}, \qquad \mathrm{Var}(X) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}, \qquad \mathrm{CV}(Y) = \frac{1}{\sqrt{1-2\xi}}.$$

The shape parameter ξ can be related to the coefficient of variation. Note that $\xi > 0$ gives $\mathrm{CV}(Y) > 1$.

For $m > 1$ the return level with period m of (B.3) is

$$x_m = \mu + \sigma\, \left[m^\xi - 1\right] / \xi$$

It can be remarked that for any fixed m the value x_m is increasing with respect to each of the three parameters μ, σ and ξ and the same is true for the expectation. Thus increasing any of the three parameters leads to a distribution with greater values.

The GPD can be said to be "stable for exceedance" in the following sense. If $X \sim \texttt{GPD}(\mu, \sigma, \xi)$ then for $u \geqslant \mu$

$$X \mid X > u \sim \texttt{GPD}(u, \sigma^\star, \xi)$$

with $\sigma^\star = \sigma + \xi(u - \mu)$. In other words, the upper tail of a GPD density is a (unnormalized) GPD density see figure B.2.

When $\xi < 1$ the GPD corresponds to a linear mean residual life

$$\mathrm{E}\left(X - v \mid X > v\right) = \frac{\sigma + \xi\, v}{1 - \xi}$$

This may be used for threshold determination in POT: replacing the expectation by a sample mean we can check that the mean excess life is linear: see Coles (2001, chap. 4).

If X is a random variable with a distribution in the domain of attraction of a GEV distribution – as in the Fisher-Tippett-Gnedenko theorem, the GPD can be shown to be the limiting distribution of $Y = X - u$ conditional on $X > u$ when u is large. Moreover the parameter ξ of the GPD coincides with that of the attracting GEV, see theorem 4.1 in Coles (2001). This property provides a justification for the traditional exclusive use of the GPD for exceedances of POT models. An illustration for the Gumbel case $\xi = 0$ is given page 8.

The GPD has an infinite variance when $\xi \geqslant 1/2$. In practice, the values used are generally in the range $-0.3 \leqslant \xi \leqslant 0.3$.

Estimation and inference

In the POT context, the parameter μ is known since it is taken as the threshold u. The exceedances $Y_i := X_i - u$ are distributed according to the GPD with location $\mu = 0$ and unknown σ and scale ξ.

Moments estimators for σ and ξ are readily available

$$\widehat{\xi}_{\texttt{mom}} = \frac{1}{2}\left[1 - \widehat{\mathrm{CV}}^{\,-2}\right], \qquad \widehat{\sigma}_{\texttt{mom}} = \frac{\bar{Y}}{2}\left[1 + \widehat{\mathrm{CV}}^{\,-2}\right].$$

ML estimation can rely on a two-dimensional maximisation. Interestingly enough, the sign of the ML estimator $\widehat{\xi}_{\texttt{ML}}$ has a simple relation with the empirical coefficient of variation $\widehat{\mathrm{CV}}$. Provided that a denominator n is used to estimate the variance[2] in (B.2), one can show that $\widehat{\xi} < 0$ is equivalent to $\widehat{\mathrm{CV}} < 1$. In other words, $\widehat{\xi}_{\texttt{mom}}$ and $\widehat{\xi}_{\texttt{ML}}$ have the same sign. This shows that the sign of the ML estimator $\widehat{\xi}_{\texttt{ML}}$ must be interpreted with care since it is not robust to outliers.

For the ordinary sample (no historical data) case, **Renext** relies on the `evd` package (Stephenson 2002) and its `fpot` estimation function.

Use in Renext

The GPD can be used in **Renext** under the name `"gpd"`. The parameters of (B.5) are named as in the `evd` package

$$\sigma \leftrightarrow \texttt{scale} \qquad \xi \leftrightarrow \texttt{shape}$$

Note that the parameter μ is used with the name `"location"` in the distribution functions, but should not be used in the POT context: it must then be equal to its default value 0, since the distribution is fitted on the exceedances Y_i.

B.3.3 Weibull

Definition

The Weibull distribution has a survival function $S(y)$ and a density function $f(y)$ given by

$$S(y) = e^{-(y/\beta)^\alpha} \qquad f(y) = \frac{\alpha}{\beta}\left[\frac{y}{\beta}\right]^{\alpha-1} e^{-(y/\beta)^\alpha} \qquad y \geqslant 0 \tag{B.6}$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ the scale parameter.

Properties

The properties of the Weibull depend on the shape parameter $\alpha > 0$.

- when $0 < \alpha < 1$ with decreasing hazard rate and increasing mean residual life MRL,
- when $\alpha = 1$ the distribution is exponential with constant hazard rate and constant MRL.
- when $\alpha > 1$ with increasing hazard rate and decreasing MRL.

[2] That is $\widehat{\mathrm{Var}}(Y) = \frac{1}{n}\sum_i (Y_i - \overline{Y})^2$.

see Bagnoli and Bergstrom (2004).

The return level of period $m > 1$ is given by $y_m = \beta\,[\log m]^{1/\alpha}$, confirming that the exponential return level curve $[\log m,\, y_m]$ is convex (concave upwards) for $0 < \alpha < 1$ and (downwards) concave for $\alpha > 1$.

The Weibull distribution is closely related to the exponential. When Y is Weibull with shape α the random variable $Z = Y^{1/\alpha}$ has an exponential distribution. Thus when Y follows a Weibull distribution $V = -\log Y$ has a Gumbel distribution.

Estimation and inference

The ML estimation is carried out by concentrating the scale parameter out of the likelihood. It can be shown that with a suitable re-parameterisation the concentrated likelihood is a log-concave function having an unique maximum easily obtained through a one-parameter maximisation. Moreover the expected information matrix can be given in closed form. These tips are used in **Renext**.

Goodness-of-fit

Specific tests exist for Weibull distributions but are not yet in **Renext**. The fit can be controlled graphically with a *Weibull plot* such as produced by the `weibplot` function.

Use in Renext

The Weibull distribution can be used in **Renext** under the name `"weibull"`. The parameters of (B.6) are named as in the `stats` package from which the distribution functions are taken

$$\beta \leftrightarrow \texttt{scale} \qquad \alpha \leftrightarrow \texttt{shape}$$

The ML estimation with likelihood concentration is available in the `fweibull` function.

This distribution can be used in `Renouv` as a special distribution. It is not necessary to provide initial values for the ML estimation since specific initial values are used then in `Renouv`.

B.3.4 Gamma

Definition

The gamma distribution has density

$$f(y) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, y^{\alpha-1} e^{-y/\beta} \qquad y \geqslant 0 \tag{B.7}$$

where $\Gamma(\alpha)$ denotes the Euler's gamma function, $\beta > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter. The distribution function $F(y)$ and the survival $S(y)$ do not have a simple expression.

Properties

Expectation, variance and coefficient of variation are given by

$$\mathrm{E}(Y) = \alpha\beta, \qquad \mathrm{Var}(Y) = \alpha\beta^2, \qquad \mathrm{CV}(Y) = \frac{1}{\sqrt{\alpha}}.$$

The shape parameter α is related to the coefficient of variation and $0 < \alpha < 1$ gives $\mathrm{CV}(Y) > 1$.

The properties of the distribution depend on the shape parameter $\alpha > 0$.

- for $0 < \alpha < 1$ the hazard rate is decreasing and the mean residual life MRL is increasing,
- for $\alpha = 1$ the distribution is the exponential with constant hazard and constant MRL,
- for $\alpha > 1$ the hazard rate is increasing and the MRL is decreasing.

see reference Bagnoli and Bergstrom (2004).

The gamma distribution is not frequently used to describe extremes. However in the decreasing hazard case $0 < \alpha < 1$, it can be considered as a continuous mixture of exponentials.

It can be shown that the gamma distribution falls in the domain of attraction of the Gumbel distribution.

Estimation

Using an ordinary sample Y_i the moment estimators are readily available

$$\widehat{\alpha}_{\texttt{mom}} = \widehat{\mathrm{CV}}^{\,-2}, \qquad \widehat{\beta}_{\texttt{mom}} = \bar{X} \times \widehat{\mathrm{CV}}^{\,2}.$$

and these could be used as initial values for a numerical likelihood maximisation.

As in the Weibull case, it is possible to concentrate the likelihood and thus to solve a one-parameter maximisation problem. Moreover, the maximisation can be reduced to that of a concave function, and the *expected* information matrix can be computed.

Use in Renext

The gamma distribution can be used in **Renext** under the name `"gamma"`. The parameters of (B.7) are named as in the `stats` package from which the distribution functions are taken

$$\beta \leftrightarrow \texttt{scale} \qquad \alpha \leftrightarrow \texttt{shape}$$

The ML estimation with likelihood concentration is available in the `fgamma` function.

It is not necessary to provide initial values for the ML estimation since specific initial values are used then in `Renouv`.

B.3.5 Log-normal

Definition

The log-normal distribution is the distribution of e^V where V is normal. It has density

$$f(y) = \frac{1}{y\,\sigma\sqrt{2\pi}}\, \exp\left\{-\frac{1}{2\sigma^2}\,[\log y - \mu]^2\right\} \qquad y > 0 \tag{B.8}$$

where μ and $\sigma > 0$ are the parameter of the normal distribution of $\log Y$. The distribution function $F(y)$ and the survival $S(y)$ do not have simple expression.

Note that these parameters are not the location nor the scale parameter since they are in the logged scale.

Properties

The expectation, variance and coefficient of variation of the log-normal distribution are

$$\mathrm{E}(Y) = e^{\mu+\sigma^2/2}, \qquad \mathrm{Var}(Y) = \left[e^{\sigma^2} - 1\right]\, e^{2\mu+\sigma^2}, \qquad \mathrm{CV}(Y) = \sqrt{e^{\sigma^2} - 1}.$$

For the log-normal distribution neither the hazard $h(y)$ nor the mean residual life $\mathrm{MRL}(y)$ are monotonous functions. The mean residual life $\mathrm{MRL}(y)$ is reputed[3] to be decreasing for large values of y.

Estimation and inference

The ML estimation from an ordinary sample is straightforward using the log transformation which leads to the normal case. Exact inference is also available for the parameters.

However, exact inference for the return levels or return periods is more complicated. Hence the standard numerical "delta method" is used in **Renext**.

Goodness-of-fit

The fit of the log-normal distribution can be assessed using the logged values and a normality test (e.g. Shapiro-Wilk). Since the log-normal is not frequently used in POT, such a test is not in computed in **Renext**.

[3] No proof of this assertion was found.

Use in Renext

The log-normal distribution can be used in **Renext** under the name `"lnorm"`. The parameters of (B.8) are named as in the `stats` package from which the distribution functions are taken

$$\mu \leftrightarrow \texttt{meanlog} \qquad \sigma \leftrightarrow \texttt{sdlog}$$

It is not necessary to provide initial values for the ML estimation since specific initial values are used then in `Renouv`.

B.3.6 Finite mixture of exponentials

Definition

The finite mixture of exponentials is a distribution with density (or survival) function obtained as a weighed mean of a finite number of exponential densities (or survivals) with different rates. For a mixture of two exponentials, the survival function $S(y)$ and density $f(y)$ are given by

$$S(y) = \alpha_1 \, e^{-\lambda_1 y} + (1 - \alpha_1) \, e^{-\lambda_2 y}, \qquad f(y) = \alpha_1 \lambda_1 \, e^{-\lambda_1 y} + (1 - \alpha_1)\lambda_2 \, e^{-\lambda_2 y}, \qquad y \geqslant 0 \tag{B.9}$$

and the parameters are α_1, λ_1 and λ_2 must verify

$$0 < \alpha_1 < 1 \qquad 0 < \lambda_1 < \lambda_2. \tag{B.10}$$

It can be preferable to use the alternative parameter vector $[\alpha_1, \lambda_1, \delta]'$ with $\delta := \lambda_2 - \lambda_1$, since the constraint $\lambda_1 < \lambda_2$ is replaced then by the simple constraint $\delta > 0$.

The usual interpretation of a mixture applies: the distribution is that of a random variable that would be randomly chosen from the exponential with rate λ_1 or from the exponential with rate λ_2 the respective probabilities being α_1 and $1 - \alpha_1$. In survival analysis the mixture components correspond to two death rates that may result from two causes of mortality or from the existence of two sub-populations.

Properties

The expectation and uncentered moments have a simple form

$$\mathrm{E}(Y^\gamma) = \alpha_1/\lambda_1^\gamma + (1 - \alpha_1)/\lambda_2^\gamma$$

for any $\gamma > 0$. The coefficient of variation is always greater than 1.

For large values of y, the survival $S(y)$ only depends on the smallest rate λ_1, since

$$S(y) \underset{y \to +\infty}{\sim} \alpha_1 \, e^{-\lambda_1 y}. \tag{B.11}$$

The survival analysis context provides a simple interpretation: after a large time y, the sub-population with smaller death rate λ_1 dominates, and the mean residual life therefore increases.

It can be shown that the hazard rate function $h(y)$ is decreasing with a limit λ_1, and that the mean excess life is increasing with a finite limit $1/\lambda_1$. This "rejuvenation effect" results from the progressive extinction of the population having the highest death rate λ_2. The cumulative hazard $H(y)$ is concave, see figure B.3.

The quantile function is not available in closed form and must be computed numerically.

Estimation and inference

Note that the model would be unidentifiable if the second constraint of (B.10) was omitted since the distribution is invariant under the transformation

$$[\alpha_1, \lambda_1, \lambda_2] \to [1 - \alpha_1, \lambda_2, \lambda_1].$$

For an ordinary sample Y_i the ML estimation can be done using Expectation-Maximisation (EM) algorithm. In this approach, each data Y_i is associated to a latent variable Z_i with value $z = 1$ or $z = 2$ indicating the group (or sub-population) for observation i and consequently the rate λ_z.

In **Renext** the standard log-likelihood maximisation is used. Initial values are computed using the moments when possible, or using (B.11): regressing $\log S(y)$ against y for large values of y give $-\log \alpha_1$ (intercept) and λ_1 (slope), see figure B.3. Then λ_2 can be deduced from the sample mean. However care is needed since these estimates may not fulfil the constraints requirements.

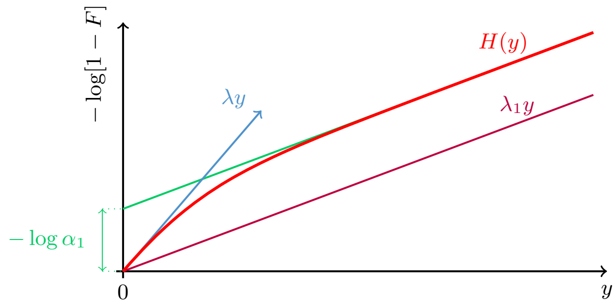


Figure B.3: Exponential plot for the distribution function of a mixture of two exponentials. The curve shows the cumulative hazard $H(y) = -\log[1 - F(y)]$. The slope of the tangent to the curve at the origin is the weighed mean rate $\lambda = \alpha_1\lambda_1 + (1 - \alpha_1)\lambda_2$. The slope of the asymptote is λ_1. Note that $\lambda_1 < \lambda < \lambda_2$.

Generalisation

A mixture of m exponentials ($m \geqslant 2$) can be defined through

$$S(y) = \sum_{i=1}^{m} \alpha_i \, e^{-\lambda_i y}, \qquad f(y) = \sum_{i=1}^{m} \alpha_i \lambda_i \, e^{-\lambda_i y}, \qquad y \geqslant 0$$

with constraints $0 < \alpha_i < 1$, $\sum_i \alpha_i = 1$ and $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$ Since the parameter α_m can be dropped as in the $m = 2$ case, the distribution depends on $2m - 1$ free parameters. The behaviour for large y results from (B.11) which still applies.

The mixture of exponentials is sometimes called *hyper-exponential distribution*.

Use in Renext

The mixture of exponential distributions can be used in **Renext** under the name `"mixexp2"`, and is currently limited to $m = 2$ exponentials. The distribution functions (including the quantile function) are provided by **Renext** and use the following names for the parameters of (B.9)

$$\alpha_1 \leftrightarrow \texttt{prob1} \qquad \lambda_1 \leftrightarrow \texttt{rate1} \qquad \delta = \lambda_2 - \lambda_1 \leftrightarrow \texttt{delta}$$

It is not necessary to provide initial values for the ML estimation since specific initial values are used then in `Renouv`.

The ML-based inference for the mixture of exponentials is well known to be difficult, and bayesian inference might be a valuable alternative.

B.3.7 Lomax

Definition

The *Lomax* distribution depends on two parameters $\beta > 0$ (scale) and $\alpha > 0$ (shape) with survival function

$$S(y) = \left[\frac{\beta}{y + \beta}\right]^{\alpha} = \left[1 + \frac{y}{\beta}\right]^{-\alpha} \qquad y > 0 \tag{B.12}$$

This distribution is also known as *Pareto distribution of the second kind* (Johnson, Kotz, and Balakrishnan 1994). When Y is a random variable following this distribution, $X = Y + \beta$ is Pareto with minimum $x_0 = \beta$ and shape α that is

$$S_X(x) = \left[\frac{x_0}{x}\right]^{\alpha} \qquad x > x_0$$

The Pareto distribution with minimum x_0 and shape α is a special case of $\texttt{GPD}(\mu, \sigma, \xi)$ with location $\mu = x_0$, shape $\xi = 1/\alpha$ (positive) and the extra constraint $\sigma/\xi = x_0$. The Lomax distribution is the special case of the Generalised Pareto $\texttt{GPD}(\mu, \sigma, \xi)$ with $\mu = 0$, $\sigma = \beta/\alpha$ and $\xi = 1/\alpha$, thus implying a positive shape parameter ξ.

We can rewrite the distribution function of Y in the form (B.15) below, with $\phi_\alpha(z) \equiv \log z$, i.e. with the Box-Cox transformation (B.16) for $\alpha = 0$. Therefore, the Lomax distribution can be considered as a limit case of the Shifted Left Truncated Weibull SLTW. We may speak of *log-exponential distribution* although the expression is ambiguous.

Properties

The quantile function is available in closed form. The expectation is finite only for $\alpha > 1$ and the variance is finite only for $\alpha > 2$. In this case

$$\mathrm{E}(Y) = \frac{\beta}{\alpha - 1}, \qquad \mathrm{Var}(Y) = \frac{\alpha\,\beta^2}{(\alpha-1)^2(\alpha-2)}, \qquad \mathrm{CV}(Y) = \sqrt{\frac{\alpha}{\alpha-2}} > 1.$$

Only the cases with $\alpha > 2$ seem practicable. Then $\mathrm{CV}(Y)$ will be close to 1 for a large shape α.

The Lomax distribution has a decreasing hazard rate and a linearly increasing Mean Residual Life.

If both α and β tend to ∞ with α/β tending to $\lambda > 0$ then the Lomax distribution tends to the exponential with rate λ.

It can be shown that this distribution is a (continuous) gamma mixture of exponentials. More precisely, the survival of (B.12) can be written as

$$S(y) = \int_0^{+\infty} g(\lambda)\, e^{-\lambda y}\, \mathrm{d}\lambda$$

where $g(\lambda)$ is the density of the gamma distribution with shape $\alpha_{\texttt{gam}} := \alpha$ and scale $\beta_{\texttt{gam}} := 1/\beta$. The survival $S(y)$ is thus the weighed mean of the exponential survivals $e^{-\lambda y}$ with the weight function $g(\lambda)$. Contrary to the finite mixture of exponentials which behaves for large return periods as does its component with the smallest rate (B.11), this continuous mixture is heavy tailed. The reason is that $g(\lambda)$ weights small rates $\lambda \approx 0$, and thus the mixture embeds exponentials with arbitrarily large means $1/\lambda$. The survival function is a *completely monotone* function (Feller 1971).

Estimation

When the two parameters $\beta > 0$ and $\alpha > 0$ are unknown, the ML estimators from an ordinary sample Y_i can be found using a one-dimensional optimisation by concentrating the shape parameter α out of the likelihood. Although the concentrated log-likelihood $\ell_{\texttt{c}}(\beta)$ is not concave, it can be proved to have a maximum[4] when the sample CV is greater than 1. Moreover the expected information matrix is available in closed form (Giles, Feng, and Godwin 2013). The ML estimates fail to exist when the sample coefficient of variation CV is less than 1. The estimation may also fail when CV is greater than, yet close to 1.

When β is known, the estimation boils down to that of the exponential distribution since $V := \log[1 + Y/\beta]$ then follows an exponential distribution with rate α.

Use in Renext

This distribution is provided in **Renext** under the name `"lomax"`. The names of the formal arguments for the parameters in the probability functions are

$$\texttt{scale} \leftrightarrow \beta \qquad \texttt{shape} \leftrightarrow \alpha$$

The ML estimation with likelihood concentration is available in the `flomax` function.

This distribution is recognized as special in `Renouv`, thus providing a simple mean to impose the constraint $\xi > 0$ for exceedances assumed to follow `GPD`$(0,\, \sigma,\, \xi)$.

Estimation and exact inference are possible in the case where the shift β is taken as the (known) threshold i.e. $\beta = u$. The exponential distribution should then be used with a logarithmic transformation as explained below in B.3.9. The two formal arguments and values to use in the `Renouv` call are `distname.y = "exponential"` and `trans.y = "log"`. Note that α is then obtained with the name `"rate"`, and its estimated value will be greater than 1.

[4] Our proof states the existence of local maximum.

B.3.8 Maxlo

Definition

Though very useful in POT models, this distribution does not seem to have deserved its own name yet. We decided to call it "maxlo" as a pun inspired by a kind of symmetry to the Lomax distribution.

The *maxlo* distribution depends on two parameters $\beta > 0$ (scale) and $\alpha > 0$ (shape). The support of the distributions is $(0,\, \beta)$ and the survival function is

$$S(y) = \left[\frac{\beta - y}{\beta}\right]^{\alpha} = \left[1 - \frac{y}{\beta}\right]^{\alpha}, \qquad 0 < y < \beta. \tag{B.13}$$

The maxlo distribution is the special case of the Generalised Pareto $\texttt{GPD}(\mu,\, \sigma,\, \xi)$ with $\mu = 0$, $\sigma = \beta/\alpha$ and $\xi = -1/\alpha$, thus implying a *negative shape* ξ.

Properties

The quantile function is available in closed form. The expectation is finite only for $\alpha > 1$ and the variance is finite only for $\alpha > 2$. In this case

$$\mathrm{E}(Y) = \frac{\beta}{\alpha + 1}, \qquad \mathrm{Var}(Y) = \frac{\alpha\,\beta^2}{(\alpha+1)^2(\alpha+2)}, \qquad \mathrm{CV}(Y) = \sqrt{\frac{\alpha}{\alpha+2}} < 1.$$

Note that $\mathrm{CV}(Y)$ will be close to 1 for large values of the shape α.

If both α and β tend to ∞ with α/β tending to $\lambda > 0$ then the maxlo distribution tends to the exponential with rate λ.

Estimation

When the two parameters $\beta > 0$ and $\alpha > 0$ are unknown, the ML estimators from an ordinary sample Y_i can be found using a one-dimensional optimisation by concentrating the shape parameter α out of the likelihood. Although the concentrated log-likelihood $\ell_{\mathtt{c}}(\beta)$ is not concave it can be proved to have a maximum[5] when the sample CV is smaller than 1, thus mirroring the property stated for the Lomax distribution. Moreover the expected information matrix is available in closed form. The ML estimates fail to exist when the sample coefficient of variation CV is greater than 1. The estimation may also fail when CV is smaller than yet close to 1.

When β is known, the estimation boils down to that of the exponential distribution since $V := -\log[1 - Y/\beta]$ follows an exponential distribution with rate α.

Use in Renext

This distribution is provided in **Renext** under the name `"maxlo"`. The names of the formal arguments for the parameters in the probability functions are

$$\texttt{scale} \leftrightarrow \beta \qquad \texttt{shape} \leftrightarrow \alpha$$

The ML estimation with likelihood concentration is available in the `fmaxlo` function.

This distribution can be used in `Renouv`, thus providing a simple mean to impose the constraint $\xi < 0$ for exceedances assumed to follow $\texttt{GPD}(0,\, \sigma,\, \xi)$.

B.3.9 Transformed Exponential distributions

Definition

This rather informal family of distributions is sometimes used in hydrology. Although we will only consider in practice the two functions $\phi(x) = x^2$ and $\phi(x) = \log x$ both for $x > 0$, a slightly more general

[5] Our proof states the existence of local maximum.

framework can be proposed as follows. Let $\phi(x)$ be a regular and strictly increasing function defined for $x > x_0$ and let u be a known value $u > x_0$. When a random variable X is such that

$$\phi(X) - \phi(u) \sim \texttt{Exp}$$

we may say that X has a *transformed exponential* distribution. The values of this distribution are the real numbers x with $x > u$. Note that the transformation needs to be one-to-one, because the distribution of X must be determinable from that of $Z = \phi(X) - \phi(u)$. Then

$$X = \psi\left(Z + \phi(u)\right)$$

where $\psi(z)$ is the reciprocal function of $\phi(x)$. As an example, the square transformation can be applied only for $x > 0$.

The survival function is given by

$$S_X(x) = \exp\left\{-\nu\left[\phi(x) - \phi(u)\right]\right\} \qquad x > u$$

where $\nu > 0$ is the rate of the exponential distribution. The density comes by derivation.

Properties

The properties of the distribution obviously depend on the choice of the transformation.

- For the square transformation $\phi(x) = x^2$ we get a shifted and truncated Weibull distribution as described below. It may be called *square-exponential* or (in french) *loi en carrés*.
- With the logarithmic transformation $\phi(x) = \log x$ we get a shifted version of the Pareto (heavy tailed) distribution called Lomax distribution and described above in B.3.7. It may be called *log-exponential*.

The quantile function is available in closed form provided that the reciprocal function $\psi(z)$ is such. This is actually the case for the two transformations considered.

Estimation and inference

As far as an ordinary sample X_i is used, the ML estimator $\widehat{\nu}$ of the rate ν is available using the mean of the transformed random variables $Z_i = \phi(X_i) - \phi(u)$

$$1/\widehat{\nu} = \overline{Z} = \overline{\phi(X)} - \phi(u)$$

Exact inference on ν is deduced from the exponential case.

Use in Renext

The package allows the use of two transformed exponential distributions with the `Renouv` function, where *u is necessarily taken as equal to the threshold.* The value given for the transformation formal argument `trans.y` can be either `"square"` or `"log"`. In both cases, the exponential distribution must be specified by giving the value `"exponential"` to the distribution argument `distname.y`.

B.3.10 Shifted Left Truncated Weibull (SLTW) distribution

Definition

We call (shifted) *left truncated Weibull* (SLTW) the following distribution for a random variable $Y > 0$.

It depends on three parameters $\delta > 0$ (shift or location), $\beta > 0$ (scale) and $\alpha > 0$ (shape) and has survival function

$$S(y) = \exp\left\{-\left[\left(\frac{y+\delta}{\beta}\right)^{\alpha} - \left(\frac{\delta}{\beta}\right)^{\alpha}\right]\right\} \qquad y > 0 \tag{B.14}$$

The density comes by derivation. This is the conditional distribution $X - \delta \mid X > \delta$ where X has Weibull distribution with shape α and scale β.

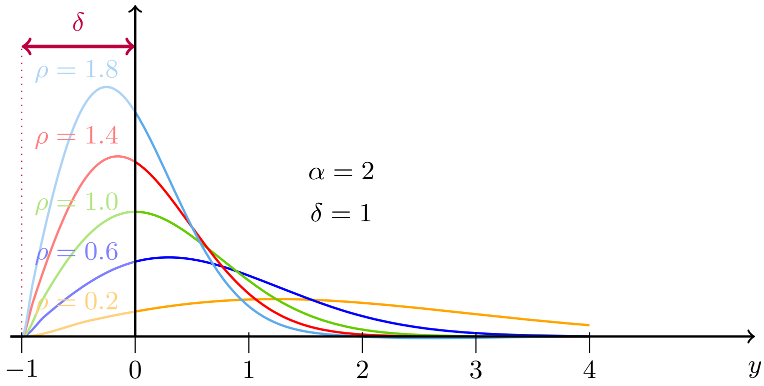


Figure B.4: "Square exponential" densities, i.e. SLTW densities with shape $\alpha = 2$. Only the part $y \geqslant 0$ of the Weibull densities is used and the normalisation is on the interval $y \geqslant 0$.

For $\alpha = 2$ we can rewrite the survival as

$$S(y) = \exp\left\{-\nu\left[(y+\delta)^2 - \delta^2\right]\right\} \qquad y > 0$$

thus the distribution is identical to the square-exponential described previously.

This three parameter family can be used for exceedances in POT, but in a general framework there is no natural choice for $\delta > 0$ in relation with a physical threshold u, though the two quantities have the same physical dimension. For some applications of POT where the random variable is positive δ is sometimes chosen as the threshold $\delta = u$.

Properties

The three parameter family is (by construction) stable by exceedance over a threshold > 0. The moments or even the expectation are not easily computed in the general case.

For $\alpha \leqslant 1$ the mode of Y is always $y = 0$. For $\alpha > 1$ the mode of Y is the positive part $y_+^\star$ of the shifted mode $y^\star$ of the Weibull i.e. $y^\star = (\alpha-1)^{1/\alpha}\beta - \delta$. Thus for a fixed α and δ we can have a mode varying with β.

The quantile function is available in closed form. The hazard and the MRL for this distribution are merely truncations of their equivalent for the Weibull distribution, e.g. the hazard is decreasing for $0 \leqslant \alpha < 1$ and increasing for $\alpha > 1$.

For $\alpha > 0$ and large δ, the distribution is close to the exponential since the Weibull distribution is in the domain of attraction of the Gumbel distribution for which the exceedances over a large threshold tend to be exponentially distributed.

Using the notation $\rho = \alpha/\beta^\alpha$ we can rewrite the survival as

$$S(y) = \exp\left\{-\rho\left[\phi_\alpha(y+\delta) - \phi_\alpha(\delta)\right]\right\} \qquad y > 0 \tag{B.15}$$

where $\phi_\alpha(z)$ is the Box-Cox transformation defined for $z > 0$ by

$$\phi_\alpha(z) = \begin{cases} (z^\alpha - 1)/\alpha & \alpha > 0 \\ \log z & \alpha = 0 \end{cases} \tag{B.16}$$

The function $\phi_\alpha(z)$ is strictly increasing with limit $+\infty$ when $z \to +\infty$ and it is regular with respect to α for $\alpha = 0$. Thus if α and β both tend to zero in such way that ρ tends to a limit $\rho^\star > 0$ the distribution tends to the Lomax distribution described above. The limit survival is (B.15) with $\alpha = 0$ and $\rho = \rho^\star$.

Estimation

In most contexts, the shift parameter δ should be known and given.

Note that when both α and δ are known and when the estimation is from an ordinary sample Y_i of size n, the ML estimator $\widehat{\rho} = \alpha/\beta^\alpha$ of ρ is available using the mean of the transformed Y_i

$$1/\widehat{\rho} = \overline{\phi_\alpha(Y+\delta)} - \phi_\alpha(\delta)$$

Exact inference on ρ or on the quantiles is then easily deduced from the exponential case.

Use in Renext

The SLTW distribution is provided in **Renext** under the name `SLTW`. The relevant probability functions share the three following formal arguments for the parameters, in correspondence with (B.14)

`delta` $\leftrightarrow \delta$ `shape` $\leftrightarrow \alpha$ `scale` $\leftrightarrow \beta$

Note that the parameter named `scale` *is not* a scale parameter in the usual statistical sense; the name only refers to the original Weibull distribution.

No specific inference method is implemented in the **Renext** POT fitting. A special case is when δ is equal to the (known) threshold u and when moreover α is known. Indeed, we then fit an exponential distribution to a transformed version $\phi_\alpha(X)$ of the level $X \equiv Y + u$. We thus can use in the special case where $\alpha = 2$ (square transformation) and the limit case where $\alpha = 0$ (log transformation) as explained above in B.3.9. In the `Renouv` function, one must then use `distname.y = "exponential"`; the transformation argument must be respectively `trans.y = "square"` and `trans.y = "log"`.

B.3.11 Other distributions

It is possible to use a quite arbitrary distribution within the `Renouv` function provided the probability functions[6] are available in R and satisfy the conditions stated in the help of the `Renouv` function.

[6]Density, distribution and quantile functions are required.

Bibliography

Bagnoli M, Bergstrom T (2004). "Log-Concave Probability and Its Applications." University of California Santa Barbara, dpt of Economics. Paper 1989D. URL `works.bepress.com/ted_bergstrom/98`.

Coles S (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer.

Cox D (1962). *Renewal Theory*. Methuen, London.

Cox D, Isham V (1980). *Point Processes*. Monograph on Applied Probability and Statistics. Chapman and Hall.

Davison A, Smith R (1990). "Models for Exceedances over High Thresholds." *J.R. Statist. Soc. B*, **52**(3), 393–442.

Embrecht P, Klüppelberg C, Mikosch T (1996). *Modelling Extremal Events for Insurance and Finance*. Springer.

Feller W (1971). *An Introduction to Probability Theory and its Applications*, volume 2. 2nd edition. Willey.

Giles D, Feng H, Godwin R (2013). "On the Bias of the Maximum Likelihood Estimator for the Two-Parameter Lomax Distribution." *Communications in Statistics - Theory and Methods*, **42**(11), 1934–1950.

Gilleland É, Katz R, Young G (2004). *extRemes: Extreme value toolkit*. R package version 1.59, URL `http://www.assessment.ucar.edu/toolkit/`.

Heffernan J, Stephenson A (2012). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.39. Original S functions by J.E. Heffernan and R port and R documentation files by A. Stephenson., URL `http://www.ral.ucar.edu/~ericg/softextreme.php`.

Johnson N, Kotz S, Balakrishnan N (1994). *Continuous Univariate Distributions*, volume 1. 2nd edition. Wiley.

Kotz S, Nadarajah S (2005). *Extreme Value Distributions, Theory and Applications*. Imperial College Press.

Miquel J (1984). *Guide pratique d'estimation des probabilités de crues*. Eyrolles.

Parent É, Bernier J (2007). *Le raisonnement bayésien: modélisation et inférence*. Coll. Statistiques et probabilités appliquées. Springer-Verlag.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ribatet M (2009). *POT: Generalized Pareto Distribution and Peaks Over Threshold*. R package version 1.1-0, URL `http://CRAN.R-project.org/package=POT`.

Stephenson A (2002). "evd: Extreme Value Distributions." *R News*, **2**(2), 0. URL `http://CRAN.R-project.org/doc/Rnews/`.

Stephenson A, Ribatet M (2008). *evdbayes: Bayesian Analysis in Extreme Value Theory*. R package version 1.0-7.

Viglione A (2009). *nsRFA: Non-supervised Regional Frequency Analysis*. R package version 0.6-9, URL `http://CRAN.R-project.org/package=nsRFA`.

Index