UNSW SYDNEY

MATH5805 – A Modern Introduction to Extreme Value Theory

Week 4 - Univariate Extremes: the Peaks-Over-Threshold approach (POT)

Dr Boris Béranger
School of Mathematics and Statistics
UNSW Sydney

CRICOS Provider Code 00098G

Lecture outline



1. Introduction
2. Limit laws
3. Using the GPD limit theorem
4. Parameter estimation
5. Threshold selection
6. Analysis of precipitation data
7. Analysis of hurricane damage data

1. Introduction



Goals



This lecture is about introducing the **Peaks-over-Threshold** approach.

In particular, we

- Present an alternative approach based on **exceedances of high thresholds**
- Introduce the **Generalised Pareto distribution**

Note that throughout this lecture we will make the assumption that $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with **common distribution function** F.

Summary of previous lecture & moving forward

The methodology introduced so far was obtained either **on the assumption that the data come from a GEV** or **belong to its maximum domain of attraction**.

We based statistical estimation of the relevant parameters on maximum likelihood or profile likelihood.

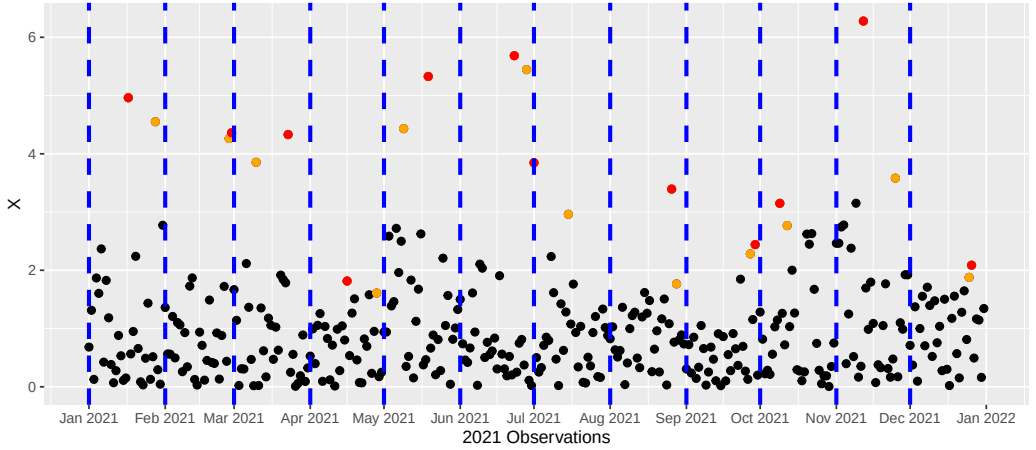
Comment 1

Modelling only block maxima can be **wasteful** if other data on extremes are available.

Example

We might consider the 2 largest observations per block to be extreme. Let's consider the daily concentration level of air pollutant example for last lecture.

Summary of previous lecture & moving forward



Simulated daily levels of an air pollutant in 2021. Red dots represent monthly maxima. **Orange dots** represent the 2nd largest observation per block.

Comment 2

Using the **r largest** observations per block can also wasteful if one block happens to contain more extreme events than another.

Looking at the previous plot, the `Apr 2021` and `Dec 2021` seem to have less extreme events than for example `Jul 2021` or `Nov 2021`.

$\Longrightarrow$ **If an entire time series is available then better use of the data is made by avoiding altogether the procedure of blocking.**

Peaks-over-Threshold - Overview

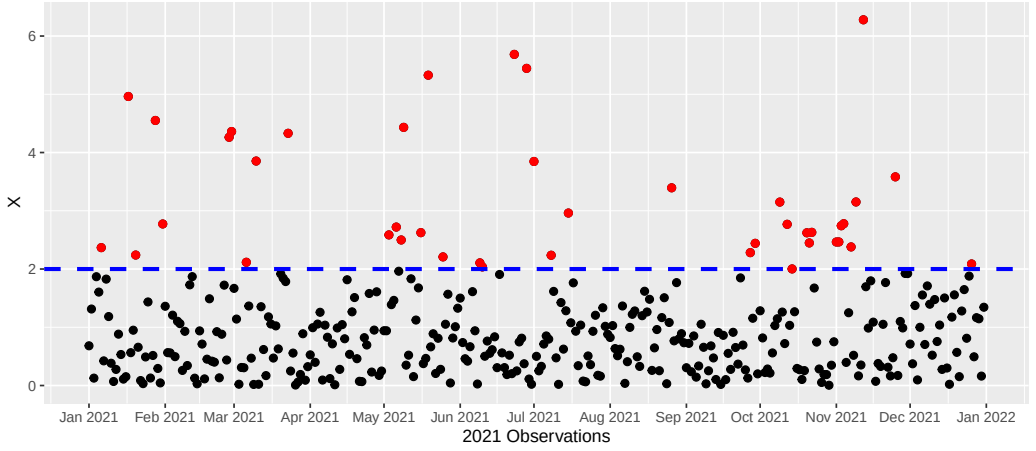
A natural way to define extreme events is to consider those that **exceed some high threshold**

Assume that X is a random variable with right end point x^*. For some **high threshold** $u < x^*$, let $Y_1, \ldots, Y_{N_u}$ **represent the exceedances of** u **by** $X_1, \ldots, X_n$ with

$$N_u = \text{card}\left\{i : i = 1, \ldots, n, X_i > u\right\}.$$

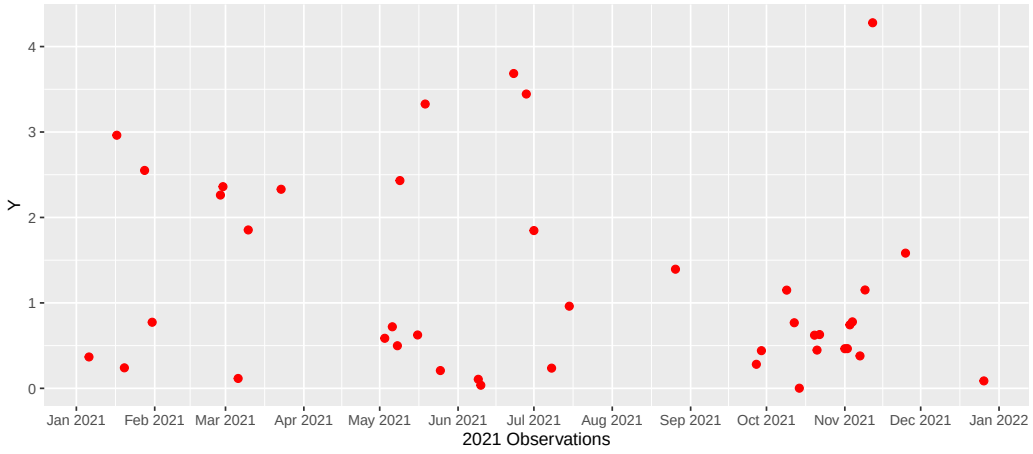
Going back again to the example of daily levels of an air pollutant, the following plot shows the observations above the threshold $u = 2$. There are $N_u = 42$ exceedances against 12 monthly maxima.

Peaks-over-Threshold - Overview



Simulated daily levels of an air pollutant in 2021. Red dots represent the observations above the threshold $u = 2$.

Peaks-over-Threshold - Overview



Exceedance of $u = 2$ by the simulated daily levels of an air pollutant in 2021.

Peaks-over-Threshold - Overview

The **excess distribution function (df)** of the rv X (of the df F) is given by:

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \mathbb{P}(Y \leq y | X > u), \quad y \geq 0. \tag{1}$$

Definition

F_u is known as the *excess-life* in reliability theory, *residual lifetime df* in medical statistics, and is usually referred to as the *excess-of-loss df* in an insurance context.

Note that we can also write:

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

and thus **if the df F were known**, the distribution of threshold exceedances (1) would also be known.



2. Limit laws

Limit laws

In practical applications the df F is unknown and therefore **approximations that are broadly applicable for high values of the threshold are sought**.

This parallels the use of GEV as an approximation to the distribution of maxima of long sequences when the parent distribution (ie. F) is unknown.

Theorem [Pickands (1975)]
The conditional distribution of the exceedances converges to the generalised Pareto distribution (GPD) as $u \to \infty$ under the same conditions as those for the convergence of sample maxima.

Before going any further, let's define the GPD...

The Generalized Pareto Distribution (GPD)

Definition [GPD]

A random variable Y is said to be distributed according to a standard generalised Pareto distribution if it has df

$$H_\xi(y) = \begin{cases} 1 - (1+\xi y)^{-1/\xi} & \xi \neq 0, \\ 1 - \exp\{-y\} & \xi = 0, \end{cases}$$

where

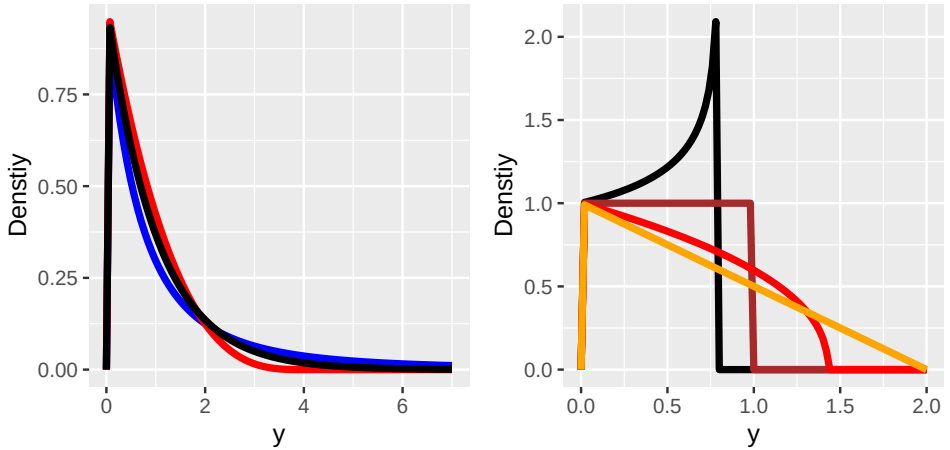
$$\text{domain}(Y) = \begin{cases} y \geq 0 & \xi \geq 0, \\ 0 \leq y \leq -1/\xi & \xi < 0. \end{cases}$$

The Generalized Pareto Distribution (GPD)

Comments

1. One can introduce the related location-scale family $H_\xi(\cdot; \nu, \tilde{\sigma})$ by replacing the argument y by $(y - \nu)/\tilde{\sigma}$ for $\nu \in \mathbb{R}, \tilde{\sigma} > 0$. The support has to be adjusted accordingly.
2. As in the case of G_0, H_0 can also be interpreted as the limit of H_ξ as $\xi \to 0$.
3. When $\xi > 0$, we obtain the ordinary **Pareto** distribution;
4. When $\xi = 0$, we obtain the **Exponential** distribution;
5. When $\xi < 0$ we obtain the **short-tailed Pareto** distribution.

The Generalized Pareto Distribution (GPD)



Left: standard GPD densities with $\xi = -0.25$, $\xi = 0$ and $\xi = 0.5$. Right: standard GPD densities with $\xi = -1.25$, $\xi = -1$, -0.7 and **-0.5**. Recall that they have compact support $[0, -1/\xi]$.

Limit laws

Theorem [Pickands, 1975; de Haan & Balkema, 1974]
Let $X_i, i = 1, \ldots, n$ with common df F such that

$$\mathbb{P}(M_n \leq y) \approx G_\xi(y; \mu, \sigma).$$

Then, **for large enough** u,

$$F_u(y) = \mathbb{P}\left(X - u \leq y | X > u\right) \approx H_\xi(y; 0, \tilde{\sigma}) \equiv H_\xi(y; \tilde{\sigma}),$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$ and where $\tilde{\sigma} = \sigma + \xi(u - \mu)$

In other words: For a large class of underlying distributions F, the excess distribution function F_u can be approximated by the GPD for an increasing threshold u.

Outline proof

Using the assumption that the **GEV limit holds** then for large n

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

for some μ, σ, ξ. Then, taking the log

$$n \log F(x) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

The Taylor expansion $\log F(z) \approx -[1-F(z)]$ for **large** z, implies that for large u:

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

For large u and $y > 0$ we then have

$$1 - F(u+y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}$$

Immediately we then have

$$\begin{aligned}
\mathbb{P}(\mathbf{X > u + y | X > u}) &= \frac{\mathbb{P}(X > u+y, X > u)}{\mathbb{P}(X > u)} \qquad (y > 0) \\
&\approx \frac{n^{-1}[1 + \xi(u+y-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u-\mu)/\sigma]^{-1/\xi}} \\
&= \left[\frac{1 + \xi(u+y-\mu)/\sigma}{1 + \xi(u-\mu)/\sigma}\right]^{-1/\xi} \\
&= \left[\mathbf{1} + \frac{\xi \mathbf{y}}{\sigma}\right]^{-\mathbf{1}/\xi}
\end{aligned}$$



3. Using the GPD limit theorem

Extremal Type Theorem - Examples

Example - Exponential(1)

The distribution function is $F(x) = 1 - e^{-x}$ for $x > 0$. By direct calculation

$$\frac{1 - F(u+y)}{1 - F(u)} = \frac{e^{-(u+y)}}{e^{-u}} = e^{-y} \qquad \text{for } y > 0,$$

and $F_u(y) = 1 - e^{-y} = H_0(y; 0, 1)$, i.e. GPD with $\nu = 0, \tilde{\sigma} = 1, \xi = 0$.

Exact result!

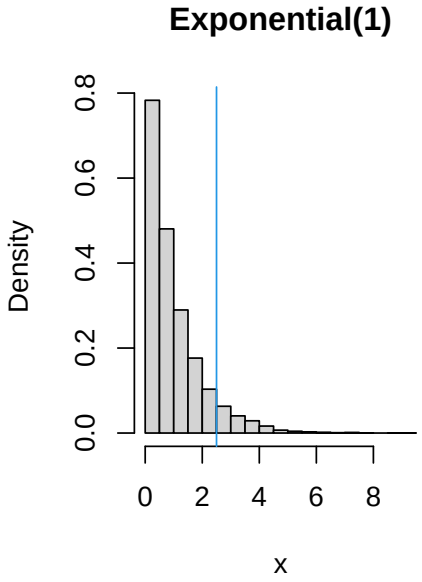
Extremal Type Theorem - Examples

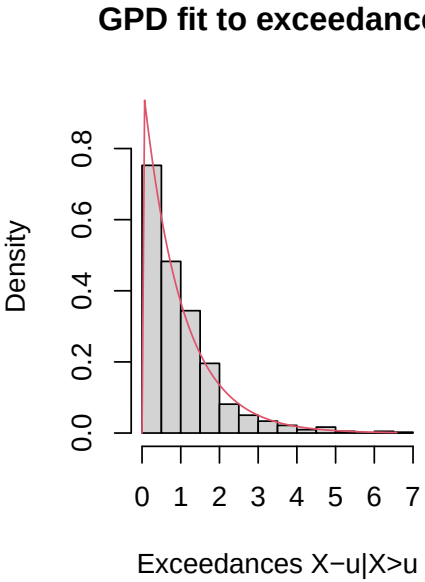
Example - Exponential(1)

Illustration:

- Generate $10,000$ samples from the Exponential(1);
- Display the histogram of the data;
- Select a threshold $u = 2.5$;
- Superimpose the GPD density to the histogram of the exceedances.

Extremal Type Theorem - Examples





Extremal Type Theorem - Examples

Example - Fréchet(1)

The distribution function is $F(x) = \exp(-1/x)$ for $x > 0$. We have

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

as $u \to \infty$.

Therefore $F_u(y) \sim 1 - \left(1 + \frac{y}{u}\right)^{-1} = H_1(y; 0, u)$, i.e. GPD with $\nu = 0, \tilde{\sigma} = u, \xi = 1$.

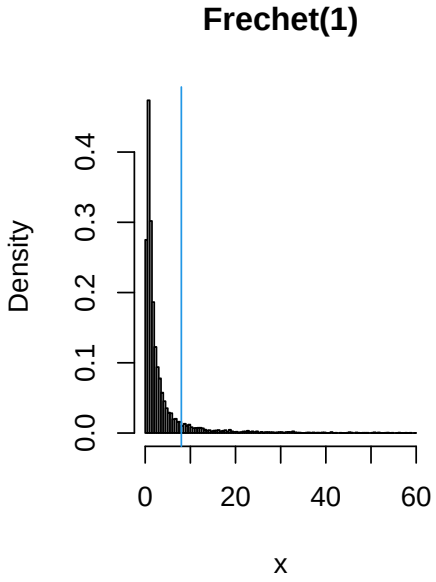
Extremal Type Theorem - Examples

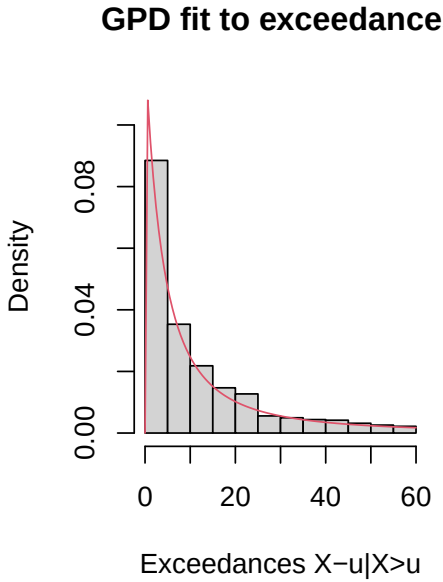
Example - Fréchet(1)

Illustration:

- Generate $10,000$ samples from the Fréchet(1);
- Display the histogram of the data;
- Select a threshold $u = 8$;
- Superimpose the GPD density to the histogram of the exceedances.

Extremal Type Theorem - Examples





Extremal Type Theorem - Examples

Example - Uniform(0,1)

The distribution function is $F(x) = x$ for $0 \leq x \leq 1$. We have

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

for $0 \leq y \leq 1 - u$.

Therefore $F_u(y) = \frac{y}{1-u} = H_{-1}(y; 0, 1-u)$, i.e. GPD with $\nu = 0, \tilde{\sigma} = 1 - u, \xi = -1$.

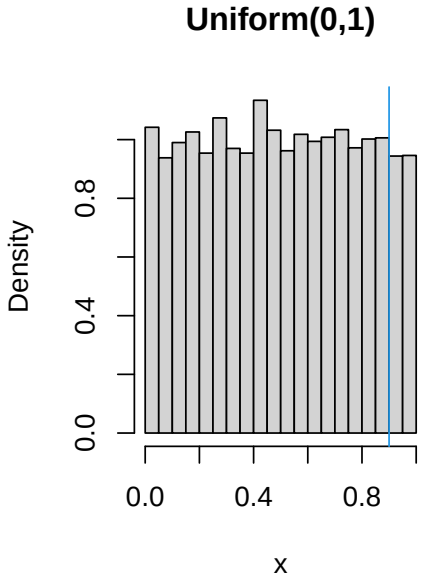
Extremal Type Theorem - Examples

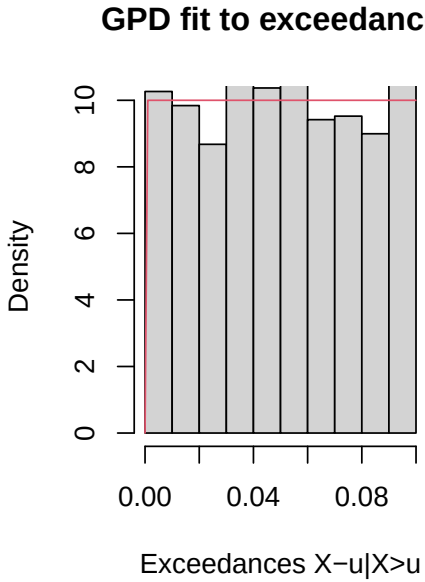
Example - Uniform(0,1)

Illustration:

- Generate $10,000$ samples from the Uniform(0,1);
- Display the histogram of the data;
- Select a threshold $u = 0.9$;
- Superimpose the GPD density to the histogram of the exceedances.

Extremal Type Theorem - Examples





4. Parameter estimation



Parameter estimation

Assuming that a threshold has been determined, the parameters of the GPD can be estimated through maximisation of the (log)-likelihood. For the N_u threshold exceedances, the log-likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{N_u} \log h(y_i; \tilde{\sigma}, \xi) \\ &= \begin{cases} -N_u \log(\tilde{\sigma}) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{N_u} \log\left(1 + \frac{\xi y_i}{\tilde{\sigma}}\right), & \xi \neq 0, \\ -N_u \log(\tilde{\sigma}) - \tilde{\sigma}^{-1} \sum_{i=1}^{N_u} y_i, & \xi = 0, \end{cases}\end{aligned}$$

where $y_i \geq 0$ for $\xi \geq 0$ and $0 \leq y_i \leq -\tilde{\sigma}/\xi$ for $\xi < 0$.

Analytical maximisation is not possible, so numerical techniques are again required.

Parameter estimation

Similar to the GEV, **the log-likelihood needs to be evaluated with care when** $|\xi| \to 0$.

[Smith, 1985]

The same problem with the MLE for block maxima retains:

- The usual properties of the MLE only hold when $\xi > -0.5$;
- The asymptotic variance does not hold when $-1 < \xi \leq -0.5$;
- The MLE does not exist when $\xi < -1$.

Again, the probability weighted moment approach can be considered (but not covered).



Estimation of $(\tilde{\sigma}, \xi)$ is only half way toward estimating tail behaviours (return level, Value-at-Risk, etc.), because the GPD is conditional on exceeding the threshold.

Parameter estimation

By construction we can write:

$$\begin{aligned}\mathbb{P}(X > u + y) &= \mathbb{P}(X > u)\mathbb{P}(X > u + y | X > u) \\ &= \mathbb{P}(X > u)\left(1 - H_\xi(y; \tilde{\sigma})\right) \\ &= \mathbb{P}(X > u)\left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}. \quad \textbf{(assuming } \xi \neq 0\textbf{)} \end{aligned} \tag{2}$$

We can estimate $\eta_u \equiv \mathbb{P}(X > u)$ empirically by $\hat{\eta}_u = N_u/n$.

Suppose that there are b_s observations per block or time period (e.g., year), the **return level** associated with a **return period** is solution to

$$\eta_u \left[1 + \frac{\xi(\text{return level} - u)}{\tilde{\sigma}}\right]^{-1/\xi} = \frac{1}{\text{return period} \times b_s}.$$

Interpretation

The **return level** is exceeded on average once every **return period** $\times$ $\mathbf{b_s}$ observations.

Rearranging the above and doing similar working assuming $\xi = 0$ in (2), we get

$$\text{return level} = \begin{cases} u + \frac{\tilde{\sigma}}{\xi}\left[(\text{return period} \times b_s \times \eta_u)^{\xi} - 1\right], & \xi \neq 0, \\ u + \tilde{\sigma}\log\left(\text{return period} \times b_s \times \eta_u\right), & \xi = 0, \end{cases}$$

provided that return period $\times$ b_s is sufficiently large to ensure that return level $> u$.

Parameter estimation

Graphical summary

Plotting return level against return period $\times\, b_s$ on a logarithmic scale produces the same qualitative features as return level plots on the GEV model:

- linearity if $\xi = 0$;
- concavity if $\xi > 0$;
- convexity if $\xi < 0$;

For $\xi > -0.5$, inferences about the return level can be done by combining inferences of η_u and $(\tilde{\sigma}, \xi)$ with the **delta method**.

Note that the empirical estimator $\hat{\eta}_u$ also corresponds to the MLE of the $\text{Bin}(n, \eta_u)$ distribution of the number of exceedances of u. The **variance-covariance matrix V of** $(\hat{\tilde{\sigma}}, \hat{\xi}, \hat{\eta}_u)$ is

$$V = \begin{bmatrix} \mathbf{V}_{(\hat{\tilde{\sigma}}, \hat{\xi})} & 0 \\ 0 & \hat{\eta}_u \left(1 - \hat{\eta}_u\right) / n \end{bmatrix},$$

where $V_{(\hat{\tilde{\sigma}}, \hat{\xi})}$ is the variance-covariance of $\left(\hat{\tilde{\sigma}}, \hat{\xi}\right)$ obtained from ML maximisation.

Parameter estimation

For ease of notation, **let z_m denote the m-observations return level, where $m =$ return period $\times\, b_s$.**

The variance of the estimated return level is thus

$$\mathrm{Var}(\hat{z}_m) \approx \Delta z_m^\top V \Delta z_m,$$

with

$$\begin{aligned}\Delta z_m^\top &= \left[\frac{\partial z_m}{\partial \tilde{\sigma}}, \frac{\partial z_m}{\partial \xi}, \frac{\partial z_m}{\partial \eta_u}\right] \\ &= \left[\frac{1}{\xi}\left\{(m\eta_u)^\xi - 1\right\}, \frac{-\tilde{\sigma}}{\xi^2}\left\{(m\eta_u)^\xi - 1\right\} + \frac{\tilde{\sigma}}{\xi}(m\eta_u)^\xi \log(m\eta_u), \tilde{\sigma} m^\xi \eta_u^{\xi-1}\right],\end{aligned}$$

evaluated at $(\hat{\tilde{\sigma}}, \hat{\xi}, \hat{\eta}_u)$.

Comments

- Interval estimators constructed from profile likelihood are preferred for their precision.
- Simple approximation ignoring the uncertainty of $\hat{\eta}_u$ can be obtained by parametrising $\tilde{\sigma}$ in terms of z_m ([C01, p.83]).

5. Threshold selection



CRICOS Provider Code 00098G

Threshold selection

Example - Fréchet(4)

Consider a random variable Y that follows the Fréchet(4) distribution, which means $\xi = 0.25$.

We simulate 36525 replicates of Y corresponding to **100 years of daily observations** between 01/01/1926 and 31/12/2025 and are interested in **estimating the high quantile**:

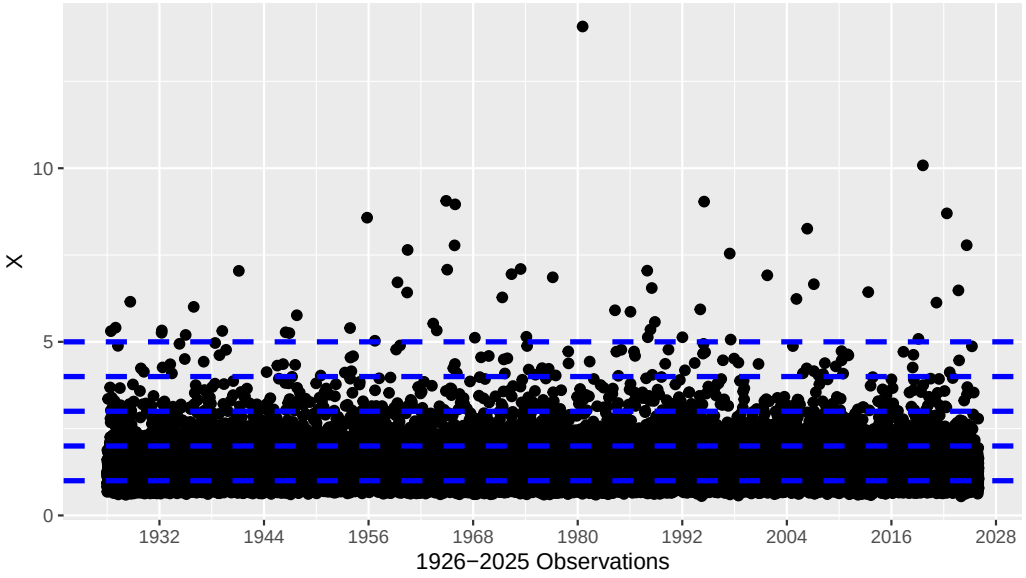
$$\mathbb{P}(Y > \mathbf{13.824}) = \frac{1}{36525}.$$

This is the true 100-year return level if observations are recorded daily.

```
n <- 36525
true.quantile <- qfrechet(1-1/n, shape=4) # qfrechet(1/n, shape=4, lower.tail=FALSE)
true.quantile
```

```
## [1] 13.82439
```

Threshold selection



Threshold selection

```r
library(extRemes)
library(extraDistr) # To generate from the Frechet distribution
# library(ggplot2) if not already loaded

set.seed(1234)

dates <- seq.Date(as.Date("1926/1/1"), by = "day", length.out = n)

dat <- data.frame(
  dates = dates,
  data = rfrechet(n, lambda=4),
  year = format(dates, format = "%Y")
)
```

Threshold selection

```
ggplot(dat, aes(dates,data)) +
  geom_point(size=2) +
  scale_x_date(date_breaks = "12 year", date_labels =  "%Y") +
  xlab("1926-2025 Observations") +
  ylab("X") +
  geom_hline(yintercept=1:5, colour="blue",
             linetype="dashed", linewidth=1.2)
```

Example - Fréchet(4)

We now fit the GPD distribution using the threshold $u = 1, 2, 3, 4$ and 5.

The table below gives the estimate of the shape parameter ξ and its standard standard error, the estimated quantile with 95% confidence interval and the true probability to exceed the threshold.

u	1	2	3	4	5	Truth
ξ	0.146	0.234	0.274	0.145	0.027	0.250
	(0.007)	(0.027)	(0.063)	**(0.108)**	**(0.141)**	
Quantile	9.361	13.061	13.889	11.994	11.472	13.824
CI	$(8.735, 9.987)$	$(10.607, 15.514)$	$(9.560, 18.219)$	$(8.530, 15.459)$	$(8.852, 14.093)$	–
$\mathbb{P}(Y > u)$	0.632	0.061	0.012	0.004	0.002	–

Threshold selection

Example - Fréchet(4)

Conclusion:

- Threshold choice involves a tradeoff between **bias** and **variance**.
- If threshold is **too low**, **bias** results becase we are not far enough in the tail for the limit result to apply.
- If threshold is **too high**, estimators have high **variance** because there is not enough data to estimate parameters well.



Find the lowest threshold u such that the tail behaves like a GPD above u.

Threshold selection

How to obtain the numbers in the table in R?

```
Thresholds <- 1:5
returnPeriod <- 100
bs <- 365.25
out <- matrix(NA, ncol=length(Thresholds), nrow=7)
rownames(out) <- c("u", "xi_hat", "xi_hat_se", "quant", "ci.low", "ci.upp"

for(i in 1:length(Thresholds)){
  fit <- fevd(x=dat$data, threshold=Thresholds[i], type="GP", span=100)
  fit.sum <- summary(fit, silent=TRUE)
  rl <-   return.level(fit, return.period=100, do.ci=TRUE)
  prob <- pfrechet(Thresholds[i], lambda=4, lower.tail=F)
```

Threshold selection

```
  out[1,i] <- Thresholds[i]
  out[2,i] <- round(fit.sum$par[2],3)
  out[3,i] <- round(fit.sum$se.theta[2],3)
  out[4:6,i] <- round(rl[c(2,1,3)],3)
  out[7,i] <- round(prob,3)
}

out
```

Threshold selection

```
##             [,1]   [,2]   [,3]   [,4]   [,5]
## u          1.000  2.000  3.000  4.000  5.000
## xi_hat     0.146  0.234  0.274  0.145  0.027
## xi_hat_se  0.007  0.027  0.063  0.108  0.141
## quant      9.361 13.061 13.889 11.994 11.472
## ci.low     8.735 10.607  9.560  8.530  8.852
## ci.upp     9.987 15.514 18.219 15.459 14.093
## p.exceed   0.632  0.061  0.012  0.004  0.002
```

Mean excess function

Corollary Mean excess function

Suppose a random variable X follows the GPD with $\xi < 1$ and $\tilde{\sigma}$. Then for $u < x^*$,

$$\mathrm{E}\left(X - u | X > u\right) = \frac{\tilde{\sigma}}{1 - \xi}.$$

In an insurance context it can be interpreted as the **expected claim size** in the unlimited layer, over priority u. In reliability or medical context it is referred to as the **mean residual life** function. In a financial risk management context, switching from the right tail to the left tail, it is referred to as the **shortfall**.

This last result gives a **nice graphical technique** for choosing the threshold u so high that an approximation of the excess df F_u by a GPD is justified.

Threshold selection

Let $\tilde{\sigma}_0$ denote the scale parameter corresponding to the excesses of the threshold u_0:

$$\mathrm{E}\,(X - u_0 | X > u_0) = \frac{\tilde{\sigma}_0}{1-\xi}.$$

If the GPD is valid for excesses over u_0, then it is also be valid **for any** $u > u_0$ and thus

$$\mathbf{E}\,(X - u | X > u) = \frac{\tilde{\sigma}}{1-\xi} = \frac{\tilde{\sigma}_0 + \xi(u - u_0)}{1-\xi}.$$

This means that for any $u > u_0$, the mean of excesses is a linear function of u!

Graphical tool for threshold selection

1. For $u < x_{\mathrm{max}}$, plot u against

$$\frac{1}{N_u} \sum_{i=1}^{N_u} (x_{(i)} - u).$$

2. Choose smallest u above which plot is linear.



How did we get that last result?

$$\mathrm{E}\left(X - u | X > u\right) = \mathrm{E}\left(\mathbf{X} + \mathbf{u_0} - \mathbf{u} - u_0 | \mathbf{X} + \mathbf{u_0} - \mathbf{u} > u_0\right)$$

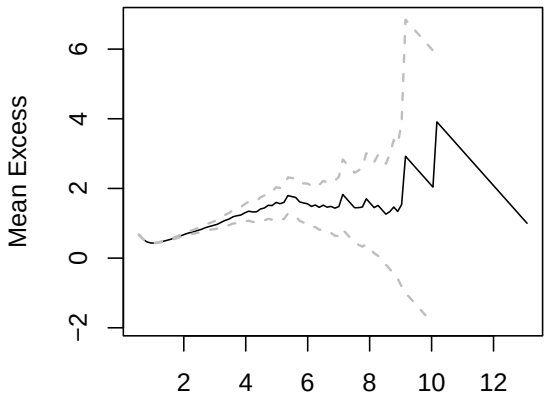
Therefore the location parameter μ in the distribution of M_n has been shifted by $u_0 - u$. We can denote the new location parameter by $\mu' = \mu + u_0 - u$.

$$\begin{aligned}
\implies \mathrm{E}\left(\mathbf{X} + \mathbf{u_0} - \mathbf{u} - u_0 | \mathbf{X} + \mathbf{u_0} - \mathbf{u} > u_0\right) &= \frac{\sigma + \xi(u_0 - \mu')}{1 - \xi} \\
&= \frac{\sigma + \xi(u_0 - \mu - \mathbf{u_0} + \mathbf{u})}{1 - \xi} \\
&= \frac{\sigma + \xi(u_0 - \mu) + \xi(u - u_0)}{1 - \xi} \\
&= \frac{\tilde{\sigma}_0 + \xi(u - u_0)}{1 - \xi}.
\end{aligned}$$

Threshold selection - Mean residual life plot

Example - Fréchet(4)

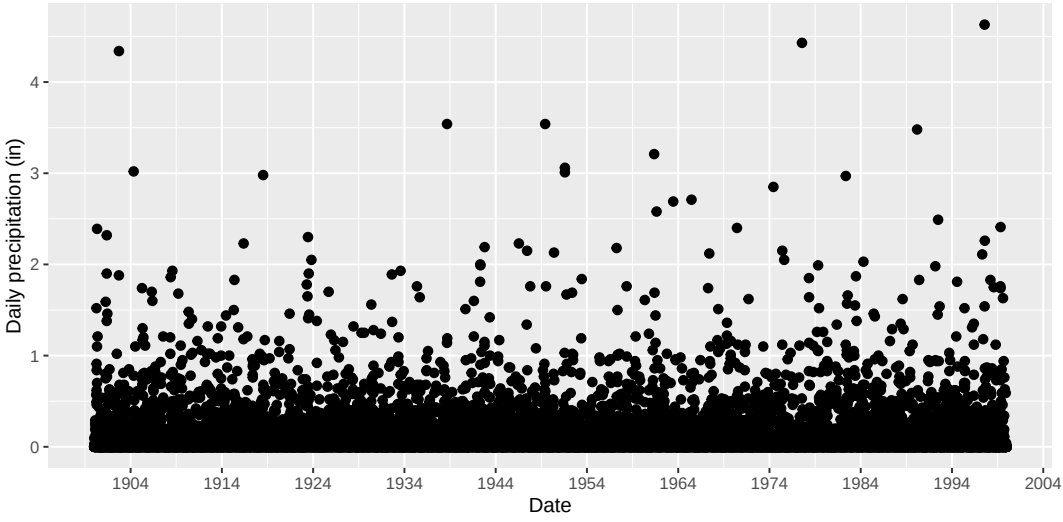
```
mrlplot(dat$data)
```



6. Analysis of precipitation data

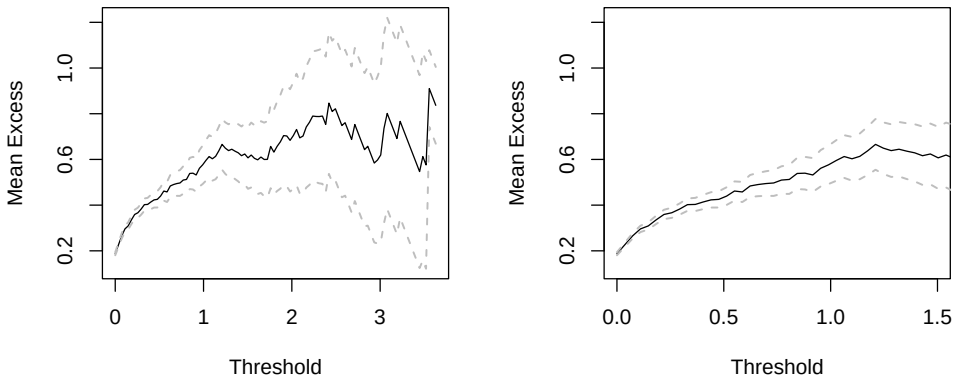


Analysis of precipitation data



Analysis of precipitation data

Step 1: Threshold selection




```
mrlplot(Fort$Prec)
mrlplot(Fort$Prec, xlim=c(0, 1.5))
```

Analysis of precipitation data

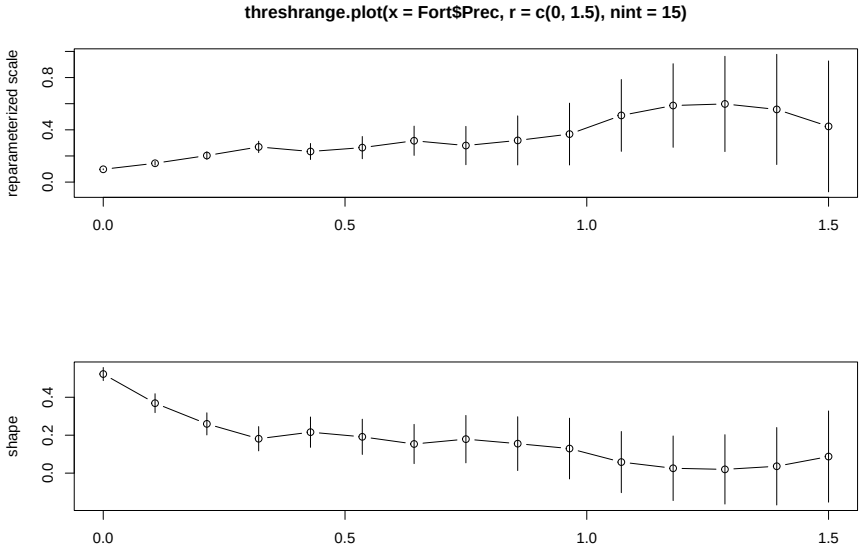
There are only 91 **observations above** 1.5 so we focus on threshold levels below 1.5.

There seem to be some linearity from about ~ 0.5 until ~ 1.25.

We can use some parameter plots to reinforce our intuition.

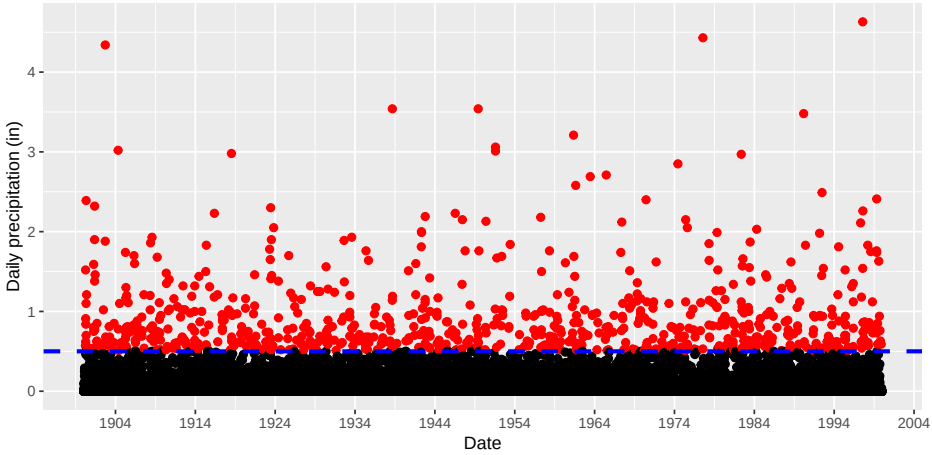
```
threshrange.plot(Fort$Prec, r=c(0,1.5), nint=15)
```

Analysis of precipitation data



Analysis of precipitation data

Selecting the threshold is an art! Here **we select** $u = 0.5$. There are 759 observations above the threshold ($\approx 9.3\%$ of the days with non-zero precipitation).



Step 2: We now fit the GPD model

```
fit.gpd <- fevd(Prec, Fort, threshold=0.5, type="GP", units="inches",
                verbose=FALSE)
# fit.gpd
ci(fit.gpd, type="parameter")
```

```
## fevd(x = Prec, data = Fort, threshold = 0.5, type = "GP", units = "inch
##     verbose = FALSE)
##
## [1] "Normal Approx."
##
##       95% lower CI  Estimate 95% upper CI
## scale    0.3206358 0.3610066    0.4013775
## shape    0.1013195 0.1886400    0.2759605
```

Analysis of precipitation data

The estimated parameters are

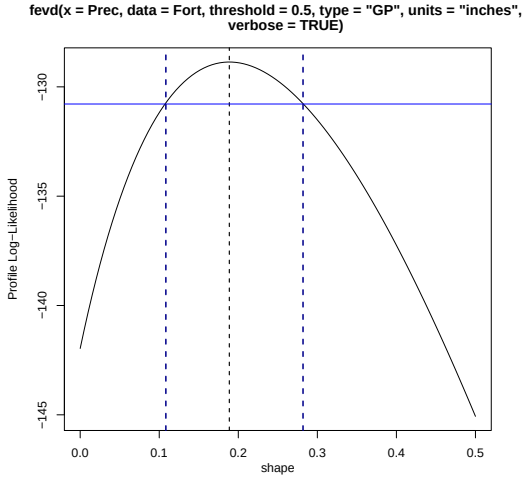
Parameter	Estimate	Std. Error	Conf. Int.
Scale ($\tilde{\sigma}$)	0.361	0.021	$(0.321, 0.628)$
Shape(ξ)	0.189	0.045	$(0.101, 0.276)$

Seems to be heavy tailed!

Let's have a closer look at the shape parameter ξ

```
ci(fit.gpd, type="parameter", which.par=2, xrange=c(0,0.5), nint=100,
   method="proflik", verbose=TRUE)
```

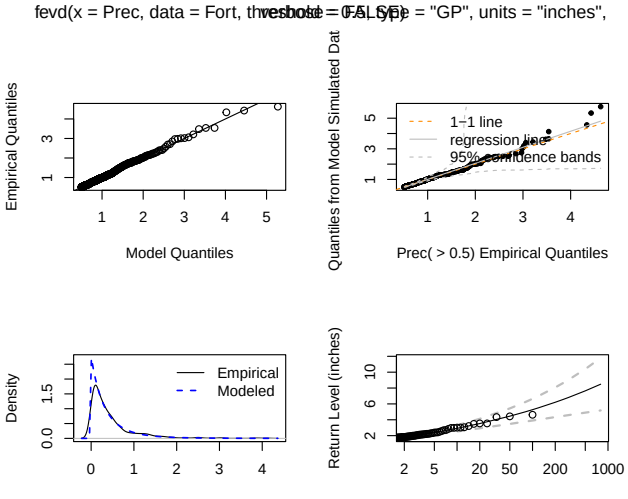
Analysis of precipitation data



Using the profile likelihood confirms that there is a heavy tail phenomena.

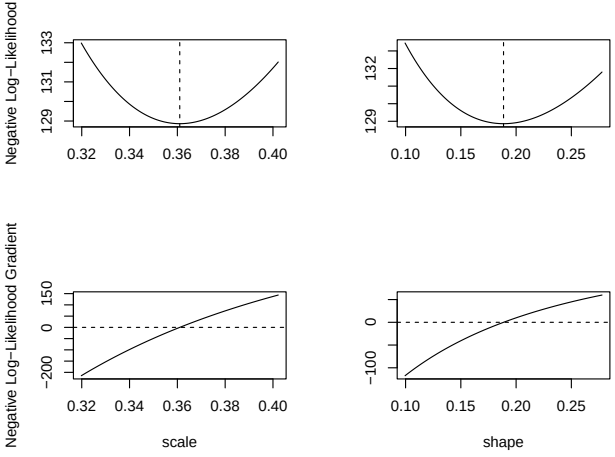
Analysis of precipitation data

We assess the quality of the fit by running some diagnostic plots



Analysis of precipitation data

fevd(x = Prec, data = Fort, threshold = 0.5, type = "GP", units = "inches",



Analysis of precipitation data

The return levels associated with return periods of 20, 50 and 100 years are

Return period	Estimate	Conf. Int.
20-year	3.522	(2.925, 4.119)
50-year	4.453	(3.477, 5.430)
100-year	5.273	(3.905, 6.641)

```
return.level(fit.gpd, return.period=c(20,50,100), do.ci=TRUE)
```

```
## fevd(x = Prec, data = Fort, threshold = 0.5, type = "GP", units = "inches",
##     verbose = FALSE)
##
## [1] "Normal Approx."
##
##                       95% lower CI Estimate 95% upper CI
## 20-year return level      2.924781 3.522081     4.119380
## 50-year return level      3.477014 4.453403     5.429792
## 100-year return level     3.905295 5.272977     6.640659
```

Comment We could fit the GPD model with $\xi = 0$ by specifying `type="Exponential"` in the `fevd()` function, and then do a likelihood ratio test using `lr.test()` to check whether ξ is significantly greater than 0.

Step 4: What is the probability of a daily precipitation above the threshold of 0.5 in to be at least as large as that during the 28 July 1997 flood (4.63 in)?

We are interested in $\mathbb{P}(X - u \geq 4.63 | X > u)$, for $u = 0.5$.

```
fit.gpd.sum <- summary(fit.gpd, silent=TRUE)
pevd(4.63, scale=fit.gpd.sum$par[1], shape=fit.gpd.sum$par[2],
     type="GP",threshold=0.5, lower.tail=FALSE)
```

```
## [1] 0.002251661
```

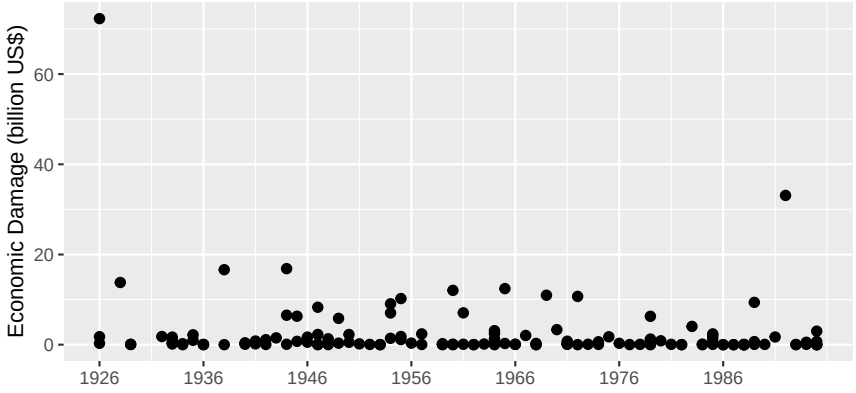
7. Analysis of hurricane damage data

Analysis of hurricane damage data

We focus on the **economic damage caused by hurricanes from 1926 to 1995**.

Trends in societal vulnerability have been removed.

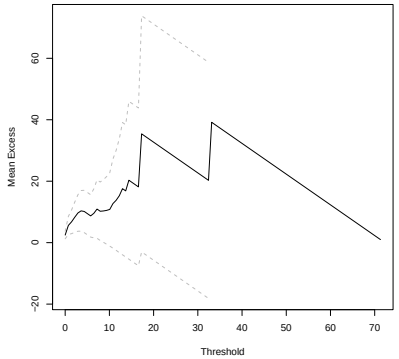
Excess threshold of $u = 6$ billion US$ (More info in **Pielke and Landsea, 1998**).

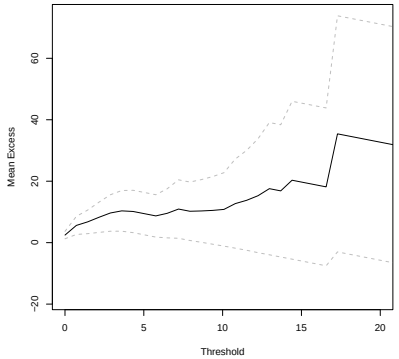


Analysis of hurricane damage data

Step 1: Threshold selection

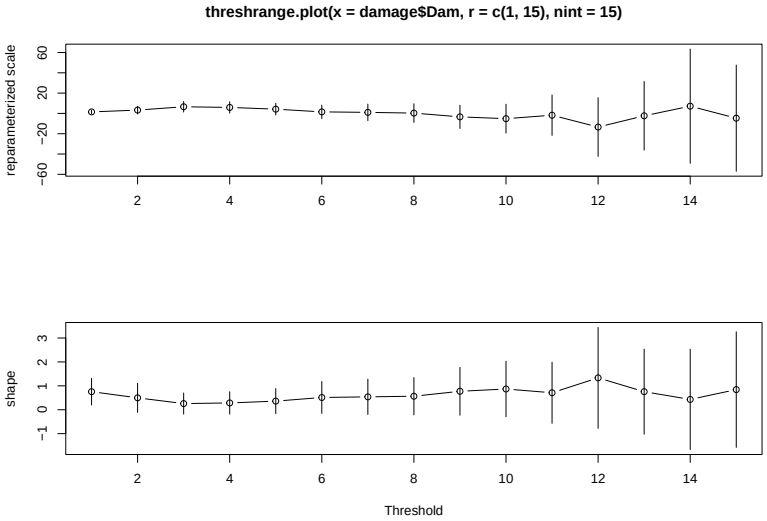
```
data(damage)
mrlplot(damage$Dam)
mrlplot(damage$Dam, xlim=c(0,20))
```





Analysis of hurricane damage data

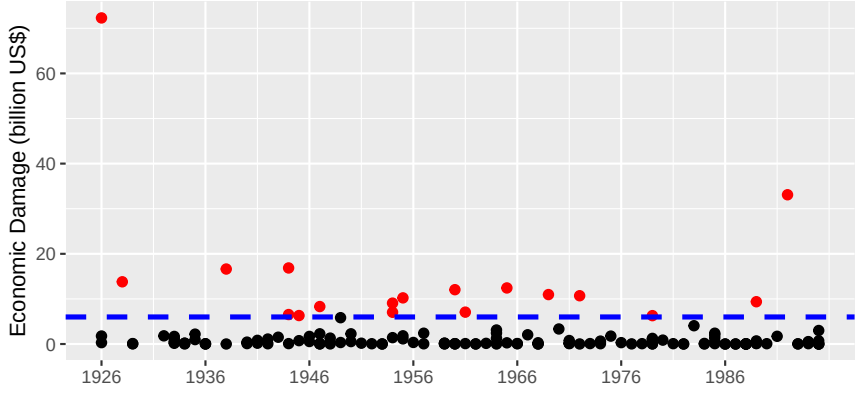
```
threshrange.plot(damage$Dam, r=c(1,15), nint=15)
```



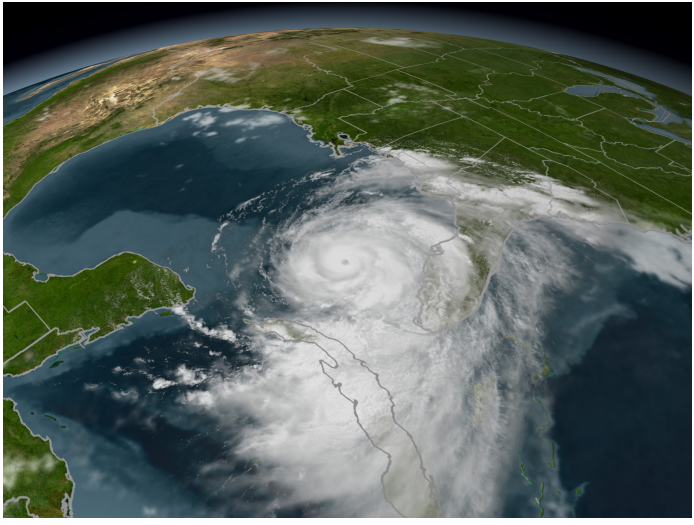
Analysis of hurricane damage data

We choose a low enough threshold to ensure that the variance is not too high, but high enough so that the bias is not too large either (and for the GPD assumptions to hold).

Looks like 6 billion US$ would work; maybe something lower could also work.



Analysis of hurricane damage data



Hurricane Dennis (2005): cause **at least 89 deaths** and **2.23 billion US$** in damage.

$\Longrightarrow$ Impactful despite being under the 6 billion US$ threshold.

Analysis of hurricane damage data

In the previous examples we had daily observations and we were interested in return levels with years as units. This was **convenient because it corresponded to the default setting in the `fevd()` function!**

Here we do not have daily hurricanes (**Luckily!**), so we need to define the `time.units` argument in `fevd()` .

It looks like we have at most 5 observations per year (**better than daily hurricanes!**), and the data is recorded over 70 years (1926 to 1995), therefore we have an **average of approx.** 2.05 **hurricanes per year**.

$$\implies b_s \approx 2.05 : \text{number of observations per time period (year)}$$

Analysis of hurricane damage data

```
nh <- table(damage$Year)
nh
```

```
##
## 1926 1928 1929 1932 1933 1934 1935 1936 1938 1940 1941 1942 1943 1944 1945 1946
##    3    1    2    1    4    3    2    3    2    2    2    2    1    3    2    2
## 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1959 1960 1961 1962 1963
##    5    3    2    2    1    1    2    3    3    1    2    4    4    2    2    1
## 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979
##    5    2    2    1    3    1    1    4    2    1    2    1    1    1    1    5
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##    1    1    2    1    2    5    2    2    5    4    1    1    1    2    3    5
```

```
ny <- max(damage$Year) - min(damage$Year) + 1
nh <- sum(nh)/ny
nh
```

```
## [1] 2.057143
```

Step 2: We now fit the Exponential model ($\xi = 0$)

```
fit.h0 <- fevd(Dam, damage, threshold=6, type="Exponential",
               time.units="2.05/year")
# fit.h0
ci(fit.h0, type="parameter")
```

```
## fevd(x = Dam, data = damage, threshold = 6, type = "Exponential",
##     time.units = "2.05/year")
##
## [1] "Normal Approx."
##
## [1] "scale: 8.948"
##
## [1] "95% Confidence Interval: (4.8146, 13.0824)"
```

Analysis of hurricane damage data

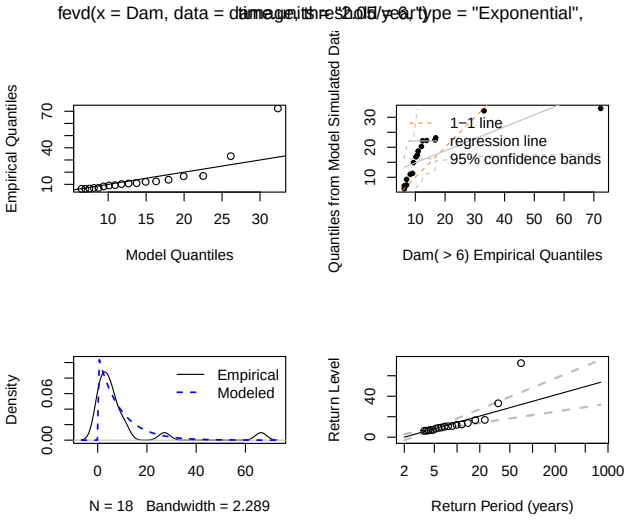
The estimated parameters are

Parameter	Estimate	Std. Error	Conf. Int.
Scale ($\tilde{\sigma}$)	8.948	2.109	$(4.825, 13.082)$

We assess the quality of the fit by running some diagnostic plots

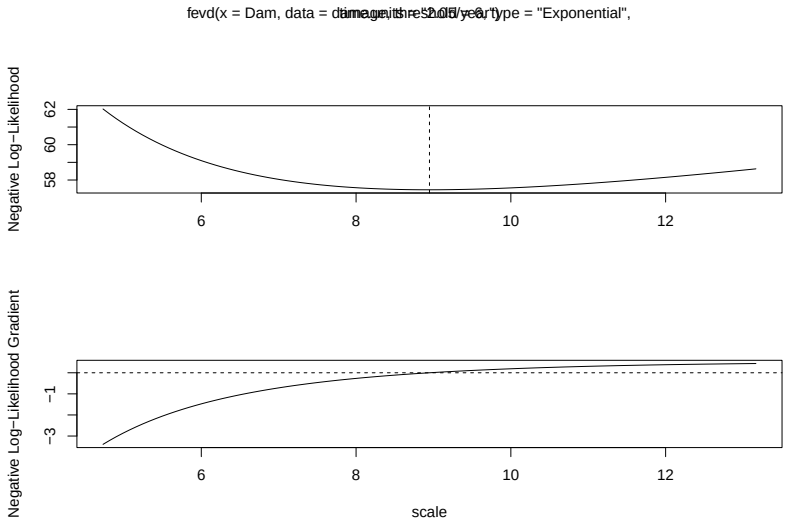
Analysis of precipitation data

```
plot(fit.h0)
```



Analysis of precipitation data

```
plot(fit.h0, "trace")
```



Analysis of precipitation data

The return levels associated with return periods of 20, 50 and 100 years are

Return period	Estimate	Conf. Int.
20-year	20.623	(13.867, 27.379)
50-year	28.822	(18.279, 39.366)
100-year	35.025	(21.616, 48.434)

```
return.level(fit.h0, return.period=c(20,50,100), do.ci=TRUE)
```

```
## fevd(x = Dam, data = damage, threshold = 6, type = "Exponential",
##     time.units = "2.05/year")
##
## [1] "Normal Approx."
##
##                       95% lower CI Estimate 95% upper CI
## 20-year return level      13.86699 20.62302     27.37904
## 50-year return level      18.27914 28.82244     39.36575
## 100-year return level     21.61641 35.02507     48.43374
```

Step 3: We fit the GPD model

```
fit.h1 <- fevd(Dam, damage, threshold=6, type="GP",
               time.units="2.05/year")
# fit1
ci(fit.h1, type="parameter")
```

```
## fevd(x = Dam, data = damage, threshold = 6, type = "GP", time.units = "2.05/year")
##
## [1] "Normal Approx."
##
##       95% lower CI  Estimate 95% upper CI
## scale     1.027072 4.5889338     8.150795
## shape    -0.154795 0.5123431     1.179481
```

Analysis of hurricane damage data

The estimated parameters are

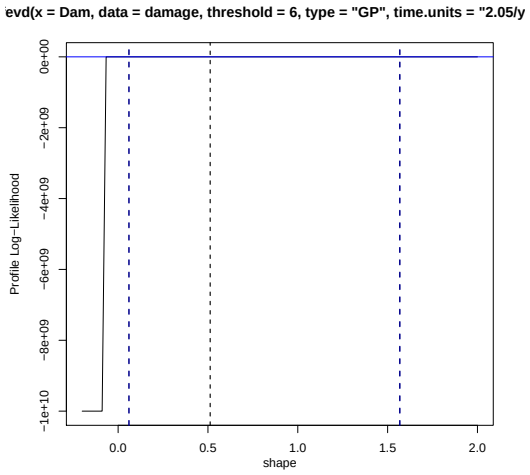
Parameter	Estimate	Std. Error	Conf. Int.
Scale ($\tilde{\sigma}$)	4.589	1.817	$(1.027, 8.151)$
Shape(ξ)	0.512	0.340	$(-0.155, 1.179)$

Another example where it seems to be heavy tailed!

Let's have a closer look at the shape parameter ξ

```
ci(fit.h1, type="parameter", which.par=2, xrange=c(-0.2,2), nint=100,
   method="proflik", verbose=TRUE)
```

Analysis of precipitation data



Using the profile likelihood confirms that there is a heavy tail phenomena.

Analysis of precipitation data

The return levels associated with return periods of 20, 50 and 100 years are

Return period	Estimate	Conf. Int.
20-year	17.733	(9.669, 25.798)
50-year	30.129	(8.530, 51.728)
100-year	44.235	(0.172, 88.299)

```
return.level(fit.h1, return.period=c(20,50,100), do.ci=TRUE)
```

```
## fevd(x = Dam, data = damage, threshold = 6, type = "GP", time.units = "2.05/year")
##
## [1] "Normal Approx."
##
##                        95% lower CI Estimate 95% upper CI
## 20-year return level       9.668635 17.73311     25.79758
## 50-year return level       8.529906 30.12888     51.72785
## 100-year return level      0.171544 44.23543     88.29932
```

Analysis of precipitation data

- The estimated the return levels seem more appropriate than those obtained from the Exponential model (**too low!**)
- The confidence intervals are **very large**.
- Might be (partly) explained by the low number of observations above the threshold (18).
- Let's consider the estimated return level obtained from the **profile likelihood**.

Return period	Estimate	Conf. Int.
20-year	17.733	(20.366, 36.087)
50-year	30.129	(19.058, 126.207)
100-year	44.235	(32.531, 353.127)

- Lower bounds of CIs seem more appropriate, very conservative upper bounds.

Analysis of precipitation data

```r
return.level(fit.h1, return.period=20, method="proflik", do.ci=TRUE)
```

```
## fevd(x = Dam, data = damage, threshold = 6, type = "GP", time.units = "2.05/yea
##
## [1] "Profile Likelihood"
##
## [1] "20-year return level: 17.733"
##
## [1] "95% Confidence Interval: (20.3665, 36.0866)"
```

```r
# return.level(fit.h1, return.period=50, method="proflik", do.ci=TRUE)
# return.level(fit.h1, return.period=100, method="proflik", do.ci=TRUE)
```

Step 4: We test if $\xi = 0$

We perform a likelihood ratio test: $H_0 : \xi = 0$ (Gumbel type) against $H_1 : \xi > 0$

```
lr.test(fit.h0, fit.h1)
```

```
##
##  Likelihood-ratio Test
##
## data:  DamDam
## Likelihood-ratio = 5.5966, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.018
## alternative hypothesis: greater
```

Since p-value ≈ 0.018 **we reject the null hypothesis** (this was expected based on the CI for ξ).

References



Sources and recommending reading

1. **[dHF06]** de Haan, L. & Ferreira, A. (2006) Extreme Value Theory: An Introduction. New York: Springer.
2. **[C01]** Coles, S. (2001) An Introduction to Statistical Modeling of Extreme Values. Springer. Chapter 1-3.
3. **Katz, R. W., Parlange, M. B. and Naveau, P. (2002)**. Statistics of extremes in hydrology, *Advances in Water Resources*, 25, pp.1287–1304.
4. **Pielke, R. A. Jr. and Landsea, C. W. (1998)**. Normalized Hurricane Damages in the United States: 1925-95, *Weather and Forecasting*, 13(3), pp.621–631.