Introduction to extreme event analysis

Stéphane Girard

Inria Grenoble Rhône-Alpes
`http://mistis.inrialpes.fr/~girard`

June 2024

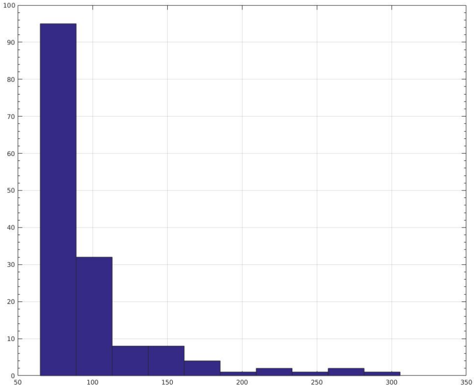
Outline

Outline

Extreme rainfall in Montpellier, France, 2014. Three hours of rainfall $= 50\%$ of mean annual rainfall (source `http://www.dailymail.co.uk`).

Nidd river (England)

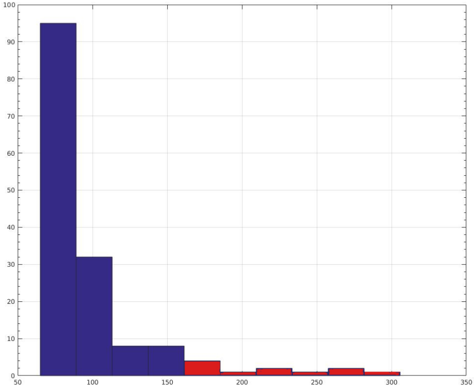
The flows of the Nidd river have been measured every three months during 38.5 years. This results in $n = 154$ measures, the maximum observed flow being about $300 m^3/s$.



Nidd river (England)

From these data, it is possible to compute some exceedance probabilities such as:

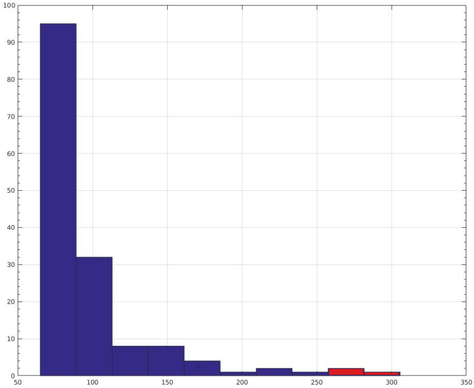
$$\mathbb{P}(X \geq 160) \simeq nb(X_i \geq 160)/n = 11/154 = 1/14$$



Nidd river (England)

From these data, it is possible to compute some exceedance probabilities such as:

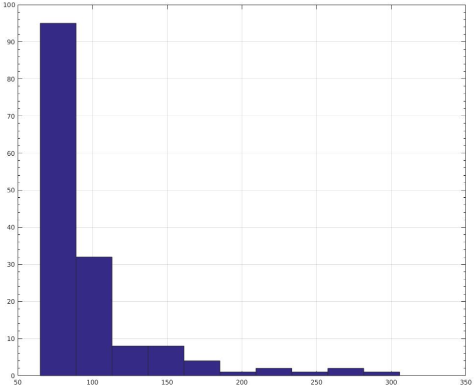
$$\mathbb{P}(X \geq 255) \simeq nb(X_i \geq 255)/n = 3/154$$



Nidd river (England)

From these data, it is possible to compute some exceedance probabilities ... but not extreme ones!

$$\mathbb{P}(X \geq 500) \simeq nb(X_i \geq 500)/n = 0$$



Statistical formulation

- Let $X_1, \ldots, X_n$ be n independent and identically distributed (iid) random variables with cumulative distribution function (cdf) F and survival function $\bar{F} = 1 - F$.
- The maximum is denoted by $X_{n,n} = \max(X_1, \ldots, X_n)$. The cdf of $X_{n,n}$ is given by $\mathbb{P}(X_{n,n} \leq x) = F^n(x)$.
- As illustrated in the previous slides, the estimation of $\bar{F}(s) = \mathbb{P}(X > s)$ for $s > X_{n,n}$ is a non-trivial problem which is referred to as the **estimation of small tail probabilities**. As shown in the next slides, the two classical statistical approaches are ill-adapted.

Nonparametric approach

- $\bar{F}(s) = \mathbb{P}(X \geq s)$ is estimated by the proportion of observations above the threshold s:

$$\hat{\bar{F}}_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq s\},$$

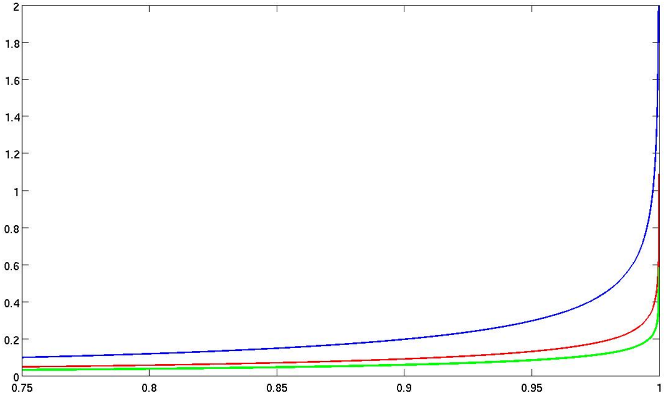
which is referred to as the **empirical survival function**.

- However, as illustrated on the Nidd river: $\hat{\bar{F}}_n(s) = 0$ if $s > X_{n,n}$.

Parametric approach

- Suppose *a priori* a parametric model for the survival function $\bar{F} \in \{\bar{F}_\theta,\ \theta \in \Theta\}$,
- Estimate θ by $\hat{\theta}_n$ and estimate $\bar{F}(s)$ by $\bar{F}_{\hat{\theta}_n}(s)$.
- However, a good fit on the sample does not necessarily lead to a good modeling above the maximum.
- This phenomena is illustrated on the next figure: Even though the Student t_ν distribution converges to a $\mathcal{N}(0,1)$ distribution as $\nu \to \infty$, their large quantiles remain different.

Illustration



Horizontally: p. Vertically: relative error between the quantile of order p estimated from $\mathcal{N}(0,1)$ and Student t_4, Student t_8, Student t_{12} distributions.

Outline

Goal

- The **Central Limit Theorem** (CLT) provides the asymptotic distribution of the mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

of a sample $\{X_1, \ldots, X_n\}$ of iid random variables:

$$\sqrt{n}\left(\frac{\bar{X}_n - \mathbb{E}(X)}{\sigma(X)}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Equivalently, in terms of cdf,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\left(\frac{\bar{X}_n - \mathbb{E}(X)}{\sigma(X)}\right) \leq x\right) = \Phi(x),$$

where Φ is the cdf of the $\mathcal{N}(0,1)$ distribution.

- The **Extreme-Value Theorem** is a simular result for the maximum.

Extreme-Value Theorem

Under mild conditions on F, there exist three parameters a_n, b_n and γ such that:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - a_n}{b_n} \leq x\right) = H_\gamma(x),$$

with

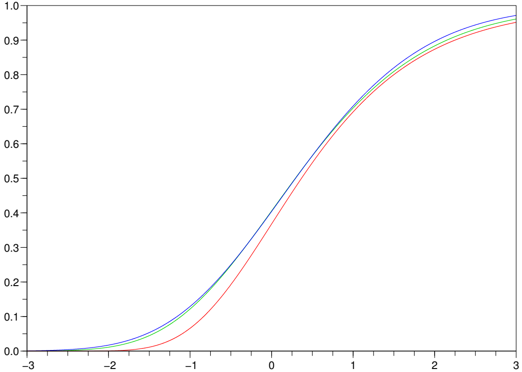
$$H_\gamma(x) = \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right)$$

and where $y_+ = \max(0, y)$.

- H_γ is the **Extreme-Value Distribution (EVD)**,
- γ is the **extreme-value parameter**,
- a_n are b_n normalization parameters.

We note $F \in \mathrm{MDA}(H_\gamma)$.

Illustration on a Gaussian distribution



Comparaison between $H_\gamma(x)$, $\mathbb{P}\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right)$ with $n = 10$ and $\mathbb{P}\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right)$ with $n = 100$.

Extreme-Value Distribution

In practice,

$$\mathbb{P}(X_{n,n} \leq x) \simeq H_\gamma\left(\frac{x - a_n}{b_n}\right),$$

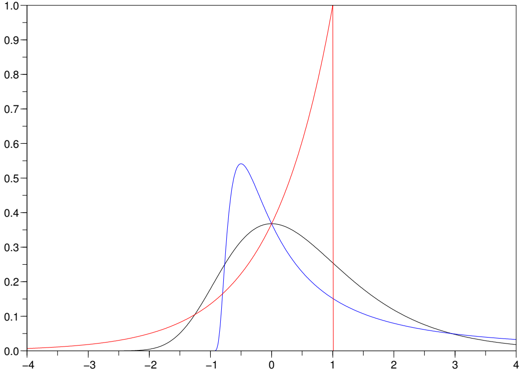
on has a three-parameter distribution

- a_n is a **location parameter**, playing the role of $\mathbb{E}(X)$ in the TCL,
- b_n is a **scale parameter**, playing the role of $\sigma(X)/\sqrt{n}$ in the TCL,
- γ is a **shape parameter**, without any equivalent in the TCL.

Three cases appear (and thus three kind of distributions):

- If $\gamma > 0$, F is said to belong to the maximum domain of attraction of Fréchet.
- If $\gamma = 0$, F is said to belong to the maximum domain of attraction of Gumbel.
- If $\gamma < 0$, F is said to belong to the maximum domain of attraction of Weibull.

Extreme-Value Distribution



Examples of densities associated with the Extreme-Value Distribution ($\gamma = 0$, $\gamma = 1$ and $\gamma = -1$).

Maximum Domain of Attractions (MDA)

MDA	**Gumbel** $\gamma = 0$	**Fréchet** $\gamma > 0$	**Weibull** $\gamma < 0$
Tail	Light	Heavy	Short
Distribution	Gaussian Exponential Lognormal Gamma Weibull	Cauchy Pareto Student	Uniform Beta

Application to extrapolation

Since $\mathbb{P}(X_{n,n} \leq x) = F^n(x)$, the Extreme-Value Theorem yields an approximation of $F(x)$ for large values of x,

$$F(x) = 1 - \bar{F}(x) \simeq H_\gamma^{1/n}\left(\frac{x - a_n}{b_n}\right).$$

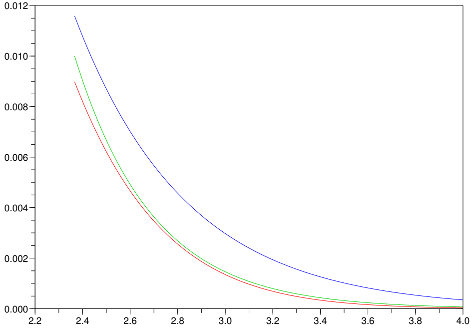
Taking the logarithm, it follows

$$\log(1 - \bar{F}(x)) \simeq \frac{1}{n} \log H_\gamma\left(\frac{x - a_n}{b_n}\right).$$

Since x is large, $\bar{F}(x)$ is small, a first order expansion of $\log(1 + u)$ entails

$$\bar{F}(x) \simeq -\frac{1}{n} \log H_\gamma\left(\frac{x - a_n}{b_n}\right) = \frac{1}{n}\left[1 + \gamma\left(\frac{x - a_n}{b_n}\right)\right]^{-1/\gamma}.$$

Illustration on a Gaussian distribution



Comparaison between $\bar{F}(x)$, $-\frac{1}{n}\log H_\gamma\left(\frac{x-a_n}{b_n}\right)$ with $n = 10$ and $-\frac{1}{n}\log H_\gamma\left(\frac{x-a_n}{b_n}\right)$ with $n = 100$.

Here, the theoretical values of a_n, b_n and γ have been used (they are known for the standard Gaussian distribution). **In practice, they are unknown (since F is unknown) and have to be estimated.**

Estimation of EVD parameters

Let $\{Y_1, \ldots, Y_k\}$ be a sample of k iid random variables from the extreme-value cdf $H_{\gamma,a,b}$. This starting point may be a problem since, usually, one does not observe directly a sample of maxima. They have to be extracted from an initial sample $\{X_1, \ldots, X_n\}$ with arbitrary cdf F. The common practice is to divide the data into k blocks and to extract the maxima of each block (*block-maxima method*).

Two drawbacks:

- Loss of information (n data $\longrightarrow$ k data),
- The maxima are not exactly distributed from an EVD distribution.

Estimation of EVD parameters

The cdf of the EVD distribution is given by:

$$H_{\gamma,a,b}(x) \stackrel{def}{=} H_\gamma\left(\frac{x-a}{b}\right) = \exp\left\{-\left[1+\gamma\left(\frac{x-a}{b}\right)\right]_+^{-1/\gamma}\right\}.$$

Maximum-likelihood estimators:

- not closed-form (optimization procedure).
- Consistency if $\gamma > -1$,
- Asymptotic normality if $\gamma > -1/2$ (Fisher information matrix available).

Selection of a sample of maxima

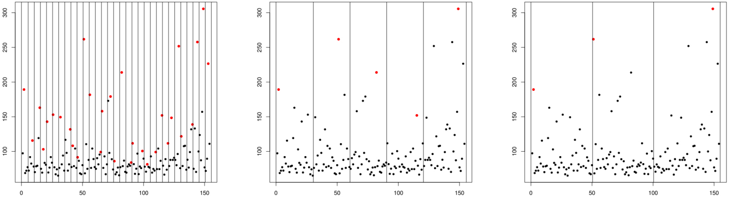


Illustration on Nidd data. Three block sizes are considered, the block maxima are depicted in red.

Application to Nidd data

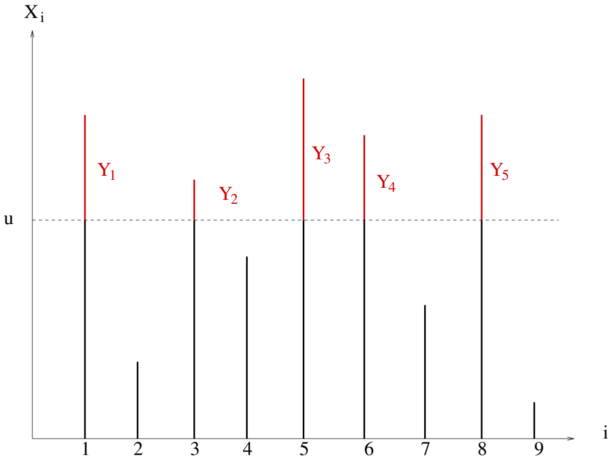
```
> # Fit GEV to the data in nidd.annual, the annual maximum water
> # levels of the River Nidd, using the "BFGS" optimization method
> gevnidd <- gev(nidd.annual, method = "BFGS", control = list(maxit = 500))

$data
 [1]  65.08  65.60  75.06  76.22  78.55  81.27  86.93  87.76  88.89  90.28
[11]  91.80  91.80  92.82  95.47 100.40 111.54 111.74 115.52 131.82 138.72
[21] 148.63 149.30 151.79 153.04 158.01 162.99 172.92 179.12 181.59 189.04
[31] 213.70 226.48 251.96 261.82 305.75

$par.ests
       xi      sigma         mu
 0.321221  36.154177 103.118249

$varcov
            [,1]       [,2]       [,3]
[1,]  0.04758274 -0.4142656 -0.7770124
[2,] -0.41426557 43.6098796 35.7316149
[3,] -0.77701236 35.7316149 58.0116406
```

MLE estimation of the GEV parameters with R using the *block-maxima method* on Nidd data. Annual blocks are considered so that $k = 35$ block-maxima are obtained. The estimated parameters are provided as well as the asymptotic covariance matrix.

Outline

Excess

Let X be random variable with cdf F. For all $u \in \mathbb{R}$ and $x \geq 0$, the cdf F_u defined by $F_u(x) = \mathbb{P}(X - u \leq x | X > u)$ is called the cdf of Y, the excess of X above u.



Generalized Pareto Distribution

The Generalized Pareto Distribution (GPD) with parameters $\gamma \in \mathbb{R}$ and $\sigma > 0$ is defined by the cdf

$$G_{\gamma,\sigma}(y) = 1 - \left(1 + \gamma \frac{y}{\sigma}\right)_+^{-1/\gamma}.$$

Particular cases:

- If $\gamma = 0$, then $G_{0,\sigma}$ is the cdf of Exp(σ),
- If $\gamma = -1$, then $G_{-1,\sigma}$ is the cdf of Unif$[0, \sigma]$,
- If $\gamma > 0$, then $G_{\gamma,\sigma}$ is the cdf of a translated Pareto distribution.

Pickands theorem

$F \in$ MDA (H_γ) if and only if there exists a positive function $a(\cdot)$ such that

$$\lim_{u \to \infty} \sup_{y>0} \left| F_u(y) - G_{\gamma, a(u)}(y) \right| = 0.$$

Interpretation:

- The excess is asymptotically distributed from a GPD if and only if the maximum is asymptotically distributed from an EVD.
- In both cases, the shape parameter γ is the extreme-value index.

Application to extrapolation

From Pickands theorem, we have, for all $y \geq 0$,

$$\bar{F}_u(y) = \frac{\bar{F}(u+y)}{\bar{F}(u)} \simeq \bar{G}_{\gamma,\sigma}(y),$$

where $\bar{G}_{\gamma,\sigma}$ is the survival function of the GPD. The change of variable $x = u + y$ yields for all $x \geq u$:

$$\bar{F}(x) \simeq \bar{F}(u)\bar{G}_{\gamma,\sigma}(x-u).$$

Finally, introducing $\alpha = \mathbb{P}(X > u) = \bar{F}(u)$, it follows

$$\bar{F}(x) \simeq \alpha \bar{G}_{\gamma,\sigma}(x - \bar{F}^{-1}(\alpha)).$$

We thus obtain an approximation of the small tail probabilities:

$$\bar{F}(x) \simeq \alpha \left[1 + \gamma \left(\frac{x - \bar{F}^{-1}(\alpha)}{\sigma}\right)\right]^{-1/\gamma}.$$

Comparison between EVD and GPD approaches

For both approaches, the approximations are the same:

$$
\begin{aligned}
\bar{F}(x) &\simeq \alpha\left[1+\gamma\left(\frac{x-\bar{F}^{-1}(\alpha)}{\sigma}\right)\right]^{-1/\gamma} \quad \text{(GPD)}\\
\bar{F}(x) &\simeq \frac{1}{n}\left[1+\gamma\left(\frac{x-a_n}{b_n}\right)\right]^{-1/\gamma} \quad \text{(EVD)}
\end{aligned}
$$

There are three parameters to be estimated:

- the extreme value index γ (shape parameter),
- a scale parameter σ (GPD) or b_n (EVD),
- a position parameter $\bar{F}^{-1}(\alpha)$ (GPD) or a_n (EVD).

Estimation of GPD parameters

Let $\{Y_1, \ldots, Y_k\}$ be a sample of k iid random variables from the Generalized Pareto cdf $G_{\gamma,\sigma}$. Similarly to EVD case, this starting point may be a problem since, usually, one does not observe direclty a sample of excesses. They have to be extracted from an initial sample $\{X_1, \ldots, X_n\}$ with arbitrary cdf F. The common practice is to choose a number of k excesses and select

$$\{Y_1, \ldots, Y_k\} := \{X_{n-k+1,n} - X_{n-k,n}, \ldots, X_{n,n} - X_{n-k,n}\}$$

(*peaks over threshold method*).

Advantages / drawbacks:

- The excesses are not exactly distributed from a GPD distribution,
- Dependence issues,
- $\alpha = k/n$, the position parameter $\bar{F}^{-1}(k/n)$ is estimated by the empirical quantile $X_{n-k,n}$. It only remains to estimate γ and σ (by Maximum likelihood for instance).

Selection of a sample of excesses

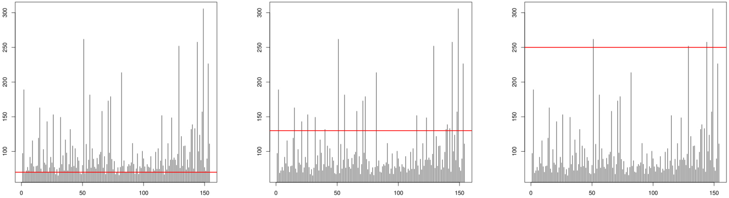


Illustration on Nidd data. Three numbers of excesses k are considered. The associated thresholds are depicted in red.

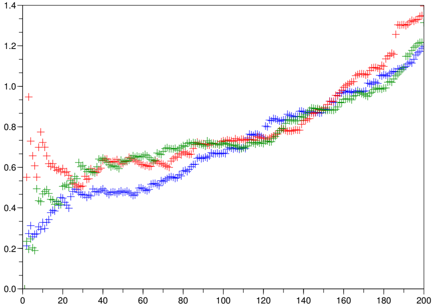
Illustration on Nidd data

```
> # Fit GPD to the data in nidd.thresh
> gpdnidd <- gpd(nidd.thresh, 100)

$data
 [1] 189.02 115.52 119.28 162.99 102.92 143.06 153.04 149.30 116.77 131.82
[11] 107.97 104.19 261.82 110.48 181.59 104.19 158.01 172.92 179.12 213.70
[21] 111.74 100.40 104.19 151.79 111.54 148.63 251.96 121.73 107.58 108.14
[31] 131.92 138.72 133.06 257.62 123.71 157.12 305.75 226.48 110.98

$threshold
[1] 100

$p.less.thresh
[1] 0.7467532

$n.exceed
[1] 39

$par.ests
          xi         beta
 0.003508321 50.608623759

$varcov
             [,1]       [,2]
[1,]  0.04562003  -2.303872
[2,] -2.30387212 182.476944
```

MLE estimation of the GPD parameters with R. The threshold $u = 100$ is considered so that $k = 39$ excesses are obtained. The estimated parameters are provided as well as the asymptotic covariance matrix.

In practice, the choice of k is difficult ...

- If k is small, $\hat{\gamma}(k)$ is based on few observations, it has therefore a **large variance**.
- If k is large, then $X_{n-k,n}$ is no longer in the distribution tail and the approximation of the excesses distribution by a GPD is no longer true, $\hat{\gamma}(k)$ has a **large bias**.



Estimation of γ on three samples of size $n = 500$ from a t_2 distribution ($\gamma = 1/2$). Horizontally: k. Vertically: $\hat{\gamma}(k)$ for $k = 1, \ldots, 200$.

Outline

Time dependent extremes

Data $X_1(\cdot), \ldots, X_n(\cdot)$ are stochastic processes.
Probabilistic problem: Find two sequences of continuous functions $a_n(\cdot)$ and $b_n(\cdot)$ such that

$$\left\{ \frac{\max_{i=1,\ldots,n} X_i(t) - a_n(t)}{b_n(t)} \right\}_{t\in[0,1]}$$

converges in distribution to a stochastic process $\{Y(t)\}_{t\in[0,1]}$. From the univariate theory, the margin at point $t \in [0,1]$ is an EVD with shape parameter $\gamma(t)$. In practice, for each time $t \in [0,1]$:

$$\mathbb{P}\left(\max_{i=1,\ldots,n} X_i(t) \leq x \right) \simeq H_{\gamma(t)} \left(\frac{x - a_n(t)}{b_n(t)} \right).$$

Two statistical problems:

- Estimations of functions $\gamma(\cdot)$, $a_n(\cdot)$ and $b_n(\cdot)$.
- Estimation of the dependence structure

$$\mathbb{P}\left(\max_{i=1,\ldots,n} X_i(t_1) \leq x_1 \cap \max_{i=1,\ldots,n} X_i(t_2) \leq x_2 \right).$$

Not discussed here, cf Chapter 9 of de Haan & Ferreira (2006).

Estimation of extreme-value parameters as functions of covariates

Parametric approach

- Chavez-Demoulin & Davison (2005): GEV + linear model of the trend (expansion on a splines basis), penalized maximum likelihood.

Semi-parametric approach

- Davison & Ramesh (2000) and Hall & Tajvidi (2000): GEV/GPD + nonlinear model of the trend (local polynomial smoothing).

Nonparametric approach

- Daouia *et al* (2013): Domain of attraction condition + kernel estimators of extreme-value parameters.
- Einmahl *et al* (2016) and Ahmad *et al* (2020): restriction to heavy-tails and $\gamma(\cdot)$ constant.

Outline

Reference books

S. Coles (2001), An introduction to statistical modeling of extreme values, *Springer.*

P. Embrechts, C. Klüppelberg and T. Mikosch (1997), Modelling extremal events, *Springer.*

L. de Haan and A. Ferreira (2006), Extreme Value Theory, Springer.

Time dependent extremes

Ahmad, A., Deme, E., Diop, A., Girard, S. and Usseglio-Carleve, A. (2020). Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model, *Electronic Journal of Statistics*, 14, 4421–4456, 2020.

Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society C*, 54, 207–222.

Daouia, A., Gardes, L. and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19, 2557–2589.

Davison, A. and Ramesh, N. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society B*, 62, 191–208.

Einmahl, J., de Haan, L. and Zhou, C. (2016). Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society B*, 78, 31–51.

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, 15, 153–167.