Statistical Learning with Extreme Values
Lecture 1

Master's program MVA, ENS Paris-Saclay

Anne Sabourin (with Stephan Clémençon)
Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

October, 2023.

Outline

Why bother about extremes?

'Il est impossible que l'improbable n'arrive jamais

Figure: *Emil Julius Gumbel, 1891-1966*

For risk management:

Measuring a risk (probability of occurrence) is the first step before implementing prevention measures

Natural Hazards

- Exceptional wave heights

Figure: Storm Xynthia, La Faute-Sur-Mer, march 1st 2010.

Natural Hazards

- torrential rain

Figure: Ouagadougou, 2009.

Natural Hazards

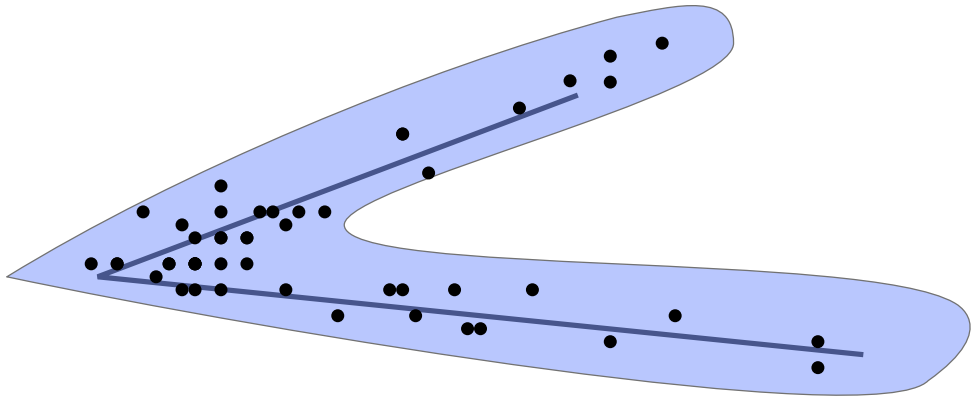
- Floods

Figure: 1934 flood at Port Pirie, Australia

Financial risk

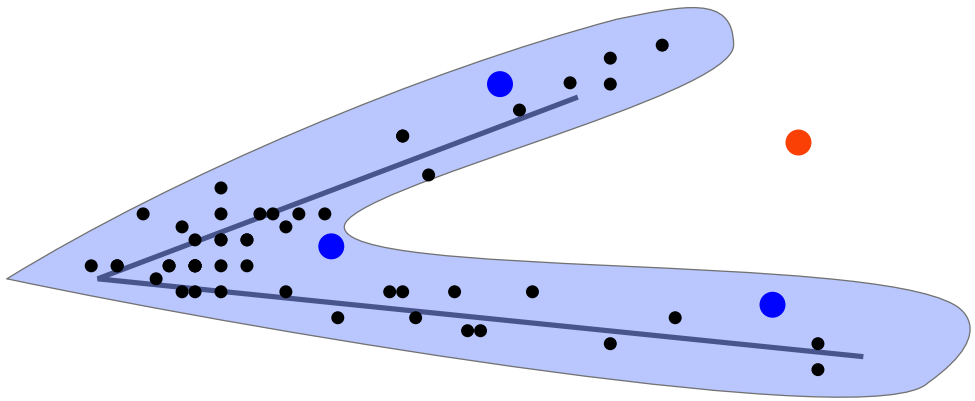
- Large losses
- Large claims (insurances), ...

Applications to anomaly detection



- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)

Applications to anomaly detection



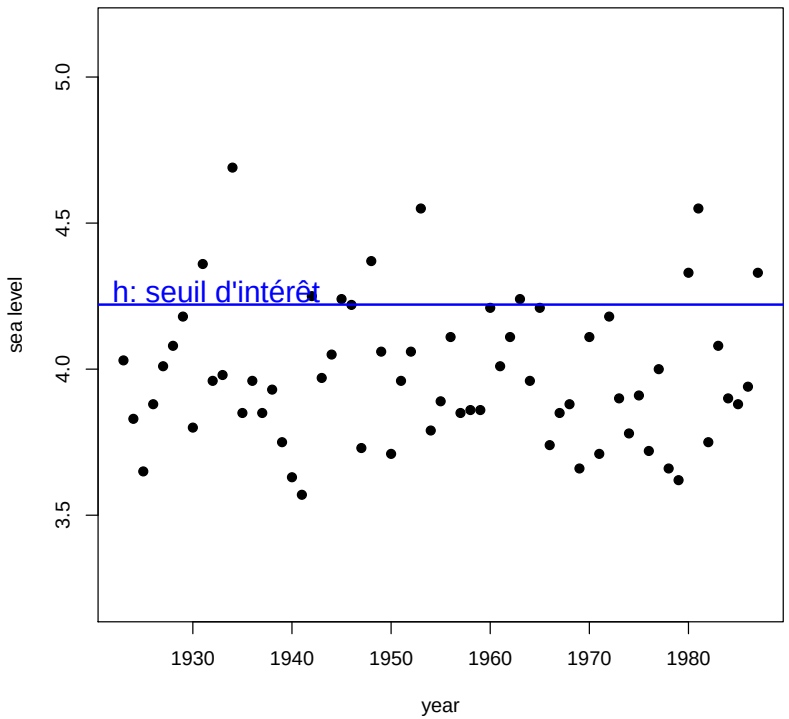
- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)
- **Prediction step:** (with new data)
  Anomalies = points outside the 'normal region'

If 'normal' data are heavy tailed, **Abnormal** $\not\Leftrightarrow$ **Extreme** .
There may be **extreme** 'normal data'.

How to distinguish between large anomalies and normal extremes?

Threshold exceedances: questions from risk management

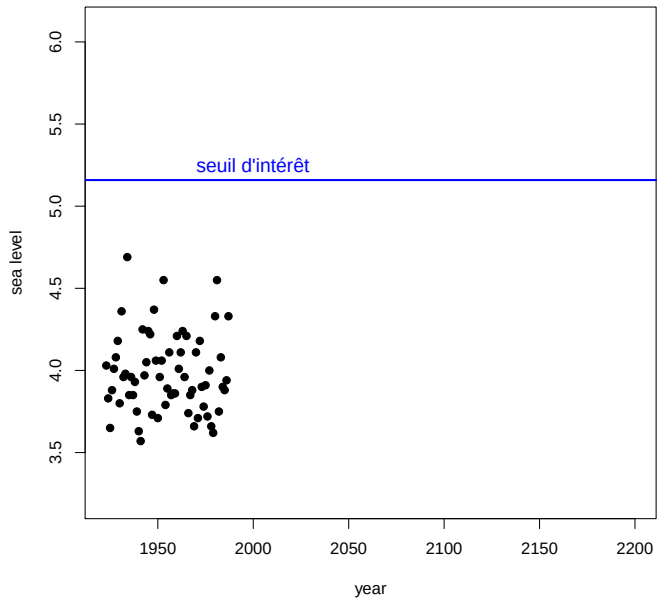
Quantity of interest: X (water level, temperature, insurance claims, ...)
$\longrightarrow$ *i.i.d.* (independent and identically distributed) time series $X_t, t \geq 0$.



- Given a high threshold h, find $p = \mathbb{P}(X \geq h)$
- Given p (e.g. $p = 10^{-4}$), find h such that $\mathbb{P}(X > h) \leq p$.
- Given a long duration T (e.g. 10^4), find $P(\max_{t \leq T} X_t \leq h)$.

Beyond the range of data

For $h \gg \max(X_{\text{obs}})$, or $T \gg T_{\text{obs}}$, or $p \ll 1/N_{\text{obs}}$ too small :



Empirical estimator $\quad \hat{P}(X > h) = \dfrac{1}{N_{\text{obs}}} \displaystyle\sum_{i=1}^{N_{\text{obs}}} \mathbb{1}_{X_i > h} = 0 \quad$!!

Need an extrapolation model

Return level / Value at Risk / Quantile

- Terminology:
  - Hydrology $\rightarrow$ 'Return level'
  - Finance: 'Value at Risk'
  - Statistics: 'Quantile'

  The return level/ VaR / Quantile associated with the (excess) probability p is the level z_p such that $\mathbb{P}(X > z_p) = p$.

- More formally (needed because such z_p need not exist nor be unique): Define $F(x) = \mathbb{P}(X \leq x)$ (c.d.f., cumulative distribution function)

$$z_p = F^{\leftarrow}(1 - p).$$

Definition 1

The generalized inverse of F, defined for $y \in [0, 1]$ is

$$F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$$

Return period τ_x for an event $\{X > x\}$ I

- Definition from the French Hydrological Society:
  - *General case*: Average duration separating two occurrences of the considered event
  - *Rare event* : (when no or very few such events have been observed before: inverse of the probability of occurrence of the consisdered even over one year.

- τ_x is the *return period* associated with the *return level* x.

Why is natural to define $\tau_x = 1/P(X > x)$?

Return period τ_x for an event $\{X > x\}$ II

- Probabilistic viewpoint: given a sampling rate (here, annual)

Definition 2

τ_x is defined as the expectancy of the *waiting time* T_x between two events:

$$T_x = \inf\{n \geq 0 : \bigvee_{t=1}^{n} X_t \geq x\} \quad \text{(a random variable)}$$

and

$$\tau_x = \mathbb{E}(T_x).$$

Exercise : Show that $\mathbb{E}(T_x) = 1/P(X > x)$.

Questions outside our scope

- Temporally dependent data (time series)
- Non-stationary data (climate change)
- In such cases the *i.i.d.* theory does not apply as it is, but refinements exist under additional weak assumption (long range independence, dynamic model)

Overview of EVT: Three complementary approaches to understand extremes

1. Block maxima

2. Excesses above a high threshold

3. Point process above a high threshold

The three approaches are equivalent in theory

Idea behind EVA

Theory: Under minimal assumptions, distributions of maxima/excesses converge to a certain class.

Modelling: Use those limits to model maxima/excesses above large thresholds.

$\mathbf{X}$: random object (variable / vector/ process) $\quad \mathbf{X}_i \overset{i.i.d.}{\sim} \mathbf{X}$.

$$\bigvee_{i=1}^{n} \mathbf{X}_i \quad \overset{d}{\approx} \quad \text{Max-stable} \qquad (n \text{ large})$$

$$\left[\mathbf{X} \;\middle|\; \|\mathbf{X}\| \geq r \right] \quad \overset{d}{\approx} \quad \text{Generalized Pareto} \qquad (r \text{ large})$$

$$\sum_{i=1}^{n} \delta_{(\frac{i}{n}, X_i)} \quad \overset{d}{\approx} \quad \text{Poisson point process}$$

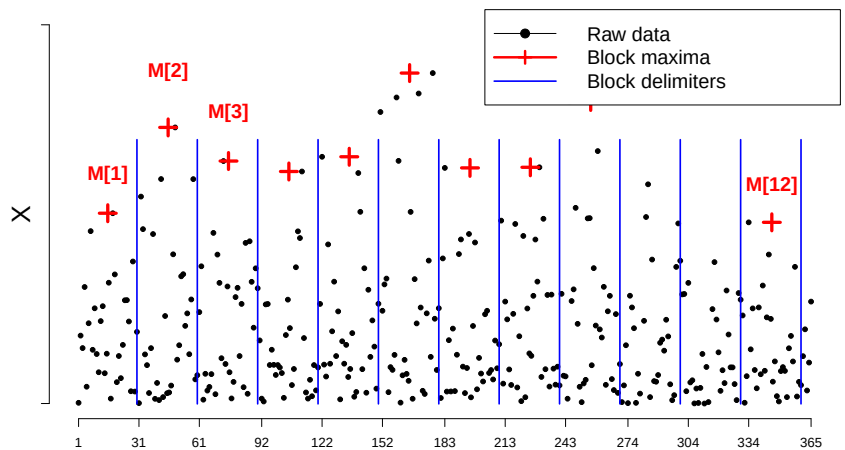
Block Maxima

- Maximum of a "block" of size n:

$$M_n = \max_{t=1,\dots,n} X_t \quad \overset{\text{notation}}{=} \quad \bigvee_{1}^{n} X_t \,.$$

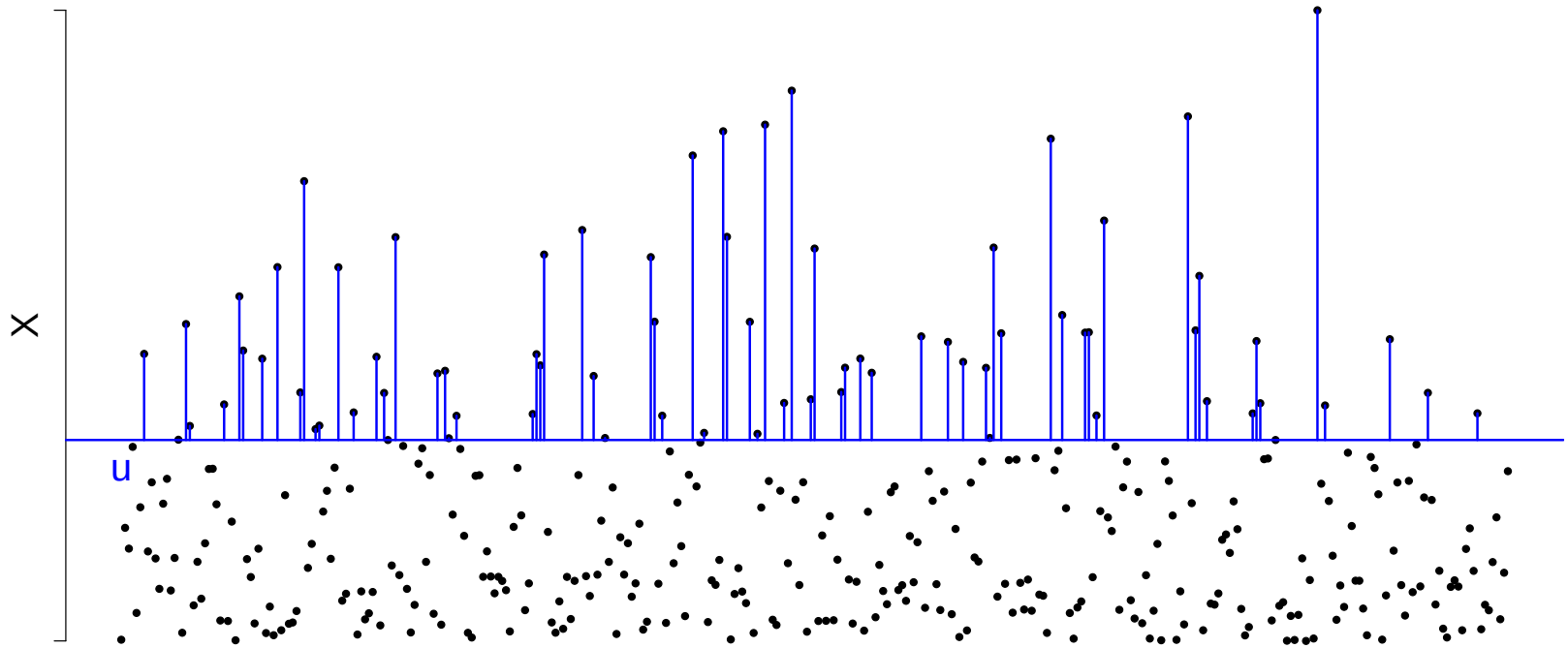
  e.g. : monthly maximum of concentration for an air pollutant.

- Dividing the dataset into m blocks $\hookrightarrow$ m maxima $(M_n[1], \dots, M_n[m])$; $M_n[i] = \bigvee_{t \in \text{ bloc } i} X_t$

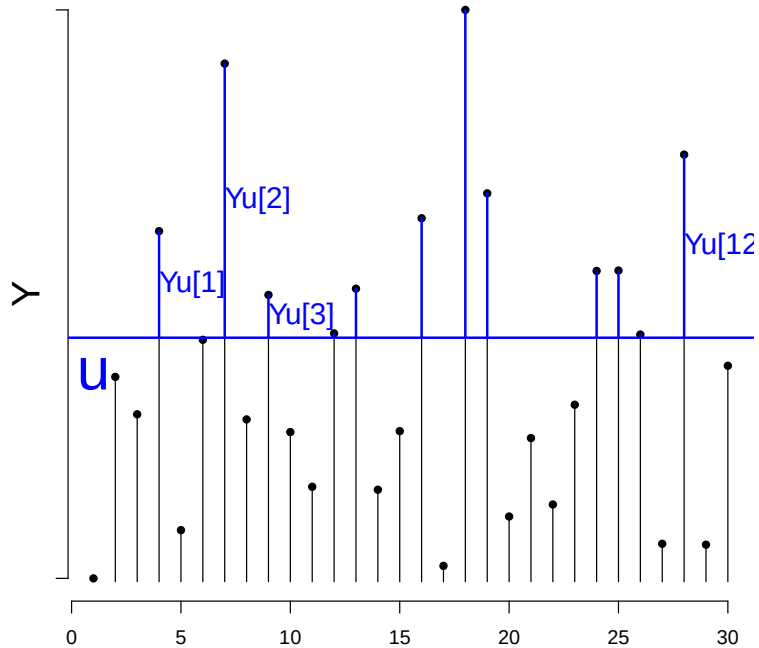


- $n * m$ data points (m blocks of size n) $\hookrightarrow$ only m maxima !

Peaks-Over-Threshold



Peaks-Over-Threshold



- *Excess* : $Y = X - u$, for $X > u$.
- *Conditional survival function*

$$\bar{F}_u(y) = P(X - u > y | X > u) = \frac{\bar{F}(u + y)}{\bar{F}(u)}$$

Outline

Limit laws and rescaling

Obvious issue:
Si $F(x) < 1$, alors $P(M_n \leq x) = F^n(x) \underset{n\to\infty}{\longrightarrow} 0$ ...

- *Maxima* : 'rescaling':

$$\tilde{M}_n = \frac{M_n - b_n}{a_n}$$

- *Excesses* : conditioning $\rightarrow$ *Conditional survival function*:

$$\bar{F}_u(y) = P(X - u > y | X > u) = \frac{\bar{F}(u+y)}{\bar{F}(u)}$$

The most famous weak convergence theorem

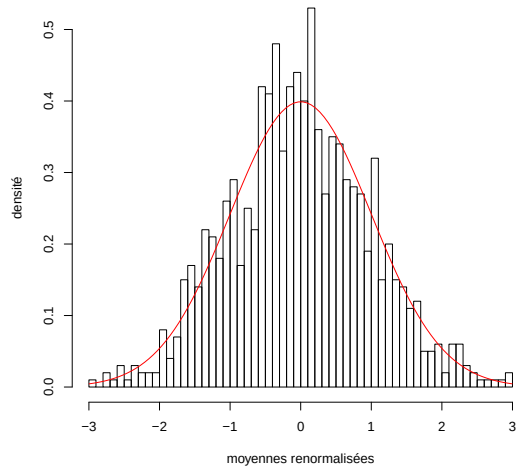
- If X_t are i.i.d. real r.v. with variance σ^2 and expectancy μ, then

$$\frac{\sum_{t=1}^{n} X_t - n\,\mu}{\sigma\sqrt{n}} \xrightarrow{w} \mathcal{N}(0,1)$$

- statement of the form $(\mathcal{O}(X_1,\ldots,X_n) - b_n)/a_n \xrightarrow{w} Y$,
  with $\mathcal{O}$ = 'sum', $b_n = n\,\mu$, $a_n = \sigma\sqrt{n}$

Is there an anologous result for $\mathcal{O} = \max$?

The most famous weak convergence theorem



- The limit law has a specific structure: a normal distribution. The sequences (a_n, b_n) can be chosen such that the limit is centered, with variance 1.

Is there an anologous result for $\mathcal{O} = \max$?

Extreme Value theorem

Theorem 3 (Fisher et Tipett, 1928 ; Gnedenko 1943)

$(X_t)_{t\geq 0}$ i.i.d random variables, $M_n = \max_{t\leq n} X_t$. If there exists sequences $(a_n)_n > 0$, $(b_n)_n \in \mathbb{R}$, and a non-degenerate r.v. Y, s.t.

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} Y,$$

then, Y follows a "Generalized Extreme Value Distribution" (GEV), i.e.

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(Y \leq x) := G_{\mu,\sigma,\xi}(x) = e^{-[(1+\xi \frac{x-\mu}{\sigma})_+]^{-1/\xi}}$$

with $\xi \in \mathbb{R}$, $y_+ = \max(0, y)$, and $G_{\mu,\sigma,0}(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}$.

GEV density

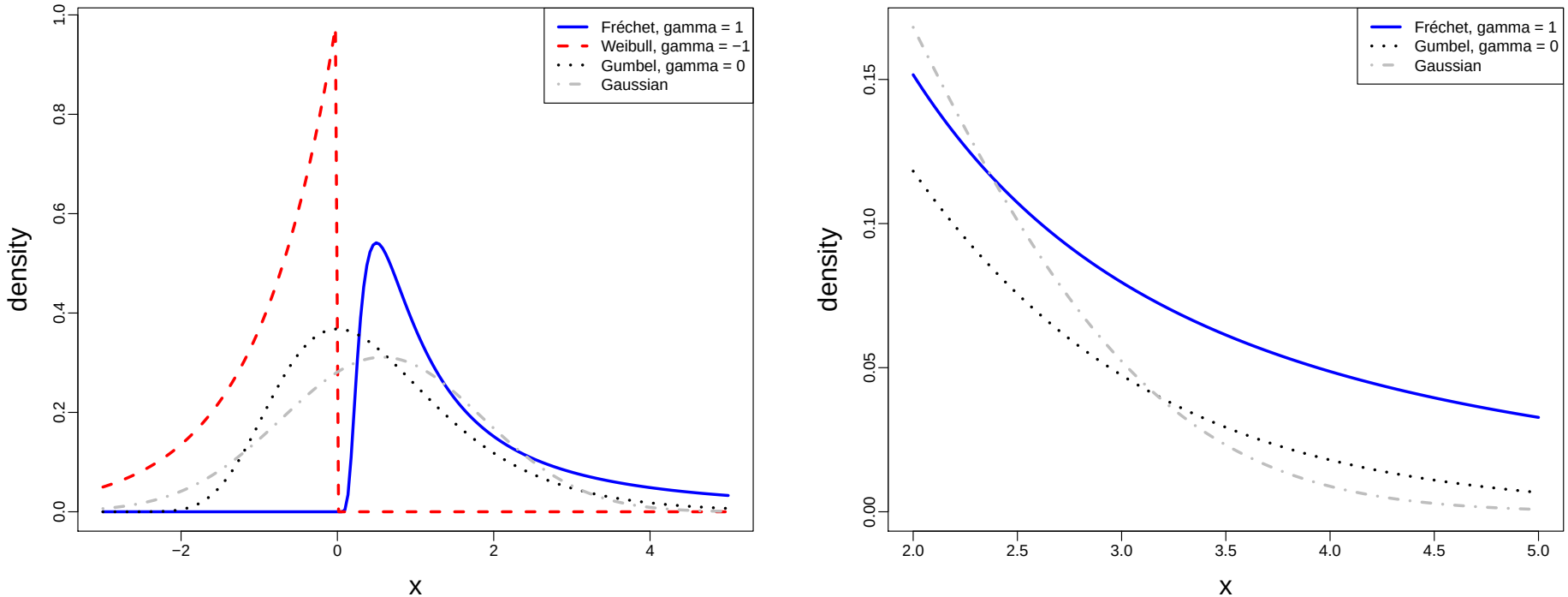


Figure: Density plot for the three extremal types, respectively $(\gamma = 1, \mu = 1, \sigma = 1)$, $(\gamma = -1, \mu = -1, \sigma = 1)$, $(\gamma = 0, \mu = 0, \sigma = 1)$; compared with the Gaussian density with same mean and variance as the Gumbel one. The right panel is a zoom on the tail.

Simulated *i.i.d.* GEV data

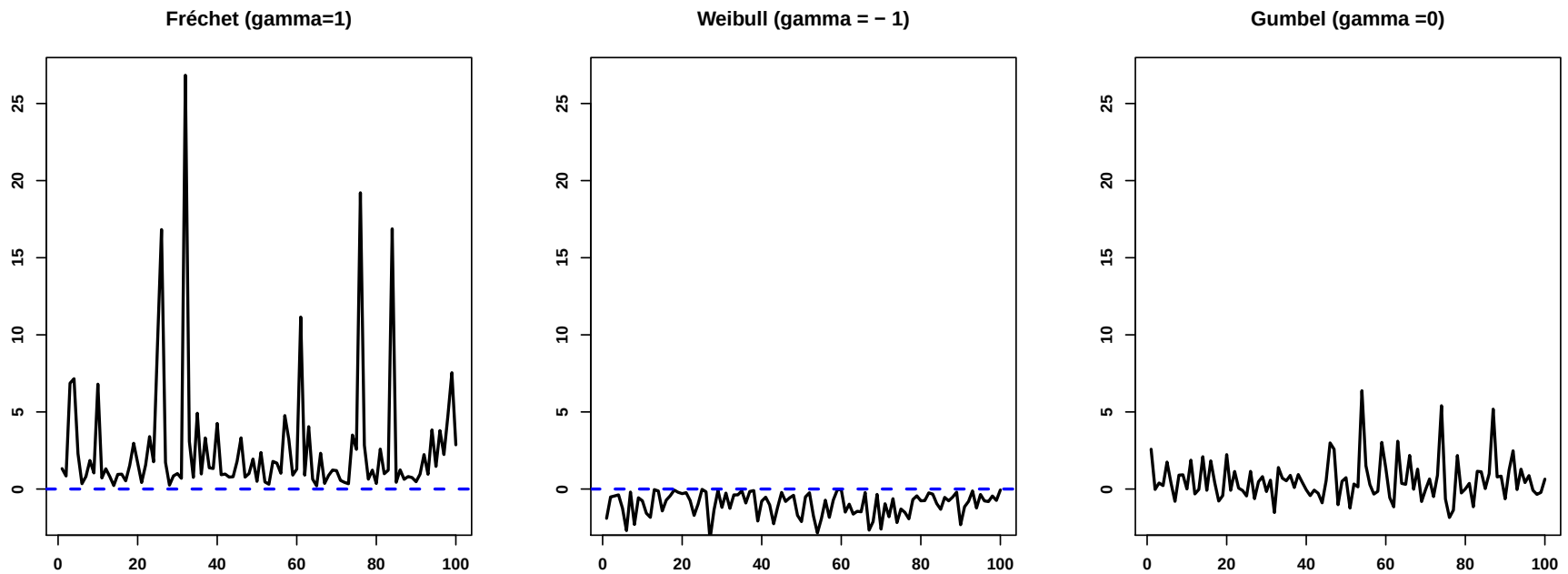


Figure: Series of *i.i.d.* random variables of the three extremal types, respectively $(\gamma = 1, \mu = 1, \sigma = 1)$, $(\gamma = -1, \mu = -1, \sigma = 1)$, $(\gamma = 0, \mu = 0, \sigma = 1)$

GEV tails

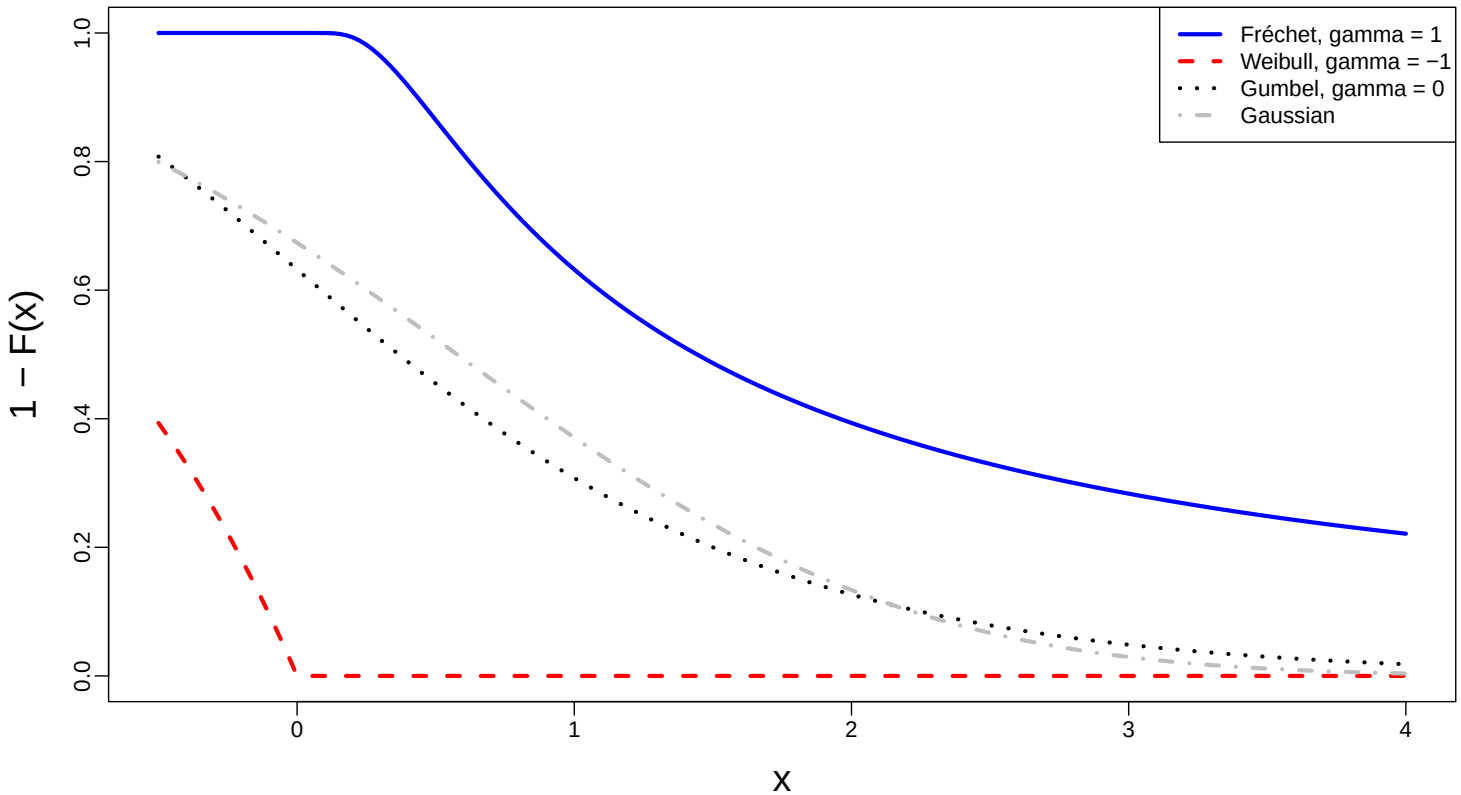
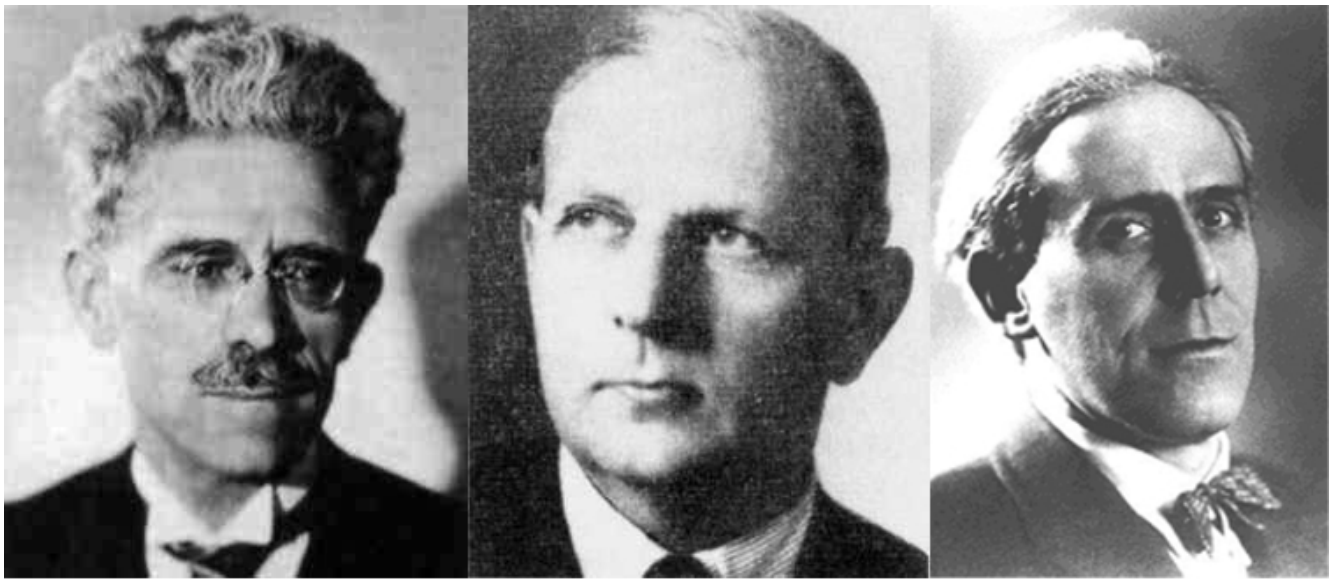


Figure: Survival function $1 - F(x)$ for the three extremal types, respectively $(\gamma = 1, \mu = 1, \sigma = 1)$, $(\gamma = -1, \mu = -1, \sigma = 1)$, $(\gamma = 0, \mu = 0, \sigma = 1)$; compared with the Gaussian survival function with same mean and variance as the Gumbel one.

EVD shapes and *sign* of ξ: Fréchet, Weibull, Gumbel.



- Maurice Fréchet (1878 - 1973): French mathematician (topology, functional analys, probability, statistics). Identifies the limit law of heavy-tailed distributions.
- Walloddi Weibull (1887- 1979, Annecy). Swedish Engineer and mathematician. Material fatigue.
- Emil Julius Gumbel (1891, Munich - 1966, New-York): German Mathematican and political essayist (pacifist). Leaves germany after his expelling from Heidelberg (32) and rejoins France then the U.S. (40). Teaches in Paris, Lyon, Columbia. Considered as the 'father' of Extreme Value Theory.

Domains of attraction: Fréchet, Weibull, Gumbel

Under the conditions of the Fisher-Tipett-Gnedenko theorem, (*i.e.*, $\exists (a_n) > 0, (b_n)$, s.t. $\forall x \in \mathbb{R}, F^n(a_n x + b_n) \underset{n\to\infty}{\longrightarrow} G_{\xi,\sigma,\mu}(x)$), the distributions *F belongs to the domain of attraction of*

- Fréchet if $\xi > 0$.
- Weibull if $\xi < 0$
- Gumbel if $\xi = 0$

Typical representants of each class:

- Fréchet distribution: $\Phi_\alpha(x) = e^{-x^{-\alpha}}$ $(x > 0)$, avec $\alpha > 0$
- Weibull law: $\Psi_\alpha(x) = e^{-(-x)^{\alpha}}$ $(x < 0)$, avec $\alpha > 0$
- Gumbel law: $\Lambda(x) = e^{-e^{-x}}$

Examples of laws in a domain of attraction, applications

- Gumbel's domain:

  ex : Gumbel law (!), Exponential laws $F(x) = 1 - e^{-\lambda x}, (\lambda > 0)$; Normal distributions, log-normal distributions.

  hydrology (River streamflow, precipitation, annual maximum of water level (dikes)

- Fréchet's domain:

  ex : Fréchet law (!), Pareto distribution $F(x) = 1 - Kx^{-\alpha}$; Cauchy $F(x) = \frac{1}{2}\frac{1}{\pi}\tan^{-1}(x)$; Student distribution $f(x) = C(1 + x^2/k)^{-\frac{k+1}{2}}$.

  precipitations, river floods in mediterranean regime, pollution peaks, financial log-returns (Dow Jones...)

- Weibull's domain:

  ex: Weibull law (!) ; Uniform distribution on a line segment ; truncated exponential.

  Material fatigue, temperatures, lifetime.

How to determine the domain of attraction of a given law?

Exercises: find the domain of attraction and suitable sequences (a_n), (b_n) for

1. Exponential law : $\forall x \geq 0, F(x) = 1 - e^{-\lambda x} \quad (\lambda > 0)$.
   Hint: use that $(1 - y/n)^n \to e^{-y}$

2. Uniform law: $F(x) = x (0 < x < 1)$

3. Pareto law $F(x) = 1 - \left(\frac{u}{x}\right)^{\alpha} \quad (x > u, \alpha > 0)$.

General case : one can check some conditions (*von Mises conditions*) on the limit behavior of the ratio $F'/(1 - F)$ (out of our scope)

Outline

Inference methods, existing R packages

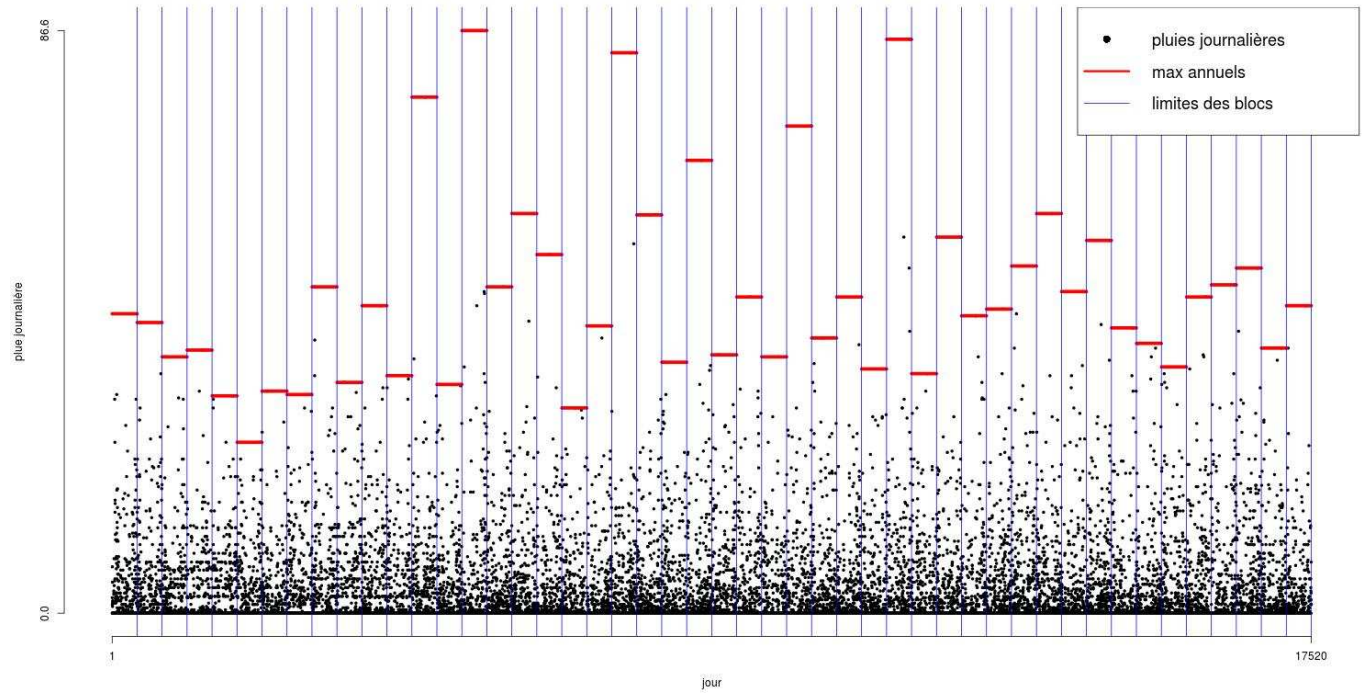
- Most popular methods: Maximum likelihood, probability weighted moments
- R packages: `ismev`, `extRemes`, `evd`, `fExtremes`, `EVIM`, `Xtremes`, `HYFRAN`, `EXTREMES` , ...

  `http://cran.r-project.org/`
- Gilleland, Ribatet, Stephenson, 2013: *A software review for extreme value analysis*
- Introductory book: Coles, 2001, An Introduction to Statistical Modeling of Extreme Values.
- Here: maximum likelihood, package evd.

Assumption behind extreme value (block maxima) models

- For n large enough, $M_n \sim G_{\mu,\sigma,\xi}$.
- goal: estimate μ, σ, ξ
- Use block maxima to learn estimates $\hat{\mu}, \hat{\sigma}, \hat{\xi}$.

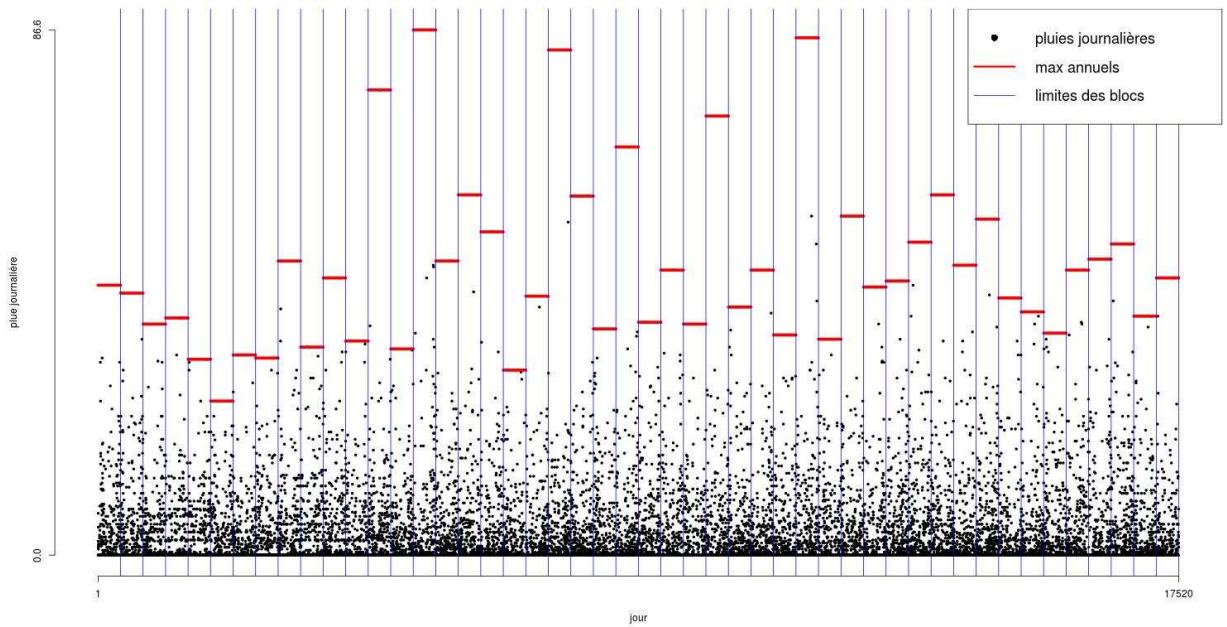
Block maxima



Modeling the max with a GEV

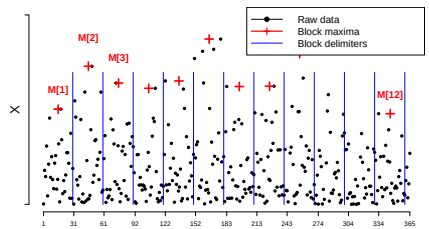
Data : block maxima, block size n,

- $(M_n[1], \ldots, M_n[m])$; $\quad M_n[i] = \bigvee_{t=(i-1)n+1}^{in} \{X_t\}$
  (*e.g.* : annual maxima of a river stream flow over 50 years, annual maxima of claims to an insurance, ...)
- For large block sizes n, in view of Fisher, Tipett & Gnedenko Theorem, the GEV family is a reasonable model for M_n



Fitting a GEV model on block maxima

- *m processed data* : $(M_n[1], \ldots, M_n[m])$, $\quad M_n[i] = \bigvee_{t=(i-1)n+1}^{in} X_t$.
  (ex: annual max of a river flow over fifty year)



- *modeling assumption*: $\frac{M_n[i]-b_n}{a_n} \sim G_{\xi,0,1}$ (n fixed, but large.) for some a_n, b_n *i.e.*

$$M_n[i] \sim G_{\xi,\mu,\sigma} \quad \text{with } \mu = b_n,\ \sigma = a_n\ .$$

- *Parametric model*:

$$\{G_{\xi,\mu,\sigma} : \xi \in \mathbb{R}, \sigma > 0, \mu \in \mathbb{R}\}$$

- Estimation problem: $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$?

Maximum likelihood method: a reminder

- Data $Y_i \overset{i.i.d.}{\sim} F_\theta(.)$
- θ : unknown model parameter. (model $= \{F_\theta, \theta \in \Theta\}$)
- Dominated model: F_θ has sdensity $f_\theta(.) = \frac{d}{dy} F_\theta(.)$.
- *Likelihood* of θ, given m observations $\mathbf{y} = y_1, \ldots, y_m$: the (product) denisty f_θ, at point $\mathbf{y}$.

$$\mathcal{L}(\theta | y_1, \ldots, y_m) = f_\theta(\mathbf{y}) \overset{\text{independence}}{=} \prod_{i=1}^{m} f_\theta(y_i) \ .$$

The maximum likelihood estimator (MLE) is the maximizer (w.r.t. θ) of the likelihood function.

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta | y_1, \ldots, y_m)$$

- Under generic assumption (model regularity): $\hat{\theta}$ is asymptotically normal, with mean θ (true) and variance $O(1/n)$
- **In the GEV model:** $\theta = (\xi, \mu, \sigma)$. Method only valid for $\xi > -0.5$ (the support depends on the parameter: the model is not regular)

Example 1: block maxima inference with R package evd

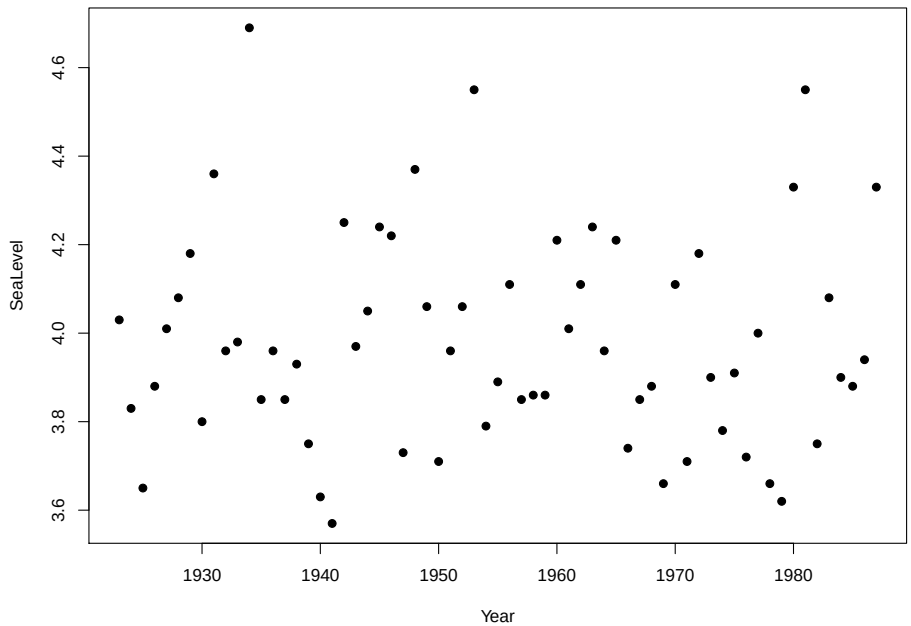


Figure: portpirie data in package evd: Annual maxima of the sea level at Port Pirie, 1923-1987

(the data are already pre-processed, only the annual maxima are available)

Example 1 Cont'd: MLE in the GEV model

```
> library(evd)
> fitgevpirie <- fgev(portpirie)
> fitgevpirie

Call: fgev(x = portpirie)
Deviance: -8.678117

Estimates
     loc    scale    shape
 3.87475  0.19805 -0.05012

Standard Errors
    loc    scale    shape
0.02793 0.02025 0.09826

Optimization Information
  Convergence: successful
  Function Evaluations: 30
  Gradient Evaluations: 8
```

Example 1 Cont'd: graphical diagnostics

```
> plot(fitgevpirie)
```

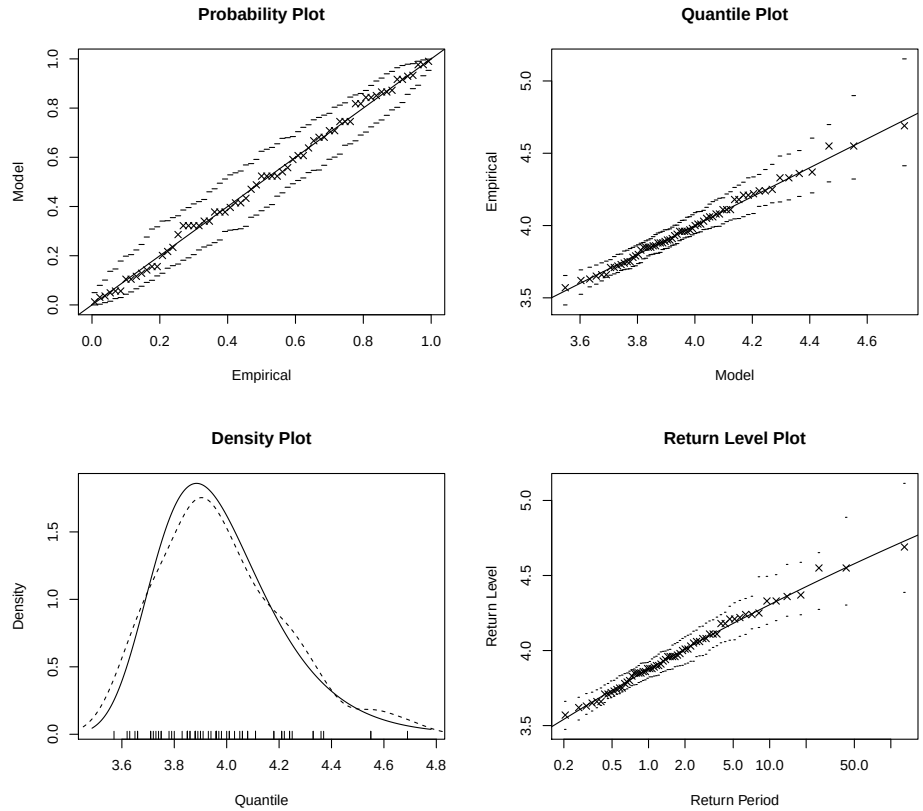


Figure: Graphical diagnostic plot for the GEV model fit on the Port Pirie dataset, as provided by R package evd.

Extreme quantile: plug-in estimation after MLE

- Return level (quantile) z_p (corresponding to a return period $1/p$):

$$z_p = G^{-1}_{\xi,\mu,\sigma}(1-p) = \begin{cases} \mu - \left[1 - \{-\log(1-p)\}^{-\xi}\right] \frac{\sigma}{\xi} & \text{if } \xi \neq 0 \\ \mu - \sigma \log(-\log(1-p)) & \text{if } \xi = 0 \end{cases}$$

- Gaussian confidence intervals on the parameters + Delta method $\Rightarrow$ Gaussian CI on r z_p

```
> fgev(portpirie, prob=0.001)
Estimates
quantile     scale     shape
 5.03508   0.19818  -0.04926

Standard Errors
quantile     scale     shape
 0.34024   0.02010   0.09904
```

- $\xi = 0$ belongs to the CI! Should one use the Gumbel model?

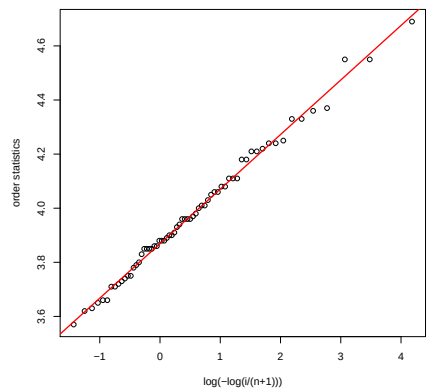
Graphical diagnostic for Gumbel domain

- Idea: If $F_n(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}$, then

$$F_n^{\leftarrow}(u) = \sigma\left[-\log(-\log(u))\right] + \mu.$$

- $F_n^{\leftarrow}(\frac{i}{n+1}) \simeq x_{(i)}$ (the i^{th} smallest observation)
- $\Rightarrow$ the graph of points $(-\log(-\log(\frac{i}{n+1})); x_{(i)i})$ is close to the diagonal.

```
xord<-sort(portpirie,decreasing=F); n <- length(portpirie)
inds=1:n/(n+1)
gbquant<--log(-log(inds))
plot(gbquant,xord)
reg.lin<-lm(xord~gbquant) ; coeff<- reg.lin$coefficients
abline(coeff[1],coeff[2], col="red",lwd=2)
```

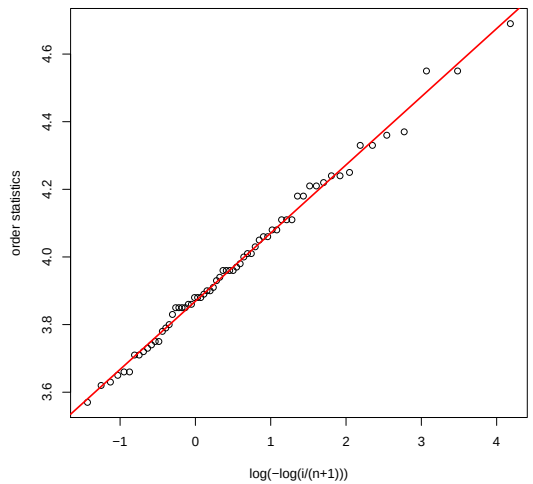


Graphical diagnostic for Gumbel domain

- Idea: If $F_n(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}$, then

$$F_n^{\leftarrow}(u) = \sigma\left[-\log(-\log(u))\right] + \mu.$$

- $F_n^{\leftarrow}(\frac{i}{n+1}) \simeq x_{(i)}$ (the i^{th} smallest observation)
- $\Rightarrow$ the graph of points $(-\log(-\log(\frac{i}{n+1})); x_{(i)i})$ is close to the diagonal.



The Gumbel model seems suitable here. Return levels are then much higher than in the Weibull model!

Discussion: block-maxima approach

- Only the block maxima are used $\rightarrow$ information loss.
- Choice of the block size: bias variance compromise, with significant impact.

- How to use the data in a different way?

$$\text{Maxima} \Leftrightarrow \text{Peaks-Over-Threshold.}$$

$$\mathbb{P}(X > u + \sigma(u)y \mid X > u) = \bar{F}_u(y) \underset{u \to \infty}{\longrightarrow} \quad ?$$

References

- Coles, Stuart, et al. An introduction to statistical modeling of extreme values. Vol. 208. London: Springer, 2001.
- Beirlant, Jan, et al. Statistics of extremes: theory and applications. John Wiley & Sons, 2006.
- Resnick, Sidney I. Heavy-tail phenomena: probabilistic and statistical modeling. Springer Science & Business Media, 2007.
- Resnick, Sidney I. Extreme values, regular variation and point processes. Springer, 2013.
- De Haan, L., & Ferreira, A. (2007). Extreme value theory: an introduction. Springer Science & Business Media.
- Goix, N., Sabourin, A., & Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. Journal of Multivariate Analysis, 161, 12-31.
- Goix, N., Sabourin, A., & Clémençon, S. (2016, March). Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. In AISTATS (pp. 75-83).