A new extreme quantile estimator based on the log-generalized Weibull-tail model

Clément ALBERT

Anne Dutfoy (EDF), Laurent Gardes (Université de Strasbourg), Stéphane Girard (INRIA)

3rd year-PhD
May 2018

Outline

1. Extreme quantile estimation
2. The framework
3. A new extreme quantile estimator
4. Application to environnemental data

Outline

1. Extreme quantile estimation
2. The framework
3. A new extreme quantile estimator
4. Application to environnemental data

Extreme quantile estimation : Principle

Let X be a random variable with distribution function

$$F(\cdot) = \mathbb{P}(X \leq \cdot)$$

and survival function

$$\overline{F} := 1 - F.$$

Starting from a $n-$sample from X, our goal is to estimate extreme quantiles $Q(\beta_n)$ of level $1 - \beta_n$ with $n\beta_n \to 0$ as $n \to \infty$, where

$$Q(\beta) := \inf\{x;\ \overline{F}(x) \leq \beta\}.$$

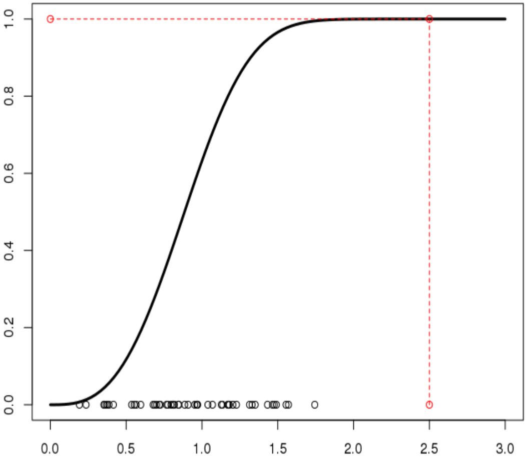


Figure: Extreme quantile estimation

The excesses above u_n are defined as $Y_i = X_i - u_n$ for all $X_i > u_n$.
Peaks Over Threshold method (POT) [Smith, 1987] relies on an approximation [Pickands, 1975] of the distribution of excesses $\overline{F}_{u_n}$ by a Generalized Pareto Distribution (GPD) :

$$\overline{F}_{u_n}(x) \approx \left| \begin{array}{ll} \left(1 + \frac{\gamma_n x}{\sigma_n}\right)^{-1/\gamma_n} & , \gamma_n \neq 0 \\ \exp\left(-\frac{x}{\sigma_n}\right) & , \gamma_n = 0 \end{array} \right.$$

where σ_n and γ_n are the scale and shape parameters of the GPD distribution.

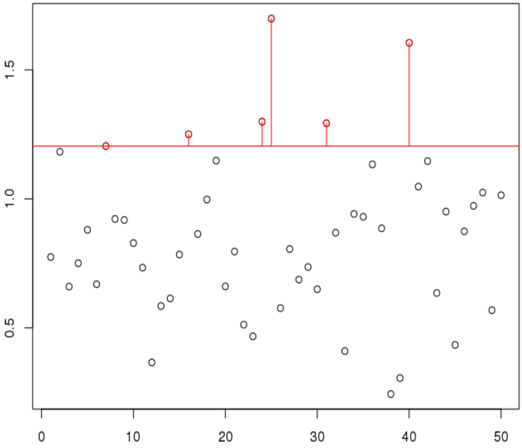


Figure: Definition of excesses

Remark

$$\overline{F}_{u_n}(x) = \mathbb{P}(Y \geq x | X \geq u_n), = \frac{\overline{F}(x+u_n)}{\overline{F}(u_n)}.$$

so that

$$\overline{F}(x+u_n) = \overline{F}(u_n)\overline{F}_{u_n}(x)$$

Let $v_n = x + u_n$, with u_n a threshold such that $u_n = Q(\alpha_n)$:

$$\overline{F}(v_n) \approx \left| \begin{array}{l} \alpha_n \left(1 + \gamma_n \frac{v_n - u_n}{\sigma_n}\right)^{-1/\gamma_n} \\ \alpha_n \exp\left(-\frac{v_n - u_n}{\sigma_n}\right) \end{array} \right.$$

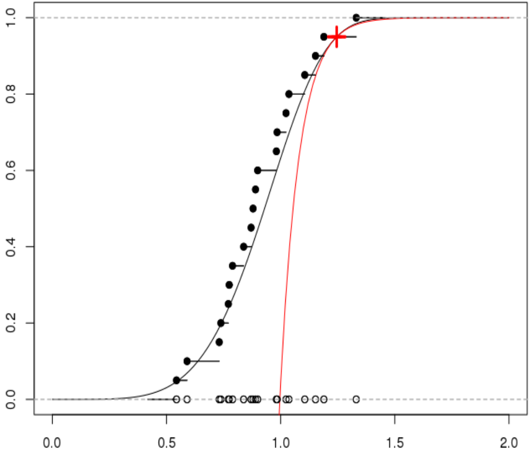


Figure: Tail approximation

As a consequence, $Q(\beta_n)$ can be in turn approximated by the deterministic term :

$$Q(\beta_n) \approx \left| \begin{array}{c} Q(\alpha_n) + \frac{\sigma_n}{\gamma_n}\left[\left(\frac{\alpha_n}{\beta_n}\right)^{\gamma_n} - 1\right] \\ Q(\alpha_n) + \sigma_n \ln\left(\frac{\alpha_n}{\beta_n}\right) \end{array} \right.$$

Extrapolation is performed in the distribution tail from $Q(\alpha_n)$ to $Q(\beta_n)$ thanks to an additive correction depending on α_n/β_n.

Then, the POT method consists in estimating the two unknown parameters σ_n and γ_n.

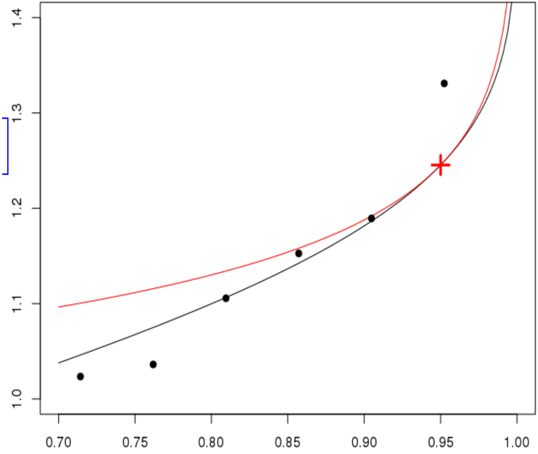


Figure: Quantile approximation

For example, if $F \in MDA(Gumbel)$ and so $\gamma_n = 0$, one can choose $\hat{Q}(\alpha_n) = X_{n-k_n+1,n}$ with $k_n = \lfloor n\alpha_n \rfloor$ and

$$\hat{\sigma}_n = \frac{1}{k_n}\sum_{i=1}^{k_n}(X_{n-i+1,n} - X_{n-k_n+1,n})$$

to obtained the so-called Exponential Tail (ET) estimator [Breiman et al, 1990] :

$$\hat{Q}(\beta_n) = \hat{Q}(\alpha_n) + \hat{\sigma}_n \ln(\alpha_n/\beta_n),$$

where $X_{1,n} \leq \cdots \leq X_{n,n}$ are the order statistics associated with $X_1, \ldots, X_n$.

Outline

In the following, The function $V(\cdot) := \ln Q(1/\exp\cdot)$ is supposed to be of extended regular variation with index $\theta \in \mathbb{R}$ ($ERV(\theta)$). More specifically, there exists a positive function a (called the auxiliary function) such that, for all $t > 0$

$$\lim_{x\to\infty} \frac{V(tx) - V(x)}{a(x)} = \int_1^t u^{\theta-1} du =: L_\theta(t). \tag{1}$$

This model is referred to as the "log-generalized Weibull-tail model" [de Valk, 2016]. A sufficient condition for (1) is

(A1) V is differentiable with derivative V' satisfying

$$\lim_{x\to\infty} \frac{V'(tx)}{V'(x)} = t^{\theta-1}.$$

Such a function V' is said to be regularly varying with index $\theta - 1$ and this property is denoted by $V' \in RV(\theta - 1)$, see [Bingham, 1987]. Moreover, under **(A1)**, a possible choice in (1) is $a(x) = xV'(x)$.

The framework

The next result provides a characterization of the tail behavior of F according to the sign of θ.

Proposition (Characterizations)

Let $x^* := \sup\{x \geq 1,\ F(x) < 1\}$ *be the endpoint of* F. *Then, under some monotonicity assumptions :*

(i) *If* $V^{\leftarrow}(\ln \cdot) \in RV(1/\beta)$, $\beta > 0$, *then* **(A1)** *holds with* $\theta = 0$.

(ii) $V^{\leftarrow} \in RV(1/\beta)$, $0 < \beta < 1$ *if and only if* **(A1)** *holds with* $\theta = \beta > 0$.

(iii) $1 \leq x^* < \infty$ *and* $V^{\leftarrow}(\ln x^* + \ln(1 - 1/\cdot)) \in RV_{-1/\beta}$, $\beta < 0$ *if and only if* **(A1)** *holds with* $\theta = \beta < 0$.

- In the case (i), F is referred to as a Weibull tail-distribution. Such distributions encompass Gaussian, Gamma, Exponential and strict Weibull distributions.
- In the case (ii) F is called a log-Weibull tail-distribution, the most popular example being the lognormal distribution.
- The case (iii) corresponds to distributions with a Weibull tail behavior in the neighborhood of a finite endpoint.

Besides, let us highlight that the domain of attraction associated with F depends on the position of θ with respect to 1:

Proposition (Domains of attraction)

Assume F is differentiable.

(i) *If* **(A1)** *holds with $\theta < 1$ then $F \in$ MDA(Gumbel).*

(ii) *If $F \in$ MDA(Fréchet) then* **(A1)** *holds with $\theta = 1$.*

(iii) *If* **(A1)** *holds with $\theta > 1$ then F does not belong to any MDA.*

It thus appears that model **(A1)** with $\theta \leq 1$ is of particular interest since it is associated with most distributions in MDA(Gumbel) $\cup$ MDA(Fréchet).

The situation $\theta > 1$ which does not correspond to any domain of attraction is sometimes referred to as super-heavy tails, see for instance [Alves, 2009].

Outline

Model inference

Let $X_1, \ldots, X_n$ be n independent copies of a random variable X distributed following the model previously introduced. The associated ordered statistics are denoted by $X_{1,n} \leq \ldots \leq X_{n,n}$. Starting from this random sample, we focus on the estimation of extreme quantiles *i.e.* $Q(u) := \overline{F}^{\leftarrow}(u) = \exp[V(\ln(1/u))]$ when $u \to 0$. Two situations for the level u are considered.

1. **Intermediate case.** If $u = \alpha_n$ where α_n is an intermediate level satisfying $\alpha_n \to 0$ and $n\alpha_n \to \infty$ as $n \to \infty$, a natural estimator is obtained by replacing Q by its empirical counterpart $\hat{Q}_n$. More precisely, $Q(\alpha_n)$ is estimated by

$$\hat{Q}_n(\alpha_n) = X_{n-\lfloor n\alpha_n \rfloor, n}.$$

2. **Extreme case.** If $u = \beta_n$ where β_n is an extreme level such that $n\beta_n \to c \geq 0$ as $n \to \infty$, a simple order statistics cannot be used. Extrapolation beyond the sample should be performed. Starting from an intermediate level $\alpha_n := k_n/n$ where $k_n \to \infty$ and $k_n/n \to 0$, we propose to estimate $Q(\beta_n)$ by

$$\hat{Q}_n(\beta_n) := \hat{Q}_n(\alpha_n) \exp\left[\hat{a}_n[\ln(n/k_n)] L_{\hat{\theta}_n}\left(\frac{\ln \beta_n}{\ln(k_n/n)}\right)\right],$$

where $\hat{\theta}_n$ and $\hat{a}_n[\ln(n/k_n)]$ are suitable estimators of θ and $a[\ln(n/k_n)]$.

Model inference

The rationale behind

$$\hat{Q}_n(\beta_n) := \hat{Q}_n\left(\alpha_n\right)\exp\left[\hat{a}_n[\ln(n/k_n)]L_{\hat{\theta}_n}\left(\frac{\ln\beta_n}{\ln(k_n/n)}\right)\right], \qquad (2)$$

is based on

$$\lim_{y\to\infty}\frac{V(ty)-V(y)}{a(y)} = \int_1^t u^{\theta-1}du =: L_\theta(t).$$

which basically means that for α close to 0 and for all $t > 0$,

$$\ln Q(t\alpha) \approx \ln Q(\alpha) + a[\ln(1/\alpha)]L_\theta\left(1+\frac{\ln(t)}{\ln(\alpha)}\right).$$

Estimator (2) is then obtained by taking $\alpha = k_n/n$ and $t = n\beta_n/k_n$ and by replacing the unknown quantities $Q(k_n/n)$, $a[\ln(n/k_n)]$ and θ by their corresponding estimators. Since k_n/n is an intermediate level, $Q(k_n/n)$ is estimated by $\hat{Q}_n(k_n/n) = X_{n-k_n,n}$.

Inference

The estimator of θ we propose is similar in spirit to the moment estimator introduced in [Dekkers et al, 1989]. Its construction is based on the following two results. Letting $\theta_+ := \theta \vee 0$ and $\theta_- := \theta \wedge 0$, for any increasing function $V \in ERV_\theta$,

$$\lim_{x\to\infty} \frac{V(x)}{a(x)} \ln \frac{V(tx)}{V(x)} = L_{\theta_-}(t),$$

locally uniformly in $(0,\infty)$, see [de Haan & Ferreira, Lemma 3.5.1]. Moreover, one has,

$$\lim_{x\to\infty} \frac{a(x)}{V(x)} = \theta_+.$$

Plugging $x := \ln(1/\alpha)$ and $t := 1 + \ln(s)/\ln(\alpha)$ yields the approximation

$$\ln_2 Q(s\alpha) - \ln_2 Q(\alpha) \approx \theta_+ L_0\left(1 + \frac{\ln s}{\ln \alpha}\right),$$

as $\alpha \to 0$ and for all $s \in (0,1)$. Integrating with respect to s on $(0,1)$ leads to

$$\int_0^1 [\ln_2 Q(s\alpha) - \ln_2 Q(\alpha)]\, ds \Big/ \int_0^1 L_0\left(1 + \frac{\ln s}{\ln \alpha}\right) ds \approx \theta_+.$$

Inference

Considering $\alpha = k_n/n$ where k_n is an intermediate sequence such that $k_n \to \infty$ and $k_n/n \to 0$ and replacing Q by its empirical estimator lead to the following estimator of θ_+:

$$\hat{\theta}_{n,+} := \frac{M_n^{(1)}}{\mu_1[\ln(n/k_n), 0]},$$

where, for $t > 0$, $b \in \mathbb{N} \setminus \{0\}$, $\zeta < 1$,

$$\mu_b(t, \zeta) := \int_0^1 \left[L_\zeta \left(1 + \frac{\ln(1/s)}{t} \right) \right]^b ds.$$

Similarly, remark that the previous equation leads to the approximation

$$\left\{ \int_0^1 [\ln_2 Q(s\alpha) - \ln_2 Q(\alpha)]\, ds \right\}^2 \Big/ \int_0^1 [\ln_2 Q(s\alpha) - \ln_2 Q(\alpha)]^2\, ds \approx \Psi_{\ln(1/\alpha)}(\theta_-),$$

as $\alpha \to 0$, where

$$\Psi_t(\zeta) := \frac{\mu_1^2(t, \zeta)}{\mu_2(t, \zeta)}.$$

Replacing again in the previous approximation α by k_n/n and Q by its empirical counterpart suggests to estimate θ_- by :

$$\hat{\theta}_{n,-} := \Psi^{-1}_{\ln(n/k_n)} \left(\frac{[M_n^{(1)}]^2}{M_n^{(2)}} \right).$$

We propose to estimate θ by :

$$\hat{\theta}_n := \hat{\theta}_{n,+} + \hat{\theta}_{n,-}.$$

To obtain an estimator of $a[\ln(n/k_n)]$, one can remark that

$$\frac{\ln Q(\alpha)}{a[\ln(1/\alpha)]} \int_0^1 \ln \frac{\ln Q(s\alpha)}{\ln Q(\alpha)} ds \approx \mu_1[\ln(1/\alpha), \theta_-],$$

for α close to 0. Replacing α by k_n/n, Q by its empirical counterpart and θ_- by $\hat{\theta}_{n,-}$ gives :

$$\hat{a}_n[\ln(n/k_n)] := \frac{\ln X_{n-k_n,n}}{\mu_1[\ln(n/k_n), \hat{\theta}_{n,-}]} M_n^{(1)}.$$

Main results

The two following results respectively provide the asymptotic behavior of the quantile estimator in the intermediate and extreme cases.

Theorem

Under the model previously introduced, assume that **(A1)** *holds. For all intermediate level α_n, one has*

$$\frac{k_n^{1/2}/\ln(n/k_n)}{a[\ln(n/k_n)]}\ln\left(\frac{\hat{Q}_n(\alpha_n)}{Q(\alpha_n)}\right)\xrightarrow{d}\mathcal{N}(0,1).$$

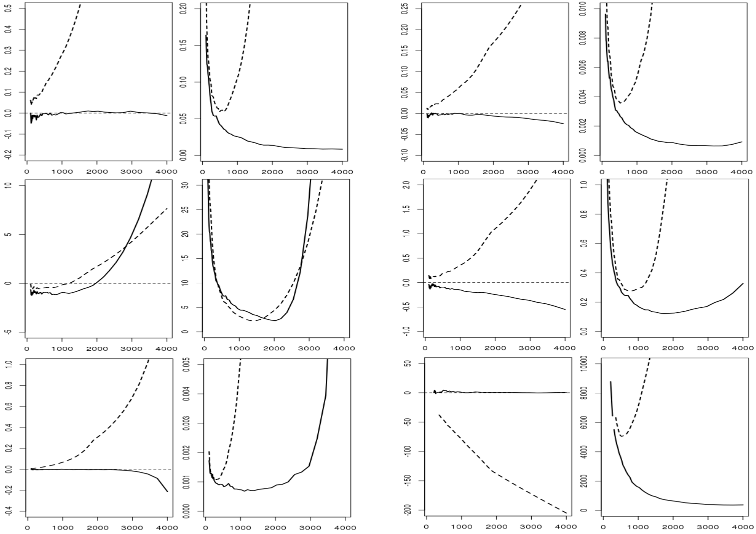
Theorem

For all extreme level β_n, under some additional second order condition on V, one has

$$\frac{k_n^{1/2}/\ln(n/k_n)}{a[\ln(n/k_n)]H_{\theta,0}(d_n)}\ln\left(\frac{\hat{Q}_n(\beta_n)}{Q(\beta_n)}\right)\xrightarrow{d}\mathcal{N}(0,1).$$

Validation on simulations

Figure: Bias (Left) and Mean Square Error (Right) associated with $\hat{Q}_n(\beta_n)$ (solid line) and with the proposal of Cees de Valk and Juan-Juan Cai (dashed line) as a function of k, for $n = 500$ and $N = 500$, N the number of replicates. From top to bottom, left to right : Gamma, Gaussian, Pareto-like, Lognormal, Finite endpoint, Super heavy tail.

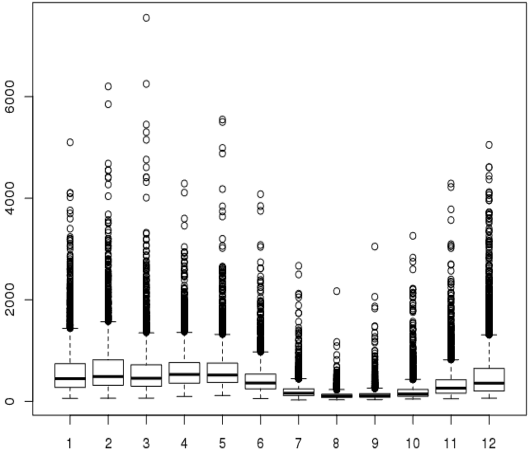


Outline

The Dataset

Figure: Left figure : first rows of the dataset. Right figure : Boxplot representing the twelve months of the year (January to December).

Date	Debit
1915-01-01	540
1915-01-02	865
1915-01-03	1140
1915-01-04	1330
1915-01-05	1750
1915-01-06	2310
1915-01-07	1920
1915-01-08	1470
1915-01-09	1230
1915-01-10	1560
1915-01-11	1830
1915-01-12	2570
1915-01-13	4020
1915-01-14	2700
1915-01-15	2260
1915-01-16	1720



We consider daily river flow measures, in m^3/s of the Rhône from 1915 to 2013. Due to seasonality aspect, only flows from December 1 to May 31 are retained leading to $n = 18043$ measures.

Estimation of the 1000 years return level

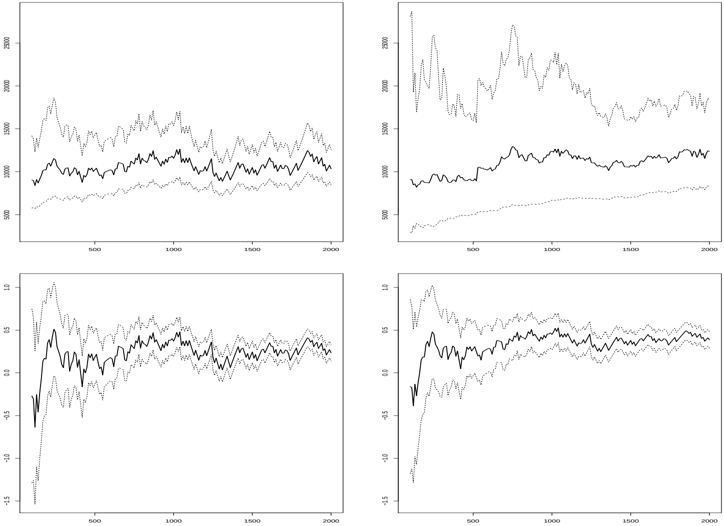


Figure: Estimates $\hat{Q}_n(\beta_n)$ (top left) and its equivalent proposed by de Valk and Cai (top right) of the 10^{-3} per year quantile ($\beta_n = 5.5\ 10^{-6}$) of river flows and their corresponding index estimates (bottom left and right) as functions of $k \in \{100, \ldots, 2000\}$. The 95% asymptotic confidence intervals are depicted by dotted lines.

Main references

- [1] **De Valk, C. (2016)**, Approximation of high quantiles from intermediate quantiles, *Extremes*, 19(4), 661-686.

- [2] **De Valk, C., & Cai, J. J. (2017)**, A high quantile estimator based on the log-generalized Weibull tail limit, *Econometrics and Statistics*, to appear.

- [3] **Albert, C., Dutfoy, A., & Girard, S. (2018)**, *Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles*, submitted, hal-01692544v2.

- [4] **Albert, C., Dutfoy, A., Gardes, L., & Girard, S. (2018)**, *An extreme quantile estimator for the log-generalized Weibull-tail model*, submitted, hal-01783929v2.