Estimation of a new parameter discriminating between Weibull tail-distributions and heavy-tailed distributions

BY

Jonathan EL METHNI

in collaboration with

Laurent GARDES & Stéphane GIRARD & Armelle GUILLOU

Tunis Thursday 26th May 2011.

Motivations

Let $X_1, \dots, X_n$ be a sample of independent and identically distributed random variables driven from X with cumulative distribution function F, and let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics associated to this sample.

- We want to estimate the extreme quantile x_{p_n} of order p_n associated to the random variable $X \in \mathbb{R}$ defined by

$$x_{p_n} = \overline{F}^{\leftarrow}(p_n) = \inf\{x, \bar{F}(x) \leq p_n\},$$

  with $p_n \to 0$ when $n \to \infty$. The function $\overline{F}^{\leftarrow}$ is the generalized inverse of the non-increasing function $\bar{F} = 1 - F$.
- Difficulty : If $np_n \to 0$ then $\mathbb{P}(x_{p_n} > X_{n,n}) \to 1$.

Principals results on extreme value theory

Fisher-Tippett-Gnedenko theorem

Under some conditions of regularity on the cumulative distribution function F, there exists a real parameter γ and two sequences $(a_n)_{n\geq 1} > 0$ and $(b_n)_{n\geq 1} \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

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

with

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right) & \text{if } \gamma \neq 0, \\ \exp\left(-e^{-x}\right) & \text{if } \gamma = 0, \end{cases}$$

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

3 domains of attraction

- $\mathcal{H}_\gamma$ is called the cumulative distribution function of the extreme value distribution.
- If F verifies the Fisher-Tippett-Gnedenko theorem, we say that F belongs to the domain of attraction of $\mathcal{H}_\gamma$.
- γ is called the extreme value index.

Fréchet ($\gamma > 0$)	Gumbel ($\gamma = 0$)	Weibull ($\gamma < 0$)
Pareto	Normal	Uniform
Student	Exponential	Beta
Burr	Log-normal	
Fréchet	Gamma	
	Weibull	

Fréchet maximum domain of attraction : heavy-tailed distributions

All cumulative functions which belong to the Fréchet maximum domain of attraction denoted by $\mathcal{D}(\textit{Fréchet})$ can be rewritten as

$$\bar{F}(x) = x^{-1/\gamma}\ell(x),$$

where $\gamma > 0$ and $\ell(x)$ is a slowly varying function *i.e.* $\ell(\lambda x)/\ell(x) \to 1$ as $x \to \infty$ for all $\lambda \geq 1$. $\bar{F}(x)$ is said to be regularly varying at infinity with index $-1/\gamma$. This property is denoted by $\bar{F} \in \mathcal{R}_{-1/\gamma}$.

Fréchet maximum domain of attraction : heavy-tailed distributions

All cumulative functions which belong to the Fréchet maximum domain of attraction denoted by $\mathcal{D}(Fréchet)$ can be rewritten as

$$\bar{F}(x) = x^{-1/\gamma}\ell(x),$$

where $\gamma > 0$ and $\ell(x)$ is a slowly varying function *i.e.* $\ell(\lambda x)/\ell(x) \to 1$ as $x \to \infty$ for all $\lambda \geq 1$. $\bar{F}(x)$ is said to be regularly varying at infinity with index $-1/\gamma$. This property is denoted by $\bar{F} \in \mathcal{R}_{-1/\gamma}$.

Gumbel maximum domain of attraction : light-tailed distributions

There is no simple representation for distributions which belong to $\mathcal{D}(Gumbel)$. We focus on an interesting sub-family called Weibull tail-distributions

$$\bar{F}(x) = \exp\left(-x^{\alpha}\ell(x)\right),$$

where α is called the Weibull tail-coefficient and $\ell(x)$ is a slowly varying function.

Model established by L. Gardes, S. Girard & A. Guillou

First order condition $(\mathbf{A}_1(\tau,\theta))$

Let us consider the family of survival distribution functions defined as

$(\mathbf{A}_1(\tau,\theta))$ $\bar{F}(x) = \exp(-K_\tau^{\leftarrow}(\log H(x)))$ for $x \geq x_*$ with $x_* > 0$ and

- $K_\tau(y) = \int_1^y u^{\tau-1} du$ where $\tau \in [0,1]$,
- H an increasing function such that $H^{\leftarrow} \in \mathcal{R}_\theta$ where $\theta > 0$.

Model established by L. Gardes, S. Girard & A. Guillou

First order condition $(\mathbf{A}_1(\tau,\theta))$

Let us consider the family of survival distribution functions defined as

$$(\mathbf{A}_1(\tau,\theta)) \quad \bar{F}(x) = \exp(-K_\tau^{\leftarrow}(\log H(x))) \text{ for } x \geq x_* \text{ with } x_* > 0 \text{ and}$$

- $K_\tau(y) = \int_1^y u^{\tau-1} du$ where $\tau \in [0,1]$,
- H an increasing function such that $H^{\leftarrow} \in \mathcal{R}_\theta$ where $\theta > 0$.

Proposition

- F verifies $(\mathbf{A}_1(0,\theta))$ if and only if F is a Weibull-tail distribution function with Weibull tail-coefficient θ.
- If F verifies $(\mathbf{A}_1(\tau,\theta))$, $\tau \in [0,1)$ and if H is twice differentiable then F belongs to the Gumbel maximum domain of attraction.
- F verifies $(\mathbf{A}_1(1,\theta))$ if and only if F is in the Fréchet maximum domain of attraction with tail-index θ.

Estimator of θ

Definition

Denoting by (k_n) an intermediate sequence of integers, the following estimator of θ is considered :

$$\widehat{\theta}_{n,\tau}(k_n) = \frac{H_n(k_n)}{\mu_\tau(\log(n/k_n))},$$

with, for all $t > 0$,

$$\mu_\tau(t) = \int_0^\infty (K_\tau(x+t) - K_\tau(t))\,\mathrm{e}^{-x}dx.$$

Definition

Let us consider (k_n) an intermediate sequence of integers such that $k_n \in \{1 \ldots n\}$ the Hill estimator is given by :

$$H_n(k_n) = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1} \log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}).$$

Model established by L. Gardes, S. Girard & A. Guillou

Definition

An estimator of the extreme quantile x_{p_n} can be deduced by :

$$\widehat{x}_{p_n,\widehat{\theta}_{n,\tau}(k_n)} = X_{n-k_n+1,n} \exp\left(\widehat{\theta}_{n,\tau}(k_n)\left(K_\tau(\log(1/p_n)) - K_\tau(\log(n/k_n))\right)\right).$$

Model established by L. Gardes, S. Girard & A. Guillou

Definition

An estimator of the extreme quantile x_{p_n} can be deduced by :

$$\widehat{x}_{p_n,\widehat{\theta}_{n,\tau}(k_n)} = X_{n-k_n+1,n} \exp\left(\widehat{\theta}_{n,\tau}(k_n)\left(K_\tau(\log(1/p_n)) - K_\tau(\log(n/k_n))\right)\right).$$

Second order condition $(\mathbf{A}_2(\rho))$

To establish the asymptotic normality of the estimators, a second-order condition on ℓ is required :

$(\mathbf{A}_2(\rho))$ There exist $\rho < 0$, a function b satisfying $b(x) \to 0$ and $|b|$ asymptotically decreasing such that uniformly locally on $\lambda > 0$

$$\log\left(\frac{\ell(\lambda x)}{\ell(x)}\right) \sim b(x) K_\rho(\lambda), \text{ when } x \to \infty.$$

It can be shown that necessarily $|b| \in \mathcal{R}_\rho$.

An intuitive justification

Objectives

1. Estimate τ independently from θ,

An intuitive justification

Objectives

1. Estimate τ independently from θ,
2. Replace τ by $\widehat{\tau}_n$ in $\widehat{\theta}_{n,\tau}(k_n)$,

An intuitive justification

Objectives

1. Estimate τ independently from θ,
2. Replace τ by $\widehat{\tau}_n$ in $\widehat{\theta}_{n,\tau}(k_n)$,
3. Replace τ by $\widehat{\tau}_n$ and $\widehat{\theta}_{n,\tau}(k_n)$ by $\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)$ in $\widehat{x}_{p_n,\widehat{\theta}_{n,\tau}(k_n)}$.

An intuitive justification

Objectives

1. Estimate τ independently from θ,
2. Replace τ by $\widehat{\tau}_n$ in $\widehat{\theta}_{n,\tau}(k_n)$,
3. Replace τ by $\widehat{\tau}_n$ and $\widehat{\theta}_{n,\tau}(k_n)$ by $\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)$ in $\widehat{x}_{p_n,\widehat{\theta}_{n,\tau}(k_n)}$.

Note that for (k_n) and (k_n') two intermediate sequences of integers such that $\widehat{\theta}_{n,\tau}(k_n) \xrightarrow{P} \theta$ and $\widehat{\theta}_{n,\tau}(k_n') \xrightarrow{P} \theta$ and $k_n' > k_n$ we have

$$\frac{\widehat{\theta}_{n,\tau}(k_n)}{\widehat{\theta}_{n,\tau}(k_n')} = \frac{H_n(k_n)}{H_n(k_n')} \frac{\mu_\tau(\log(n/k_n'))}{\mu_\tau(\log(n/k_n))} \xrightarrow{P} 1.$$

Then,

$$\frac{H_n(k_n)}{H_n(k_n')} \overset{P}{\sim} \frac{\mu_\tau(\log(n/k_n))}{\mu_\tau(\log(n/k_n'))} = \psi(\tau; \log(n/k_n), \log(n/k_n')),$$

where

$$\psi(x; t, t') = \frac{\mu_x(t)}{\mu_x(t')} \quad \text{is a bijection from} \quad \mathbb{R} \quad \text{to} \quad (-\infty, exp(t - t')) .$$

Estimator of τ

Definition

Denoting by (k_n) and (k_n') two intermediate sequences of integers such that $k_n' > k_n$, the following estimator of τ is considered :

$$\widehat{\tau}_n = \begin{cases} \psi^{-1}\left(\frac{H_n(k_n)}{H_n(k_n')}; \log(n/k_n), \log(n/k_n')\right) & \text{if} \quad \frac{H_n(k_n)}{H_n(k_n')} < \frac{k_n'}{k_n}, \\ u & \text{if} \quad \frac{H_n(k_n)}{H_n(k_n')} \geq \frac{k_n'}{k_n}. \end{cases}$$

where u is the realization of a standard uniform distribution.

Asymptotic properties

Asymptotic normality of $\widehat{\tau}_n$

Suppose that $(\mathbf{A}_1(\tau,\theta))$ and $(\mathbf{A}_2(\rho))$ hold. Let (k_n) and (k_n') be two intermediate sequences of integers such that

$$(H_1) \quad k_n \to \infty, \quad k_n'/n \to 0, \quad k_n/k_n' \to 0, \quad \sqrt{k_n'}\, b(\exp K_\tau(\log n/k_n')) \to 0,$$

$$(H_2) \quad \log(n/k_n')\left(\log_2(n/k_n) - \log_2(n/k_n')\right) \to \infty,$$

$$\sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right) \to \infty.$$

we have :

$$\sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right)(\widehat{\tau}_n - \tau) \xrightarrow{d} \mathcal{N}(0,1).$$

where $\log_2 = \log(\log)$.

Asymptotic properties

Replacing τ by $\widehat{\tau}_n$ we obtain

$$\widehat{\theta}_{n,\widehat{\tau}_n}(k_n) = \frac{H_n(k_n)}{\mu_{\widehat{\tau}_n}(\log(n/k_n))}.$$

Asymptotic properties

Replacing τ by $\widehat{\tau}_n$ we obtain

$$\widehat{\theta}_{n,\widehat{\tau}_n}(k_n) = \frac{H_n(k_n)}{\mu_{\widehat{\tau}_n}(\log(n/k_n))}.$$

Asymptotic normality of $\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)$

Suppose that **($\mathbf{A}_1(\tau,\theta)$)** and **($\mathbf{A}_2(\rho)$)** hold. Let (k_n) and (k_n') be two intermediate sequences of integers such that (H_1), (H_2) hold with

$$(H_3) \quad \left(\log_2(n/k_n) - \log_2(n/k_n')\right) / \log_2(n/k_n) \to 0,$$

$$(H_4) \sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right) / \log_2(n/k_n) \to \infty.$$

we have :

$$\frac{\sqrt{k_n}\left(\log_2(n/k_n) - \log_2(n/k_n')\right)}{\log_2(n/k_n)}\left(\widehat{\theta}_{n,\widehat{\tau}_n}(k_n) - \theta\right) \xrightarrow{d} \mathcal{N}(0,\theta^2).$$

Asymptotic properties

Replacing τ by $\widehat{\tau}_n$ and $\widehat{\theta}_{n,\tau}(k_n)$ by $\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)$ we obtain

$$\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)} = X_{n-k_n+1,n} \exp\left(\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)\left(K_{\widehat{\tau}_n}(\log(1/p_n)) - K_{\widehat{\tau}_n}(\log(n/k_n))\right)\right).$$

Asymptotic properties

Replacing τ by $\widehat{\tau}_n$ and $\widehat{\theta}_{n,\tau}(k_n)$ by $\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)$ we obtain

$$\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)} = X_{n-k_n+1,n} \exp\left(\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)\left(K_{\widehat{\tau}_n}(\log(1/p_n)) - K_{\widehat{\tau}_n}(\log(n/k_n))\right)\right).$$

Asymptotic normality of $\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)}$

Suppose that $(\mathbf{A}_1(\tau,\theta))$ and $(\mathbf{A}_2(\rho))$ hold. Let (k_n) and (k_n') be two intermediate sequences of integers such that (H_1), (H_2), (H_3), (H_4) hold with

$$(\log(n/k_n))^{1-\tau}(K_\tau(\log(1/p_n)) - K_\tau(\log(n/k_n))) \to \infty,$$

$$(\log_2(1/p_n))/\sqrt{k_n}(\log_2(n/k_n) - \log_2(n/k_n')) \to 0,$$

$$\frac{\log_2(n/k_n)(K_\tau(\log(1/p_n)) - K_\tau(\log(n/k_n)))}{\int_{\log(n/k_n)}^{\log(1/p_n)} \log(u)u^{\tau-1}du} \to 0,$$

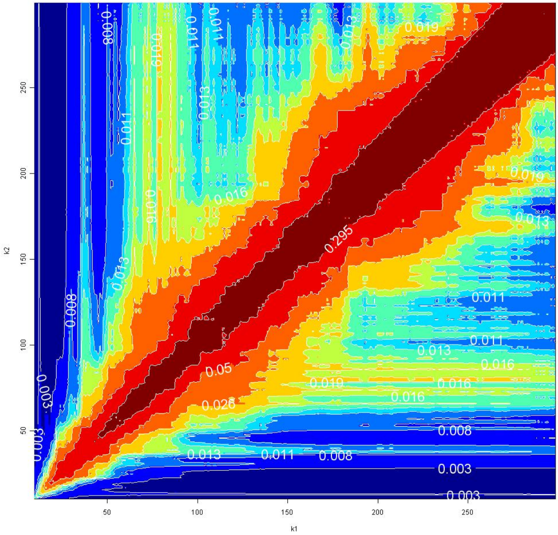
we have :

$$\frac{\sqrt{k_n}\,(\log_2(n/k_n) - \log_2(n/k_n'))}{\int_{\log(n/k_n)}^{\log(1/p_n)} \log(u)u^{\tau-1}du}\left(\frac{\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)}}{x_{p_n}} - 1\right) \xrightarrow{d} \mathcal{N}(0,\theta^2).$$

Simulations

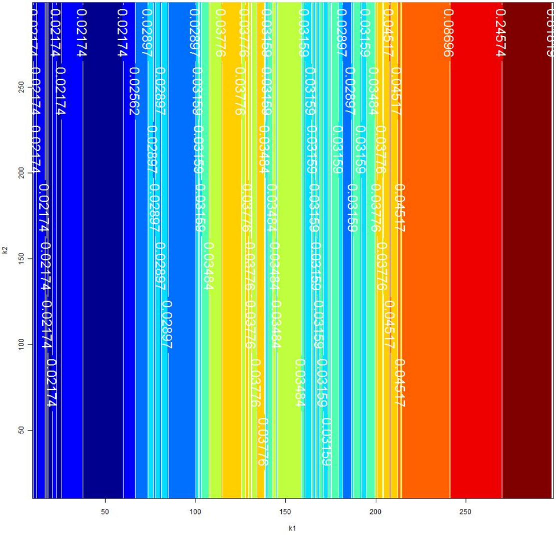
- We generate $N = 100$ samples $(\mathcal{X}_{n,i})_{i=1,\dots,N}$ of size $n = 300$.
- On each sample $(\mathcal{X}_{n,i})$, the estimator $\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)}$ is computed for $k_n = 2, \dots, 299$ and $k'_n = k_n, \dots, 300$.
- In what follows we show simulation results for quantiles corresponding to $p_n = 1/2n = 1.6 * 10^{-3}$.
- The associated deciles of the empirical Mean-Squared Error $\mathcal{MSE}$ are plotted.
- Comparison with an estimator of A. L. M. Dekkers, J.H.J. Einmahl & L. de Haan.

Gamma distribution for $\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)}$ / $\mathcal{D}(Gumbel)$



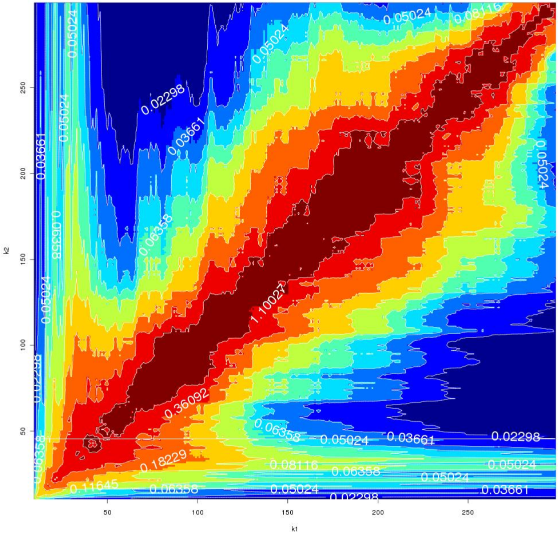
Empirical deciles of the Mean-Square Error

Gamma distribution for an estimator of A. L. M. Dekkers et al.



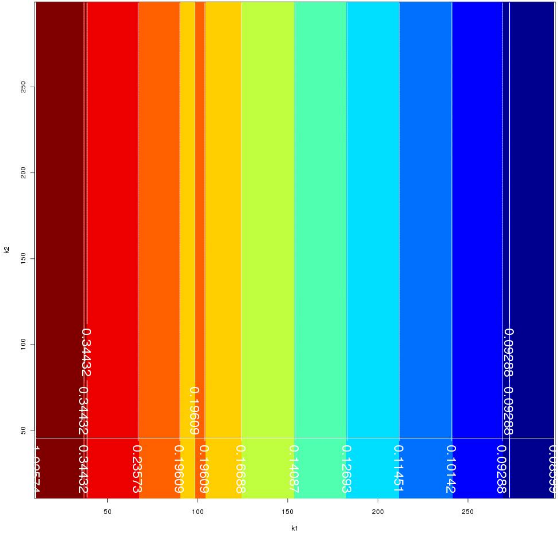
Empirical deciles ot the Mean-Square Error

Pareto distribution for $\widehat{x}_{p_n,\widehat{\theta}_{n,\widehat{\tau}_n}(k_n)}$ / $\mathcal{D}(Fréchet)$



Empirical deciles of the Mean-Square Error

Pareto distribution for an estimator of A. L. M. Dekkers et al.



Empirical deciles of the Mean-Square Error

Concluding remarks

Conclusions and Further Work

- The choice of the parameters k_n and k_n' in practice.

Concluding remarks

Conclusions and Further Work

- The choice of the parameters k_n and k_n' in practice.
- Adapt our results to the case $\tau > 1$ and investigate the possible link with super-heavy tails.

Concluding remarks

Conclusions and Further Work

- The choice of the parameters k_n and k'_n in practice.
- Adapt our results to the case $\tau > 1$ and investigate the possible link with super-heavy tails.
- Extend this work to random variable $Y = \varphi(X)$ where X has a parent distribution satisfying $(\mathbf{A}_1(\tau, \theta))$.

Concluding remarks

Conclusions and Further Work

- The choice of the parameters k_n and k'_n in practice.
- Adapt our results to the case $\tau > 1$ and investigate the possible link with super-heavy tails.
- Extend this work to random variable $Y = \varphi(X)$ where X has a parent distribution satisfying $(\mathbf{A}_1(\tau, \theta))$.
- For instance, choosing $\varphi(x) = x^* - 1/x$ would allow to consider distributions in the Weibull maximum domain of attraction (with finite endpoint x^*).

Main references

Dekkers, A. L. M., Einmahl, J. H. J., De Haan, L., (1989). A Moment Estimator for the Index of an Extreme-Value Distribution, *The Annals of Statistics*, **17**, 1833–1855.

Gardes, L., Girard, S., Guillou, A., (2011). Weibull tail-distributions revisited : a new look at some tail estimators, *Journal of Statistical Planning and Inference*, **141**, 429–444.

Gnedenko, B.V., (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *The Annals of Mathematics*, **44**, 423–453.

Hill, B.M., (1975). A simple general approach to inference about the tail of a distribution, *The Annals of Statistics*, **3**, 1163–1174.

Thank you for your attention