



Statistical inference for Weibull tail-distributions

Stéphane Girard

► **To cite this version:**

Stéphane Girard. Statistical inference for Weibull tail-distributions. Journées Extrêmes à Lille, Mar 2006, Lille, France. <hal-00985807>

HAL Id: hal-00985807

<https://hal.inria.fr/hal-00985807>

Submitted on 30 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

STATISTICAL INFERENCE FOR WEIBULL TAIL-DISTRIBUTIONS

Stéphane Girard

SMS/LMC-IMAG, Université Grenoble 1, INRIA

Outline

1. Weibull tail-distributions.
2. Kernel estimators of the Weibull tail-coefficient.
3. Bias-reduced estimator of the Weibull tail-coefficient.
4. Estimation of extreme quantiles.
5. Simulation study.
6. Nidd river data.

1. Weibull tail-distributions.

Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with cumulative distribution function F such that

$$(A1) \quad 1 - F(x) = \exp(-H(x)), \quad H^{-1}(t) = \inf\{x, H(x) \geq t\} = t^\theta \ell(t),$$

where

- $\theta > 0$ is the Weibull tail-coefficient,
- ℓ is a slowly varying function *i.e.*

$$\ell(\lambda x)/\ell(x) \rightarrow 1 \text{ as } x \rightarrow \infty \text{ for all } \lambda > 0.$$

The inverse failure rate function H^{-1} is said to be regularly varying at infinity with index θ and this property is denoted by $H^{-1} \in \mathcal{R}_\theta$.

In the following, we also assume a second order condition on ℓ :

(A2) There exist $\rho \leq 0$ and $b(x) \rightarrow 0$ such that uniformly locally on $\lambda \geq 1$

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

with

$$K_\rho(\lambda) = \int_1^\lambda u^{\rho-1} du.$$

It can be shown that necessarily $|b| \in \mathcal{R}_\rho$. The second order parameter $\rho \leq 0$ tunes the rate of convergence of $\ell(\lambda x)/\ell(x)$ to 1. The closer ρ is to 0, the slower is the convergence.

Examples:

	θ	$\ell(x)$	$b(x)$	ρ
Absolute Gaussian $ \mathcal{N} (\mu, \sigma^2)$	$1/2$	$2^{1/2}\sigma - \frac{\sigma}{2^{3/2}} \frac{\log x}{x} + O(1/x)$	$\frac{1}{4} \frac{\log x}{x}$	-1
Gamma $\Gamma(\alpha \neq 1, \beta)$	1	$\frac{1}{\beta} + \frac{\alpha - 1}{\beta} \frac{\log x}{x} + O(1/x)$	$(1 - \alpha) \frac{\log x}{x}$	-1
Weibull $\mathcal{W}(\alpha, \lambda)$	$1/\alpha$	λ	0	$-\infty$

Related work: Berred (1991), Broniatowski (1993), Beirlant, Broniatowski, Teugels, Vynckier, (1995), Beirlant, Bouquiaux, Werker (2006).

2. Kernel estimators of the Weibull tail-coefficient.

Principle: Our approach is based on the following approximation. Denoting by $q(t)$ the quantile function

$$q(t) = F^{-1}(1 - t) = H^{-1}(\log(1/t)) = (\log(1/t))^\theta \ell(\log(1/t)),$$

we obtain for t and s small:

$$\begin{aligned} \log(q(t)) - \log(q(s)) &= \theta(\log_2(1/t) - \log_2(1/s)) + \log\left(\frac{\ell(\log(1/t))}{\ell(\log(1/s))}\right) \\ &\simeq \theta(\log_2(1/t) - \log_2(1/s)), \end{aligned} \tag{1}$$

where $\log_2(x) = \log \log(x)$ and since $\ell \in \mathcal{R}_0$. Considering $t = i/n$, $s = k_n/n$ and replacing F by its empirical counterpart yield

$$\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) \simeq \theta(\log_2(n/i) - \log_2(n/k_n)),$$

for $i = 1, \dots, k_n - 1$ and where k_n is an intermediate sequence.

Our method: Estimation via linear combination of upper order statistics.

$$\hat{\theta}_n(\alpha) = \sum_{i=1}^{k_n-1} \alpha_{i,n} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \bigg/ \sum_{i=1}^{k_n-1} \alpha_{i,n} (\log_2(n/i) - \log_2(n/k_n)),$$

where

- $\alpha_{i,n} = W(i/k_n) + \varepsilon_{i,n}$,
- $\varepsilon_{i,n}$, $i = 1, \dots, k_n - 1$ is a non-random sequence, and
- $W : [0, 1] \rightarrow \mathbb{R}$ is a smooth score function, verifying

(A3) W has a continuous derivative W' on $(0, 1)$,

(A4) There exist $M > 0$, $0 \leq q < 1/2$ and $p < 1$ such that, for all $x \in (0, 1)$:

$$|W(x)| \leq Mx^{-q} \text{ and } |W'(x)| \leq Mx^{-p-q}.$$

Asymptotic normality: Suppose **(A1)**–**(A4)** hold. Then

$$k_n^{1/2}(\hat{\theta}_n(\alpha) - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta, W)),$$

for any sequence (k_n) satisfying $k_n \rightarrow \infty$ and

$$k_n^{1/2} \max\{b(\log(n/k_n)), 1/\log(n/k_n), \|\varepsilon\|_{n,\infty}\} \rightarrow 0,$$

where we have defined:

$$\begin{aligned} \|\varepsilon\|_{n,\infty} &= \max_{i=1,\dots,k_n-1} |\varepsilon_{i,n}|, \\ \mu(W) &= \int_0^1 W(x) \log(1/x) dx, \\ \sigma^2(W) &= \int_0^1 \int_0^1 W(x)W(y) \frac{\min(x,y)(1 - \max(x,y))}{xy} dx dy, \\ \sigma(\theta, W) &= \theta \frac{\sigma(W)}{\mu(W)}. \end{aligned}$$

Example 1. Constant weights $\alpha_{i,n} = 1$ for all $i = 1, \dots, k_n - 1$ yield an existing estimator (Girard, 2004):

$$\hat{\theta}_n^G = \frac{\sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}))}{\sum_{i=1}^{k_n-1} (\log_2(n/i) - \log_2(n/k_n))}.$$

We found back the same limiting result: If **(A1)** and **(A2)** hold then

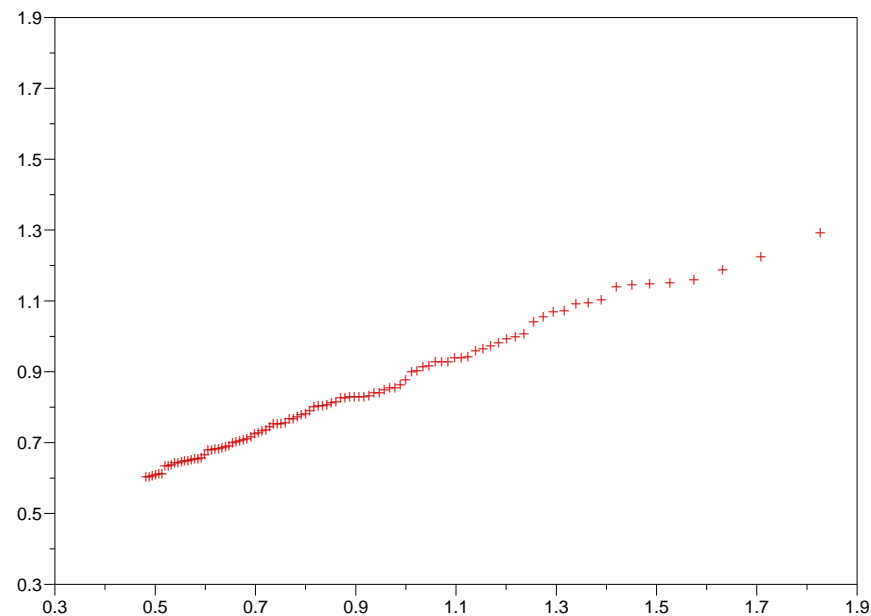
$$k_n^{1/2}(\hat{\theta}_n^G - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2),$$

for any sequence (k_n) satisfying $k_n \rightarrow \infty$ and

$$k_n^{1/2} \max\{b(\log(n/k_n)), 1/\log(n/k_n)\} \rightarrow 0.$$

Example 2: A new estimator of the Weibull tail-coefficient based on a QQ-plot.

- Drawing the pairs $(\log_2(n/i), \log(X_{n-i+1,n}))$ for $i = 1, \dots, k_n$ gives a graph which is approximatively linear (with slope θ).
- Example : $|\mathcal{N}|(0, 1)$ distribution, $n = 500$, $k_n = 100$.



- $\hat{\theta}_n^Z$ is the least square estimator of θ based on the k_n largest observations:

$$\hat{\theta}_n^Z = \frac{\sum_{i=1}^{k_n-1} (\log_2(n/i) - \tau_n) \log(X_{n-i+1,n})}{\sum_{i=1}^{k_n-1} (\log_2(n/i) - \tau_n) \log_2(n/i)},$$

where

$$\tau_n = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} \log_2(n/i).$$

- Similar to the Zipf estimator for the extreme value index proposed by Kratz, Resnick (1996) and Schultze, Steinebach (1996).
- Particular case of $\hat{\theta}_n(\alpha)$ with $W(x) = -(\log(x) + 1)$. Thus, under **(A1)** and **(A2)**,

$$k_n^{1/2}(\hat{\theta}_n^Z - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2),$$

for any sequence (k_n) satisfying $k_n \rightarrow \infty$ and

$$k_n^{1/2} \max\{b(\log(n/k_n)), \log^2(k_n)/\log(n/k_n)\} \rightarrow 0.$$

3. Bias-reduced estimator of the Weibull tail-coefficient.

Principle: We focus on the case where the convergence in **(A2)** is slow, *i.e.*

(A5) $x|b(x)| \rightarrow \infty$ as $x \rightarrow \infty$.

Let us note that this condition implies $\rho \geq -1$. Gamma and (absolute) Gaussian distribution fulfill **(A5)** whereas Weibull distribution do not. Condition **(A2)** can be used to precise approximation **(1)**:

$$\log(q(t)) - \log(q(s)) = \theta(\log_2(1/t) - \log_2(1/s)) + b(\log(1/s))K_\rho \left(\frac{\log(1/t)}{\log(1/s)} \right) (1 + o(1)). \quad (2)$$

Exponential regression models:

- Define $Z_i = i \log(n/i)(\log X_{n-i+1,n} - \log X_{n-i,n})$, $i = 1, \dots, k_n$. Then, under **(A1)**, **(A2)**, **(A5)**,

$$\sup_{1 \leq i \leq k_n} \left| Z_i - \left(\theta + b(\log(n/k_n)) \left(\frac{\log(n/i)}{\log(n/k_n)} \right)^\rho \right) f_i \right| = o_{\mathbb{P}}(b(\log(n/k_n))), \quad (3)$$

for any sequence (k_n) such that $k_n \rightarrow \infty$ and $\log k_n / \log n \rightarrow 0$, and where (f_1, \dots, f_{k_n}) is a vector of independent and standard exponentially distributed random variables.

- Similar to the ones proposed by Beirlant, Dierckx, Goegebeur, Matthys (1999), Feuerverger, Hall (1999) and Beirlant, Dierckx, Guillou, Starica (2002) in the case of Pareto-type distributions.
- One can plug the canonical choice $\rho = -1$ in the regression model (3) without perturbing the approximation so that

$$\sup_{1 \leq i \leq k_n} \left| Z_i - \left(\theta + b(\log(n/k_n)) \frac{\log(n/k_n)}{\log(n/i)} \right) f_i \right| = o_{\mathbb{P}}(b(\log(n/k_n))). \quad (4)$$

Application 1: Bias-reduced estimator. Estimation of θ and $b(\log(n/k_n))$ by a Least-Square method after substituting ρ with the value -1 :

$$\hat{\theta}_n^R = \bar{Z}_{k_n} - \hat{b}(\log(n/k_n))\bar{x}_{k_n}, \quad \hat{b}(\log(n/k_n)) = \frac{\sum_{i=1}^{k_n} (x_i - \bar{x}_{k_n})Z_i}{\sum_{i=1}^{k_n} (x_i - \bar{x}_{k_n})^2}, \quad (5)$$

where $x_i = \log(n/k_n)/\log(n/i)$, $\bar{x}_{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n} x_i$ and $\bar{Z}_{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n} Z_i$.

Asymptotic normality under **(A1)**, **(A2)**, **(A5)**:

$$\frac{k_n^{1/2}}{\log(n/k_n)} (\hat{\theta}_n^R - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$$

for any sequence (k_n) such that $k_n \rightarrow \infty$ and

$$k_n^{1/2} b(\log(n/k_n)) / \log(n/k_n) \rightarrow \Lambda \neq 0.$$

Application 2: Adaptive selection of k_n . Adapted from Matthys, Beirlant (2003) in the context of extreme-value index estimation. Neglecting the bias correction in (5) yields the Maximum Likelihood estimator:

$$\hat{\theta}_n^{ML} = \bar{Z}_{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n} i \log(n/i) (\log X_{n-i+1,n} - \log X_{n-i,n}).$$

From (4), the Asymptotic Mean Squared Error (AMSE) associated to $\hat{\theta}_n^{ML}$ is given by

$$AMSE(\hat{\theta}_n^{ML}) = \frac{\theta^2}{k_n} + \left(b(\log(n/k_n)) \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{\log(n/k_n)}{\log(n/i)} \right)^2,$$

and can be estimated by

$$\widehat{AMSE}(\hat{\theta}_n^{ML}) = \frac{(\hat{\theta}_n^R)^2}{k_n} + \left(\hat{b}(\log(n/k_n)) \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{\log(n/k_n)}{\log(n/i)} \right)^2.$$

The minimization of $\widehat{AMSE}(\hat{\theta}_n^{ML})$ with respect to k_n gives rise to an adaptive selection procedure.

4. Estimation of extreme quantiles.

Principle: An extreme quantile x_{p_n} of order $p_n < 1/n$ is defined by:

$$1 - F(x_{p_n}) = p_n.$$

Recall that, for small t and s

$$\frac{q(t)}{q(s)} = \frac{H^{-1}(\log(1/t))}{H^{-1}(\log(1/s))} = \left(\frac{\log(1/t)}{\log(1/s)} \right)^\theta \frac{\ell(\log(1/t))}{\ell(\log(1/s))} \simeq \left(\frac{\log(1/t)}{\log(1/s)} \right)^\theta.$$

Considering $t = p_n$, $s = k_n/n$, replacing F by its empirical counterpart and θ by an estimator $\hat{\theta}_n$ yield the following estimator

$$\hat{x}_{p_n}(\hat{\theta}_n) = X_{n-k_n+1,n} \left(\frac{\log(1/p_n)}{\log(n/k_n)} \right)^{\hat{\theta}_n} = X_{n-k_n+1,n} \exp \left(\hat{\theta}_n \log \tau_n \right),$$

where we have defined $\tau_n = \log(1/p_n)/\log(n/k_n)$.

Asymptotic normality: Based on the following result (Gardes, Girard, 2005). Suppose **(A1)** and **(A2)** hold. If moreover

$$1 \leq \liminf \tau_n \leq \limsup \tau_n < \infty,$$

and

$$k_n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

then

$$\frac{\log(n/k_n)k_n^{1/2}}{\log(k_n/(np_n))} \left(\frac{\hat{x}_{p_n}(\hat{\theta}_n)}{x_{p_n}} - 1 \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

for any sequence (k_n) satisfying $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ and

$$k_n^{1/2}b(\log(n/k_n)) \rightarrow 0.$$

As a consequence, we obtain the asymptotic normality of $\hat{x}_{p_n}(\hat{\theta}_n^G)$, $\hat{x}_{p_n}(\hat{\theta}_n^Z)$ and $\hat{x}_{p_n}(\hat{\theta}_n^{\text{BBTV}})$ where

$$\hat{\theta}_n^{\text{BBTV}} = \frac{\log(n/k_n)}{X_{n-k_n+1,n}} \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} (X_{n-i+1,n} - X_{n-k_n+1,n})$$

is the estimator introduced by Beirlant, Broniatowski, Teugels, Vinckier (1995).

Bias-reduced estimator: Basing on the refined approximation (2) of $\log q(t) - \log q(s)$, it is natural to introduce

$$\hat{x}_{p_n}^R = X_{n-k_n+1,n} \exp \left(\hat{\theta}_n^R \log \tau_n + \hat{b}(\log(n/k_n)) K_{-1}(\tau_n) \right).$$

Asymptotic normality under **(A1)**, **(A2)**, **(A5)**:

$$\frac{k_n^{1/2}}{\log(n/k_n)} \left(\frac{\hat{x}_{p_n}^R}{x_{p_n}} - 1 \right) \xrightarrow{d} \mathcal{N}(\Lambda \mu(\tau), \theta^2 \sigma^2(\tau))$$

for any sequence (k_n) such that $k_n \rightarrow \infty$, $\tau_n \rightarrow \tau > 1$ and

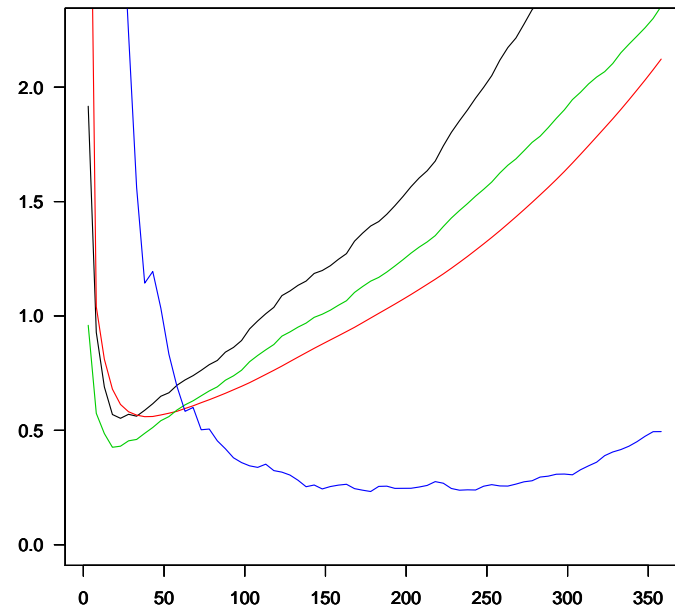
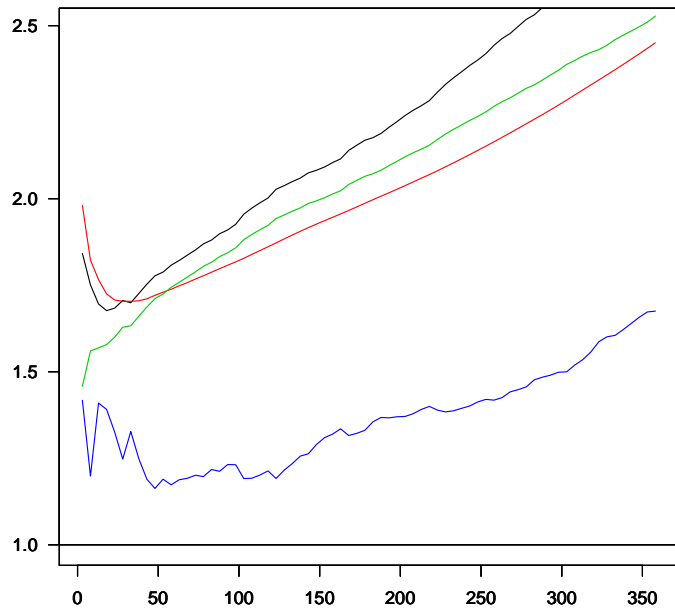
$$k_n^{1/2} b(\log(n/k_n)) / \log(n/k_n) \rightarrow \Lambda \neq 0,$$

where $\mu(\tau) = (K_{-1}(\tau) - K_\rho(\tau))$ and $\sigma^2(\tau) = (K_{-1}(\tau) - \log(\tau))^2$.

5. Simulation study.

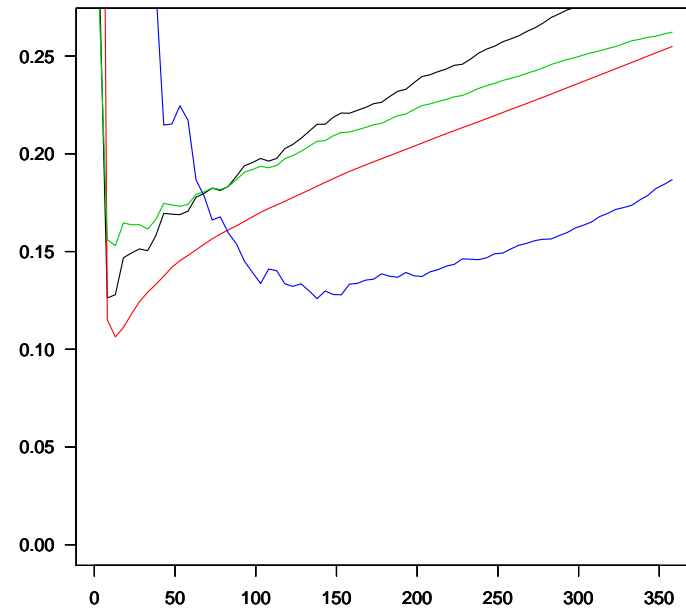
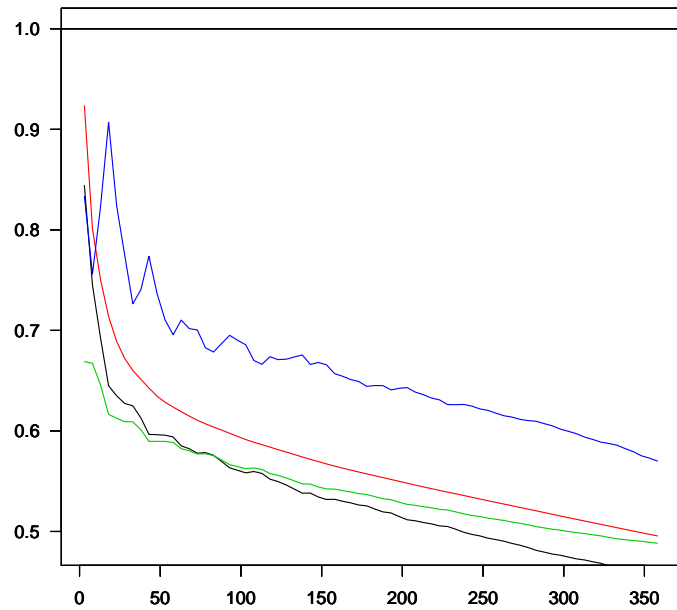
- Comparison of $\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue) to the true θ (black horizontal line).
- Simulated distributions : Absolute Gaussian $|\mathcal{N}|(0, 1)$, Gamma $\Gamma(0.25, 1)$, $\Gamma(4, 1)$, and Weibull $\mathcal{W}(4, 4)$, $\mathcal{W}(0.25, 0.25)$.
- Sample size $n = 500$, $k_n \in \{2, \dots, 360\}$, 100 replications.
- Computation of the mean estimate (Hill plot, left pannel) and of the Mean Square Error (MSE, right pannel).

Gamma $\Gamma(0.25, 1) \longrightarrow \theta = 1, b(x) > 0$



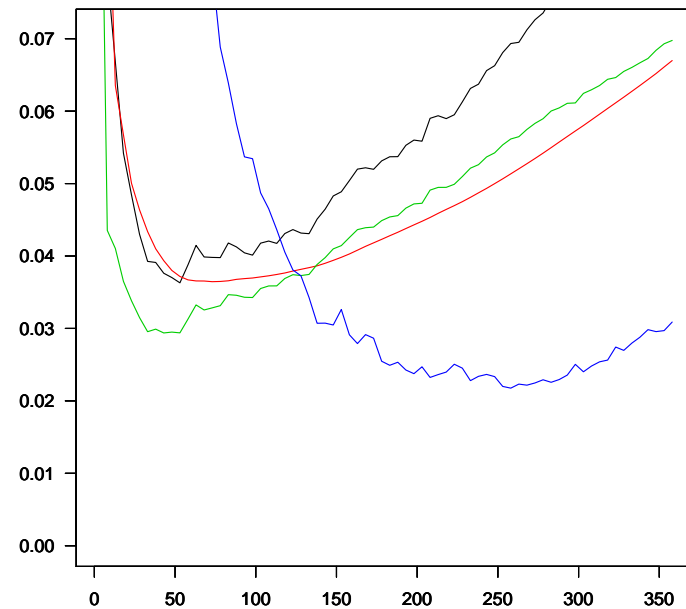
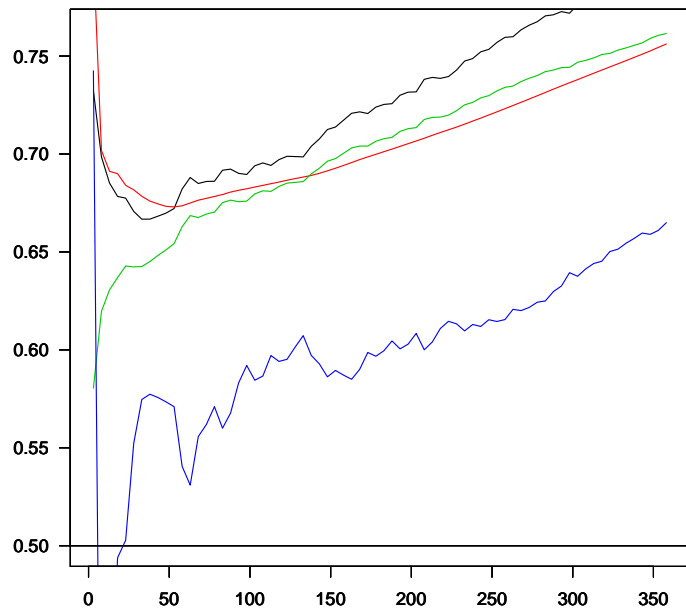
$\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue)

Gamma $\Gamma(4, 1) \longrightarrow \theta = 1, b(x) < 0$



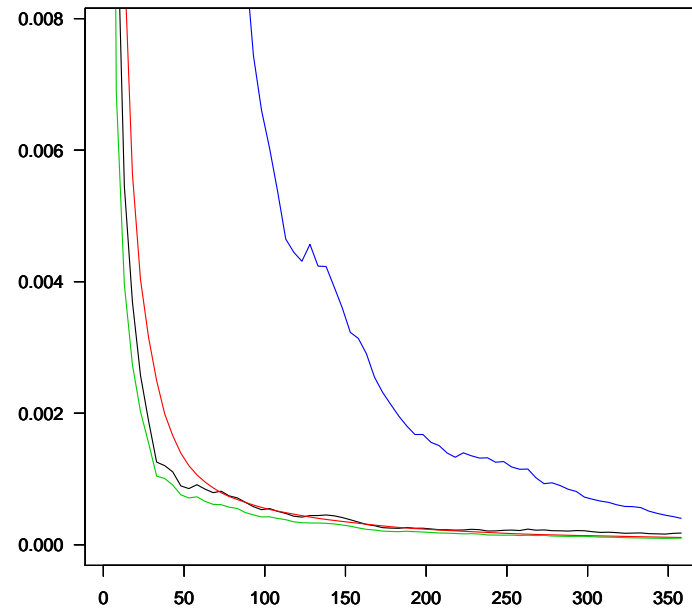
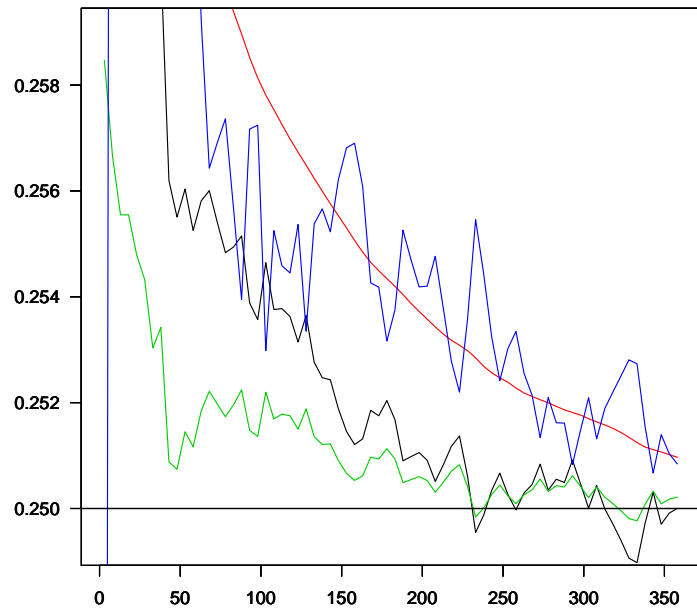
$\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue)

Absolute Gaussian $|\mathcal{N}|(0, 1) \longrightarrow \theta = 1/2, b(x) > 0$



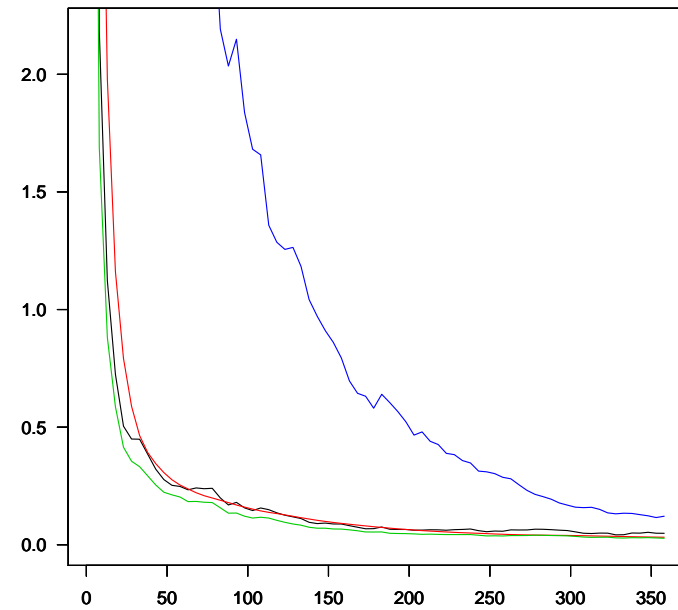
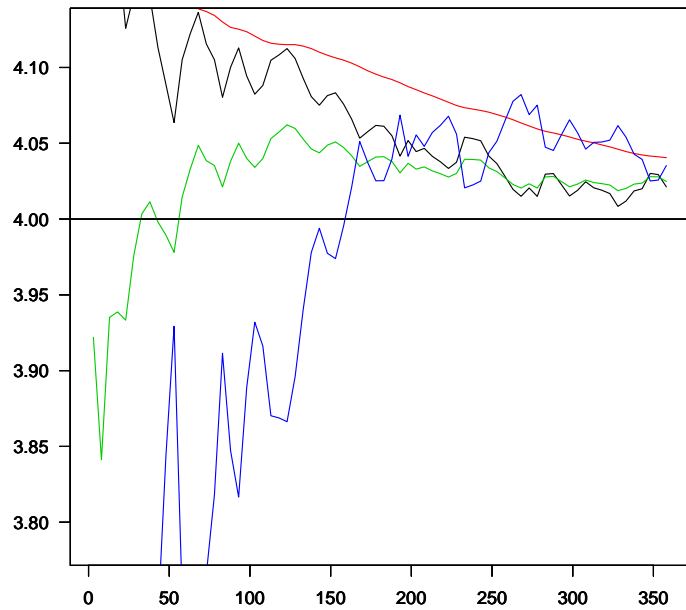
$\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue)

Weibull $\mathcal{W}(4, 4) \longrightarrow \theta = 0.25, b(x) = 0$



$\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue)

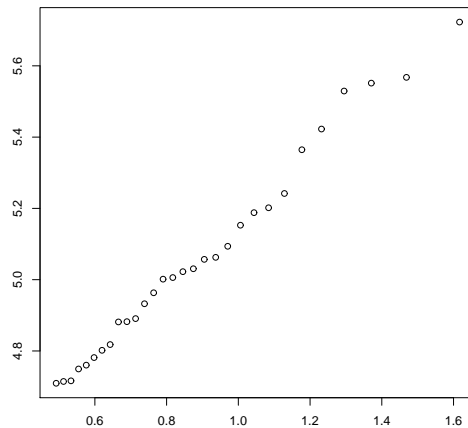
Weibull $\mathcal{W}(0.25, 0.25) \longrightarrow \theta = 4, b(x) = 0$



$\hat{\theta}_n^G$ (in black), $\hat{\theta}_n^Z$ (in red), $\hat{\theta}_n^{ML}$ (in green) and $\hat{\theta}_n^R$ (in blue)

6. Nidd river data.

- 154 exceedances of the level $65 \text{ m}^3\text{s}^{-1}$ by the river Nidd (Yorkshire, England) during the period 1934-1969 (35 years).
- Widely used in extreme value studies *i.e.* Hosking, Wallis, Wood (1985) and Davison, Smith (1990).
- The N -year return level is the water level which is exceeded on average once in N years.



QQ-plot

$$\hat{k}_n = 29,$$

$$\hat{\theta}_n^{ML} \simeq 0.89,$$

Estimation of the 100-year return level:

$$\hat{x}_{p_n}(\hat{\theta}_n^{ML}) = 366 \text{ m}^3\text{s}^{-1}$$

References

- L. Gardes & S. Girard. Estimation de quantiles extrêmes pour les lois queue de type Weibull : une synthèse bibliographique, *Journal de la Société Française de Statistique*, 154, 98–118, 2013.
- J. El Methni, L. Gardes, S. Girard & A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions, *Journal of Statistical Planning and Inference*, 142(10), 2735–2747, 2012.
- L. Gardes, S. Girard & A. Guillou. Weibull tail-distributions revisited: a new look at some tail estimators, *Journal of Statistical Planning and Inference*, 141(1), 429–444, 2011.
- J. Diebolt, L. Gardes, S. Girard & A. Guillou. Bias-reduced extreme quantiles estimators of Weibull-tail distributions, *Journal of Statistical Planning and Inference*, 138, 1389–1401, 2008.
- L. Gardes & S. Girard. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics, *Journal of Statistical Planning and Inference*, 138, 1416–1427, 2008.

References

- J. Diebolt, L. Gardes, S. Girard & A. Guillo. Bias-reduced estimators of the Weibull tail-coefficient, *Test*, 17, 311–331, 2008.
- L. Gardes & S. Girard. Comparison of Weibull tail-coefficient estimators, *REVSTAT - Statistical Journal*, 4(2):163-188, 2006.
- L. Gardes & S. Girard. Estimating extreme quantiles of Weibull tail-distributions, *Communication in Statistics - Theory and Methods*, 34, 1065-1080, 2005.
- S. Girard. A Hill type estimate of the Weibull tail-coefficient, *Communication in Statistics - Theory and Methods*, 33(2), 205-234, 2004