

Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model

Stéphane GIRARD⁽¹⁾

joint work with

Aboubacrène Ag AHMAD^(2,3), Aliou DIOP⁽³⁾, El Hadji DEME⁽³⁾ and Antoine USSEGLIO-CARLEVE^(1,4)

⁽¹⁾Inria, Univ. Grenoble Alpes, France.

⁽²⁾Univ. des Sciences, des Techniques et des Technologies de Bamako, Mali.

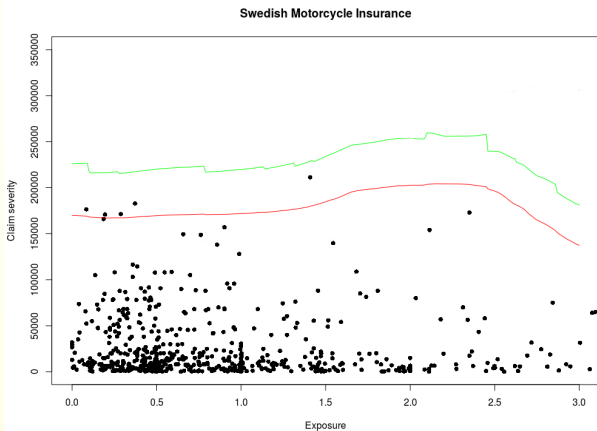
⁽³⁾Univ. Gaston Berger de Saint-Louis, LERSTAD, Sénégal,

⁽⁴⁾Ensaï, Univ. Rennes, France.

This research was partially supported by the French National Research Agency under the grant ANR-19-CE40-0013/ExtremReg project. S. Girard gratefully acknowledges the support of the Chair Stress Test, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas.

Goal: estimation of extreme conditional quantiles

Quantile of extreme level $\alpha \in (0, 1)$ associated with a response variable $Y \in \mathbb{R}$ given a covariate $x \in \mathbb{R}^d$.



Outline

- 1 Location-dispersion regression model
- 2 Inference
- 3 Asymptotic results
- 4 Validation on simulated data
- 5 Application to tsunami data

Location-dispersion regression model for heavy-tailed distributions

Assume the following regression model between a random response variable $Y \in \mathbb{R}$ and a deterministic covariate vector $x \in \Pi \subset \mathbb{R}^d$, $d \geq 1$:

$$Y = a(x) + b(x) Z,$$

where

- $a: \Pi \rightarrow \mathbb{R}$: (unknown) **regression** / **location** function,
- $b: \Pi \rightarrow \mathbb{R}_+ \setminus \{0\}$: (unknown) **dispersion** / **scaling** function,
- Z is a heavy-tailed random variable with tail-index $\gamma > 0$, *i.e.* with survival function $\bar{F}_Z(z) = z^{-1/\gamma} L(z)$, with L is a slowly-varying function such that for all $t > 0$,

$$\lim_{z \rightarrow \infty} \frac{L(tz)}{L(z)} = 1.$$

☞ **Consequence:** Conditional survival function of Y :

$$\bar{F}_Y(y | x) := \mathbb{P}(Y > y | x) = \bar{F}_Z\left(\frac{y - a(x)}{b(x)}\right) = \left(\frac{y - a(x)}{b(x)}\right)^{-1/\gamma} L\left(\frac{y - a(x)}{b(x)}\right),$$

The tail-index of the response variable does not depend on the covariate.

☞ **Identifiability issue:** Let $(\mu_1, \mu_2, \mu_3) \in (0, 1)^3$ such that $q_Z(\mu_2) = 0$ and $q_Z(\mu_3) - q_Z(\mu_1) = 1$ where $q_Z(\cdot)$ is the quantile associated with the survival function of Z .

We thus have

$$a(x) = q_Y(\mu_2 | x) \text{ and } b(x) = q_Y(\mu_3 | x) - q_Y(\mu_1 | x),$$

for all $x \in \Pi$ and where $q_Y(\cdot | x)$ is the conditional quantile (associated with the survival function) of Y .

☞ **Data: Multidimensional fixed design setting**

$\{(Y_1, x_1), \dots, (Y_n, x_n)\}$ a n -sample from the location-dispersion regression model $Y_i = a(x_i) + b(x_i)Z_i$, where Z_1, \dots, Z_n are iid from a heavy-tailed distribution. The design points x_i are all distinct from each other and included in Π , a compact subset of \mathbb{R}^d . Let $\{\Pi_i, i = 1, \dots, n\}$ be a partition of Π such that $x_i \in \Pi_i$.

☞ **Goal: Estimation of extreme conditional quantiles**

$$q_Y(\alpha_n | x) = a(x) + b(x)q_Z(\alpha_n) \text{ as } \alpha_n \rightarrow 0.$$

Such quantiles are asymptotically located outside the convex hull of the sample.

☞ **Inference: Three step estimation procedure**

Inference

Step 1: Estimation of regression and dispersion functions

- Kernel estimator of the conditional survival function $\bar{F}_Y(y | x)$:

$$\hat{\bar{F}}_{n,Y}(y | x) = \sum_{i=1}^n \mathbb{1}_{\{Y_i > y\}} \int_{\Pi_i} K_h(x - t) dt,$$

with $K_h(\cdot) := K(\cdot/h)/h^d$ where K is a density function on \mathbb{R}^d and $h = h_n$ is a bandwidth such that $h_n \rightarrow 0$ as $n \rightarrow \infty$.
(Muller & Prewitt, 1993).

- Kernel estimator of (non-extreme) conditional quantiles $q_Y(\alpha | x)$ for all $(x, \alpha) \in \Pi \times (0, 1)$:

$$\hat{q}_{n,Y}(\alpha | x) := \hat{F}_{n,Y}^{\leftarrow}(\alpha | x) = \inf\{y, \hat{\bar{F}}_{n,Y}(y | x) \leq \alpha\}.$$

- Estimators of position and dispersion functions:

$$\hat{a}_n(x) = \hat{q}_{n,Y}(\mu_2 | x) \quad \text{and} \quad \hat{b}_n(x) = \hat{q}_{n,Y}(\mu_3 | x) - \hat{q}_{n,Y}(\mu_1 | x)$$

for all $x \in \Pi$.

Step 2: Estimation of the tail-index

Estimation of the residuals:

$$\hat{Z}_i = (Y_i - \hat{a}_n(x_i)) / \hat{b}_n(x_i), \quad i = 1, \dots, n.$$

Due to **boundary effects associated with kernel estimators**, residuals \hat{Z}_i close to the boundary of Π are not reliable. (Kyung-Joon & Schucany, 1998).

Focus on the “interior” points:

Let $\tilde{\Pi}^{(n)} := \{x \in \mathbb{R}^d, B(x, h) \subset \Pi\}$ the erosion of Π by $B(0, h)$ and $I_n := \{i \in \{1, \dots, n\}, x_i \in \tilde{\Pi}^{(n)}\}$. We note $m_n := \text{card}(I_n)$.

Hill-type estimator of the tail-index:

$$\hat{\gamma}_n = \frac{1}{k_n} \sum_{i=0}^{k_n-1} \log \hat{Z}_{m_n-i, m_n} - \log \hat{Z}_{m_n-k_n, m_n},$$

where $\hat{Z}_{m_n-k_n, m_n} \leq \dots \leq \hat{Z}_{m_n, m_n}$ are the k_n order statistics associated with the estimated residuals $\hat{Z}_i, i \in I_n$. (Hill, 1975).

Step 3: Estimation of extreme conditional quantiles

☞ From the location-dispersion regression model,

$$\tilde{q}_{n,Y}(\alpha_n | x) = \hat{a}_n(x) + \hat{b}_n(x) \hat{q}_{n,Z}(\alpha_n),$$

where $\hat{a}_n(x)$ and $\hat{b}_n(x)$ are defined as previously and $\hat{q}_{n,Z}(\alpha_n)$ is a Weissman-type estimator of the extreme quantiles of Z :

$$\hat{q}_{n,Z}(\alpha_n) = \hat{Z}_{m_n - k_n, m_n}(\alpha_n m_n / k_n)^{-\hat{\gamma}_n},$$

(Weissman, 1978), with $\hat{\gamma}_n$ an estimator of the tail-index γ .

☞ **Remark:** Y and Z have same tail-index. Thus, γ can be estimated either from the estimated residuals \hat{Z}_i (as proposed) or from the original response variables Y_i (would yield a **high bias**, see numerical results).

Asymptotic results

We consider four assumptions.

(A.1) Model. $(Y_1, x_1), \dots, (Y_n, x_n)$ are independent observations from the above defined location-dispersion regression model for heavy-tailed distributions in the fixed design setting of (*Muller & Prewitt, 1993*):

$$\begin{aligned} \max_{i=1, \dots, n} \left| \lambda(\Pi_i) - \frac{\lambda(\Pi)}{n} \right| &= o(1/n), \\ \max_{i=1, \dots, n} \sup_{(s, t) \in \Pi_i^2} \|s - t\| &= O\left(n^{-1/d}\right), \end{aligned}$$

(A.2) Regularity conditions.

- $a(\cdot)$ and $b(\cdot)$ are twice continuously differentiable on Π ,
- $b(\cdot)$ is lower bounded on Π ,
- $\bar{F}_Z(\cdot)$ is twice continuously differentiable on \mathbb{R} .

Under **(A.2)**, the density $f_Z(\cdot)$ exists and we let $H_Z(\cdot) := 1/f_Z(q_Z(\cdot))$ the quantile density function and $U_Z(\cdot) = q_Z(1/\cdot)$ the tail quantile function of Z .

(A.3) Assumptions on the kernel. K is a bounded and even density with symmetric support $S \subset B(0, 1)$ and verifying the Lipschitz property:

$$\exists c_K > 0, \forall (u, v) \in S^2, |K(u) - K(v)| \leq c_K \|u - v\|.$$

(A.4) Second order condition. For all $t > 0$, as $z \rightarrow \infty$,

$$\frac{U_Z(tz)}{U_Z(z)} - t^\gamma \sim A(z) t^\gamma \frac{t^\rho - 1}{\rho},$$

where $\gamma > 0$, $\rho \leq 0$ and A is a positive or negative function such that $A(z) \rightarrow 0$ as $z \rightarrow \infty$. The second-order parameter ρ tunes the rate of convergence of most extreme-value estimators (*de Haan & Ferreira, 2006*).

In the following, we set

$$\kappa(d) := \begin{cases} 4 & \text{if } 1 \leq d \leq 4 \\ 2d/(d-2) & \text{if } d \geq 4. \end{cases}$$

Theorem 1 (Joint asymptotic normality of \hat{a}_n and \hat{b}_n)

Assume **(A.1)**-**(A.3)** hold and $f_Z(q_Z(\mu_j)) > 0$ for all $j \in \{1, 2, 3\}$. If $nh^d \rightarrow \infty$ and $nh^{d+\kappa(d)} \rightarrow 0$ as $n \rightarrow \infty$, then, for all sequence $(t_n) \subset \tilde{\Pi}^{(n)}$,

$$\frac{\sqrt{nh^d}}{b(t_n)} \begin{pmatrix} \hat{a}_n(t_n) - a(t_n) \\ \hat{b}_n(t_n) - b(t_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}_{\mathbb{R}^2}, \lambda(\Pi) \|\mathbf{K}\|_2^2 \Sigma),$$

where Σ is a covariance matrix depending on μ_1, μ_2, μ_3 and $H_Z(\cdot)$.

Theorem 2 (Asymptotic normality of $\hat{\gamma}_n$ and $\hat{q}_{n,Z}$)

Assume **(A.1)**-**(A.4)** hold. Let $k_n \rightarrow \infty$ be a sequence of integers such that $nh^d/(k_n \log n) \rightarrow \infty$, $nh^{d+\kappa(d)}/\log n \rightarrow 0$ and $\sqrt{k_n}A(n/k_n) \rightarrow \beta \in \mathbb{R}$ as $n \rightarrow \infty$. Then,

$$\Rightarrow \sqrt{k_n}(\hat{\gamma}_n - \gamma) \xrightarrow{d} \mathcal{N}(\beta/(1-\rho), \gamma^2),$$

\Rightarrow For all sequence $(\alpha_n) \subset (0, 1)$ such that $n\alpha_n/k_n \rightarrow 0$ and $\log(n\alpha_n)/\sqrt{k_n} \rightarrow 0$,

$$\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{n\alpha_n}\right)} \left(\log \hat{q}_{n,Z}(\alpha_n) - \log q_Z(\alpha_n) \right) \xrightarrow{d} \mathcal{N}(\beta/(1-\rho), \gamma^2).$$

If $\rho \geq -\kappa(d)/(2d)$, then the rate of convergence of $\hat{\gamma}_n$ is $n^{\rho/(1-2\rho)}$ which coincides with the usual rate of convergence for the estimation of the tail-index in the non-conditional setting.

- If $d=1$, this rate is reached for $\rho \geq -2$.
- If $d=2$, this rate is reached for $\rho \geq -1$.

Theorem 3 (Asymptotic normality of the estimator of extreme conditional quantiles)

Assume **(A.1)**-**(A.4)** hold an $f_Z(q_Z(\mu_j)) > 0$ for all $j \in \{1, 2, 3\}$. Let $k_n \rightarrow \infty$ be a sequence of integers. Suppose $nh^d / (k_n \log n) \rightarrow \infty$, $nh^{d+\kappa(d)} \rightarrow 0$ and $\sqrt{k_n}A(n/k_n) \rightarrow \beta \in \mathbb{R}$ as $n \rightarrow \infty$.

Then, for all sequences $(t_n) \subset \tilde{\Pi}^{(n)}$ and $(\alpha_n) \subset (0, 1)$ such that $n\alpha_n/k_n \rightarrow 0$ and $\log(n\alpha_n)/\sqrt{k_n} \rightarrow 0$ as $n \rightarrow \infty$,

$$\frac{\sqrt{k_n}}{\log\left(\frac{k_n}{n\alpha_n}\right)} \left(\frac{\tilde{q}_{n,Y}(\alpha_n | t_n)}{q_Y(\alpha_n | t_n)} - 1 \right) \xrightarrow{d} \mathcal{N}(\beta/(1-\rho), \gamma^2).$$

In the case of purely nonparametric estimators of extreme conditional quantiles, the term $\sqrt{k_n}$ is replaced by $\sqrt{k_n h_n^d}$ (Gardes & Girard, 2008), a consequence of **the curse of dimensionality**.

Validation on simulated data

Experimental design

- Bi-dimensional setting: $d = 2$, $\Pi = [0, 1]^2$.
- Fixed design: The x_i are chosen on a regular grid of Π .
- Norm: $\|x\| = \max(|x^{(1)}|, |x^{(2)}|)$ leading to $\tilde{\Pi}^{(n)} = [h, 1-h]^2$.
- Location and dispersion functions: $a(x) = 1 - \cos(\pi(x^{(1)} + x^{(2)}))$ and $b(x) = \exp(-(x^{(1)} - 0.5)^2 - (x^{(2)} - 0.5)^2)$.
- $\mu_1 = 3/4$, $\mu_2 = 1/2$ and $\mu_3 = 1/4$.
- Two distributions for Z : Student(ν) with $\nu \in \{1, 2, 4\}$ df ($\gamma = 1/\nu$, $\rho = -2/\nu$) and Burr(α) with $\alpha \in \{1, 2, 4\}$ as shape parameter ($\gamma = 1/\alpha$, $\rho = -1$).
- Sample size $n = 10,000$.

Implementation of the estimators

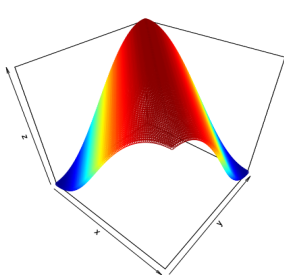
- The kernel K is the product of two univariate quartic kernels:

$$K(u, v) = \left(\frac{15}{16}\right)^2 (1 - u^2)^2 (1 - v^2)^2 \mathbb{1}_{\{|u| \leq 1\}} \mathbb{1}_{\{|v| \leq 1\}},$$

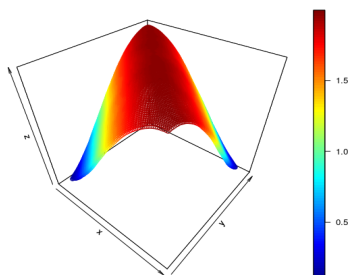
- The bandwidth $h = h_n = \sigma n^{-1/6}$, where $\sigma = 12^{-1/2}$ is the standard deviation of the coordinates of the design points (optimal choice for density estimation in the Gaussian case).
- The sequence k_n is chosen by minimizing the asymptotic mean squared error.
- Order of the extreme conditional quantile: $\alpha_n = 1/n$.
- $N = 100$ replications.

Estimation of the location function (Student distribution)

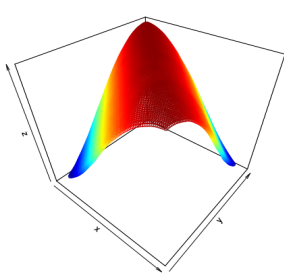
Theoretical a



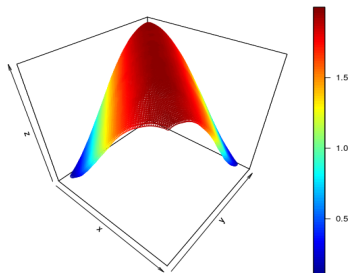
Estimated $a - 1$ df



Estimated $a - 2$ df

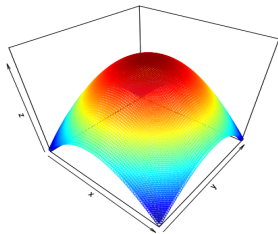


Estimated $a - 4$ df

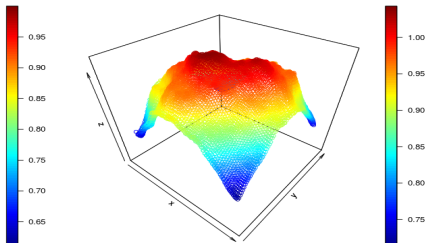


Estimation of the dispersion function (Student distribution)

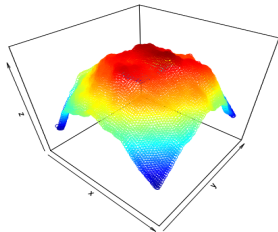
Theoretical b



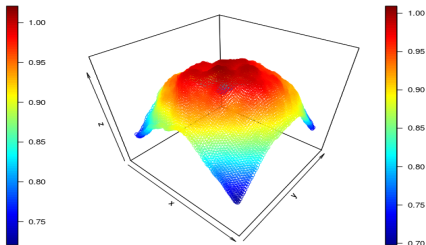
Estimated b - 1 df



Estimated b - 2 df

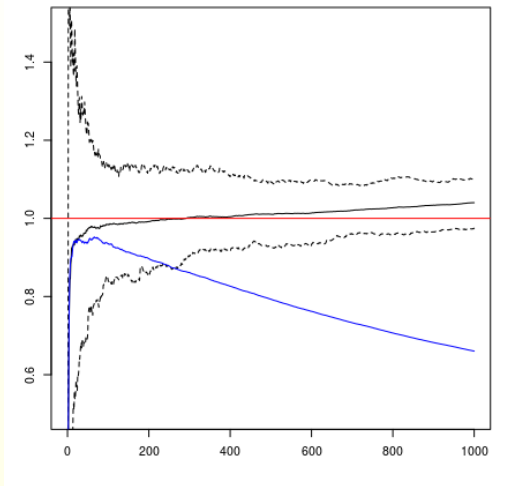


Estimated b - 4 df



Estimation of the tail-index (Student distribution)

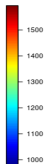
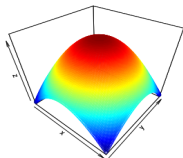
$$\nu = 1 \text{ df}$$



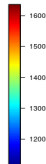
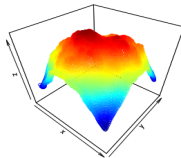
True tail-index, tail-index estimated on the residuals, tail-index estimated on original response variables (as functions of k_n).

Estimation of extreme conditional quantiles (Student distribution)

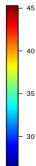
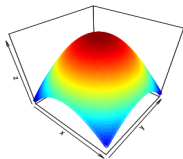
Theoretical quantile - 1 df



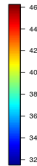
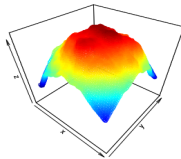
Estimated quantile - 1 df



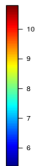
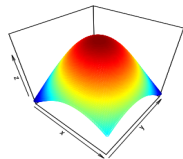
Theoretical quantile - 2 df



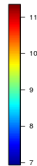
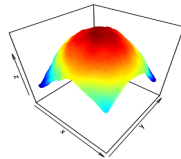
Estimated quantile - 2 df



Theoretical quantile - 4 df



Estimated quantile - 4 df



Relative MSE: comparison with two purely nonparametric estimators

n	Student, $\nu = 1$	Student, $\nu = 2$	Student, $\nu = 4$
400	0.547 (0.890, 0.976)	0.129 (0.643, 0.630)	0.062 (0.442, 0.458)
1,600	0.138 (0.867, 0.893)	0.065 (0.533, 0.458)	0.020 (0.284, 0.352)
3,600	0.145 (0.855, 0.837)	0.048 (0.477, 0.431)	0.012 (0.226, 0.306)
6,400	0.061 (0.845, 0.776)	0.032 (0.456, 0.454)	0.011 (0.206, 0.253)
10,000	0.045 (0.820, 0.723)	0.026 (0.425, 0.435)	0.013 (0.184, 0.222)
n	Burr, $\alpha = 1$	Burr, $\alpha = 2$	Burr, $\alpha = 4$
400	0.525 (0.746, 0.588)	0.197 (0.329, 0.285)	0.104 (0.129, 0.176)
1,600	0.182 (0.796, 0.637)	0.068 (0.348, 0.260)	0.038 (0.124, 0.168)
3,600	0.157 (0.825, 0.625)	0.056 (0.333, 0.264)	0.023 (0.118, 0.149)
6,400	0.096 (0.827, 0.591)	0.054 (0.311, 0.271)	0.020 (0.107, 0.122)
10,000	0.070 (0.845, 0.563)	0.030 (0.301, 0.262)	0.023 (0.102, 0.107)

- 👉 The relative MSE increases with the tail heaviness.
- 👉 Unsurprisingly, the semi-parametric estimator performs better than the nonparametric ones (*Gardes & Girard, 2008*).

Application to tsunami data

“Tsunami Causes and Waves” dataset, <https://www.kaggle.com/noaa/seismic-waves>.
Maximum wave height recorded at stations where a tsunami occurred.
We focus on the 2011 Tohoku tsunami, in Japan. This earthquake was the cause of the Fukushima nuclear disaster: A wave height $\geq 15m$ flooded the nuclear plant, protected by a seawall of only $5.7m$.

Data:

- Maximum wave height Y_1, \dots, Y_n (in m) recorded at $n = 5364$ stations. The values are ranging from 0 to $55.88m$.
- Latitudes $x_1^{(1)}, \dots, x_n^{(1)}$ and longitudes of stations: $x_1^{(2)}, \dots, x_n^{(2)}$.

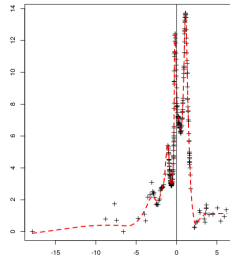
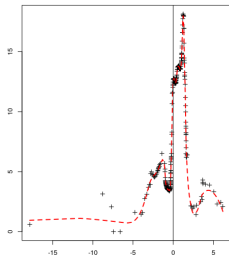
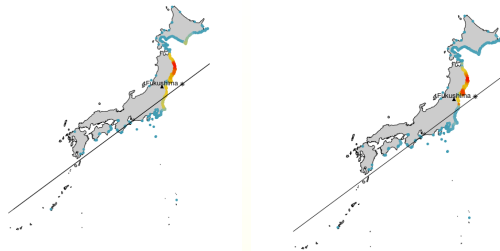
Goal: Estimation of return levels of wave heights associated with small probability.

Maximum wave height recorded at each station



2011 Tohoku tsunami, Japan, * is the epicenter.

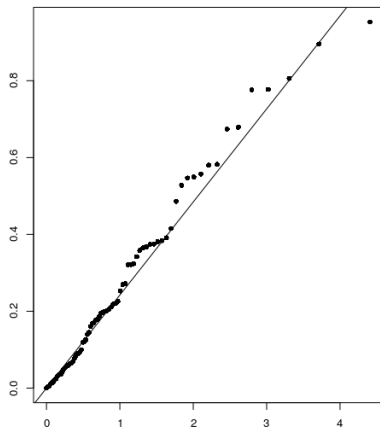
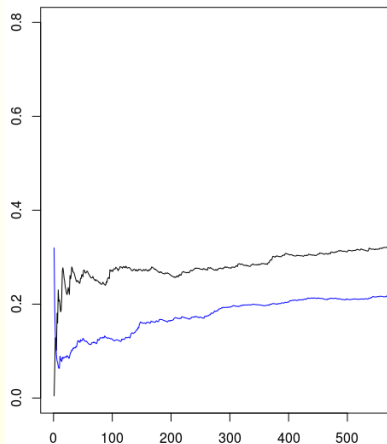
Estimation of location and dispersion functions



Left: location, Right: dispersion. Bottom: projections on the principal axis (of the station locations) depicted as a straight line on the top panel. The vertical line is the epicenter.

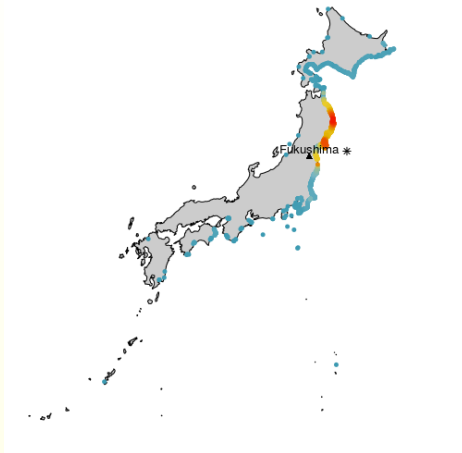
Estimation of the tail-index

Test for constant tail-index (*Einmahl, de Haan & Zhou, 2016*).



Left: Hill estimator (as a function of k_n) computed on the residuals and on the original response variable. Right: quantile-quantile plot associated with $k_n = 82$.

Estimation of the extreme conditional quantile of order $\alpha_n = 10/n$.







The estimated quantiles of the maximum wave height are ranging from 0 to 60.53m, with largest values close to the epicenter.





Further work

- Extension to random design,
- Extension to over domains of attraction (exponential tails, ...)
- Other risk measures (expected shortfall, expectiles, ...)

References I

-  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, *Electron. J. Stat.*, to appear.
-  Einmahl, J. H. J., de Haan, L. and Zhou, C. (2016). Statistics of heteroscedastic extremes. *JRSS B*, 78, 31–51.
-  Gardes, L., Girard, S., 2008. A moving window approach for nonparametric estimation of the conditional tail index. *J. Multivariate Anal.*, 99, 2368–2388.
-  de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, New York, Springer.

References II

-  Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, 3, 1163–1174.
-  Kyung-Joon, C. and Schucany, W. (1998). Nonparametric kernel regression estimation near endpoints. *J. Stat. Plan. Infer.*, 66, 289–304.
-  Müller, H. G. and Prewitt, K. (1993). Multiparameter bandwidth processes and adaptive surface smoothing. *J. Multivariate Anal.*, 47, 1–21.
-  Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *J. Am. Stat. Assoc.*, 73, 812–815.