Estimation of conditional extreme quantiles with random censoring

Jean-François DUPUY

joint work with Pathé NDAO and Aliou DIOP (Saint Louis University, Senegal)

SCHOOL "MATHEMATICAL METHODS OF STATISTICS"
Angers, 20 June 2016

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

IRMAR

Outline

1. Introduction
2. Estimation of the conditional extreme value index with censoring
3. Estimation of conditional extreme quantiles with censoring
4. Simulations
5. Perspectives

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

The framework

- statistics of extremes : estimate extreme quantiles of a random variable (r.v.) Y, which are defined as

$$\mathbb{P}(Y > q(\alpha)) = \alpha$$

with $\alpha \to 0$

- "conditional" extreme value statistics : we consider estimation of conditional extreme quantiles, defined as

$$\mathbb{P}(Y > q(\alpha, x) | X = x) = \alpha$$

with $\alpha \to 0$, where $X \in \mathbb{R}^p$ is a covariate vector (or explanatory variable) associated with Y

$\hookrightarrow$ regression setting : we are interested in just one variable (*response* variable) and we want to study how its distribution (and in particular, its conditional tail characteristics) depends on a set of variables (*explanatory* variables)

The framework

Some examples :

- magnitude of earthquakes given their location (Pisarenko et Sornette, 2003)
- amount of production of a firm given available inputs (*e.g.*, labor, capital) (Daouia *et al.*, 2010)
- analysis of extreme rainfalls given the geographical location (Gardes et Girard, 2010)
- analysis of survival of patients with HIV given their age at diagnosis (Ndao *et al.*, 2014 ; Ameraoui *et al.*, 2016)

The framework

Difficulty : estimating the survival function

$$\bar{F}(y) := 1 - F(y) = \mathbb{P}(Y > y)$$

(or conditional survival function $\bar{F}(y|x) = \mathbb{P}(Y > y|X = x)$ when covariates are present) beyond the maximum observed value $Y_{(n)} := \max(Y_1, \ldots, Y_n)$.

One cannot merely use the edf (or any version adapted to presence of covariates).

Why ? Consider a sample $Y_1, \ldots, Y_n$ of n i.i.d. r.v. and let $Y_{(1)} \le \ldots \le Y_{(n)}$ be the ordered data. Let

$$Q(p) := \inf\{y : F(y) \ge p\}$$

be the quantile function.

The framework

To estimate $F(\cdot)$, one can use the empirical distribution function

$$\hat{F}_n(y) = \frac{i}{n} \text{ if } y \in [Y_{(i)}, Y_{(i+1)}),$$

where $Y_{(i)}$ is the i-th order sample value. Usual estimate of $Q(\cdot)$ is the empirical quantile function

$$\hat{Q}_n(p) = \inf\{y : \hat{F}_n(y) \geq p\}.$$

Problems arise when considering high quantiles $Q(1-\alpha)$ with $\alpha < \frac{1}{n}$. One cannot simply assume that such values of Y are impossible.

$\Rightarrow$ these observations show that it is necessary to develop special techniques to investigate extreme quantiles of a distribution

Asymptotic distribution of the sample maximum

Theorem (Fisher-Tippett, 1928 ; Gnedenko, 1943)

Let $(Y_n) \overset{i.i.d.}{\sim} F(\cdot)$. If there exist norming sequences $(a_n > 0), (b_n)$ and some non degenerate cdf H_γ (with γ a real value) such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{Y_{(n)} - b_n}{a_n} \leq y\right) = H_\gamma(y),$$

then H_γ is of the form

$$H_\gamma(y) = \begin{cases} \exp\left(-(1+\gamma y)_+^{-1/\gamma}\right) & \text{si } \gamma \neq 0, \\ \exp(-\exp(-y)) & \text{si } \gamma = 0, \end{cases}$$

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

- $H_\gamma(\cdot)$ is known as the (generalized) extreme value distribution
- the parameter γ is called the extreme value index (EVI)

Extreme value index

According to the sign of γ, three cases can be distinguished :

- If $\gamma > 0$, $F(\cdot)$ is said to belong to Fréchet domain of attraction (DA) (or to be "of Fréchet-Pareto type" or a "heavy-tailed" distribution). Recall that Fréchet distribution has d.f. $H_\gamma(y) = \exp(-y^{-1/\gamma})$, $y > 0$.

  Roughly speaking, the survival function $\bar{F}(y) = 1 - F(y) \to 0$ at a polynomial speed, that is, as $y^{-1/\gamma}$ when $y \to \infty$.

  Example : Cauchy, Pareto, Student, F-distribution

- If $\gamma = 0$, $F(\cdot)$ is said to belong to Gumbel DA as the maxima are attracted to Gumbel d.f. $H_0(y) = \exp(-e^{-y})$ (exponential decrease of the tail of $\bar{F}$) $\Rightarrow$ "light-tailed" distributions

  Example : normal, exponential, Gamma, lognormal

Extreme value index

- If $\gamma < 0$, $F(\cdot)$ is said to belong to Weibull DA : $\bar{F}(y) = 0$ for $y > y_F$ (right end-point).

  Example : uniform, Beta

Extreme value index

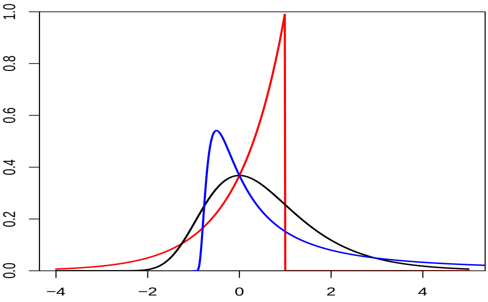

FIGURE 1 – Examples of distributions belonging to Weibull $(\gamma = -1)$, Gumbel $(\gamma = 0)$ and Fréchet $(\gamma = 1)$ domains of attraction.

$\hookrightarrow$ the EVI is closely related to the tail behaviour of a cdf. Thus, knowledge of γ is crucial for estimating extreme quantiles.

Fréchet domain of attraction ($\gamma > 0$)

The d.f. $F(\cdot)$ belongs to Fréchet DA if and only if there exists a slowly varying function $\ell(\cdot)$, that is, a function satisfying

$$\forall t > 1, \quad \lim_{y\to\infty} \frac{\ell(ty)}{\ell(y)} = 1,$$

such that $\bar{F}(y) = y^{-1/\gamma}\ell(y)$. Then :

$$\forall t > 1, \quad \lim_{y\to\infty} \frac{\bar{F}(ty)}{\bar{F}(y)} = t^{-1/\gamma} \lim_{y\to\infty} \frac{\ell(ty)}{\ell(y)} = t^{-1/\gamma},$$

and $F(\cdot)$ is said to be a regular varying function.

Remark 1

The tail becomes heavier with increasing value of γ. In other words, the dispersion is larger and large values become more likely. For this reason, Fréchet-Pareto type distributions are useful for modeling data with large outliers.

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

Conditional extreme value index

- assume that some covariate vector $X \in \mathbb{R}^p$ (with pdf g) is recorded at the same time as Y
- a natural approach to tail analysis in the presence of covariate information is to model the EVI as a function $\gamma_Y : \mathbb{R}^p \mapsto \mathbb{R}$ of the covariates :

$$x \mapsto \gamma_Y(x),$$

  which is called conditional EVI (of Y given $X = x$)
- **References** (Hill/moment estimators ; MLE under the assumption that $\gamma_Y(x) = h(x; \beta)$ for some completely specified function h and β an unknown regression parameter ; various DA ; functional covariate) :

  Gardes and Girard (2008, 2010, 2012), Daouia *et al.* (2011), Stupfler (2013), Gardes and Stupfler (2014), Goegebeur *et al.* (2014), Ndao *et al.* (2014, 2016) . . .

Conditional extreme value index

- the conditional distribution $F(\cdot|x)$ of $Y|X = x$ belongs to Fréchet DA, *i.e.* there exists a positive function $\gamma_Y(\cdot)$ of the covariate x such that :

$$\bar{F}(y|x) := 1 - F(y|x) = y^{-1/\gamma_Y(x)}\ell(y|x),$$

where $\ell(\cdot|x)$ is a slowly varying function :

$$\forall t > 1, \quad \lim_{y\to\infty} \frac{\ell(ty|x)}{\ell(y|x)} = 1.$$

- estimation of $\gamma_Y(x)$: let $(Y_i, X_i), i = 1, \ldots, n$ be independent copies of the pair (Y, X)

  Goegebeur *et al.* (2014) propose a kernel version of Hill estimator of $\gamma_Y(x)$, adapted from Hill estimator (1975) of the EVI in the univariate case.

Hill estimator of the EVI

Recall that for a heavy-tailed distribution :

$$\frac{\bar{F}(ty)}{\bar{F}(t)} \longrightarrow y^{-1/\gamma} \text{ as } t \to \infty \text{ for any } y > 1,$$

which can be interpreted as

$$\mathbb{P}(Y/t > y | Y > t) \approx y^{-1/\gamma} \text{ for } t \text{ large, } y > 1.$$

Hence, it appears natural to associate a Pareto distribution (with survival function $y^{-1/\gamma}$) to the distribution of the relative excess $E := Y/t$ over a high threshold t conditionally on $Y > t$.

Hill estimator of the EVI

Assume that we observe n i.i.d. $Y_1, \ldots, Y_n$ and let $E_i := Y_i/t$ be the i-th exceedance in the original sample, where $i = 1, \ldots, N_t$.

The log-likelihood of γ based on excesses $E_1, \ldots, E_{N_t}$ is

$$\ell(\gamma; E_1, \ldots, E_{N_t}) = -N_t \ln \gamma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{N_t} \ln E_i.$$

Solving the likelihood equation

$$0 = \frac{\partial \ell(\gamma; E_1, \ldots, E_{N_t})}{\partial \gamma} = -\frac{N_t}{\gamma} + \frac{1}{\gamma^2} \sum_{i=1}^{N_t} \ln E_i$$

yields Hill estimator of the EVI :

$$\hat{\gamma}_t^H = \frac{1}{N_t} \sum_{i=1}^{N_t} \ln E_i = \frac{\sum_{i=1}^n (\ln Y_i - \ln t) 1_{\{Y_i > t\}}}{\sum_{i=1}^n 1_{\{Y_i > t\}}}.$$

A Hill-type estimator of the conditional EVI

Goegebeur *et al.* (2014) propose :

$$\widehat{\gamma}^{H}_{t_n}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)(\ln Y_i - \ln t_n)1_{\{Y_i > t_n\}}}{\sum_{i=1}^{n} K_h(x - X_i)1_{\{Y_i > t_n\}}}$$

where

- $h := h_n$ and t_n are non-random sequences such that $h \to 0$ and $t_n \to \infty$ as $n \to \infty$,
- $K_h(x) := h^{-p} K(x/h)$ and K is a density on $\mathbb{R}^p$.

Theorem (Goegebeur *et al.*, 2014)

Under regularity conditions, $\widehat{\gamma}^{H}_{t_n}(x)$ is a consistant estimator of $\gamma_Y(x)$ and $\sqrt{nh^p \bar{F}(t_n|x)}(\widehat{\gamma}^{H}_{t_n}(x) - \gamma_Y(x))$ is asymptotically normal.

$\hookrightarrow$ no estimator of extreme quantiles is provided

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

The problem

We observe n independent triplets :

$$\mathcal{D}_n := (X_i, \delta_i, Z_i), i = 1, \ldots, n$$

where

- $Z_i = \min(Y_i, C_i)$ and C_i is a censoring r.v.,
- $\delta_i = 1_{\{Y_i \leq C_i\}}$,
- X_i is a covariate with density g on $\mathbb{R}^p$.

Objective : estimate $\gamma_Y(\cdot)$ and $q(\alpha, \cdot)$ from the sample $\mathcal{D}_n$.

The problem

We assume that

- the conditional distribution function $G(\cdot|x)$ of C given x belongs to Fréchet DA, with conditional EVI $\gamma_C(x)$
- Y and C are independent given x

$\Longrightarrow$ the conditional distribution function $H(\cdot|x)$ of Z given $X = x$ belongs to Fréchet DA and has conditional EVI

$$\gamma_Z(x) = \frac{\gamma_Y(x)\gamma_C(x)}{\gamma_Y(x) + \gamma_C(x)} = \gamma_Y(x)p_x \neq \gamma_Y(x), \quad \text{where}$$

$$p_x = \frac{\gamma_C(x)}{\gamma_Y(x) + \gamma_C(x)} = \lim_{z\to\infty} \frac{\bar{H}^1(z|x)}{\bar{H}(z|x)} = \lim_{z\to\infty} \frac{\mathbb{P}(Z > z, \delta = 1|X = x)}{\mathbb{P}(Z > z|X = x)}$$

In the literature...

Without covariates, Einmahl *et al.* (2008) propose to estimate $\gamma_Y := \gamma_Y(\cdot)$ by $\frac{\hat{\gamma}_{Z,k}}{\hat{p}_k}$, where

$$\hat{p}_k = \frac{1}{k}\sum_{j=1}^{k}\delta_{(n-j+1)}$$

and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the δ_i corresponding to $Z_{(1)}, \ldots, Z_{(n)}$.

References : Gomes and Oliveira (2003), Einmahl *et al.* (2008), Brahimi *et al.* (2013), Worms and Worms (2014)...

$\hookrightarrow$ idea is to correct for censoring by using an appropriate weight : "inverse-probability-of-censoring" method (same idea used in missing data problem)

Estimating $\gamma_Y(x)$

- recall that

$$p_x = \lim_{z\to\infty} \frac{\bar{H}^1(z|x)}{\bar{H}(z|x)} = \lim_{z\to\infty} \frac{\mathbb{P}(Z > z, \delta = 1|X = x)}{\mathbb{P}(Z > z|X = x)}$$

- we estimate respectively $\bar{H}^1(z|x)$ and $\bar{H}(z|x)$ by

$$\frac{\sum_{i=1}^{n} K_h(x - X_i) 1_{\{Z_i > z, \delta_i = 1\}}}{\sum_{i=1}^{n} K_h(x - X_i)} \quad \text{and} \quad \frac{\sum_{i=1}^{n} K_h(x - X_i) 1_{\{Z_i > z\}}}{\sum_{i=1}^{n} K_h(x - X_i)}$$

- then we construct

$$\widehat{p}_{t_n}(x) = \sum_{i=1}^{n} B_i(x) 1_{\{Z_i > t_n, \delta_i = 1\}} \Big/ \sum_{i=1}^{n} B_i(x) 1_{\{Z_i > t_n\}}$$

where $B_i(x) = K_h\left(x - X_i\right) \Big/ \sum_{j=1}^{n} K_h\left(x - X_j\right)$

Estimating $\gamma_Y(x)$

Finally, we estimate $\gamma_Y(x)$ by :

$$\boxed{\widehat{\gamma}_{t_n}^{(c,H)}(x) = \frac{\widehat{\gamma}_{t_n}^{H}(x)}{\widehat{p}_{t_n}(x)}}$$

Regularity hypothesis

- if $(x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^p$, we denote by $d(x_1, x_2)$ the distance between x_1 and x_2
- Lipschitz conditions : there exist positive constants c_γ, c_g, c_ℓ and y_0 such that

$$\begin{aligned} \left|\frac{1}{\gamma(x_1)} - \frac{1}{\gamma(x_2)}\right| &\leq c_\gamma d(x_1, x_2) \\ |g(x_1) - g(x_2)| &\leq c_g d(x_1, x_2) \\ \sup_{y \geq y_0} \left|\frac{\ln \ell(y|x_1)}{\ln y} - \frac{\ln \ell(y|x_2)}{\ln y}\right| &\leq c_\ell d(x_1, x_2) \end{aligned}$$

Asymptotics

Proposition (PN, AD & JFD, 2016)

Let (t_n) be a positive sequence such that as $n \to \infty : t_n \to \infty$, $nh^p \bar{H}(t_n|x) \to \infty$ and $nh^{p+2} \bar{H}(t_n|x)(\log t_n)^2 \to 0$. Let x be such that $g(x) > 0$. Then, as $n \to \infty$,

$$\sqrt{nh^p \bar{H}(t_n|x)}(\widehat{p}_{t_n}(x) - p_x) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{p_x(1-p_x)\|K\|_2^2}{g(x)}\right),$$

with $\|K\|_2^2 = \int K^2(u)du$.

Asymptotics

Theorem (PN, AD & JFD, 2016)

Let (t_n) be a positive sequence such that as $n \to \infty : t_n \to \infty$, $nh^p \bar{H}(t_n|x) \to \infty$ and $nh^{p+2}\bar{H}(t_n|x)(\log t_n)^2 \to 0$. Let x be such that $g(x) > 0$. Then, as $n \to \infty$,

$$\sqrt{nh^p \bar{H}(t_n|x)} \left(\widehat{\gamma}_{t_n}^{(c,H)}(x) - \gamma_Y(x) \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\gamma_Y^3(x)}{\gamma_Z(x)} \frac{\|K\|_2^2}{g(x)}\right).$$

Remark 2 (Asymptotic variance (a.v.))

- additional term $\|K\|_2^2/g(x)$, compared to the censored case without covariate (Beirlant *et al.*, 2007)
- in the absence of censoring, our a.v. reduces to the a.v. in Goegebeur *et al.* (2014)
- consistant estimator of the a.v. $\Rightarrow$ IC for $\gamma_Y(x)$

Outline of the proof

We decompose

$$\sqrt{nh^p \bar{H}(t_n|x)}\left(\widehat{\gamma}_{t_n}^{(c,H)}(x) - \gamma_Y(x)\right) = \frac{1}{p_x}\sqrt{nh^p \bar{H}(t_n|x)}\left(\widehat{\gamma}_{t_n}^{H}(x) - \gamma(x)\right) - \frac{\gamma_Y(x)}{p_x}\sqrt{nh^p \bar{H}(t_n|x)}\left(\widehat{p}_{t_n}(x) - p_x\right) + o_{\mathbb{P}}(1).$$

We prove asymptotic normality of $\mathbb{X}_n(x) :=$

$$\sqrt{\frac{nh^p}{g(x)^2 \bar{H}(t_n|x)}}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_h(x-X_i)1_{\{Z_i>t_n\}} - \bar{H}(t_n|x)g(x) \\ \frac{1}{n}\sum_{i=1}^n K_h(x-X_i)1_{\{Z_i>t_n,\delta_i=1\}} - \bar{H}^1(t_n|x)g(x) \\ \frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\log\left(\frac{Z_i}{t_n}\right)1_{\{Z_i>t_n\}} - \int_{t_n}^{\infty}\frac{\bar{H}(z|x)g(x)}{z}dz \end{pmatrix}$$

and then apply the delta-method. To prove asymptotic normality of $\mathbb{X}_n(x)$: Cramér-Wold and CLT for triangular arrays.

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

Non-censored case (fixed $\alpha \in (0,1)$)

- suppose we want to estimate the conditional quantile $q(\alpha, x)$ defined by

$$\mathbb{P}(Y > q(\alpha, x) | X = x) = \alpha$$

- kernel estimator of the conditional survival function :

$$\tilde{\bar{F}}_n(y|x) = \sum_{i=1}^{n} K_h(x - X_i) 1_{\{Y_i > y\}} \Big/ \sum_{i=1}^{n} K_h(x - X_i)$$

we consider its generalized inverse :

$$\widehat{q}_n(\alpha, x) = \tilde{\bar{F}}_n^{\leftarrow}(\alpha|x) = \inf\{y, \tilde{\bar{F}}_n(y|x) \leq \alpha\}.$$

References : Stone (1977), Stute (1986), Samanta (1989), Berlinet *et al.* (2001)

Non-censored case ($\alpha_n \to 0$ as $n \to \infty$)

- conditional extreme quantile : we want to estimate $q(\alpha_n, x)$ such that

$$\mathbb{P}(Y > q(\alpha_n, x) | X = x) = \alpha_n$$

  with $\alpha_n \to 0$ as $n \to \infty$

- generalized inverse of $\tilde{\bar{\bar{F}}}_n$:

$$\widehat{q}_n(\alpha_n, x) = \tilde{\bar{\bar{F}}}_n^{\leftarrow}(\alpha_n | x) = \inf\{y, \tilde{\bar{\bar{F}}}_n(y|x) \leq \alpha_n\}$$

- $\sqrt{nh^p\alpha_n}\left(\frac{\widehat{q}_n(\alpha_n,x)}{q(\alpha_n,x)} - 1\right)$ is asymptotically zero-mean normal, under some conditions which entail :

$$\alpha_n > \log^p(n)/n$$

  $\Rightarrow$ restriction on the order of the estimable extreme quantiles

Non-censored case : Weissman estimator

- kernel Weissman estimator : an adaptation of Weissman estimator (1978)of extreme quantiles to the conditional case

$$\hat{q}_n^W(\alpha_n, x) = \widehat{q}_n(\beta_n, x) \left(\frac{\beta_n}{\alpha_n}\right)^{\hat{\gamma}_n(x)}$$

where $\widehat{q}_n(\beta_n, x)$ is the kernel estimator of $q(\beta_n, x)$

Remark 3

The term $(\beta_n/\alpha_n)^{\hat{\gamma}_n(x)}$ is an extrapolating term which allows to estimate conditional extreme quantiles of arbitrarily small order α_n.

Censored case

- kernel Kaplan-Meier estimator (Beran, 1981)

$$\widehat{\bar{F}}_n(t|x) = \begin{cases} \prod_{i=1}^n \left[1 - \frac{B_i(x)}{\sum_{j=1}^n 1_{\{Z_j \geq Z_i\}} B_j(x)}\right]^{1_{\{Z_i \leq t, \delta_i = 1\}}} & \text{if } t \leq Z_{(n)} \\ 0 & \text{if } t > Z_{(n)} \end{cases}$$

(which reduces to $\tilde{\bar{F}}_n(t|x)$ in the absence of censoring). Its generalized inverse :

$$\widehat{q}_n^c(\alpha, x) = \widehat{\bar{F}}_n^{\leftarrow}(\alpha|x) = \inf\{t, \widehat{\bar{F}}_n(t|x) \leq \alpha\}.$$

- kernel Weissman estimator : conditional case with censoring

$$\boxed{\widehat{q}_n^{(c,W)}(\alpha, x) = \widehat{q}_n^c\left(\widehat{\bar{F}}_n(Z_{(n-k)}|x), x\right)\left(\frac{\widehat{\bar{F}}_n(Z_{(n-k)}|x)}{\alpha}\right)^{\widehat{\gamma}^{(c,H)}_{Z_{(n-k)}}(x)}}$$

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

Simulation design

- 500 samples $\{(X_i, \delta_i, Z_i), i = 1, \ldots, n\}$ of size $n = 200$, 400, 600, 800 with $Y|X = x$ distributed as Pareto with

$$\mathbb{P}(Y > y|X = x) = y^{-1/\gamma_Y(x)}$$

and

$$\gamma_Y(x) = 0.5\left(0.1 + \sin(\pi x) \times \left(1.1 - 0.5 \exp\left(-64\,(x - 0.5)^2\right)\right)\right)$$

- proportion of censored data : $10\%, 25\%, 40\%$
- **Objective** : estimate $\gamma_Y(\cdot)$ and $q(1/1000, \cdot)$ on $[0, 1]$
- kernel : $K(x) = \frac{15}{16}(1 - x^2)^2 1_{\{-1 \leq x \leq 1\}}$
- comparison with so-called "complete-case" method

Choosing the bandwidth h and threshold t_n

- we select the bandwidth h using the a cross-validation criterion (Daouia *et al.*, 2011 ; Gardes et Girard, 2012. . .) :

$$h^* := \arg\min_h \sum_{i=1}^{n} \sum_{j=1}^{n} \left(1_{\{Z_i > Z_j\}} - \widehat{\bar{\bar{F}}}_{n,-i}(Z_j | X_i) \right)^2,$$

where $\widehat{\bar{\bar{F}}}_{n,-i}$ is the kernel conditional Kaplan-Meier estimator

$$\widehat{\bar{\bar{F}}}_n(t|x) = \begin{cases} \prod_{i=1}^n \left[1 - \frac{B_i(x)}{\sum_{j=1}^n 1_{\{Z_j \geq Z_i\}} B_j(x)} \right]^{1_{\{Z_i \leq t, \delta_i = 1\}}} & \text{si } t \leq Z_{(n)} \\ 0 & \text{si } t > Z_{(n)} \end{cases}$$

(depending on h) calculated on the subsample of observations $\{(X_j, \delta_j, Z_j), 1 \leq j \leq n, j \neq i\}$,

- threshold t_n selection : we consider $t_n = Z_{(n-k)}$ and we select k as follows :

Choosing the bandwidth h and threshold t_n

1. we calculate $\widehat{\gamma}^{(c,H)}_{Z_{(n-k)}}(x)$ for $k = 1, \ldots, n-1$,
2. we form successive "blocks" of estimates $\widehat{\gamma}^{(c,H)}_{Z_{(n-k)}}(x)$ (one block for $k \in \{1, \ldots, 15\}$, a second block for $k \in \{16, \ldots, 30\}$ and so on),
3. we calculate the standard deviation of the $\widehat{\gamma}^{(c,H)}_{Z_{(n-k)}}(x)$ within each block,
4. we consider the block with minimal standard deviation and take the median value k^* of the k in the block.

Finally, we estimate $\gamma_Y(x)$ by calculating

$$\widehat{\gamma}^{(c,H)}_{t_n}(x) = \frac{\widehat{\gamma}^{H}_{t_n}(x)}{\widehat{p}_{t_n}(x)}$$

with $(h, k) = (h^*, k^*)$.

Simulation results for conditional EVI $(n = 200, 400)$

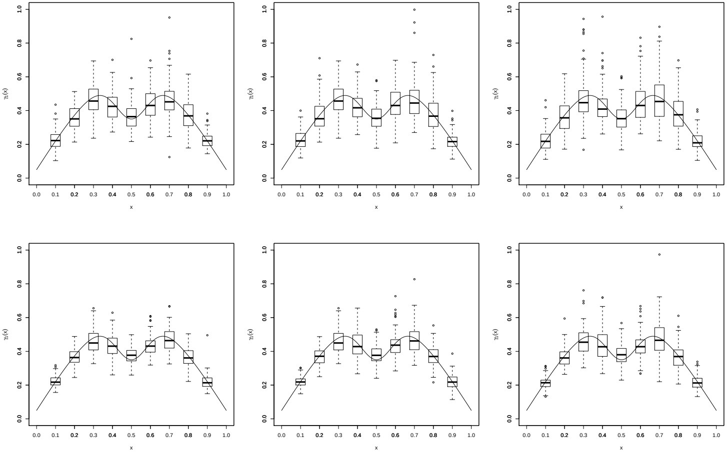

FIGURE 2 – Left : 10% censoring, middle : 25%, right : 40%.

Simulation results for conditional EVI $(n = 600, 800)$

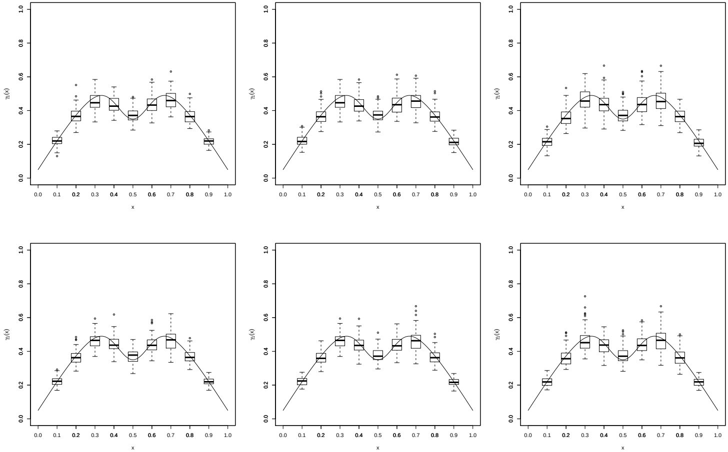

FIGURE 3 – Left : 10% censoring, middle : 25%, right : 40%.

Simulation results for conditional extreme quantiles $(n = 200, 400)$

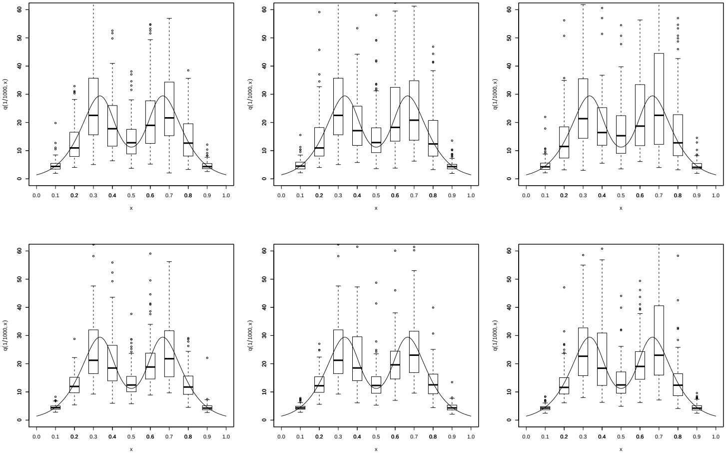

FIGURE 4 – Left : 10% censoring, middle : 25%, right : 40%.

Simulation results for conditional extreme quantiles $(n = 200, 400)$

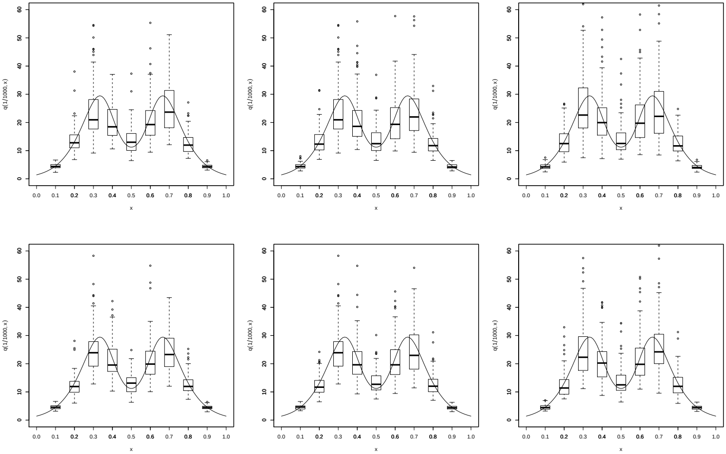

FIGURE 5 – Left : 10% censoring, middle : 25%, right : 40%.

1 Introduction
- Statistics of extremes
- Conditional extreme value index

2 Estimation of the conditional extreme value index with censoring

3 Estimation of conditional extreme quantiles with censoring

4 Simulations

5 Discussion

Discussion

- asymptotics for kernel Weissman estimator in presence of censoring $\widehat{q}_n^{(c,W)}(\alpha, x)$
- uniform results w.r.t. x
- weakening of the assumption of independent censoring

Some references

- Ameraoui A., Boukhetala K., Dupuy J.-F., 2016. Bayesian estimation of the tail index of a heavy-tailed distribution under random censoring. To appear in COMPUTATIONAL STATISTICS & DATA ANALYSIS.
- Beirlant J., Goegebeur Y., Teugels J., Segers J., 2004. Statistics of Extremes : Theory and Applications. John Wiley & Sons, Ltd.
- Daouia A., Gardes L., Girard S., Lekina A., 2011. Kernel estimators of extreme level curves. TEST 20(2), 311-333.
- Einmahl J.H.J., Fils-Villetard A., Guillou A., 2008. Statistics of extremes under random censoring. BERNOULLI 14(1), 207-227.

Some references

- Gardes L., Girard S., 2010. Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels. EXTREMES 13(2), 177-204.
- Goegebeur Y., Guillou A., Schorgen A., 2014. Nonparametric regression estimation of conditional tails : the random covariate case. STATISTICS 48(4), 732-755.
- Ndao P., Diop A., Dupuy J.-F., 2016. Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. JOURNAL OF STATISTICAL PLANNING AND INFERENCE 168, 20-37.
- Ndao P., Diop A., Dupuy J.-F., 2014. Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. COMPUTATIONAL STATISTICS & DATA ANALYSIS 79, 63-79.

Grandes lignes de la démonstration

- soit $\ell = (\ell_1, \ell_2, \ell_3)^\top \in \mathbb{R}^3$, $\ell \neq 0$. On a :

$$\ell^\top \mathbb{X}_n(x) := \sum_{i=1}^{n} T_{i,n}$$

  où pour chaque n, les $T_{1,n}, \ldots, T_{n,n}$ sont indépendants centrés. Notons $s_{n,x}^2 = \mathrm{var}(\ell^\top \mathbb{X}_n(x))$.
- condition de Lyapounov : il existe $\delta > 0$ tel que

$$\frac{1}{s_{n,x}^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}(|T_{i,n}|^{2+\delta}) \longrightarrow 0 \text{ quand } n \to \infty.$$

- Alors

$$\frac{\ell^\top \mathbb{X}_n(x)}{s_{n,x}} \xrightarrow{d} \mathcal{N}(0,1).$$