Nonparametric estimation of the conditional tail index and extreme quantiles estimation under random censoring

Pathé NDAO[1] & Aliou DIOP[1] & Jean-François DUPUY [2]

[1] *Université Gaston Berger, LERSTAD, UFR SAT, B.P. 234, Saint- Louis, Sénégal.*

[2] *INSA de Rennes, 20 Avenue des Buttes de Coesmes, CS 70839, 35708 Rennes cedex 7, France.*

Abstract.
In this talk, we consider the problem of estimation of the conditional extreme value index and extreme conditional quantiles in the presence of censoring. This work is motivated by statistical applications where the distribution functions have heavy tails : hydrology, biology, finance, telecommunications, etc. Einmahl et al. [2] considered the similar problem in the censored case without covariates. In our work, we adapt the estimators proposed in [3] to the case of right-censored observations with covariates. We establish the theoretical properties of our estimators, and we investigate their finite-sample behavior via simulations.

Keywords: random censoring, conditional tail index, conditional quantiles, Kaplan-Meier estimator.

1 Model

We consider the problems of estimation of the conditional extreme value index and extreme conditional quantiles from right-censored data. Let $(X_i)_{1\leq i\leq n}$, $(Y_i)_{1\leq i\leq n}$ and $(C_i)_{1\leq i\leq n}$ be i.i.d. copies of the random variables X, Y and C respectively, and let T, F and G be the distribution functions of X, Y and C. Let $Z_i = Y_i \wedge C_i$ and $\delta_i = \mathbb{1}_{(Y_i\leq C_i)}$, where $\mathbb{1}_{(.)}$ denotes the indicator function. We suppose Y and C are independent conditionally to X. We suppose in the sequel that only $(Z_1, \delta_1, X_1), \dots, (Z_n, \delta_n, X_n)$ are observed. The distribution function of Z will be denoted by H.
We assume that the conditional distribution functions of Y and C given x are:

$$F(y|x) = 1 - y^{-1/\gamma_1(x)} L_1(y, x),$$

$$G(y|x) = 1 - y^{-1/\gamma_2(x)} L_2(y, x),$$

where $\gamma_1(.)$ and $\gamma_2(.)$ are unknown positive functions of the covariate x and, for x fixed, $L_1(., x)$ and $L_2(., x)$ are a slowly varying functions, i.e. for $\lambda > 0$,

$$\lim_{y\to\infty} \frac{L_i(\lambda y, x)}{L_i(y, x)} = 1, i = 1, 2.$$

We have,

$$\overline{H}(.|x) = \overline{F}(.|x)\overline{G}(.|x)$$

because we suppose that Y and C are independents conditionally to X.

2 Defining the estimators

Given a sample $(Z_1, \delta_1, x_1), \ldots, (Z_n, \delta_n, x_n)$, our aim is to build a point-wise estimator of the function γ_1. More precisely, for a given x, we want to estimate $\gamma_1(x)$, focusing on the case where the design points (x_i) are nonrandom.
We use the moving window approach as in [1] and define $B(x, r)$ the ball centered at point x with radius r

$$B(x, r) = \{t, d(x, t) \leq r\}.$$

$m_{n,x} = n\phi(h_{n,x})$ is the number of observations in $(0, \infty) \times B(x, h_{n,x})$, where

$$\phi(h_{n,x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \in B(x, h_{n,x})\}}.$$

Let $\left\{Z^x_{(i)}, i = 1, \ldots, m_{n,x}\right\}$ be the observed variables Z_i's for which the associate covariates belong to the ball $B(x, h_{n,x})$. Let, $Z^x_{(1)} \leq \ldots \leq Z^x_{(m_{n,x})}$, the corresponding order statistics. If the censoring is not taken into account, then an estimator (Hill, for example) of the function $\gamma(.)$ is given by

$$\widehat{\gamma}^{(H)}_{k_x, m_{n,x}}(x) = \frac{1}{k_x}\sum_{i=1}^{k_x} i log\left(\frac{Z^x_{(m_{n,x}-i+1)}}{Z^x_{(m_{n,x}-i)}}\right)$$

To adapt this estimator to censoring, we divide this estimator by the proportion of non-censored observations in the k_x largest Z^x 's:

$$\widehat{\gamma}^{(c,H)}_{m_{n,x}, k_x}(x) = \frac{\widehat{\gamma}^{(H)}_{m_{n,x}, k_x}(x)}{\widehat{p}_x}$$

where $\widehat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x} \delta^x_{(m_{n,x}-i+1)}$ estimates $\frac{\gamma_2(x)}{\gamma_1(x)+\gamma_2(x)}$, with $\delta^x_{(1)}, \ldots, \delta^x_{(m_{n,x})}$ being the δ's corresponding to $Z^x_{(1)}, \ldots, Z^x_{(m_{n,x})}$, respectively.
And the formula is as follows:

$$\widehat{\gamma}^{(c,.)}_{k_x, m_{n,x}}(x) = \frac{\widehat{\gamma}^{(.)}_{k_x, m_{n,x}}(x)}{\widehat{p}_x}$$

where $\widehat{\gamma}^{(.)}_{k_x,m_{n,x}}(x)$ can be defined by: Hill estimator(1975), Moment estimator(1989), UH estimator(1996), etc. The overall objective is not simply an estimate of the index but also the determination of quantile. Thus, using the above notation we can estimate the corresponding quantile.The problem amounts to solve the following equation:

$$\mathbb{P}(Z > q(\alpha_{m_{n,x}}, x)|X = x) = \alpha_{m_{n,x}}$$

where $m_{n,x}$ is probability, $\alpha_{m_{n,x}} \longrightarrow 0 \quad if \quad n \longrightarrow +\infty$ (see [3])
and if we assume that the distribution function of Z is estimated by the Kaplan-Meier estimator as follows:

$$1 - \widehat{F}_{m_{n,x}}(t) = \prod_{\substack{(Z^x_{(i)} \leq t) \\ 1 \leq i \leq m_{n,x}}} \left(\frac{m_{n,x} - i}{m_{n,x} - i + 1} \right)^{\delta^x_{(i)}}$$

The quantile estimator is

$$\widehat{q}^{(c,.)}(\alpha_{m_{n,x}}, x) = Z^x_{(m_{n,x}-k_x)} \left(\frac{1 - \widehat{F}_{m_{n,x}}(Z^x_{(m_{n,x}-k_x)})}{\alpha_{m_{n,x}}} \right)^{\widehat{\gamma}^{(c,.)}_{k_x,m_{n,x}}(x)}$$

We prove the asymptotic normality of the proposed estimators of the conditional tail index and extrem quantiles.

Bibliographie

[1] Gardes, L. and Girard, S. A moving window approach for nonparametric estimation of the conditional tail index. Journal of Multivariate Analysis, 99:2368-2388, 2008.
[2] John H.J. Einmahl and Amelie Fils-Villetard, A. and Armelle Guillou. Statistics of extremes under random censoring. Bernoulli, 14:207-227, 2008.
[3] Lekina, A. Estimation non-parametrique des quantiles extrêmes conditionnels. Thèse de doctorat, Université de Grenoble, 2010.