Biometrical Letters
Vol. 48 (2011), No. 1, 1-22

Estimation of the extreme value index for randomly censored data

M. Ivette Gomes[1], **M. Manuela Neves**[2]

[1]Universidade de Lisboa, DEIO, CEAUL and FCUL, e-mail: ivette.gomes@fc.ul.pt
[2]Universidade Técnica de Lisboa, Instituto Superior de Agronomia and CEAUL,
Portugal, e-mail: manela@isa.utl.pt

Summary

In the field of Statistics of Extremes, the most common assumption on any set of univariate data is to consider them as a complete sample of either independent and identically distributed or weakly dependent and stationary observations, from an unknown distribution function F. However, in the analysis of lifetime data, observations are usually randomly censored. We assume the case of random censorship, and dedicate our attention to the estimation of the extreme value index (EVI), the primordial parameter of extreme events. Such a parameter measures the heaviness of the right tail and its estimation has been widely studied in the literature, for complete samples. If we are under a random censoring scheme, any of the common EVI estimators needs to be slightly modified in order to be consistent. We pay special attention to such estimation, making use of an adequate set of semi-parametric EVI estimators, among which we select second-order reduced-bias estimators. The performance of those estimators is illustrated through the use of Monte Carlo simulations and the application of the methodology to a few sets of survival data, available in the literature.

Key words: extreme value index estimation, censoring schemes.

1. Introduction, motivation and scope of the paper

In *Statistics of Extremes* we deal essentially with the estimation of parameters of extreme or even rare events. The most common assumption on any set of univariate data, $(X_1, X_2, \ldots, X_n)$, is to consider them as a *complete* sample of size n, with observations either independent and identically distributed (i.i.d.) or weakly dependent and stationary, from an unknown distribution function (d.f.) $F \equiv F_X$. There is a large variety of parameters of extreme events, but in all applications of extreme value theory (EVT), the estimation of the *extreme value index* (EVI), usually denoted γ, is of primordial importance and is the basis for the estimation of all other parameters of extreme events. Among the most relevant parameters of extreme events, and assuming that we are interested in large values, i.e., in the right tail of the underlying model F, we mention:

- the *probability of exceedance* of a high level $x \equiv x_H$, $p_x := \mathbb{P}(X > x) = 1 - F(x) =: \overline{F}(x)$,
- the *return period* of a high level x, which is given by $r_x := 1/(1-F(x))$ in any i.i.d. scheme,
- the *right endpoint* of an underlying model F, $x^* \equiv x^F := \sup\{x : F(x) < 1\}$, and
- a *high quantile* of probability $1 - p$, p small, situated in the border or even beyond the range of the available data, defined as $\chi_{1-p} := \inf\{x : F(x) \geqslant 1 - p\} =: F^{\leftarrow}(1 - p),\ p < 1/n$.

However in many real situations, censored observations can occur. For example, and among other cases, censored observations appear

- in the analysis of lifetime data or reliability data and
- in the analysis of some physical phenomena such as wind speeds, earthquake intensities or floods, where extreme measurements are sometimes not available due to damage to the instruments.

We shall pay special attention here to the estimation of the extreme value index γ_X under *random censorship*, where apart from two recent papers by Einmahl *et al.* (2008) and Gomes and Neves (2010), there is only, as far as we know, a brief reference to the topic in Reiss and Thomas (1997, Section 6.1) and a paper by Beirlant *et al.* (2007).

In Section 2 of this paper, we provide a few details on the EVI and max-domains of attraction. Next, in Section 3, we introduce a set of semi-parametric EVI estimators, valid for complete samples, the *Hill* (Hill, 1975), the *moment* (Dekkers *et al.*, 1989), the *generalized Hill* (Beirlant *et al.*, 1996), a *minimum-variance reduced-bias* (MVRB) *Hill* (Caeiro *et al.*, 2005) and the *mixed moment* (Fraga Alves *et al.*, 2009) EVI-estimators, providing some details on their asymptotic non-degenerate behavior. In Section 4, we illustrate the effect of random censorship on the EVI of the potential, non-available sample $\mathbf{X} = (X_1, \ldots, X_n)$. This is done in order to motivate the functional expression of the EVI estimators for randomly censored data, provided in Section 4.2. We shall now anticipate some of the recommendations. For heavy tails, i.e., for a positive EVI, we strongly advise the use of any of the recent MVRB EVI-estimators introduced and studied in Caeiro *et al.* (2005) and Gomes *et al.* (2007, 2008b). For a general tail, and particularly if we have a clear indication that the right tail is light, we suggest the use of second-order reduced-bias EVI-estimators associated with either the generalized Hill or the mixed moment EVI-estimators. This kind of estimators has not yet been considered in the literature, essentially due to the difficulties of estimating second-order parameters or functionals for a general tail, a topic that deserves further attention, but is outside the scope of this paper. In Section 5, we present the results of a small-scale Monte Carlo simulation devised to obtain the behavior of the EVI-estimators in Sections 3 and 4.2. In Section 6, we illustrate the behavior of the same EVI-estimators for a few sets of survival data, available in the literature, providing some further hints for adequate EVI-estimation. Finally, in Section 7, we provide some concluding remarks and mention a few items of future research in the topic.

2. Max-domains of attraction and the EVI

For large values, the EVI measures essentially the heaviness of the right tail $\overline{F} = 1 - F$ of an underlying model F. It is in fact the real parameter γ in the extreme value (EV) d.f.,

$$EV_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), \ 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0 \\ \exp(-\exp(-x)) & \text{if} \quad \gamma = 0. \end{cases} \tag{1}$$

This d.f. appears as the limiting distribution of the sequence of maximum values, linearly normalized, whenever such a non-degenerate limit does exist. We then say that F is in the *domain of attraction* (for maxima) of EV_γ, and write $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$.

As mentioned before, and illustrated in Figure 1, the *extreme value index* γ measures essentially the weight of the right tail $\overline{F}$.

- If $\gamma < 0$, the right tail is light, and F has a finite *right endpoint* $(x^* < +\infty)$;

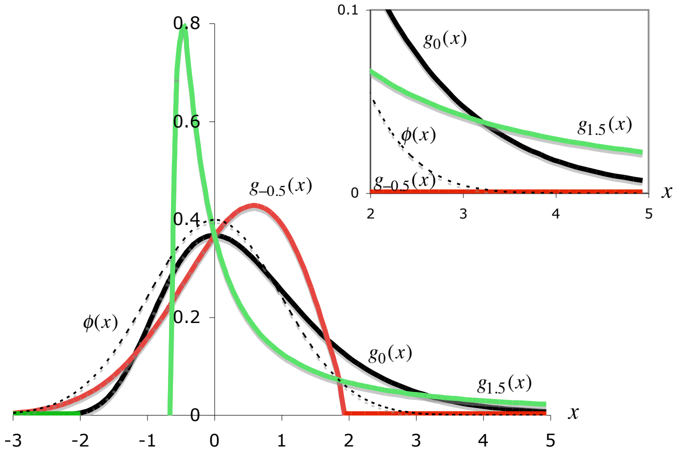

Figure 1. P.d.f. $g_\gamma(x) = dEV_\gamma(x)/dx$, for $\gamma = -0.5$, $\gamma = 0$ and $\gamma = 1.5$, together with the normal p.d.f., $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi},\ x \in \mathbb{R}$.

- If $\gamma > 0$, the right tail is heavy, of negative polynomial type, and F has an infinite *right endpoint*;

- If $\gamma = 0$, the right tail is of exponential type. The *right endpoint* can then be either finite or infinite.

2.1. First and second-order conditions in EVT

The following *extended regular variation* property (de Haan, 1984) is a well-known necessary and sufficient condition for $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$:

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{if} \quad \gamma \neq 0 \\ \ln x & \text{if} \quad \gamma = 0, \end{cases} \tag{2}$$

for every $x > 0$ and some positive measurable function a, with U standing for a quantile-type function associated with F and defined by

$$U(t) := \Big(\frac{1}{1-F}\Big)^{\leftarrow}(t) = \inf\Big\{x : F(x) \geqslant 1 - \tfrac{1}{t}\Big\}, \quad t \geqslant 1.$$

Apart from the first-order condition, in (2), we often need a second-order condition, specifying the rate of convergence in the first-order condition. More restrictively than $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$, we then assume the existence of a function A, possibly not changing in sign, but necessarily tending to zero, as $t \to \infty$, such that

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^\gamma-1}{\gamma}}{A(t)} = H_{\gamma,\rho}(x) := \frac{1}{\rho}\Big(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^\gamma-1}{\gamma}\Big), \qquad (3)$$

for all $x > 0$, where $\rho \leqslant 0$ is a *second-order* parameter controlling the speed of convergence of maximum values, linearly normalized, towards the extreme value limit law, in (1). Then $|A|$ is regularly varying with an index of regular variation equal to ρ (see Bingham *et al.*, 1997, for details on regular variation). Even slightly more restrictively, we often assume that $\rho < 0$. We can then choose $A(t) = \gamma\beta t^\rho$ in the second-order condition in (3).

3. Semi-parametric EVI estimation for complete data

For complete samples, now denoted $\mathbf{Z} = (Z_1, \ldots, Z_n)$, and for a general EVI estimation, i.e. for $\gamma \equiv \gamma_Z \in \mathbb{R}$, we shall mention the *moment* (Dekkers *et al.*, 1989), the *generalized Hill* (Beirlant *et al.*, 1996) and the *mixed moment* (Fraga Alves *et al.*, 2009) estimators. For $j \geqslant 1$, $1 \leqslant k < n$, and with $Z_{i:n}$, $1 \leqslant i \leqslant n$ denoting, as usual, the set of ascending order statistics (o.s.'s) associated with the sample $(Z_1, Z_2, \ldots, Z_n)$, let us denote

$$L_{k,n}^{(j)} \equiv L_{k,n}^{(j)}(\mathbf{Z}) := \frac{1}{k}\sum_{i=1}^{k} \{1 - Z_{n-k:n}/Z_{n-i+1;n}\}^j$$

and

$$M_{k,n}^{(j)} \equiv M_{k,n}^{(j)}(\mathbf{Z}) := \frac{1}{k}\sum_{i=1}^{k} \{\ln Z_{n-i+1:n} - \ln Z_{n-k:n}\}^j \,.$$

The *moment* (M) estimator is given by

$$M = M_k \equiv \widehat{\gamma}_{k,n}^{M}(\mathbf{Z}) := M_{k,n}^{(1)} + \tfrac{1}{2}\Big\{1 - \Big(M_{k,n}^{(2)}/[M_{k,n}^{(1)}]^2 - 1\Big)^{-1}\Big\}, \qquad (4)$$

and the *mixed moment* (MM) estimator has the functional form,

$$\begin{aligned} MM = MM_k \equiv \widehat{\gamma}_{k,n}^{MM}(\mathbf{Z}) &:= \tfrac{\widehat{\varphi}_{k,n}-1}{1+2\min\left(\widehat{\varphi}_{k,n}-1,0\right)}, \\ \widehat{\varphi}_{k,n} &:= \tfrac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{\left(L_{k,n}^{(1)}\right)^2}. \end{aligned} \qquad (5)$$

For heavy tails, i.e. whenever $\gamma \equiv \gamma_{_Z} > 0$, we mention the classical *Hill* estimator (Hill, 1975), with the functional expression

$$H = H_k \equiv \widehat{\gamma}^{H}_{k,n}(\mathbf{Z}) := \frac{1}{k}\sum_{i=1}^{k} \ln Z_{n-i+1:n} - \ln Z_{n-k:n} \equiv M^{(1)}_{k,n}(\mathbf{Z}), \qquad (6)$$

$1 \leqslant k < n$, and one of the most recent *minimum-variance reduced-bias* (MVRB) estimators of the extreme value index (Caeiro *et al.*, 2005), given by

$$\overline{H} = \overline{H}_k \equiv \widehat{\gamma}^{\overline{H}}_{k,n}(\mathbf{Z}) := H_k \left(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})\right), \qquad (7)$$

with $(\hat{\beta}, \hat{\rho})$ adequate consistent estimators of the vector (β, ρ) of second order parameters, involved in the function $A(t) = \gamma\beta t^{\rho}$, in (3) (see Gomes and Pestana, 2007, for an algorithm related to such estimation and application to the estimation of high quantiles).

The *generalized Hill* (GH) estimator is also defined for $k = 2, \ldots, n-1$, and is a generalization of the Hill estimator, in (6), being valid for a general $\gamma \equiv \gamma_{_Z}$, like the estimators in (4) and (5). It is given by

$$\begin{aligned} &GH = GH_k \equiv \widehat{\gamma}^{GH}_{k,n}(\mathbf{Z}) := \tfrac{1}{k}\textstyle\sum_{j=1}^{k} \ln UH_{j,n} - \ln UH_{k,n}, \\ &UH_{j,n} := Z_{n-j:n} H_j, \ 1 \leqslant j \leqslant k, \end{aligned} \qquad (8)$$

with H_k defined in (6). To enhance the similarity between the M estimator, in (4), and the GH estimator, in (8), we can also write an asymptotically equivalent expression for GH_k, given by

$$GH^{*}_{k} := H_k + \frac{1}{k}\sum_{i=1}^{k}\{\ln H_i - \ln H_k\}.$$

In all these papers the available sample is complete, and for the abovementioned estimators, as well as for other EVI estimators, we can prove consistency, i.e. convergence in probability to $\gamma \equiv \gamma_{_Z}$ in the domain of attraction where they are valid, for any intermediate k, i.e. whenever $k = k_n \to \infty$ and $\frac{k}{n} \to 0$, as $n \to \infty$. Under the additional validity of the second-order condition in (3), we can guarantee the asymptotic normality of $\hat{\gamma}^{\bullet}_{k,n}$, i.e., for any $k = 2, \ldots, n-1$, there exist a standard normal random variable (r.v.) $P^{\bullet}_{k}$ and real functions $\sigma^{\bullet} = \sigma^{\bullet}(\gamma)$ and $b^{\bullet} = b^{\bullet}(\gamma, \rho)$ such that

$$\widehat{\gamma}^{\bullet}_{k,n} \overset{d}{=} \gamma + \frac{\gamma\sigma^{\bullet}P^{\bullet}_{k}}{\sqrt{k}} + b^{\bullet}A(n/k) + o_p(A(n/k)).$$

For the MVRB corrected-Hill estimator $\overline{H}$, in (7), $b^{\overline{H}} \equiv 0$ and $\sigma^{\overline{H}} = \sigma^H = \gamma$ for all $\gamma > 0$, i.e. $\overline{H}_k$ outperforms H_k for all k.

4. Random censoring and EVI estimation

Let us now assume that, with $F_X \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_1})$, $\gamma_1 \equiv \gamma_X$, we are under a framework of random censorship, i.e., there is a r.v. Y such that $F_Y \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_2})$ and only $Z = X \wedge Y$ and $\delta = I_{\{X \leqslant Y\}}$ are observed. The indicator variable δ determines whether X has been censored or not. Let us denote $\gamma \equiv \gamma_Z$, the EVI associated with Z, i.e., let us assume that $F_Z \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$. We thus have access to the random complete sample (Z_i, δ_i), $1 \leqslant i \leqslant n$, of independent copies of (Z, δ), but our goal is to make inference on the right tail of the unknown lifetime distribution of X, i.e. on $\overline{F}_X(x) := \mathbb{P}(X > x) = 1 - F_X(x)$, while F_Y, the d.f. of Y, is considered to be a nonparametric nuisance parameter. As mentioned in Einmahl *et al.* (2008), all the EVI estimators need to be slightly modified in order to be consistent for the estimation of $\gamma_1 \equiv \gamma_X$ in the whole domain of attraction $\mathcal{D}_{\mathcal{M}}(EV_{\gamma_X})$, and, with τ_X and τ_Y denoting the right endpoints of F_X and F_Y, respectively, the cases of interest are:

- Case 1: $\gamma_1 > 0$, $\gamma_2 > 0$.
- Case 2: $\gamma_1 < 0$, $\gamma_2 < 0$ and $\tau_X = \tau_Y$.
- Case 3: $\gamma_1 = 0$, $\gamma_2 = 0$ and $\tau_X = \tau_Y = +\infty$.

In all the above mentioned cases, we have $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$, with the notation $\gamma = 0$, in Case 3.

4.1. A few naive comments on random censoring

As mentioned above, let us consider that instead of a potential sample $(X_1, X_2, \ldots, X_n)$ (non-observed), and assuming Y independent of X, we observe (Z_i, δ_i), $\;1 \leqslant i \leqslant n$, with $Z = \min(X, Y)$, $\delta = I_{\{X \leqslant Y\}}$. Let us assume that $F_X \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_1})$, with $\gamma_1 \equiv \gamma_X$ the parameter we really want to estimate, and that $F_Y \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_2})$. To motivate how to estimate $\gamma_1 \equiv \gamma_X$, let us begin with some examples.

Example 1. We shall first simplify the problem, assuming that we are in Case 1 and X and Y are Pareto(γ_1) and Pareto(γ_2), respectively, i.e.

for all $x \geqslant 1$, $F_X(x) = 1 - x^{-1/\gamma_1}$ and $F_Y(x) = 1 - x^{-1/\gamma_2}$, $\gamma_1, \gamma_2 > 0$. Consequently,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(\min(X,Y) \leqslant z) = 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= 1 - z^{-1/\gamma_1} z^{-1/\gamma_2} = 1 - z^{-\frac{\gamma_1+\gamma_2}{\gamma_1\gamma_2}}, \end{aligned}$$

i.e. $Z \sim$ Pareto$(\gamma_1\gamma_2/(\gamma_1+\gamma_2))$, and we can use the available semi-parametric estimators in Section 3 to estimate $\gamma = \gamma_1\gamma_2/(\gamma_1+\gamma_2)$.

On the other hand, with $f_\bullet(\cdot)$ denoting the probability density function associated with $F_\bullet(\cdot)$,

$$\begin{aligned} p \equiv p_z &:= \mathbb{P}(X \leqslant Y | Z = z) = \mathbb{P}(\delta = 1 | Z = z) = \frac{f_X(z)(1 - F_Y(z))}{f_Z(z)} \\ &= \frac{\frac{1}{\gamma_1} z^{-1/\gamma_1 - 1}\ z^{-1/\gamma_2}}{(\frac{1}{\gamma_1} + \frac{1}{\gamma_2}) z^{-1/\gamma_1 - 1/\gamma_2 - 1}} = \frac{\gamma_2}{\gamma_1 + \gamma_2}. \end{aligned}$$

Consequently, the quotient between any estimator of $\gamma \equiv \gamma_Z$ and an estimator of $p \equiv p_Z$ will provide an estimator of $\gamma_X \equiv \gamma_1$, the parameter of interest.

Example 2. Let us assume now that we are in Case 2, and that X and Y are both generalized Pareto (GP) r.v.'s, with shape parameters $\gamma_1 < 0$, $\gamma_2 < 0$, respectively, and the same right endpoint. We can then consider, without loss of generality, $F_X(x) = 1 - (1 + \gamma_1 x)^{-1/\gamma_1}$, $0 \leqslant x < -1/\gamma_1$ and $F_Y(x) = 1 - (1 + \gamma_1 x)^{-1/\gamma_2}$, $0 \leqslant x < -1/\gamma_1$. Then

$$\begin{aligned} F_Z(z) &= 1 - (1 + \gamma_1 z)^{-1/\gamma_1}(1 + \gamma_1 z)^{-1/\gamma_2} = 1 - (1 + \gamma_1 z)^{-(\gamma_1+\gamma_2)/(\gamma_1\gamma_2)}, \\ f_Z(z) &= \frac{\gamma_1 + \gamma_2}{\gamma_2}(1 + \gamma_1 z)^{-(\gamma_1+\gamma_2)/(\gamma_1\gamma_2) - 1}, \end{aligned}$$

and

$$\begin{aligned} p \equiv p_z &:= \mathbb{P}(X \leqslant Y | Z = z) \\ &= \frac{\gamma_1(1 + \gamma_1 z)^{-1/\gamma_1 - 1}(1 + \gamma_1 z)^{-1/\gamma_2 - 1}}{(\gamma_1 + \gamma_2)(1 + \gamma_1 z)^{-(\gamma_1+\gamma_2)/(\gamma_1\gamma_2) - 1}} = \frac{\gamma_2}{\gamma_1 + \gamma_2}, \end{aligned}$$

as happens for Pareto heavy tailed models.

Example 3. For heavy tails again, but for the GP models in Example 2, let $F_X(x) = 1 - (1 + \gamma_1 x)^{-1/\gamma_1}$ and $F_Y(x) = 1 - (1 + \gamma_2 x)^{-1/\gamma_2}$, for all $x \geqslant 0$, $\gamma_1, \gamma_2 > 0$. We then get

$$\begin{aligned} F_Z(z) &= 1 - (1 + \gamma_1 z)^{-1/\gamma_1}(1 + \gamma_2 z)^{-1/\gamma_2}, \\ f_Z(z) &= (1 + \gamma_1 z)^{-1/\gamma_1 - 1}(1 + \gamma_2 z)^{-1/\gamma_2}\left(1 + \frac{1 + \gamma_1 z}{1 + \gamma_2 z}\right), \end{aligned}$$

and

$$\begin{aligned} p \equiv p_z &:= \mathbb{P}(X \leqslant Y | Z = z) \\ &= \frac{(1+\gamma_1 z)^{-1/\gamma_1 - 1}(1+\gamma_2 z)^{-1/\gamma_2}}{(1+\gamma_1 z)^{-1/\gamma_1 - 1}(1+\gamma_2 z)^{-1/\gamma_2}(1+(1+\gamma_1)/(1+\gamma_2 z)} \\ &= \frac{1}{1+\frac{1+\gamma_1 z}{1+\gamma_2 z}}, \end{aligned}$$

no longer equal to $\gamma_2/(\gamma_1+\gamma_2)$. But

$$p_z = \frac{1}{1+\frac{1+\gamma_1 z}{1+\gamma_2 z}} \underset{z\to\infty}{\longrightarrow} \frac{\gamma_2}{\gamma_1+\gamma_2}. \tag{9}$$

Indeed, much more generally, and for all the above-mentioned Cases 1-2-3, $F_X \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_1}),\ F_Y \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_2}) \implies F_{Z=\min(X,Y)} \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$ with $\gamma = \gamma_1\gamma_2/(\gamma_1+\gamma_2)$.

Also, the above-mentioned limiting result on $p_z = \mathbb{P}(X \leqslant Y|Z=z)$, provided in (9), holds generally, for any $F_X \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_1})$, $F_Y \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma_2})$ and the Cases 1-2-3 (see Einmahl *et al.*, 2008). Consequently, any functional based on the $k+1$ top o.s.'s in the observed sample $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$, devised for the estimation of the EVI in complete samples, and generically denoted $\widehat{\gamma}^{\bullet}_{k,n}(\mathbf{Z})$, will converge towards $\gamma_1\gamma_2/(\gamma_1+\gamma_2)$ for intermediate k, i.e. whenever

$$k = k_n \to \infty \quad \text{and} \quad \frac{k}{n} \to 0, \text{ as } n \to \infty.$$

Moreover, and for the same intermediate k-sequences,

$$\frac{\widehat{\gamma}^{\bullet}_{k,n,Z}}{p} \underset{n\to\infty}{\overset{p}{\longrightarrow}} \gamma_1 \equiv \gamma_X. \tag{10}$$

4.2. Semi-parametric EVI estimation for randomly censored data

On the basis of (10), if we want to estimate the EVI in randomly censored samples, we merely need to find an estimator of $p \equiv p_Z$, where $1-p$ is the *percentage of censoring* in the *right tail* of F_X. A possible, and the most simple, semi-parametric consistent estimator of p has been provided in Einmahl *et al.* (2008), and is merely given by

$$\hat{p}^C = \hat{p}^C_k \equiv \hat{p}^C_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1]}, \tag{11}$$

where $\delta_{[n-i+1]}$, $1 \leqslant i \leqslant n$, are the induced or concomitant o.s.'s associated with the ascending ordering of $(Z_1, Z_2, \ldots, Z_n)$, i.e. $\delta_{[n-i+1]}$ is the concomitant value of δ associated with $Z_{n-i+1:n}$, $1 \leqslant i \leqslant n$. An obvious semi-parametric estimator of γ_X, based on the observed (Z_i, δ_i), $1 \leqslant i \leqslant n$ is thus

$$\widehat{\gamma}^{\bullet}_{k,n}(\mathbf{X}) := \frac{k\ \widehat{\gamma}^{\bullet}_{k,n}(\mathbf{Z})}{\sum_{i=1}^{k} \delta_{[n-i+1]}}.$$

As counterparts, for randomly censored data $\mathbf{X}$, of the estimators in Section 3, we shall now consider such simple modifications. With T denoting any of the estimators M, MM, H, $\overline{H}$ and GH, in (4), (5), (7) and (8), respectively, we shall consider the estimators

$$T^C \equiv T^C_k \equiv \widehat{\gamma}^T_{k,n}(\mathbf{X}) := \widehat{\gamma}^T_{k,n}(\mathbf{Z})/\widehat{p}^C, \tag{12}$$

with $\widehat{p}^C$ as given in (11). We shall study here the behavior of these estimators through Monte Carlo simulation, and shall use them for the EVI estimation associated with a few sets of survival data, available in the literature, providing some hints for adequate EVI estimation.

5. A small-scale Monte Carlo simulation

For heavy right–tails, we have considered $X \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_1})$, with $\gamma_1 = 0.25$ censored by $Y \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_2})$, with $\gamma_2 = \gamma_1 p/(1-p)$ (> 0) for $p = 0.35(0.10)0.95$. This means that we allow a percentage of censoring in the right tail ranging from 65% to 5%. We then get $Z \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$, with $\gamma = \gamma_1\gamma_2/(\gamma_1+\gamma_2)$, again > 0. For light right-tails, we considered $X \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_1})$, with $\gamma_1 = -0.25$ censored by $Y \in \mathcal{D}_{\mathcal{M}}(G_{\gamma_2})$, with $\gamma_2 = \gamma_1 p/(1-p)$ (< 0), again for $p = 0.35(0.10)0.95$. As before, we get $Z \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$, with $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$ (< 0). In the following table we present the values of $|\gamma_2|$ and $|\gamma|$ associated with the different values of p:

p	0.95	0.85	0.75	0.65	0.55	0.45	0.35
$\lvert\gamma_2\rvert$	4.7500	1.4167	0.7500	0.4643	0.3056	0.2045	0.1346
$\lvert\gamma\rvert$	0.2375	0.2125	0.1875	0.1625	0.1375	0.1125	0.0875

We performed Monte Carlo simulations, based on 1000 runs, for the following parents:

- X and Y are both *Fréchet* r.v.'s, with heavy right-tails. The Fréchet(γ) d.f. is given by $F(x) = \exp(-x^{-1/\gamma})$, $x \geqslant 0$, $\gamma > 0$;

- X and Y are both *Reversed-Burr* (RB) r.v.'s, with light right-tails. The RB$(\gamma, \beta, \lambda, x^*)$ d.f. is $F(x) = 1-\big(\beta/\big(\beta+(x^*-x)^{1/(\lambda\gamma)}\big)\big)^\lambda$, $x \leqslant x^*$.

5.1. Fréchet underlying parents

In Figure 2 we picture, for a sample size $n = 1000$, the mean values of the different estimators under analysis, when we consider $X \sim$ Fréchet$(\gamma_1 = 0.25)$ censored by $Y \sim$ Fréchet$(\gamma_2 = 0.75)$. We thus have $p = 0.75$, i.e., we have a censoring in the right tail equal to 25%, and we get for the observed Z an extreme value index $\gamma = \gamma_Z = 0.1875$.

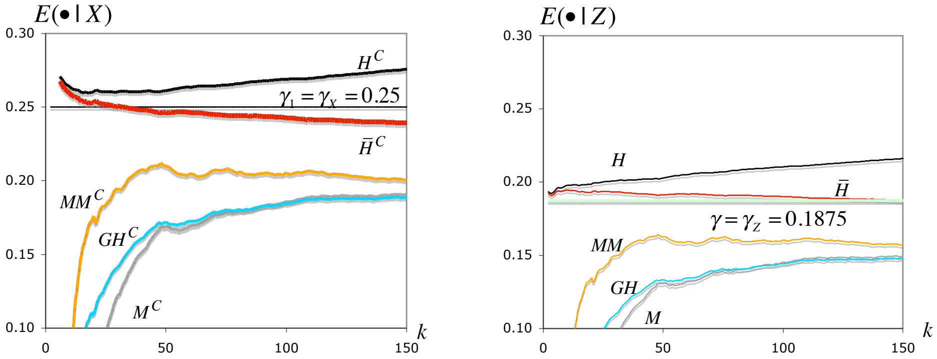

Figure 2. Mean values of the Hill, bias-corrected Hill, moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = 0.25$ (*left*) and of $\gamma = \gamma_Z = 0.1875$ (*right*), based on observed Z-samples of size $n = 1000$ of $X \sim$ Fréchet$(\gamma_1 = 0.25)$ censored by $Y \sim$ Fréchet$(\gamma_2 = 0.75)$ — $p = 0.75$ (25% censoring in the right tail).

In Figure 3 we picture, at the left, the root mean squared errors (RMSEs) of the estimators in Figure 2, and at the right, the mean values and RMSEs of the estimators, in (11), for the estimation of the known value $p = 0.75$.

If the censoring in the right tail is not too heavy, let us say smaller than or equal to 25%, the results obtained for a sample size as small as $n = 100$ are still interesting, as can be seen in Figure 4 and Figure 5, equivalent to Figure 2 and Figure 3 respectively, with the same percentage of censoring in the right tail as before, but for $n = 100$.

For the same *Fréchet* model as before, but with a higher percentage of censoring in the right tail, equal to 55%, we next present Figures 6, 7, 8 and 9, similar to Figures 2, 3, 4 and 5 respectively, but for $p = 0.45$.

We have now to pay special attention to the underestimation associated with the M, GH and MM EVI-estimators, whenever n is small. Such an underestimation leads to M and GH negative estimates of γ, for all k.

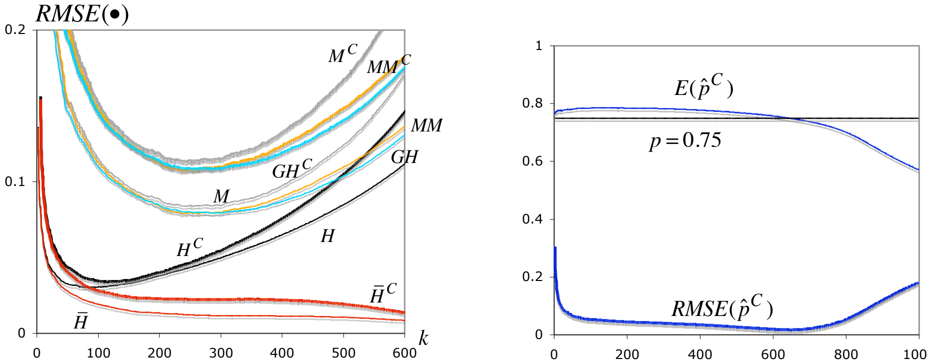

Figure 3. RMSEs of the estimators in Figure 2 (*left*) and mean value (E) and RMSE of the estimator of $p = 0.75$ (*right*).

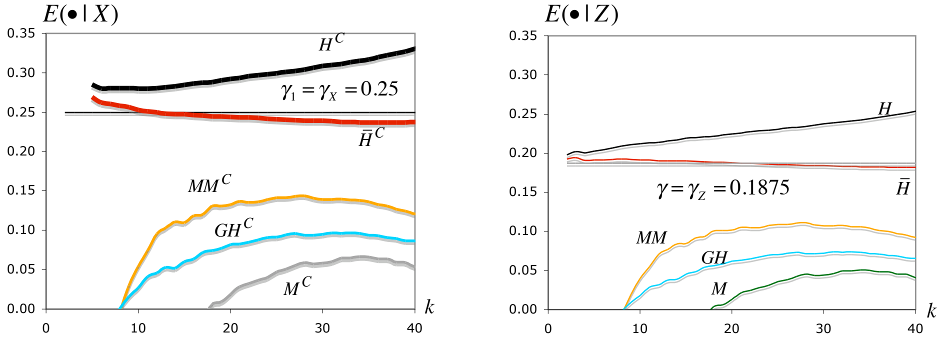

Figure 4. Mean values of the Hill, bias-corrected Hill, moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = 0.25$ (*left*) and of $\gamma = \gamma_Z = 0.1875$ (*right*), based on observed Z-samples of size $n = 100$ of $X \sim$ Fréchet($\gamma_1 = 0.25$) censored by $Y \sim$ Fréchet($\gamma_2 = 0.75$) — $p = 0.75$ (25% of censoring in the right tail).

This is the reason why we mentioned before, at the end of Section 1, the need for bias-correction, already available in Gomes *et al.* (2008a) for these heavy-tailed models, and these estimators. The bias-correction would lead to positive estimates, and a clear improvement of the EVI estimation not only for complete samples but also for censored samples, while performance is currently very poor. Anyway, in all the cases presented, it is clear that overall best performance comes from $\overline{H}$, for complete samples, and $\overline{H}^C$ for randomly censored samples.

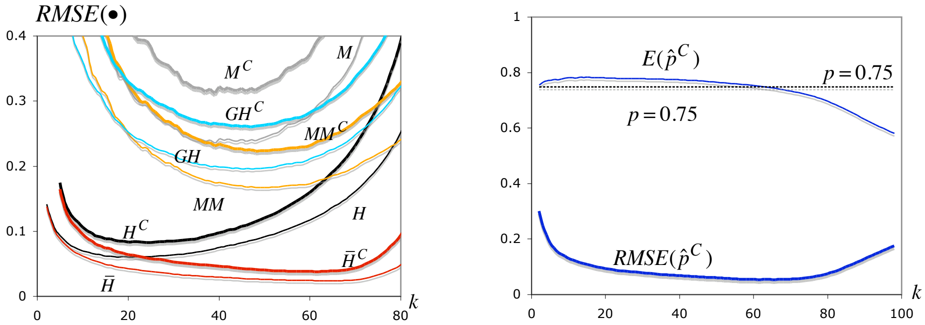

Figure 5. RMSE's of the estimators in Figure 4 (*left*) and E and RMSE of the estimator of $p = 0.75$ (*right*).

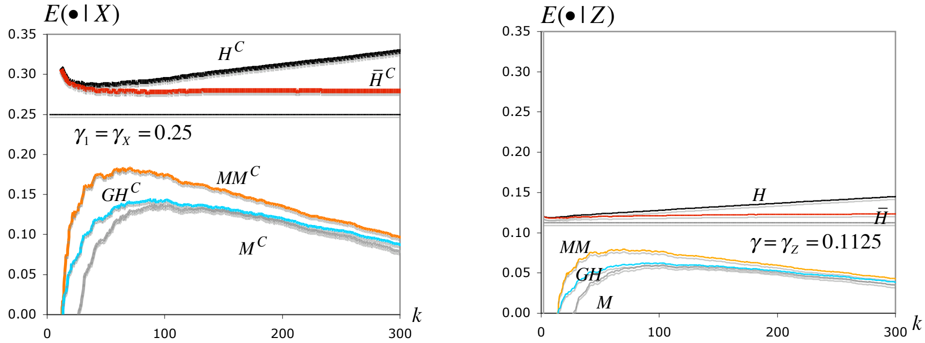

Figure 6. Mean values of Hill, bias-corrected Hill, moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = 0.25$ (*left*) and $\gamma = \gamma_Z = 0.1125$ (*right*), based on the observed Z-samples of size $n = 1000$ of a $X \sim$ Fréchet($\gamma_1 = 0.25$) censored by $Y \sim$ Fréchet($\gamma_2 = 0.2045$) — $p = 0.45$ (55% of censoring in the right tail).

5.2. Reversed-Burr underlying parents

The following figures, from 10 to 17 are equivalent to figures from 2 to 9, but for *Reversed-Burr* models. In these simulations, and due to the fact that the EVI is negative, we have considered neither H, in (6), nor $\overline{H}$, in (7).

As happened before, for heavy right tails, and as expected, the EVI-estimators for randomly censored samples always exhibit poorer behavior than the corresponding ones associated with complete samples. Such poor behavior becomes clearer when either the percentage of censoring in the right tail increases or the sample size decreases. For these *Reversed-Burr* parents, GH and GH^C exhibit the best performance among all the

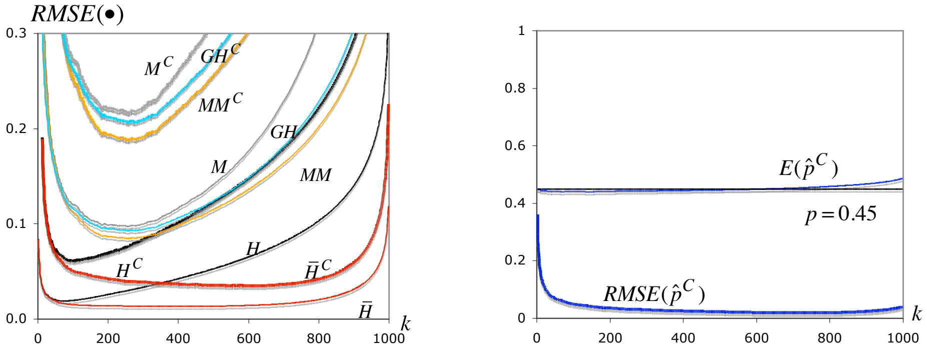

Figure 7. RMSE's of the estimators in Figure 6 (*left*) and E and RMSE of the estimator of $p = 0.45$ (*right*).

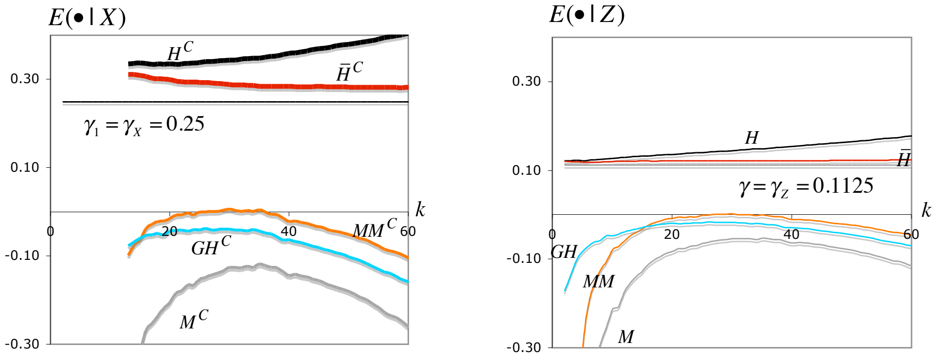

Figure 8. Mean values of Hill, bias-corrected Hill, moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = 0.25$ (*left*) and $\gamma = \gamma_Z = 0.1125$ (*right*), based on the observed Z-samples of size $n = 100$ of a $X \sim$ Fréchet($\gamma_1 = 0.25$) censored by $Y \sim$ Fréchet($\gamma_2 = 0.2045$) — $p = 0.45$ (55% of censoring in the right tail).

estimators considered in this paper, and bias-corrected M, GH and MM EVI-estimators, valid for a general $\gamma \in \mathbb{R}$, are surely advisable and welcome.

6. Applications to survival data sets

We have analyzed several data sets, available in Klein and Moeschberger (2005), among which we mention:

D1. Data on 80 males diagnosed with cancer of the tongue, with Z denoting time to death or on-study time, in weeks (Section 1.11; Sickle-Santanello *et al.*, 1988);

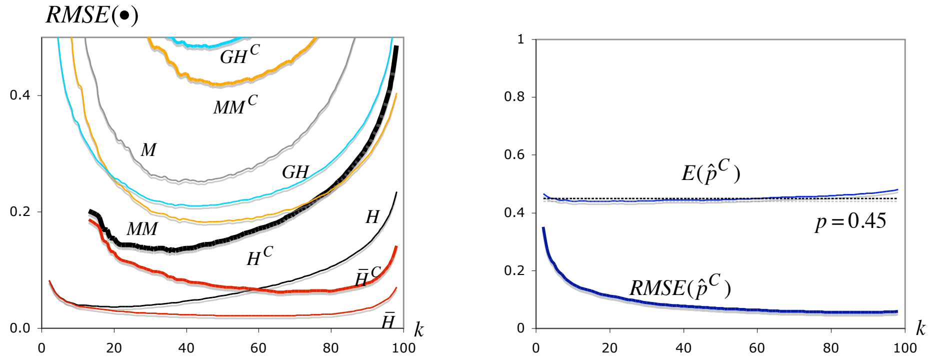

Figure 9. RMSE's of the estimators in Figure 8 (*left*) and E and RMSE of the estimator of $p = 0.45$ (*right*).

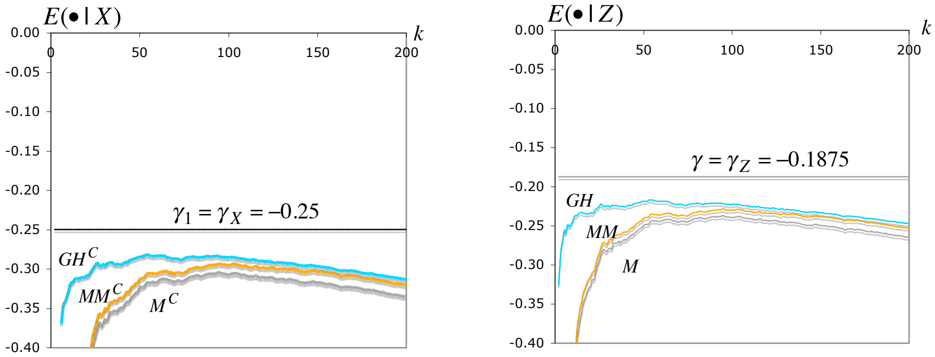

Figure 10. Mean values of moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = -0.25$ (*left*) and of $\gamma = \gamma_Z = -0.1875$ (*right*), based on observed Z-samples of size $n = 1000$ of $X \sim \mathrm{RB}(-0.25, 1, 0.5, 10)$ censored by $Y \sim \mathrm{RB}(-0.75, 1, 0.5, 10)$ — $p = 0.75$ (25% of censoring in the right tail).

D2. Data on 50 allotransplant and 51 autotransplants. The leukemia-free survival indicator is set at 0 whenever the person is alive without relapse and set at 1 if the person is dead or with relapse (Section 1.9);

D3. Data on 90 males with larynx cancer, with Z denoting again time to death or on-study time, in months (Section 1.8; Kardaun, 1983).

Although counterintuitive, there was an indication of a right tail clearly heavier than the normal, and also heavier than the Gumbel, for the data set in **D1.** (cancer of the tongue), where, due to the size of the sample ($n = 80$), we have jointly considered both kinds of tumor. However, the same

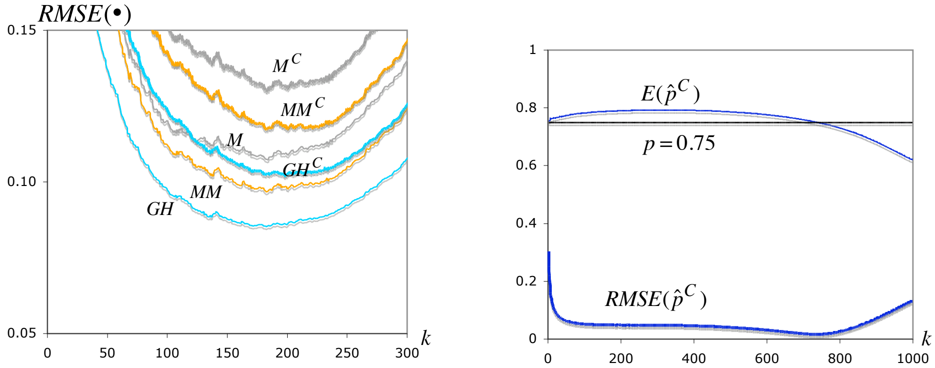

Figure 11. RMSE's of the estimators in Figure 10 (*left*) and E and RMSE of the estimator of $p = 0.75$ (*right*).

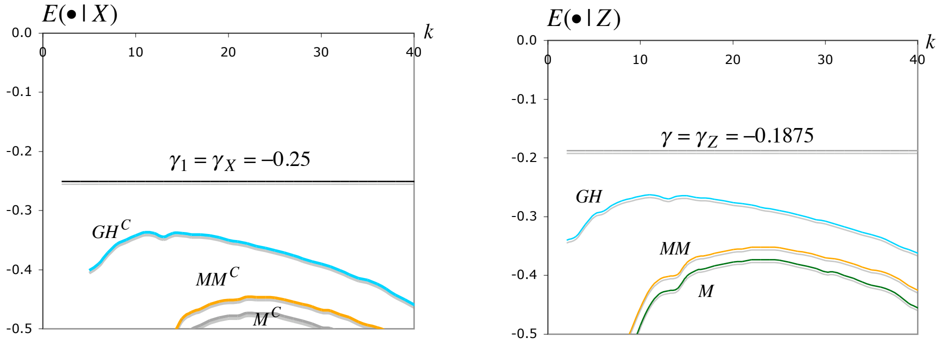

Figure 12. Mean values of moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = -0.25$ (*left*) and of $\gamma = \gamma_Z = -0.1875$ (*right*), based on observed Z-samples of size $n = 100$ of $X \sim \mathrm{RB}(-0.25, 1, 0.5, 10)$ censored by $Y \sim \mathrm{RB}(-0.75, 1, 0.5, 10)$ — $p = 0.75$ (25% of censoring in the right tail).

comment applies if we distinguish the two kinds of tumors (aneuploid and diploid). For the data in **D3.** (larynx cancer), we obtained no evidence of a heavy right tail. Indeed the right tail of the model underlying this data set can be considered exponential, i.e. there is no reason to reject the hypothesis $\gamma_Z = 0$. All other data sets analyzed, and in particular the data set in **D2.** (leukemia), clearly provide an indication of a light right tail, i.e. that $\gamma_Z < 0$. We shall next provide the estimation of the extreme value index γ_X for these three data sets.

In Figure 18 we provide estimates for γ_Z and γ_X on the left figure and estimates of p in the right figure, as a function of k, the number of top o.s.'s used in the estimation.

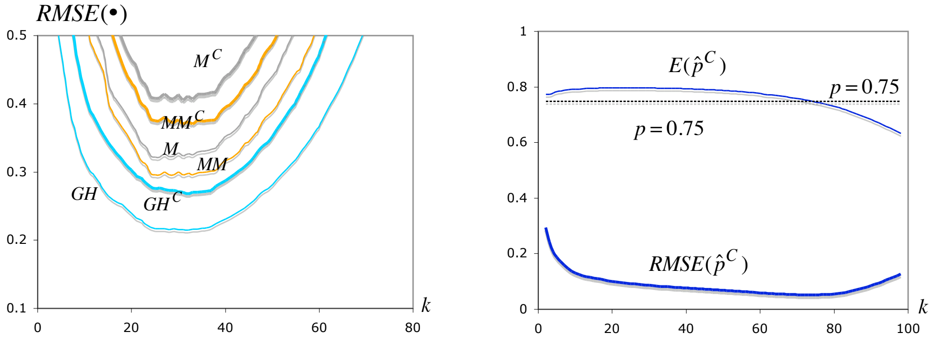

Figure 13. RMSE's of the estimators in Figure 12 (*left*) and E and RMSE of the estimator of $p = 0.75$ (*right*).

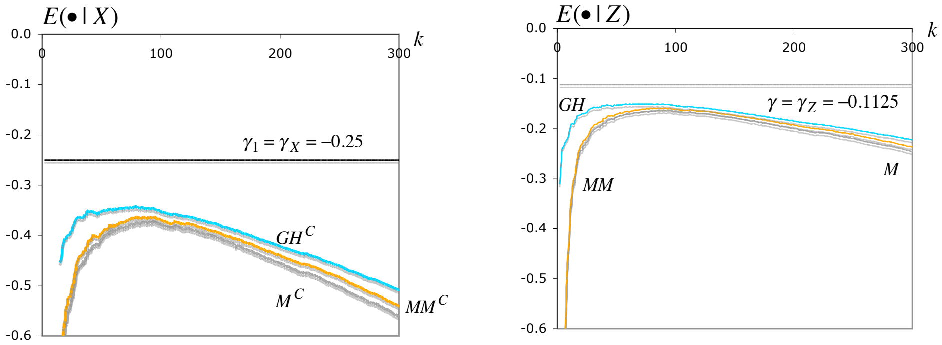

Figure 14. Mean values of moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_X = -0.25$ (*left*) and of $\gamma = \gamma_Z = -0.1125$ (*right*), based on observed Z-samples of size $n = 1000$ of $X \sim \mathrm{RB}(-0.25, 1, 0.5, 10)$ censored by $Y \sim \mathrm{RB}(-0.2045, 1, 0.5, 10)$ — $p = 0.45$ (55% of censoring in the right tail).

On the basis of any adequate stability criteria, we get the estimate $\hat{p} = 0.4$, for the parameter $p = \mathbb{P}(\delta = 1|Z = z)$. We have thus estimated a reasonably high censoring in the right tail, around 60%. The consecutive k values that lead to a difference $|\hat{p}_k^C - \hat{p}| \leqslant 0.05$ are $15 \leqslant k \leqslant 30$, the region pictured in Figure 18, top right. We were then led to the choice $\hat{k} := \arg\min_k |\hat{p}_k^C - \hat{p}| = 25$. For γ_Z, the use of any of the EVI estimators under consideration leads us to the estimate $\hat{\gamma}_Z = 0.4$. The estimate of γ_X at $\hat{k}$ is $\hat{\gamma}_X = 0.9$. The best decision, again taking into account stability criteria for the estimate sample paths, is the choice of $\overline{H}_k$, in (7), for γ_Z and the corresponding $\overline{H}_k^C$, for γ_X. We were then led to $\hat{\gamma} \equiv \hat{\gamma}_Z = \overline{H}_{\hat{k}} = 0.35$, and $\hat{\gamma}_1 \equiv \hat{\gamma}_X = \overline{H}_{\hat{k}}^C = 0.87$, the values pictured in Figure 18.

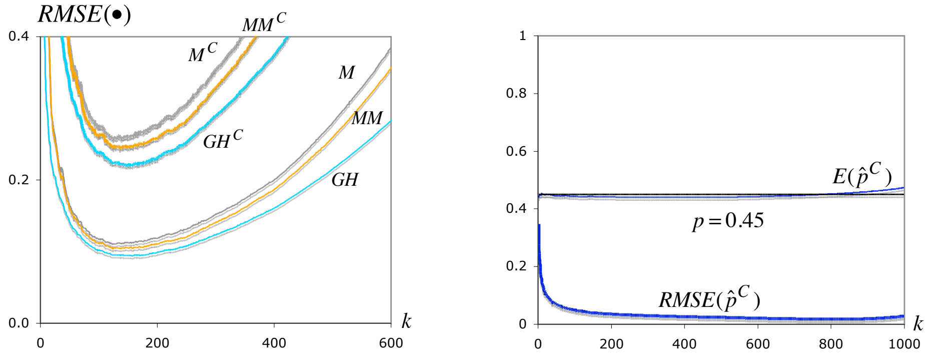

Figure 15. RMSE's of the estimators in Figure 14 (*left*) and E and RMSE of the estimator of $p = 0.45$ (*right*).

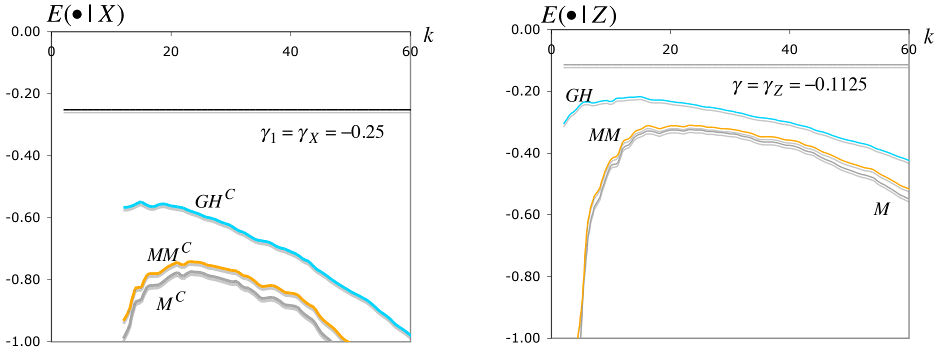

Figure 16. Mean values of moment, generalized Hill and mixed moment estimators of $\gamma_1 = \gamma_{_X} = -0.25$ (*left*) and of $\gamma = \gamma_{_Z} = -0.1125$ (*right*), based on observed Z-samples of size $n = 100$ of $X \sim \mathrm{RB}(-0.25, 1, 0.5, 10)$ censored by $Y \sim \mathrm{RB}(-0.2045, 1, 0.5, 10)$ — $p = 0.45$ (55% of censoring in the right tail).

Figure 19 is similar to Figure 18, but for the data in **D3.** Note that we were led to $\hat{p} = 0.3$, i.e., a censoring in the right tail around 70%, even higher than before. The same arguments as before led us to $20 \leqslant k \leqslant 50$, and $\hat{k} = 37$. In this region of k-values, all estimates under consideration, and valid for a general tail, are negative, i.e., the right-tail of the model underlying the data is light, and we thus discarded the Hill (H) and the MVRB ($\overline{H}$) estimators, valid only for heavy right tails. The final EVI estimates were obtained through GH_k, for the complete data, and GH_k^C, for the randomly censored data. We were led to $\hat{\gamma} \equiv \hat{\gamma}_{_Z} = GH_{\hat{k}} = -0.28$, and $\hat{\gamma}_1 \equiv \hat{\gamma}_{_X} = GH_{\hat{k}}^C = -0.94$, the values pictured in Figure 19.

Figure 20 is also similar to Figure 18, for the data mentioned in **D2.**

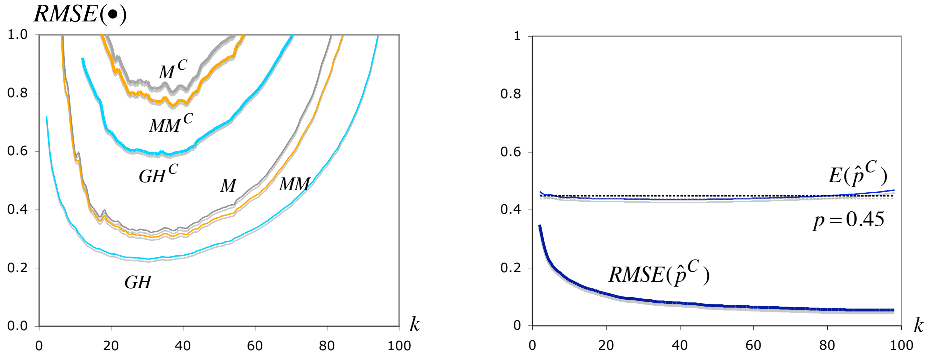

Figure 17. RMSE's of the estimators in Figure 16 (*left*) and E and RMSE of the estimator of $p = 0.45$ (*right*).

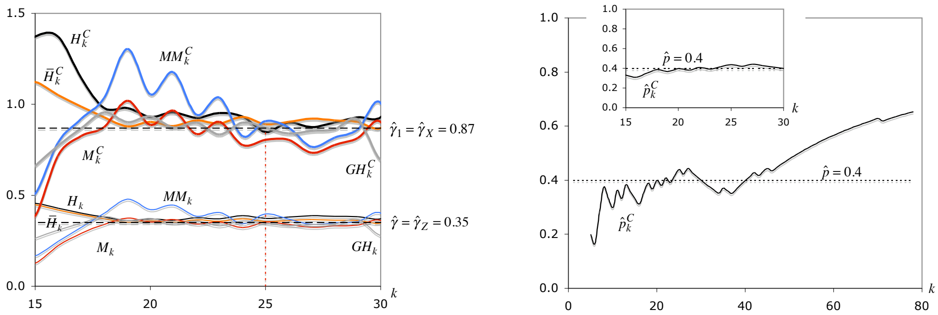

Figure 18. Estimation of parameters associated with data relating to cancer of the tongue.

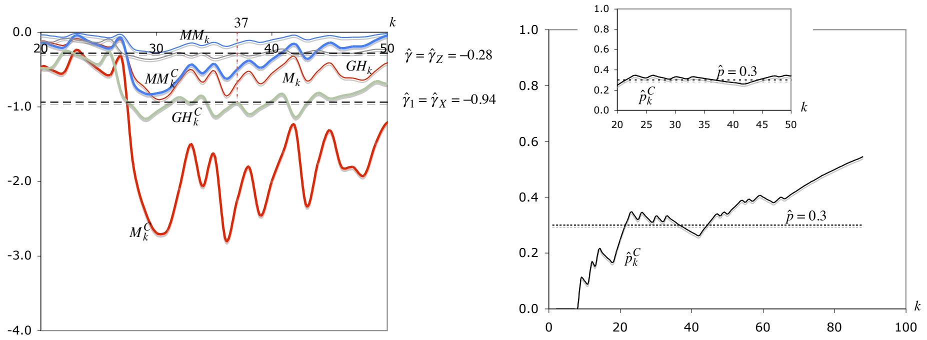

Figure 19. Estimation of parameters associated with data relating to larynx cancer.

For this data set, the percentage of censoring in the right-tail is even higher than before, around 80%. The same considerations as before led us to

$42 \leqslant k \leqslant 57$, and $\hat{k} = 45$. Now, the MM-estimates are positive in this region of k-values, but close to zero. The GH-estimates are slightly negative, but quite close to zero. The M-estimates exhibit a neat fluctuating behavior around zero, and we thus decided on the choice of M_k and M_k^C. We were then led to $\hat{\gamma} \equiv \hat{\gamma}_Z = M_{\hat{k}} = 0.01$, and $\hat{\gamma}_1 \equiv \hat{\gamma}_X = M_{\hat{k}}^C = -0.00$, and we see no reason to consider any deviation from an exponential right-tail, i.e., to consider that $\gamma_X = \gamma_Z = 0$.

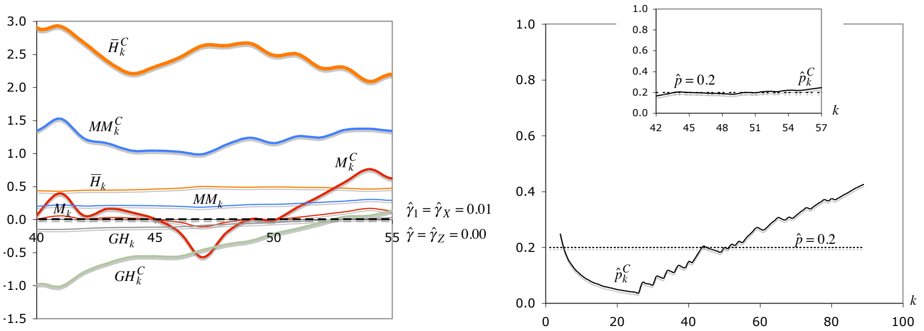

Figure 20. Estimation of parameters associated with data relating to leukemia.

7. Concluding remarks and future research on the topic

We conclude by making the following remarks:

- Whenever dealing with randomly censored data and a semi-parametric framework, i.e., whenever assuming that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$, for a certain γ, a first test on the sign of γ, the EVI, is without doubt a sensible task. In many areas where extreme events are relevant, the simplest case $\gamma = 0$ is often considered. But if we clearly come to the conclusion that $\gamma < 0$ or that $\gamma > 0$, we have specific procedures for the estimation of γ, possibly more reliable than the procedures valid for a general $\gamma \in \mathbb{R}$. Prior to a deeper semi-parametric analysis of the tail associated with any type of data, it thus seems sensible to test

$$\begin{aligned} &H_0 : F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma=0} \ \Big(\text{or } F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma \geqslant 0}\Big) \quad \text{versus} \\ &H_1 : F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)_{\gamma<0}, \end{aligned}$$

through the use of any semi-parametric test statistic. Several test procedures are available for complete data (see Neves and Fraga Alves, 2008, for an overview and recent approaches), and similar procedures need to be developed for randomly censored data.

- As mentioned before, if confronted with a heavy right tail, i.e., a positive EVI, we strongly advise the use of any of the recent MVRB EVI-estimators introduced and studied in Caeiro *et al.* (2005) and Gomes *et al.* (2007, 2008b). For a general tail, and particularly if we have a clear indication that the right tail is light, i.e., the EVI is negative, we suggest the use of a second-order reduced-bias mixed moment or generalized Hill estimator, not yet considered in the literature, essentially due to the difficulties of estimating the functional $A(t)$, in (3). The estimation of the second-order parameter ρ, for a general tail has been provided in Fraga Alves *et al.* (2003). The estimation of $A(\cdot)$ or of an adequate scale second-order parameter in the above mentioned function A is still a topic for further research.
- A third open and relevant topic of related research is the development of a reduced-bias estimation of the percentage of censoring in the right tail of the underlying model. This will certainly lead to more precise EVI-estimators under the schemes under consideration, i.e., under randomly censored schemes. The non-degenerate asymptotic and finite-sample behavior of those estimators is then needed, and constitutes another open topic of research in the field.

Acknowledgments

To Professor João Tiago Mexia, another token of friendship.

This research was partially supported by FCT/OE, POCI 2010 and PTDC/FEDER.

References

Beirlant J., Vynckier P., Teugels J. (1996): Tail index estimation, Pareto quantile plots, and regression diagnostics. J. Amer. Statist. Assoc. 91: 1659–1667.

Beirlant J., Guillou A., Dierckx G., Fils-Viletard A. (2007): Estimation of the extreme value index and extreme quantiles under random censoring. Extremes 10(3): 151–174.

Bingham N., Goldie C.M., Teugels J.L. (1987): Regular Variation. Cambridge Univ. Press, Cambridge.

Caeiro F., Gomes M.I., Pestana D. (2005): Direct reduction of bias of the classical Hill estimator. Revstat 3(2): 113–136.

Dekkers A., Einmahl J., de Haan L. (1989): A moment estimator for the index of an extreme-value distribution. Ann. Statist. 17: 1833–1855.

Einmahl J.H.J, Fils-Villetard A., Guillou A. (2008): Statistics of extremes under random censoring. Bernoulli 14(1): 207–227.

Fraga Alves M.I., de Haan L., Lin T. (2003): Estimation of the parameter controlling the speed of convergence in extreme value theory. Mathematical Methods of Statistics 12(2): 155–176.

Fraga Alves M.I., Gomes M.I., de Haan L., Neves C. (2009): Mixed moment estimator and location invariant alternatives. Extremes 12(2): 149–185.

Gomes M.I., Neves M.M. (2010): A note on statistics of extremes for censoring schemes on a heavy right tail. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of ITI 2010, SRCE Univ. Computing Centre Editions: 539–544.

Gomes M.I., Pestana D. (2007): A sturdy reduced bias extreme quantile (VaR) estimator. J. Amer. Statist. Assoc. 102(477): 280–292.

Gomes M.I., Martins M.J., Neves M.M. (2007): Improving second order reduced bias extreme value index estimation. Revstat 5(2): 177–207.

Gomes M.I., Fraga Alves M.I., Araújo Santos P. (2008a): PORT Hill and moment estimators for heavy-tailed models. Communications in Statistics – Simulation and Computation 37(7): 1281–1306.

Gomes M.I., de Haan L., Henriques Rodrigues L. (2008b): Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. J. Royal Statistical Society B70(1): 31–52.

de Haan L. (1984): Slow variation and characterization of domains of attraction. In Tiago de Oliveira, ed., Statistical Extremes and Applications: 31–48, D. Reidel, Dordrecht, Holland.

Hill B. (1975): A simple general approach to inference about the tail of a distribution. Ann. Statist. 3: 1163–1174.

Kardaun O. (1983): Statistical survival analysis of male larynx cancer patients – a case study. Statistica Neerlandica 37: 103–125.

Klein J.P., Moeschberger M.L. (2005): *Datasets for Survival Analysis – Techniques for Censored and Truncated Data*. Second Edition, Springer.

Neves C., Fraga Alves M.I. (2008): Testing extreme value conditions – an overview and recent approaches. Revstat 6(1): 83–100.

Reiss R.D., Thomas M. (1997): *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields*. Birkhäuser Verlag, Basel.

Sickle-Santanello B.J., Farrar W.B., DeCenzo J.F., Keyhani-Rofagha S., Klein J., Pearl D., Laufman H., O'Toole R.V. (1988): Technical and statistical improvements for flow cytometric DNA analysis of paraffin-embedded tissue. Cytometry 9(6): 594-9.