New estimators of the extreme value index under random right censoring, for heavy-tailed distributions

Julien Worms, Rym Worms

New estimators of the extreme value index under random right censoring, for heavy-tailed distributions

Julien Worms · Rym Worms

Uploaded in HAL : 11th december 2013

Abstract This paper presents new approaches for the estimation of the extreme value index in the framework of randomly censored (from the right) samples, based on the ideas of Kaplan-Meier integration and the synthetic data approach of S.Leurgans (1987). These ideas are developed here in the heavy tail case and for the adaptation of the Hill estimator, for which the consistency is proved under first order conditions. Simulations show good performance of the two approaches, with respect to the only existing adaptation of the Hill estimator in this context.

Keywords Extreme value index · Tail inference · Random censoring · Kaplan-Meier integration

Mathematics Subject Classification (2000) 62G32 (Extreme value statistics) · 62N02 (Estimation for censored data)

1 Introduction

Estimating the extreme value index is an important problem in extreme value statistics. A distribution function (d.f.) F is said to be in the maximum domain of attraction of H_γ (noted $F \in D(H_\gamma)$) with

$$H_\gamma(x) := \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & \text{for } \gamma \neq 0 \text{ and } 1+\gamma x > 0 \\ \exp(-\exp(-x)) & \text{for } \gamma = 0 \text{ and } x \in \mathbb{R}\ , \end{cases}$$

Julien Worms
Université de Versailles-Saint-Quentin-en-Yvelines, Laboratoire de Mathématiques de Versailles (CNRS UMR 8100), F-78035 Versailles Cedex, France, E-mail: julien.worms@uvsq.fr

Rym Worms
Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées (CNRS UMR 8050), UPEMLV, UPEC, F-94010, Créteil, France, E-mail: rym.worms@u-pec.fr

if there exist two normalizing sequences $(a_n) \subset \mathbb{R}^+$ and $(b_n) \subset \mathbb{R}$ such that (for every $x \in \mathbb{R}$)

$$F^n(a_n x + b_n) \overset{n\to\infty}{\longrightarrow} H_\gamma(x).$$

If we observe a sample $(X_i)_{i\leq n}$ with common distribution function $F \in D(H_{\gamma_X})$, with $\gamma_X > 0$, a classical estimator of the extreme value index γ_X is the so-called Hill estimator

$$\hat{\gamma}_{X,Hill} := \frac{1}{k_n}\sum_{i=1}^{k_n} \log\left(\frac{X_{n-i+1,n}}{X_{n-k_n,n}}\right)$$

where $X_{1,n} \leq \cdots \leq X_{n,n}$ are the ascending order statistics associated to the X-sample and k_n the sample fraction to keep from this sample.

However, in a certain number of applications, such as survival analysis, reliability theory, insurance ..., the variable of interest X is not necessarily completely available. This is the case in the presence of random right censoring. Examples of censored data with apparent heavy tails can be found in [Gomes and Neves (2011)] and [Einmahl et. al. (2008)].

The usual way to model this situation is to introduce a random variable C, independent of X, such that only

$$Z = X \wedge C \quad \text{and} \quad \delta = \mathbb{I}_{X\leq C}$$

are observed. The observed variable δ determines whether X has been censored or not. It is common sense that any classical estimator of the extreme value index (such as the Hill or the Moment estimator) is not consistent for estimating γ_X if it is naively computed from the Z-sample (indeed, it estimates the extreme value index associated to the Z-sample, denoted by γ in the sequel).

Recently, [Beirlant et. al. (2007)] and [Einmahl et. al. (2008)] proposed an adaptation of classical extreme value index estimators in the case of right random censoring, therefore providing (to the best of our knowledge) the first methodological papers on this subject; their method will be presented in subsection 2.1.

In this paper, we propose two different approaches to deal with the estimation of γ_X, relying on more natural heuristics in this randomly censored sample framework : one of these amounts to consider Kaplan-Meier integrals, and the other on ideas coming from censored regression. Given the combination of difficulties coming from extreme values and censoring, we will restrict ourselves to the application of these approaches to the adaptation of the Hill estimator, in the heavy tailed case, and to the consistency of these adapted estimators. It is however more than likely that our ideas adapt to other, more efficient, estimators of the extreme value index, and to other domains of attraction.

In Subsection 2.1, we define the framework more precisely and recall the existing methodology cited above. In Subsections 2.2 and 2.3, we present our methodology and state the consistency results for the adaptation of the Hill estimator. Section 3 is devoted to a small simulation study, and Section 5 to

the proofs. A conclusion is provided in Section 4.

Notations : in the whole paper, the sign := denotes an equality defining the quantity on the left side, $f(t^-)$ denotes $\lim_{s\uparrow t} f(s)$, and $f^{\leftarrow}$ the general inverse of the function f. $\mathbb{I}_A(x)$ is equal to 1 if $x \in A$ and 0 if $x \notin A$. The definition of a regularly varying function f of order α (noted $f \in RV_\alpha$) is recalled in the Appendix.

2 Methodology

2.1 The framework and a general existing methodology

We consider in this paper two independent i.i.d. non-negative samples $(X_i)_{i\leq n}$ and $(C_i)_{i\leq n}$ with respective continuous distribution functions F and G (with end-points τ_F and τ_G, where $\tau_F := \sup\{x,\ F(x) < 1\}$). In the context of randomly right-censored observations, one actually only observes, for $i \leq n$,

$$Z_i = X_i \wedge C_i \quad \text{and} \quad \delta_i = \mathbb{I}_{X_i \leq C_i}.$$

We denote by H the distribution function of the Z-sample, satisfying

$$1 - H = (1 - F)(1 - G)$$

and by $Z_{1,n} \leq \cdots \leq Z_{n,n}$ the associated order statistics. In the whole paper, $\delta_{1,n}, \ldots, \delta_{n,n}$ denote the δ's corresponding to $Z_{1,n}, \ldots, Z_{n,n}$, respectively ([Stute (1995)] call them "concomitant" to the order statistics).

If F and G are assumed to be in the maximum domains of attraction $D(H_{\gamma_X})$ and $D(H_{\gamma_C})$ respectively, where γ_X and γ_C are real numbers, then this implies that $H \in D(H_\gamma)$, for some $\gamma \in \mathbb{R}$. [Einmahl et. al. (2008)] considered the following three most interesting cases :

$$\begin{array}{lll}
\text{case 1:} & \gamma_X > 0\ ,\ \gamma_C > 0 & \text{in this case } \gamma = \dfrac{\gamma_X \gamma_C}{\gamma_X + \gamma_C} \\
\text{case 2:} & \gamma_X < 0\ ,\ \gamma_C < 0\ ,\ \tau_F = \tau_G & \text{in this case } \gamma = \dfrac{\gamma_X \gamma_C}{\gamma_X + \gamma_C} \\
\text{case 3:} & \gamma_X = \gamma_C = 0\ ,\ \tau_F = \tau_G = +\infty & \text{in this case } \gamma = 0.
\end{array}$$

The general existing method, appeared first in [Beirlant et. al. (2007)] and developed in [Einmahl et. al. (2008)], is to consider any consistent estimator $\hat{\gamma}_Z$ of the extremal index γ applied to the Z-sample and divide it by the proportion $\hat{p}$ of non-censored observations in the tail (*i.e.* in the k_n largest observations of the Z-sample). That is, an adaptation of an extreme value index estimator in the presence of random right censoring is :

$$\hat{\gamma}_X^c := \frac{\hat{\gamma}_Z}{\hat{p}}, \quad \text{where} \quad \hat{p} := \frac{1}{k_n} \sum_{j=1}^{k_n} \delta_{n-j+1,n}. \tag{1}$$

It is proved in [Einmahl et. al. (2008)] that $\hat{p}$ consistently estimates $p := \frac{\gamma_C}{\gamma_X+\gamma_C}$ and therefore $\hat{\gamma}_X^c$ consistently estimates $\gamma/p = \gamma_X$ (obtaining as well the asymptotic normality, if it holds for the sequence $\hat{\gamma}_Z$). This method provides flexibility as to the choice of the estimator of γ_Z. Up to now, the only alternative to it, in this context, is the adaptation of the ML estimator based on the excesses over a threshold, developed in [Beirlant et. al. (2010)], which we shall however not detail here.

We now present an alternative path for estimating the extreme value index γ_X, based on ideas which are well-known in the survival analysis literature.

2.2 First approach

The starting point of the first new approach is the following well known result, which is the basis of censored regression methods (for instance, an early reference is [Koul et. al. (1981)]) : if ϕ is some nonnegative real function,

$$\mathbb{E}\left[\frac{\delta}{1-G(Z)}\phi(Z)\right] = \mathbb{E}[\phi(X)]. \tag{2}$$

It is readily proved : since $Z = X$ when $\delta = 1$,

$$\begin{aligned}\mathbb{E}\left[\frac{\delta}{1-G(Z)}\phi(Z)\right] &= \iint \mathbb{I}_{x\le c}\frac{\phi(x)}{1-G(x)}\,dF(x)\,dG(c)\\ &= \int \phi(x)(1-G(x))^{-1}\left(\int_x^{+\infty} dG(c)\right)dF(x)\\ &= \int \phi(x)\,dF(x) \;=\; \mathbb{E}[\phi(X)].\end{aligned}$$

In the context of extreme value statistics, the idea is to take advantage of this property and of the fact that some tail parameters of the distribution of X can be approached by the expectation of some function of X, therefore opening the way to their estimation. In this paper we will illustrate it in the context of heavy-tailed distributions, and for the estimation of the extreme value index, assuming that we are in the first of the three situations presented in paragraph 2.1 :

$$F \in D(H_{\gamma_X}) \;\;,\;\; G \in D(H_{\gamma_C}) \;\text{ with } \gamma_X > 0 \text{ and } \gamma_C > 0 \tag{3}$$

which, as noted earlier, implies that $H \in D(H_\gamma)$ with

$$\gamma = \frac{\gamma_X\gamma_C}{\gamma_X+\gamma_C}.$$

In this case, it is well known that (see Remark 1.2.3 in [Haan and Ferreira (2006)] for instance)

$$\lim_{t\to+\infty} \mathbb{E}\left[\log(X/t)\,|\,X>t\right] = \gamma_X. \tag{4}$$

If (k_n) is a sequence of integers satisfying, as n tends to $+\infty$,

$$k_n \to +\infty \ \text{ and } \ k_n = o(n) \tag{5}$$

then we can define a random version of $\phi(x) = (1-F(t))^{-1}\log(x/t)\mathbb{I}_{x>t}$, with random threshold $t = Z_{n-k_n,n}$

$$\hat{\phi}_n(x) := \frac{1}{1-\hat{F}_n(Z_{n-k_n,n})}\log\left(\frac{x}{Z_{n-k_n,n}}\right)\mathbb{I}_{x>Z_{n-k_n,n}}. \tag{6}$$

Consequently, by combining (2) and (4) with this function $\hat{\phi}_n$, our first adaptation of the Hill estimator comes, valid in situation (3),

$$\hat{\gamma}^{KM}_{X,Hill} := \frac{1}{n(1-\hat{F}_n(Z_{n-k_n,n}))}\sum_{i=1}^{k_n}\frac{\delta_{n-i+1,n}}{1-\hat{G}_n(Z^-_{n-i+1,n})}\log\left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}\right), \tag{7}$$

where $\hat{F}_n$ and $\hat{G}_n$ (the Kaplan-Meier estimators of F and G, respectively) are defined as follows : for $t < Z_{n,n}$,

$$1-\hat{F}_n(t) = \prod_{Z_{i,n}\le t}\left(\frac{n-i}{n-i+1}\right)^{\delta_{i,n}} \ \text{ and } \ 1-\hat{G}_n(t) = \prod_{Z_{i,n}\le t}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_{i,n}}.$$

Note that we take $\hat{G}_n(Z^-_{n-i+1,n})$ instead of $\hat{G}_n(Z_{n-i+1,n})$, in the definition of $\hat{\gamma}^{KM}_{X,Hill}$, because $1-\hat{G}_n(Z_{n,n})$ can be zero or undefined.

The following theorem provides the consistency of this estimator. For this purpose, we need two additional assumptions on the behavior of the function $p\circ H^{\leftarrow}$: if $p(z) := \mathbb{P}(\delta = 1|Z = z)$,

$$\frac{1}{k_n}\sum_{i=1}^{k_n}\left|p\left(H^{\leftarrow}\left(1-\frac{i}{n}\right)\right)-p\right| \ \xrightarrow{\mathbb{P}} \ c\in\mathbb{R} \tag{8}$$

$$\sup_{(s,t)\in C_n}|p(H^{\leftarrow}(t))-p(H^{\leftarrow}(s))| \ \to 0, \text{ for all } C>0 \tag{9}$$

where $C_n = \{(s,t) \text{ such that } s<1\ ,\ 1-k_n/n\le t<1\ ,\ |t-s|\le C\sqrt{k_n}/n\ \}$.

Theorem 1 *Under assumptions* (3), (5), (8), (9), *if we additionally assume that*

$$-\log(k_n/n)/k_n = O(n^{-\delta}), \tag{10}$$

for some $\delta>0$ *if* $\gamma_X<\gamma_C$ *and for some* $\delta\ge\dfrac{\gamma_X-\gamma_C}{\gamma_X+\gamma_C}$ *if* $\gamma_X\ge\gamma_C$, *then, as* n *tends to* $+\infty$,

$$\hat{\gamma}^{KM}_{X,Hill} \ \xrightarrow{\mathbb{P}} \ \gamma_X\,.$$

Remark 1 *Inspection of the proof (see the treatment of P_n at the end of subsection 5.1.2) reveals that any sequence (k_n) such that $k_n \geq cn^a$ (for some $a \in]0,1[$, some constant $c > 0$, and n large) is suitable for the theorem to hold (whether $\gamma_X < \gamma_C$ or not, and without assuming (10)).*

Remark 2 *Assumptions (8) and (9) have similarities with (but are weaker than) those used in [Einmahl et. al. (2008)]. The latter have been proved, in [Brahimi et. al. (2013)], to be unnecessary for obtaining the asymptotic normality of the adapted Hill estimator $\hat{\gamma}^c_{X,Hill} := \hat{\gamma}_{Z,Hill}/\hat{p}$ (see (1)). The empirical process techniques used in [Brahimi et. al. (2013)] do not seem to be applicable in our setting, since our estimator is a nonconstantly-weighted version of the Hill estimator*

$$\hat{\gamma}^{KM}_{Z,Hill} = \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{1 - \hat{G}_n(Z_{n-k_n,n})}{1 - \hat{G}_n(Z_{n-i+1,n})} \delta_{n-i+1,n} \log\left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}\right)$$

whereas

$$\hat{\gamma}^c_{Z,Hill} = \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{1}{\hat{p}} \log\left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}\right)$$

Remark 3 *We have made the choice of a random threshold and fixed number of relative excesses, because it seemed closer to what is done in practice and, secondly, the other path went with its own difficulties. This other choice was to consider a deterministic threshold t_n, and then (up to the estimation of F at t_n, which causes no problem) write $\hat{\gamma}^{KM}_X$ as a proper Kaplan-Meier integral $\int \phi_n(x)\, d\hat{F}_n(x)$, with deterministic $\phi_n(x) = (1-F(t_n))^{-1} \log(x/t_n) \mathbb{I}_{x>t_n}$ (with our choice, the function $\hat{\phi}_n$ is random). However, this function ϕ_n is intrinsically unbounded, has a "sliding towards infinity" support, and is dependent on n : we found no way to deal with this, using the Kaplan-Meier integration tools known in the literature.*

2.3 Second approach

Our second approach, alternative to the Kaplan-Meier integral approach presented in the previous paragraph, is based on ideas of [Leurgans (1987)], who developed a "synthetic data" strategy in censored regression problems (see [Delecroix et. al. (2008)] for a more recent reference to this method). The starting point of this second approach is the following result :

if ϕ and ψ are two nonnegative $\mathbb{R}_+ \to \mathbb{R}$ functions such that $\int_0^x \psi(t)\, dt = \phi(x)$, then

$$\mathbb{E}\left[\int_0^Z \frac{\psi(t)}{1 - G(t)}\, dt\right] = \mathbb{E}[\phi(X)]. \tag{11}$$

Indeed,

$$\mathbb{E}\left[\int_0^Z \frac{\psi(t)}{1-G(t)}\,dt\right] = \int_0^{+\infty}\left(\int_t^{+\infty} dH(z)\right)\frac{\psi(t)}{1-G(t)}\,dt$$
$$= \int_0^{+\infty}(1-F(t))\psi(t)\,dt \;=\; \int_0^{+\infty}\left(\int_t^{+\infty} dF(x)\right)\psi(t)\,dt$$
$$= \int_0^{+\infty}\left(\int_0^{x}\psi(t)\,dt\right)dF(x) \;=\; \mathbb{E}[\phi(X)].$$

Consequently, if $\phi(x)=\int_0^x \psi(z)\,dz$, then

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^{Z_i}\frac{\psi(z)}{1-\hat{G}_n(z^-)}\,dz$$

should correctly estimate $\mathbb{E}[\phi(x)]$. This estimator can be rewritten using the special form of function ψ and the piecewise constant form of the Kaplan-Meier estimator $\hat{G}_n$: noting $Z_{0,n}=0$, and $rk(Z_i)$ the (ascending order) rank of the observation Z_i in the Z-sample, we have indeed

$$\int_0^{Z_i}\frac{\psi(z)}{1-\hat{G}_n(z^-)}\,dz = \sum_{j=1}^{rk(Z_i)}\int_{]Z_{j-1,n},Z_{j,n}]}\frac{\psi(z)}{1-\hat{G}_n(z^-)}\,dz$$
$$= \sum_{j=1}^{rk(Z_i)}\frac{1}{1-\hat{G}_n(Z_{j-1,n})}\int_{Z_{j-1,n}}^{Z_{j,n}}\psi(z)\,dz$$
$$= \sum_{j=1}^{rk(Z_i)}\frac{\phi(Z_{j,n})-\phi(Z_{j-1,n})}{1-\hat{G}_n(Z_{j-1,n})}.$$

Considering, once again, the function $\hat{\phi}_n$ introduced in (6), we can now define our second new adaptation of the Hill estimator, valid in situation (3)

$$\hat{\gamma}_{X,Hill}^{Leurg} \;:=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{i}\frac{\hat{\phi}_n(Z_{j,n})-\hat{\phi}_n(Z_{j-1,n})}{1-\hat{G}_n(Z_{j-1,n})}$$

which turns out to be, after some simplifications,

$$\hat{\gamma}_{X,Hill}^{Leurg} \;=\; \frac{1}{n(1-\hat{F}_n(Z_{n-k_n,n}))}\sum_{i=1}^{k_n}\frac{1}{1-\hat{G}_n(Z^-_{n-i+1,n})}\,i\log\left(\frac{Z_{n-i+1,n}}{Z_{n-i,n}}\right). \tag{12}$$

We note that, while $\hat{\gamma}_{X,Hill}^{KM}$ appeared as a weighted version of the classical form of the Hill estimator (mean of the log relative excesses $\log(Z_{n-i+1,n}/Z_{n-k_n,n})$), our second candidate $\hat{\gamma}_{X,Hill}^{Leurg}$ is a weighted version (but with weights which are always non null) of the mean of the so-called log spacings $i\log(Z_{n-i+1,n}/Z_{n-i,n})$, *i.e.* the other form of the Hill estimator.

The following theorem provides the consistency of this estimator, under less restrictive conditions than Theorem 1.

Theorem 2 *Under assumptions* (3), (5) *and* (10), *then, as* n *tends to* $+\infty$,

$$\hat{\gamma}_{X,Hill}^{Leurg} \xrightarrow{\mathbb{P}} \gamma_X \,.$$

3 Finite sample behaviour

In this section, we present some graphs (issued from an extensive study in the heavy tail framework) corresponding to the comparison, in terms of observed bias and mean squared error (MSE) of our new estimators $\hat{\gamma}_{X,Hill}^{KM}$ and $\hat{\gamma}_{X,Hill}^{Leurg}$ (defined by (7) and (12)) with other adapted estimators for γ_X (see (1)) : the adapted Hill estimator $\hat{\gamma}_{X,Hill}^{c}$ and the adapted version of the bias-corrected Hill estimator (introduced and studied in [Caeiro et. al. (2005)] and [Gomes et. al. (2008)]) $\hat{\gamma}_{X,MVRB}^{c}$, where MVRB stands for minimum-variance reduced-bias.

We consider two classes of heavy-tailed distributions :

- Burr(β, τ, λ) with d.f. $1 - (\frac{\beta}{\beta + x^{\tau}})^{\lambda}$, which extreme value index is $\frac{1}{\lambda\tau}$.
- Frechet(γ) with d.f. $\exp(-x^{-1/\gamma})$, which extreme value index is γ.

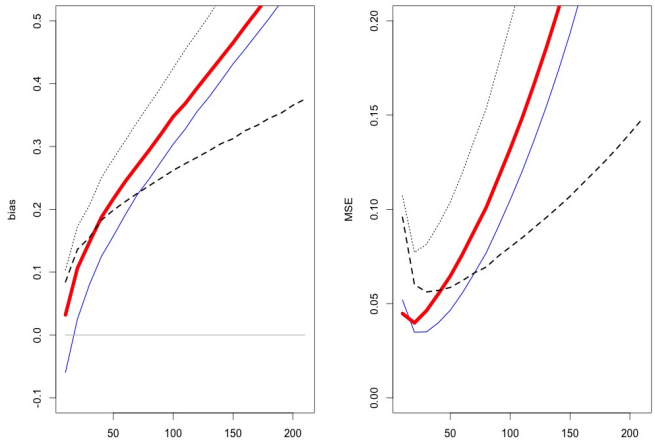
For each considered distribution, 2000 random samples of length $n = 500$ were generated ; median bias and MSE of the four above-mentioned estimators are plotted against different values of k_n, the number of excesses used.

We considered three cases : a Burr distribution censored by another Burr distribution (Fig.1), a Fréchet distribution censored by another Fréchet distribution (Fig.2) and a Fréchet distribution censored by a Burr distribution (Fig.3). In each case, we considered a situation with $\gamma_X < \gamma_C$ (subfigure (a)), which corresponds to a weak censoring in the tail, and the reverse situation with $\gamma_X > \gamma_C$ (subfigure (b)), which corresponds to a strong censoring.

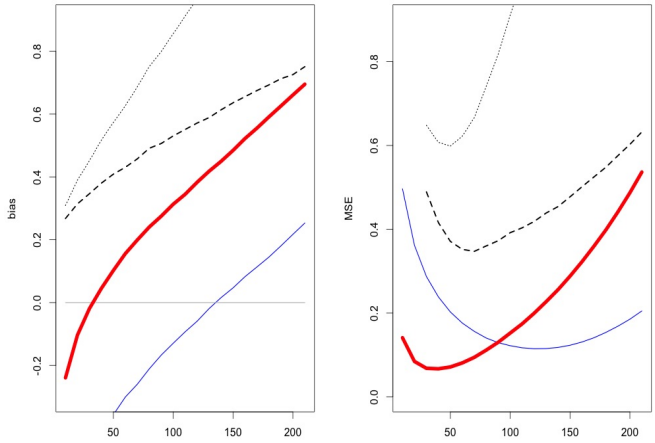
It seemed more natural to present separately, below, the comparison of our estimators $\hat{\gamma}_{X,Hill}^{KM}$ and $\hat{\gamma}_{X,Hill}^{Leurg}$ with $\hat{\gamma}_{X,Hill}^{c}$ on one hand, and with $\hat{\gamma}_{X,MVRB}^{c}$ on the other hand.

From the three situations presented above, it seems that our new estimators perform better than the former adapted Hill estimator, in the weak censoring case, both in term of bias and MSE. It is not surprising that, in the strong censoring case, results become worse for all estimators but (clearly) to a lesser extent for ours. Moreover, in the strong censoring situation, $\hat{\gamma}_{X,Hill}^{Leurg}$ seems to have systematically the best behavior. Other simulations not presented here confirm this phenomenon.

Turning now to the comparison of our estimators with the adaptation of the MVRB estimator, first note that intuitively this should end with a better performance for the latter, since in the uncensored situation it possesses better theoretical and empirical properties than the Hill estimator, on which our estimators are based. It is indeed the case when the bias-reduction is successful (as in Figure 2(a)), but surprisingly, our adaptations of the Hill estimator



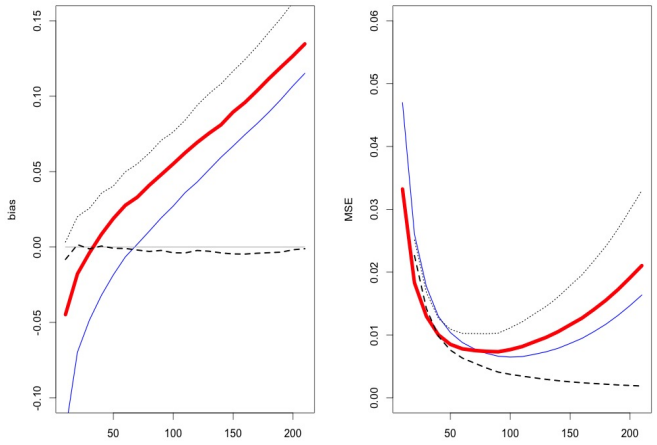
(a) Burr$(10, 1, 2)$ censored by Burr$(10, 1, 1)$



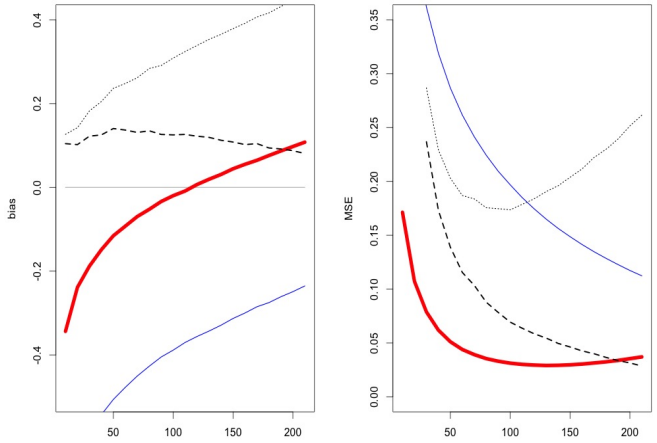
(b) Burr$(10, 1, 1)$ censored by Burr$(10, 1, 2)$

Fig. 1 Comparison of bias and MSE for $\hat{\gamma}^{KM}_{X,Hill}$ (solid), $\hat{\gamma}^{Leurg}_{X,Hill}$ (thick), $\hat{\gamma}^{c}_{X,Hill}$ (dotted) and $\hat{\gamma}^{c}_{X,MVRB}$ (dashed) for a Burr distribution censored by another Burr distribution : (a) $\gamma_X = 1/2$ and $\gamma_C = 1$ (weak censoring), (b) $\gamma_X = 1$ and $\gamma_C = 1/2$ (strong censoring)

compete quite well (especially the "Leurgans" one) or sometimes even outperform the MVRB usual adaptation (in term of MSE), particularly in the strong censoring framework. Even if these comments are only based on a rather small simulation study, they are nonetheless quite encouraging, and motivate the extension of our methodology to other estimators.

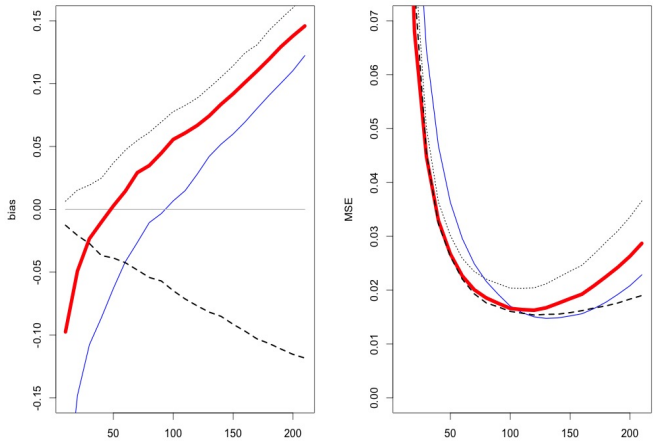


(a) Fréchet(1/2) censored by Fréchet(1)

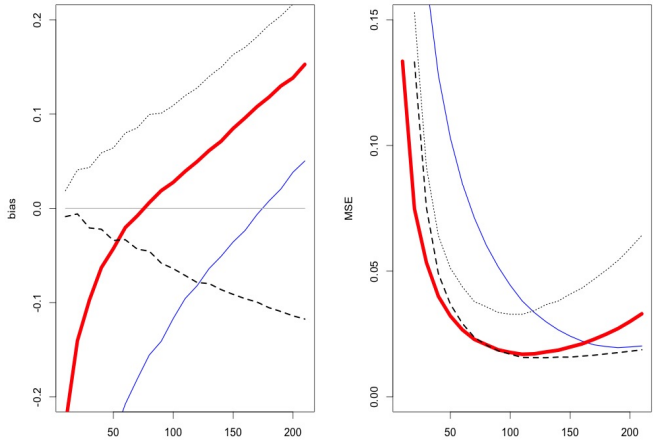


(b) Fréchet(1) censored by Fréchet(1/2)

Fig. 2 Comparison of bias and MSE for $\hat{\gamma}^{KM}_{X,Hill}$ (solid), $\hat{\gamma}^{Leurg}_{X,Hill}$ (thick) $\hat{\gamma}^{c}_{X,Hill}$ (dotted) and $\hat{\gamma}^{c}_{X,MVRB}$ (dashed) for a Fréchet distribution censored by another Fréchet distribution : (a) $\gamma_X = 1/2$ and $\gamma_C = 1$ (weak censoring), (b) $\gamma_X = 1$ and $\gamma_C = 1/2$ (strong censoring)



(a) Fréchet(1) censored by Burr$(10, 1, 1/2)$



(b) Fréchet(1) censored by Burr$(10, 1, 3/2)$

Fig. 3 Comparison of bias and MSE for $\hat{\gamma}^{KM}_{X,Hill}$ (solid), $\hat{\gamma}^{Leurg}_{X,Hill}$ (thick) $\hat{\gamma}^{c}_{X,Hill}$ (dotted) and $\hat{\gamma}^{c}_{X,MVRB}$ (dashed) for a Fréchet distribution censored by a Burr distribution : (a) $\gamma_X = 1$ and $\gamma_C = 2$ (weak censoring), (b) $\gamma_X = 1$ and $\gamma_C = 2/3$ (strong censoring)

4 Conclusion

In this paper, we have introduced two new approaches for the estimation of the extreme value index in the case of randomly censored observations, based on ideas coming from the censored regression literature. The estimation problem of the e.v.i. in the censoring framework had been addressed in very few papers before, and our methodology, though we have applied it to the adaptation of the Hill estimator in the heavy tail case only, has some potential for more applications, either for other estimators (moment based estimators on first place) and other maximum domains of attraction, or maybe for the estimation of other tail parameters. This work thus forms a basis for future research in this recently studied area of censored extremes, which can prove much useful in applications, as was showed in the review paper [Gomes and Neves (2011)]. For the moment, technical problems prevent us from obtaining asymptotic normality results and rigorous evaluation of the variance, but simulations show that our two versions of the Hill estimator perform quite well with respect to the existing version, in terms of MSE, even in the apparently less favorable case of heavy censoring in the tail ($\gamma_X > \gamma_C$).

5 Proofs

First, note that in several occasions in the next pages, reference will be made to Proposition 1 : it is stated in the Appendix.

5.1 Proof of Theorem 1

Since $(1-\hat{F}_n(t))(1-\hat{G}_n(t)) = k_n/n$ for $t = Z_{n-k_n,n}$, by introducing

$$A_n := \frac{1-\hat{G}_n(Z_{n-k_n,n})}{1-G(Z_{n-k_n,n})} \quad \text{and} \quad C_{in} := \frac{1-G(Z_{n-i+1,n})}{1-\hat{G}_n(Z^-_{n-i+1,n})}$$

we can write

$$\hat{\gamma}^{KM}_{X,Hill} = A_n \frac{1}{k_n}\sum_{i=1}^{k_n} C_{in} W_{in},$$

where

$$W_{in} := \delta_{n-i+1,n} \log\left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}\right) \frac{1-G(Z_{n-k_n,n})}{1-G(Z_{n-i+1,n})}.$$

Therefore, we have the decomposition $\gamma^{KM}_{X,Hill} = A_n(\overline{W}_n + R_n)$ where

$$\overline{W}_n := \frac{1}{k_n}\sum_{i=1}^{k_n} W_{in} \quad \text{and} \quad R_n := \frac{1}{k_n}\sum_{i=1}^{k_n}(C_{in}-1)W_{in}.$$

First of all, relying on Theorem 2 in [Csörgő (1996)], continuity of G entails that $A_n \xrightarrow{\mathbb{P}} 1$, as $n \to +\infty$ (please note that, unfortunately, this theorem

will not be sufficient for controlling the quantities C_{in}, because they involve really extreme observations ; in section 5.1.2 we show how this difficulty is circumvented).

Consequently, it remains to prove that $\overline{W}_n \xrightarrow{\mathbb{P}} \gamma_X$ and $R_n \xrightarrow{\mathbb{P}} 0$: this is the purpose of the next two subsections.

5.1.1 Proof of $\overline{W}_n \xrightarrow{\mathbb{P}} \gamma_X$

Let us first introduce the following notation, used throughout the rest of the proof,

$$\tilde{Z}_{i,n} := \frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}.$$

Under assumption (3) (which implies that $1-G \in RV_{-1/\gamma_C}$ and $1-H \in RV_{-1/\gamma}$), setting $\epsilon > 0$, we can apply Potter bounds (29) stated in Proposition 1, to the function $1/(1-G) \in RV_{\gamma_C^{-1}}$, and to $t = Z_{n-k_n,n} \xrightarrow{\mathbb{P}} +\infty$ and $x = \tilde{Z}_{i,n} \geq 1$. We thus obtain for n sufficiently large, the following bounds for W_{in} :

$$(1-\epsilon)\delta_{n-i+1,n}\log(\tilde{Z}_{i,n})(\tilde{Z}_{i,n})^{\gamma_C^{-1}-\epsilon} \leq W_{in} \leq (1+\epsilon)\delta_{n-i+1,n}\log(\tilde{Z}_{i,n})(\tilde{Z}_{i,n})^{\gamma_C^{-1}+\epsilon}. \tag{13}$$

Therefore, it remains to prove that both the mean of the lower bound, and that of the upper bound, converges in probability to a quantity arbitrary close to γ_X when ϵ is taken close to 0. We consider the case of the upper bound only, the lower bound being similar.

Recall that

$$p := \frac{\gamma_C}{\gamma_X + \gamma_C}$$

which is (see [Einmahl et. al. (2008)]) the limit of $p(z) = \mathbb{P}(\delta = 1 | Z = z)$ when $z \to \infty$. We intend to rely on the closeness of the $\delta_{n-i+1,n}$ to i.i.d. Bernoulli(p) random variables, independent of the log-spacings $\tilde{Z}_{i,n}$.

Mimicking what was proposed in [Einmahl et. al. (2008)], we use the fact that the original $(Z_i, \delta_i)_{i\leq n}$ are identically distributed as $(Z_i', \delta_i')_{i\leq n}$, where $\delta_i' = \mathbb{I}_{U_i \leq p(Z_i')}$ and $(U_i)_{i\leq n}$ denotes an i.i.d. sequence of standard uniform variables, independent of a given sequence $(Z_i')_{i\leq n}$ of i.i.d. variables having H as their c.d.f. We thus carry on the proof by considering now that δ_i is related to Z_i by

$$\delta_i = \mathbb{I}_{U_i \leq p(Z_i)}.$$

We then define (where the U_i below is the same as the one in the above definition of δ_i)

$$\tilde{\delta}_i = \mathbb{I}_{U_i \leq p}$$

which are Bernoulli(p) distributed and independent of the sequences (Z_i) and $(\tilde{Z}_{i,n})$. Note that we define $(\tilde{\delta}_{1,n}, \ldots, \tilde{\delta}_{n,n})$ as the rearrangements of the $\tilde{\delta}_i$

corresponding to the order induced by the order statistics $(Z_{1,n},\dots,Z_{n,n})$: these are however still independent of the sequences (Z_i) or $(\tilde{Z}_{i,n})$.

According to (13), it comes

$$\begin{aligned}(1+\epsilon)^{-1}\overline{W}_n \;\le\; & \frac{1}{k_n}\sum_{i=1}^{k_n}\tilde{\delta}_{n-i+1,n}\log(\tilde{Z}_{i,n})(\tilde{Z}_{i,n})^{\gamma_C^{-1}+\epsilon} \\ & + \frac{1}{k_n}\sum_{i=1}^{k_n}(\delta_{n-i+1,n}-\tilde{\delta}_{n-i+1,n})\log(\tilde{Z}_{i,n})(\tilde{Z}_{i,n})^{\gamma_C^{-1}+\epsilon} \\ =:\; & J_n^1 \;+\; J_n^2 \end{aligned} \tag{14}$$

(i) Let us first prove that $J_n^1 \xrightarrow{\mathbb{P}} g(\epsilon)$, with $\lim_{\epsilon\downarrow 0} g(\epsilon)=\gamma_X$.

Let $Y_{1,n},\dots,Y_{n,n}$ be the ascending order statistics of n i.i.d standard Pareto random variables $Y_1,\dots,Y_n$ with distribution function $1-1/x$, for $x>1$. By independence of the $\tilde{\delta}_{j,n}$ and the $\tilde{Z}_{i,n}$, if U is the quantile function associated to H (i.e. $U(t)=H^{\leftarrow}(1-1/t)$) then

$$\tilde{Z}_{i,n}\;\log(\tilde{Z}_{i,n}) \;\overset{d}{=}\; \frac{U(Y_{n-i+1,n})}{U(Y_{n-k_n,n})}\log\left(\frac{U(Y_{n-i+1,n})}{U(Y_{n-k_n,n})}\right). \tag{15}$$

Under assumption (3), U is regularly varying with index γ and Proposition 1 can be applied to U. Taking for each $1\le i\le k_n$, $t=Y_{n-k_n,n}$ and $x=Y_{n-i+1,n}/Y_{n-k_n,n}$, bounds (29) and their logarithm yield, for some given $\epsilon'>0$ and n sufficiently large,

$$(1-\epsilon')^{\gamma_C^{-1}+\epsilon}H_{in}^{\epsilon} \;\le\; \left(\frac{U(Y_{n-i+1,n})}{U(Y_{n-k_n,n})}\right)^{\gamma_C^{-1}+\epsilon}\log\left(\frac{U(Y_{n-i+1,n})}{U(Y_{n-k_n,n})}\right) \;\le\; (1+\epsilon')^{\gamma_C^{-1}+\epsilon}K_{in}^{\epsilon} \tag{16}$$

where , setting $\alpha^-:=(\gamma-\epsilon')(\gamma_C^{-1}+\epsilon)$ and $\alpha^+:=(\gamma+\epsilon')(\gamma_C^{-1}+\epsilon)$,

$$H_{in}^{\epsilon}=\left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)^{\alpha^-}\left(\log(1-\epsilon')+(\gamma-\epsilon')\log\left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)\right)$$

$$K_{in}^{\epsilon}=\left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)^{\alpha^+}\left(\log(1+\epsilon')+(\gamma+\epsilon')\log\left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)\right)$$

However, it is known that

$$(Y_{n-i+1,n}/Y_{n-k_n,n})_{1\le i\le k_n}\overset{d}{=}(\tilde{Y}_{i,k_n})_{1\le i\le k_n}, \tag{17}$$

where $\tilde{Y}_{1,k_n},\dots,\tilde{Y}_{k_n,k_n}$ are the ascending order statistics of k_n i.i.d random variables $\tilde{Y}_1,\dots,\tilde{Y}_{k_n}$ with standard Pareto distribution. Thanks to the independence of the $\tilde{\delta}_{j,n}$ and the Y_i, it follows that J_n^1 is bounded above by some variable which equals in distribution

$$\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{I}_{U_i\le p}\tilde{Y}_i^{\alpha^+}(\log(1+\epsilon')+(\gamma+\epsilon')\log(\tilde{Y}_i)). \tag{18}$$

Independence of the (U_i) and $(\tilde{Y}_i)$ and the law of large numbers then yields that (since $0 < \alpha^+ < 1$) $\limsup J_n^1$ is bounded above, in probability, by

$$p \times \left(\log(1+\epsilon') \frac{1}{1-\alpha^+} + (\gamma+\epsilon') \frac{1}{(1-\alpha^+)^2} \right).$$

Dealing now with the lower bound of (16), one can handle H_{in}^{ϵ} similarly and obtain a lower bound in probability for $\liminf J_n^1$: straightforward computations show that both bounds converge to γ_X as ϵ and ϵ' tend to 0. This thus concludes part (i).

(ii) It remains to prove that $J_n^2 \xrightarrow{\mathbb{P}} 0$.

Let us put $T_{n-i+1,n} := \log(\tilde{Z}_{i,n})(\tilde{Z}_{i,n})^{\gamma_C^{-1}+\epsilon}$ and recall that

$$J_n^2 = \frac{1}{k_n} \sum_{i=1}^{k_n} (\delta_{n-i+1,n} - \tilde{\delta}_{n-i+1,n}) T_{n-i+1,n}.$$

Let $p > 1$ and $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Hölder's Inequality gives :

$$\begin{aligned} |J_n^2| &\leq \left(\frac{1}{k_n} \sum_{i=1}^{k_n} \left| \delta_{n-i+1,n} - \tilde{\delta}_{n-i+1,n} \right|^q \right)^{1/q} \left(\frac{1}{k_n} \sum_{i=1}^{k_n} (T_{n-i+1,n})^p \right)^{1/p} \\ &= \left(\frac{1}{k_n} \sum_{i=1}^{k_n} \left| \delta_{n-i+1,n} - \tilde{\delta}_{n-i+1,n} \right| \right)^{1/q} \left(\frac{1}{k_n} \sum_{i=1}^{k_n} (T_{n-i+1,n})^p \right)^{1/p} \end{aligned}$$

So it remains to prove that $\frac{1}{k_n} \sum_{i=1}^{k_n} |\delta_{n-i+1,n} - \tilde{\delta}_{n-i+1,n}| = o_{\mathbb{P}}(1)$ and that $\frac{1}{k_n} \sum_{i=1}^{k_n} (T_{n-i+1,n})^p = O_{\mathbb{P}}(1)$, for an appropriate $p > 1$.

On one hand, according to (15), (16) and (17) (as in part (i)), we have to prove that

$$\frac{1}{k_n} \sum_{i=1}^{k_n} \left(\tilde{Y}_i^{\alpha^+} (\log(1+\epsilon') + (\gamma+\epsilon') \log(\tilde{Y}_i)) \right)^p = O_{\mathbb{P}}(1)$$

This is the case, using the law of large numbers, as soon as we take $p < 1 + \frac{\gamma_C}{\gamma_X}$ (in this case $\alpha^+ p < 1$, so $\mathbb{E}((\tilde{Y}_1^{\alpha^+} \log(\tilde{Y}_1))^p)$ is finite).

On the other hand, recall that $\delta_{i,n} = \mathbb{I}_{U_i \leq p(Z_i)}$ and $\tilde{\delta}_{i,n} = \mathbb{I}_{U_i \leq p}$. Then,

$$\begin{aligned} \frac{1}{k_n} \sum_{i=1}^{k_n} \left| \delta_{n-i+1,n} - \tilde{\delta}_{n-i+1,n} \right| \leq\ & \frac{1}{k_n} \sum_{i=1}^{k_n} \left| \mathbb{I}_{U_i \leq p(Z_{n-i+1,n})} - \mathbb{I}_{U_i \leq p(H^{\leftarrow}(1-\frac{i}{n}))} \right| \\ & + \frac{1}{k_n} \sum_{i=1}^{k_n} \left| \mathbb{I}_{U_i \leq p(H^{\leftarrow}(1-\frac{i}{n}))} - \mathbb{I}_{U_i \leq p} \right| \\ =:\ & T_{1,k} + T_{2,k}. \end{aligned}$$

Following the same lines as in [Einmahl et. al. (2008)] (p218) for the treatment of their terms $T_{1,k}$ and $T_{2,k}$ (which are different from ours in the rates but nevertheless quite similar), we show using a result of [Chow and Teicher (1997)] (p356), that both $T_{1,k}$ and $T_{2,k}$ tend to 0 in probability, thanks to assumptions (8) and (9). This concludes the proof of (ii) and therefore of $\bar{W}_n \xrightarrow{\mathbb{P}} \gamma_X$.

5.1.2 Proof of $R_n \xrightarrow{\mathbb{P}} 0$

As mentioned at the beginning of the proof, difficulties arise as to the control of the Kaplan-Meier estimate of G in the tail (here it takes the form of the variables $C_{i,n}$, not to be mistaken with the censoring variables C_i) : we will circumvent them via a device known in the survival analysis literature. Let us define, for some $\epsilon' > 0$,

$$\tilde{C}(t) := \int_0^t \frac{dG(x)}{(1-G(x))^2(1-F(x))} \quad \text{and} \quad h_{in} := (\tilde{C}(Z_{n-i+1,n}))^{-\frac{1}{2}-\epsilon'}.$$

We readily have $|R_n| \leq T_n^1 T_n^2$, where

$$T_n^1 := \sup_{1\leq i\leq k_n} \sqrt{n}|h_{in}(C_{in}-1)| \quad \text{and} \quad T_n^2 := \frac{1}{k_n}\sum_{i=1}^{k_n} W_{in} h_{in}^{-1} n^{-\frac{1}{2}}$$

and we are going to prove that $T_n^1 = O_{\mathbb{P}}(1)$ and $T_n^2 = o_{\mathbb{P}}(1)$.

First remind that $C_{i,n}$ is the value of the function $t \mapsto (1-G(t))/(1-\hat{G}_n(t^-))$ at $t = Z_{n-i+1,n}$, and consequently, by continuity of G and $\tilde{C}$

$$\begin{aligned} T_n^1 &\leq \sup_{t\leq Z_{n,n}} \left| \sqrt{n}(\tilde{C}(t))^{-\frac{1}{2}-\epsilon'} \frac{\hat{G}_n(t^-) - G(t)}{1-\hat{G}_n(t^-)} \right| \\ &\leq \sup_{t< Z_{n,n}} \left| \sqrt{n}(\tilde{C}(t))^{-\frac{1}{2}-\epsilon'} \frac{\hat{G}_n(t) - G(t)}{1-\hat{G}_n(t)} \right|. \end{aligned}$$

Since $\int_0^{+\infty} h^2(t) d\tilde{C}(t) < \infty$ for the function $h(t) = (\tilde{C}(t))^{-\frac{1}{2}-\epsilon'}$, Theorem 2.1 of [Gill (1983)] applies and therefore the process

$$\left(\sqrt{n}(\tilde{C}(t))^{-\frac{1}{2}-\epsilon'} \frac{\hat{G}(t) - G(t)}{1-\hat{G}(t)} \right)_{t<Z_{n,n}}$$

converges in distribution. As a consequence, $T_n^1 = O_{\mathbb{P}}(1)$.

It remains to prove that $T_n^2 = o_{\mathbb{P}}(1)$. First, from the definition of h_{in} and $\tilde{C}$, since $(1-H) = (1-F)(1-G)$ we clearly have

$$h_{in}^{-1} < \left(\frac{-\log(1-G(Z_{n-i+1,n}))}{1-H(Z_{n-i+1,n})} \right)^{\frac{1}{2}+\epsilon'}.$$

Moreover, according to (13), for some given $\epsilon > 0$, we have $T_n^2 \leq (1+\epsilon) P_n Q_n$ (n large) where

$$P_n := n^{-\frac{1}{2}} \left(\frac{-\log(1-G(Z_{n-k_n,n}))}{1-H(Z_{n-k_n,n})} \right)^{\frac{1}{2}+\epsilon'}$$

$$Q_n := \frac{1}{k_n} \sum_{i=1}^{k_n} \left(\frac{1-H(Z_{n-k_n,n})}{1-H(Z_{n-i+1,n})} \right)^{\frac{1}{2}+\epsilon'} \left(\frac{\log(1-G(Z_{n-i+1,n}))}{\log(1-G(Z_{n-k_n,n}))} \right)^{\frac{1}{2}+\epsilon'} \log(\tilde{Z}_{i,n}) \left(\tilde{Z}_{i,n} \right)^{\gamma_C^{-1}+\epsilon}$$

Under assumption (3), $1/(1-H)$ is regularly varying with index $1/\gamma$ and $-\log(1-G)$ is slowly varying : therefore, for some given $\epsilon'' > 0$ and n sufficiently large, the application of Proposition 1 to these functions yields the following upper bounds

$$\frac{1-H(Z_{n-k_n,n})}{1-H(Z_{n-i+1,n})} \leq (1+\epsilon'') \tilde{Z}_{i,n}^{\gamma^{-1}+\epsilon''}$$

$$\frac{\log(1-G(Z_{n-i+1,n}))}{\log(1-G(Z_{n-k_n,n}))} \leq (1+\epsilon'') \tilde{Z}_{i,n}^{\epsilon''}$$

and consequently

$$Q_n \leq (1+\epsilon'')^{1+2\epsilon'} \frac{1}{k_n} \sum_{i=1}^{k_n} \tilde{Z}_{i,n}^{\beta} \log(\tilde{Z}_{i,n}), \tag{19}$$

where $\beta = (2\gamma)^{-1} + \gamma_C^{-1} + \epsilon'''$, for some $\epsilon''' > 0$ (arbitrarily small when ϵ and ϵ' are closer to 0).

Concerning P_n, we have $(n/k_n)(1-H(Z_{n-k_n,n})) \xrightarrow{\mathbb{P}} 1$ and therefore

$$1-H(Z_{n-k_n,n}) = \frac{k_n}{n} U_n, \quad \text{and} \quad -\log(1-H(Z_{n-k_n,n})) = -\log\left(\frac{k_n}{n}\right) V_n$$

for some sequences U_n and V_n tending to 1, in probability, as $n \to +\infty$. Using the inequality $1-H \leq 1-G$, it follows that, for some $W_n \xrightarrow{\mathbb{P}} 1$,

$$P_n \;\leq\; n^{-\frac{1}{2}} \{-\log(k_n/n)/(k_n/n)\}^{\frac{1}{2}+\epsilon'} W_n \;=\; n^{\epsilon'} \{-\log(k_n/n)/k_n\}^{\frac{1}{2}+\epsilon'} W_n. \tag{20}$$

In order to prove that $P_n Q_n = o_{\mathbb{P}}(1)$, we have to distinguish the case $\gamma_X < \gamma_C$ from the case $\gamma_X \geq \gamma_C$ (respectively weak and strong censoring in the tail).

(i) Case $\gamma_X < \gamma_C$

Proceeding as in subsection 5.1.1 by using (15), (16) and (17), we obtain that the upper bound of Q_n in (19) is $O_{\mathbb{P}}(1)$ via the law of large numbers as soon as $\mathbb{E}(\tilde{Y_1}^{\beta(\gamma+\epsilon')} \log(\tilde{Y_1}))$ is finite, *i.e.* $\beta(\gamma+\epsilon') < 1$ (which is the case since $\gamma_X < \gamma_C$). This proves $Q_n = O_{\mathbb{P}}(1)$.

Finally, under assumption (10) on k_n, (20) implies that $P_n \xrightarrow{\mathbb{P}} 0$ as soon as $\epsilon' < \delta/(4-2\delta)$. This concludes the proof of $P_n Q_n = o_{\mathbb{P}}(1)$ in this case.

(*ii*) Case $\gamma_X \geq \gamma_C$

We use, as in (i), equations (15), (16) and (17) to treat Q_n . But, since $\beta(\gamma+\epsilon')$ is not < 1 but ≥ 1, the upper bound of Q_n in (19) is no longer $O_{\mathbb{P}}(1)$. In this case, we rely on the Marcinkiewicz-Zygmund law of large numbers to obtain

$$\frac{1}{k_n^q}\sum_{i=1}^{k_n} \tilde{Y}_i^{\beta(\gamma+\epsilon')}\log(\tilde{Y}_i) = o_{\mathbb{P}}(1),$$

where $q := \beta(\gamma+\epsilon') + \tilde{\delta}$, for some $\tilde{\delta} > 0$. This proves that $Q_n = o_{\mathbb{P}}(k_n^{q-1})$. It remains to prove that $k_n^{q-1}P_n = O_{\mathbb{P}}(1)$. From (20) and condition (10) on k_n, we have

$$k_n^{q-1}P_n \leq \left(\frac{k_n}{n}\right)^{q-1} \; n^{\epsilon'+q-1-\frac{\delta}{2}-\delta\epsilon'}.$$

So, ϵ' and $\tilde{\delta}$ being arbitrary close to 0, $k_n^{q-1}P_n = o_{\mathbb{P}}(1)$ as soon as $\delta \geq \frac{\gamma_X-\gamma_C}{\gamma_X+\gamma_C}$, thanks to assumption (5).
Note that if k_n is of the order of n^a, for some $a \in]0,1[$, the control of P_n does not require condition (10). □

5.2 Proof of Theorem 2

Similarly to $\hat{\gamma}_{X,Hill}^{KM}$, the estimator $\hat{\gamma}_{X,Hill}^{Leurg}$ can be written as follows :

$$\hat{\gamma}_{X,Hill}^{Leurg} = A_n(\overline{W}_n + R_n),$$

where A_n,C_{in} and R_n are defined as before (see Subsection 5.1) but now

$$W_{in} := \xi_{in}\frac{1-G(Z_{n-k_n,n})}{1-G(Z_{n-i+1,n})} \quad \text{and} \quad \xi_{in} := i\log\left(\frac{Z_{n-i+1,n}}{Z_{n-i,n}}\right).$$

Recall that $A_n \xrightarrow{\mathbb{P}} 1$ (see beginning of the proof of Theorem 1). In Section 5.2.2, we prove that $R_n \xrightarrow{\mathbb{P}} 0$. Let us first deal with $\overline{W}_n$.

5.2.1 Proof of $\overline{W}_n \xrightarrow{\mathbb{P}} \gamma_X$

Recall first the notation used in Subsection 5.1

$$\tilde{Z}_{i,n} := \frac{Z_{n-i+1,n}}{Z_{n-k_n,n}}.$$

Let $\eta > 0$ and $\epsilon > 0$. Using Potter bounds (29) for $1/(1-G)$, which is regularly varying of order γ_C^{-1}, from the definition of the W_{in} we first obtain for ϵ small enough

$$\mathbb{P}(\overline{W}_n - \gamma_X > \eta) \leq \mathbb{P}(k_n^{-1} \sum_{i=1}^{k_n} \tilde{Z}_{i,n}^{\gamma_C^{-1}+\epsilon} \xi_{in} - \gamma_X > \tfrac{\eta}{2}),$$
$$\mathbb{P}(\gamma_X - \overline{W}_n > \eta) \leq \mathbb{P}(\gamma_X - k_n^{-1} \sum_{i=1}^{k_n} \tilde{Z}_{i,n}^{\gamma_C^{-1}-\epsilon} \xi_{in} > \tfrac{\eta}{2}).$$

Let us now consider constants $c > 1$ and $c' < 1$, both close to 1, and $\alpha > 0$ and $\alpha' > 0$ both close to γ/γ_C, which values will be specified in the proof of Lemma 1 below. Using positivity of ξ_{in}, it comes

$$\mathbb{P}(\overline{W}_n - \gamma_X > \eta) \leq \mathbb{P}\left(\max_{i \leq k_n} \frac{\tilde{Z}_{i,n}^{\gamma_C^{-1}+\epsilon}}{c u_i^{-\alpha}} > 1 \right) + \mathbb{P}\left(c\ k_n^{-1} \sum_{i=1}^{k_n} u_i^{-\alpha} \xi_{in} - \gamma_X > \frac{\eta}{2} \right) \quad (21)$$

$$\mathbb{P}(\gamma_X - \overline{W}_n > \eta) \leq \mathbb{P}\left(\min_{i \leq k_n} \frac{\tilde{Z}_{i,n}^{\gamma_C^{-1}-\epsilon}}{c' u_i^{-\alpha'}} < 1 \right) + \mathbb{P}\left(\gamma_X - c'\ k_n^{-1} \sum_{i=1}^{k_n} u_i^{-\alpha'} \xi_{in} > \frac{\eta}{2} \right) \quad (22)$$

where $u_i = u_{in} := i/(k_n + 1)$, for $1 \leq i \leq k_n$.

Convergence in probability of $\overline{W}_n$ to γ_X thus comes from the combination of (21), (22) and the following two Lemmas, by letting ϵ go to 0 in the end : Lemma 1 is applied with $\theta = \gamma_C^{-1} + \epsilon$ and $\theta' = \gamma_C^{-1} - \epsilon$, whereas Lemma 2 is applied twice with $a = \alpha$ and $a = \alpha'$ (both close to γ/γ_C which is < 1), noticing that $\gamma/(1 - \gamma/\gamma_C)$ equals γ_X. □

Lemma 1 *Let θ and $\theta' > 0$. There exist constants $c > 1$, $c' < 1$ both arbitrarily close to 1, and $\alpha > 0$, $\alpha' > 0$, arbitrarily close to $\gamma\theta$ and to $\gamma\theta'$ respectively, such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i \leq k_n} \frac{\tilde{Z}_{i,n}^{\theta}}{c u_i^{-\alpha}} > 1 \right) = \lim_{n\to\infty} \mathbb{P}\left(\min_{i \leq k_n} \frac{\tilde{Z}_{i,n}^{\theta'}}{c' u_i^{-\alpha'}} < 1 \right) = 0.$$

Lemma 2 *If $0 < a < 1$, then*

$$\frac{1}{k_n} \sum_{i=1}^{k_n} u_i^{-a} \xi_{in} \ \xrightarrow{\mathbb{P}}\ \frac{\gamma}{1-a}.$$

If $a > 1$, then for any $\delta' > 0$,

$$\frac{1}{k_n^{a+\delta'}} \sum_{i=1}^{k_n} u_i^{-a} \xi_{in} \ \xrightarrow{\mathbb{P}}\ 0.$$

Lemmas 1 and 2 are proved one after the other below. Note that Lemma 2 is proved by using what could be called a first order version of the techniques used in [Beirlant et. al. (2007)] (where a second order approximation of the log-spacings $\xi_{i,n}$ is obtained).

Proof of Lemma 1

Once again (as in Subsection 5.1.1), we introduce $Y_{1,n}, \ldots, Y_{n,n}$ the ascending order statistics of n i.i.d standard Pareto random variables, in order to have

$$\tilde{Z}_{i,n} \stackrel{d}{=} \frac{U(Y_{n-i+1,n})}{U(Y_{n-k_n,n})}.$$

Applying Potter bounds (29) to $U \in RV_\gamma$, it comes, for some given $\epsilon' > 0$, and n large enough,

$$(U(Y_{n-i+1,n})/U(Y_{n-k_n,n}))^\theta \leq (1+\epsilon')^\theta \left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)^\alpha$$
$$(U(Y_{n-i+1,n})/U(Y_{n-k_n,n}))^{\theta'} \geq (1-\epsilon')^{\theta'} \left(\frac{Y_{n-i+1,n}}{Y_{n-k_n,n}}\right)^{\alpha'}$$

where

$$\alpha = (\gamma+\epsilon')\theta \quad \text{and} \quad \alpha' = (\gamma-\epsilon')\theta'.$$

Using (17) and introducing $V_{1,k_n} \leq \cdots \leq V_{k_n,k_n}$ the order statistics of k_n i.i.d Uniform $[0,1]$ random variables (which have the same distribution as $1/Y_1$), it comes

$$\mathbb{P}\left(\max_{i\leq k_n} \frac{\tilde{Z}_{i,n}^\theta}{c u_i^{-\alpha}} > 1\right) \leq \mathbb{P}\left((1+\epsilon')^\theta \max_{i\leq k_n}(V_{i,k_n}/u_i)^{-\alpha} > c\right),$$
$$\mathbb{P}\left(\min_{i\leq k_n} \frac{\tilde{Z}_{i,n}^{\theta'}}{c' u_i^{-\alpha'}} < 1\right) \leq \mathbb{P}\left((1-\epsilon')^{\theta'} \min_{i\leq k_n}(V_{i,k_n}/u_i)^{-\alpha'} < c'\right).$$

Relying on $\max_{1\leq i\leq k_n} |V_{i,k_n} - u_i| \stackrel{\mathbb{P}}{\longrightarrow} 0$ (uniform consistency of the uniform empirical quantile process), we readily have, for any given $\beta > 1$,

$$\lim_{n\to\infty} \mathbb{P}(E_n) = 1 \quad \text{where} \quad E_n := \{\forall\, 1 \leq i \leq k_n\ ,\ \beta^{-1}u_i \leq V_{i,k_n} \leq \beta u_i\} \quad (23)$$

Therefore, for given constants $\epsilon' > 0$ and $\beta > 1$, if we set

$$c = \beta^\alpha (1+\epsilon')^\theta \quad \text{and} \quad c' = \beta^{-\alpha}(1-\epsilon')^{\theta'},$$

then both probabilities appearing in Lemma 1 are bounded above by $1-\mathbb{P}(E_n)$, which achieves the proof in view of (23). As announced, by choosing appropriate values of ϵ' and β, the constants c, c', α, α', are respectively arbitrarily close to 1, 1, $\gamma\theta$ and $\gamma\theta'$. □

Proof of Lemma 2

We proceed very similarly as in [Beirlant et. al. (2002)], therefore some details will be ommited. Let $E_1, \ldots, E_n$ be i.i.d. $\mathcal{E}xp(1)$ random variables. Then $\tilde{Z}_{i,n} \stackrel{d}{=} U(\exp(E_{n-i+1,n}))/U(\exp(E_{n-i,n}))$. The first order Potter-type bounds for $U \in RV_\gamma$ stated in Proposition 1 thus yield : for some given $\epsilon > 0$, n large enough and $1 \leq i \leq k_n$,

$$\begin{aligned} \xi_{i,n} \;=\; i \log \frac{Z_{n-i+1,n}}{Z_{n-i,n}} &\stackrel{d}{=} \gamma\, i(E_{n-i+1,n} - E_{n-i,n}) + i(B_{k,n}(i) - B_{k,n}(i+1)) \\ &\stackrel{d}{=:} \gamma\, \xi_i \;+\; \beta_{i,n} \end{aligned} \tag{24}$$

where

$$\log(1-\epsilon) - \epsilon(E_{n-i+1,n} - E_{n-k_n,n}) \leq B_{k,n}(i) \leq \log(1+\epsilon) + \epsilon(E_{n-i+1,n} - E_{n-k_n,n})$$

and the Rényi representation was used to derive (24), with $\xi_1, \ldots, \xi_{k_n}$ denoting independent $\mathcal{E}xp(1)$ variables.

Using the fact that $(E_{n-i+1,n} - E_{n-k_n,n})_{1 \leq i \leq k_n} \stackrel{d}{=} (-\log V_{i,k_n})_{1 \leq i \leq k_n}$, where $V_1, \ldots, V_{k_n}$ are independent standard uniform variables, and using relation (23), we obtain (this is in fact the first order version of Theorem 2.1 in [Beirlant et. al. (2002)]), uniformly in $i \in \{1, ..., k_n\}$,

$$\left| \textstyle\sum_{j=i}^{k_n} \frac{\beta_{j,n}}{j} \right| \;=\; |B_{k,n}(i)| \;=\; o_{\mathbb{P}}(\log_+(1/u_i)) \tag{25}$$

(with $\log_+(x) = \max(1, \log x)$ and recalling that u_i stands for $i/(k+1)$).
Regular application of the law of large numbers for triangular arrays of independent random variables yields (cf [Chow and Teicher (1997)] ; details are ommited), when $0 < a < 1$

$$\frac{1}{k_n} \sum_{i=1}^{k_n} u_i^{-a} \xi_i \stackrel{\mathbb{P}}{\longrightarrow} \frac{1}{1-a}$$

and, when $a > 1$ and $\delta' > 0$ is arbitrary small,

$$\frac{1}{k_n^{a+\delta'}} \sum_{i=1}^{k_n} u_i^{-a} \xi_i \stackrel{\mathbb{P}}{\longrightarrow} 0.$$

Therefore, according to (24), Lemma 2 is proved as soon as we have

$$\frac{1}{k_n} \sum_{i=1}^{k_n} u_i^{-a} \beta_{i,n} \stackrel{\mathbb{P}}{\longrightarrow} 0 \quad \text{when } \; 0 < a < 1, \tag{26}$$

$$\frac{1}{k_n^{a+\delta'}} \sum_{i=1}^{k_n} u_i^{-a} \beta_{i,n} \stackrel{\mathbb{P}}{\longrightarrow} 0 \quad \text{when } \; a > 1. \tag{27}$$

Suppose first that $0 < a < 1$. The trick to show the negligibility (26) is to write (where $u_0 = 0$)

$$\begin{aligned}
\left|\tfrac{1}{k_n+1}\sum_{i=1}^{k_n} u_i^{-a}\beta_{i,n}\right| &= \left|\sum_{i=1}^{k_n} \tfrac{\beta_{i,n}}{i} u_i^{1-a}\, ds\right| \\
&= (1-a)\left|\sum_{i=1}^{k_n} \tfrac{\beta_{i,n}}{i}\int_0^{u_i} s^{-a}\, ds\right| \\
&\leq (1-a)\sum_{j=1}^{k_n}\left|\sum_{i=j}^{k_n}\tfrac{\beta_{i,n}}{i}\right|\int_{u_{j-1}}^{u_j} s^{-a}\, ds \\
&= o_{\mathbb{P}}(1)\sum_{j=1}^{k_n}\log_+(1/u_j)(u_j^{1-a}-u_{j-1}^{1-a}) \\
&\leq o_{\mathbb{P}}(1)\tfrac{1}{k_n}\sum_{j=1}^{k_n}\log_+(1/u_j)u_j^{-a} \\
&= o_{\mathbb{P}}(1).
\end{aligned}$$

Suppose now that $a > 1$ and let us prove similarly (27) : we need to be more cautious since $\int_0^u s^{-a}ds$ is no longer defined. We have

$$\begin{aligned}
&\left|\tfrac{1}{k_n+1}\sum_{i=1}^{k_n} u_i^{-a}\beta_{i,n}\right| \\
&= |1-a|\left|\sum_{i=1}^{k_n}\tfrac{\beta_{i,n}}{i}\left(\int_{u_1}^{u_i} s^{-a}\, ds \;+\; u_1^{1-a}\right)\right| \\
&\leq |1-a|\left|\sum_{i=2}^{k_n}\left(\tfrac{\beta_{i,n}}{i}\sum_{j=2}^{i}\int_{u_{j-1}}^{u_j} s^{-a}\, ds\right)\right| \;+\; |1-a|(k_n+1)^{a-1}\left|\sum_{i=1}^{k_n}\tfrac{\beta_{i,n}}{i}\right| \\
&\leq o_{\mathbb{P}}(1)\left(\tfrac{1}{k_n}\sum_{j=2}^{k_n}\log_+(1/u_j)u_j^{-a} \;+\; k_n^{a-1}\log(k_n+1)\right)
\end{aligned}$$

hence

$$\begin{aligned}
\left|\tfrac{1}{k_n^{a+\delta'}}\sum_{i=1}^{k_n} u_i^{-a}\beta_{i,n}\right| &\leq o_{\mathbb{P}}(1)\left(\tfrac{1}{k_n^{a+\delta'}}\sum_{j=1}^{k_n} u_j^{-a-\delta'} \;+\; k_n^{-\delta'}\log(k_n+1)\right) \\
&= o_{\mathbb{P}}(1).
\end{aligned}$$

□

5.2.2 Proof of $R_n \xrightarrow{\mathbb{P}} 0$

Most of the proof is identical to the case of the first theorem. As in Subsection 5.1.2, we have $|R_n| \leq T_n^1 T_n^2$ where T_n^1 is left unchanged and is $O_{\mathbb{P}}(1)$. The factor T_n^2 is bounded by $(1+\epsilon)P_nQ_n$, for some $\epsilon > 0$, where P_n is defined as in 5.1.2 but, in the definition of Q_n, the factor $\log \tilde{Z}_{i,n}$ needs to be replaced by ξ_{in}. The same arguments as before allow us to write, for some given $\epsilon, \epsilon' > 0$ and n sufficiently large,

$$Q_n \leq (1+\epsilon'')^{(1+2\epsilon')}\frac{1}{k_n}\sum_{i=1}^{k_n}\tilde{Z}_{i,n}^{\beta}\xi_{in},$$

where, as in Subsection 5.1.2, $\beta = (2\gamma)^{-1} + \gamma_C^{-1} + \epsilon'''$, for some $\epsilon''' > 0$ (as small as needed when ϵ and ϵ' are set close to 0). We now proceed as in Subsection 5.2.1 to control T_n^2 and end the proof of $R_n = o_{\mathbb{P}}(1)$.

Let $\eta > 0$ and consider constants $c > 1$ close to 1 and $\alpha > 0$ close to $\frac{1}{2} + \frac{\gamma}{\gamma_C}$.

$$\mathbb{P}(T_n^2 > \eta) \leq \mathbb{P}\left(\max_{i \leq k_n} \frac{\tilde{Z}_{i,n}^{\beta}}{c u_i^{-\alpha}} > 1\right) + \mathbb{P}\left(P_n \; k_n^{-1} \sum_{i=1}^{k_n} u_i^{-\alpha} \xi_{in} > \frac{\eta}{c}\right) \tag{28}$$

The fact that $T_n^2 = o_{\mathbb{P}}(1)$, then, comes from the combination of (28) and Lemmas 1 and 2, with delails given below.

First, Lemma 1 is applied with $\theta = \beta$, which implies that $\alpha = (\gamma + \epsilon')\beta$ and thus the first term of the right hand-side of (28) tends to 0.

Next, Lemma 2 is applied with $a = \alpha$ and needs to be combined with some rate for the factor P_n : as in subsection 5.1.2, we need here to distinguish the case $\gamma_X < \gamma_C$ from the case $\gamma_X \geq \gamma_C$.

(i) case $\gamma_X < \gamma_C$

In this case, $\alpha < 1$ and therefore Lemma 2 implies that $\frac{1}{k_n} \sum_{i=1}^{k_n} u_i^{-\alpha} \xi_{in} = O_{\mathbb{P}}(1)$. Moreover, we have already proved in subsection 5.1.2 that $P_n = o_{\mathbb{P}}(1)$, and consequently the second term of the right hand-side of (28) tends to 0.

(ii) case $\gamma_X \geq \gamma_C$

In this case, $\alpha > 1$ and therefore Lemma 2 implies that, for any given $\delta' > 0$, $\frac{1}{k_n^{\alpha+\delta'}} \sum_{i=1}^{k_n} u_i^{-\alpha} \xi_{in} = o_{\mathbb{P}}(1)$. Therefore, it remains to prove that $k_n^{\alpha+\delta'-1} P_n = O_{\mathbb{P}}(1)$: this has already been done in subsection 5.1.2 (where, there, $1/p$ denoted $\beta(\gamma + \epsilon') + \tilde{\delta}$ where $\tilde{\delta}$ was arbitrary close to 0, as is δ' here). □

6 Appendix

Definition 1 *An ultimately positive function $f : \mathbb{R}^+ \to \mathbb{R}$ is regularly varying (at infinity) of order $\alpha \in \mathbb{R}$, if*

$$\lim_{t \to +\infty} \frac{f(tx)}{f(t)} = x^{\alpha}, \; x > 0.$$

This is noted $f \in RV_\alpha$. If $\alpha = 0$, f is said to be slowly varying.

Proposition 1 *(See [Haan and Ferreira (2006)] Proposition B.1.9) Suppose $f \in RV_\alpha$. If $x > 0$ and $\delta_1, \delta_2 > 0$ are given, then there exists $t_0 = t_0(\delta_1, \delta_2)$ such that for $t \geq t_0$ and $tx \geq t_0$,*

$$(1 - \delta_1) x^{\alpha} \min(x^{\delta_2}, x^{-\delta_2}) < \frac{f(tx)}{f(t)} < (1 + \delta_1) x^{\alpha} \max(x^{\delta_2}, x^{-\delta_2}).$$

If $x \geq 1$ and $\epsilon > 0$, then there exists $t_0 = t_0(\epsilon)$ such that for every $t \geq t_0$,

$$(1 - \epsilon) x^{\alpha - \epsilon} < \frac{f(tx)}{f(t)} < (1 + \epsilon) x^{\alpha + \epsilon}. \tag{29}$$

References

[Beirlant et. al. (2002)] J. Beirlant, G. Dierckx, A. Guillou and C. Stărică . On exponential representations of log spacings of order statistics. In *Extremes* **5**, pages 157-180 (2002)

[Beirlant et. al. (2007)] J. Beirlant, G. Dierckx, A. Guillou and A. Fils-Villetard . Estimation of the extreme value index and extreme quantiles under random censoring. In *Extremes* **10**, pages 151-174 (2007)

[Beirlant et. al. (2010)] J. Beirlant, A. Guillou and G. Toulemonde . Peaks-Over-Threshold modeling under random censoring. In *Comm. Stat. : Theory and Methods* **39**, pages 1158-1179 (2010)

[Brahimi et. al. (2013)] B. Brahimi, D. Meraghni and A. Necir . Asymptotic normality of the adapted Hill estimator to censored data. working paper (2013)

[Caeiro et. al. (2005)] F. Caeiro, M.I. Gomes and D. Pestana . Direct reduction of bias of the classical Hill estimator. In *Revstat* **3/2**, pages 113-136 (2005)

[Chow and Teicher (1997)] Y.S. Chow and H. Teicher . Probability theory. Independence, interchangeability, martingales. *Springer* (1997)

[Csörgő (1996)] S. Csörgő . Universal Gaussian approximations under random censorship. In *Annals of statistics* **24 (6)**, pages 2744-2778 (1996)

[Haan and Ferreira (2006)] L. de Haan and A. Ferreira . Extreme Value Theory : an Introduction. *Springer Science + Business Media* (2006)

[Delecroix et. al. (2008)] M. Delecroix, O. Lopez and V. Patilea . Nonlinear censored regression using synthetic data. In *Scandinavian J. of Statistics* **35**, pages 248-265 (2008)

[Einmahl et. al. (2008)] J. Einmahl, A. Fils-Villetard and A. Guillou . Statistics of extremes under random censoring. In *Bernoulli* **14**, pages 207-227 (2008)

[Gill (1983)] R. Gill . Large sample behaviour of the product-limit estimator on the whole line. In *Annals of statistics* **11 (1)**, pages 49-58 (1983)

[Gomes et. al. (2008)] M.I. Gomes, L. de Haan and L. Henriques Rodrigues. Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. In *J. Royal Statistical Society* **B70:1**, pages 31-52 (2008)

[Gomes and Neves (2011)] M.I. Gomes and M.M. Neves . Estimation of the extreme value index for randomly censored data. In *Biometrical Letters* **48 (1)**, pages 1-22 (2011)

[Koul et. al. (1981)] H. Koul, V. Susarla and J. Van Ryzin . Regression analysis with randomly right-censored data. In *Annals of statistics* **9 (6)**, pages 1276-1288 (1981)

[Leurgans (1987)] S. Leurgans . Linear models, random censoring and synthetic data. In *Biometrika* **74**, pages 301-309 (1987)

[Stute (1995)] W. Stute . The central limit theorem under random censorship. In *Annals of statistics* **23 (2)**, pages 422-439 (1995)