Computational Statistics and Data Analysis 57 (2013) 518–535

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A simple generalisation of the Hill estimator

M. Fátima Brilhante [a], M. Ivette Gomes [b,*], Dinis Pestana [b]

[a] CEAUL and DM, Universidade dos Açores, Portugal

[b] CEAUL and DEIO, FCUL, Universidade de Lisboa, Portugal

ARTICLE INFO

Article history:
Received 15 March 2012
Received in revised form 21 July 2012
Accepted 22 July 2012
Available online 24 July 2012

Keywords:
Bias estimation
Bootstrap methodology
Heavy tails
Optimal levels
Semi-parametric estimation
Statistics of extremes

ABSTRACT

The classical Hill estimator of a positive extreme value index (EVI) can be regarded as the logarithm of the geometric mean, or equivalently the logarithm of the mean of order $p = 0$, of a set of adequate statistics. A simple generalisation of the Hill estimator is now proposed, considering a more general mean of order $p \geq 0$ of the same statistics. Apart from the derivation of the asymptotic behaviour of this new class of EVI-estimators, an asymptotic comparison, at optimal levels, of the members of such class and other known EVI-estimators is undertaken. An algorithm for an adaptive estimation of the tuning parameters under play is also provided. A large-scale simulation study and an application to simulated and real data are developed.

1. The new class of estimators and scope of the paper

Let us consider a sample of size n of independent, identically distributed (i.i.d.) random variables (r.v.'s), $X_1, \ldots, X_n$, with a common distribution function (d.f.) F. Let us denote by $X_{1:n} \leq \cdots \leq X_{n:n}$ the associated ascending order statistics (o.s.'s) and let us assume that there exist sequences of real constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that the maximum, linearly normalised, i.e. $(X_{n:n} - b_n)/a_n$, converges in distribution to a non-degenerate r.v. Then, the limit distribution is necessarily of the type of the general *extreme value* (EV) d.f., given by

$$\mathrm{EV}_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & 1+\gamma x > 0 \text{ if } \gamma \neq 0 \\ \exp(-\exp(-x)), & x \in \mathbb{R} \text{ if } \gamma = 0. \end{cases} \quad (1)$$

The d.f. F is said to belong to the max-*domain of attraction* of EV_γ, and we use the notation $F \in \mathcal{D}_\mathcal{M}(\mathrm{EV}_\gamma)$. The parameter γ is the *extreme value index* (EVI), the primary parameter of extreme events.

Let us denote by RV_a the class of regularly varying functions at infinity, with an index of regular variation equal to $a \in \mathbb{R}$, i.e. positive measurable functions $g(\cdot)$ such that for all $x > 0$, $g(tx)/g(t) \to x^a$, as $t \to \infty$ (see Bingham et al., 1987). The EVI measures the heaviness of the right *tail function*

$$\overline{F}(x) := 1 - F(x),$$

and the heavier the right tail, the larger γ is. In this paper we work with Pareto-type underlying d.f.'s, with a positive EVI, or equivalently, models such that $\overline{F}(x) = x^{-1/\gamma}L(x)$, $\gamma > 0$, with $L \in \mathrm{RV}_0$, a slowly varying function at infinity, i.e. a regularly varying function with an index of regular variation equal to zero. These heavy-tailed models are quite common in many

* Correspondence to: Departamento de Estatística e Investigação Operacional, Faculdade de Ciências de Lisboa, Bloco C6, Piso 4, 1749-016 Lisboa, Portugal. Tel.: +351 217500041.

E-mail addresses: fbrilhante@uac.pt (M. Fátima Brilhante), ivette.gomes@fc.ul.pt (M. Ivette Gomes), dinis.pestana@fc.ul.pt (D. Pestana).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.07.019

areas of application, like computer science, telecommunications, insurance, finance, bibliometrics and biostatistics, among others.

For Pareto-type models, the classical EVI-estimators are the Hill estimators (Hill, 1975), which are the averages of the log-excesses, given by

$$V_{ik} := \ln \frac{X_{n-i+1:n}}{X_{n-k:n}}, \quad 1 \le i \le k < n. \tag{2}$$

We thus have

$$\widehat{\gamma}_n^H(k) \equiv H(k) := \frac{1}{k}\sum_{i=1}^{k} V_{ik}, \quad 1 \le k < n. \tag{3}$$

Note that with $F^{\leftarrow}(x) := \inf\{y : F(y) \ge x\}$ denoting the generalised inverse function of F, and

$$U(t) := F^{\leftarrow}(1 - 1/t), \quad t \ge 1,$$

the reciprocal quantile function, we can write the distributional identity $X = U(Y)$, with Y a unit Pareto r.v., i.e. a r.v. with d.f. $F_Y(y) = 1 - 1/y$, $y \ge 1$. For the o.s.'s associated with a random Pareto sample $(Y_1, \ldots, Y_n)$, we have the distributional identity $Y_{n-i+1:n}/Y_{n-k:n} = Y_{k-i+1:k}$, $1 \le i \le k$. Moreover, $kY_{n-k:n}/n \underset{n\to\infty}{\overset{p}{\longrightarrow}} 1$, i.e. $Y_{n-k:n} \overset{p}{\sim} n/k$. Consequently, and provided that $k = k_n$, $1 \le k < n$, is an intermediate sequence of integers, i.e. if

$$k = k_n \to \infty \quad \text{and} \quad k_n = o(n), \text{ as } n \to \infty, \tag{4}$$

we get

$$U_{ik} := \frac{X_{n-i+1:n}}{X_{n-k:n}} = \frac{U(Y_{n-i+1:n})}{U(Y_{n-k:n})} = \frac{U(Y_{n-k:n}Y_{k-i+1:k})}{U(Y_{n-k:n})} = Y_{k-i+1:k}^{\gamma}(1 + o_p(1)), \tag{5}$$

i.e. $U_{ik} \overset{p}{\sim} Y_{k-i+1:k}^{\gamma}$. Hence, we have the approximation $\ln U_{ik} \approx \gamma \ln Y_{k-i+1:k} = \gamma E_{k-i+1:k}$, $1 \le i \le k$, with E denoting a standard exponential r.v. The log-excesses, $V_{ik} = \ln U_{ik}$, $1 \le i \le k$, in (2), are thus approximately the k top o.s.'s of a sample of size k from an exponential parent with mean value γ. This justifies the Hill EVI-estimator, in (3).

We can write

$$H(k) = \sum_{i=1}^{k} \ln \left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^{1/k} = \ln \left(\prod_{i=1}^{k} \frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^{1/k}, \quad 1 \le i \le k < n,$$

the logarithm of the *geometric mean* of the statistics U_{ik}, given in (5). More generally, we now consider as basic statistics for the EVI estimation, the *mean of order p* (MOP) of U_{ik}, i.e. the class of statistics

$$A_p(k) = \begin{cases} \left(\dfrac{1}{k}\displaystyle\sum_{i=1}^{k} U_{ik}^p\right)^{1/p} & \text{if } p > 0 \\ \left(\displaystyle\prod_{i=1}^{k} U_{ik}\right)^{1/k} & \text{if } p = 0. \end{cases} \tag{6}$$

From (5), we can write $U_{ik}^p = Y_{k-i+1:k}^{\gamma p}(1 + o_p(1))$. Since

$$\mathbb{E}(Y^a) = \frac{1}{1-a} \quad \text{if } a < 1, \tag{7}$$

the law of large numbers enables us to say that if $p < 1/\gamma$,

$$A_p(k) \underset{n\to\infty}{\overset{p}{\longrightarrow}} \left(\frac{1}{1-\gamma p}\right)^{1/p}, \quad \text{i.e. } \frac{1 - A_p^{-p}(k)}{p} \underset{n\to\infty}{\overset{p}{\longrightarrow}} \gamma.$$

Hence the reason for the new class of MOP EVI-estimators,

$$\widehat{\gamma}_n^{H_p}(k) \equiv H_p(k) := \begin{cases} \left(1 - A_p^{-p}(k)\right)/p & \text{if } p > 0 \\ \ln A_0(k) = H(k) & \text{if } p = 0, \end{cases} \tag{8}$$

with $A_p(k)$ given in (6), and with $H_0(k) \equiv H(k)$, given in (3). This class of MOP EVI-estimators depends on this tuning parameter $p \ge 0$, which makes it very flexible, and even able to overpass one of the simplest and one of the most efficient EVI-estimators in the literature, the corrected-Hill (CH) estimator in Caeiro et al. (2005), to be introduced in Section 2.2.