The Role of Bootstrap Methodologies in the Estimation of a Negative Extreme Value Index

Frederico Caeiro[1] and M. Ivette Gomes[2]
[1] *CMA and FCT, Universidade Nova de Lisboa, PORTUGAL*
[2] *CEAUL and DEIO, FCUL, Universidade de Lisboa, PORTUGAL*
Corresponding author: Frederico Caeiro, e-mail: fac@fct.unl.pt

Abstract: In this note we deal with the estimation, under a semi-parametric framework, of a negative extreme value index, the primary parameter in statistics of extremes. We consider a recent class of generalized negative moment estimators of a negative extreme value index. Apart from the usual integer parameter k, related to the number of top order statistics involved in the estimation, the estimator depend on an extra real parameter θ, which makes it highly flexible and possibly second-order unbiased for a large variety of models. We are interested on the study of the bootstrap method in Gomes *et al.* (2013) for the adaptive choice of the parameters.

Key Words: Bootstrap methods, semi-parametric estimation, statistics of extremes.

1 Introduction

One of the main results in extreme value theory is the possible limiting laws of maximum values, $X_{n:n} := \max(X_1, X_2, \ldots, X_n)$, of either independent, identically distributed random variables (r.v.'s) or possibly weakly dependent and stationary from a model F. We know that if the maximum $X_{n:n}$, linearly normalized, converges to a non-degenerate r.v., then there exist real constants $\{a_n\}_{n\geq 1}$ $(a_n > 0)$ and $\{b_n\}_{n\geq 1}$, the so-called *attraction coefficients* of F, such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = EV_\gamma(x),$$

for some $\gamma \in \mathbb{R}$, with $EV_\gamma(x)$ given by

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

We then say that F is in the *domain of attraction* (for maxima) of EV_γ and we use the notation $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$. The parameter γ is the *extreme value index* (EVI) and measures the heaviness of the right *tail function* $\overline{F} := 1 - F$. The heavier the tail, the larger the EVI is. The EVI is one of the basis of other parameters of extreme and large events, like a *high quantile* of probability $1-p$, with p small, the *right endpoint* of the model F underlying the data, $x^F := \sup\{x : F(x) < 1\}$, whenever finite, and the *return period* of a high level, among others.

We will work with the $k+1$ top o.s.'s associated to the n available observations, assuming only that, for a certain $\gamma < 0$, the model F underlying the data is in $\mathcal{D}_{\mathcal{M}}(G_\gamma)$. Most of the classical semi-parametric estimators of any parameter of extreme events have a strong bias for moderate up to large values of k, including the optimal k, in the sense of minimal mean squared error (MSE). Accommodation of bias of classical estimators of parameters of extreme events has been deeply considered in the recent literature. For the

negative EVI-estimation ($\gamma < 0$), we refer the recent *negative moment* estimator (Caeiro and Gomes, 2010),

$$\hat{\gamma}_{k,n}^{NM(\theta)} := \frac{1}{2}\left\{1 - \left(\frac{M_{k,n}^{(2)}}{(M_{k,n}^{(1)})^2} - 1\right)^{-1}\right\} + \theta\, M_{k,n}^{(1)}, \quad \theta \in \mathbb{R}. \tag{2}$$

with

$$M_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\left\{\ln X_{n-i+1:n} - \ln X_{n-k:n}\right\}^j, \quad j \geq 1, \quad X_{n-k:n} > 0,$$

and $X_{i:n}$ denotes the i-th ascending order statistic.

Apart from the usual integer parameter k, related to the number of top order statistics involved in the estimation, the estimator depend on an extra real parameter θ, which makes it flexible and possibly second-order unbiased for a large variety of models in $\mathcal{D}_{\mathcal{M}}(EV_\gamma)_{\gamma<0}$. Indeed, for a negative EVI, adequate conditions on k and θ (see Caeiro and Gomes 2010, for details), and with $\mathcal{N}(\mu, \sigma^2)$ denoting a normal r.v. with mean value μ and variance σ^2, we get a null bias, even for moderate values of k, i.e.,

$$\sqrt{k}(\hat{\gamma}_{k,n}^{NM(\theta)} - \gamma) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0,\ \sigma_{NM}^2 = \frac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}\right).$$

In this paper, we are interested on the adaptive choice of the *tuning* parameters k and θ. We will study, computationally, the bootstrap method in Gomes *et al.* (2013) for the adaptive choice of such parameters.

2 Adaptive selection of the tuning parameters

The adaptive selection of θ and k was already adressed in Gomes *et al.* (2013). We shall next present the algorithm which is based on the auxiliary statistic

$$\begin{aligned} T_{k,n}(\theta) &:= \gamma_{[k/2],n}^{NM(\theta)} - \gamma_{k,n}^{NM(\theta)} = \left(\hat{\gamma}_{[k/2],n}^{NM(0)} - \hat{\gamma}_{k,n}^{NM(0)}\right) + \theta\left(M_{[k/2],n}^{(1)} - M_{k,n}^{(1)}\right) \\ &=: r_k + \theta s_k, \qquad k = 2, \ldots, n-1, \end{aligned} \tag{3}$$

where $[x]$ is the integer part of x.

2.1 Adaptive selection of the tuning parameter θ

The stability of $T_{k,n}(\theta)$ around zero for moderate values of k, say $k \in [k_1, k_2]$, with $k_1 := [n^{0.25}] + 1$ and $k_2 := [n^{0.95}]$, enable us to choose

$$\hat{\theta} \equiv \hat{\theta}(k_1, k_2) := \arg\min_{\theta} \sum_{k=k_1}^{k_2} (r_k + \theta s_k)^2 = -\sum_{k=k_1}^{k_2} r_k s_k \Big/ \sum_{k=k_1}^{k_2} s_k^2, \tag{4}$$

where r_k and s_k have been defined in (3).

2.2 Adaptive selection of k

The choice of k for the EVI-estimation is next done on the basis of the bootstrap methodology, in a way similar to the one in Danielson *et al.* (2001), Draisma *et al.* (1999), Gomes and Oliveira (2001) and more recently in Gomes *et al.* (2012), and it is written algorithmically in the following steps:

1. Compute $\hat{\theta}$ defined in (4).

2. Next, consider sub-sample sizes $n_1 = o(n)$ and $n_2 = [n_1^2/n] + 1$.

3. For l from 1 until B (for example $B = 250$), generate independently B bootstrap samples $(x_1^*, \ldots, x_{n_2}^*)$ and $(x_1^*, \ldots, x_{n_2}^*, x_{n_2+1}^*, \ldots, x_{n_1}^*)$, of sizes n_2 and n_1, respectively, from the empirical d.f., $F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}$, associated with the observed sample $(x_1, \ldots, x_n)$.

4. Denoting $T_{k,n}^*(\hat{\theta})$ the bootstrap counterpart of $T_{k,n}(\hat{\theta})$, with $T_{k,n}(\theta)$ defined in (3), obtain $(t_{k,n_1,l}^*,\ 2 < k < n_1 - 1)$, $(t_{k,n_2,l}^*,\ 2 < k < n_2 - 1)$, $1 \le l \le B$, the observed values of the statistics T_{k,n_i}^*, $i = 1, 2$. For $k = 2, \ldots, n_i - 1$, compute

$$MSE^*(n_i, k) = \frac{1}{B}\sum_{l=1}^{B} \left(t_{k,n_i,l}^*\right)^2, \quad i = 1, 2.$$

and obtain

$$\hat{k}_{0|T}^*(n_i) := \underset{2<k<n_i-1}{\arg\min}\, MSE^*(n_i, k), \quad i = 1, 2. \tag{5}$$

5. For the estimation of the second-order parameter ρ (see [2], for details), consider the bootstrap estimator given by

$$\hat{\rho}^* := \ln \hat{k}_{0|T}^*(n_1) / \left(2\ln(\hat{k}_{0|T}^*(n_1)/n_1)\right). \tag{6}$$

6. Compute the threshold estimate

$$\hat{k}_0^* \equiv \hat{k}_0^*(n; n_1) := \left[(1 - 2^{\hat{\rho}^*})^{2/(1-2\hat{\rho}^*)} \left(\hat{k}_{0|T}^*(n_1)\right)^2 / \hat{k}_{0|T}^*(n_1^2/n)\right] + 1, \tag{7}$$

with $\hat{k}_{0|T}^*(n_i)$ and $\hat{\rho}^*$ given in (5) and (6), respectively (see equation (29) and Section 4. of [6] for the theoretical details). If $\hat{k}_0^* > n - 1$ then go back to STEP 3.

7. Obtain $\hat{\gamma}^* \equiv \hat{\gamma}^*(n; n_1) := \hat{\gamma}_{\hat{k}_0(n;n_1),n}^{NM(\hat{\theta})}$, with $\hat{\gamma}_{k,n}^{NM(\theta)}$, $\hat{\theta}$ and $\hat{k}_0(n; n_1)$ given in (2), (4) and (7), respectively.

3 Study of the Adaptive Algorithm

3.1 Aplication to a simulated dataset

We shall next consider a simulated sample of size $n = 6000$ from the $EV_\gamma \equiv EV(-\gamma)$ model, in eq. (1), with $\gamma = -0.5$. Due to the nature of the estimators, we can only use the 3801 positive values of the sample. Figure 1 (left) illustrates, for $\theta = 0$, 1, 1.5 and 2, the behaviour of $\hat{\gamma}_{k,n}^{NM(\theta)}$ as function of k. Notice that the value θ has a big influence on the sample path of $\hat{\gamma}_{k,n}^{NM(\theta)}$. The application of the adaptive choice of θ, in (4), with $k_1 = [n^{0.25}] + 1 = 8$ and $k_2 = [n^{0.95}] = 2517$, led us to $\hat{\theta} = 1.55$. In Figure 1 (right), we picture the values of $\hat{\theta}(k_1, k_2)$, as function of k_2, with $k_1 \le k_2 \le n - 1$. Notice the resistance of the method to the choice of k_2.

The use of the Algorithm, in Sect. 2.2, with $n_1 = [3801^{0.9}] = 1667$ and $B = 250$ led us to $\hat{k}_0^* = 2376$ and to the adaptive EVI-estimate $\hat{\gamma}^* = -0.5129$, which is very close to the target value $\gamma = -0.5$.

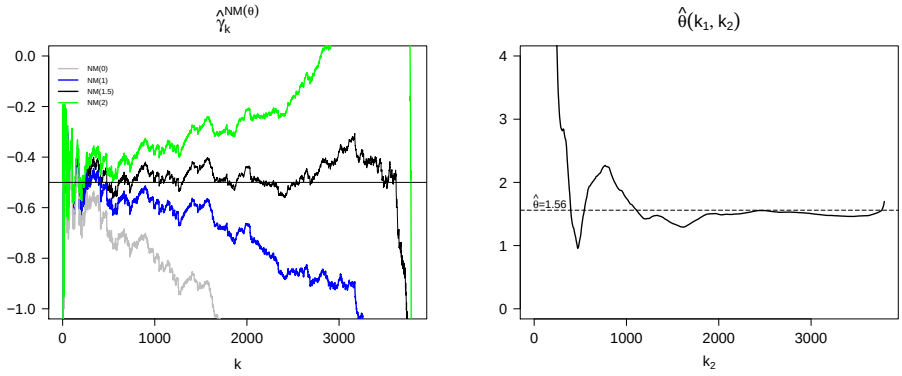

Figure 1: Estimates of the EVI (left) provided by the estimator under consideration, for the simulated EV(-0.5) sample and estimates $\hat{\theta}(k_1; k_2)$ (*left*), as function of k_2, $k_1 \leq k_2 \leq n-1$.

3.2 Sensitivity of the Algorithm to the Choice of B

Working with the previous sample, we shall now study the sensitivity of the Algorithm in Sect. 2.2 to the choice of the number of bootstrap samples B. A study to the sensitivity to the choice of n_1 can be found in Gomes *et al.* (2013). As an illustration, we aplied the algorithm 100 times to the same EV(-0.5) simulated data set with $n_1 = [3801^{0.95}] = 2517$, $B = 250$, 1000 and 5000. Figure 2 presents, in a box and whisker plot, the results of the bootstrap estimates of the optimal sample fractions, $\hat{k}_0^*/n$, (left) and of the EVI (right).

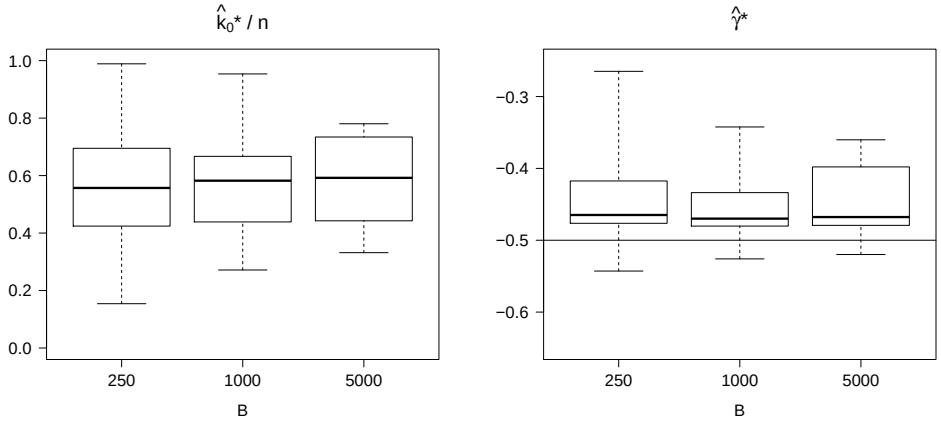

Figure 2: Bootstrap estimates of the optimal sample fractions (left) and of the EVI (right), for the simulated EV(-0.5) sample.

We can draw the following conclusions:

- Regarding the optimal sample fractions, and for an underlying EV model, with $\gamma = -0.5$, the variability decreases as we increase the value of B. Also, with $B = 250$ we can get bootstrap estimates of the optimal sample fraction very close to 1, which is a region where the asymptotic bias should be larger.

- Regarding the EVI bootstrap estimates, we have a clear improvement in the results only when we go from $B = 250$ to $B = 1000$. Also, the bootstrap EVI-estimates with $B = 5000$ have a much larger interquartile range than in the other two cases ($B = 250$ and $B = 1000$).

- To improve the precision of the EVI estimates it is advisable to apply the Algorithm m times and choose the estimate $\hat{\gamma}^*$ that corresponds to the median of the

m bootstrap EVI estimates.

3.3 A small-scale Monte Carlo simulation study of the Algorithm

Here we are interested in the distributional properties of the Algorithm in Sect. 2.2 for finite sample sizes. The study is based on a multi-sample Monte Carlo simulation with 400 runs for the following underlying parents:

- the EV model in (1) with $\gamma = -0.5$ and samples of size $n = 2500$, 6000 and 15000.
- the *generalized Pareto* (GP) distribution with d.f. $GP_\gamma(x) = 1 + \ln EV_\gamma(x)$, $1 + \gamma x > 0,\ x > 0$. We have choosed $\gamma = -0.5$ and samples of size $n = 2000$, 5000 and 10000.

For each sample, we have applied the Algorithm $m = 50$ times with $B = 250$ and $n_1 = [n^{0.9}]$. Then we choosed the estimate $\hat{\gamma}^*$ that corresponds to the median of the $m = 50$ bootstrap EVI estimates. We have simulated the mean value and the mean squared error of the EVI bootstrap estimator. We present, in Figs. 3 and 4, a box and whisker plot with the estimates $\hat{\theta}$ (left) and $\hat{\gamma}^*$ (right) of the 400 samples from the EV and GP model, respectively. In Table 1, we present the simulated mean values and root mean squared error of the EVI estimator, computed through the Algorithm in Sect. 2.2.

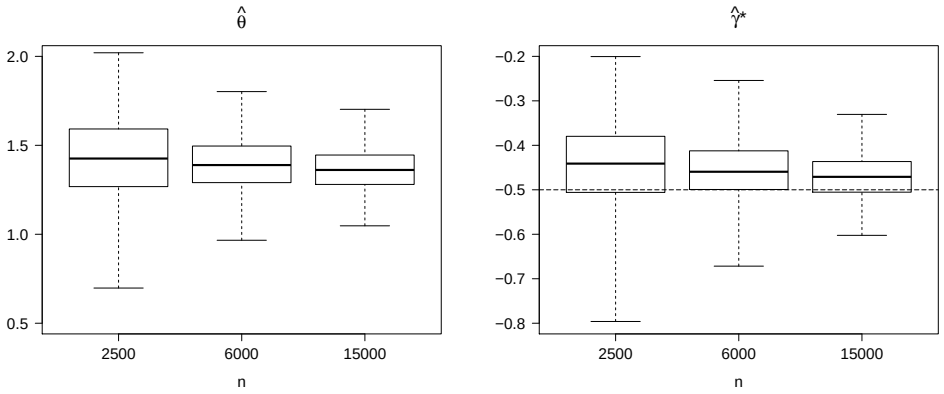

Figure 3: Simulated values of $\hat{\theta}$ in (4) (left) and Bootstrap estimates of the EVI (right), for the 400 samples from the EV model with $\gamma = -0.5$.

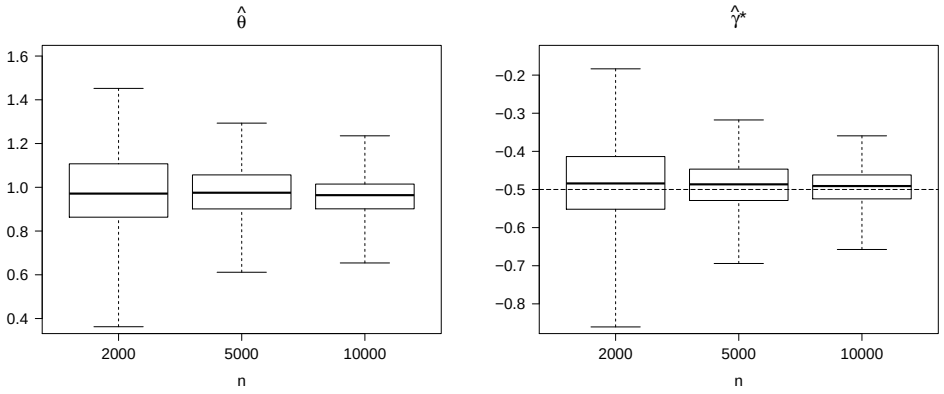

Figure 4: Simulated values of $\hat{\theta}$ in (4) (left) and Bootstrap estimates of the EVI (right), for the 400 samples from the GP model with $\gamma = -0.5$.

Table 1: Simulated mean values / root mean squared errors.

	n =2500	n =6000	n =15000
$EV_{-0.5}$	-0.4476 / 0.1094	-0.4590 / 0.0788	-0.4696 / 0.0572
	n =2000	n =5000	n =10000
$GP_{-0.5}$	-0.4840 / 0.1012	-0.4873 / 0.0646	-0.4929 / 0.0475

Some conclusions:

- As expected, the precision of the Algorithm improves as the sample size increases.
- The volatility of the bootstrap estimates is very high when we have samples with a few thousand observations.
- For the $EV_{-0.5}$ parents, we tend to overestimate γ. This could be related to the estimation of the tuning parameter θ or with the choice of the value n_1 in the Algorithm.

Acknowledgements. Research partially supported by National Funds through FCT– Fundação para a Ciência e a Tecnologia, projects PEst-OE /MAT /UI0006 /2011 (CEAUL), PEst-OE/MAT /UI0297/2011 (CMA/UNL) and EXTREMA, PTDC/MAT/101736/2008.

References

[1] Caeiro, F. and Gomes, M.I. (2010). An asymptotically unbiased moment estimator of a negative extreme value index. Discussiones Mathematica: Probability and Statistics **30**(1), 5–19.

[2] Danielsson, J., Haan, L. de, Peng, L., Vries, C.G. de (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. Journal of Multivariate Analysis **76**, 226–248.

[3] Draisma, G., Haan, L. de, Peng, L., Themido Pereira, T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme value index. Extremes **2**(4), 367–404.

[4] Gomes, M.I., Figueiredo, F. and Neves, M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action. Extremes **15**, 463–489.

[5] Gomes, M.I., Henriques-Rodrigues, L. and Caeiro, F. (2013). Refined Estimation of a Light Tail: an Application to Environmental Data. Accepted in Torelli, N., Pesarin, F., Bar-Hen, A. (Eds.), Advances in Theoretical and Applied Statistics, Springer.

[6] Gomes, M.I., Oliveira, O. (2001). The bootstrap methodology in Statistics of Extremes — choice of the optimal sample fraction. Extremes **4**(4), 331–358.