Abstract: We consider a class of semi-parametric estimators of a negative or null extreme value index, parameterized in a tuning or control parameter. Such tuning parameter enables us to have access to an estimator with a null dominant component of asymptotic bias, and we are able to achieve a high efficiency relatively to other classical estimators. In this work, we compare three adaptive choices of the tuning parameter through a Monte Carlo simulation study.

Key Words: Extreme value index, semi-parametric estimation, bias reduction.
AMS: 62G20, 62G32.

1 Introduction

Let us consider the common set-up of independent, identically distributed (i.i.d.) random variables (r.v.’s) \(X_1, X_2, \ldots, X_n\), with a common distribution function (d.f.) \(F\) and denote the associated ascending order statistics (o.s.) by \(X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}\). Let us assume that there exist sequences of real constants \(\{a_n > 0\}\) and \(\{b_n \in \mathbb{R}\}\) such that the normalized maximum, i.e. \((X_{n:n} - b_n)/a_n\), converges in distribution towards a non-degenerate r.v.. Then \(F\) belongs to the max-domain of attraction of the extreme value (EV) d.f., \(EV_\gamma(x) := \exp\{-(1 + \gamma x)^{-1/\gamma}\}, 1 + \gamma x > 0, \gamma \in \mathbb{R}\), and we write \(F \in \mathcal{D}_M(EV_\gamma)\). A necessary and sufficient condition for \(F \in \mathcal{D}_M(G)\) is (de Haan, 1984):

\[
\lim_{t \to \infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \quad \forall x > 0,
\]

for some measurable positive function \(a(t)\) and with \(U(t)\) standing for the quantil function defined by \(U(t) := F^*(1 - 1/t), t \geq 1\) with \(F^*(x) := \inf\{y : F(y) \geq x\}\).

The parameter \(\gamma\) is the extreme value index (EVI), a measure of the heaviness of the right tail function \(F := 1 - F\), and the heavier the tail, the larger \(\gamma\) is. The EVI needs to be estimated in a precise way, because such an estimation is one of the basis for the estimation 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 the model \(F\) underlying the data is in \(\mathcal{D}_M(G_\gamma)\), for a certain \(\gamma \leq 0\). Most of the classical semi-parametric EVI-estimators 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). To improve the performance of classical EVI-estimators we have to deal with bias reduction techniques (see Beirlant et al., 2012 and references within for more details). For the negative or null EVI-estimation, we refer the recent estimator in Caeiro and Gomes (2010),

\[ z_{NM}^{k,n}(\theta) := \frac{1}{2} \left\{ 1 - \left( \frac{M_{[k/2],n}^{(2)}}{M_{k,n}^{(1)}} \right)^{2} - 1 \right\} + \theta M_{k,n}^{(1)}, \quad \theta \in \mathbb{R}. \tag{2} \]

with

\[ M_{k,n}^{(j)} := \frac{1}{k} \sum_{i=1}^{k} \{ \ln X_{n-i+1:n} - \ln X_{n-k:n} \}^{j}, \quad j \geq 1. \]

Apart from the usual integer parameter \( k \), related with the number of top order statistics involved in the estimation, the estimator depend on an extra tuning parameter \( \theta \), which makes it flexible and possibly second-order unbiased for a large variety of models in \( D_{M}(EV_{\gamma}, \gamma < 0) \). To derive the asymptotic behaviour of the EVI-estimator, we shall assume the following second order condition:

\[ \lim_{t \to \infty} \frac{U(tx) - U(t) - \frac{\gamma - 1}{\gamma} t^{\gamma - 1}}{A(t)\gamma} = \frac{1}{\rho} \left( \frac{x^{\gamma+\rho} - 1 - x^{\gamma-1}}{\gamma} \right), \]

for all \( x > 0 \), where \( \rho \leq 0 \) is a second order parameter controlling the speed of convergence of the first order condition in (1) and \( |A(t)| \in RV_{\rho} \), and \( RV_{\rho} \) stands for the class of regularly varying functions with index of regular variation \( \rho \), i.e. positive measurable functions \( g \) such that \( \lim_{t \to \infty} g(tx)/g(t) = x^{\rho} \), for all \( x > 0 \).

Under the second order condition with \( \gamma \leq 0 \), adequate conditions on \( k \) and \( \theta \) (see Caeiro and Gomes, 2010, for more details), and with \( \mathcal{N}(\mu, \sigma^{2}) \) denoting a normal r.v. with mean value \( \mu \) and variance \( \sigma^{2} \), we get a null bias, even for moderate values of \( k \), i.e.,

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

If \( \theta \) is not appropriated chosen, then the mean value in eq. (3) will be non-null.

In this work, we are interested on the adaptive choice of the tuning parameter \( \theta \) in the semi-parametric estimation of the EVI. We shall study and compare the performance of several adaptive choices of \( \theta \) through a Monte Carlo simulation study.

## 2 Adaptive selection of the tuning parameter

For the adaptive selection of \( \theta \), we shall consider the same auxiliary statistic used in Gomes et al. (2013),

\[ T_{k,n}(\theta) := \gamma_{[k/2],n}^{NM(\theta)} - \gamma_{k,n}^{NM(\theta)} = (\hat{\gamma}_{[k/2],n}^{NM(\theta)} - \hat{\gamma}_{k,n}^{NM(\theta)}) + \theta (M_{[k/2],n}^{(1)} - M_{k,n}^{(1)}) \]

\[ =: r_{k} + \theta s_{k}, \quad k = 2, \ldots, n - 1, \tag{4} \]

where \( [x] \) is the integer part of \( x \). The stability of \( T_{k,n}(\theta) \) around zero for moderate values of \( k \), say \( k \in [k_{1}, k_{2}] \), with \( 2 \leq k_{1} < k_{2} \leq n - 1 \), let Gomes et al. (2013) to chose \( \theta \) as the value that minimizes the sum of squared bias of \( T \), i.e.,

\[ \hat{\theta} \equiv \hat{\theta}(k_{1}, k_{2}) := \arg \min_{\theta} \sum_{k=k_{1}}^{k_{2}} (r_{k} + \theta s_{k})^{2} = - \sum_{k=k_{1}}^{k_{2}} r_{k}s_{k}/ \sum_{k=k_{1}}^{k_{2}} s_{k}^{2}, \tag{5} \]
where $r_k$ and $s_k$ have been defined in (4). Since $r_k$ and $s_k$, and consequently $\hat{\theta}$, can be affected by an asymptotic bias, other adaptive choices of $\theta$ can be considered. Here we shall introduce two new adaptative choices of $\theta$. The first is the adaptive choice $\tilde{\theta}$ given by the least absolute value

$$
\tilde{\theta} \equiv \tilde{\theta}(k_1, k_2) := \arg \min_{\theta} \sum_{k=k_1}^{k_2} |r_k + \theta s_k|.
$$

(6)

Notice that if $k_2 > k_1$, the solution of eq. (6) can only be achieved thought a numerical method. The other adaptative choice of $\theta$, introduced in this work, is a statistic consistent in probability to the value $\theta_0$ such that the auxiliary statistic $T_{k,n}(\theta_0)$, in eq. (4), is asymptotic unbiased. This value is given by

$$
\tilde{\theta} \equiv \tilde{\theta}(k_1, k_2) := - \sum_{k=k_1}^{k_2} r_k / \sum_{k=k_1}^{k_2} s_k.
$$

(7)

Remark: Since $|r_k|$ can take very high values if $k$ is very small, we advise the choice

$$
k_1 = [n^{0.5}] + 1 \quad \text{and} \quad k_2 = [n^{0.95}].
$$

(8)

3 Comparative simulation study of the adaptive choices of the tuning parameter

We are now interested in the comparative behaviour of the adaptive choices of $\theta$ given in (5), (6) and (7), in the estimation of the EVI through the estimator in (2), for finite sample sizes. The study is based on a multi-sample Monte Carlo simulation with 5000 runs, with $k_1$ and $k_2$ given in (8), for samples of size $n = 1000$ and $n = 5000$, and for the following underlying parents:

- Arcsin model with d.f.

$$
F(x) = \frac{2}{\pi} \arcsin(\sqrt{x}), \quad 0 < x < 1 \quad (\gamma = -2);
$$

- Generalized Pareto (GP) distribution with d.f.

$$
GP_\gamma(x) = 1 + \ln \text{EV}_\gamma(x), \quad 1 + \gamma x > 0, \quad x > 0
$$

and $\gamma = -0.25$.

We also considered the choices $\theta = 0.45$ and $\theta = 0.9$, for the Arcsin and GP models, respectively. Those values where obtained from previous simulation studies (Caeiro and Gomes, 2010; Gomes et al., 2013). We present, in Figures 1 and 2, the simulated mean value (E) and root mean squared error (RMSE), as function of $k$, for an underlying Arcsin parents. Figures 3 and 4 have the same simulated quantities, as function of $k$, for $\text{GP}_{-0.25}$ parents.

The results here presented allow us to present some conclusions:

- As expected, the precision of the Algorithm improves as the sample size increases.

- For the same sample size, the simulated mean values have larger bias for the GP parents than for Arcsin parents.
Figure 1: Patterns of simulated mean values of $\hat{\gamma}_{k,n}^{NM(\theta)}$, as functions of $k$, for an underlying Arcsin parent and $n = 1000$ (left), $n = 5000$ (right).

Figure 2: Patterns of simulated RMSE of $\hat{\gamma}_{k,n}^{NM(\theta)}$, as functions of $k$, for an underlying Arcsin parent and $n = 1000$ (left), $n = 5000$ (right).

Figure 3: Patterns of simulated mean values (E) of $\hat{\gamma}_{k,n}^{NM(\theta)}$, as functions of $k$, for an underlying $GP_{-0.25}$ parent and $n = 1000$ (left), $n = 5000$ (right).
If we compare the RMSE, which is an important measure of the precision of the EVI-estimator, we have usually

$$RMSE(\hat{\gamma}^{NM}_{k,n}) < RMSE(\hat{\gamma}^{NM}_{k,n}) < RMSE(\hat{\gamma}^{NM}_{k,n}),$$

for $k$ not very small, neither very high.

**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**


