

Refined Estimation of a Light Tail: an Application to Environmental Data

M. Ivette Gomes, Lúgia Henriques-Rodrigues and Frederico Caeiro

Abstract In this paper, we consider a recent class of *generalized negative moment* estimators of a negative extreme value index, the primary parameter in *statistics of extremes*. Apart from the usual integer parameter k , related with the number of top order statistics involved in the estimation, these estimators depend on an extra real parameter, which makes them highly flexible and possibly second-order unbiased for a large variety of models. In this paper, we are interested not only on the adaptive choice of the *tuning* parameters, but also on an application of these semi-parametric estimators to the analysis of sets of environmental and simulated data.

Key words: statistics of extremes, semi-parametric estimation, bootstrap methods.

1 Introduction and outline of the paper

Whenever interested in extreme large events, the most common models in *statistics of extremes* are semi-parametric or even non-parametric in nature, with the imposition of a few “regularity conditions” in the right-tail, $\bar{F}(x) := 1 - F(x)$, as $x \rightarrow +\infty$, of an unknown model F underlying the available data. The primordial parameter is the *extreme value index* (EVI). For large values, the EVI is the shape parameter γ in the distribution function (d.f.)

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

M. Ivette Gomes
CEAUL and FCUL, Campo Grande, 1749-016, Lisboa, e-mail: ivette.gomes@fc.ul.pt

Lúgia Henriques-Rodrigues
CEAUL and IPT, Tomar, Portugal, e-mail: Ligia.Rodrigues@aim.estt.ipt.pt

Frederico Caeiro
CMA and Universidade Nova de Lisboa, e-mail: fac@fct.unl.pt

the (unified) *extreme value* (EV) distribution.

The parameter γ 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.

In this paper we are interested in an application of *statistics of extremes* to a set of environmental data. Due to the specificity of the data, we pay special attention to the estimation of the *extreme value index* γ , in (1), based on a quite recent estimator of a negative EVI, the *generalized negative moment* estimator ([1]).

In Section 2, we present some preliminary results in *extreme value theory* (EVT). In Section 3, we introduce the EVI-estimators under consideration, giving special emphasis to the *generalized negative moment* estimator. In Section 4, we provide an *Algorithm* for the adaptive choice of the *tuning* parameters under play in the semi-parametric estimation of the EVI through such an estimator. Finally, in Section 5, we apply the *Algorithm*, in Section 4, to the analysis of a set of environmental data, the daily average wind speeds in knots (one nautical mile per hour), collected in Dublin airport, in the period 1961-1978, as well as to a set of simulated data. Due to the seasonality of *Wind* data, we restrict ourselves to the Spring and Summer months.

2 Preliminary results in EVT

One of the main results in EVT is related with the possible limiting laws, as $n \rightarrow \infty$, of the sequence $X_{n:n} := \max(X_1, X_2, \dots, X_n)$, of maximum values of either independent, identically distributed random variables (r.v.'s) or possibly weakly dependent and stationary from a model F . In order to obtain a possible non-degenerate behaviour for $X_{n:n}$, we need to normalize it. Similarly to the *central limit theorem* for sums or means, 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

$$\lim_{n \rightarrow \infty} \mathbb{P}((X_{n:n} - b_n)/a_n \leq x) = EV_\gamma(x), \quad (2)$$

for some $\gamma \in \mathbb{R}$, with $EV_\gamma(x)$ given in (1) ([6]; [8]). We then say that F is in the *domain of attraction* (for maxima) of EV_γ and we use the notation $F \in \mathcal{D}_M(EV_\gamma)$.

The *extreme value index* γ , in (1), measures essentially the weight of the right-tail $\bar{F} = 1 - F$. If $\gamma < 0$, the right tail is *light*, i.e., F has a finite right endpoint ($x^F < +\infty$). If $\gamma > 0$, the right-tail is *heavy*, of a *negative polynomial* type, i.e., F has an infinite right endpoint. If $\gamma = 0$, the right tail is of an *exponential* type and the right endpoint can be either finite or infinite.

Let us define

$$U(t) := F^{\leftarrow}(1 - 1/t) \quad (t > 1), \quad \text{with} \quad F^{\leftarrow}(x) := \inf\{y : F(y) \geq x\} \quad (3)$$

denoting the generalized inverse function of F . It is possible to prove (see [9] and [10]) that, with $U(\cdot)$ defined in (3), and for all $x > 0$,

$$F \in \mathcal{DM}(EV_\gamma) \iff \lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ \ln x & \text{if } \gamma = 0. \end{cases} \quad (4)$$

We have $a(t) \equiv a_t \equiv a_{[t]}$, $[t]$ = integer part of t , with a_n the scale attraction coefficient in (2). Moreover, the location attraction coefficient b_n , also in (2), comes similarly from the relation $b_t = U(t)$, with $U(\cdot)$ defined in (3).

In order to derive the asymptotic behaviour of semi-parametric EVI-estimators, we need to assume, for all $x > 0$, the validity of a second-order condition, like

$$\lim_{t \rightarrow \infty} \frac{\frac{U(tx) - U(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} = H_{\gamma, \rho}(x) := \frac{1}{\rho} \left(\frac{x^{\gamma+\rho} - 1}{\gamma + \rho} - \frac{x^\gamma - 1}{\gamma} \right), \quad (5)$$

where $\rho \leq 0$ is a second order parameter controlling the speed of convergence in the first-order condition in (4) and $|A(t)| \in RV_\rho$, with RV_α standing for the class of regularly varying functions at infinity with an index of regular variation α , i.e. positive measurable functions g such that $\lim_{t \rightarrow \infty} g(tx)/g(t) = x^\alpha$, for all $x > 0$.

3 Semi-parametric estimation of some parameters of rare events

On the basis of the available random sample, $(X_1, \dots, X_n)$, and the sample $(X_{1:n} \leq \dots \leq X_{n:n})$ of associated ascending order statistics (o.s.'s), let us see how to estimate the extreme value index γ , the primordial parameter in *statistics of extremes*, and the right endpoint x^F . For any integer $j \geq 1$, let us denote

$$M_{k,n}^{(j)} := \frac{1}{k} \sum_{i=1}^k \{\ln X_{n-i+1:n} - \ln X_{n-k:n}\}^j. \quad (6)$$

For the estimation of γ , we shall first refer one estimator, valid for all $\gamma \in \mathbb{R}$, the *moment* (M) estimator ([3]), with the functional form

$$\hat{\gamma}_{k,n}^M := M_{k,n}^{(1)} + \frac{1}{2} \left\{ 1 - \left(M_{k,n}^{(2)} / [M_{k,n}^{(1)}]^2 - 1 \right)^{-1} \right\}, \quad (7)$$

$M_{k,n}^{(j)}$, $j = 1, 2$ defined in (6). Note that the statistic $M_{k,n}^{(1)}$, in (6), which plays also a main role in the *moment* estimator, in (7), is the well-known *Hill* estimator ([11]), often denoted $\hat{\gamma}_{k,n}^H \equiv M_{k,n}^{(1)}$, the average of the log-excesses and valid, i.e. consistent, only for $\gamma \geq 0$. We can thus write the *moment* estimator in (7), as

$$\hat{\gamma}_{k,n}^M = \hat{\gamma}_{k,n}^H + \hat{\gamma}_{k,n}^{NM}, \quad \hat{\gamma}_{k,n}^{NM} := \frac{1}{2} \left\{ 1 - \left(M_{k,n}^{(2)} / [M_{k,n}^{(1)}]^2 - 1 \right)^{-1} \right\}, \quad (8)$$

with NM standing for *negative moment* estimator. Indeed, whereas the H-estimator, $\hat{\gamma}_{k,n}^H$, is consistent for the estimation of $\gamma_+ := \max(0, \gamma)$ in the whole $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$, the NM-estimator, in (8), is consistent for the estimation of $\gamma_- := \min(0, \gamma)$ in the whole $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$, provided that $k = k_n$ is an intermediate sequence, i.e., a sequence of integers such that $k = k_n \rightarrow \infty$ and $k_n = o(n)$, as $n \rightarrow \infty$. Under these same conditions, the M-estimator, in (7), is consistent for the estimation of any real γ , which can be written as $\gamma = \gamma_+ + \gamma_-$.

Most of the classical EVI-estimators have usually a high variance for small values of k and a high bias when k is large. This problem affects both the *moment* and the *Hill* estimators, and leads to a difficult choice of the “optimal” k , in the sense of the value k that minimizes the mean squared error (MSE). On the basis of a comment in [5], where it is noticed that when $\gamma > 0$, $\hat{\gamma}_{k,n}^H$ has a smaller asymptotic variance than $\hat{\gamma}_{k,n}^M$, and when $\gamma < 0$, $\hat{\gamma}_{k,n}^{NM}$ and $\hat{\gamma}_{k,n}^M$ have the same asymptotic variance, the authors in [1] were led to the introduction of a semi-parametric class of consistent estimators for $\gamma < 0$, which generalizes both the M-estimator, in (7), and the NM-estimator, in (8). Such a class, denoted $NM(\theta)$, is given by

$$\hat{\gamma}_{k,n}^{NM(\theta)} := \hat{\gamma}_{k,n}^{NM} + \theta M_{k,n}^{(1)}, \quad \theta \in \mathbb{R}. \quad (9)$$

The dependence on the tuning parameter $\theta \in \mathbb{R}$ makes such a class highly flexible, even able to become second-order reduced-bias for a great variety of underlying models in $\mathcal{D}_{\mathcal{M}}(EV_\gamma)_{\gamma < 0}$. Note that we get the *negative moment* estimator in (8) for $\theta = 0$ and the *moment estimator*, in (7), for $\theta = 1$. With the appropriate choice of θ , $\hat{\gamma}_{k,n}^{NM(\theta)}(k)$ enables us to have access to an estimator of a negative EVI with a smaller asymptotic bias and the same asymptotic variance as the M-estimator.

Figure 1 illustrates, for a few values of θ , $\theta = 0, 0.5, 0.9$ and 1 , the behaviour of $\hat{\gamma}_{k,n}^{NM(\theta)}$, in (9), versus k , for a sample of size $n = 6574$ (the size of the whole sample of real data analysed in Section 4) from the *generalized Pareto* (GP) distribution with $\gamma = -0.25$. The GP d.f. is strongly related with the EV d.f. in (1), through the equation $GP_\gamma(x) = 1 + \ln EV_\gamma(x)$, $1 + \gamma x > 0$, $x > 0$.

For a large variety of models, and under mild second-order conditions, like the one in (5), we can guarantee the asymptotic normality of all the above mentioned estimators and build approximate confidence intervals (CI's) for γ , as well as for all other parameters of extreme events discussed next.

Let us consider the estimation of the right endpoint, for $\gamma < 0$. As mentioned before, we have $b_t \equiv b(t) = U(t)$, with $U(\cdot)$ defined in (3). On another side, the universal uniform transformation enables us to guarantee that for all F , unknown and underlying the r.v. X , $X \stackrel{d}{=} U(Y)$, with Y a unit Pareto r.v., i.e. a r.v. with d.f. $F_Y(y) = 1 - 1/y$, $y \geq 1$. Consequently, $X_{n-k:n} \stackrel{d}{=} U(Y_{n-k:n})$, and since $Y_{n-k:n} \stackrel{p}{\sim} n/k$ as $n \rightarrow \infty$, it is sensible to consider $\hat{b} = \hat{b}_{k,n} = \hat{U}(n/k) = X_{n-k:n}$. And for any γ -estimator,

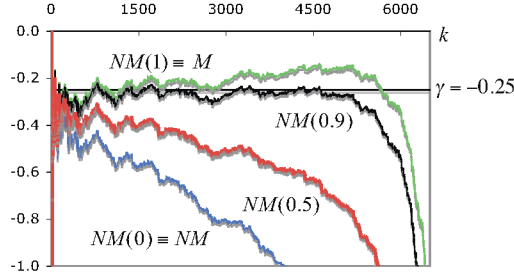


Fig. 1 Sample paths of $\hat{\gamma}_{k,n}^{NM(\theta)}$ in (9), denoted $NM(\theta)$, against k , for one sample of size $n = 6574$ from the GP_γ model, with $\gamma = -0.25$.

$\hat{\gamma}^\bullet \equiv \hat{\gamma}_{k,n}^\bullet$, we can consider (see [10]) $\hat{a}^\bullet = \hat{a}_{k,n}^\bullet = X_{n-k:n} M_{k,n}^{(1)} (1 - \min(0, \hat{\gamma}^\bullet))$, with $M_{k,n}^{(1)}$ given in (6). Moreover, for large values of t and $\gamma \neq 0$, we can write the approximation, $U(tx) \approx U(t) + a(t) (x^\gamma - 1)/\gamma$. But $x^F = U(\infty)$ and for all $\gamma < 0$, $(x^\gamma - 1)/\gamma \rightarrow -1/\gamma$, as $x \rightarrow \infty$. If we consider $t = n/k$, with k intermediate, we can thus guarantee that, whenever $\hat{\gamma}^\bullet < 0$,

$$x^F \approx U(n/k) - a(n/k)/\gamma \implies \hat{x}_\bullet^F := \hat{b} - \hat{a}^\bullet/\hat{\gamma}^\bullet. \quad (10)$$

For further details on the subject, see Chapters 1 and 4 of [10].

4 Adaptive choice of the tuning parameters under play

The objective of this Section is to provide an *Algorithm* for the choice of k and θ for the EVI-estimation through the *generalized negative moment* estimator in (9). Due to the fact that the main idea underlying this estimator is related with a possible reduction of bias, and γ is unknown, it is sensible to use any auxiliary statistic, strongly and directly related with the estimator under consideration, but going to the known value zero. This can be done in a way similar to the one considered in [7] for the choice of k in the *Hill* estimator, through the bootstrap methodology. Indeed, in the present set-up, the most obvious auxiliary statistic is

$$T_{k,n}(\theta) := \gamma_{[k/2],n}^{NM(\theta)} - \gamma_{k,n}^{NM(\theta)} = (\hat{\gamma}_{[k/2],n}^{NM} - \hat{\gamma}_{k,n}^{NM}) + \theta (M_{[k/2],n}^{(1)} - M_{k,n}^{(1)}) \quad (11)$$

$$=: r_k + \theta s_k, \quad k = 2, \dots, n-1. \quad (12)$$

Given the objectives, stability of $T_{k,n}(\theta)$ around zero for moderate values of k , say $k_1 := \lceil n^{0.05} \rceil + 1 \leq k \leq k_2 := \lceil n^{0.95} \rceil$, it seems sensible to choose

$$\hat{\theta} \equiv \hat{\theta}(k_1, k_2) := \arg \min_{\theta} \sum_{k=k_1}^{k_2} (r_k + \theta s_k)^2 = - \sum_{k=k_1}^{k_2} r_k s_k / \sum_{k=k_1}^{k_2} s_k^2, \quad (13)$$

where r_k and s_k have been defined in (12).

The choice of k for the estimation of the *extreme value index* is next done on the basis of bootstrap methodology, in a way similar to the one in [4], [2] and [7], and it is written algorithmically in the following. For further details, see the above mentioned papers.

Algorithm

1. Compute $\hat{\gamma}_{k,n}^{NM(\hat{\theta})}$, $k = 1, 2, \dots, n-1$, with $\hat{\gamma}_{k,n}^{NM(\theta)}$ and $\hat{\theta}$ defined in (9) and (13), respectively.
2. Next, consider sub-sample sizes $n_1 = o(n)$ and $n_2 = \lfloor n_1^2/n \rfloor + 1$.
3. For l from 1 until $B = 250$, generate independently B bootstrap samples $(x_1^*, \dots, x_{n_2}^*)$ and $(x_1^*, \dots, x_{n_2}^*, x_{n_2+1}^*, \dots, x_{n_1}^*)$, of sizes n_2 and n_1 , respectively, from the empirical d.f., $F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}}$, associated with the observed sample $(x_1, \dots, x_n)$.
4. Denoting $T_{k,n}^*(\hat{\theta})$ the bootstrap counterpart of $T_{k,n}(\hat{\theta})$, with $T_{k,n}(\theta)$ defined in (11), obtain $(t_{k,n_1,l}^*, 1 < k < n_1)$, $(t_{k,n_2,l}^*, 1 < k < n_2)$, $1 \leq l \leq B$, the observed values of the statistics T_{k,n_i}^* , $i = 1, 2$. For $k = 1, 2, \dots, n_i - 1$, compute

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

5. Obtain

$$\hat{k}_{0|T}^*(n_i) := \arg \min_{1 < k < n_i} MSE^*(n_i, k), \quad i = 1, 2. \quad (14)$$

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

$$\hat{\rho}^* := \ln \hat{k}_{0|T}^*(n_1) / (2 \ln(\hat{k}_{0|T}^*(n_1)/n_1)). \quad (15)$$

7. Compute the threshold estimate

$$\hat{k}_0^* \equiv \hat{k}_0^*(n; n_1) := \left[(1 - 2^{\hat{\rho}^*})^{2/(1-2\hat{\rho}^*)} (\hat{k}_{0|T}^*(n_1))^2 / \hat{k}_{0|T}^*(n_1^2/n) \right] + 1, \quad (16)$$

with $\hat{k}_{0|T}^*(n_i)$ and $\hat{\rho}^*$ given in (14) and (15), respectively.

8. 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 (9), (13) and (16), respectively.

Remarks:

1. If there are negative elements in the sample, the value of n should be replaced by $n^+ = \sum_{i=1}^n I_{\{X_i > 0\}}$ (the number of positive values in the sample).
2. The use of the sample of size n_2 , $(x_1^*, \dots, x_{n_2}^*)$, and of the extended sample of size n_1 , $(x_1^*, \dots, x_{n_2}^*, x_{n_2+1}^*, \dots, x_{n_1}^*)$, lead us to increase the precision of the result with

a smaller B , the number of bootstrap samples generated in Step 3. This is quite similar to the use of the simulation technique of “*Common Random Numbers*” in comparison problems, when we want to decrease the variance of a final answer to $z = y_1 - y_2$, inducing a positive dependence between y_1 and y_2 .

3. The Monte Carlo procedure in Steps 3. up to 7. or 3. up to 8. of the *Algorithm* in this Section can be replicated if we want to associate easily a standard error to the estimated parameters. The value of B can also be adequately chosen.
4. Although aware of the fact that k must be intermediate, we have used the entire region of search of $\arg \min_k MSE(k)$, the region $1 < k < n$, with n possibly denoting n_1 or n_2 and $MSE(k)$ standing for any mean squared error structure dependent on k . Indeed, whenever working with small values of n , $k = 1$ or $k = n - 1$ are also possible candidates to the value of an intermediate sequence for such n . Here we have again the old controversy between theoreticians and practitioners: $k_n = \lfloor c \ln n \rfloor$ is intermediate for every constant c , and if we take for instance $c = 1/5$, $k_n = 1$ for every $n \leq 22026$.
5. The main practical question which can be raised under this set-up is related with the dependence of the method on the choice of the sub-sample size n_1 , in Step 2. of the *Algorithm*. Although aware of the need to have $n_1 = o(n)$, it seems that, like in the classical semi-parametric estimation, we often get good results up till n , for moderate sample sizes. A possible choice seems then to be for instance $n_1 = \lfloor n^{0.90} \rfloor$ or $n_1 = \lfloor n^{0.95} \rfloor$. The first one will be the one used later on in Section 5. If the sample size n is large, values of n_1 larger than $\lfloor n^{0.95} \rfloor$ do not seem convenient, as we shall see later on in Section 6.

5 Applications to simulated and environmental data sets

We shall first provide an application to a simulated data sample, a case where we know the values of both γ and x^F . We next provide an application not only to a subset related with the periods from March 21 until June 20 of the *Wind* data set mentioned in Section 1, the so-called *Spring-Wind* data, but also to the *Summer-Wind* data set.

Simulated GP data

The analysis of a randomly simulated sample from a GP model with $\gamma = -0.25$, with a size $n = 6574$, the size of the *Wind* data set, led us to the choice $\hat{\theta} = 0.899$, as shown in Figure 2. It is sensible to note the resistance of the method to the choice of k , something that does not happen for all data sets analyzed. For this data set, we have got $\hat{\theta} = 0.9$ for $k \in [2793, 5884]$.

The application of the *Algorithm*, in Section 4, for $n_1 = \lfloor n^{0.9} \rfloor$, led then to $\hat{k}_0^* = 3817$ and the adaptive EVI-estimate $\hat{\gamma}^* = -0.267$, quite close to the true value $\gamma = -0.25$ and pictured in Figure 3. In this case there is not a big discrepancy between the M and the NM*-estimators.

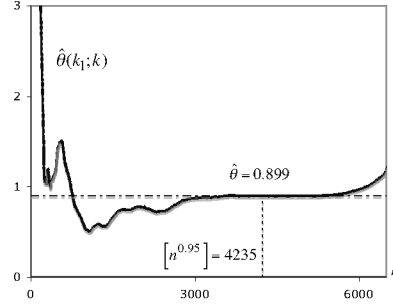


Fig. 2 Estimates $\hat{\theta}(k_1; k)$, with $\hat{\theta}(k_1; k_2)$ given in (13), for the simulated GP(−0.25) sample.

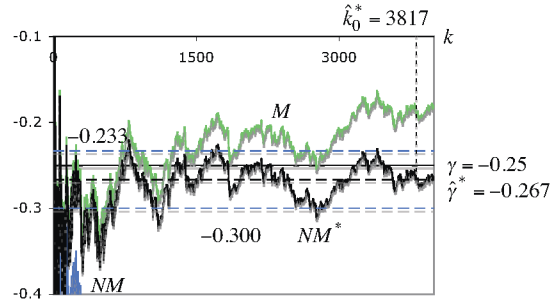


Fig. 3 Estimates of the EVI provided by the estimators under consideration for the simulated GP(−0.25) sample.

For a negative EVI and adequate conditions on k (see [1], for details), and with $\mathcal{N}(\mu, \sigma^2)$ denoting a normal r.v. with mean value μ and variance σ^2 , we get

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

An approximate 95% confidence interval (CI) for γ is given by $(\hat{\gamma}^* - 1.96 \hat{\sigma}_{NM}/\sqrt{k}, \hat{\gamma}^* + 1.96 \hat{\sigma}_{NM}/\sqrt{k})$. In this case, we have got $\hat{\sigma}_{NM} = 1.058$ and the approximate 95% CI, $(-0.300, -0.233)$, also pictured in Figure 3. The associated estimate of the right endpoint was $\hat{x}^F = 3.976$, quite close to the right endpoint $x^F = 4$.

Wind data

On the basis of the sample of daily average wind speeds (in knots), collected in Dublin airport, in the period 1961-1978, with a size $n = 6574$, and in order to avoid seasonality, we have considered the *Spring-Wind* data set, with a size $n = 1656$, as well as the *Summer-Wind* data set, with a size $n = 1692$.

Spring-Wind data. For this data set, we were led to the choice $\hat{\theta} = 1.411$. The application of the *Algorithm*, in Section 4, for $n_1 = [n^{0.9}]$, led then to $\hat{k}_0^* = 948$ and the adaptive EVI-estimate $\hat{\gamma}^* = -0.201$, the value pictured in Figure 4. In this case there is again not a big discrepancy between the M and the NM^* -estimators, but this is not the usual situation, as will be illustrated later on, with the consideration of the *Summer-Wind*. We have got $\hat{\sigma}_{NM} = 1.005$ and the approximate 95% CI, $(-0.265, -0.137)$, also pictured in Figure 4 and clearly indicating a light right-tail.

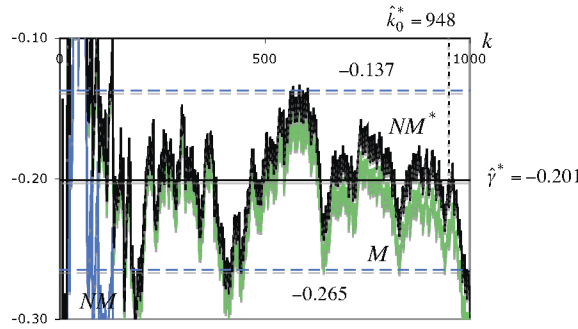


Fig. 4 Estimates of the EVI provided by the estimators under consideration, for the wind data (Spring period).

Summer-Wind data. For this data set, we were led to the choice $\hat{\theta} = 1.655$. The application of the *Algorithm*, in Section 4, for $n_1 = [n^{0.9}]$, led then to $\hat{k}_0^* = 1067$ and the adaptive EVI-estimate $\hat{\gamma}^* = -0.022$, the value pictured in Figure 5.

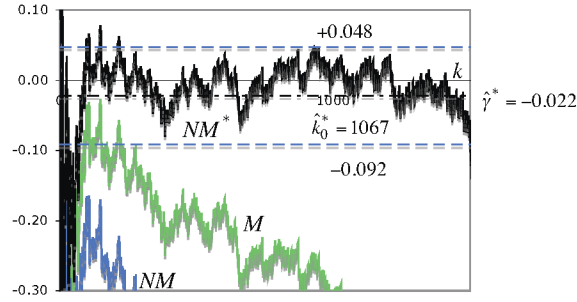


Fig. 5 Estimates of the EVI provided by the estimators under consideration, for the wind data (Summer period).

In this case, we have got $\hat{\sigma}_{NM} = 0.982$ and the approximate 95% CI, $(-0.092, +0.048)$, also pictured in Figure 4, indicating now the possibility of an exponential-type tail, i.e. of an extreme value index $\gamma = 0$.

6 Sensitivity of the Algorithm to the choice of n_1

We shall now briefly address the sensitivity of the *Algorithm* in Section 4 to the choice of n_1 . In Tables 1, 2 and 3, we present estimates of $\hat{k}_0^*$ and $\hat{\gamma}^*$, for the analyzed data sets, the *GP*, the *Spring-Wind* and the *Summer-Wind*, respectively. Those estimates, as functions of the sub-sample size n_1 , for n_1 running from $[n^{0.8}]$ until $[n^{0.999}]$, are pictured in Figures 6, 7 and 8, where we also picture horizontal line(s), with the overall median of all values or the medians of different stability regions.

GP data

p	$n_1 = \lfloor n^p \rfloor$	$n_2 = n_1^2/n$	$\hat{k}_{0 T}^*(n_1)$	$\hat{k}_{0 T}^*(n_2)$	$\hat{k}_0^*$	$\hat{\gamma}^*$
0.8	1133	195	662	116	3770	-0.22
0.9	2729	1132	1576	650	3817	-0.24
0.95	4235	2728	2427	1562	3768	-0.22
0.975	5276	4234	1970	1539	2481	-0.19
0.990	6020	5512	3442	3162	3744	-0.20
0.999	6516	6458	2169	2160	2128	-0.22

Table 1 Estimates associated to a GP_γ model, with $\gamma = -0.25$.

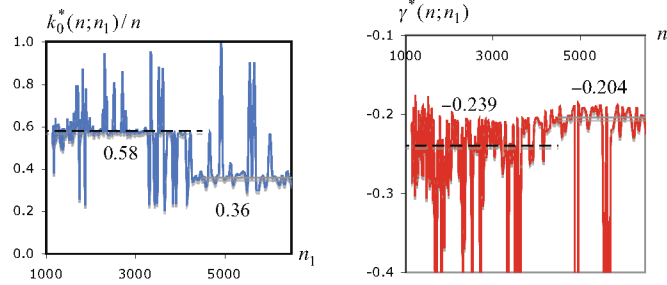


Fig. 6 Bootstrap estimates of the optimal sample fraction (*left*) and of the extreme value index γ (*right*), as a function of the sub-sample size n_1 , for the *GP* data.

In this case of an underlying GP model, with $\gamma = -0.25$, it seems sensible to advise the choice of sub-sample sizes $n_1 \leq [n^{0.95}]$. We then get a median equal to -0.239 (slightly over-estimating the true γ). For larger values of n_1 there appears a clear over-estimation of the extreme value index γ .

Spring-Wind data

For this data set we have got the following results.

p	$n_1 = \lfloor n^p \rfloor$	$n_2 = n_1^2/n$	$\hat{k}_{0 T}^*(n_1)$	$\hat{k}_{0 T}^*(n_2)$	$\hat{k}_0^*$	$\hat{\gamma}^*$
0.8	376	85	223	54	917	-0.078
0.9	789	375	476	226	1001	-0.145
0.95	1143	788	687	470	1003	-0.137
0.975	1375	1141	832	690	1002	-0.141
0.990	1537	1426	917	850	988	-0.126
0.999	1643	1630	955	950	959	-0.066

Table 2 Estimates associated to *Spring-Wind* data.

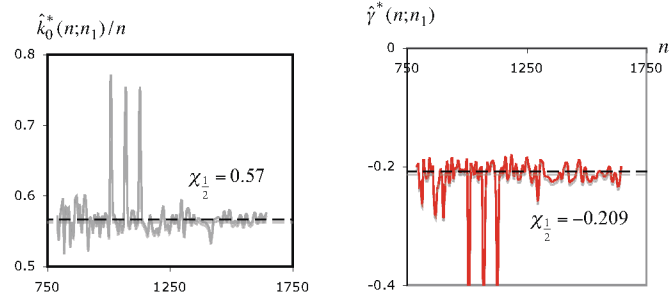


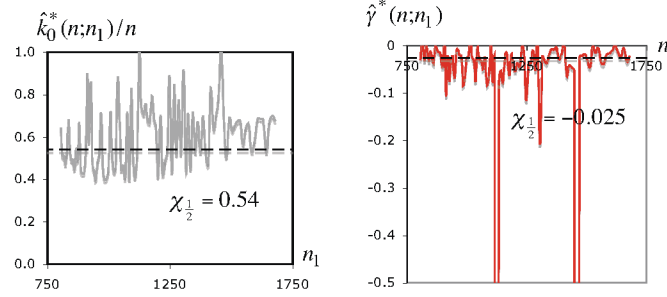
Fig. 7 Bootstrap estimates of the optimal sample fraction (*left*) and of the extreme value index γ (*right*), as a function of the sub-sample size n_1 , for the *Spring-Wind* data.

It seems that we can now go up to sub-sample sizes $n_1 = \lfloor n^{0.999} \rfloor$. The median of the bootstrap EVI-estimates associated to all these sub-sample sizes is equal to -0.209 . There are however a few discrepant values, related with a weak estimation of $\hat{k}_{0|T}^*(n_2)$.

Summer-Wind data

For the *Summer-Wind* data, there is a high volatility in the estimation of the optimal sample fractions, which vary from 40% until 99%. A few sub-sample sizes n_1 (around 6.5% of all sub-sample sizes considered) lead even to positive values of $\hat{\gamma}^*$, and there were two extremely small estimates of γ , the estimates -1.35 and -7.69 , neatly discrepant from all the other estimates. The volatility of the bootstrap estimates of γ is also quite high. These results claim for a simulation study of the *Algorithm* in Section 4, a topic out of the scope of this paper.

p	$n_1 = \lfloor n^p \rfloor$	$n_2 = n_1^2/n$	$\hat{k}_{0T}^*(n_1)$	$\hat{k}_{0T}^*(n_2)$	$\hat{k}_0^*$	$\hat{\gamma}^*$
0.8	382	86	200	52	760	-0.045
0.9	804	382	502	236	1067	-0.030
0.95	1166	803	777	545	1108	-0.015
0.975	1405	1166	786	784	786	0.002
0.990	1570	1456	1018	948	1093	-0.039
0.999	1679	1666	1111	1102	1120	-0.022

Table 3 Estimates associated to *Summer-Wind* data.**Fig. 8** Bootstrap estimates of the optimal sample fraction (*left*) and of the extreme value index γ (*right*), as a function of the sub-sample size n_1 , for the *Summer-Wind* data.

Acknowledgements Research partially financed by FCT / POCTI, POCI and PPCDT / FEDER.

References

1. Caeiro, F., Gomes, M.I.: An asymptotically unbiased moment estimator of a negative extreme value index. *Discussiones Mathematica: Probability and Statistics*, accepted (2010)
2. Danielsson, J., Haan, L. de, Peng, L., Vries, C.G. de: Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivar. Anal.* **76**, 226–248 (2001)
3. Dekkers, A.L.M., Einmahl, J.H.J., Haan, L. de: A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics* **17**(4), 1833–1855 (1989)
4. Draisma, G., Haan, L. de, Peng, L., Themido Pereira, T.: A bootstrap-based method to achieve optimality in estimating the extreme value index. *Extremes* **2**(4), 367–404 (1999)
5. Fraga Alves, M.I.: Weiss-Hill estimator. *Test* **10**, 203–224 (2001)
6. Gnedenko, B.V.: Sur la distribution limite du terme maximum d’une srie alatoire. *Ann. Math.* **44**, 423–453 (1943)
7. Gomes, M.I., Oliveira, O.: The bootstrap methodology in Statistics of Extremes — choice of the optimal sample fraction. *Extremes* **4**:4, 331–358 (2001)
8. Haan, L. de: On Regular Variation and its Application to the Weak Convergence of Sample Extremes. *Mathematical Centre Tract 32*, Amsterdam (1970)
9. Haan, L. de: Slow variation and characterization of domains of attraction. In: Tiago de Oliveira (ed.) *Statistical Extremes and Applications*, pp. 31–48, D. Reidel, Dordrecht, Holland (1984)
10. Haan, L. de, Ferreira, A.: *Extreme Value Theory: an Introduction*. Springer, New York (2006)
11. Hill, B.M.: A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* **3**(5), 1163–1174 (1975)