Available online at www.sciencedirect.com
SCIENCE DIRECT®

C. R. Acad. Sci. Paris, Ser. I 337 (2003) 287–292

COMPTES RENDUS
MATHEMATIQUE

Statistics/Probability Theory

Double-thresholded estimator of extreme value index

Laurent Gardes

Laboratoire de probabilités et statistique, Université Montpellier 2, place Eugène Bataillon, 34 095 Montpellier cedex 5, France

Received 1 July 2002; accepted after revision 24 June 2003

Presented by Paul Deheuvels

Abstract

The purpose of this Note is to propose an estimator of the extreme value index constructed by using only the number of points exceeding random thresholds. We prove the weak consistency and the asymptotic normality of this estimator. We deduce from this last result that the rate of convergence of our estimator is in a power of the sample size. To our knowledge, this rate of convergence is not reached by any other estimate of the extreme value index. Through a simulation, we compare our estimator to the moment estimator (Dekkers et al., Ann. Statist. 17 (1989) 1833–1855). ***To cite this article: L. Gardes, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Résumé

Un estimateur de l'indice des valeurs extrêmes à double seuil. Dans cette Note, nous proposons un estimateur de l'indice des valeurs extrêmes construit en utilisant uniquement le nombre de points qui dépassent des seuils aléatoires. On démontre qu'il est faiblement consistant et asymptotiquement normal. Du résultat de convergence en loi, on déduit que la vitesse de convergence de notre estimateur est une puissance de la taille de l'échantillon. A notre connaissance, cette vitesse n'est atteinte par aucun autre estimateur de l'indice des valeurs extrêmes. A l'aide de simulations, nous comparons notre estimateur à l'estimateur des moments (Dekkers et al., Ann. Statist. 17 (1989) 1833–1855). ***Pour citer cet article : L. Gardes, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Version française abrégée

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même fonction de répartition F. On note $X_{1,n} \leqslant \cdots \leqslant X_{n,n}$ l'échantillon ordonné. Alors que l'estimation de l'indice des valeurs extrêmes (noté ξ) a été intensivement étudié dans le cas du domaine d'attraction de Fréchet ($\xi > 0$), nous allons dans cette Note traiter l'estimation de ξ dans le cas où F appartient au domaine d'attraction de Weibull ($\xi < 0$). Nous avons en projet d'utiliser cet estimateur pour l'estimation de support (voir Gijbels et al. [6,7], Hall et al. [8], Härdle et al. [9], ...). En effet, dans ce type de question, les lois sous-jacentes appartiennent, pour leur grande majorité, au domaine

E-mail address: gardes@stat.math.univ-montp2.fr (L. Gardes).

1631-073X/$ – see front matter
doi:10.1016/S1631-073X(03)00329-7

d'attraction de Weibull. On suppose donc dans la suite que F appartient au domaine d'attraction de Weibull. Autrement dit, on suppose que $M = \sup\{x \mid F(x) < 1\} < \infty$ et que

$$F(x) = 1 - (M - x)^{-1/\xi} \ell\big[(M - x)^{-1}\big],$$

où ℓ est une fonction à variations lentes, c'est à dire telle que $\ell(\lambda x)/\ell(x) \to 1$ lorsque $x \to \infty$ pour tout $\lambda > 0$. La question de l'estimation de l'indice des valeurs extrêmes connait un intérêt croissant. Deux premiers estimateurs ont été proposés en 1975 par Hill [10] et Pickands [13]. Des améliorations de ceux-ci ont été proposées notamment par Dekkers, Einmahl et de Haan [1] et Drees [2]. Ces estimateurs sont construits en utilisant l'ensemble des k plus grandes observations de l'échantillon. Une autre méthode pour construire des estimateurs de l'indice des valeurs extrêmes consiste à utiliser les observations qui dépassent un seuil déterministe. Cette méthode est notamment utilisée par Falk [3,4] et Marohn [11,12] pour proposer un estimateur basé sur le nombre de points qui dépassent le seuil déterministe. L'inconvénient de cet estimateur réside dans le fait que, pour être consistant, le seuil déterministe doit tendre vers le point terminal de la distribution, ce point terminal pouvant ne pas être connu. Notre idée a alors été de travailler avec un seuil aléatoire défini par $U_n^* = u_n X_{n,n}$, où (u_n) est une suite de $]0,1[$ tendant vers 1 lorsque la taille de l'échantillon n tend vers l'infini. Ce seuil aléatoire a l'avantage de tendre en probabilité vers le point terminal.

Dans la suite, τ_{u_n} désigne le nombre de points qui dépassent le seuil U_n^*. Nous proposons un estimateur de l'indice ξ construit en utilisant le rapport de nombre de points τ_{u_n}/τ_{v_n} où (v_n) est une suite de $]0,1[$, différente de (u_n) et tendant vers 1 lorsque n tend vers l'infini. Plus précisément, on propose comme estimateur de ξ :

$$\hat{\xi}_n = -\frac{\ln[(1 - u_n)/(1 - v_n)]}{\ln[\tau_{u_n}/\tau_{v_n}]}.$$

L'estimateur proposé a l'avantage d'être toujours strictement négatif. On montre que sous certaines conditions, cet estimateur est convergent en probabilité. De plus, sous certaines conditions supplémentaires (en particulier si $\xi < -1/2$), la loi de S_n avec $S_n = \sqrt{n[1 - F(u_n M)]}(\hat{\xi}_n - \xi)$, converge vers une loi normale. On déduit de ce dernier résultat que, dans le cas $\xi < -1/2$, la vitesse de convergence de $\hat{\xi}_n$ est une puissance de n. A notre connaissance, cette vitesse n'est atteinte par aucun autre estimateur de ξ.

1. Definition and assumptions

Let $X_1, \ldots, X_n$ be a sample of n independent random variables with common distribution function F. We denote by $X_{1,n} \leqslant \cdots \leqslant X_{n,n}$ the ordered sample. Our goal is to propose an estimator of the extreme value index which will be used in the future for estimating support frontier. Thus, naturally, we suppose that F belongs to the domain of attraction of Weibull. Recall that this is equivalent to $M = \sup\{x \mid F(x) < 1\} < \infty$ and:

$$F(x) = 1 - (M - x)^{-1/\xi} \ell\big[(M - x)^{-1}\big],$$

where ℓ is a slowly varying function, i.e., such that $\ell(\lambda x)/\ell(x) \to 1$ as $x \to \infty$ for all $\lambda > 0$. This Note is devoted to the study of an estimator of the extreme value index $\xi < 0$. This estimator is based on the ratio of two random numbers of points exceeding two random thresholds. They are defined by $U_n^* = u_n X_{n,n}$ where $u_n \in]0,1[$ with $u_n \to 1$ as $n \to \infty$ and $V_n^* = v_n X_{n,n}$ where $v_n \in]0,1[$ with $v_n \neq u_n$ and $v_n \to 1$ as $n \to \infty$. We denote by τ_{u_n} (respectively τ_{v_n}) the number of points exceeding U_n^* (respectively V_n^*). More precisely, this estimator is defined by:

$$\hat{\xi}_n = -\frac{\ln[(1 - u_n)/(1 - v_n)]}{\ln[\tau_{u_n}/\tau_{v_n}]}.$$

Remark 1. It is readily seen that $X_{n,n} \leqslant 0$ entails $\tau_{u_n} = 0$. Consequently, we must suppose in what follows that $M > 0$ to insure that $\mathrm{P}[X_{n,n} \leqslant 0] \to 0$ as $n \to \infty$. This assumption is not a restrictive one since it can be achieved by a simple shift.

The following assumptions will be considered in the sequel:

(A1) The slowly varying function ℓ is normalized, i.e.,

$$\ell(x) = c \exp\left\{ \int_1^x t^{-1}\varepsilon(t)\,\mathrm{d}t \right\},$$

where c is a positive constant and $\varepsilon(x) \to 0$ as $x \to \infty$, which implies that ℓ is differentiable.

This last assumption implies in particular that F admits a derivative denoted by f.

(A2) ℓ verifies the second order condition:

$$\ln\left(\frac{\ell(\lambda t)}{\ell(t)}\right) \sim b(t)K_\rho(\lambda) \quad \text{uniformly in } \lambda \geqslant 1,$$

where $\rho \leqslant 0$, $b(t) \to 0$ as $t \to \infty$ and $K_\rho(\lambda) = \int_1^\lambda u^{\rho-1}\,\mathrm{d}u$. It can be shown [6] that necessarily $|b|$ is regularly varying with index ρ.

(B1) In the case $\limsup_{y\uparrow M} f(y) < +\infty$, we assume that f is bounded.

(B2) In the case $\limsup_{y\uparrow M} f(y) = +\infty$, we assume that there exists $\varepsilon > 0$ such that f is bounded on $[0, M - \varepsilon]$ and f is an increasing function on $[M - \varepsilon, M[$.

In the following (B) denotes one of the assumptions (B1) or (B2). Note that this last assumption is verified for example by a Generalized Pareto Distribution ($F(x) = 1 - (1 + \xi x/\sigma)^{-1/\xi}$, $\sigma > 0$ and $\xi < 0$), by the Weibull Extreme Value Distribution ($F(x) = \exp[-(1 + \xi(x - \mu)/\sigma)^{-1/\xi}]$, $\xi < 0$, $\sigma > 0$ and $\mu \in \mathbb{R}$), by a reversed Burr distribution ($F(x) = 1 - [\beta/(\beta + (M - x)^{-\tau})]^\lambda$, $(\beta, \tau, \lambda) \in (\mathbb{R}^+)^3$), ...

(C) $\xi < -1/2$.

Let us denote by $\theta_{1,\xi} = \min\{\xi^2/2, \xi^2/(1-\xi)\}$ and $\theta_{2,\xi} = \min\{\xi^2/2, \xi^2/(1-\xi), 2\xi^2 + \xi\}$.

(D1) $n^{-\theta_{1,\xi}} = \mathrm{o}(1 - u_n)$. (D2) $n^{-\theta_{1,\xi}} = \mathrm{o}(1 - v_n)$. (D3) $n^{-\theta_{2,\xi}} = \mathrm{o}(1 - u_n)$. (D4) $n^{-\theta_{2,\xi}} = \mathrm{o}(1 - v_n)$.

Remark that assumption (D3) implies assumption (D1) and that assumption (D4) implies assumption (D2).

2. Main results

In this section we state that, under some conditions, $\hat{\xi}_n$ converges in probability. Furthermore, under some additionnal conditions, we show that the distribution of S_n with $S_n = \sqrt{n[1 - F(u_n M)]}(\hat{\xi}_n - \xi)$, converges to a Gaussian distribution. The proofs are postponed to Section 4.

Theorem 2.1. *Under assumptions* (A1), (B), (D1), (D2) *and if*

$$0 < \liminf_{n\to\infty} \frac{1 - u_n}{1 - v_n} \leqslant \limsup_{n\to\infty} \frac{1 - u_n}{1 - v_n} < 1, \tag{1}$$

then, $\hat{\xi}_n \xrightarrow{\mathrm{P}} \xi$.

Remark 2. If we assume that $\ell(x) \to C > 0$ as $x \to \infty$, then one can substitute (1) by the weaker condition

$$\limsup_{n\to\infty} \frac{1 - u_n}{1 - v_n} < 1. \tag{2}$$

Theorem 2.2. *Suppose assumptions* (A1), (A2), (B), (C), (D3) *and* (D4) *hold. If there exist* $c \in]0,1[$ *and a sequence* (η_n) *such that* $(1-u_n)/(1-v_n) = c + \eta_n$ *and*

$$\sqrt{n\big[1-F(u_n M)\big]}\max\big(\eta_n, b\big[M^{-1}(1-u_n)^{-1}\big]\big) \to 0,$$

as $n \to \infty$ *then*

$$\sqrt{n\big[1-F(u_n M)\big]}(\hat{\xi}_n - \xi) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{(1-c^{-1/\xi})\xi^4}{\ln^2(c)}\right).$$

Remark 3. (1) Suppose that $1-u_n \sim dn^{-\gamma}$ with $d>0$ and $\gamma>0$. Since $|b|$ is regularly varying with index ρ, it is easily seen that if

$$-\frac{\xi}{1+2\xi\rho} < \gamma < \theta_{2,\xi}, \tag{3}$$

with $\xi < -1/2$, then u_n verifies the assumptions of Theorem 2.2. Condition (3) can be verified if and only if $-\xi/(1+2\xi\rho) < \theta_{2,\xi}$, i.e., if and only if $(-1/\xi, -\rho) \in D = \{(x,y) \in \mathbb{R}^2 \mid 0 < x < 2 \text{ and } y > g(x)\}$, where $g(x) = x^2/2$ if $0 < x \leqslant 1$, $g(x) = x^2 - x/2$ if $1 < x \leqslant 3/2$ and $g(x) = (x^2-x)/(2-x)$ if $x > 3/2$. Thus, if $(-1/\xi, -\rho) \in D$, the best rate of convergence, which is given by the normalizing coefficient $\sqrt{n[1-F(u_n M)]}$, is obtained by taking γ as small as possible i.e., by taking $\gamma = -\xi/(1+2\xi\rho) + \varepsilon$, with $0 < \varepsilon < \theta_{2,\xi} + \xi/(1+2\xi\rho)$ small. This choice of γ leads to the rate of convergence:

$$\sqrt{n\big[1-F(u_n M)\big]} \sim dM^{-1/(2\xi)} n^{\varepsilon/(2\xi)} \big\{\ell\big[M^{-1}(1-u_n)^{-1}\big]\big\}^{1/2} n^{1/[2+1/(\xi\rho)]} = \mathrm{o}\big(n^{1/[2+1/(\xi\rho)]}\big).$$

We know nothing of the asymptotic distribution of $\sqrt{n[1-F(u_n M)]}(\hat{\xi}_n - \xi)$ if $(-1/\xi, -\rho) \notin D$.

(2) The problem of choosing the sequence (v_n) is quite similar to the one of choosing the number k of order statistics for classical estimators such that Hill [10], moment estimator [1], Pickands [13], ... Concerning the choice of the sequence (u_n), the variance (u_n must be choosen not too large) and the bias (u_n must be choosen not too small) of the estimator should be taken into account. These problems are difficult and are not treated in this Note.

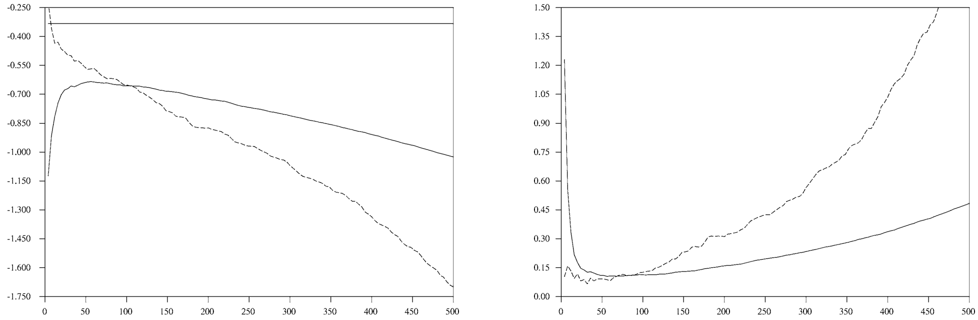
3. An example of simulation

We compare the mean and the mean squared error (MSE) of $\hat{\xi}_n$ and the moment estimator. More precisely, let $F(x) = 1 - [1-(10-x)^{-1}]^{-3}$ (here, $\xi = -1/3$). We simulate, under this distribution, $N = 100$ samples of size $n = 1000$. Let us denote by $k(v_n)$ (resp. $k(u_n)$) the mean of the N values of τ_{v_n} (resp. τ_{u_n}) ($k(v_n)$ is the mean of the number of points used to compute $\hat{\xi}_n$). We choose the sequence (v_n) such that $k(v_n) = k$ where k is the number of order statistics used to compute the moment estimator. Recall that the moment estimator is defined by $\hat{\xi}_n^M(k) = H_{k,n} + 1 - 1/2(1 - H_{k,n}^2/S_{k,n})^{-1}$, where $H_{k,n} = 1/k\sum_{j=1}^k \ln(X_{n-j+1,n}) - \ln(X_{n-k,n})$ and $S_{k,n} = 1/k\sum_{j=1}^k [\ln(X_{n-j+1,n}) - \ln(X_{n-k,n})]^2$. We choose the sequence (u_n) such that $k(u_n) = k/4$ (this choice is arbitrary). We then compute the mean (Fig. 1(a)) and the MSE (Fig. 1(b)) for different values of k ($k = 4, 8, 12, \ldots, 500$). More detailed simulations can be found in [5].

4. Proofs

The first lemma is dedicated to the control of the mathematical expectation of τ_{u_n}. The behaviour of τ_{u_n} in probability is given in Lemma 4.2 which in turn gives the behaviour of $\hat{\xi}_n$ in probability.

Lemma 4.1. *Under assumptions* (A1), (B) *and* (D1) *we have* $\mathrm{E}(\tau_{u_n}) \sim n[1-F(u_n M)]$.



(a) (b)

Fig. 1. Comparison between $\hat{\xi}_n$ (continuous line) and the moment estimator (dashed line). In (a), the straight line represent the true value of ξ. (a) Mean. (b) MSE.

Fig. 1. Comparaison entre $\hat{\xi}_n$ (ligne continue) et l'estimateur des moments (pointillé). Dans (a) la ligne horizontale représente la vraie valeur de ξ.

Proof. Let $p_{n,\xi} = \mathrm{P}[X_1 \geqslant U_n^*]$. By calculating the distribution of τ_{u_n}, we show that $\mathrm{E}(\tau_{u_n}) = np_{n,\xi}$. It is easy to prove that $p_{n,\xi} = [1 - F(u_n M)] + \int_0^M u_n f(u_n t)[F(t)]^{n-1}\,\mathrm{d}t$. Thus, it remains to demonstrate that:

$$\int_0^M u_n f(u_n t)\big[F(t)\big]^{n-1}\,\mathrm{d}t = \mathrm{o}\big[1 - F(u_n M)\big].$$

Under assumption (B1), one can find $B > 0$ such that:

$$\int_0^M u_n f(u_n t)\big[F(t)\big]^{n-1}\,\mathrm{d}t \leqslant u_n B \int_0^M \big[F(t)\big]^{n-1}\,\mathrm{d}t \leqslant u_n B n^{\xi} L(n),$$

where L is a slowly varying function. This last inequality is a consequence of the equivalence $M - \mathrm{E}(X_{n,n}^+) \sim n^{\xi} L(n)$, where $X_{n,n}^+ = \max(0, X_{n,n})$ (see [14, Proposition 2.1(ii)]). Consequently,

$$\frac{\int_0^M u_n f(u_n t)[F(t)]^{n-1}\,\mathrm{d}t}{1 - F(u_n M)} = \mathrm{O}\left[\frac{(1-u_n)^{1/\xi} n^{\xi} L(n)}{\ell[M^{-1}(1-u_n)^{-1}]}\right].$$

Assumption (D1) and a well known property of slowly varying function conclude the first part of the proof. Consider now assumption (B2). Since $u_n \to 1$ as $n \to \infty$, we have for n sufficiently large,

$$\int_0^M u_n f(u_n t)\big[F(t)\big]^{n-1}\,\mathrm{d}t \leqslant f(u_n M) \int_0^M \big[F(t)\big]^{n-1}\,\mathrm{d}t \leqslant f(u_n M) n^{\xi} L(n).$$

As a consequence of Lemma 4.1, $f(u_n M) \sim -1/(M\xi)(1-u_n)^{-1}[1 - F(u_n M)]$ and thus

$$\frac{\int_0^M u_n f(u_n t)[F(t)]^{n-1}\,\mathrm{d}t}{1 - F(u_n M)} = \mathrm{O}\big[(1-u_n)^{-1} n^{\xi} L(n)\big].$$

Assumption (D1) and a well known property of slowly varying function conclude the proof. □

Similarly to the calculus of the mathematical expectation, one can calculate the variance of τ_{u_n} and then show that $\mathrm{Var}(\tau_{u_n}) = \mathrm{o}\{[\mathrm{E}(\tau_{u_n})]^2\}$. Hence, the Bienaymé–Tchebichev inequality leads to the following result which is essential to prove Theorem 2.1:

Lemma 4.2. *Under assumptions* (A1), (B) *and* (D1), *we have*:

$$\frac{\tau_{u_n}}{n[1-F(u_n M)]} \xrightarrow{\mathrm{P}} 1.$$

Proof of Theorem 2.1. Remarking that

$$\frac{1-v_n}{1-u_n}\left[\frac{\tau_{u_n}}{\tau_{v_n}}\right]^{-\xi} = \left[\frac{\tau_{u_n}}{n[1-F(u_n M)]}\right]^{-\xi}\left[\frac{\tau_{v_n}}{n[1-F(v_n M)]}\right]^{\xi}\frac{\ell[M^{-1}(1-u_n)^{-1}]}{\ell[M^{-1}(1-v_n)^{-1}]},$$

it is easily seen, with Lemma 4.2, that

$$\frac{1-v_n}{1-u_n}\left[\frac{\tau_{u_n}}{\tau_{v_n}}\right]^{-\xi} \xrightarrow{\mathrm{P}} 1. \tag{4}$$

By taking the logarithm in (4) and by dividing by $\ln[(1-u_n)/(1-v_n)]$, we find that $1-\xi/\hat{\xi}_n \xrightarrow{\mathrm{P}} 0$. □

We now prove Theorem 2.2. To this end, let $\alpha_n = \sqrt{n[1-F(u_n M)]}$ and $\beta_n = \sqrt{n[1-F(v_n M)]}$. Under assumptions of Theorem 2.2, a second-order Taylor development of the caracteristic function of $(\tau_{u_n}/\alpha_n - \alpha_n, \tau_{v_n}/\beta_n - \beta_n)$ leads to

$$\left(\frac{\tau_{u_n}}{\alpha_n} - \alpha_n, \frac{\tau_{v_n}}{\beta_n} - \beta_n\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

with

$$\Sigma = \begin{bmatrix} 1 & c^{-1/(2\xi)} \\ c^{-1/(2\xi)} & 1 \end{bmatrix}.$$

An application of the δ-method on this last result leads to Theorem 2.2.

References

[1] A.L.M. Dekkers, J.H.J. Einmahl, L. de Haan, A moment estimator for the index of an extreme-value distribution, Ann. Statist. 17 (1989) 1833–1855.
[2] H. Drees, Refined Pickands estimator of the extreme value index, Ann. Statist. 23 (1995) 2059–2080.
[3] M. Falk, On testing the extreme value index via the POT-method, Ann. Statist. 23 (1995) 2013–2035.
[4] M. Falk, Local asymptotic normality of truncated empirical processes, Ann. Statist. 26 (1998) 692–718.
[5] L. Gardes, Estimation de l'indice de valeur extrême, Rapport de Recherche ENSAM-INRA-UM2 02-06, 2002.
[6] I. Gijbels, E. Mammen, B.U. Park, L. Simar, On estimation of monotone and concave frontier functions, J. Amer. Statist. Assoc. 94 (1999) 220–228.
[7] I. Gijbels, L. Peng, Estimation of a support curve via order statistics, Extremes 3 (1999) 251–277.
[8] P. Hall, M. Nussbaum, S.E. Stern, On the estimation of a support curve of indeterminate sharpness, J. Multivariate Anal. 62 (1997) 204–232.
[9] W. Härdle, B.U. Park, A.B. Tsybakov, Estimation of non-sharp boundaries, J. Multivariate Anal. 55 (1995) 205–218.
[10] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (1975) 1163–1174.
[11] F. Marohn, Testing the Gumbel hypothesis via the P.O.T. method, Extremes 1 (2) (1998) 191–213.
[12] F. Marohn, Local asymptotic normality of truncated models, Statist. Decisions 17 (1999) 237–253.
[13] J. Pickands III, Statistical inference using extreme-order statistics, Ann. Statist. 3 (1975) 119–131.
[14] S.I. Resnick, Extreme Values, Regular Variation, and Point Process, Springer-Verlag, New York, 1987.