On Blocks and Runs Estimators of the Extremal Index

J. Statist. Planning Inference, v. 66, No 2, 281–288.

S. Yu. Novak

Novosibirsk State University
Novosibirsk 630090
Russia

I. Weissman

Faculty of Industrial Engineering and Management
Technion — Israel Institute of Technology
Haifa 32000, Israel

ABSTRACT

Given a sample from a stationary sequence of random variables, we study the blocks and runs estimators of the extremal index. Conditions are given for consistency and asymptotic normality of these estimators. We show that moment restrictions assumed by Hsing (1991, 1993) may be relaxed if a stronger mixing condition holds. The CLT for the runs estimator seems to be proven for the first time.

1 Introduction

Let $\{X_i : i \geq 1\}$ be a strictly stationary sequence of random variables (r.v.'s) with a marginal distribution function (df) F. For $0 \leq m \leq n$ and $i, r \geq 1$ we define

$$M_{m,n} = \max_{m<i\leq n} X_i \ , \ M_n = M_{0,n} \ , \ M_r^{(i)} = M_{(i-1)r,ir}.$$

We suppose that sequence $\{X_i\}$ possesses an *extremal index* $\theta \in (0, 1]$. Namely, if a threshold level $u_n = u_n(\tau)$ is chosen so that

$$n(1 - F(u_n)) \to \tau > 0 \tag{1.1}$$

as $n \to \infty$, then

$$\mathbb{P}\{M_n \le u_n\} \to e^{-\theta\tau}. \tag{1.2}$$

This means that for large n, one can use the approximation

$$\mathbb{P}\{M_n \le u_n\} \approx F^{n\theta}(u_n) \tag{1.3}$$

for the distribution of the r.v. M_n. Hence, the extremal index θ is a key parameter for the distribution of sample extremes.

The present paper is concerned with the estimation of θ.

We base our results on two types of approximations for θ. The first one was introduced by O'Brien (1974, 1987), who showed that θ may be approximated by

$$\theta^R(r,u) = P\{M_{1,r} \le u | X_1 > u\} \ .$$

The second type of approximation for θ is based on Leadbetter's (1983) results: θ may be approximated by

$$\theta^B(r,u) = P\{M_r > u\}/rP\{X_1 > u\} \ .$$

Both $\theta^R(r,u)$ and $\theta^B(r,u)$ converge to θ under suitable choices of $r = r_n \to \infty$ and $u = u_n \to x_* = \sup\{x : F(x) < 1\}$. This motivates the use of their sample analogs

$$\hat\theta_n^R = \frac{\sum_{i=1}^n \mathbb{I}\{X_i > u,\ M_{i,i+r-1} \le u\}}{\sum_{i=1}^n \mathbb{I}\{X_i > u\}}\ , \quad \hat\theta_n^B = \frac{\sum_{i=1}^k \mathbb{I}\{M_r^{(i)} > u\}}{\sum_{i=1}^{kr} \mathbb{I}\{X_i > u\}}\ ,$$

where $k = [n/r]$, as *runs* and *blocks estimators* of the extremal index.

In this paper we suggest simple sifficient conditions for consistency and asymptotic normality of those estimators. The results are given in Section 2; they are illustrated by an example. Proofs are given in Section 3.

2 Consistency and Asymptotic Normality

It is assumed throughout, that the threshold $u = u_n$ and the integer $r = r_n$ are chosen so that

$$n\mathbb{P}\{X_1 > u_n\} \to \infty\ , \qquad r_n\mathbb{P}\{X_1 > u_n\} \to 0 \tag{2.1}$$

as $n \to \infty$. Note that (2.1) implies that $r_n = o(n)$. We need the following notation:

$$\begin{aligned} p_n &= \mathbb{P}\{X_1 > u_n\}\ ,\ q_n = \mathbb{P}\{M_r > u_n\}\ ,\ q_n^* = \mathbb{P}\{X_1 > u_n,\ M_{1,r} \le u_n\}\ , \\ \mathbb{I}_{i,r} &= \mathbb{I}\{M_r^{(i)} > u_n\}\ ,\ \mathbb{I}_i = \mathbb{I}\{X_i > u_n\}\ ,\ \mathbb{I}_i^* = \mathbb{I}\{X_i > u_n,\ M_{i,i+r-1} \le u_n\}\ . \end{aligned}$$

For $1 \le m \le n$ we define $\mathcal{F}_{m,n}(u) = \sigma\{\mathbb{I}\{X_i > u\} : m \le i < n\}$, the σ-field generated by the variables involved, and let

$$\varphi(k) := \varphi(k,u) = \sup |\mathbb{P}\{B|A\} - \mathbb{P}\{B\}|\ , \tag{2.2}$$

where the supremum is taken over all sets $A \in \mathcal{F}_{1,m}(u),\ B \in \mathcal{F}_{m+k,\infty}(u)$ such that $\mathbb{P}\{A\}\mathbb{P}\{B\} > 0$ and $m > 1$.

Theorem 1 *Suppose that as $n \to \infty$,*

$$\theta_n^R := \theta^R(r_n, u_n) = q_n^*/p_n \to \theta \tag{2.3}$$

and

$$\gamma_n := \sum_{i=1}^{n}(1 - i/n)\left(\mathbb{P}\{\mathbb{1}_{i+1}^* = 1 | \mathbb{1}_1^* = 1\} - q_n^*\right) = o(np_n) \tag{2.4}$$

for $r = r_n$ and $r = 1$. Then $\quad \hat{\theta}_n^R \underset{p}{\to} \theta \quad$ *as $n \to \infty$.*

Condition (2.4) is weaker than the corresponding one in Theorem 2.1 of Hsing (1993) (where the factor $(1 - i/n)$ seems to be missing):

$$\limsup_{n\to\infty} \left| \sum_{i=1}^{n} \left(\mathbb{P}\{\mathbb{1}_{i+1}^* = 1 | \mathbb{1}_1^* = 1\} - q_n^*\right) \right| < \infty \, .$$

Note that $|\gamma_n| \le \sum_{i=r}^{n} \varphi(1 + i - r)$. Hence, a sufficient condition for (2.4) is the following one:

$$\sum_{i=1}^{n} \varphi(i, u_n) = o(np_n) \ . \tag{2.5}$$

Theorem 2 *Suppose that*

$$\theta_n^B := \theta^B(r_n, u_n) = q_n/rp_n \to \theta \tag{2.6}$$

as $n \to \infty$ and $(k = [n/r])$

$$\delta_n := \sum_{i=1}^{k}(1 - i/k)(\mathbb{P}\{\mathbb{1}_{i+1,r} = 1 \,|\, \mathbb{1}_{1,r} = 1\} - q_n) = o(np_n) \tag{2.7}$$

for $r = r_n$ and $r = 1$. Then $\quad \hat{\theta}_n^B \underset{p}{\to} \theta, \quad$ *as $n \to \infty$.*

Note that $|\delta_n| \le \sum_{i=1}^{k} \varphi(1 + (i-1)r) \le \sum_1^n \varphi(i)$. Hence, (2.7) holds if (2.5) is true.

We allow $\delta_n \to \infty$ though it seems to be bounded in most cases. Moreover, Smith and Weissman (1994), following Hsing et al. (1988) (i.e., assuming all the assumptions needed for compound Poisson convergence of $\sum_1^n \mathbb{1}_i$) argue that $\mathrm{Var} \sum_{i=1}^{k} \mathbb{1}_{i,r} \approx \mathbb{E} \sum_{i=1}^{k} \mathbb{1}_{i,r} \sim np_n\theta$, which means that $\delta_n \to 0$ (and, similarly, $\gamma_n \to 0$).

Consistency of the blocks estimator $\hat{\theta}_n^B$ is proved in Hsing (1991) under more complicated assumptions. Besides (2.1), Hsing assumed that

$$\begin{aligned} \beta_n(l_n; u_n)/r_n p_n + k_n \beta_n(r_n; u_n) &\to 0 \ \text{ for some } \ l_n = o(r_n) \\ \mathbb{E}\, T_r \mathbb{1}\{T_r > np_n\}/r_n p_n &\to 0 \\ \mathbb{E}\, T_r^2 \mathbb{1}\{T_r \le np_n\}/nr_n p_n^2 &\to 0 \end{aligned} \tag{2.8}$$

as $n \to \infty$, where $T_r = \sum_{i=1}^{r} \mathbb{I}\{X_i > u_n\}$ and $\beta(i, u_n)$ is a Rosenblatt strong mixing coefficient for the sequence $\{\mathbb{I}\{X_i > u_n\} : i \geq 1\}$. Conditions (2.3) and (2.6) are necessary and sufficient for $\{X_i\}$ to possess the extremal index θ.

Now we present conditions for the asymptotic normality of $\hat{\theta}_n^R$ and $\hat{\theta}_n^B$. We need the following notation:

$$Y_i = \mathbb{I}_i^* - \theta_n^R \mathbb{I}\{X_i > u_n\} \ , \ Z_i = \mathbb{I}_{i,r} - \theta_n^B \sum_{j=1+(i-1)r}^{ir} \mathbb{I}_i \ .$$

Observe that $\mathbb{E}Y_i = \mathbb{E}Z_i = 0$, Var $Y_i = \theta_n^R(1-\theta_n^R)p_n$.

Theorem 3 *Suppose that $\theta < 1$, conditions (2.3), (2.4) hold and*

$$\sup_n \varphi(kr_n, u_n) \to 0 \qquad (k \to \infty) \ . \tag{2.9}$$

If $(\mathit{Var}\ \sum_{i=1}^n Y_i)/(n\,\mathit{Var}\ Y_1) \to \sigma_R^2$ and $r_n^2 = o(np_n)$ as $n \to \infty$, then

$$\sqrt{np_n}(\hat{\theta}_n^R - \theta_n^R) \Rightarrow \mathcal{N}(0, \sigma_R^2 \theta(1-\theta)) \ . \tag{2.10}$$

Theorem 4 *Suppose that conditions (2.6), (2.7), (2.9) hold, $r_n^4 = o(np_n)$ and $(\mathit{Var} \sum_1^k Z_i)/np_n \to \sigma_B^2$ as $n \to \infty$. Then*

$$\sqrt{np_n}(\hat{\theta}_n^B - \theta_n^B) \Rightarrow \mathcal{N}(0, \sigma_B^2) \ . \tag{2.11}$$

Note that if $(\theta_n^R - \theta) = o(\sqrt{np_n})$ and/or $(\theta_n^B - \theta) = o(\sqrt{np_n})$ then θ_n^R and/or θ_n^B can be replaced by θ in (2.10) and (2.11), respectively.

Hsing (1991) proved the asymptotic normality of $\hat{\theta}_n^B$ under more complicated restrictions. Besides (2.1) and (2.8), he imposed the following assumptions:

$$\begin{array}{ll} \mathbb{E}\{T_r^2 \mathbb{I}\{T_r^2 > \varepsilon\, np_n\} \,|\, T_r > 0\} \to 0 & (\forall\ \varepsilon > 0) \ , \\ \mathbb{E}\{T_r^2 | T_r > 0\} \to \sigma_H^2 & \textit{for some } \sigma_H^2 > 0 \ . \end{array}$$

The last one means that $\mathrm{Var}\{T_r | T_r > 0\} \to \sigma^2 = \sigma_H^2 - \theta^{-2}$ which is the asymptotic variance of a cluster size.

The asymptotic normality of the runs estimator seems to be proven for the first time.

Example Let $\{\xi_i\}, \{\alpha_i\}$ be independent sequences of i.i.d.r.v.'s, $\mathbb{P}\{\xi_i \leq x\} = F(x)$, $\mathbb{P}\{\alpha_i = 1\} = 1 - \mathbb{P}\{\alpha_i = 0\} = 1 - \psi \in (0,1)$. The sequence $\{X_i : i \geq 1\}$ is defined as follows: $X_1 = \xi_1$ and for $i \geq 2$, $X_i = \alpha_i \xi_i + (1-\alpha_i) X_{i-1}$. It is easy to check that the marginal df of $\{X_i\}$ is F, the cluster sizes are geometric with mean $1/(1-\psi)$, hence $\theta = 1 - \psi$. Furthermore, with $\bar{F} = 1 - F$, we have

$$\begin{aligned} \theta^R(r,u) &= \mathbb{P}\{X_1 > u,\ M_{1,r} \leq u,\ \alpha_2 = 1\}/\bar{F}(u) \\ &= \theta \mathbb{P}\{M_{1,r} \leq u\} = \theta F(u) \mathbb{E} F^V(u) \\ &= \theta F(u)[1 - \theta \bar{F}(u)]^{r-2}; \end{aligned} \tag{2.12}$$

here $V = \sum_{i=3}^{r} \alpha_i$ stands for a binomial r.v. with parameters $(r-2, \theta)$. Similarly one has

$$\theta^B(r,u) = \{1 - F(u)(1-\theta\bar{F}(u))^{r-1}\}/r\bar{F}(u) \,. \tag{2.13}$$

Under (2.1),

$$\theta_n^R = \theta - \theta^2 r p_n + O(p_n) \tag{2.14}$$

and

$$\theta_n^B = \theta - \frac{1}{2}\,\theta^2 r p_n + \frac{1-\theta}{r} \;+\; o(1/r + rp_n) \,. \tag{2.15}$$

Now, for the function $\varphi(k)$ we claim that

$$\varphi(k) \le \psi^k \qquad (k \ge 1) \,. \tag{2.16}$$

Indeed, suppose $A \in \sigma(X_1, \ldots, X_m),\ B \in \sigma(X_{m+k}, \ldots)$ and let ζ be the length of a 0-run starting at α_{m+1} (we put $\zeta = 0$ if $\alpha_{m+1} = 1$). Then

$$\mathbb{P}\{B, \zeta < k | A\} - \mathbb{P}\{B, \zeta < k\} = 0$$

and

$$\mathbb{P}\{B, \zeta \ge k | A\} \le \mathbb{P}\{\zeta \ge k | A\} = \mathbb{P}\{\zeta \ge k\} = \psi^k.$$

This implies (2.16).

One can verify that $EY_iY_{i+j} = 0$ $(i, j \ge 1)$ and $\sigma_R^2 = 1$. If we choose $r = r_n,\ u = u_n$ to satisfy (2.1), (2.3) and

$$r_n^2 = o(np_n),\ nr^2p_n^3 = o(1) \ , \tag{2.17}$$

all the assumptions of Theorems 2.1 and 2.3 are satisfied. Thus, $\hat{\theta}_n^R$ is consistent, asymptotically unbiased and

$$\sqrt{np_n}(\hat{\theta}_n^R - \theta) \Rightarrow \mathcal{N}(0, \theta(1-\theta)) \,. \tag{2.18}$$

Similar calculations show that Var $\sum_1^k Z_i = np_n\theta(1-\theta) + o(np_n)$, hence $\sigma_B^2 = \theta(1-\theta)$. In view of (2.14) and (2.15), θ_n^R is a better approximation for θ than θ_n^B. Moreover, under (2.17) one has $\sqrt{np_n}(\theta_n^B - \theta) \to \infty$. Hence, one cannot replace θ_n^B by θ in (2.11). Smith and Weissman (1994) also conclude that the runs estimator is preferred based on bias considerations.

Hsing (1993) argues that for a large class of processes

$$\theta^R(r,u) - \theta = L(\bar{F}(u))(\bar{F}(u))^\rho, \tag{2.19}$$

where $L(\cdot)$ varies slowly at 0, $r \ge 2$ is a constant and $\rho > 0$. Smith and Weissman (1994) suppose that for a wide class of processes

$$\theta^R(r,u) - \theta = O(r\bar{F}(u) + \beta^r) \qquad\qquad (\exists \beta \in (0,1)) \tag{2.20}$$

if $r\bar{F}(u) \to 0$. In our example, (2.20) holds with $\beta = 0$ and (2.19) with $r = 2:$

$$\theta^R(2,u) = \theta(1 - \bar{F}(u)) \,. \tag{2.21}$$

Note that in the special situation considered by Novak (1993), θ was in fact calculated up to $O(p_n)$. This together with (2.21) allows one to expect that in many cases $\theta^R(r,u) - \theta = O(\bar{F}(u))$.

3 Proofs

Proof of Theorem 2. Note that $\mathbb{E}\sum_{i=1}^{k}\mathbb{I}_{i,r} = k\theta_n^B r p_n = kq_n$ and $\mathbb{E}\sum_{i=1}^{n}\mathbb{I}_i = np_n$ — these are the expectations of the numerator and denominator of $\hat{\theta}_n^B$. We calculate

$$\begin{aligned}
\mathrm{Var}\sum_{i=1}^{k}\mathbb{I}_{i,r} &= \sum_{i=1}^{k}\mathrm{Var}\,\mathbb{I}_{i,r} + 2\sum\sum_{i<j}\mathrm{Cov}\,\mathbb{I}_{i,r}\mathbb{I}_{j,r} \\
&= kq_n\left(1 - q_n + 2\sum_{j=1}^{k}(1-j/k)\left(\mathbb{P}\{M_r^{(j+1)} > u | M_r^{(1)} > u\} - q_n\right)\right) \\
&= kq_n(1 - q_n + 2\delta_n)\,. \qquad (3.1)
\end{aligned}$$

Since $k = [n/r]$,

$$q_n = \theta_n^B r p_n \to 0,\ \ kq_n \sim \theta_n^B np_n \to \infty,\ \ \delta_n = o(np_n)\,,$$

the right-hand side of (3.1) is $o((np_n)^2)$. Thus, by Chebychev inequality,

$$\sum_{i=1}^{k}\mathbb{I}_{i,r}/kq_n \underset{p}{\to} 1 \qquad (3.2)$$

as $n \to \infty$. When $r = 1$, (3.2) implies $\sum_{i=1}^{n}\mathbb{I}_i/np_n \underset{p}{\to} 1$. Hence,

$$\hat{\theta}_n^B \underset{p}{\sim} kq_n/np_n \sim \theta_n^B \to \theta$$

as $n \to \infty$. □

Proof of Theorem 1. Note first that $\mathbb{E}\sum_{i=1}^{n}\mathbb{I}_i^* = nq_n^* = n\theta_n^R p_n$. Similarly to (3.1) we show that

$$\begin{aligned}
\mathrm{Var}\sum_{1}^{n}\mathbb{I}_i^* &= nq_n^*\left(1 - q_n^* + 2\sum_{i=1}^{n}(1-i/n)\left(\mathbb{P}\{\mathbb{I}_{1+i}^* = 1 \,|\, \mathbb{I}_1^* = 1\} - q_n^*\right)\right) \\
&= nq_n^*(1 - q_n^* + 2\gamma_n)\ . \qquad (3.3)
\end{aligned}$$

The rest of the proof follows as before, since we assume $\gamma_n = o(np_n)$. □

Proof of Theorem 3. The proof is based on the following result of Utev (1990):

Let $\{\xi_{i,n} : 1 \le i \le k_n\}_{n\ge1}$ be a triangular array of r.v.'s, $S_n = \sum_{i=1}^{k_n}\xi_{i,n}$, $\sigma_n^2 = Var\ S_n$. Let $\varphi_n(l)$ be the corresponding mixing coefficient. If for some sequence of integers $\{j_n\}$

$$\sup_n \varphi_n(lj_n) \to 0 \qquad (l \to \infty) \qquad (3.4)$$

and

$$\lim_{n\to\infty} j_n\sigma_n^{-2}\sum_{i=1}^{k_n}\mathbb{E}\xi_{i,n}^2\mathbb{I}\{|\xi_{i,n}|>\varepsilon\sigma_n/j_n\}=0 \qquad (\forall\,\varepsilon>0) \tag{3.5}$$

then

$$S_n/\sigma_n \Rightarrow \mathcal{N}(0,1) \qquad (n\to\infty)\,. \tag{3.6}$$

Consider the identity

$$\frac{\sqrt{np_n}(\hat\theta_n^R-\theta_n^R)}{np_n/\sum_1^n\ \mathbf{1}_i} \;=\; \frac{\sum_1^n Y_i}{\sqrt{np_n}} \quad . \tag{3.7}$$

In view of Theorem 1, it is enough to show that

$$\sum_1^n Y_i/\sqrt{np_n} \;\Longrightarrow\; \mathcal{N}(0,\sigma_R^2\theta(1-\theta)) \tag{3.8}$$

as $n\to\infty$. Recall that

$$\mathrm{Var}\ Y_1 = p_n\theta_n^R(1-\theta_n^R)\ ,\ \sigma_n^2 = \mathrm{Var}\ \sum_1^n Y_i \sim np_n\theta_n^R(1-\theta_n^R)\sigma_R^2\ .$$

Conditions (3.4), (3.5) hold by assumption with $j_n=r_n, k_n=n, \xi_{i,n}=Y_i$. Hence, (3.6) implies (3.8). This completes the proof. □

Proof of Theorem 4. The condition $r_n^4=o(np_n)$ as $n\to\infty$ implies (3.5). The rest of the proof runs along similar lines. □

Acknowledgement This research was supported by Grant No. 9200008 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel. The article was written when S.Novak visited the Technion in 1995. The support and hospitality of the Technion and BSF are greatly appreciated.

References

[1] Chernick, M.R., Hsing, T. and McCormick, W.P. (1991), Calculating the extremal index for a class of stationary sequences, *Adv. Appl. Probab.* **23**, 835–850.

[2] Hsing, T. (1991), Estimating the parameters of rare events, *Stochastic Processes Appl.* **37** (1), 117–139.

[3] Hsing, T. (1993), Extremal index estimation for a weakly dependent stationary sequence, *Ann. Statist.* **21** (4), 2043–2071.

[4] Hsing, T., Hüsler, J. and Leadbetter, M.R. (1988), On the exceedence point process for stationary sequence, *Probab. Theory and Related Fields* **78**, 97–112.

[5] Ibragimov, I.A. and Linnik, Yu.V. (1969), *Independent and Stationary Sequences of Random Variables*, Groningen: Wolters-Noordhoff.

[6] Leadbetter, M.R. (1983), Extremes and local dependence in stationary sequences, *Z. Wahr. Ver. Geb.* **65**, 291–306.

[7] Leadbetter, M.R. and Rootzèn, H. (1983), Extremal theory for stochastic processes, *Ann. Probab.* **16** (2), 431–438.

[8] Leadbetter, M.R., Lindgren, G. and Rootzèn, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer Verlag, 366 pp.

[9] Leadbetter, M.R., Weissman, I., de Haan, L. and Rootzèn, H. (1989), On clustering of high levels in statistically stationary series, in: *Proc. 4th Intern. Meet. Statistical Climatology* (J. Samson, ed.), Willington: New Zealand Meteorological Service.

[10] Novak, S. Yu. (1993), On the asymptotic distribution of the number of random variables exceeding a given level, *Siberian Adv. Math.* **3** (4), 108–122.

[11] O'Brien, G.L. (1974), The maximum term of uniformly mixing sequence, *Z. Wahr. Ver. Geb.* **30**, 57–63.

[12] O'Brien, G.L. (1987), Extreme values for stationary and Markov sequences, *Ann. Probab.* **15** (1), 281–291.

[13] Smith, R.L. and Weissman, I. (1994), Estimating the extremal index, *J. R. Statist. Soc. B* **56** (3), 515–528.

[14] Utev, S.A. (1990), On the central limit theorem for φ-mixing triangle arrays of random variables, *Theory Probab. Appl.* **35**, 131-139.