Department of Economics
School of Social Sciences

A New Family of Estimators for the Extremal Index

José Olmo[1]
City University

Department of Economics
Discussion Paper Series
No. 06/01

[1] Department of Economics, City University, Northampton Square, London, EC1V 0HB, UK. Email: j.olmo@city.ac.uk

A New family of Estimators for the Extremal Index

Jose Olmo [*†‡]

Department of Economics

City University, London.

This version, January 2006

Abstract

The extremal index (θ) is the key parameter for extending extreme value theory results from *IID* to stationary sequences. It determines the extent of clustering found in the largest observations of a stationary sequence $\{X_i\}$. This paper introduces an alternative interpretation of θ as a ratio of the limiting expected value of two random variables defined by *extreme levels* u_n, v_n and a partition of the stationary sequence into blocks. These random variables consist on elements of the sequence of block maxima exceeding such levels. The estimator of θ derived from this interpretation is simple and follows a binomial distribution. This estimator is asymptotically unbiased in contrast to other estimators for θ (blocks method and runs method). Under certain conditions this methodology can be extended to *moderately high* levels $\widetilde{u}_n$ and $\widetilde{v}_n$. The estimator obtained in this context is consistent. Furthermore, it has a binomial distribution that converges to a normal distribution with mean θ. This family of estimators outperform the rest of candidates commonly used to estimate θ. Some simulation experiments reinforce these findings. These experiments highlight the importance of block size selection and provide some guidance to proceed in practice with the estimation of the extremal index.

[*]AMS 2000 subject classification. Primary-62F12; secondary-62G05

[†]Key words. Extremal index, extreme value theory, order statistic

[‡]Address for correspondence: City University, London, Social Sciences Building, Room D309, Northampton Square, EC1V0HB London. E-mail: j.olmo@city.ac.uk.

1 Introduction

Consider an *iid* random sample, of size n, from an unknown distribution, F, and let G be the limiting distribution of the sample maximum, $M_{1,n}$. Under some regularity conditions on the tail of F, and for some suitable constants $a_n > 0$, b_n,

$$P\{a_n^{-1}(M_{1,n} - b_n) \leq x\} \to G(x), \tag{1}$$

where G must be one of the following types (see de Haan (1976)),

Type I: (Gumbel) $\quad G(x) = e^{-e^{-x}}, \quad -\infty < x < \infty.$

Type II: (Fréchet) $\quad G(x) = \begin{cases} 0 & x \leq 0, \\ e^{-x^{-\frac{1}{\xi}}} & x > 0, \xi > 0. \end{cases}$

Type III: (Weibull) $\quad G(x) = \begin{cases} 1 & x \geq 0, \\ e^{-(-x)^{-\frac{1}{\xi}}} & x < 0, \xi < 0. \end{cases}$

This important result may be extended to study the maximum of a wide class of dependent processes. We concentrate here on stationary sequences where the dependence is restricted by different distributional *mixing* conditions. We distinguish two types of dependence: long range and short range dependence. To limit the first type of dependence we assume the distributional mixing condition $D(u_n)$ of Leadbetter (1983). This mixing condition is said to hold for a sequence $\{u_n\}$ if for any integers $1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_{p'} \leq n$ for which $j_1 - i_p \geq l$, we have

$$D(u_n): \quad \left| F_{i_1,\ldots,i_p,j_1,\ldots,j_{p'}}(u_n) - F_{i_1,\ldots,i_p}(u_n) F_{j_1,\ldots,j_{p'}}(u_n) \right| \leq \alpha_{n,l}, \tag{2}$$

where $\alpha_{n,l_n} \to 0$ as $n \to \infty$ for some $l_n = o(n)$, and $F_{i_1,\ldots,i_p}(u_n)$ denotes $P\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\}$. This condition entails that

$$\begin{aligned} |P\left\{(X_{i_1} > u_n \; or \ldots or \; X_{i_p} > u_n) \bigcap (X_{j_1} > u_n \; or \ldots or \; X_{j_{p'}} > u_n)\right\} - \\ P\left\{X_{i_1} > u_n \; or \ldots or \; X_{i_p} > u_n\right\} P\left\{X_{j_1} > u_n \; or \ldots or \; X_{j_{p'}} > u_n\right\}| \to 0 \text{ as } n \to \infty. \end{aligned}$$

This condition only concerns events of the form $\{X_i > u_n\}$ in contrast to more restrictive mixing conditions, for example the strong mixing condition introduced in Rosenblatt (1956). $D(u_n)$ alone is sufficient to extend the central result given in (1) to stationary sequences. In particular for weak short range dependence, $a_n > 0$ and b_n are the same of the *iid* case. In this case stationary sequences satisfy a more restrictive mixing condition, denoted $D'(u_n)$ in Leadbetter (1983). This condition precludes the presence of clustering in the extreme values.

It is as follows,

$$D'(u_n): \quad \underset{n\to\infty}{limsup} \quad n\sum_{j=2}^{[n/k_n]} P\{X_1 > u_n, X_j > u_n\} \to 0 \quad as \quad k_n \to \infty, \tag{3}$$

with k_n a sequence that defines a partition of the sample, and $[\cdot]$ denoting integer value. More generally, for a stationary sequence $\{X_i\}$ satisfying only $D(u_n)$ with $u_n = a_n x + b_n$, we have

$$P\{M_{1,n} \leq u_n\} \to G^{\theta}(x), \tag{4}$$

with $0 \leq \theta \leq 1$ the extremal index.

There are different interpretations of the extremal index. This concept, originated in papers by Loynes (1965), O'Brien (1974) and developed in detail by Leadbetter (1983), reflects the effect of clustering of extreme observations on the limiting distribution of the maximum.

Loynes (1965) under mixing conditions different from $D(u_n)$ and $D'(u_n)$ found that

$$P\{M_{1,n} \leq u_n\} = F^{n\theta}(u_n). \tag{5}$$

O'Brien (1987) showed that the presence of clustering affected the limiting distribution of block maxima. He found that

$$P\{M_{2,r_n} \leq u_n | X_1 > u_n\} \to \theta, \tag{6}$$

with $M_{2,r_n} = \max\{X_2, \ldots, X_{r_n}\}$, and r_n such that $r_n \to \infty$ and $r_n = o(n)$.

Alternatively Leadbetter (1983) showed that for stationary sequences exhibiting short range dependence the inverse of the extremal index is the limiting mean number of exceedances of u_n in an interval of length r_n. This mathematically reads as follows

$$E\left[\sum_{j=1}^{r_n} I(X_j > u_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1\right] \to \theta^{-1}, \tag{7}$$

with $I(X > u_n)$ the indicator function. By stationarity this property is satisfied for any block of r_n consecutive elements defined in the sequence.

Finally, Hsing (1993) and Ferro and Segers (2003) use a reinterpretation of (4),

$$P\{M_{1,n} \leq u_n\} \to e^{-\theta\tau(x)}, \; 0 < \tau(x) < \infty, \tag{8}$$

to provide two more characterizations of the extremal index. Hsing shows that the distribution of $n(1 - F(M_{1,n}))$ is well approximated by an exponential distribution with mean θ^{-1}, and

Ferro and Segers find that the process of interexceedance times determined by observations exceeding u_n follows asymptotically the exponential distribution $Exp(\theta^{-1})$.

This paper presents an alternative characterization of the extremal index that permits to present an intuitive estimation procedure. In contrast to other estimators for θ the family of estimators introduced herein allow to derive statistical inference about the parameter. The finite-sample and asymptotic distributions of these estimators are found under weak conditions. A byproduct of these findings is that it is possible testing the presence of serial clustering of extreme values in stationary sequences.

The paper is structured as follows. Section 2 introduces a characterization of the extremal index as a ratio of the limiting expected value of two random variables defined by extreme levels and derived from the asymptotic properties of the distribution of the maximum. This characterization of θ is extended to cover the case of exceedances of lower levels denoted herein moderately high levels. Section 3 introduces natural estimators for this parameter based on these characterizations of θ. The finite-sample as well as the limiting distributions of these estimators are derived. This section also reviews some statistical properties (bias and variance) of other well known estimators of θ: logs, blocks and runs method. A simulation experiment for time series exhibiting clustering of extreme values is conducted in Section 4. In particular the analysis of coverage probabilities derived from gaussian confidence intervals for these new estimators of θ. Finally some conclusions and guidelines for further research are found in Section 5.

2 Characterization of the extremal index

Let $\{X_i, i \geq 1\}$ be an *iid* sequence of n observations with marginal distribution function F and let $M_{1,n} = max\{X_1, \ldots, X_n\}$ be the sample maximum of the sequence. This sequence satisfies condition (1) if and only if

$$\lim_{x \uparrow x_F} \frac{1 - F(x)}{1 - F(x^-)} = 1, \tag{9}$$

with $x_F = \sup\{x | F(x) < 1\} \leq +\infty$ denoting the right end point of F, and $F(x_F^-) = \lim_{x \uparrow x_F} F(x)$. This condition precludes the existence of jumps in the right tail of the distribution.

If (9) holds condition (1) is equivalent to

$$n(1 - F(u_n)) \to \tau(x) \quad as \quad n \to \infty, \tag{10}$$

with $u_n = a_n x + b_n$ sufficiently high. The proof of this result is immediately derived from

$$P\{M_{1,n} \leq u_n\} = F^n(u_n) = \left(1 - \frac{n(1 - F(u_n))}{n}\right)^n.$$

If u_n is sufficiently high, $1 - F(u_n) \to 0$, conditions (9) and (10) are sufficient to define a random variable $Z_{u_n} = \sum_{j=1}^{n} I(X_j > u_n)$ that converges in distribution to a Poisson random variable with mean $\tau(x)$, see Hodges and Le Cam (1960) or Lehman (1999, p. 105.)

Suppose now $\{X_i, i \geq 1\}$ is a stationary sequence. If $D(u_n)$ and $D'(u_n)$ hold the above results for $M_{1,n}$ and Z_{u_n} still hold. However if condition $D'(u_n)$ is relaxed the limiting distribution of $M_{1,n}$ is

$$P\{M_{1,n} \leq u_n\} \to e^{-\theta\tau(x)},\ 0 < \tau(x) < \infty, \tag{11}$$

and we can construct a partition of the sequence $\{X_i\}$ of length n in k_n blocks of size r_n with $k_n \to \infty$, $k_n = o(n)$, $k_n l_n = o(n)$ with l_n introduced in (2), and $r_n = [n/k_n]$ such that

$$k_n \left(1 - F_{1,\ldots,r_n}(u_n)\right) \to \theta\tau(x). \tag{12}$$

It can be seen that this condition is sufficient to show the existence of the extremal index, see Leadbetter (1983). This author also shows the equivalence of (11) and (12) provided by the approximation of $P\{M_{1,n} \leq u_n\}$ by $P^{k_n}\{M_{1,r_n} \leq u_n\}$ under $D(u_n)$.

We will suppose hereafter that $D'(u_n)$ does not hold. In this context the random variable Z_{u_n} does not consist on independent elements and in general no longer converges in distribution to a Poisson law. Nonetheless this random variable can be *thinned* to eliminate the presence of serial dependence in the extremes. The thinning process consists on dividing the sequence in k_n blocks of size r_n and choosing the block maxima that exceed the level u_n. This method allows to define a new random variable denoted $Z^*_{u_n} = \sum_{j=1}^{k_n} I(M_{(j-1)r_n+1,jr_n} > u_n)$. This random variable follows a binomial distribution for n sufficiently high. By (12) this distribution converges to a Poisson distribution with parameter $\theta\tau(x)$. Note that u_n is really a family of sequences $u_n(x)$. If one considers certain sequence u_n x is fixed and $\tau(x)$ takes a constant value τ.

Leadbetter (1983) uses this thinning to define a point process $N_t^{(u_n)}$ on the interval $(0, 1]$ consisting on the elements of $Z^*_{u_n}$ indexed by $t = j/k_n$, $j = 1, \ldots, k_n$. This point process converges to a Poisson process N with mean $\theta\tau$, see Leadbetter (1983) and Leadbetter et al (1983). The core of this result is that

$$E\left[\sum_{j=1}^{r_n} I(X_j > u_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1\right] \to \theta^{-1}.$$

Using similar arguments we can define an *extreme level* v_n characterized by the following condition,

$$E\left[\sum_{j=1}^{r_n} I(X_j > v_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1\right] \to 1. \tag{13}$$

It is immediate to see that $v_n \geq u_n$. Furthermore,

$$E\left[\sum_{j=1}^{r_n} I(X_j > v_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1\right] = \frac{r_n P\{X_j > v_n\}}{P\left\{\bigcup_{j=1}^{r_n} (X_j > u_n)\right\}} \to 1.$$

It follows that

$$n\,(1 - F(v_n)) \to \theta\tau(x), \quad with \quad 0 < \tau(x) < \infty \quad as \quad n \to \infty. \tag{14}$$

$D(v_n)$ holds for $v_n \geq u_n$. Then (14) implies

$$P\{M_{1,n} \leq v_n\} \to e^{-\theta^2 \tau(x)} \quad as \quad n \to \infty, \tag{15}$$

see (10) and (11). For appropriate sequences k_n and r_n this is equivalent to

$$k_n\,(1 - F_{1,\ldots,r_n}(v_n)) \to \theta^2\tau(x) \quad as \quad n \to \infty. \tag{16}$$

This condition determines a second *thinning* of Z_{u_n}. This is determined by the extreme level v_n that defines a new random variable $Z^*_{v_n} = \sum_{j=1}^{k_n} I(M_{(j-1)r_n+1,jr_n} > v_n)$ following a binomial distribution. This random variable determines a point process $N_t^{(v_n)}$ that converges in distribution to a Poisson process with intensity $\theta^2\tau$.

Definition 2.1. *The extremal index is the ratio of the limiting expected value of the point processes $N_t^{(v_n)}$ and $N_t^{(u_n)}$. The extremal index reads as*

$$\theta = \lim_{n\to\infty} \frac{E[N_t^{(v_n)}]}{E[N_t^{(u_n)}]}. \tag{17}$$

In terms of random variables,

$$\theta = \lim_{n\to\infty} \frac{E[Z^*_{v_n}]}{E[Z^*_{u_n}]}. \tag{18}$$

This characterization of the extremal index can be extended to lower levels denoted hereafter moderately high levels. The counterpart of u_n is denoted $\widetilde{u}_n$. The variable $Z^*_{\widetilde{u}_n}$ is hence defined as $Z^*_{\widetilde{u}_n} = \sum_{j=1}^{k_n} I(M_{(j-1)r_n+1,jr_n} > \widetilde{u}_n)$.

Assumptions.- We will assume throughout that $\widetilde{u}_n$ satisfies the following:

A.1. $D(\widetilde{u}_n)$.

A.2. $k_n(1 - F_{1,\ldots,r_n}(\widetilde{u}_n)) \to \infty$ as $n \to \infty$.

A.3. $1 - F_{1,\ldots,r_n}(\widetilde{u}_n) = (1 - F_{1,\ldots,r_n}(u_n))s_n$ with $s_n = o\left(\frac{1}{1-F_{1,\ldots,r_n}(u_n)}\right)$.

A.4. $1 - F(\widetilde{u}_n) = (1 - F(u_n))s'_n$ with $s'_n = o\left(\frac{1}{1-F(u_n)}\right)$.

A.5. $\frac{s'_n}{s_n} \to 1 \quad as \quad n \to \infty$.

A sequence $\widetilde{u}_n$ satisfying A.1.-A.5. is denominated *moderately high level*.

Result 2.1. *Suppose $\widetilde{u}_n$ is a level satisfying **A.1.-A.5.** and let c_n be a realization of $Z^*_{\widetilde{u}_n}$. If c_n satisfies*

$$\frac{c_n - \theta\tau s_n}{\left(\theta\tau s_n F_{1,\ldots,r_n}(\widetilde{u}_n)\right)^{1/2}} \to \lambda, \tag{19}$$

then

$$P\left\{Z_{\widetilde{u}^*_n} \leq c_n\right\} \to \Phi(\lambda) \quad as \quad n \to \infty, \tag{20}$$

with $\Phi(\cdot)$ a standard normal distribution.

Proof.- If A.1. holds individual contributions to $Z^*_{\widetilde{u}_n}$ are *almost* independent (converge to an *iid* sequence as n increases). Each contribution is a bernoulli random variable with probability of success $1 - F_{1,\ldots,r_n}(\widetilde{u}_n)$. Hence the finite-sample distribution of the sum of $I(M_{(j-1)r_n+1,jr_n} > \widetilde{u}_n)$ with $j = 1, \ldots, k_n$ is well approximated by a binomial distribution of k_n observations and parameter $1 - F_{1,\ldots,r_n}(\widetilde{u}_n)$ denoted hereafter $Bin(1 - F_{1,\ldots,r_n}(\widetilde{u}_n), k_n)$. Furthermore if A.2. and A.3. hold the Berry-Essen bound applies (see Feller (Vol 2) (1966)). It follows that

$$\left| P\left\{Z_{\widetilde{u}^*_n} \leq c_n\right\} - \Phi\left(\frac{c_n - \theta\tau s_n}{\left(\theta\tau s_n F_{1,\ldots,r_n}(\widetilde{u}_n)\right)^{1/2}}\right)\right| \leq \frac{C}{\sqrt{k_n(1 - F_{1,\ldots,r_n}(\widetilde{u}_n))F_{1,\ldots,r_n}(\widetilde{u}_n)}}$$

with $C > 0$. This converges to zero as n increases. □

The event $\{Z_{\widetilde{u}^*_n} \leq c_n\}$ is equivalent to $\{M_{c_n+1:k_n} \leq \widetilde{u}_n\}$ with $M_{c_n+1:k_n}$ an element of the sequence of order statistics $M_{1:k_n} \geq \ldots \geq M_{k_n:k_n}$. Hence its limiting distribution is

$$P\{M_{c_n+1:k_n} \leq \widetilde{u}_n\} \to \Phi(\lambda) \quad as \quad n \to \infty. \tag{21}$$

This limiting distribution characterizes an intermediate order statistic, see Leadbetter et al. (1983, p. 44). Hence the name *moderately high level* for $\widetilde{u}_n$.

Following the same notation the counterpart of the level v_n is denoted $\widetilde{v}_n$. This sequence is chosen to be a *moderately high level*. It is characterized by the following properties.

Properties.-

B.1. $k_n(1-F_{1,\ldots,r_n}(\widetilde{v}_n)) \to \infty$ as $n \to \infty$.

B.2. $1-F_{1,\ldots,r_n}(\widetilde{v}_n) = (1-F_{1,\ldots,r_n}(v_n))t_n$ with $t_n = o\left(\frac{1}{1-F_{1,\ldots,r_n}(v_n)}\right)$.

B.3. $1-F(\widetilde{v}_n) = (1-F(v_n))t'_n$ with $t'_n = o\left(\frac{1}{1-F(v_n)}\right)$.

B.4. $\frac{t'_n}{t_n} \to 1 \quad as \quad n \to \infty$.

B.5.

$$E\left[\sum_{j=1}^{r_n} I(X_j > \widetilde{v}_n) | \sum_{j=1}^{r_n} I(X_j > \widetilde{u}_n) \geq 1\right] \to 1.$$

Property **B.5.** implies

$$\frac{n(1-F(\widetilde{v}_n))}{k_n(1-F_{1,\ldots,r_n}(\widetilde{u}_n))} \to 1 \quad as \quad n \to \infty.$$

This yields $\frac{t'_n}{s_n} \to 1$ and in turn $\frac{t_n}{s_n} \to 1$ as $n \to \infty$. Therefore

$$\frac{k_n(1-F_{1,\ldots,r_n}(\widetilde{v}_n))}{k_n(1-F_{1,\ldots,r_n}(\widetilde{u}_n))} \to \theta \quad as \quad n \to \infty. \tag{22}$$

From the previous properties it is clear that $\widetilde{v}_n \geq \widetilde{u}_n$ and condition $D(\widetilde{v}_n)$ holds. This level determines $Z^*_{\widetilde{v}_n} = \sum_{j=1}^{k_n} I(M_{(j-1)r_n+1,jr_n} > \widetilde{v}_n)$.

Result 2.2. *Suppose $\widetilde{u}_n$ is a level satisfying assumptions **A.1.-A.5.** and for some c_n condition (19) holds. If $\widetilde{v}_n$ satisfies **B.1.-B.5.** then*

$$P\left\{Z_{\widetilde{v}^*_n} \leq \theta c_n\right\} \to \Phi(\lambda) \quad as \quad n \to \infty, \tag{23}$$

with $\Phi(\cdot)$ a standard normal distribution.

Proof.- The methodology is identical to the proof in result (2.1).

$$\left| P\left\{Z_{\widetilde{v}^*_n} \leq \theta c_n\right\} - \Phi\left(\frac{\theta c_n - \theta^2 \tau t_n}{(\theta^2 \tau t_n F_{1,\ldots,r_n}(\widetilde{v}_n))^{1/2}}\right)\right| \leq \frac{C}{\sqrt{k_n(1-F_{1,\ldots,r_n}(\widetilde{v}_n))F_{1,\ldots,r_n}(\widetilde{v}_n)}}$$

with $C > 0$.

Operating in (22) we obtain $F_{1,\ldots,r_n}(\widetilde{v}_n) - ((1-\theta) + \theta F_{1,\ldots,r_n}(\widetilde{u}_n)) \to 0$. Therefore the limit of $\frac{\theta c_n - \theta^2 \tau t_n}{(\theta^2 \tau t_n F_{1,\ldots,r_n}(\widetilde{v}_n))^{1/2}}$ can be written as

$$\frac{1}{\frac{(\theta^2 \tau t_n (1-\theta) + \theta^3 \tau t_n F_{1,\ldots,r_n}(\widetilde{u}_n))^{1/2}}{\theta c_n - \theta^2 \tau t_n}}.$$

This expression is of the same order than $\frac{\theta(c_n - \theta \tau s_n)}{\theta(\theta \tau s_n F_{1,\ldots,r_n}(\widetilde{u}_n))^{1/2}}$ that converges to λ if $\widetilde{u}_n$ satisfies

(19). Then

$$P\left\{Z_{\widetilde{v}_n^*} \leq \theta c_n\right\} \to \Phi(\lambda) \quad as \quad n \to \infty. \qquad \square$$

These results extend the characterization of the extremal index given in (18) to *moderately high levels*. This is

$$\theta = \lim_{n\to\infty} \frac{E[Z^*_{\widetilde{v}_n}]}{E[Z^*_{\widetilde{u}_n}]}. \tag{24}$$

3 Estimation of the extremal index

The extremal index provides a measure of the clustering of the largest observations of a stationary sequence. If there is clustering the distribution of $M_{1,n}$ is $F^{n\theta}(u_n)$ instead of $F^n(u_n)$. This result generates the first estimator of the extremal index. For appropriate sequences k_n, r_n it holds that $P^{k_n}\{M_{1,r_n} \leq u_n\}$ approximates $P\{M_{1,n} \leq u_n\}$. Taking logs in both expressions we have $\theta = \frac{logP\{M_{1,r_n} \leq u_n\}}{r_n logF(u_n)}$. Thus a natural estimator for the extremal index is

$$\hat{\theta}_n^{(1)} = \frac{log(1 - Z^*_{u_n}/k_n)}{r_n log(1 - Z_{u_n}/n)}. \tag{25}$$

The empirical distribution Z_{u_n}/n is a simple estimator of $1 - F(u_n)$, and $Z^*_{u_n}/k_n$ of $1 - F_{1,\ldots,r_n}(u_n)$. This estimator of θ is denoted the logs method.

Alternatively, the concept of extremal index introduced by Leadbetter (1983), θ^{-1} the limiting mean cluster size of the exceedances, yields the blocks method

$$\hat{\theta}_n^{(2)} = \frac{Z^*_{u_n}}{Z_{u_n}}. \tag{26}$$

This estimator can be regarded as an approximation of $\hat{\theta}_n^{(1)}$ using the first order expansions of the logarithm for numerator and denominator.

The characterization of θ in O'Brien (1987) and in Hsing (1993) motivate a different method to estimate the parameter. It is as follows

$$\overline{\theta}_n = \frac{W_{u_n}}{Z_{u_n}} \tag{27}$$

where $W_{u_n} = \sum_{i=1}^{n-r_n} I(X_i > u_n)(1 - I(X_{i+1} > u_n)) \cdot \cdot (1 - I(X_{i+r_n} > u_n))$. This method gives rise to the runs estimator.

The methodology introduced herein yields very straightforward estimators for θ. We will

use $\widetilde{\theta}_n$ to denote the estimator of θ based on extreme levels. It takes this expression

$$\widetilde{\theta}_n = \frac{Z^*_{v_n}}{Z^*_{u_n}}. \tag{28}$$

For the case of *moderately high levels* the estimator takes this form

$$\widetilde{\widetilde{\theta}}_n = \frac{Z^*_{\widetilde{v}_n}}{Z^*_{\widetilde{u}_n}}. \tag{29}$$

We will develop in detail the first estimator. The empirical counterpart of (13) takes this form

$$\frac{1}{Z^*_{u_n}} \sum_{j=1}^{k_n} \sum_{i=(j-1)r_n+1}^{jr_n} I(X_i > v_n) \to 1 \quad as \quad n \to \infty. \tag{30}$$

This can be written as

$$\frac{Z_{v_n}}{Z^*_{u_n}} \to 1 \quad as \quad n \to \infty. \tag{31}$$

In practice this relationship is exactly satisfied for $v_n = X_{Z^*_{u_n}+1:n}$, with $X_{Z^*_{u_n}+1:n}$ an order statistic of $\{X_i\}$. This statistic is an extreme order statistic by definition of u_n. An appropriate candidate for this level if (12) exactly holds is $u_n = M_{c+1:k_n}$ with $c = \theta\tau$ fixed.

The algorithm for $\widetilde{\theta}_n$ is sketched as follows.

Algorithm 3.1. .

1. *Consider appropriate sequences k_n and r_n.*
2. *Construct $M_{1,r_n}, M_{r_n+1,2r_n}, \ldots, M_{(k_n-1)r_n+1,n}$ from $\{X_i\}$ with $i = 1, \ldots, n$.*
3. *u_n is an extreme level. Suppose $u_n = M_{c+1:k_n}$ for some fixed c (c small).*
4. *$Z^*_{u_n} = c$.*
5. *From (30), $Z_{v_n} = Z^*_{u_n}$ with $Z_{v_n} = \sum_{i=1}^{n} I(X_i > v_n)$.*
6. *Then $v_n = X_{c+1:n}$.*
7. *Compute $Z^*_{v_n} = \sum_{j=1}^{k_n} I(M_{(j-1)r_n+1,jr_n} > v_n)$.*
8. *$\widetilde{\theta}_n = \frac{Z^*_{v_n}}{Z^*_{u_n}}$.*

This estimator may be interpreted as a refinement of the blocks method where the level u_n in (26) is replaced by v_n.

The procedure for $\widetilde{\widetilde{\theta}}_n$ is similar. In this case $Z_{\widetilde{v}_n} = Z^*_{\widetilde{u}_n}$ and $\widetilde{v}_n = X_{Z^*_{\widetilde{u}_n}+1:n}$ with $X_{Z^*_{\widetilde{u}_n}+1:n}$ an intermediate order statistic of $\{X_i\}$. The level $\widetilde{u}_n$ is determined by conditions A.1.-A.5. An adequate choice of this level is $M_{c_n+1:k_n}$ with $c_n \to \infty$ and $c_n = o(k_n)$.

In practice the exact choice of the base levels u_n and $\widetilde{u}_n$ is not important as long as the levels v_n and $\widetilde{v}_n$ are chosen properly to satisfy (12) and (16). The difference between

these estimators of θ lies on the limiting distribution of their components. In a nutshell, Z_{u_n} converges to a Poisson distribution while $Z_{\widetilde{u}_n}$ satisfies the central limit theorem. This becomes important for the inference about θ.

3.1 Statistical Inference

Standard methods for estimating the extremal index rely on the choice of an extreme level u_n. For appropriate partitions of the stationary sequence this level determines the block cluster size. By definition of extreme level the number of exceedances entering into a cluster is roughly constant although n increases. Furthermore by the properties of the Poisson distribution the variance of the cluster size converges to a constant different from zero. Therefore estimators based on these levels are not successful at providing more accurate estimates of the extremal index as n increases. This together with the presence of dependence in $\{X_i\}$ make difficult to find the distribution of the estimators commonly used for θ.

On the other hand the extension of these estimators to moderately high levels is not straightforward. Hsing (1988) shows that the distribution of clusters of exceedances defined by extreme levels converges to a geometric distribution. If the level is lowered to achieve consistency the number of exceedances entering into the cluster increases with n and no longer converges to a distribution function. In order to solve this problem Hsing (1991) introduces a lower level defined by a sequence, say y_n, that converges to infinity. Increasing cluster sizes determined by the lower level are standardized by y_n in order to obtain a random variable. Hsing in that paper proposes a variant of the blocks method estimator for estimating θ that is asymptotically normal.

The characterization of θ in this paper as a limiting ratio determined by two levels makes possible statistical inference about the parameter. Under $D(u_n)$ or alternatively $D(\widetilde{u}_n)$ and for n sufficiently high, numerator and denominator of $\widetilde{\theta}_n$ and $\widetilde{\widetilde{\theta}}_n$ are well approximated by a binomial distribution. These estimators only differ in their limiting behavior.

We will study first the distribution of $\widetilde{\theta}_n$ assuming u_n and v_n are the levels of interest. In order to do that we assume the conditional distribution of $Z^*_{v_n}$ given $Z^*_{u_n} = z^*_{u_n}$ is binomial with parameter $p_n = \frac{1-F_{1,\ldots,r_n}(v_n)}{1-F_{1,\ldots,r_n}(u_n)}$. The probability of $Z^*_{v_n}$ can be expressed as

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \sum_{z^*_{u_n}=k}^{k_n} P\{Z^*_{v_n} = k \mid Z^*_{u_n} = z^*_{u_n}\} \; P\{Z^*_{u_n} = z^*_{u_n}\}.$$

Then

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \sum_{z^*_{u_n}=k}^{k_n} \binom{z^*_{u_n}}{k} p_n^k (1-p_n)^{z^*_{u_n}-k} \binom{k_n}{z^*_{u_n}} (1 - F_{1,\ldots,r_n}(u_n))^{z^*_{u_n}} F_{1,\ldots,r_n}^{k_n - z^*_{u_n}}(u_n). \tag{32}$$

This distribution can be written as

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} =$$
$$\sum_{k=0}^{z^*_{v_n}} \binom{k_n}{k} [p_n(1 - F_{1,\ldots,r_n}(u_n))]^k \sum_{z^*_{u_n}=k}^{k_n} \binom{k_n - k}{z^*_{u_n} - k} [(1-p_n)(1 - F_{1,\ldots,r_n}(u_n))]^{z^*_{u_n}-k} F_{1,\ldots,r_n}^{k_n - z^*_{u_n}}(u_n).$$

By Newton's formula, $(x+y)^t = \sum_{k=0}^{t} \binom{t}{k} x^k y^{t-k}$,

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \binom{k_n}{k} [p_n(1 - F_{1,\ldots,r_n}(u_n))]^k [1 - p_n(1 - F_{1,\ldots,r_n}(u_n))]^{k_n - k}.$$

Replacing p_n by its value yields

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \binom{k_n}{k} [(1 - F_{1,\ldots,r_n}(v_n))]^k [F_{1,\ldots,r_n}(v_n)]^{k_n - k}.$$

This result implies that $Z^*_{v_n} \mid Z^*_{u_n} = z^*_{u_n}$ is a binomial distribution of the form $Bin(p_n, z^*_{u_n})$.

Using the same methodology for the asymptotic distributions of $Z^*_{u_n}$ and $Z^*_{v_n}$ we obtain the asymptotic distribution of the conditional distribution of $Z^*_{v_n}$. The procedure is as follows

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \sum_{z^*_{u_n}=k}^{\infty} \binom{z^*_{u_n}}{k} \theta^k (1-\theta)^{z^*_{u_n}-k} \exp^{-(\theta\tau)} \frac{(\theta\tau)^{z^*_{u_n}}}{z^*_{u_n}\,!}.$$

Under some algebra this probability becomes

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \exp^{-(\theta\tau)} \frac{(\theta^2\tau)^k}{k!} \sum_{z^*_{u_n}=k}^{\infty} \frac{[\theta\tau(1-\theta)]^{z^*_{u_n}-k}}{(z^*_{u_n}-k)!}.$$

Therefore

$$P\{Z^*_{v_n} \leq z^*_{v_n}\} = \sum_{k=0}^{z^*_{v_n}} \exp^{-(\theta^2\tau)} \frac{(\theta^2\tau)^k}{k!}. \tag{33}$$

Provided that $Z^*_{u_n}$ and $Z^*_{v_n}$ follow a Poisson distribution asymptotically we find that $Z^*_{v_n} \mid Z^*_{u_n} = z^*_{u_n}$ has a binomial limiting distribution of the form $Bin(\theta, z^*_{u_n})$.

In order to derive the unconditional first moments of $\widetilde{\theta}_n$ we will calculate the conditional expected value and variance. This is immediate from the conditional distribution of $Z^*_{v_n}$ given

$Z^*_{u_n}$. Then

$$E[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}] = p_n, \tag{34}$$

and the conditional variance takes this form

$$V[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}] = \left(\frac{1 - F_{1,\ldots,r_n}(v_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right)\left(1 - \frac{1 - F_{1,\ldots,r_n}(v_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right)\frac{1}{z^*_{u_n}}. \tag{35}$$

By the law of iterated expectations,

$$E[\widetilde{\theta}_n] = E\left[E[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}]\right] = p_n \quad with \quad p_n \to \theta \quad as \quad n \to \infty. \tag{36}$$

The unconditional variance can be decomposed into two different terms,

$$V[\widetilde{\theta}_n] = V\left[E[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}]\right] + E\left[V[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}]\right].$$

By the Taylor expansion of $E[1/Z^*_{u_n}]$ about $E[Z^*_{u_n}]$ we obtain that

$$E[V[\widetilde{\theta}_n|Z^*_{u_n} = z^*_{u_n}]] = p_n(1 - p_n)\left(\frac{1}{E[Z^*_{u_n}]} + \frac{V[Z^*_{u_n}]}{E^3[Z^*_{u_n}]}\right). \tag{37}$$

The unconditional variance reads as

$$V[\widetilde{\theta}_n] = p_n(1 - p_n)\left(\frac{1}{\theta\tau} + O(1)\right) = \frac{1 - \theta}{\tau} + O(1). \tag{38}$$

The variance converges to a constant different from zero for τ constant. Although $\widetilde{\theta}_n$ is asymptotically unbiased the estimator is not consistent for the uncertainty does not diminish as the sample size increases.

The choice of $\widetilde{\widetilde{\theta}}_n$ consisting on a ratio of exceedances of moderately high levels is motivated by the lack of consistency of $\widetilde{\theta}_n$. The factors defining this estimator are $Z^*_{\widetilde{u}_n}$ and $Z^*_{\widetilde{v}_n}$. The distribution of $Z^*_{\widetilde{v}_n} \mid Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}$ is binomial of parameters $Bin(\widetilde{p}_n, z^*_{u_n})$ with $\widetilde{p}_n = \frac{1-F_{1,\ldots,r_n}(\widetilde{v}_n)}{1-F_{1,\ldots,r_n}(\widetilde{u}_n)}$. The proof is identical to the extreme levels case.

Proceeding as before we have

$$E[\widetilde{\widetilde{\theta}}_n] = \widetilde{p}_n \quad with \quad \widetilde{p}_n \to \theta \quad as \quad n \to \infty, \tag{39}$$

see (22). Operating as in (37) the unconditional variance reads as

$$V[\widetilde{\widetilde{\theta}}_n] = \widetilde{p}_n(1 - \widetilde{p}_n)\left(\frac{1}{E[Z^*_{\widetilde{u}_n}]} + \frac{V[Z^*_{\widetilde{u}_n}]}{E^3[Z^*_{\widetilde{u}_n}]}\right).$$

By definition of the level $\widetilde{u}_n$ we have

$$V[\widetilde{\widetilde{\theta}}_n] \to 0 \quad as \quad n \to \infty. \tag{40}$$

Conditions (39) and (40) are sufficient to assert the consistency of $\widetilde{\widetilde{\theta}}_n$. Mathematically,

$$\widetilde{\widetilde{\theta}}_n \xrightarrow{p} \theta \quad as \quad n \to \infty. \tag{41}$$

The proof of this result is immediate by applying Chebyshev inequality.

In practice to avoid uncertainty in $Z^*_{\widetilde{u}_n}$ the level $\widetilde{u}_n$ is assumed to be an intermediate order statistic $\widetilde{u}_n = M_{c_n+1:k_n}$ with $c_n \to \infty$ and $c_n = o(k_n)$. The binomial distribution of $Z^*_{\widetilde{v}_n} \mid Z^*_{\widetilde{u}_n} = c_n$ is well approximated by a normal distribution $N(\widetilde{p}_n c_n, \widetilde{p}_n(1-\widetilde{p}_n)c_n)$.

Hence for n sufficiently high

$$\widetilde{\widetilde{\theta}}_n \overset{w}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{c_n}\right) \tag{42}$$

with $(\sim)$ denoting approximation. In this case standard inference is straightforward. The asymptotic confidence intervals for θ are

$$\theta \in \left[\widetilde{\widetilde{\theta}}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widetilde{\widetilde{\theta}}_n(1-\widetilde{\widetilde{\theta}}_n)}{c_n}} \quad \right] \tag{43}$$

with $z_{1-\alpha/2}$ the quantile of $\Phi(\cdot)$.

Testing the existence of clustering in the largest observations becomes an attainable objective given that it is possible testing the true value of the extremal index. If $D'(\widetilde{u}_n)$ is violated there exists clustering of observations in the tails; otherwise $\theta = 1$. We can devise one-sided confidence intervals to test the clustering of exceedances of $\widetilde{u}_n$ in stationary sequences satisfying $D(\widetilde{u}_n)$. The null hypothesis is $\theta_0 = 1$ against $\theta_0 < 1$. Testing this equals to check if the value 1 is contained in the interval

$$\left(-\infty, \widetilde{\widetilde{\theta}}_n + z_{1-\alpha}\sqrt{\frac{\widetilde{\widetilde{\theta}}_n(1-\widetilde{\widetilde{\theta}}_n)}{c_n}} \quad \right].$$

3.2 A comparison between different estimators

In this section we calculate the order of bias and variance for different estimators of the extremal index. In particular for the logs method, blocks method and runs method. These estimators are defined based on a single extreme level u_n. We use the results found in Smith and Weissman (1994). These authors found that the logs method is asymptotically unbiased.

In particular

$$E[\hat{\theta}_n^{(1)}] = \theta + O\left(\frac{\tau}{k_n}\right).$$

They also show that the variance of this estimator is of order $O(\frac{1}{\tau})$. For the blocks method these authors reinforce the results derived in Hsing (1991). They find that

$$E[\hat{\theta}_n^{(2)}] = \theta + O(\frac{1}{\tau}),$$

and

$$V[\hat{\theta}_n^{(2)}] = O(\frac{1}{\tau}).$$

For u_n an extreme level τ is constant and $O(\frac{1}{\tau})$ amounts to $O(1)$ (see (12)). Therefore in terms of mean square error both estimators, $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$, are identical. It is worth observing however than the logs method is asymptotically unbiased provided that k_n increases with n in contrast to $\hat{\theta}_n^{(2)}$. This can be observed in the simulation experiments that are presented in the next section.

Results for the runs method are similar. By the law of iterated expectations $E[\overline{\theta}_n] = E[E[\frac{W_{u_n}}{Z_{u_n}} \mid Z_{u_n}]]$. The expected value of the numerator takes this expression

$$E[W_{u_n}] = (n - r_n)P\{X_{i+1} \leq u_n, \ldots, X_{i+r_n} \leq u_n \mid X_i > u_n\}P\{X_i > u_n\}.$$

These authors define $\theta(r_n + 1, u_n) = P\{X_{i+1} \leq u_n, \ldots, X_{i+r_n} \leq u_n \mid X_i > u_n\}$. Then

$$E[\overline{\theta}_n] = (n - r_n)(1 - F(u_n))\theta(r_n + 1, u_n)E\left[\frac{1}{Z_{u_n}}\right].$$

That is

$$E[\overline{\theta}_n] = \theta(r_n + 1, u_n) - \frac{\theta(r_n + 1, u_n)}{k_n} + O\left(\frac{1}{\tau}\right)$$

and the bias of the runs estimator is of order $O(1)$ for $\theta(r_n + 1, u_n) \to \theta$ as $n \to \infty$.

For the unconditional variance it is sufficient to analyze $E\left[V[\overline{\theta}_n \mid Z_{u_n}]\right]$. In order to that we derive the conditional variance. This is

$$V[\overline{\theta}_n \mid Z_{u_n}] = \frac{(n - r_n)(1 - F(u_n))\,\theta(r_n + 1, u_n)\,[1 - \theta(r_n + 1, u_n)(1 - F(u_n))]}{Z_{u_n}^2}.$$

Then

$$V[\overline{\theta}_n] = \frac{(n - r_n)(1 - F(u_n))\,\theta(r_n + 1, u_n)\,[1 - \theta(r_n + 1, u_n)(1 - F(u_n))]}{E[Z_{u_n}]^2} + O(1).$$

To summarize this section we can say that in terms of mean square error any estimator of the extremal index based on extreme levels provides the same kind of disappointing results. Neither of them is consistent for the variance converges to a constant different from zero. Nonetheless the estimator of this type introduced herein, $\widetilde{\theta}_n$ defined by two extreme levels, outperforms the rest of estimators in the sense that it is possible to obtain its finite-sample distribution as well as its limiting distribution. Hence statistical inference is plausible. The extension of this estimator to moderately high levels is successful at overcoming both problems. $\widetilde{\widetilde{\theta}}_n$ is consistent, bias and variance converge to zero, and statistical inference is straightforward.

4 Simulations: Some examples

This section studies some examples of stationary sequences exhibiting short range dependence in the extremes. Consider the example due to Chernick (1981) for $\{X_i\}$ a strictly stationary first order autoregressive sequence driven by

$$X_i = \frac{1}{r} X_{i-1} + \varepsilon_i, \tag{44}$$

with $r \geq 2$, an integer, ε_i discrete uniforms on $\{0, 1/r, \ldots, (r-1)/r\}$, and ε_i independent of X_{i-1}. The random variable X_i has a uniform distribution on $[0, 1]$. In this example the extremal index is $\theta = \frac{r-1}{r}$.

Figure A.1 displays estimates of θ by different techniques for several extreme levels determined by $u_n = x_{c+1:n}$ with $c = 5, 15, 25, 35$ and for $n = 200$. By construction, the blocks and the runs method provide underestimates of θ as r_n increases for the number of elements in the numerator of these estimators decreases as the block size increases. Estimates given by the logs method are very accurate for extreme levels. For lower levels however, $\hat{\theta}_n^{(1)}$ exhibits problems derived from the fact that every single block has an exceedance ($Z_{u_n}^* = k_n$). In this case the estimator is not defined. In contrast $\widetilde{\theta}_n$ shows reliable estimates of θ across all the levels. The same results are observed for moderately high levels defined by $\widetilde{u}_n = x_{c_n+1:n}$ with $c_n = n^{2/3}$. In this case the logs method is as reliable as $\widetilde{\widetilde{\theta}}_n$. The plot of this case is not presented but can be obtained from the author upon request.

Figure A.2 shows a sample of coverage probabilities corresponding to the asymptotic gaussian distribution. The plot includes both types of levels. It is shown in the core of the paper that $\widetilde{\widetilde{\theta}}_n$ was devised to converge to a normal distribution with $n \to \infty$. $\widetilde{\theta}_n$ however followed a binomial distribution even for large sample sizes. Surprisingly, the left plot of the figure shows that for a sample of n=1000 observations the gaussian approximation of the binomial distribution works for u_n an extreme level. This result vanishes as n increases. For moderately

high levels the right plot of figure A.2 describes a curious phenomenon. The asymptotic theory developed only works for certain partitions of the sequence. For unnecessary large blocks the asymptotic normal approximation does not work due to misleading estimates of the variance of $\widetilde{\widetilde{\theta}}_n$. In a sense this method provides a technique to find out appropriate blocks sizes.

The failure of the coverage probability based on gaussian confidence intervals to approximate the actual $\alpha = 0.05$ for large block sizes is also analyzed in the following example (Chernick model with $r = 2$). In this case the level of clustering is $\theta = 0.5$. Figure A.3 shows similar results to figure A.1 about the estimates of θ. The empirical coverage probability however shows interesting results (figure A.4). The estimates of the actual coverage $1 - \alpha$ produced by extreme levels are far from the actual value (left plot of A.4). In contrast coverage probabilities corresponding to $\widetilde{\widetilde{\theta}}_n$ yield very nice convergence results for both n and r_n increasing. This suggests that inference about θ for processes with high clustering in the extremes requires larger blocks sizes to eliminate serial dependence. This phenomenon is also studied in the following example.

This is the doubly stochastic model. Let $\{\xi_i, i \geq 1\}$ be *iid* with distribution function F, and suppose that $Y_1 = \xi_1$, and for $i > 1$,

$$Y_i = \begin{cases} Y_{i-1} & \text{with probability} \quad \psi, \\ \xi_i & \text{with probability} \quad 1 - \psi, \end{cases}$$

the choice being made independently for each i. The doubly stochastic sequence $\{X_i\}$ is defined by

$$X_i = \begin{cases} Y_i & \text{with probability} \quad \eta, \\ 0 & \text{with probability} \quad 1 - \eta, \end{cases}$$

independently of anything else. In this example the extremal index is $\theta = \frac{1-\psi}{1-\psi+\psi\eta}$. Smith and Weissman (1994) compare different estimators of θ for this example. For $\Psi = 0.9$ and $\eta = 0.7$ ($\theta = 0.137$) these authors find the runs method superior to the rest of competing estimators. Figure A.5 is consistent with their results. $\widetilde{\theta}_n$ seems to be however a very good competitor of $\overline{\theta}_n$ for every single level. This result is also observed for moderately high levels though is not reported for sake of space. Furthermore $\widetilde{\theta}_n$ outperforms the logs and the blocks estimators across all levels. The empirical coverage probability (figure A.6) exhibits a poor approximation of the normal distribution for any sample size. On the other hand for moderately high levels the empirical coverage seems to converge to the theoretical value $1 - \alpha$ for large blocks sizes ($r_n > 20$). This may reflect the large amount of clustering in this doubly stochastic process.

Finally, to assess the performance of the runs method versus $\widetilde{\theta}_n$ and $\widetilde{\widetilde{\theta}}_n$ we also estimate the extremal index of this process for $\Psi = 0.5$ and $\eta = 0.5$ ($\theta = 0.667$.) The runs and blocks method exhibit the same declining pattern observed before for increasing blocks sizes

(figure A.7). $\widetilde{\theta}_n$ and $\widetilde{\widetilde{\theta}}_n$ are superior to the rest of estimators. For moderately high levels the results are alike. Within the competitors only the logs method exhibits a similar performance. The empirical coverage probabilities for extreme levels and moderately high levels show the same patterns than for the Chernick model with $r = 5$. Both processes exhibit little clustering in the extremes. This entails choices of the block size commensurate with the extent of dependence within the blocks. Large values of r_n would imply spurious clustering of the largest observations within the blocks.

5 Conclusion

Measuring serial dependence in the extremes of stationary sequences boils down to assess the extent of clustering in these observations. This phenomenon is observed in a number of fields studying time series and concerned about the occurrence of extreme events. Serve as illustration fields as risk management, hydrology or climatology.

The extent of this extremal dependence is summarized in one single parameter, the extremal index. Standard statistical techniques involving the estimation of θ present some serious shortcomings derived from the lack of consistency and the use of a different type of technology (statistics of extremes).

In fact, it is even difficult to disentangle the distribution function of most of these estimators for θ. To overcome this, we have introduced a family of estimators of this parameter. The first estimator, $\widetilde{\theta}_n$ is a ratio of two binomial random variables defined by extreme levels. This estimator is asymptotically unbiased and follows a binomial distribution that converges to a Poisson distribution. In turn it is not consistent by construction and shares the type of problems of usual estimators as logs method, blocks method and the runs method. The natural extension of $\widetilde{\theta}_n$ to lower levels (moderately high levels) yields a very appealing estimator $\widetilde{\widetilde{\theta}}_n$. This estimator is consistent and follows a binomial distribution. It differs from the other in what its asymptotic distribution is normal and enables the use of standard statistical inference.

From the asymptotic theory and the simulation experiments we have developed we can extract some interesting results about how to proceed to derive pointwise estimates and confidence intervals for θ. For small sample sizes if a stationary sequence exhibits low clustering in the extremes the distribution of both estimators can be well approximated by a normal distribution. If the level of clustering is high we should explore alternative confidence intervals derived from binomial distributions.

For large sample sizes $\widetilde{\widetilde{\theta}}_n$ is a safer choice. For appropriate partitions of the sequence its asymptotic distribution is normal. However for sequences with low clustering of extremes

blocks sizes excessively large can yield misleading estimates of the variance and in turn wrong confidence intervals. This is due to the occurrence of spurious clustering within the blocks. On the other hand the presence of high clustering in the extremes requires the use of larger blocks sizes to eliminate such dependence.

These results suggest two strategies when estimating the extremal index. We can proceed with a preliminary inspection of the data to determine roughly the amount of clustering in the extremes. For small sample sizes and little clustering use $\widetilde{\theta}_n$ and the normal approximation. If the amount of clustering is high consider $\widetilde{\widetilde{\theta}}_n$ estimated for large blocks. For large sample sizes and low clustering use $\widetilde{\widetilde{\theta}}_n$ determined by moderate partitions, and for high clustering use $\widetilde{\widetilde{\theta}}_n$ determined by large blocks sizes.

Acknowledgements. The author is deeply grateful to the Department of Statistics and Operations Research of UNC at Chapel Hill, especially to Ross Leadbetter for his helpful comments. The author also acknowledges suggestions of Peter Burridge and comments received during the 3^{rd} Symposium on Extreme Value Analysis: Theory and Practice celebrated in Aveiro (Portugal) as well as its scientific committee for the C.E.A.U.L. prize for young researchers.

References

[1] Chernick, M.R., (1981): A limit theorem for the maximum of autorregresive processes with uniform marginal distribution, Annals of Probability 9, 145 – 149.

[2] Feller, W., (1966): *An introduction to Probability Theory and its Applications.* Ed. John Wiley & Sons, New York. (Volume 2).

[3] Ferro, C.A., and Segers, J., (2003): Inference for clusters of extremes. Journal of the Royal Statistical Society *B*, 65, 545 – 556.

[4] Haan, L. de, (1976): Sample extremes: an elementary introduction. Statist. Neerlandica, 30, 161 – 172.

[5] Hodges, J.L., and Le Cam, L., (1960): The Poisson approximation to the Poisson binomial distribution. Annals of Mathematical Statistics, 31, 737-740.

[6] Hsing, T., (1988): On the weak convergence of extreme order statistics, Stochastic Processes and Applications, 29, 155 – 169.

[7] Hsing, T., (1991): Estimating the parameters of rare events. Stochastic Processes and Applications, 37, 117 – 139.

[8] Hsing, T., (1993): Extremal Index Estimation for a Weakly Dependent Stationary Sequence. Annals of Statistics, 21, 2043 – 2071.

[9] Leadbetter, M. R., (1983): Extremes and Local Dependence in Stationary Sequences, Z. Wahrscheinlichkeitstheorie verw. Gebiete 65, 291 – 306.

[10] Leadbetter, M. R., Lindgren, G., and Rootzén, H., (1983): *Extremes and Related Properties of Random Sequences and Processes.* Ed. Springer-Verlag, New York.

[11] Lehman, E.L., (1999): *Elements of Large-Sample Theory.* Ed. Springer-Verlag, New York.

[12] Loynes, R.M., (1965): Extreme Values in Uniformly Mixing Stationary Stochastic Processes. Annals of Mathematical Statistics, 36, 993 – 999.

[13] O'Brien, G.L., (1974):The maximum term of uniformly mixing stationary sequences. Z. Wahrscheinlichkeitstheorie verw. Gebiete 30, 57 – 63.

[14] O'Brien, G.L., (1987): Extreme Values for Stationary and Markov Sequences. Annals of Probability, 15, 281 – 291.

[15] Rosenblatt, M., (1956): A central limit theorem and a strong mixing condition. Proc. Nat. Acad. Sci. USA, 42, 43 – 47.

[16] Smith, R.L., Weissman, I., (1994): Estimating the Extremal index. Journal of the Royal Statistical Society *B*, 56, 515 – 528.

Appendix: List of figures

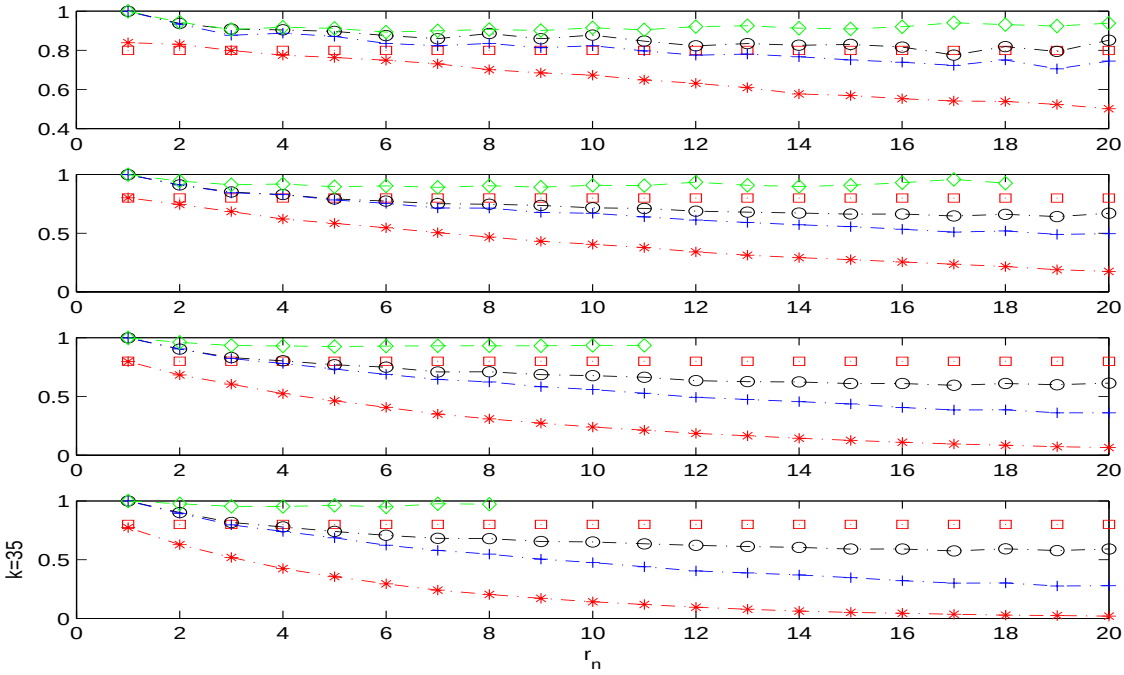

Figure A.1. *Sample mean for different estimators of* θ *(*$m = 100$ *Monte-Carlo simulations) for different levels* u_n *defined by* $x_{c+1:n}$*, c=5,15,25,35, and n=200. The process is the Chernick model with* $r = 5$ *and* $\theta = 0.8$*.* r_n *moves along the interval* $[1, 20]$*. The extremal index is plotted with* $\square$*.* $\widetilde{\theta}_n$ *is represented by* $-o$*.* $\hat{\theta}_n^{(1)}$ *by* $-\diamond$*.* $\hat{\theta}_n^{(2)}$ *by* $-+$*, and* $\overline{\theta}_n$ *by* $-*$ *line.*

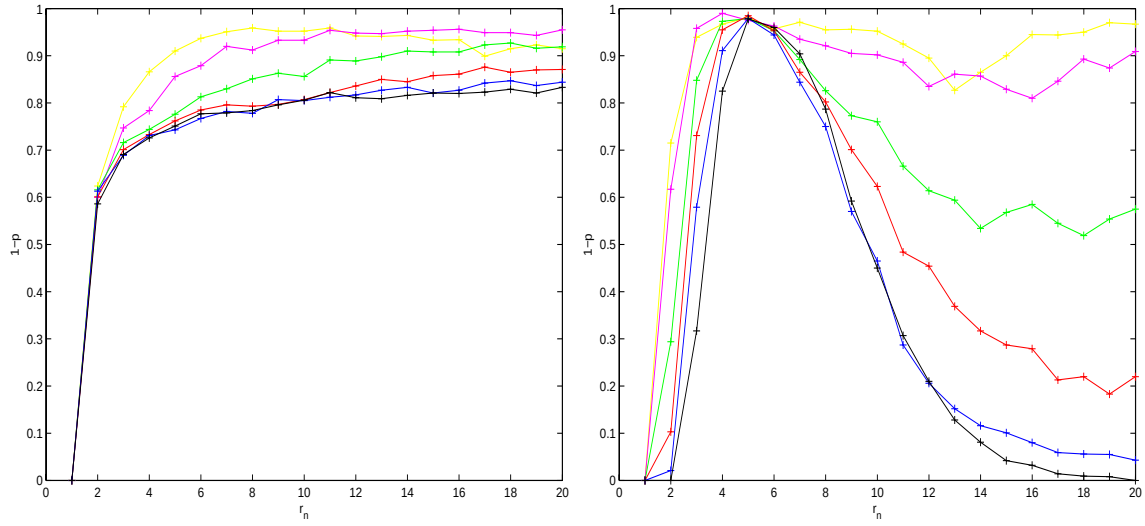

Figure A.2. *Coverage probabilities (c.p.) derived from* $\widetilde{\theta}_n \sim N(\theta, (\theta(1-\theta))/Z^*_{u_n})$ *for the Chernick model,* $r = 5$*.* $m = 1000$ *simulations. The left plot displays* $u_n = x_{c+1:n}$*, c=10 (extreme levels). The right plot,* $\widetilde{u}_n = x_{c_n+1:n}$*,* $c_n = n^{2/3}$ *(moderately high levels).* $n = 100, 200, 500, 1000, 2000, 5000$*. In both cases c.p. is decaying for higher n. For* u_n*, c.p. increases in* r_n*. For* $\widetilde{u}_n$*, c.p. converges to its actual value* $1 - \alpha = 0.95$ *for* $r_n \in [3, 9]$*.*

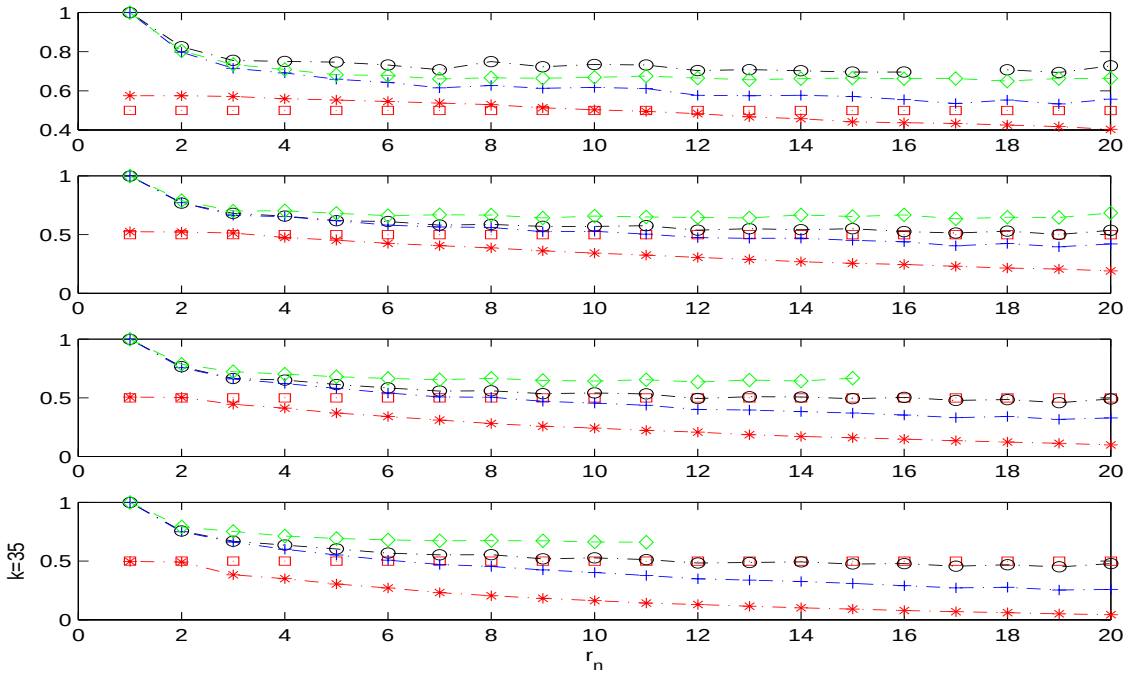

Figure A.3. *Sample mean for different estimators of* θ *(*$m = 100$ *Monte-Carlo simulations) for different levels* u_n *defined by* $x_{c+1:n}$*, c=5,15,25,35, and n=200. The process is the Chernick model with* $r = 2$ *and* $\theta = 0.5$*.* r_n *moves along the interval* $[1, 20]$*. The extremal index is plotted with* $\square$*.* $\widetilde{\theta}_n$ *is represented by* $(-o)$*.* $\hat{\theta}_n^{(1)}$ *by* $(-\diamond)$*.* $\hat{\theta}_n^{(2)}$ *by* $(-+)$*, and* $\overline{\theta}_n$ *by* $(-*)$ *line.*

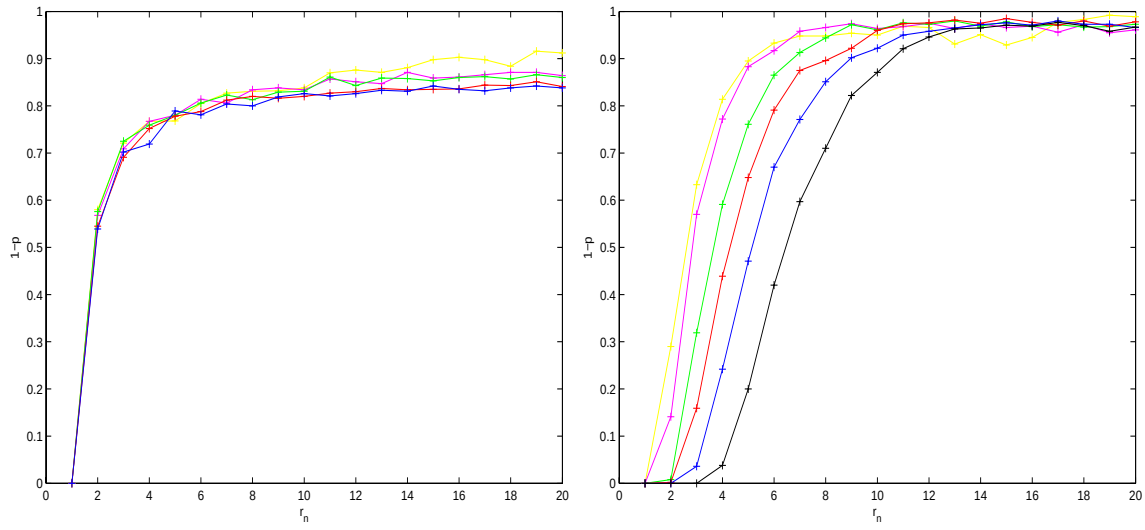

Figure A.4. *Coverage probabilities (c.p.) derived from* $\widetilde{\theta}_n \sim N(\theta, (\theta(1-\theta))/Z^*_{u_n})$ *for the Chernick model,* $r = 2$*.* $m = 1000$ *simulations. The left plot displays the case* $u_n = x_{c+1:n}$*, c=10 (extreme levels). The right plot,* $\widetilde{u}_n = x_{c_n+1:n}$*,* $c_n = n^{2/3}$ *(moderately high levels).* $n = 100, 200, 500, 1000, 2000, 5000$*. In both cases c.p. is decaying for higher n. This however increases with* r_n*. For* $\widetilde{u}_n$ *c.p. converges to its actual value* $1 - \alpha = 0.95$*.*

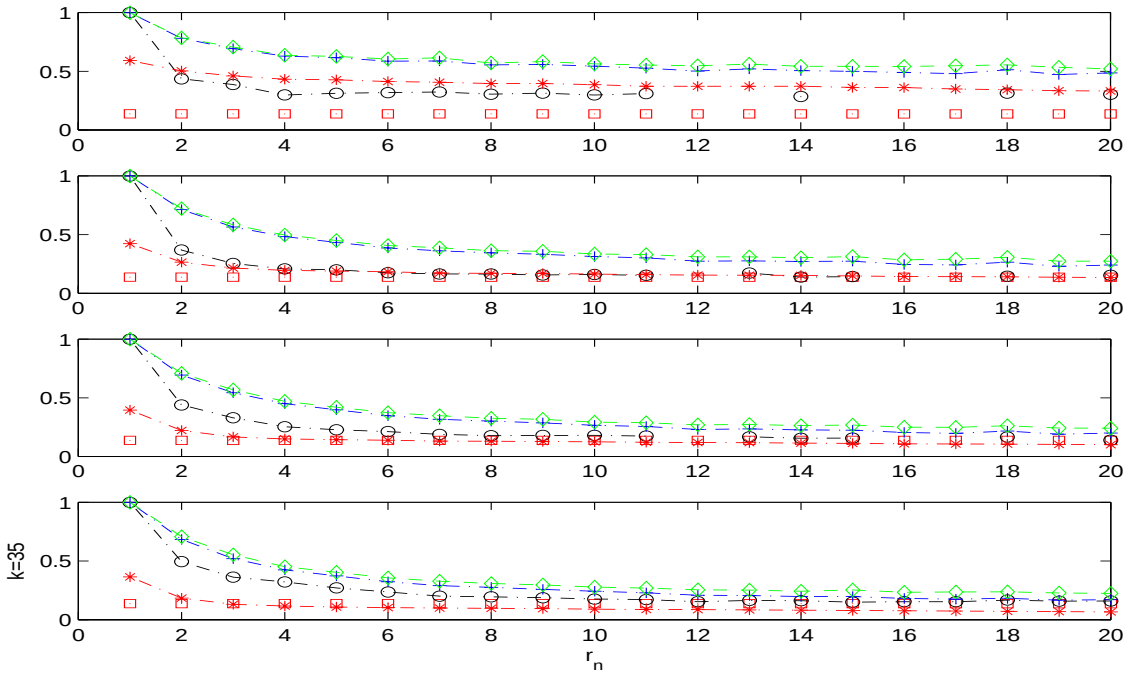

Figure A.5. *Sample mean for different estimators of* θ *(*$m = 100$ *Monte-Carlo simulations) for different levels* u_n *defined by* $x_{c+1:n}$*, c=5,15,25,35, and n=200. The process is the doubly stochastic model with* $\Psi = 0.9$ *and* $\eta = 0.7$*, and* $\theta = 0.137$*.* r_n *moves along the interval* $[1, 20]$*. The extremal index is plotted with* $\square$*.* $\widetilde{\theta}_n$ *is represented by* $(-o)$*.* $\hat{\theta}_n^{(1)}$ *by* $(-\diamond)$*.* $\hat{\theta}_n^{(2)}$ *by* $(-+)$*, and* $\overline{\theta}_n$ *by* $(-*)$ *line.*

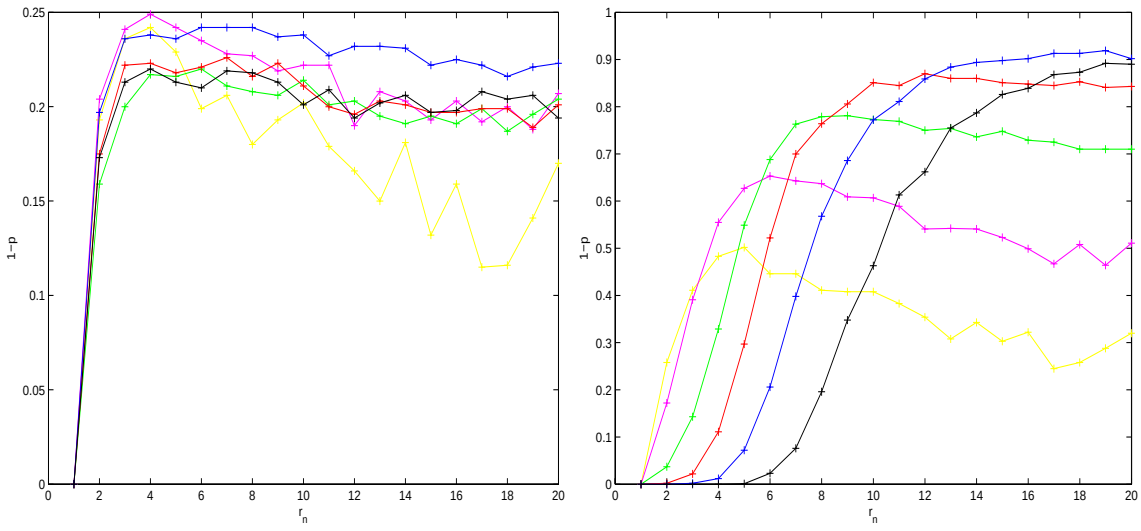

Figure A.6. *Coverage probabilities (c.p.) derived from* $\widetilde{\theta}_n \sim N(\theta, (\theta(1-\theta))/Z^*_{u_n})$ *for the doubly stochastic model with* $\Psi = 0.9$ *and* $\eta = 0.7$*.* $m = 1000$ *simulations. The left plot displays the case* $u_n = x_{c+1:n}$*, c=10 (extreme levels). The right plot,* $\widetilde{u}_n = x_{c_n+1:n}$*,* $c_n = n^{2/3}$ *(moderately high levels).* $n = 100, 200, 500, 1000, 2000, 5000$*. In both cases c.p. is decaying for higher n. For* u_n *c.p. decreases with* r_n*. For* $\widetilde{u}_n$ *c.p. converges very slowly to its actual value* $1 - \alpha = 0.95$ *as* r_n *increases.*

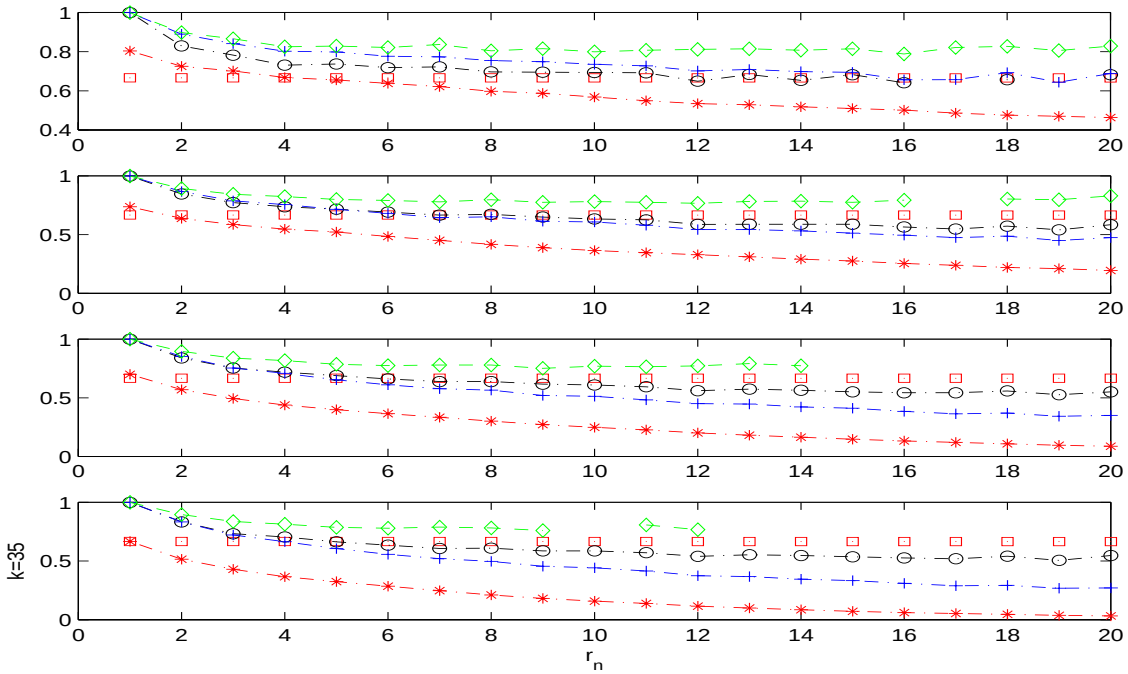

Figure A.7. *Sample mean for different estimators of* θ *(*$m = 100$ *Monte-Carlo simulations) for different levels* u_n *defined by* $x_{c+1:n}$*, c=5,15,25,35, and n=200. The process is the doubly stochastic model with* $\Psi = 0.5$ *and* $\eta = 0.5$*, and* $\theta = 0.667$*.* r_n *moves along the interval* $[1, 20]$*. The extremal index is plotted with* $\square$*.* $\widetilde{\theta}_n$ *is represented by* $(-o)$*.* $\hat{\theta}_n^{(1)}$ *by* $(-\diamond)$*.* $\hat{\theta}_n^{(2)}$ *by* $(-+)$*, and* $\overline{\theta}_n$ *by* $(-*)$ *line.*

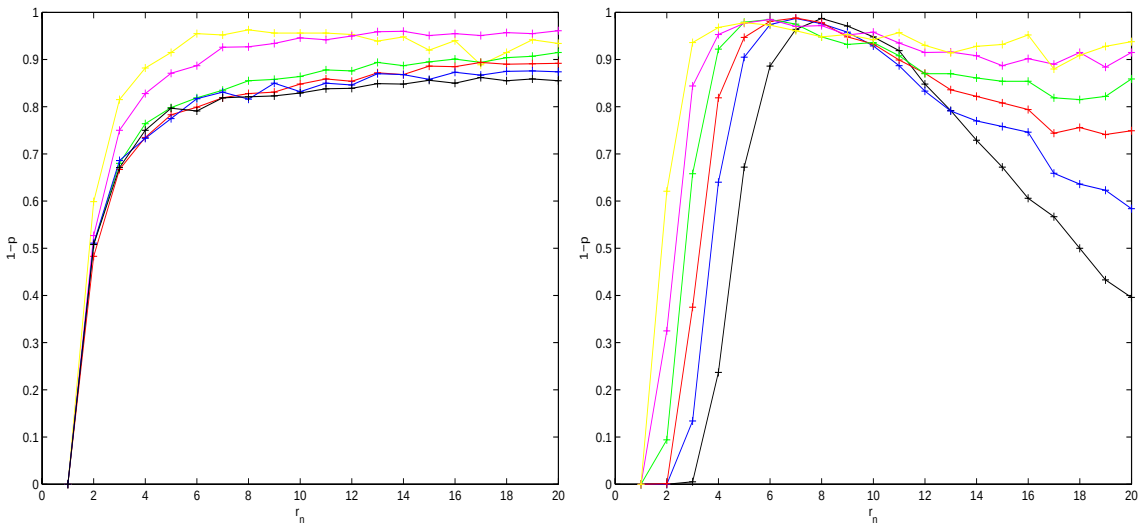

Figure A.8. *Coverage probabilities (c.p.) derived from* $\widetilde{\theta}_n \sim N(\theta, (\theta(1-\theta))/Z^*_{u_n})$ *for the doubly stochastic model with* $\Psi = 0.5$ *and* $\eta = 0.5$*.* $m = 1000$ *simulations. The left plot displays the case* $u_n = x_{c+1:n}$*, c=10 (extreme levels). The right plot,* $\widetilde{u}_n = x_{c_n+1:n}$*,* $c_n = n^{2/3}$ *(moderately high levels).* $n = 100, 200, 500, 1000, 2000, 5000$*. In both cases c.p. is decaying for higher n. For* u_n*, c.p. increases in* r_n*. For* $\widetilde{u}_n$*, c.p. converges to its actual value* $1 - \alpha = 0.95$ *for* $r_n \in [6, 12]$*. For higher values* $1 - p$ *decreases.*