CentER

Discussion Paper

No. 2003–91

APPROXIMATE DISTRIBUTIONS OF CLUSTERS OF EXTREMES

By J. Segers

October 2003

ISSN 0924-7815

Tilburg University

APPROXIMATE DISTRIBUTIONS OF CLUSTERS OF EXTREMES

JOHAN SEGERS*
Tilburg University & Catholic University Leuven

October 8, 2003

Abstract. In a stationary sequence of random variables, high-threshold exceedances may cluster together. Two approximations of such a cluster's distribution are established. These justify and generalize sampling schemes for clusters of extremes already known for Markov chains.

Key words: cluster of extremes; extremal index; stationary sequence; threshold exceedance; maximum

JEL: C13, C14

AMS 2000: 60G70, 62G32

1 Introduction

Let $\{X_n\}_{n\geq 1}$ be a stationary sequence of random variables. Let $\{u_n\}$ be a real sequence such that $\Pr[X_1 > u_n] > 0$ but $\Pr[X_1 > u_n] \to 0$ as $n \to \infty$. We are interested in exceedances over the threshold u_n among the variables $X_1, \ldots, X_{r_n}$, where $\{r_n\}$ is a positive integer sequence tending to infinity. We require that the expected number of exceedances tends to zero, that is,

$$\mathrm{E}\left[\sum_{i=1}^{r_n} \mathbf{1}(X_i > u_n)\right] = r_n \Pr[X_1 > u_n] \to 0, \qquad n \to \infty, \tag{1}$$

with $\mathbf{1}(A)$ denoting the indicator function of the event A.

For positive integer m, let $M_m = \max_{i=1,\ldots,m} X_i$. By (1), the probability that M_{r_n} exceeds u_n will tend to zero as well. However, if it should happen that $M_{r_n} > u_n$, then there might be more than one exceedance. Now think

*Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, the Netherlands (e-mail: jsegers@uvt.nl). The author is Post-Doctoral Research Fellow of the Fund for Scientific Research, Flanders.

of all the exceedances in the block as one single *cluster*. For instance, if the X_i represent rainfall measurements on a single site at equidistant time points and if u_n denotes a particularly high value relevant for some hydrological construction like a dike, dam, or drainage system, then all the exceedances, if any, may be thought of as pertaining to the same storm.

The distribution of a cluster depends on the distribution of $(X_1, \ldots, X_{r_n})$ conditionally on $\{M_{r_n} > u_n\}$, which is in general not easy to compute. Rather, we would like to write the cluster distribution in terms of the distribution of $(X_1, \ldots, X_m)$ conditionally on $\{X_1 > u_n\}$ for some large enough m. For instance, if $\{X_n\}$ is a Markov chain, then we can get access to this conditional distribution through simulation from the limiting transition probabilities, which can be written in terms of a random walk, called the forward *tail chain* (Smith, 1992; Perfekt, 1994; Yun, 1998). Alternatively, if the backward tail chain is available as well, then Smith el al. (1997) propose a simulation scheme based on the conditional distribution of $(X_1, \ldots, X_{2m})$ conditionally on X_{m+1} being the *cluster peak*, that is, $\{X_{m+1} = M_{2m} > u_n\}$.

For higher-order Markov chains, Yun (2000) showed how to find the limit distribution of so-called *cluster functionals* in terms of the forward tail chain. An extension to general stationary sequences was described by Segers (2003). Such sequences are the setting of the present paper as well, but the attention is shifted from cluster functionals to the cluster distribution itself. As will turn out, this leads to a more transparant theory and much shorter proofs. Moreover, a distributional approximation is established to justify the approach of Smith et al. (1997) as well.

In Section 2, the notion of cluster of extremes is given a formal definition. The tail chain approximation is established in Section 3, and the cluster peak approximation in Section 4. A brief discussion of possible statistical applications is given in Section 5.

2 Clusters of extremes

We need a formal definition of what we intuitively described as a cluster. For positive integer r, let $\mathbb{A}_r := \mathbb{R}^r \setminus (-\infty, 0]^r$ be the set of all real r-tuples $(x_1, \ldots, x_r)$ with at least one positive entry. Define $\mathbb{A} = \bigcup_{r \geq 1} \mathbb{A}_r$ to be the set of all such tuples, of arbitrary dimension. The *cluster map* C acts on $(x_1, \ldots, x_r) \in \mathbb{A}$ as follows: if $\alpha = \min\{i : i = 1, \ldots, r,\, x_i > 0\}$ and $\omega = \max\{i : i = 1, \ldots, r,\, x_i > 0\}$ denote the indices of the first and the last positive entry respectively, then

$$C(x_1, \ldots, x_r) = (x_\alpha, \ldots, x_\omega).$$

Observe that the dimension of the vector $(x_\alpha, \ldots, x_\omega)$ is $\omega - \alpha + 1$, that its first and last entries are positive, but that not all entries need to be positive.

Now on the event $\{M_{r_n} > u_n\}$ we simply define the cluster of exceedances by

$$C_n := C(X_1 - u_n, \ldots, X_{r_n} - u_n).$$

We are interested in the distribution of C_n. To this end, we will need to consider sub-blocks $\{X_i, \ldots, X_k\}$. Specifically, put $M_{i,k} := \max_{i \le j \le k} X_j$; also, on the event $\{M_{i,k} > u_n\}$, put

$$C_n(i,k) := C(X_i - u_n, \ldots, X_k - u_n).$$

In particular, $M_{1,m} = M_m$ and $C_n(1, r_n) = C_n$.

Related to the distribution of a cluster is the *expected cluster size*,

$$\mathrm{E}\left[\sum_{i=1}^{r_n} \mathbf{1}(X_i > u_n) \,\middle|\, M_{r_n} > u_n\right] = \frac{r_n \Pr[X_1 > u_n]}{\Pr[M_{r_n} > u_n]} =: \frac{1}{\theta_n}. \tag{2}$$

We shall need to require the expected cluster size to be uniformly bounded. This will be accomplished by the following condition:

$$\lim_{m\to\infty} \limsup_{n\to\infty} \sum_{i=m+1}^{r_n} \Pr[X_i > u_n \mid X_1 > u_n] = 0. \tag{3}$$

Condition (3) was originally proposed by Smith (1992). It is typically fulfilled in the context of (higher-order) Markov chains.

We show that (3) indeed forces $\liminf \theta_n > 0$. By stationarity,

$$\left.\begin{array}{l} \Pr[M_{m+1,r_n} > u_n \mid X_1 > u_n] \\ \Pr[M_{r_n - m} > u_n \mid X_{r_n} > u_n] \end{array}\right\} \le \sum_{i=m+1}^{r_n} \Pr[X_i > u_n \mid X_1 > u_n]. \tag{4}$$

Now since

$$\begin{aligned} \Pr[M_{r_n} > u_n] &\ge \sum_{k=0}^{\lfloor r_n/m \rfloor - 1} \Pr[X_{km+1} > u_n,\, M_{(k+1)m+1,r_n} \le u_n] \\ &\ge \left\lfloor \frac{r_n}{m} \right\rfloor \{\Pr[X_1 > u_n] - \Pr[X_1 > u_n, M_{m+1,r_n} > u_n]\}, \end{aligned}$$

equation (3) and the first inequality of (4) imply $\liminf \theta_n > 0$.

3 Tail chain approximation

We derive an approximation of the distribution of C_n conditionally on $\{M_{r_n} > u_n\}$ in terms of the distribution of $(X_1, \ldots, X_m)$ conditionally on $\{X_1 > u_n\}$. For Markov chains, the limit of this conditional distribution, after an appropriate normalization, is called the tail chain (Smith, 1992; Perfekt, 1994; Yun, 1998). Therefore, we refer to the following approximation as the tail chain approximation.

Theorem 3.1 *Let $\{X_n\}_{n\geq 1}$ be a stationary sequence and let u_n and r_n be as in (1). For measurable $B \subset \mathbb{A}$, define*

$$\begin{aligned} a_{n,m}(B) &= \Pr[C_n(1,m) \in B \mid X_1 > u_n] \\ &\quad - \Pr[C_n(2,m) \in B,\, M_{2,m} > u_n \mid X_1 > u_n] \\ \theta_{n,m} &= \Pr[M_{2,m} \leq u_n \mid X_1 > u_n], \end{aligned}$$

If (3), then

$$\lim_{m\to\infty} \limsup_{n\to\infty} |\theta_n - \theta_{n,m}| = 0 \tag{5}$$

and

$$\lim_{m\to\infty} \limsup_{n\to\infty} \sup_{B} \left|\Pr[C_n \in B \mid M_{r_n} > u_n] - \theta_{n,m}^{-1} a_{n,m}(B)\right| = 0. \tag{6}$$

Proof. Let $B \subset \mathbb{A}$ be measurable. We have

$$\begin{aligned} \Pr[C_n \in B \mid M_{r_n} > u_n] &= \frac{\Pr[C_n \in B,\, M_{r_n} > u_n]}{\Pr[M_{r_n} > u_n]} \\ &= \frac{\Pr[C_n \in B,\, M_{r_n} > u_n]}{\theta_n r_n \Pr[X_1 > u_n]}. \end{aligned} \tag{7}$$

Split the event $\{M_{r_n} > u_n\}$ according to the smallest index $j = 1, \ldots, r_n$ for which $X_j > u_n$:

$$\Pr[C_n \in B,\, M_{r_n} > u_n] = \sum_{j=1}^{r_n} \Pr[C_n(j, r_n) \in B,\, M_{j-1} \leq u_n,\, X_j > u_n], \tag{8}$$

where $M_0 = -\infty$. Fix a positive integer m and let n be large enough such that $r_n \geq 2m + 1$. For integer j such that $1 \leq j \leq r_n - m$, we have $C_n(j, r_n) = C_n(j, j + m - 1)$ unless $M_{j+m,r_n} > u_n$. Hence, for such j,

$$\begin{aligned} &\Big|\Pr[C_n(j, r_n) \in B,\, M_{j-1} \leq u_n,\, X_j > u_n] \\ &- \Pr[C_n(j, j+m-1) \in B,\, M_{j-1} \leq u_n,\, X_j > u_n]\Big| \\ &\leq \Pr[X_j > u_n,\, M_{j+m,r_n} > u_n]. \end{aligned} \tag{9}$$

Now for positive integer j, we have by stationarity

$$\begin{aligned} &\Pr[C_n(j, j+m-1) \in B,\, M_{j-1} \leq u_n,\, X_j > u_n] \\ &\quad = \Pr[C_n(1,m) \in B,\, X_1 > u_n] \\ &\qquad - \Pr[C_n(j, j+m-1) \in B,\, M_{j-1} > u_n,\, X_j > u_n]. \end{aligned} \tag{10}$$

If $j \geq m + 1$, then by stationarity

$$\begin{aligned} &\Big|\Pr[C_n(j, j+m-1) \in B,\, M_{j-1} > u_n,\, X_j > u_n] \\ &- \Pr[C_n(m+1, 2m) \in B,\, M_m > u_n,\, X_{m+1} > u_n]\Big| \\ &\leq \Pr[M_{j-m-1} > u_n,\, X_j > u_n]. \end{aligned} \tag{11}$$

Decompose the event $\{M_m > u_n\}$ according to the largest $l = 1, \ldots, m$ such that $X_l > u_n$. This leads to the following chain of equalities:

$$\begin{aligned}
&\Pr[C_n(m+1, 2m) \in B,\, M_m > u_n,\, X_{m+1} > u_n] \\
&\quad = \sum_{l=1}^{m} \Pr[C_n(l+1, 2m) \in B,\, X_l > u_n,\, M_{l+1,m} \le u_n,\, X_{m+1} > u_n] \\
&\quad = \sum_{k=1}^{m} \Pr[C_n(2, m+k) \in B,\, X_1 > u_n,\, M_{2,k} \le u_n,\, X_{k+1} > u_n], \quad (12)
\end{aligned}$$

the second equality being justified by stationarity and a change of summation index, $k = m - l + 1$. On the other hand,

$$\begin{aligned}
&\Pr[C_n(2, m) \in B,\, X_1 > u_n,\, M_{2,m} > u_n] \\
&\quad = \sum_{k=1}^{m-1} \Pr[C_n(2, m) \in B,\, X_1 > u_n,\, M_{2,k} \le u_n,\, X_{k+1} > u_n]. \quad (13)
\end{aligned}$$

Now $C_n(2, m+k) = C_n(2, m)$ unless $M_{m+1,m+k} > u_n$. Hence, by comparing term by term in (12) and (13), we get

$$\begin{aligned}
&\Big| \Pr[C_n(m+1, 2m) \in B,\, M_m > u_n,\, X_{m+1} > u_n] \\
&\quad - \Pr[C_n(2, m) \in B,\, X_1 > u_n,\, M_{2,m} > u_n] \Big| \\
&\quad \le \sum_{k=1}^{m-1} \Pr[X_1 > u_n,\, M_{2,k} \le u_n,\, X_{k+1} > u_n,\, M_{m+1,m+k} > u_n] \\
&\qquad + \Pr[X_1 > u_n,\, M_{2,m} \le u_n,\, X_{m+1} > u_n] \\
&\quad \le \Pr[X_1 > u_n,\, M_{m+1,2m} > u_n]. \quad (14)
\end{aligned}$$

Now combine (8–11) and (14) to see that

$$\begin{aligned}
&\Big| \Pr[C_n \in B,\, M_{r_n} > u_n] \\
&\quad - r_n \Big\{ \Pr[C_n(1, m) \in B,\, X_1 > u_n] \\
&\qquad\qquad - \Pr[C_n(2, m) \in B,\, X_1 > u_n,\, M_{2,m} > u_n] \Big\} \Big| \\
&\quad \le 4m \Pr[X_1 > u_n] + 2r_n \Pr[X_1 > u_n,\, M_{m+1,r_n} > u_n] \\
&\qquad + r_n \Pr[M_{r_n - m} > u_n,\, X_{r_n} > u_n]. \quad (15)
\end{aligned}$$

Divide both sides of (15) by $r_n \Pr[X_1 > u_n]$ and set $B = \mathbb{A}$ to obtain (5). Combine (5), (7), and (15) to obtain (6). □

4 Cluster peak approximation

Next we derive an approximation of the distribution of C_n conditionally on $\{M_{r_n} > u_n\}$ in terms of the conditional distribution of $(X_1, \ldots, X_{2m})$

conditionally on X_{m+1} being the cluster peak, that is, $X_{m+1} = M_{2m} > u_n$. Allowing for ties, we actually require $m+1$ to be the first time that the maximum is reached, that is, $u_n \vee M_m < X_{m+1} = M_{2m}$, where $x \vee y := \max(x, y)$. The approximation provides theoretical underpinning for the simulation scheme proposed by Smith et al. (1997).

Theorem 4.1 *Let $\{X_n\}_{n\geq 1}$ be a stationary sequence and let u_n and r_n be as in (1). For measurable $B \subset \mathbb{A}$, define*

$$\begin{aligned}\varepsilon_{m,n}(B) &= \Big|\Pr[C_n \in B \mid M_{r_n} > u_n] \\ &\qquad - \Pr[C_n(1,2m) \in B \mid u_n \vee M_m < X_{m+1} = M_{2m}]\Big|.\end{aligned}$$

If (3), then

$$\lim_{m\to\infty} \limsup_{n\to\infty} |\theta_n - \Pr[M_m < X_{m+1} = M_{2m} \mid X_{m+1} > u_n]| = 0 \tag{16}$$

and

$$\lim_{m\to\infty} \limsup_{n\to\infty} \sup_B \varepsilon_{n,m}(B) = 0. \tag{17}$$

Proof. Let B be an arbitrary measurable subset of $\mathbb{A}$. We start from equation (7) in the proof of Theorem 3.1. This time, split the event $\{M_{r_n} > u_n\}$ according to the first time $j = 1, \ldots, r_n$ that the maximum is reached:

$$\Pr[C_n \in B,\, M_{r_n} > u_n] = \sum_{j=1}^{r_n} \Pr[C_n \in B,\, u_n \vee M_{j-1} < X_j = M_{r_n}], \tag{18}$$

where $M_0 = -\infty$. Fix a positive integer m and let n be large enough so that $r_n \geq 2m+1$. For integer j such that $m+1 \leq j \leq r_n - m$, we have $C_n(1, r_n) = C_n(j-m, j+m-1)$ unless $M_{j-m-1} > u_n$ or $M_{j+m,r_n} > u_n$. Hence, for such j,

$$\begin{aligned}&\Big|\Pr[C_n \in B,\, u_n \vee M_{j-1} < X_j = M_{r_n}] \\ &\quad - \Pr[C_n(j-m, j+m-1) \in B,\, u_n \vee M_{j-m,j-1} < X_j = M_{j-m,j+m-1}]\Big| \\ &\leq \Pr[M_{j-m+1} > u_n,\, X_j > u_n] + \Pr[X_j > u_n,\, M_{j+m,r_n} > u_n] \\ &\leq \Pr[M_{r_n-m+1} > u_n,\, X_{r_n} > u_n] + \Pr[X_1 > u_n,\, M_{m+1,r_n} > u_n] \\ &\leq 2\Pr[X_1 > u_n] \sum_{i=m+1}^{r_n} \Pr[X_i > u_n \mid X_1 > u_n],\end{aligned} \tag{19}$$

where we used stationarity and (4). By stationarity, the second term on the left-hand side of (19) is the same for all $j \geq m+1$. Hence, (18) and (19)

together imply

$$\begin{aligned}
&\Big| \Pr[C_n \in B,\, M_{r_n} > u_n] \\
&- r_n \Pr[C_n(1,2m) \in B,\, u_n \vee M_m < X_{m+1} = M_{2m}] \Big| \\
&\leq \quad 4m \Pr[X_1 > u_n] + 2r_n \Pr[X_1 > u_n] \sum_{i=m+1}^{r_n} \Pr[X_i > u_n \mid X_1 > u_n]. \quad (20)
\end{aligned}$$

Define $\delta_{n,m}(B)$ by

$$\begin{aligned}
&\Big| \Pr[C_n \in B \mid M_{r_n} > u_n] \\
&\quad - \quad \theta_n^{-1} \Pr[C_n(1,2m) \in B,\, M_p < X_{m+1} = M_{2m} \mid X_{m+1} > u_n] \Big|.
\end{aligned}$$

By (7) and (20),

$$\sup_B \delta_{n,m}(B) \leq \theta_n^{-1} \left(\frac{4m}{r_n} + 2 \sum_{i=m+1}^{r_n} \Pr[X_i > u_n \mid X_1 > u_n] \right).$$

Since $\liminf \theta_n > 0$, we arrive at

$$\lim_{m\to\infty} \limsup_{n\to\infty} \sup_B \delta_{n,m}(B) = 0. \quad (21)$$

The choice $B = \mathbb{A}$ yields (16). This, together with (21), yields (17), finishing the proof. □

Remark. In Theorems 3.1 and 4.1, condition (3) can be replaced by the weaker one,

$$\left.\begin{array}{l} \lim\limits_{m\to\infty} \limsup\limits_{n\to\infty} \Pr[M_{m+1,r_n} > u_n \mid X_1 > u_n] \\ \lim\limits_{m\to\infty} \limsup\limits_{n\to\infty} \Pr[M_{r_n-m} > u_n \mid X_{r_n} > u_n] \end{array}\right\} = 0.$$

The gain is small, however, since typically the easiest way to check this condition is by actually proving (3).

5 Discussion

If the long-range dependence in the sequence $\{X_n\}$ is sufficiently weak at extreme levels, then the limit of θ_n, provided it exists, is called the *extremal index*, $\theta \in [0,1]$ (Leadbetter, 1983). The extremal index is a summary measure for the propensity of extremes to cluster in the limit. It provides the proper way to connect the marginal distribution of the sequence to the asymptotic distribution of the sample maximum, and extreme quantiles of

the marginal distribution to high return levels of the sequence. The characterization (2) in terms of block maxima was derived by Leadbetter (1983), while the characterization (5) in terms of a run of non-exceedances immediately following an exceedance was established by O'Brien (1974, 1987).

In this paper, the connection between the two characterizations of the extremal index has been shown using only the condition (3). Moreover, a new characterization (16) has been established in terms of cluster peaks. In analogy to the blocks and runs estimators, which are the sample versions of (2) and (5) respectively (Hsing, 1991 and 1993), the empirical counterpart of (16) suggests what could be called the *peaks estimator* of the extremal index. More generally, the cluster distribution may be estimated by the sample analogues of the expressions given in the two theorems.

References

Hsing, T., 1991. Estimating the parameters of rare events. Stochast. Process. Applic. 37, 117–139.

Hsing, T., 1993. Extremal index estimation for a weakly dependent stationary sequence. Ann. Statist. 21, 2043–2071.

Leadbetter, M.R., 1983. Extremes and local dependence in stationary sequences. Z. Wahrscheinlichkeitsth. Verw. Geb. 65, 291–306.

O'Brien, G.L., 1974. The maximum term of uniformly mixing stationary processes. Z. Wahrscheinlichkeitsth. Verw. Geb. 30, 57–63.

O'Brien, G.L., 1987. Extreme values for stationary and Markov sequences. Ann. Probab. 15, 281–291.

Perfekt, R., 1994. Extremal behaviour of stationary Markov chains with applications. Ann. Appl. Probab. 4, 529–548.

Segers, J., 2003. Functionals of Clusters of Extremes. Adv. Appl. Probab. 35, to appear.

Smith, R.L., 1992. The extremal index for a Markov chain. J. Appl. Probab. 29, 37–45.

Smith, R.L., Tawn, J.A., Coles, S.G., 1997. Markov chain models for threshold exceedances. Biometrika 84, 249–268.

Yun, S., 1998. The extremal index of a higher-order stationary Markov chain. Ann. Appl. Probab. 8, 408–437.

Yun, S., 2000. The distributions of cluster functionals of extreme events in a dth-order Markov chain. J. Appl. Probab. 37, 29–44.