A regeneration-based runs estimator for the extremal index in the Markov setup

Patrice Bertail, Stéphan Clémençon, Jessica Tressou

A regeneration-based runs estimator for the extremal index in the Markov setup

Patrice Bertail[1], Stéphan Clémençon[2,3], and Jessica Tressou[3]

[1] MODAL'X - Université Paris X Nanterre et CREST-LS
(e-mail: pbertail@u-paris10.fr)

[2] LTCI UMR GET/CNRS 5141- Telecom Paris
(e-mail: stephan.clemencon@enst.fr)

[3] Unité Met@risk - INRA
(email: jessica.tressou@agroparistech.fr)

Abstract. It is the purpose of this paper to introduce a novel estimator for the extremal index of an *instantaneous function* $\{f(X_n)\}_{n\in\mathbb{N}}$ of a regenerative Harris Markov chain X, based on the renewal properties of the latter. The estimate proposed may be viewed as a "regenerative version" of the *runs estimator*, insofar as it measures the clustering tendency of high threshold exceedances within regeneration cycles. Strong consistency of this estimator is established under mild stochastic stability assumptions and a simulation result is displayed in the case when the underlying chain is the waiting process related to a simple M/M/1 queue.
Keywords. Regenerative Markov Chain, Extremal Index, Runs Estimator.

1 Introduction

A key parameter in the extremal behavior analysis of (approximately) stationary sequences $Y = \{Y_n\}_{n\in\mathbb{N}}$ of dependent r.v.'s, when well defined, is the *extremal index* $\theta_Y \in (0,1)$, measuring to which extent extreme values tend to come in "small clusters" (refer to Embrechts et al. (1997), Coles (2001), Finkenstadt and Rootzén (2003) for an account of this notion). Indeed, assuming that Y is ergodic with limiting probability measure μ, it allows to connect the distribution of the sample maximum to its counterpart in the case where the Y_n's would be i.i.d. with common distribution μ:

$$\mathbb{P}(\max_{1\leq k\leq n} Y_k \leq u) \approx \mu(]-\infty,\ u])^{n\theta_Y},\ \text{as } u\uparrow\infty. \tag{1}$$

As a continuation of the results established in Bertail et al. (2007), this paper is devoted to introduce a novel statistical methodology for estimating this parameter in the case where the sequence of interest is an *instantaneous function* of a time-homogeneous regenerative Markov chain $X = \{X_n\}_{n\in\mathbb{N}}$ with state space $(E,\mathcal{E})$, *i.e.* a sequence of the form $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$ where $f : E \to \mathbb{R}$ is a measurable function.

Various extremal index estimators have been recently proposed in the statistical literature (see Ancona-Navarette and Tawn (2000), Laurini and Tawn (2003), Hsing (1993) for instance), which generally rely on *blocking techniques*, where data segments of fixed (deterministic) length are considered in order to account for the dependence structure within the observations, whereas we propose here a methodology specifically tailored for regenerative sequences. Roughly speaking, data blocks correspond here to *cycles* (of random length) in between successive regeneration times and our procedure boils down to counting how many times over the observed sample path, within a cycle, solely the first observation exceeds a given high threshold u and then dividing the result by the number of cycles with a first observation above u. First developed in the seminal work of Rootzén (1988), the idea of exploiting X's renewal properties for extremal values analysis has recently been revisited in Bertail et al. (2007) from a statistical perspective.

The paper is structured as follows. Notation are set out in section 2, together with a list of required assumptions. The *regenerative runs estimator* for the extremal index of a sequence $f(X)$ is then defined in the next section, where its strong consistency is established under mild hypotheses. Eventually, a simulation result is briefly presented in section 4, while technical details are postponed to the Appendix.

2 Notation and assumptions

Here and throughout $X = \{X_n\}_{n\in\mathbb{N}}$ is a Harris recurrent time-homogeneous Markov chain, valued in a measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$ and initial distribution ν (see Revuz (1984) for an account of the Markov chain theory). Recall that Harris recurrence boils down to assuming the existence of a positive measure ψ (namely, a *maximal irreducibility measure*) such that, for any measurable set $B \in \mathcal{E}$, the condition "$\psi(B) > 0$" entails that it is visited by the chain infinitely many times with probability one, no matter what the initial state.
A Markov chain is said *regenerative* when it possesses an accessible atom, *i.e.*, a measurable set A such that $\psi(A) > 0$ and $\Pi(x, .) = \Pi(y, .)$ for all x, y in A. Denote then by $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ the hitting time on A, by $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\}$ for $j \geq 2$ the successive return times to A, also termed *regeneration times*, insofar as they are times at which X forgets its past. Indeed, it follows from the *strong Markov property* that the data blocks determined by the latter (namely, the *regeneration cycles*)

$$\mathcal{B}_1 = (X_{\tau_A(1)+1}, ...,\ X_{\tau_A(2)}), ...,\ \mathcal{B}_j = (X_{\tau_A(j)+1}, ...,\ X_{\tau_A(j+1)}), ...,$$

are i.i.d., valued in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$.

Denote by $\mathbb{P}_\nu$ (resp. $\mathbb{P}_A$) the probability measure on the underlying space such that $X_0 \sim \nu$ (resp. $X_0 \in A$) and by $\mathbb{E}_\nu[.]$ (resp. $\mathbb{E}_A[.]$) the corresponding expectation.

In the regenerative setup, stochastic stability properties classically boil down to checking conditions related to the speed of return times to the regenerative set. It is well-known for instance that X is *positive recurrent* if and only if $\alpha = \mathbb{E}_A[\tau_A] < \infty$ (see Theorem 10.2.2 in Meyn and Tweedie (1996)), and its (unique) invariant probability distribution μ is then the Pitman's occupation measure given by $\mu(B) = \alpha^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\}]$ for all $B \in \mathcal{E}$.

Under adequate conditions related to the distribution of the regenerative blocks, standard limit theorems can be classically derived from the application of the corresponding results in the i.i.d. setting to the $\mathcal{B}_j$'s blocks (see Smith (1992)). The following assumptions are involved in the analysis below.
Assumptions. Let $\kappa \geq 1$. $\mathcal{H}(\kappa)$: $\mathbb{E}_A[\tau_A^\kappa] < \infty$ and $\mathcal{H}(\nu, \kappa)$: $\mathbb{E}_\nu[\tau_A^\kappa] < \infty$.

3 Regeneration-based estimation of the extremal index

Let $f : E \to \mathbb{R}$ be measurable. It is well-known that, when X is positive recurrent with limiting distribution μ, the sequence $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$ fulfills *Leadbetter's mixing condition* and its extremal index $\theta(= \theta(f))$ consequently exists (see O'Brien (1987), Leadbetter and Rootzén (1988)). Precisely, we have $\mathbb{P}_\mu(\max_{1\leq i\leq n} f(X_i)) \leq u_n) \underset{n\to\infty}{\sim} F(u_n)^{n\theta}$, for any sequence u_n such that $n(1-F(u_n)) \to \eta < \infty$, denoting by $F(x) = \alpha^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) \leq x\}]$ the cdf of $f(X_1)$ in steady-state (*i.e.* under $\mathbb{P}_\mu$).

Using the regenerative method, it has been proved in Rootzén (1988) that θ may be expressed as a limiting conditional probability:

$$\theta = \lim_{n\to\infty} \mathbb{P}_A(\max_{2\leq i\leq \tau_A} f(X_i) \leq u_n \mid X_1 > u_n). \tag{2}$$

Based on a path $X_1, \ldots, X_n$, the natural empirical counterpart of (2) is

$$\hat{\theta}_n(u) = \frac{\sum_{j=1}^{l_n-1} \mathbb{I}\{\max_{2+\tau_A(j)\leq i\leq \tau_A(j+1)} f(X_i) \leq u < f(X_{1+\tau_A(j)})\}}{\sum_{j=1}^{l_n-1} \mathbb{I}\{f(X_{1+\tau_A(j)}) > u\}}, \tag{3}$$

where $l_n = \sum_{i=1}^{n} \mathbb{I}\{f(X_i) \in A\}$ (with the usual convention regarding empty summation and $\frac{0}{0} = 0$). Insofar as (2) measures the clustering tendency of high threshold exceedances within regeneration cycles only, it should be seen as a "regenerative version" of the *runs estimator*

$$\hat{\theta}_n^{(r)}(u) = \frac{\sum_{j=1}^{n-r} \mathbb{1}\{\max_{j+1\leq i\leq j+r} f(X_i) \leq u < f(X_j)\}}{\sum_{j=1}^{n-r} \mathbb{1}\{f(X_j) > u\}}, \tag{4}$$

obtained by averaging over overlapping data segments of fixed length r.

Beyond its practical advantage (blocks are here entirely determined by the data), the estimator (3) may be proved *strongly consistent* as stated in the next theorem, while only weak consistency has been established for (4) (but for a wider class of weakly dependent sequences, see Hsing (1993)).

Theorem 1. *Let $r_n \uparrow \infty$ in a way that $r_n = o(\sqrt{n/\log\log n})$ as $n \to \infty$. Assume that $\mathcal{H}(\nu,\ 1)$ and $\mathcal{H}(2)$ are fulfilled. Considering $(v_n)_{n\in\mathbb{N}}$ such that $r_n(1-F(v_n)) \to \eta < \infty$ as $n \to \infty$, we then have $\hat{\theta}_n(v_n) \to \theta$ $\mathbb{P}_\nu$-a.s. .*

Remark 1. (Extension to the pseudo-regenerative case) Following the data-driven approach developped in Bertail and Clémençon (2006), in the general Harris setting, one may consider the estimator built from pseudo-regeneration times (approximating the regeneration times of a Nummelin extension) replacing the renewal times by their approximate versions in (3). In spite of the approximation step, the resulting estimator may be still proved consistent, under additional mild hypotheses.

4 A simulation result

Numerical experiments have been carried out from a sequence drawn as the waiting time process X related to a standard M/M/1 queue: $X_{n+1} = \max\{X_n + U_n - \Delta T_{n+1},\ 0\}$ where inter-arrivals and service times, $(\Delta T_n)_{n\geq 1}$ and $(U_n)_{n\geq 1}$, are assumed independent from each other and i.i.d. with exponential distributions of respective intensities λ and μ. If the *load condition* "$\lambda/\mu < 1$" holds, X is classically positive recurrent with the empty file $\{0\}$ as atom. Besides, it is known that X's extremal index is then $\theta = (1-\lambda/\mu)^2$ (see Hooghiemstra and Meester (1995)). Using threshold levels u corresponding

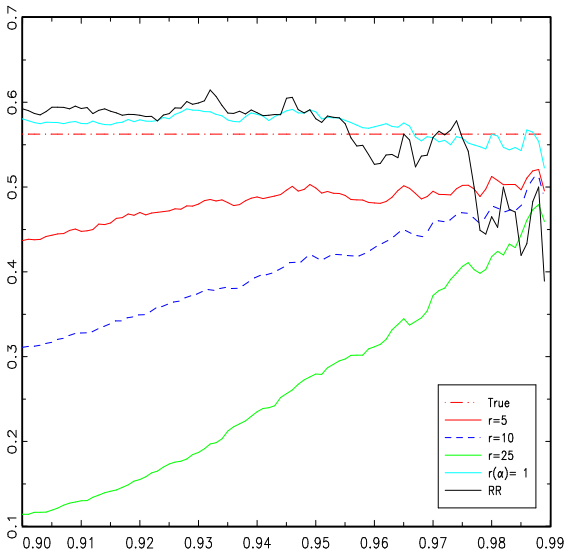

Fig. 1. Estimation of the extremal index in the M/M/1 queue with parameters $\lambda = 0.2$, $\mu = 0.8$, $\theta = 0.56$.

to high percentiles of the X_n's with $n = 10000$ (represented along the x-axis in Fig. 1), $\hat{\theta}_n(u)$ is plotted (y-axis in Fig. 1), together with the standard runs estimates for various lengths r. We observe that the accuracy of our estimator generally surpasses the one of (4), except in the case when a runs length is taken approximately equal to the mean blocklength $r = \lfloor n/l_n \rfloor$, for which value the latter estimate behaves similarly to the regenerative version (3).

A Proof of Theorem 1

Consider the empirical counterparts of the theoretical probabilities $F_1(u) = \mathbb{P}_A(f(X_1) \leq u)$ and $H_1(u) = \mathbb{P}_A(\max_{2 \leq i \leq \tau_A} f(X_i) \leq u < f(X_1))$

$$F_{1,l_n}(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} \mathbb{I}\{f(X_{1+\tau_A(j)}) > u\},$$

$$H_{1,l_n}(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} \mathbb{I}\{\max_{2+\tau_A(j) \leq i \leq \tau_A(j+1)} f(X_i) \leq u < f(X_{1+\tau_A(j)})\}.$$

Equipped with this notation, $\hat{\theta}_n(u) = H_{1,l_n}(u)/(1 - F_{1,l_n}(u))$. The fact that $\hat{\theta}_n(u_n) \to \theta = \lim_{n\to\infty} \frac{H_1(u_n)}{1-F_1(u_n)}$ immediately follows from the decomposition

$$\hat{\theta}_n(u) - \theta = \frac{1 - F(u)}{1 - F_{1,l_n}(u)} \cdot \{\frac{H_{1,l_n}(u) - H_1(u)}{1 - F(u)} - \theta \frac{F_{1,l_n}(u) - F_1(u)}{1 - F(u)} + \frac{H_1(u) - \theta(1 - F_1(u))}{1 - F(u)}\}.$$

combined with the next lemma (of which proof is a slight modification of the one of Lemma 6 in Bertail et al. (2007) and is thus omitted) and the fact that we choose $r_n = o(n/\log\log n)$ as $n \to \infty$.

Lemma 1. *(*LIL FOR FUNCTIONALS OF POSITIVE CHAINS*) Let X be a regenerative chain, fulfilling assumptions $\mathcal{H}(\nu, 1)$ and $\mathcal{H}(2)$. We then have*

1. $\limsup_{n\to\infty} \frac{\sup_{u\in\mathbb{R}} |F_{1,l_n}(u) - F_1(u)|}{\sqrt{(2\sigma^2_{F_1} \log\log n)\alpha/n}} = +1$ $\mathbb{P}_\nu$*-a.s., with* $\sigma^2_{F_1} = \sup_{u\in\mathbb{R}} \sigma^2_{F_1}(u)$ *and, for all* $u \in \mathbb{R}$, $\sigma^2_{F_1}(u) = F_1(u)(1 - F_1(u))$.
2. $\limsup_{n\to\infty} \frac{\sup_{u\in\mathbb{R}} |H_{1,l_n}(u) - H_1(u)|}{\sqrt{(2\sigma^2_{H_1} \log\log n)\alpha/n}} = +1$ $\mathbb{P}_\nu$*-a.s., with* $\sigma^2_{H_1} = \sup_{u\in\mathbb{R}} \sigma^2_{H_1}(u)$ *and, for all* $u \in \mathbb{R}$, $\sigma^2_{H_1}(u) = H_1(u)(1 - H_1(u))$.

References

Ancona-Navarette, M. A. and Tawn, J. A. (2000): A comparison of methods for estimating the extremal index. *Extremes*, *3*, 5–38.

Bertail, P. and Clémençon, S. (2006): Regenerative-block bootstrap for Markov chains. *Bernoulli*, *12*, 689–712.

Bertail, P., Clémençon, S. and Tressou, J. (2007): Extreme value statistics for Markov chains via the (pseudo-)regenerative method. Available at `http://hal.archives-ouvertes.fr/hal-00165652`.

Coles, S. (2001): *An introduction to statistical modelling of Extreme Values*. Springer series in Statistics. Springer.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Springer-Verlag.

Finkenstadt, B. and Rootzén, H. (2003): *Extreme values in Finance, Telecommunications and the Environment*. volume 99 of *Monograph on Statistics and Applied Probability*. Chapman & Hall.

Hooghiemstra, G. and Meester, L. E. (1995): Computing the extremal index of special Markov chains and queues. *Stoch. Proc. Appl.*, *65*, 171–185.

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

Laurini, F. and Tawn, J. A. (2003): New estimators for the extremal index and other cluster characteristics. *Extremes*, *6*, 189–211.

Leadbetter, M. R. and Rootzén, H. (1988): Extremal theory for stochastic processes. *Ann. Probab.*, *16*, 431–478.

Meyn, S. P. and Tweedie, R. L. (1996): *Markov Chains and Stochastic Stability*. Springer-Verlag.

O'Brien, G. L. (1987): Extreme values for stationnary and Markov sequences. *Ann. Probab.*, *15*, 281–291.

Revuz, D. (1984): *Markov Chains*. 2nd edition, North-Holland.

Rootzén; H. (1988): Maxima and exceedances of stationary Markov chains. *Adv. Appl. Probab.*, *20*, 371–390.

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