# Extreme values statistics for Markov chains with applications to Finance and Insurance

Patrice Bertail, MODAL'X, université Paris-Ouest
Stéphan Clémençon, TSI, TelecomParisTech
Charles Tillier, MODAL'X, université Paris-Ouest

July 9, 2015

**Abstract**

We review in this paper several statistical methods, specifically tailored for Markov processes with a view towards their extremal behavior. Precisely, this paper proposes some statistical inference tools for extremal events from a regeneration theory angle.Indeed, Harris Markov chains may be decomposed into independent *regeneration cycles*, namely data segments between consecutive regeneration times $\tau_1$, $\tau_2$, ... (*i.e.* random times at which the chain forgets its past). Working on this approach, the methodology proposed in this paper boils down to split up the observed sample path into regeneration data blocks (or into data blocks drawn from a distribution approximating the regeneration cycle's distribution, in the general case when regeneration times cannot be observed). Then, the analysis boils down to examining the sequence of maxima over the resulting data segments, as if they were i.i.d. We focus on the estimation of the *extremal dependence index* and the *tail index*. We illustrate the method on two examples taken from the insurance and finance literature, ruin models and times series exhibiting smooth threshold and/or strong conditional heteroscedasticity. An illustration of the estimation methods to the CAC40 shows the potential of regenerative tools for real data applications.

**Keywords and phrases:** regenerative Markov chain, Nummelin splitting technique, extreme values statistics, cycle submaximum, Hill estimator, extremal index, ruin theory.
**AMS 2000 Mathematics Subject Classification:** 60G70, 60J10, 60K20.

## 1 Introduction

Extremal events for (strongly or weakly) dependent data have received an increasing attention in the statistical literature in the last past years (see Newell (1964); Loynes (1965), O'Brien (1974, 1987),Hsing (1988, 1991, 1993); Resnick and Stărică (1995); Rootzén (2006) for instance). A major issue for evaluating risks and understanding extremes and their possible replications is to take into account some dependencies. Indeed, whereas extreme values naturally occur in an isolated fashion in the identically independent distributed (i.i.d.)

setup, since extreme values may be highly correlated, they generally tend to take place in small clusters for weakly dependent sequences. Most methods for statistical analysis of extremal events in weakly dependent setting rely on (fixed length) *blocking-techniques*, which consist, roughly speaking, in dividing an observed data series into (overlapping or non overlapping) blocks of fixed length. Examine how extreme values occur over these data segments allows to capture the tail and the dependency structure of extreme values.

As originally pointed out in Rootzén (1988), the extremal behavior of instantaneous functionals $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$ of a Harris recurrent Markov chain $X$ may be described through the regenerative properties of the underlying chain. The present paper emphasizes the importance of renewal theory and regeneration from the perspective of statistical inference for extremal events. Indeed, as observed by Rootzén (1988) (see also Asmussen (1998a,b); Haiman et al. (1995); Hansen and Jensen (2005)), certain parameters of extremal behavior features of Harris Markov chains may be also expressed in terms of *regeneration cycles*, namely data segments between consecutive regeneration times $\tau_1$, $\tau_2$, ..., *i.e.* random times at which the chain completely forgets its past. Following in the footsteps of the seminal contribution of Rootzén (1988) (see also Asmussen (1998a)), Bertail et al. (2009) and Bertail et al. (2013) have recently investigated the performance of regeneration-based statistical procedures for estimating key parameters, related to the extremal behavior analysis in a Markovian setup. In the spirit of the works of Bertail and Clémençon (2006b) (refer also to Bertail and Clémençon (2004a), Bertail and Clémençon (2004b) Bertail and Clémençon (2006a)), they developed a statistical methodology, called the "pseudo-regenerative method", based on approximating the pseudo-regeneration properties of general Harris Markov chains, for tackling various estimation problems in a Markovian setup. Most of their works deal with regular differentiable functionals like the mean (see Bertail and Clémençon (2004a), Bertail and Clémençon (2007)), the variance, quantiles, L-statistics and their robustified versions (Bertail et al. (2014)), as well as $U$-statistics (Bertail et al. (2011)). Bootstrap versions of these estimates have also been proposed . For regular functionals, they possess the same nice second order properties as the bootstrap in the i.i.d case, that is the rate of the convergence of the bootstrap distribution which is close to $n^{-1}$, for regular Markov chains, instead of $n^{-1/2}$ for the asymptotic (Gaussian) benchmark (see Bertail and Clémençon (2006b)).

The purpose of this paper is to review and give some extensions of this approach in the framework of extreme values for general Markov chains. The proposed methodology consists in splitting up the observed sample path into regeneration data blocks (or into data blocks drawn from a distribution approximating the regeneration cycle's distribution, in the general case when regeneration times cannot be observed). We mention that the estimation principle exposed in this paper is by no means restricted to the sole Markovian setup, but indeed applies to any process for which a regenerative extension can be constructed and simulated from available data (see chap. 10 in Thorisson (2000)). Then, statistical tools are built over the sequence of maxima over the resulting data segments, as if these maxima were i.i.d.. In order to illustrate the interest of this technique, we focus on the question of estimating the sample maximum's tail, the *extremal dependence index* and the *tail index*

by means of the (pseudo-) regenerative method. To motivate this approach in financial and insurance applications (as well as queuing or inventory models), we illustrate how these tools may be used in order to estimate ruin probabilities or extremal index, in ruin models with a dividend barrier, exhibiting some regenerative properties. Such applications have also straightforward extensions (for continuous Markov chains) in the field of finance, for instance for put option pricing (for which the "strike" plays here the role of the ruin level).

## 2 On the (pseudo-) regenerative approach for Markovian data

Here and throughout, $X = (X_n)_{n\in\mathbb{N}}$ denotes a $\psi$-irreducible aperiodic time-homogeneous Markov chain, valued in a (countable generated) measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$ and initial distribution $\nu$. We recall that the Markov property means that, for any set $B$, such that $\psi(B) > 0$, for any sequence $(x_n, x_{n-1}, .....)$ in $E$,

$$\begin{aligned} P(X_{n+1} \in B \ \mid \{X_j = x_j, j \leq n\}) &= P(X_{n+1} \in B \ \mid X_n = x_n) \\ &= \Pi(x_n, B). \end{aligned}$$

For homogeneous Markov chains, the transition probability does not depend on $n$. Refer to Revuz (1984) and Meyn and Tweedie (1996), for basic concepts of the Markov chain theory. For sake of completeness, we specify the two following notions :

- The chain is *irreductible* if there exists a $\sigma$-finite measure $\psi$ such that for all set $B \in \mathcal{E}$, when $\psi(B) > 0$, the chain visits $B$ with a strictly positive probability, no matter what the starting point.
- Assuming $\psi$-irreducibility, there is $d' \in \mathbb{N}^*$ and disjointed sets $D_1, ...., D_{d'}$ ($D_{d'+1} = D_1$) weighted by $\psi$ such that $\psi(E \backslash \cup_{1\leqslant i\leqslant d'} D_i) = 0$ and $\forall x \in D_i, \Pi(x, D_{i+1}) = 1$. The *period* of the chain is the greatest common divisor $d$ of such integers. It is *aperiodic* if $d = 1$.
- The chain is said to be recurrent if any set $B$ with positive measure $\psi(B) > 0$, i.f.f the set $B$ is visited an infinite number of times.

The first notion formalizes the idea of a communicating structure between subsets and the second notion considers the set of time points at which such communication may occur. Aperiodicity eliminates deterministic cycles. If the chain satisfies these three properties, it is said to be **Harris recurrent.**

In what follows, $\mathbb{P}_\nu$ (respectively, $\mathbb{P}_x$ for $x$ in $E$) denotes the probability measure on the underlying space such that $X_0 \sim \nu$ (resp., conditioned upon $X_0 = x$), $\mathbb{E}_\nu[.]$ the $\mathbb{P}_\nu$-expectation (resp. $\mathbb{E}_x[.]$ the $\mathbb{P}_x(.)$-expectation) and $\mathbb{I}\{\mathcal{A}\}$ the indicator function of any event $\mathcal{A}$. We assume further that $X$ is positive recurrent and denote by $\mu$ its (unique) invariant probability distribution.

## 2.1 Markov chains with regeneration times : definitions and examples

A Markov chain $X$ 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 $A$. A recurrent Markov chain taking its value in a finite set is always atomic since each visited point is itself an atom. Queuing systems or ruin models visiting an infinite number of time the value $0$ (the empty queue) or a given level (for instance a barrier in the famous Cramér-Lundberg model, see Embrechts et al. (1997) and the examples below) are also naturally atomic. Refer also to Asmussen (2003) for regenerative models involved in queuing theory, see also the examples and the applications below.

Denote then by $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ the hitting time on $A$ or first return time to $A$. Put also $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\},\ j \geq 2$ for the so called **successive return times** to $A$, corresponding to the time of successive visits to the set $A$.

In the following $\mathbb{E}_A[.]$ denotes the expectation conditioned on the event $\{X_0 \in A\}$. When the chain is Harris recurrent, for any starting distribution, the probability of returning infinitely often to the atom $A$ is equal to one. Then, for any initial distribution $\nu$, by the *strong Markov property*, the sample paths of the chain may be divided into **i.i.d. blocks** of random length corresponding to consecutive visits to $A$, generally called *regeneration cycles*:

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

taking their values in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$. The renewal sequence $\{\tau_A(j)\}_{j\geq 1}$ defines successive times at which the chain forgets its past, termed *regeneration times*.

**Example 1 : Queuing system or storage process with an empty queue.**

We consider here a storage model (or a queuing system), evolving through a sequence of *input times* $(T_n)_{n\in\mathbb{N}}$ (with $T_0 = 0$ by convention), at which the storage is refilled. Such models appear naturally in many domains like hydrology, operation research, but also for modeling computer CPU occupancy.

Let $X_n$ be the size of the input into the storage system at time $T_n$. Between each input time, it is assumed that withdrawals are done from the storage system at a constant rate $r$. Then, in a time period $[T, T + \Delta T]$, the amount of stored contents which disappears is equal to $r\Delta T$. If $X_n$ denotes the amount of contents immediately before the input time $T_n$, we have for all $n \in \mathbb{N}$,

$$X_{n+1} = (X_n + U_{n+1} - r\Delta T_{n+1})_+ ,$$

with $(x)_+ = \sup(x, 0)$, $X_0 = 0$ by convention and $\Delta T_n = T_n - T_{n-1}$ for all $n \geq 1$ and $T_0 = 0$. $\Delta T_n$ is sometimes called the waiting time period.

This model can be seen as a reflected random walk on $\mathbb{R}^+$. Assume that, conditionally to $X_1, ..., X_n$, the amounts of input $U_1, ..., U_n$ are independent from each other and independent from the inter-arrival times $\Delta T_1, ..., \Delta T_n$ and that the distribution of $U_i$ is given by $K(X_i, .)$, for $0 \leqslant i \leqslant n$. Under the further assumption that $(\Delta T_n)_{n\geqslant 1}$ is an i.i.d.

sequence, independent from $U = (U_n)_{n\in\mathbb{N}}$, the storage process $X$ is a Markov chain. The case with exponential input-output has been extensively studied in Asmussen (1998a).

It is known that the chain $\Pi$ is irreducible as soon as $K(x, .)$ has an infinite tail for all $x \geqslant 0$ and if in addition $EU_n - rE\Delta T_{n+1} < 0$, $\{0\}$ is an accessible atom of the chain $X_n$. Moreover, if $U_{n+1} - r\Delta T_{n+1}$ has exponential tails, then the chain is exponentially geometrically ergodic. The case with heavy tails has been studied in details by Asmussen (1998b) and S. Asmussen and Höjgaard (2000). Under some technical assumptions, the chain is recurrent positive and the times at which the storage process $X$ reaches the value $0$ are regeneration times. This property allows to define regeneration blocks dividing the sample path into independent blocks, as shown below. Figure 1 represents the storage process with $\Delta T_i$ and $X_i$ with $\gamma(1)$ distribution and $r = 1.05$. The red line corresponds to the atom $A = \{0\}$ and the green lines are the corresponding renewal times (visit to the atom).

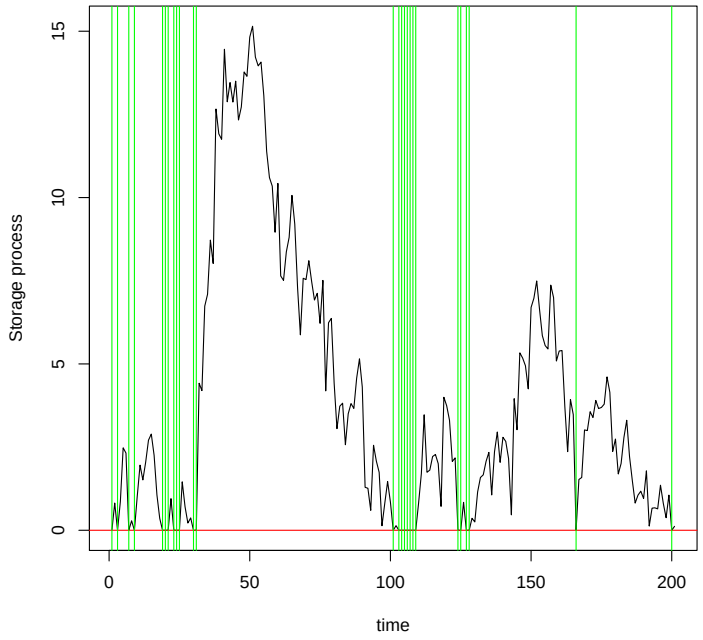


Figure 1: Splitting a reflected random walk, with an atom at $\{0\}$; vertical lines corresponds to regeneration times, at which the chain forgets its past. A block is a set of observations between two lines (it may be reduced to $\{0\}$ in some case)

Notice that the blocks are of random size. Some are rather long (corresponding to large excursion of the chain), others reduce to the point $\{0\}$ if the chain stays at $0$ for several

periods. In this example, for the given values of the parameters, the mean length of a block is close to 50.5. It is thus clear that we need a lot of observations to get enough blocks. The behavior of the maximum of this process for subexponential arrivals has been studied at length in Asmussen (1998b).

**Example 2 : Cramér-Lundberg with a dividend barrier.**

Ruin models, used in insurance, are dynamic models in continuous time which describe the behavior of the reserve of a company as a function of

i) its initial reserve $u$ (which may be chosen by the insurer),

ii) the claims which happen at some random times (described by a arrival claim process),

iii) the premium rate which is the price paid by customers per unit of time.

In the classical Cramér-Lundberg model, the claim arrival process $\{N(t), t \geq 0,\ N(0) = 0\}$ is supposed to be an homogeneous Poisson process with rate $\lambda$, modeling the number of claims in an interval $[0, t]$. The claims sizes $U_i,\ i = 1,....\infty$ , that an insurance company has to face, are assumed to be strictly positive and independent, with cumulative distribution function (cdf) $F$. The premium rate is supposed to be constant equal to $c$. Then, the total claim process, given by $S(t) = \sum_{i=1}^{N(t)} U_i$ is a compound Poisson process. Starting with an initial reserve $U(0) = u$, the reserve of the company evolves as

$$\begin{aligned} R(t) &= u + ct - S(t) \\ &= u + \sum_{n=1}^{N(t)} (c\Delta T_n - U_n). \end{aligned}$$

One of the major problems in ruin models for insurance company is how to choose the initial amount to avoid the ruin or at least ensure that the probability of ruin over a finite horizon (or an infinite one) is small, equal to some given error of first kind, for instance $10^{-3}$. The probability of ruin for an initial reserve $u$ over an horizon $[0, T]$ is given by

$$\psi(u, T) = P(\inf_{t\in[0,T]}(R(t)) < 0).$$

Notice that this model is very close to the queuing process considered in example 1. The *input times* $(T_n)_{n\in\mathbb{N}}$ correspond here to the times of the claims. It is easy to see that under the given hypotheses, the inter-arrival times $(\Delta T_n)_{n\in\mathbb{N}}$ are i.i.d with exponential distribution $\gamma(1, \lambda)$ (with $E\Delta T_n = 1/\lambda$). However, most of the time, for a given company, we only observe (at most) one ruin (since it is an absorbing state), and the reserve is not allowed to grow over a given barrier. Actually, if the process $R(t)$ crosses a given threshold $b$, the money is redistributed in some way to the shareholders of the company. This threshold is called a **dividend barrier**. In this case the process of interest is rather

$$X(t) = (u + ct - R(t)) \wedge b,$$

where $a \wedge b$ designs the infimum between $a$ and $b$. Of course, the existence of a barrier reinforces the risk of ruin especially if the claim size may be large in particular if their

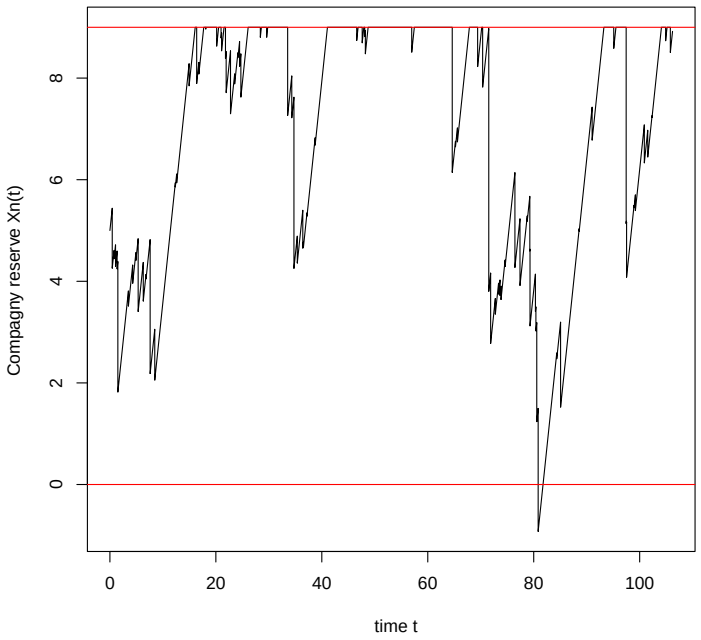


Figure 2: Cramér-Lundberg model with a dividend barrier at b=9 (where the chain is reflected); ruin occurs at t=81 when the chain goes below 0.

distributions have a fat tail. The embedded chain is defined as the value of $X(t)$ at the claim times, say $X_n = X(T_n)$ then it is easy to see that we have

$$X_{n+1} = \inf(X_n + c\Delta T_n - U_{n+1}, b) \text{ with } X_0 = u.$$

Otherwise, the probability of no ruin is clearly linked to the behavior of $Max_{1\leq i\leq n}(-X_n)$.

In comparison to example 1, this model is simply a mirror process, with this time, an atom at $\{b\}$ instead of $\{0\}$ as shown in the two graphics below. In this example, the $(\Delta T_n)_{n\in\mathbb{N}}$ are exponential and the claims with exponential tails, the initial reserve is 5 and the barrier at 9. In this simulation the "ruin" is attained at time t=81.

The embedded chain shows that the barrier is attained several times and allows to build regeneration times (in green) and independent blocks just as in the first example. Because of the choice of the parameters (fat tail for the claims), the number of blocks is small on this short period but in practical insurance applications, we may hope to have more regenerations...

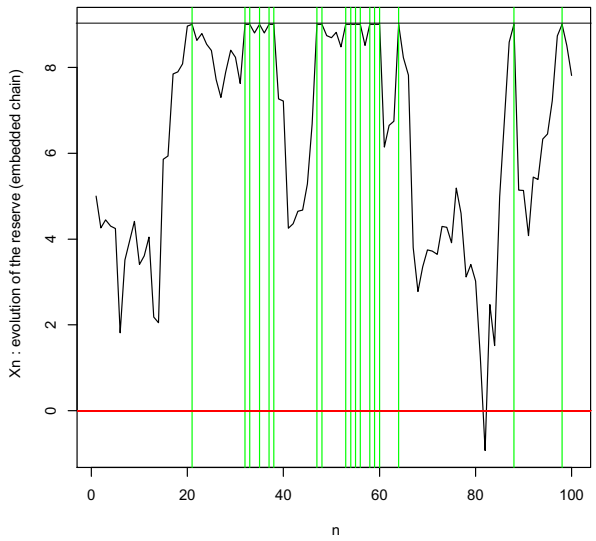


Figure 3: Splitting the embedded chain of a Cramér-Lundberg model with a dividend barrier. Vertical lines corresponds to regeneration times (when the chain attains the barrier b=9). The blocks of observations between two vertical lines are independent.

## 2.2 Basic regeneration properties

When an accessible atom exists, the *stochastic stability* properties of the chain are reduced to properties concerning the speed of return time to the atom only. Theorem 10.2.2 in Meyn and Tweedie (1996) shows for instance that the chain $X$ is positive recurrent if and only if (i.f.f.) $\mathbb{E}_A[\tau_A] < \infty$. The (unique) invariant probability distribution $\mu$ is then the Pitman's occupation measure given by

$$\mu(B) = \frac{1}{\mathbb{E}_A[\tau_A]} \mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\}], \text{ for all } B \in \mathcal{E}. \tag{1}$$

In the case $\mathbb{E}_A[\tau_A] = \infty$, if there exists $0 < \beta < 1$ such that $E\tau_A^\beta < \infty$ and $E\tau_A^{\beta+\eta} = \infty$, for any $\eta > 0$, the chain is said $\beta - \mathrm{null\ recurrent}$ and there exists an invariant measure (not a probability) for the chain. The splitting into independent blocks still holds (see for instance Tjöstheim (1990) and Karlsen and Tjöstheim (2001)). This includes for instance the case of the random walk (with $\beta = 1/2$) and such procedure may be useful for studying the properties of the maximum for Markovian processes which have somehow the same kind of behavior as long range memory processes. We will not consider this more technical case here. For atomic chains, limit theorems can be derived from the application of the corresponding results to the i.i.d. blocks $(\mathcal{B}_n)_{n\geq 1}$ (see Smith (1992) and the references therein). For instance, using this kind of techniques, Meyn and Tweedie (1996) have proved the Law of Large Number (LLN), the Central Limit Theorem (CLT) and Laws of Iterated Logarithm (LIL) for Markov chains. Bolthausen (1980) obtained a Berry-Esseen type theorem and Malinovskiĭ (1985), Malinovskiĭ (1987, 1989); Bertail and Clémençon (2006b) have proved other refinements of the CLT in particular Edgeworth expansions. The same technique can also be applied to establish moment and probability inequalities, which are not asymptotic results (see Clémençon (2001); Bertail and Clémençon (2010)).

Recall that a set $S \in \mathcal{E}$ is said to be *small* for $X$ if there exist $m \in \mathbb{N}^*$, $\delta > 0$ and a probability measure $\Phi$ supported by $S$ such that, for all $x \in S$, $B \in \mathcal{E}$,

$$\Pi^m(x, B) \geq \delta\Phi(B), \tag{2}$$

denoting by $\Pi^m$ the $m$-th iterate of the transition kernel $\Pi$. In the sequel, (2) is referred to as the *minorization condition* $\mathcal{M}(m, S, \delta, \Phi)$. Recall that accessible small sets always exist for $\psi$-irreducible chains : any set $B \in \mathcal{E}$ such that $\psi(B) > 0$ contains such a set (*cf* Jain and Jamison (1967)). In many models of interest $m = 1$ but even if it is not the case it is possible to vectorize the Markov chains to reduce the study of this condition to $m = 1$. Even if it entails replacing the initial chain $X$ by the chain $\{(X_{nm}, ..., X_{n(m+1)-1})\}_{n\in\mathbb{N}}$, we now suppose $m = 1$. From a practical point of view, the minorizing probability measure may be chosen by the user. For instance, $\Phi(B)$ may be the uniform distribution over a given small set, typically a compact set which is often visited by the chain, then in this

case $\delta$ may simply be seen as the minimum of the $\Pi(x, B)$ over $S$. Of course in practice $\Pi$ is unknown but easily estimable so that plug-in estimators of these quantities may be easily constructed (see below).

### 2.3 The Nummelin splitting trick and a constructive approximation

We now precise how to construct the atomic chain onto which the initial chain $X$ is embedded. Suppose that $X$ satisfies $\mathcal{M} = \mathcal{M}(m, S, \delta, \Phi)$ for $S \in \mathcal{E}$ such that $\psi(S) > 0$. The sample space is expanded so as to define a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter $\delta$ by defining the joint distribution $\mathbb{P}_{\nu,\mathcal{M}}$ whose construction relies on the following randomization of the transition probability $\Pi$ each time the chain hits $S$. If $X_n \in S$ and

- if $Y_n = 1$ (with probability $\delta \in ]0, 1[$), then $X_{n+1} \sim \Phi$,
- if $Y_n = 0$, then $X_{n+1} \sim (1 - \delta)^{-1}(\Pi(X_{n+1}, .) - \delta\Phi(.))$.

The key point of the construction relies on the fact that $A_S = S \times \{1\}$ is an atom for the bivariate Markov chain $(X, Y)$, which inherits all its communication and stochastic stability properties from $X$ (refer to Chapt. 14 in Meyn and Tweedie (1996)).

Here we assume further that the conditional distributions $\{\Pi(x, dy)\}_{x\in E}$ and the initial distribution $\nu$ are dominated by a $\sigma$-finite measure $\lambda$ of reference, so that $\nu(dy) = f_\nu(y)\lambda(dy)$ and $\Pi(x, dy) = \pi(x, y)\lambda(dy)$ for all $x \in E$. For simplicity, we suppose that condition $\mathcal{M}$ is fulfilled with $m = 1$. Hence, $\Phi$ is absolutely continuous with respect to $\lambda$ too, and, setting $\Phi(dy) = \phi(y).\lambda(dy)$,

$$\forall x \in S,\ \pi(x, y) \geq \delta\phi(y),\ \lambda(dy)\text{-almost surely.} \tag{3}$$

If we were able to generate binary random variables $Y_1, ..., Y_n$, so that $((X_1, Y_1), ..., (X_n, Y_n))$ be a realization of the split chain described above, then we could divide the sample path $X^{(n)} = (X_1, ..., X_n)$ into regeneration blocks. Given the sample path $X^{(n+1)}$, it may be shown that the $Y_i$'s are independent random variables and the conditional distribution of $Y_i$ is the Bernoulli distribution with parameter

$$\frac{\delta\phi(X_{i+1})}{\pi(X_i, X_{i+1})} \cdot \mathbb{I}\{X_i \in S\} + \delta \cdot \mathbb{I}\{X_i \notin S\}. \tag{4}$$

Therefore, knowledge of $\pi$ over $S^2$ is required to draw $Y_1, ..., Y_n$ by this way.

A natural way of mimicking the Nummelin splitting construction consists in computing first an estimate $\widehat{\pi}_n(x, y)$ of the transition density over $S^2$, based on the available sample path and such that $\widehat{\pi}_n(x, y) \geq \delta\phi(y)$ a.s. for all $(x, y) \in S^2$, and then generating independent Bernoulli random variables $\widehat{Y}_1, \ldots, \widehat{Y}_n$ given $X^{(n+1)}$, the parameter of $\widehat{Y}_i$ being

obtained by plugging $\widehat{\pi}_n(X_i, X_{i+1})$ into (4) in place of $\pi(X_i, X_{i+1})$. We point out that, from a practical point of view, it actually suffices to draw the $\widehat{Y}_i$'s only at times $i$ when the chain hits the small set $S$. $\widehat{Y}_i$ indicates whether the trajectory should be cut at time point $i$ or not. Proceeding this way, one gets the sequence of *approximate regeneration times*, namely the successive time points at which $(X, \widehat{Y})$ visits the set $A_S = S \times \{1\}$. Setting $\widehat{l}_n = \sum_{1 \leq k \leq n} \mathbb{I}\{(X_k, \widehat{Y}_k) \in S \times \{1\}\}$ for the number of splits (that is the number of visits of the approximated split chain to the artificial atom), one gets a sequence of *approximate renewal times*,

$$\widehat{\tau}_{A_S}(j+1) = \inf\{n \geq 1 + \widehat{\tau}_{A_S}(j) / \ (X_n, \widehat{Y}_n) \in S \times \{1\}\}, \text{ for } 1 \leq j \leq \widehat{l}_n - 1, \tag{5}$$

with $\widehat{\tau}_{A_S}(0) = 0$ by convention and forms the *approximate regeneration blocks* $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n - 1}$.

The knowledge of the parameters $(S, \delta, \phi)$ of condition (3) is required for implementing this approximation method. A practical method for selecting those parameters in a fully data-driven manner is described at length in Bertail and Clémençon (2007). The idea is essentially to select a compact set around the mean of the time series and to increase its size. Indeed, if the small set is too small, then there will be no data in it and the Markov chain could not be split. On the contrary, if the small set is too large, the minimum $\delta$ over the small set will be very small and there is little change that the we observe $Y_i = 1$. As the size increases, the number of regenerations increases up to an optimal value and then decreases, the choice of the small set and of the corresponding splitting is then entirely driven by the observations. To illustrate these ideas, we apply the method to a financial time series assuming that it is Markovian (even if there are some structural changes, the Markovian nature still remains).

**Example 3 : Splitting a non regenerative financial time series**.

Many financial time series exhibit some nonlinearities and structural changes both in level and variance. To illustrate how it is possible to divide such kind of data into "almost" independent blocks, we will study a particular model exhibiting such behavior.

Consider the following SETAR(1)-ARCH(1) model (Smooth Exponential Threshold AutoRegressive Model with AutoRegressive Conditional Heteroscedasticity) defined by

$$X_{t+1} = (\alpha_1 + \alpha_2 e^{-X_t^2}) X_t + (1 + \beta X_t^2)^{1/2} \varepsilon_{t+1} \, ,$$

where the noise $(\varepsilon_t)_{t=1,...,n}$ are i.i.d with variance $\sigma^2$. See Fan and Yao (2003) for a detailed description of these kinds of non-linear models. It may be used to model log-returns or log-prices. Notice that this Markov chain (of order 1) may be seen as a continuous approximation of a threshold model. Assume that $|\alpha_1| < 1$, then for large values of $|X_t|$, it is easy to see that in mean $X_{t+1}$ behaves like a simple $AR(1)$ model with coefficient $\alpha_1$ (ensuring that the process will come back to its mean, equal to 0). Conversely, for small values of $X_t$ (close to 0), the process behaves like an $AR(1)$ model with coefficient

$\alpha_1+\alpha_2$ (eventually explosive if $\alpha_1+\alpha_2>1$). This process is thus able to engender bursting bubbles. The heteroscedastic part implies that the conditional variance $Var(X_{t+1}|X_t) = \sigma^2(1+\beta X_t^2)^{1/2}$ may be strongly volatile when large values (the bubble) of the series occur. To ensure stationarity, we require $0<\beta<1$.

In the following simulation, we choose $n = 200$, $\alpha_1 = 0.60$, $\alpha_2 = 0.45$, $\beta = 0.35$ and $\sigma^2 = 1$. The follows graph panel shows the Nadaraya estimator of the transition density, the number of blocks obtained as the size of the small set increases. For a small set of the form $[-0.8, 0.8]$, we obtain $\widehat{l}_n = 21$ pseudo-blocks and the mean length of a block is close to 10. The estimated lower bound for the density over the small set $\widehat{\delta_n}$ is 0.15. The third graphic shows the level sets of the density and the corresponding optimal small set (containing the possible points at which the times series may be split). The last graph shows the original time series and the corresponding pseudo-blocks obtained for an optimal data driven small set.

Beyond the consistency property of the estimators that we will later study, this method has an important advantage that makes it attractive from a practical perspective : blocks are here entirely determined by the data (up to the approximation step), in contrast to standard blocking-techniques based on fixed length blocks. Indeed, it is well known that the choice of the block length is crucial to obtain satisfactory results and is a difficult technical task.

### 2.4 Some hypotheses

The validity of this approximation has been tackled in Bertail and Clémençon (2006a) using a *coupling approach*. Precisely, the authors established a sharp bound for the deviation between the distribution of $((X_i, Y_i))_{1\leq i\leq n}$ and the one of the $((X_i, \widehat{Y}_i))_{1\leq i\leq n}$ in the sense of Wasserstein distance. The coupling "error" essentially depends on the rate of the *mean squared error* (MSE) of the estimator of the transition density

$$\mathcal{R}_n(\hat{\pi}_n,\ \pi) = \mathbb{E}[(\sup_{(x,y)\in S^2} |\hat{\pi}_n(x,y) - \pi(x,y)|)^2], \tag{6}$$

with the sup norm over $S\times S$ as a loss function, under the next conditions :

**A1.** the parameters $S$ and $\phi$ in (3) are chosen so that $\inf_{x\in S}\phi(x) > 0$,

**A2.** $\sup_{(x,y)\in S^2}\pi(x,y) < \infty$ and $\mathbb{P}_\nu$-almost surely $\sup_{n\in\mathbb{N}}\sup_{(x,y)\in S^2}\hat{\pi}_n(x,y) < \infty$.

Throughout the next sections, $f$ denotes a fixed real valued measurable function defined on the state space $E$. To study the properties of the block, we will also need the following usual moment conditions on the time return.

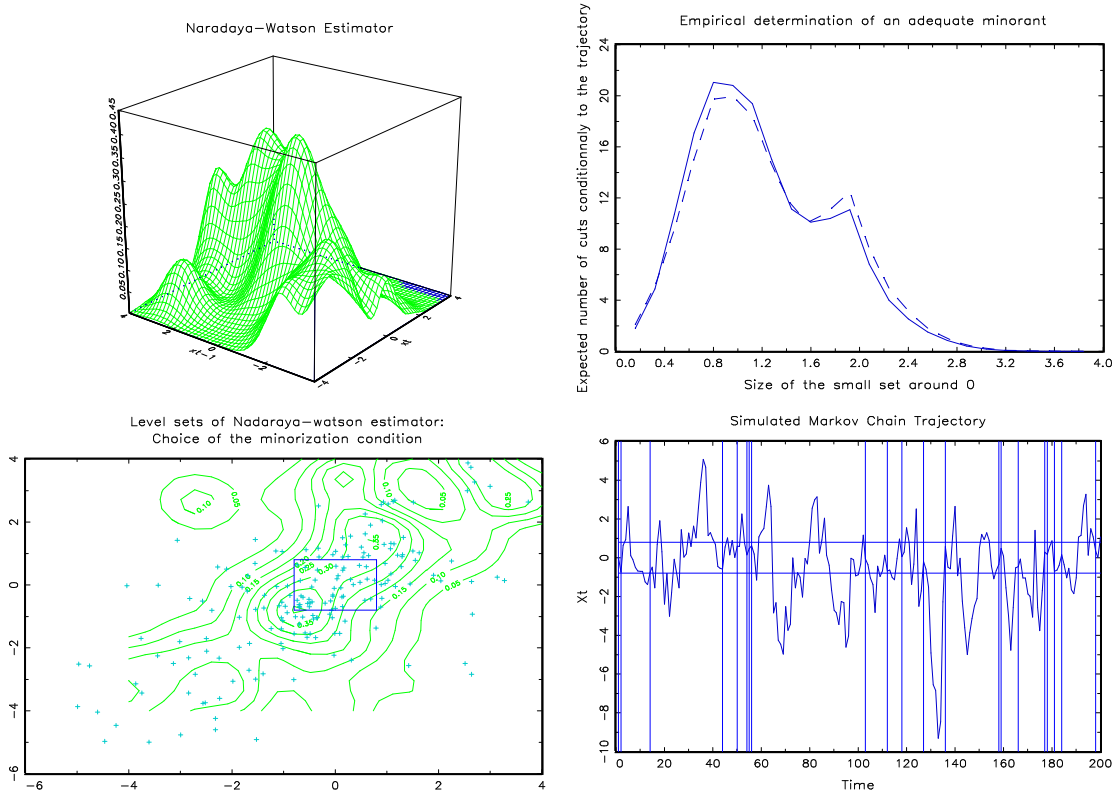


Figure 4: Splitting a Smooth Exponential Threshold Arch time-series, with $n = 200$, $\alpha_1 = 0.60$, $\alpha_2 = 0.45$, $\beta = 0.35$ and $\sigma^2 = 1$. Left upper side : estimator of the transition density. Left down side : visit of the chain to the small set $[-\varepsilon, \varepsilon]^2$ and the level sets of the transition density estimator : the optimal small set should contain a lot of points in a region with high density. Right upper side: number of regenerations according to the size $\varepsilon$ of the small set, optimal for $\varepsilon_{opt} = 0.8$. Right down side : splitting (vertical bars) of the original time series, with horizontal bars corresponding to the optimal small set.

**A3** (Regenerative case)

$$\begin{array}{lll} \mathcal{H}(\kappa) & : & \mathbb{E}_A[\tau_A^\kappa] < \infty, \\ \mathcal{H}(\nu,\kappa) & : & \mathbb{E}_\nu[\tau_A^\kappa] < \infty. \end{array}$$

and their analog versions in the non regenerative case

**A4** (General Harris recurrent case)

$$\begin{array}{lll} \tilde{\mathcal{H}}(\kappa) & : & \sup_{x\in S} \mathbb{E}_x[\tau_S^\kappa] < \infty, \\ \tilde{\mathcal{H}}(\nu,\kappa) & : & \sup_{x\in S} \mathbb{E}_x[\tau_S^\kappa] < \infty. \end{array}$$

# 3 Preliminary results

Here we begin by briefly recalling the connection between the (pseudo-) regeneration properties of a Harris chain $X$ and the extremal behavior of sequences of type $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$, firstly pointed out in the seminal contribution of Rootzén (1988) (see also Asmussen (1998b) and Hansen and Jensen (2005)).

## 3.1 Cycle submaxima for regenerative Markov chains.

We first consider the case when $X$ possesses a known accessible atom $A$. In the following we denote $\alpha = \mathbb{E}_A[\tau_A]$. For $j \geq 1$, define the *submaximum* over the $j$-th cycle of the sample path:

$$\zeta_j(f) = \max_{1+\tau_A(j)\leq i\leq \tau_A(j+1)} f(X_i). \tag{7}$$

In the following $l_n = \sum_{i=1}^n \mathbb{I}\{X_i \in A\}$ denotes the number of visits of $X$ to the regeneration set $A$ until time $n$. $\zeta_0(f) = \max_{1\leq i\leq \tau_A} f(X_i)$ denotes the maximum over the first cycle (starting from an initial distribution $\nu$). Because of the "initialization" phase, its distribution is different from the others and essentially depends on $\nu$. $\zeta_{l_n}^{(n)}(f) = \max_{1+\tau_A(l_n)\leq i\leq n} f(X_i)$ denotes the maximum over the last non-regenerative data block (meaning by that it may be an incomplete block, since we may not observe the return to the the atom A) with the usual convention that maximum over an empty set equals to $-\infty$.

With these definitions, it is easy to understand that the maximum value $M_n(f) = \max_{1\leq i\leq n} f(X_i)$, taken by the sequence $f(X)$ over a trajectory of length $n$, may be naturally expressed in terms of *submaxima* over cycles

$$M_n(f) = \max\{\zeta_0(f), \max_{1\leq j\leq l_n-1} \zeta_j(f), \zeta_{l_n}^{(n)}(f)\}. \tag{8}$$

By the strong Markov property and independence of the blocks, the $\zeta_j(f)$'s are i.i.d. random variables with common distribution function (df) $G_f(x) = \mathbb{P}_A(\max_{1\leq i\leq \tau_A} f(X_i) \leq x)$. Moreover, by Harris recurrence, the number of blocks is of order $l_n \sim n/\alpha$ $\mathbb{P}_\nu$-almost

surely as $n \to \infty$. Thus, $M_n(f)$ behaves like the maximum of $n/\alpha$ i.i.d. rv's. The following result established in Rootzén (1988) shows that the limiting distribution of the sample maximum of $X$ is entirely determined by the tail behavior of the df $G_f$ and relies on this crucial asymptotic independence of the blocks.

**Proposition 1** *(Rootzén, 1988)*
*Let $\alpha = \mathbb{E}_A[\tau_A]$ be the mean return time to the atom $A$. Under the assumption that the first block does not affect the extremal behavior, that is to say that*

$$\mathbb{P}_\nu(\zeta_0(f) > \max_{1 \le k \le l} \zeta_k(f)) \to 0 \text{ as } l \to \infty, \tag{9}$$

*we have then*

$$\sup_{x \in \mathbb{R}} |\mathbb{P}_\nu(M_n(f) \le x) - G_f(x)^{n/\alpha}| \to 0 \text{ as } n \to \infty. \tag{10}$$

In the terminology of O'Brien (see O'Brien (1974, 1987)), $G_f(x)^{1/\alpha}$ may be seen as a so called "phantom distribution", that is an artificial distribution which gives the same distribution for the maximum as in the i.i.d. case. Indeed the preceding theorem shows that the distribution of the maximum behaves exactly as if the observations were independent with distribution $G_f(x)^{1/\alpha}$. As a consequence, the limiting behavior of the maximum in this dependent setting may be simply retrieved by using the famous Fischer-Typett-Gnedenko theorem (obtained in the i.i.d. case), with the marginal distribution replaced by the phantom distribution $G_f(x)^{1/\alpha}$. Then, the asymptotic behavior of the sample maximum is entirely determined by the tail properties of the df $G_f(dx)$. In particular, the limiting distribution of $M_n(f)$ (for a suitable normalization) is the generalized extreme value distribution function $H_\xi(dx)$ with parameter $\xi \in \mathbb{R}$ , given by

$$H_\xi(x) = \begin{cases} \exp(-x^{-1/\xi})\mathbb{I}\{x > 0\}, & \text{when } \xi > 0. \\ \exp(-\exp(-x)), & \text{when } \xi = 0. \\ \exp(-(-x)^{-1/\xi})\mathbb{I}\{x < 0\}, & \text{if } \xi < 0. \end{cases}$$

In the following $\xi$ will be referred as extreme value index. When $\xi > 0$, we will also call it the tail index, corresponding to a Pareto like distribution. The smaller $\xi$, the heavier the tail is.

**Remark 1** *To explain the link between the transition and the behavior of the submaximum, consider the case where A is reduced to a point (which will be the case in our applications). Here to simplify $A = \{0\}$ and positive r.v.'s $X_k, k \in \mathbb{N}$, it is easy to see that*

$$G_f(x) = \mathbb{P}_A(\max_{1 \le i \le \tau_A} X_i \le x) = \sum_{k=1}^{\infty} a_k$$
$$a_k = \mathbb{P}_A(\max_{1 \le i \le \tau_A} X_i \le x, \tau_A = k)$$

*but for* $k \geq 2$

$$\begin{aligned} a_k &= P_A(X_i \leq x, X_i > 0,\ i = 1, \ldots k-1,\ ,\ X_k = 0) \\ &= P_A(0 < X_1 \leq x | X_0 = 0) \prod_{i=2}^{k-1} P_A(0 < X_i \leq x | 0 < X_{i-1} < x) P_A(X_k = 0 | 0 < X_{k-1} \leq x) \\ &= \Pi(0, ]0, x]) \Pi(]0, x], ]0, x])^{k-2} \Pi(]0, x], 0) \end{aligned}$$

*so that*

$$G_f(x) = \Pi(0, 0) + \frac{\Pi(0, ]0, x]) \Pi(]0, x], 0)}{1 - \Pi(]0, x], ]0, x])}.$$

*Thus, it follows that the tail of* $G_f$ *essentially depends on the behavior of* $\Pi(,)$ *for large values of* $x$*. The invariant measure depends itself on this quantity.*

In the following, we assume that $G_f$ belongs to the maximum domain of attraction $H_\xi(x)$ say $MDA(H_\xi)$ (refer to Resnick (1987) for basics in extreme value theory). Then, there exist some sequences $a_n$ and $c_n$ such that $G_f(a_n x + c_n)^n \to H_\xi(x)$ as $n \to \infty$ and we have $\mathbb{P}_\nu(M_n(f) \leq a'_n x + c_n) \to H_\xi(x)$ as $n \to \infty$, with $a'_n = a_n / \alpha^\xi$.

### 3.1.1 Estimation of the cycle submaximum cumulative distribution function

In the atomic case, the cdf $G_f$ of the cycle submaxima, $\zeta_j(f)$ with $j \geq 1$, may be naturally estimated by the empirical counterpart df from the observation of a random number $l_n - 1$ of complete regenerative cycles, namely

$$G_{f,n}(x) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} \mathbb{I}\{\zeta_j(f) \leq x\}, \tag{11}$$

with $G_{f,n} \equiv 0$ by convention when $l_n \leq 1$. Notice that the first and the last (non regenerative blocks) are dropped in this estimator. As a straightforward consequence of Glivenko-Cantelli's theorem for i.i.d. data, we have that

$$\Delta_n = \sup_{x \in \mathbb{R}} |G_{f,n}(x) - G_f(x)| \to 0,\ \mathbb{P}_\nu\text{-almost surely.} \tag{12}$$

Furthermore, by the LIL, we also have $\Delta_n = O(\sqrt{\log\log(n)/n})$ a.s.

### 3.1.2 Estimation of submaxima in the pseudo-regenerative case

Cycles submaxima of the split chain are generally not observable in the general Harris case, since Nummeling extension depends on the true underlying transition probability.

However, our regeneration-based statistical procedures may be directly applied to the submaxima over the approximate regeneration cycles. Define the pseudo-regenerative block maxima by

$$\hat{\zeta}_j(f) = \max_{1+\hat{\tau}_{A_S}(j)\leq i\leq \hat{\tau}_{A_S}(j+1)} f(X_i), \tag{13}$$

for $i = 1, \ldots, \hat{l}_n - 1$. The empirical df counterpart is now given by

$$\hat{G}_{f,n}(x) = \frac{1}{\hat{l}_n - 1} \sum_{j=1}^{\hat{l}_n - 1} \mathbb{I}\{\hat{\zeta}_j(f) \leq x\}, \tag{14}$$

with, by convention, $\hat{G}_{f,n} \equiv 0$ if $\hat{l}_n \leq 1$. As shown by the next theorem, using the approximate cycle submaxima instead of the 'true' ones does not affect the convergence, under assumption A1. Treading in the steps of Bertail and Clémençon (2004a), the proof essentially relies on a *coupling argument*.

**Theorem 2** *Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be a measurable function. Suppose that conditions (3), **A1** and **A2** are fulfilled by the chain $X$. Assume further that $\mathcal{R}_n(\hat{\pi}_n, \pi) \to 0$ as $n \to \infty$. Then, $\hat{G}_{f,n}(x)$ is a consistent estimator of $G_f(x) = \mathbb{P}_{A_S}(\max_{1\leq i\leq \tau_{A_S}} f(X_i) \leq x)$, uniformly over $\mathbb{R}$, as $n \to \infty$,*

$$\widehat{\Delta}_n = \sup_{x\in\mathbb{R}} |\hat{G}_{f,n}(x) - G_f(x)| = O_{\mathbb{P}_\nu}(\mathcal{R}_n(\hat{\pi}_n, \pi)^{1/2}). \tag{15}$$

For smooth Markov chains with smooth $C^\infty$ transition kernel density, the rate of convergence of $\widehat{\Delta}_n$ will be close to $n^{-1/2}$. Under standard Hölder constraints of order $s$, the typical rate for the MSE (6) is of order $n^{-s/(s+1)}$ so that $\widehat{\Delta}_n = O_{\mathbb{P}_\nu}(n^{-s/(2(s+1))})$.

# 4 Regeneration-based statistical methods for extremal events

The core of this paragraph is to show that, in the regenerative setup, consistent statistical procedures for extremal events may be derived from the application of standard inference methods introduced in the i.i.d. setting.

In the case when assumption (9) holds, one may straightforwardly derive from (10) estimates of $H^{(b_n)}(x) = \mathbb{P}_\nu(M_{b_n}(f) \leq x)$ as $n \to \infty$ and $b_n \to \infty$ based on the observation of (a random number of) submaxima $\zeta_j(f)$ over a sample path of length $n$, as proposed in Glynn and Zeevi (2000). Because of the estimation step, we will require that $\frac{b_n}{n} \to 0$. Indeed, if we want to obtain convergent estimators of the distribution of the maximum, we need to subsample the size of the maximum to ensure that the empirical estimation procedure does not alter the limiting distribution. For this, put

$$H_{n,\,l}(x) = (G_{f,n}(x))^l, \tag{16}$$

with $l \geq 1$. The next limit result establishes the asymptotic validity of estimator (16) for an adequate choice of $l$ depending both on the number of regenerations and of the size $b_n$, extending this way Proposition 3.6 of Glynn and Zeevi (2000). If computations are carried out with the pseudo-regeneration cycles, under some additional technical assumptions taking into account the order of the approximation of the transition kernel, the procedure remains consistent. In this case, one would simply consider estimates of the form $\hat{H}_{n,\,l}(x) = (\hat{G}_{f,n}(x))^l$. The following theorem is a simple adaption of a theorem given in Bertail et al. (2009).

**Proposition 3** *(i) (Regenerative case) Suppose that assumption (9) holds. Assume that $b_n \to \infty$ is chosen in such a way that $b_n = o(\sqrt{n/\log\log n})$. Let $(u_n)_{n\in\mathbb{N}}$ be a (deterministic) sequence of real numbers such that $b_n(1 - G_f(u_n))/\alpha \to \eta < \infty$ as $n \to \infty$. Then, we have*

$$H^{(b_n)}(u_n) \to \exp(-\eta) \text{ as } n \to \infty. \tag{17}$$

*In the regenerative setup, suppose furthermore that $\mathcal{H}(\nu, 1)$ is fulfilled. If we choose $l = l(b_n) \sim \frac{b_n}{\alpha}$ as $n \to \infty$. Then,*

$$H_{n,\,l(b_n)}(u_n)/H^{(b_n)}(u_n) \to 1 \tag{18}$$

*in $\mathbb{P}_\nu$- probability, as $n \to \infty$.*

*(ii) (General Harris recurrent case), suppose that* **A1**, **A2** *and $\tilde{\mathcal{H}}(\nu, 1)$ hold and $\mathcal{R}_N(\hat{\pi}_n, \pi) = O(n^{-1+\epsilon})$, as $n \to \infty$ for some $\epsilon \in ]0, 1[$. If $l$ is chosen so that, as $b_n \to \infty$, $l \sim \hat{l}_{b_n}$ and $b_n = o(n^{(1-\epsilon)/2})$ , then*

$$\hat{H}_{n,\,l(n)}(u_n)/H^{(b_n)}(u_n) \to 1 \text{ in } \mathbb{P}_\nu\text{- probability, as } n \to \infty. \tag{19}$$

*(iii) The same results hold if the deterministic threshold is replaced by an estimator based on the empirical distribution for instance $u_n = G_{f,n}^{-1}(1 - \eta\alpha/b_n)$.*

This result indicates that, in the most favorable case, we can recover the behavior of the maximum only over $b_n$ observations with $b_n$ much smaller than $n$. However, it is still possible to estimate the tail behavior of $M_n(f)$, by extrapolation techniques (as it is done for instance in Bertail et al. (2004)). If in addition, one assumes that $G_f$ belongs to some specific domain of attraction $MDA(H_\xi)$, for instance to the Fréchet domain with $\xi > 0$ , it is possible to use classical inference procedures (refer to § 6.4 in Embrechts et al. (1997) for instance) based on the submaxima $\zeta_1(f), \ldots, \zeta_{l_n-1}(f)$ or the estimated submaxima over pseudo-cycles to estimate the shape parameter $\xi$, as well as the norming constants $a_n$ and $c_n$.

## 5 The extremal index

The problem of estimating the extremal index of some functionals of this quantity has been the subject of many researches in the strong mixing framework (see for instanceHsing (1993), Ferro and Segers (2003) and more recently Robert (2009), Robert et al. (2009)). However, we will show that in a Markov chain setting, the estimators are much more simpler to study. Recall that $\alpha = \mathbb{E}_A[\tau_A]$ is the mean return to the atom $A$. In the following, when the regenerative chain $X$ is positive recurrent, we denote $F_\mu(x) = \alpha^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) \leq x\}]$, the empirical distribution function of the limiting stationary measure $\mu$ given by (1). It has been shown (see Leadbetter and Rootzén (1988) for instance) that there exists some index $\theta \in [0,1]$, called the *extremal index* of the sequence $\{f(X_n)\}_{n\in\mathbb{N}}$ , such that

$$\mathbb{P}_\mu(M_n(f) \leq u_n) \underset{n\to\infty}{\sim} F_\mu(u_n)^{n\theta}, \tag{20}$$

for any sequence $u_n = u_n(\eta)$ such that $n(1-F_\mu(u_n)) \to \eta$. Once again, $F_\mu(.)^\theta$ may be seen as an another phantom distribution. The inverse of the extremal index measures the clustering tendency of high threshold exceedances and how the extreme values cluster together. It is a very important parameter to estimate in risk theory, since it indicates somehow, how many times (in mean) an extremal event will reproduce, due to the dependency structure of the data.

As notice in Rootzén (1988), because of the non-unicity of the phantom distribution, it is easy to see from Proposition 1 and (20) that

$$\theta = \lim_{n\to\infty} \frac{\log(G_f(u_n))/\alpha}{\log(F_\mu(u_n))} \tag{21}$$

$$= \lim_{n\to\infty} \frac{\log(1-\overline{G_f}(u_n))/\alpha}{\log(1-\overline{F_\mu}(u_n))} \tag{22}$$

$$= \lim_{n\to\infty} \frac{\overline{G_f}(u_n))}{\alpha\overline{F_\mu}(u_n)}. \tag{23}$$

The last equality following by a simple Taylor expansion. In the i.i.d. setup, by taking the whole state space as an atom ($A = \mathcal{X}$, so that $\tau_A = \alpha \equiv 1,\ G_f = F_\mu$), one immediately finds that $\theta = 1$. In the dependent case, the index $\theta$ may be interpreted as the proportionality constant between the probability of exceeding a sufficiently high threshold within a regenerative cycle and the mean time spent above the latter between consecutive regeneration times.

It is also important to notice that Proposition 1 combined with (20) also entail that, for all $\xi$ in $\mathbb{R}$, $G_f$ and $F_\mu$ belong to the same domain of attraction (when one of them is in a domain attraction of the maximum). Their tail behavior only differs from the slowly varying functions appearing in the tail behavior. We recall that a slowly varying function is a function $L$ such that $L(tx)/L(x) \to 1$ as $x \to \infty$ for any $t > 0$ . For instance $\log$, iterated logarithm, $1/\log,\ 1 + 1/x^\beta,\ \beta > 0$ are slowly varying functions.

Suppose that $G_f$ and $F_\mu$ belong to the Fréchet domain of attraction, then it is known (*cf* Theorem 8.13.2 in Bingham et al. (1987)) that there exist $\xi > 0$ and two slowly varying functions $L_1(x)$ and $L_2(x)$ such that $\overline{G_f}(x) = L_1(x) \cdot x^{-\frac{1}{\xi}}$ and $\overline{F_\mu}(x) = L_2(x) \cdot x^{-\frac{1}{\xi}}$. In this setup, the extremal index is thus simply given by the limiting behavior of

$$\theta(u) = \frac{L_1(u)}{\alpha L_2(u)}.$$

However, estimating slowly varying functions is a difficult task, which requires a lot of data (see Bertail et al. (2004)). Some more intuitive empirical estimators of $\theta$ will be proposed below.

In the regenerative case, a simple estimator of $\theta$ is given by the empirical counterpart of expression 23. $F_n(x) = n^{-1} \sum_{1 \leq i \leq n} \mathbb{I}\{f(X_i) \leq x\}$ is a natural a.s. convergent empirical estimate of $F_\mu(x)$. Recalling that $\frac{n}{l_n} \to \alpha$ a.s., define for a given threshold $u$,

$$\theta_n(u) = \frac{\sum_{j=1}^{l_n - 1} \mathbb{I}\{\zeta_j(f) > u\}}{\sum_{i=1}^{n} \mathbb{I}\{f(X_i) > u\}}, \tag{24}$$

with the convention that $\theta_n(u) = 0$ if $M_n(f) < u$. For general Harris chains, the empirical counterpart of 23 computed from the approximate regeneration blocks is now given by

$$\hat{\theta}_n(u) = \frac{\sum_{j=1}^{\hat{l}_n - 1} \mathbb{I}\{\hat{\zeta}_j(f) > u\}}{\sum_{i=1}^{n} \mathbb{I}\{f(X_i) > u\}}, \tag{25}$$

with $\hat{\theta}_n(u) = 0$ by convention when $M_n(f) < u$. The following result has been recently proved in Bertail et al. (2013). Other estimators based on fixed length blocks in the framework of strong mixing processes are given in Robert (2009) and Robert et al. (2009).

**Proposition 4** *Let* $(r_n)_{n \in \mathbb{N}}$ *be increasing to infinity in a way that* $r_n = o(\sqrt{n / \log\log n})$ *as* $n \to \infty$. *And consider* $(v_n)_{n \in \mathbb{N}}$ *such that* $r_n(1 - G_f(v_n)) \to \eta < \infty$ *as* $n \to \infty$.

*(i) In the regenerative case, suppose that* $\mathcal{H}(v, 1)$ *and* $\mathcal{H}(2)$ *are fulfilled. Then,*

$$\theta_n(v_n) \to \theta \ \mathbb{P}_v\text{-almost surely, as } n \to \infty. \tag{26}$$

*Moreover we have*

$$\sqrt{n/r_n}\,(\theta_n(v_n) - \theta(v_n)) \Rightarrow \mathcal{N}(0, \theta^2/\eta), \text{ as } n \to \infty. \tag{27}$$

*(ii) In the general case, assume that* $A1 - A3$, $\tilde{\mathcal{H}}(v, 1)$ *and* $\tilde{\mathcal{H}}(4)$ *are satisfied. Then,*

$$\hat{\theta}_n(v_n) \to \theta \text{ in } \mathbb{P}_v\text{-probability, as } n \to \infty. \tag{28}$$

*We also have the following central limit theorem :*

$$\sqrt{n/r_n}\left(\widehat{\theta}_n(v_n) - \theta(v_n)\right) \Rightarrow \mathcal{N}(0, \theta^2/\eta) \text{ as } n \to \infty. \tag{29}$$

**Remark 2** *In practice, the levels $v_n$ are unknown since they are defined as upper quantiles of the true underlying sub-maximum distribution. However, if these thresholds are chosen empirically by taking $r_n$ equal to $G_n^{-1}(1-\eta/r_n)$ in the regenerative case or $\widehat{G}_n^{-1}(1-\eta/r_n)$ in the pseudo-regenerative case, then the limiting results remain valid. Because of the condition $r_n = o(\sqrt{n/\log\log n})$ , notice that the best attainable rate with our method is close to $n^{1/4}$ in the regenerative case.*

**Remark 3** *(*The extremal index $\theta$ seen as a limiting conditional probability*) Rootzén (1988) also showed that the extremal index may also be defined as*

$$\theta = \lim_{n\to\infty} \mathbb{P}_A(\max_{2\le i\le \tau_A} f(X_i) \le u_n \mid X_1 > u_n). \tag{30}$$

*for any sequence $u_n$ defined as before. This may be seen as a regenerative version of the so-called runs representation of the extremal index (see Hsing (1993) for instance). This also indicates that the extremal index measures the clustering tendency of high threshold exceedences within regeneration cycles only. An empirical estimator is then simply given by the empirical counterpart based on blocks (or pseudo-blocks)*

$$\theta'_n(u) = \frac{\sum_{j=1}^{l_n-1} \mathbb{I}\{f(X_{1+\tau_A(j)}) > u,\ \max_{2+\tau_A(j)\le i\le \tau_A(j+1)} f(X_i) \le u\}}{\sum_{j=1}^{l_n-1} \mathbb{I}\{f(X_{1+\tau_A(j)}) > u\}}, \tag{31}$$

*for a properly chosen level $u > 0$. The same kind of results may be obtained for this estimator : however, a moderate sample simulation proposed in Bertail et al. (2013) shows that our first estimator outperforms this long-run version as far as coverage of confidence intervals are concerned. This is probably due to the second order properties of these estimators which may be quite difficult to investigate.*

**Remark 4** *Notice that the recentering value in the preceding theorem is $\theta(v_n)$, which converges asymptotically to $\theta$. To control the induced bias (which is a common phenomenon in extreme value parameter estimation), some additional second order conditions are needed. Typically, if one assumes some second-order Hall-type conditions say*

$$L_i(x) = \lim_{y\to\infty} L_i(y) + C_i \cdot x^{-\beta_i} + o(x^{-\beta_i})$$

*as $x \to \infty$ where $C_i < \infty$ and $\beta_i > 0$, $i = 1, 2$, then it can be shown that $\theta(v_n)$ converges to $\theta$ at the rate $v_n^{-\beta}$ with $\beta = \beta_1 \wedge \beta_2$ and $v_n \sim r_n^{1/\beta_1}$. Hence, as soon as $r_n$ is chosen such that $n/r_n^{1+2\beta/\beta_1} \to 0$, we have that $\sqrt{n/r_n}\,(\theta_n(v_n) - \theta) \Rightarrow \mathcal{N}(0, \theta^2/\eta)$ as $n \to \infty$. Similar result holds true in the pseudo-regenerative case. From a practical point of view, to control for the bias a graphical based techniques is generally used for choosing the level, by screening different values of $u_n$ and detecting the region of stability of the estimator (see the simulations below).*

## 6 The regeneration-based Hill estimator

As pointed out in section 5, provided that the extremal index of $\{f(X_n)\}_{n\in\mathbb{N}}$ exists and is strictly positive, the equivalence $G_f \in MDA(H_\xi) \Leftrightarrow F_\mu \in MDA(H_\xi)$ holds true, in particular in the Fréchet case, for $\xi > 0$ namely. Classically, the df $F$ belongs to $MDA(H_\xi)$ if and only if it fulfills the tail regularity condition

$$1 - F(x) = L(x)x^{-1/\xi}, \tag{32}$$

where $L(x)$ is a slowly varying function. Statistical estimation of the tail risk index $\xi > 0$ of a regularly varying df based on i.i.d. data has been the subject of a good deal of attention since the seminal contribution of Hill (1975). Most methods boil down to computing a certain functional of an increasing sequence of upper order statistics, have been proposed for dealing with this estimation problem, just like the celebrated *Hill estimator*, which can be viewed as a conditional maximum likelihood approach. Given i.i.d. observations $Z_1, ..., Z_n$ with common distribution $F(dx)$, the Hill estimator is

$$H^Z_{k,n} = k^{-1} \sum_{i=1}^{k} \log \frac{Z_{(i)}}{Z_{(k+1)}}, \text{ with } 1 \leq k < n, \tag{33}$$

where $Z_{(i)}$ denotes the $i$-th largest order statistic of the data sample $Z^{(n)} = (Z_1, ..., Z_n)$. The asymptotic behavior of this estimator has been extensively investigated when stipulating that $k = k_n$ goes to $\infty$ at a suitable rate. Strong consistency is proved when $k_n = o(n)$ and $\log\log n = o(k_n)$ as $n \to \infty$ in Deheuvels et al. (1988). Its asymptotic normality is established in Goldie (1991) : under further conditions on $L$ (referred to as *second order regular variation*) and $k_n$, we have the convergence in distribution $\sqrt{k_n}(H^Z_{k_n,n} - \xi) \Rightarrow \mathcal{N}(0, \xi^2)$.

The *regeneration-based Hill estimator* based on the observation of the $l_n - 1$ submaxima $\zeta_1(f), ..., \zeta_{l_n-1}(f)$, denoting by $\zeta_{(j)}(f)$ the $j$-th largest submaximum, is naturally defined as

$$\xi_{n,\,k} = k^{-1} \sum_{i=1}^{k} \log \frac{\zeta_{(i)}(f)}{\zeta_{(k+1)}(f)}, \tag{34}$$

with $1 \leq k \leq l_n - 1$ when $l_n > 1$. Observing that, as $n \to \infty$, $l_n \to \infty$ with $\mathbb{P}_\nu$ probability one, limit results holding true for i.i.d. data can be immediately extended to the present setting (*cf* assertion (i) of Proposition 5). In the general Harris situation, an estimator of exactly the same form can be used, except that approximate submaxima are involved in the computation:

$$\widehat{\xi}_{n,\,k} = k^{-1} \sum_{i=1}^{k} \log \frac{\widehat{\zeta}_{(i)}(f)}{\widehat{\zeta}_{(k+1)}(f)}, \tag{35}$$

with $1 \leq k \leq \widehat{l}_n - 1$ when $\widehat{l}_n > 1$. As shown by the next result, the approximation stage does not affect the consistency of the estimator, on the condition that the estimator $\widehat{\pi}$

involved in the procedure is sufficiently accurate. For the purpose of building Gaussian asymptotic confidence intervals in the non-regenerative case, the estimator $\widehat{\xi}_{n,\,k}$ is also considered, still given by Eq. (35).

**Proposition 5** *Suppose that* $F_\mu \in \mathrm{MDA}(H_\xi)$ *with* $\xi > 0$. *Let* $\{k(n)\}$ *be an increasing sequence of integers such that:* $k(n) < n$, $k(n) = o(n)$ *and* $\log\log n = o(k(n))$ *as* $n \to \infty$.

*(i) Then the regeneration-based Hill estimator is strongly consistent*

$$\xi_{n,\,k(l_n)} \to \xi \ \mathbb{P}_\nu\text{- almost surely, as } n \to \infty. \tag{36}$$

*Under the further assumption that* $F_\mu$ *satisfies the Von Mises condition and that* $k(n)$ *is chosen accordingly (cf Goldie (1991)), it is moreover asymptotically normal in the sense that*

$$\sqrt{k(l_n)}(\xi_{n,\,k(l_n)} - \xi) \Rightarrow \mathcal{N}(0,\ \xi^2) \text{ under } \mathbb{P}_\nu, \text{ as } n \to \infty. \tag{37}$$

*(ii) In the general Harris case, if* ***A1*** *and* ***A2*** *are furthermore fulfilled, and* $k = k(n)$ *is chosen accordingly to the Von Mises conditions and is such that* $\mathcal{R}_n(\hat{\pi}_n, \pi)^{1/2} n \log n = o(k(n))$, *then*

$$\widehat{\xi}_{n,k(\hat{l}_n)} \to \xi \text{ in } \mathbb{P}_\nu\text{- probability, as } n \to \infty. \tag{38}$$

*(iii) Under* ***A1*** *and* ***A2***, *if one chooses a sequence* $(m_n)_{n\in\mathbb{N}}$ *of integers increasing to infinity such that* $m_n \mathcal{R}_n(\hat{\pi}_n, \pi)^{1/2}/\sqrt{k(m_n)} \to 0$ *as* $n \to \infty$, *then,*

$$\sqrt{k(\hat{l}_{m_n})}(\widehat{\xi}_{m_n,\,k(\hat{l}_{m_n})} - \xi) \Rightarrow \mathcal{N}(0,\ \xi^2) \text{ under } \mathbb{P}_\nu, \text{ as } n \to \infty. \tag{39}$$

Before showing how the extreme value regeneration-based statistics reviewed in the present article practically perfom on several examples, a few comments are in order.

The tail index estimator (34) is proved strongly consistent under mild conditions in the regenerative setting, whereas only (weak) consistency has been established for the alternative method proposed in Resnick and Stărică (1995) under general strong mixing assumptions. The condition stipulated in assertion (ii) may not be satisfied for some $k(n)$. When the slowly varying function $L(u) = \bar{G}_f(u)/u^{-1/\xi}$ equals for instance $\log(.)$, it can not be fulfilled. Indeed in this case, $k(n)$ should be chosen of order $o(\log(n))$ according to the von Mises conditions. In contrast, choosing a subsampling size $m_n$ such that the conditions stipulated in assertion (iii) holds is always possible.The issue of picking $m_n$ in an optimal fashion in this case remains open.

Given the number $l > 2$ ($l_n$ or $\hat{l}_n$) of (approximate) regeneration times observed within the available data series, the tuning parameter $k \in 1, \ldots, l-1$ can be selected by means of standard methods in the i.i.d. context. A possible solution is to choose $k$ so as to minimize the estimated Mean Square Error

$$\widehat{\gamma}^2_{n,\,k}/k + (a_{n,\,k} - \widehat{\gamma}_{n,\,k})^2,$$

where $\widehat{\gamma}_{n,\,k}$ is a bias corrected version of the Hill estimator. Either the Jackknife method or else an analytical method (see Feuerverger and Hall (1999) or Beirlant et al. (1999)) can be used for this purpose. The randomness of the number of submaxima is the sole difference here.

# 7 Applications to ruin theory and financial time series

As an illustration, we now apply the inference methods described in the previous section to two models from the insurance and the financial fields.

## 7.1 Cramér-Lundberg model with a barrier : example 2

Considering example 2, we apply the preceding results to obtain an approximation of the distribution of the subminimum, of the global minimum (that is the probability of ruin over a given period) and the extremal index. We will not consider here the subexponential case (heavy tail claims) for which it is known that the extremal index is equal to $\theta = 0$, corresponding to infinite clusters of extreme values (see Asmussen (1998b)). Recall that the continuous process of interest is given by

$$X(t) = (u + ct - R(t)) \wedge b$$

and that the embedded chain satisfies

$$X_{n+1} = \inf(X_n + c\Delta T_n - U_{n+1}, b) \text{ with } X_0 = u.$$

Notice that if the barrier $b$ is too high in comparison to the initial reserve $u$, then the chain will regenerate very rarely (unless the price $c$ is very high) and the method will not be useful. But if the barrier is attained at least one time, then the probability of ruin will only depend on $b$ not on $u$. Assume that $\Delta T_n$ is $\gamma(\lambda)$ and the claims are distributed as $\gamma(1/\mu)$ with $EW_n = \mu$ . The safety loading is then given by $\rho = \frac{c}{\lambda\mu} - 1$ and is assumed to be non negative to ensure that the probability of ruin is not equal to 1 a.s.

Using well known results in the case of i.i.d. exponential inputs and outputs, the extremal index is given by $\theta = (1 - \frac{1}{1+\rho})^2 = (1 - \frac{\lambda\mu}{c})^2$. In our simulation we choose $\mu = 0.2$ and $c = 0.3\lambda$ with $\lambda = 10^{-2}$ so that the extremal index is given here by $\theta = 0.111$. We emphasize the fact that we need to observe the times series over a very long period (5000 days) so as to observe enough cycles. The barrier is here at $b = 44$ with a initial reserve $u = 43$.

For $n = 5000$ and if we choose $b_n$ of order $\sqrt{n} \approx 70.7$, with proposition 3 by calculating the quantile of $G_{f,n}$ of order $1+\log(\eta)\alpha/b_n$ for $\eta = 0.95$, we obtain that $\mathrm{Prob}(\min_{1\le i\le\sqrt{n}}(X_i) \le 4.8) = 5\%$. This is an indicator that in the next 70 days there is a rather high probability of being ruined. Inversely, some straightforward inversions (here $G_{f,n}(b) = \Pr(\min_{1\le i\le\tau_A}(X_i) \ge$

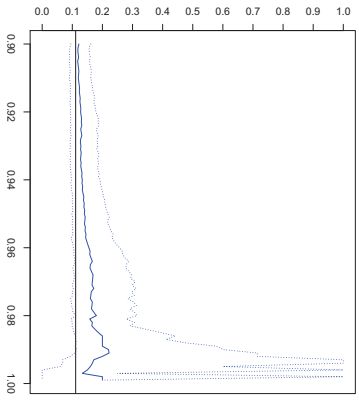


Figure 5: Estimator (continuous line) and bootstrap confidence interval (dotted lines) of the extremal index $\theta(\nu_n)$, for a sequence of high values of the threshold $\nu_n$ (seen as a quantile of the x-coordinate). True value of $\theta = 0.111$.

$0) = P(\min_{1\le i\le \tau_A}(X_i - b) \ge -b) = P(\max_{1\le i\le \tau_A}(b - X_i) \le b))$ shows that the probability of ruin

$$P(\min_{1\le i\le \sqrt{n}}(X_i) \le 0) = 1 - H_{n,\, l(b_n)}(b)$$

and that

$$\simeq 1 - (G_{f,n}(b))^{b_n/\alpha} = 1 - 0.9583 \simeq 4.2$$

This strongly suggests that the dividend barrier and the initial reserve are too low.

As far as the extremal index is concerned, we obtain a rather good estimator of $\theta$ as shown in Figure 5 (see also the simulation results in Bertail et al. (2013) in a slightly different setting (M/M/1 queues)). It represents the value of $\theta(\nu_n)$ for a sequence of high value of the threshold. The stable part of $\theta(\nu_n)$ for a large range of value of levels corresponding to $\nu_n$ is very close to the true value. It should be noticed that when $\nu_n$ is too high, the quantiles of $G_{f,n}$ are badly estimated, resulting in a very bad estimation of $\theta$. Although we did not present in this paper the validity of the regenerative bootstrap (that is bootstrapping regenerative blocks) as shown in Bertail et al. (2013), we represent the corresponding bootstrap confidence intervals on the graphics. It is also interesting to notice that the change in width of the confidence interval is a good indicator in order to choose the adequate level $\nu_n$.

## 7.2 Pseudo regenerative financial time series : extremal index and tail estimation

We will consider the model exhibited in example 3 for a much more longer stretch of observations. Recall that the process is given by the nonlinear autoregressive form

$$X_{t+1} = (\alpha_1 + \alpha_2 e^{-X_t^2})X_t + (1 + \beta X_t^2)^{1/2}\varepsilon_{t+1}\,,\ t = 0, ..., n-1.$$

Indeed the methods used here will only be of interest when $\sqrt{n/\alpha}$ and the number of pseudo regeneration are not too small. The rate of convergence of the Hill estimator is also strongly influenced by the presence of the slowly varying function (here in the distribution of the sub-maxima). Recall that if the slowly varying function belongs to the Hall's family, i.e., is of the form, for some $D \in \mathbb{R}$ and $\beta > 0$,

$$L(x) = 1 + Dx^{-\beta}(1 + o(1)),$$

then the optimal rate of convergence of the Hill estimator is of order at most $n^{\beta/(2\beta+1/\xi)}$ (see Goldie (1991)). Thus, if $\beta$ is small, the rate of convergence of the Hill estimator may be very slow. In practice, we rarely estimate the slowly varying function, but the index is determined graphically by looking at range $k_n$ of extreme values, where the index is quite stable. We also use the bias correction methods (Feuerverger and Hall (1999) or Beirlant et al. (1999)) mentioned before, which greatly improve the stability of the estimators.

We now present in Figure 6 a path of an SETAR-ARCH process, with a large value of $n = 5000$. We choose $\alpha_1 = 0.6$ , $\alpha_2 = 0.45$ and $\beta = 0.35$, which insures stationarity of the process. This process clearly exhibits the features of many log-returns encountered in finance. The optimal small set (among those of the form $[-c, c]$) is given by $[-1.092, 1.092]$, which is quite large, because of the variability of the time-series, with a corresponding value of $\delta = 0.145$.

The true value of $\theta$ (obtained by simulating several very long time series $N = 10^7$) is close to 0.50. This means that maxima clusterize by pair. Figure 11 presents the dependence index estimator, for a range of values of the threshold (the level of the quantile is given on the axe). The estimator is rather unstable for large quantiles, but we clearly identify a zone of stability near the true value of $\theta$. Bootstrap confidence intervals lead to an estimator of $\theta$,between $0.428$ and $0.587$ at the level 95% (which is in the range of the limit theorem given before). The problem of choosing the optimal value of $k_n$ in this case is still an open problem.

Figure 7 presents the regenerative Bootstrap distribution (Bertail and Clémençon (2006b)) of the Hill estimator, with a choice of the optimal fraction $k(\widehat{1}_{m_n})$ obtained by minimizing the mean-square-error. This leads to a confidence interval for the tail (with a error rate of 5%) of the distribution given by $[0.090, 0.345]$. This suggests that for this process that the tail may be quite heavy, since even the moment of order 3 may not exist.

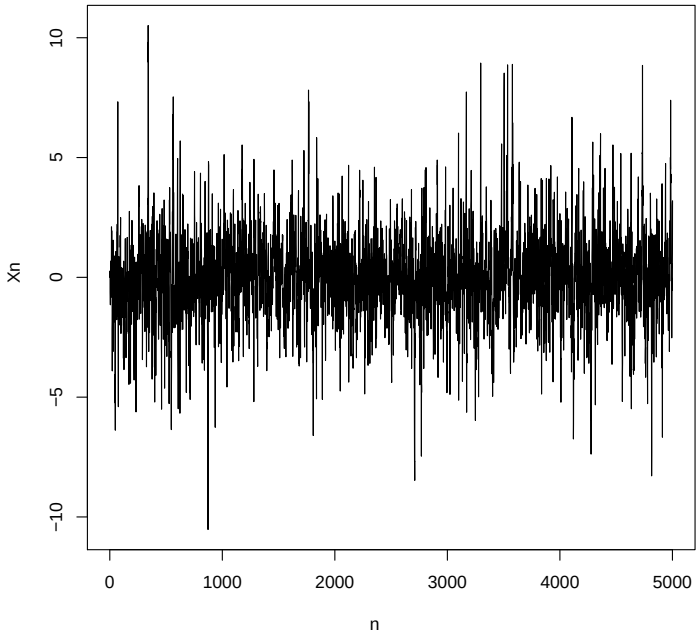


Figure 6: Simulation of the SETAR-ARCH process for $n = 5000$, $\alpha_1 = 0.6$ , $\alpha_2 = 0.45$ and $\beta = 0.35$,exhibiting strong volatility and large excursions.

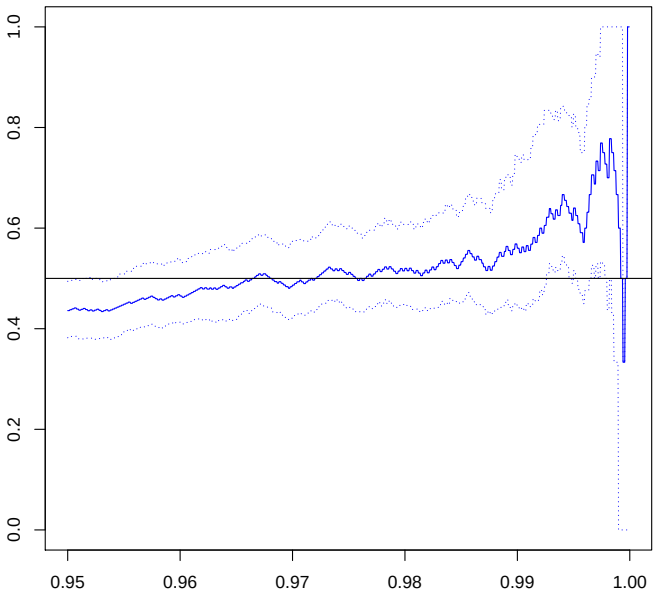


Figure 7: Estimator (continuous line) and confidence intervals (dotted lines) of the extremal index as a function of the quantile level $\theta(\nu_n)$, for a sequence of high values of the threshold $\nu_n$ (seen as a quantile of the x-coordinate). True value of $\theta$ close to 0.5.

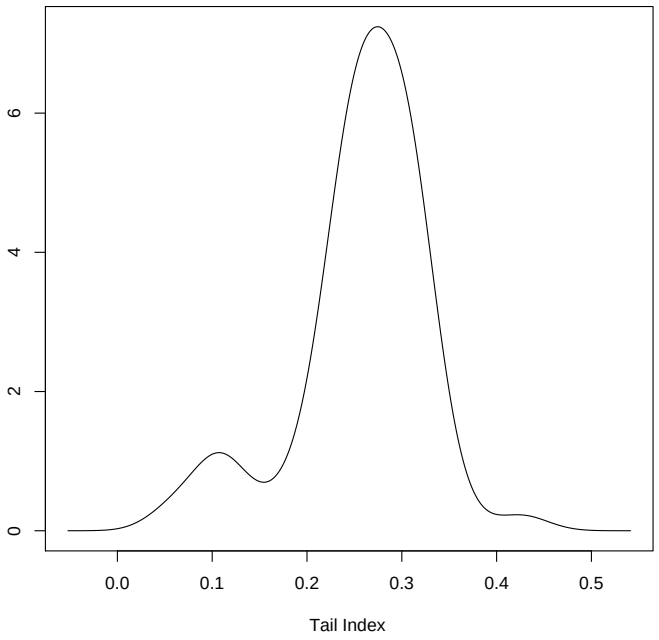


Figure 8: Bootstrap distribution of the pseudo-regenerative Hill estimator (smoothed with a Gaussian kernel), based on $B = 999$ bootstrap replications. Mode around **2.8.**

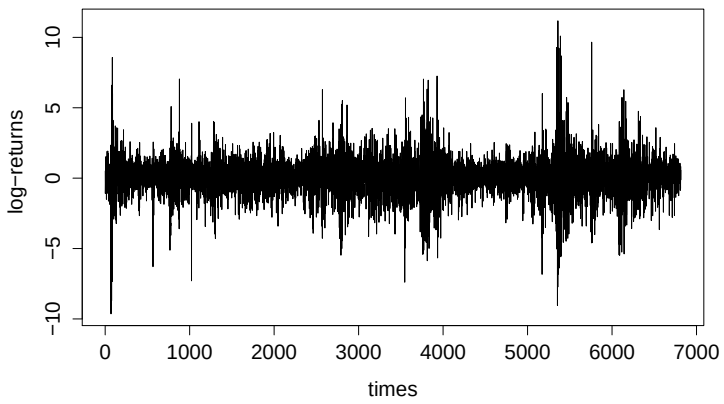


Figure 9: Log-returns of the CAC40, from 10/07/1987 to 06/16/2014.

# 8 An application to the CAC40

We will apply our method to the daily log-return of the CAC40, from 10/07/1987 to 06/16/2014 assuming that this time series follows a Markov chain. Such hypothesis has been tested by several authors (using discretization methods) on other periods, suggesting that the usual stochastic volatility models (Black and Scholes) may not be appropriate in this case (see for instance McQueen and Thorley (1991), Jondeaua and Rockinger (2003), Bhat and Kuma (2010) and Cont (2000)). The log-returns are plotted in figure 8. Notice that the time-series exhibits the same features as the SETAR-ARCH model studied before. However, we do not assume here any specific model for the underlying Markov chain. We observe a lot of regeneration blocks (1567 over 6814 observations) in a small set of optimal size close to $\varepsilon = 0.985$ (a minorizing constant close to $\delta = 0.188$), yielding blocks of mean size 4.35.

We have used two different specifications for the Markov chains, a Markov chain of order 1 and 2. The results are very similar and we thus present only the results for a specification of a Markov chain of order 1. We distinguish between the behavior of the Markov chain for the miniminum and the maximum, for which both the tail index and the extremal index may be different, leading to an asymmetric behavior between gains and losses. The following tables summarizes the main results : we give the value of the estimators of the tail and extremal index as well as Bootstrap confidence intervals (CI) respectively for the minimum and the maximum of the time series.

| Estimators/left and right tail | Min (left tail) | Max (right tail) |
|---|---|---|
| Hill Tail Index estimator | 0.307 | 0.328 |
| Lower CI Tail Index 2.5% | 0.242 | 0.273 |
| Upper CI Tail Index 97.5% | 0.361 | 0.389 |
| Extremal Index estimator | 0.440 | 0.562 |
| Lower CI Extremal Index 2.5% | 0.359 | 0.494 |
| Upper CI Extremal Index 97.5% | 0.520 | 0.614 |

Estimators and confidence intervals for tails and extremal indexes.

The extremal index estimators are very stable when the threshold $u$ is changed, yielding very robust estimators. We emphasize that the tail of the process is very heavy since we are close to the non-existence of the moment of order 4. A simple test based on the Bootstrap confidence intervals allows us to accept the hypothesis that $H_0 : \xi < 1/3$ against $\xi > 1/3$ but reject the the existence of the moment of order four, $H_0 : \xi < 1/4$ against $\xi > 1/4$ for a type I error of 5%.

A striking feature of these results is seemingly some asymmetry in the times series between the minimum and the maximum log returns. In both case, the process has heavy tails with a much more heavy tail for positive log-returns, but with a dynamic which creates smaller clusters of extremum values for maximum (of mean size 1.78) than for minimun (of mean size 2.27). This means that losses may be less strong than gains but may be more persistent. However, a simple test (consisting in comparing the confidence regions) yields that we do not reject the hypothesis of similar tail. This goes in the same direction as Jondeau, Rockinger (2003) on a different period with different method, rather based on the notion of weak dependence.

# 9 Conclusion

Given the ubiquity of the Markov assumption in time-series modeling and applied probability models, we review in this paper several statistical methods, specifically tailored for the Markovian framework with a view towards the extremal behavior of such processes. Precisely, this paper looks at statistical inference for extremal events from the renewal theory angle. We recalled that certain extremal behavior features of Harris Markov chains may be also expressed in terms of *regeneration cycles*, namely data segments between consecutive regeneration times $\tau_1$, $\tau_2$, ... (*i.e.* random times at which the chain forgets its past). Working on this approach, the methodology proposed in this paper boils down to split up the observed sample path into regeneration data blocks (or into data blocks drawn from a distribution approximating the regeneration cycle's distribution, in the general case when regeneration times cannot be observed). Then the analysis boils down to examining the sequence of maxima over the resulting data segments, as if they were i.i.d., via standard statistical methods. In order to illustrate the interest of this technique, we have concentrated on several important inference problems concerning, the question of estimating the sample maximum's tail, the *extremal dependence index* and the *tail index*. However many

other parameters of interest may be investigated in the same manner. The two examples given here, ruin models in insurance and times series exhibiting threshold and/or strong conditional heteroscedasticity clearly show the potentiality of such methods in these fields. An illustration of the estimation methods to the CAC40 shows the potential of the method for real data applications.

## References


Asmussen, S. (1998a). Extreme value theory for queues via cycle maxima. *Extremes*, 1(2):137–168.

Asmussen, S. (1998b). Subexponential asymptotics for stochastic processes: Extremal behavior, stationary distributions and first passage probabilities. *Adv. Appl. Probab.*, 8(2):354–374.

Asmussen, S. (2003). *Applied Probability and Queues*. Springer-Verlag, New York.

Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200.

Bertail, P. and Clémençon, S. (2004a). Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields*, 130(3):388–414.

Bertail, P. and Clémençon, S. (2004b). Note on the regeneration-base bootstrap for atomic Markov chains. *TEST*, 16:109–122.

Bertail, P. and Clémençon, S. (2006a). Regeneration-based statistics for Harris recurrent Markov chains. In Bertail, P., Doukhan, P., and Soulier, P., editors, *Probability and Statistics for dependent data*, volume 187 of *Lecture notes in Statistics*, pages 3–54. Springer.

Bertail, P. and Clémençon, S. (2006b). Regenerative-block bootstrap for Markov chains. *Bernoulli*, 12(4).

Bertail, P. and Clémençon, S. (2007). Approximate regenerative block-bootstrap for Markov chains. *Computational Statistics and Data Analysis*, 52(5):2739–2756.

Bertail, P. and Clémençon, S. (2010). Sharp bounds for the tails of functionals of Markov chains. Available at http://hal.archives-ouvertes.fr/hal-00140591/.

Bertail, P., Clémençon, S., and Tressou, J. (2009). Extreme value statistics for Markov chains via the (pseudo-)regenerative method. *Extremes*, 12:327–360.

Bertail, P., Clémençon, S., and Tressou, J. (2011). A renewal approach to markovian u-statistics. *Mathematical Methods of Statistics*, 20:79–105.

Bertail, P., Clémençon, S., and Tressou, J. (2013). Regenerative block-bootstrap confidence intervals for the tail and extremal indexes. *Electronic Journal of Statistics*, 7:1224–1248.

Bertail, P., Clémençon, S., and Tressou, J. (2014). Bootstrapping robust statistics for markovian data applications to regenerative r- and l-statistics. *submitted J. of Times Series Analysis*.

Bertail, P., Haeffke, C., Politis, D., and White, H. (2004). A subsampling approach to estimating the distribution of diverging statistics with applications to assessing financial market risks. *Journal of Econometrics*, 120:295–326.

Bhat, H. and Kuma, N. (2010). Markov tree options pricing. *Proceedings of the Fourth SIAM Conference on Mathematics for Industry*.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Encyclopedia of Mathematics and its applications. Cambridge Univ Press, Cambridge.

Bolthausen, E. (1980). The Berry-Esseen theorem for functionals of discrete Markov chains. *Z. Wahrsch. Verw. Geb.*, 54(1):59–73.

Clémençon, S. (2001). Moment and probability inequalities for sums of bounded additive functionals of a regular Markov chains via the Nummelin splitting technique. *Statistics and Probability Letters*, 55:227–238.

Cont, R. (2000). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, pages 223–236.

Deheuvels, P., Häusler, E., and Mason, D. (1988). Almost sure convergence of the Hill estimator. *Math. Proc. Cambridge Philos. Soc.*, 104:371–381.

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

Ferro, C. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Statist. Soc.*, 65(2):545–556.

Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto Distribution. *Ann. Statist.*, 27:760–781.

Glynn, P. and Zeevi, A. (2000). Estimating tail probabilities in queues via extremal statistics. In McDonald, D. and Turner, S., editors, *Analysis of Communication Networks: Call Centres, Traffic, and Performance*, pages 135–158, Providence, Rhode Island.

Goldie, C. (1991). Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.*, 1:126–166.

Haiman, G., Kiki, M., and Puri, M. (1995). Extremes of Markov sequences. *Journal of Statistical Planning Inference*, 45:185–201.

Hansen, N. and Jensen, A. (2005). The extremal behaviour over regenerative cycles for Markov additive processes with heavy tails. *Stoch. Proc. Appl.*, 115:579–591.

Hill, B. (1975). A simple approach to inference about the tail of a distribution. *Ann. Statist.*, 3:1163–1174.

Hsing, T. (1988). On the extreme order statistics for a stationary sequence. *Stoch. Proc. Appl.*, 29(1):155–169.

Hsing, T. (1991). On tail estimation using dependent data. *Ann. Statist.*, 19:1547–1569.

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

Jain, J. and Jamison, B. (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.*, 8:19–40.

Jondeaua, E. and Rockinger, M. (2003). Testing for differences in the tails of stock-market returns. *Journal of Empirical Finance*, pages 559–581.

Karlsen, H. A. and Tjöstheim, D. (2001). Nonparametric estimation in null recurrent time series. *Ann. Statist.*, 29:372–416.

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

Loynes, R. (1965). Extreme values in uniformly mixing stochastic processes. *Ann. Math. Stat.*, 36:993–999.

Malinovskiĭ, V. (1985). On some asymptotic relations and identities for Harris recurrent Markov chains. In *Statistics and Control of Stochastic Processes*, pages 317–336.

Malinovskiĭ, V. (1987). Limit theorems for Harris Markov chains I. *Theory Prob. Appl.*, 31:269–285.

Malinovskiĭ, V. (1989). Limit theorems for Harris Markov chains II. *Theory Prob. Appl.*, 34:252–265.

McQueen, G. and Thorley, S. (1991). Are stock returns predictable? a test using markov chains. *The Journal of Finance*, 46:239–63.

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

Newell, G. (1964). Asymptotic extremes for $\mathfrak{m}$-dependent random variables. *Ann. Math. Stat.*, 35:1322–1325.

O'Brien, G. (1974). The maximum term of uniformly mixing stationary processes. *Z. Wahr. verw. Gebiete*, 30:57–63.

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

Resnick, S. (1987). *Extreme Values, Point Processes and Regular Variation.* Springer-Verlag, New York.

Resnick, S. and Stărică, C. (1995). Consistency of Hill's estimator for dependent data. *J. Appl. Probab.*, 32:139–167.

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

Robert, C., Segers, J., and Ferro, C. (2009). A sliding blocks estimator for the extremal index. *Electronic Journal of Statistics*, 3:993–1020.

Robert, C. Y. (2009). Inference for the limiting cluster size distribution of extreme values. *Ann.Statist.*, 37(1):271–310.

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

Rootzén, H. (2006). Weak convergence of the tail empirical process for dependent sequences. Available at http://www.math.chalmers.se/ rootzen/papers/tail_empirical060816.pdf.

S. Asmussen, K. B. and Höjgaard, B. (2000). Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6:303–322.

Smith, R. (1992). The extremal index for a Markov chain. *J. Appl. Probab.*, 29(4):37–45.

Thorisson, H. (2000). *Coupling Stationarity and Regeneration.* Probability and its applications. Springer.

Tjöstheim, D. (1990). Non-linear time series and markov chains. *Adv. Appl. Probab.*, 22:587–611.