Moments of cluster characteristics of time series

Natalia Markovich

Institute of Control Sciences
of Russian Academy of Sciences
Moscow, Russia

8-11 September, EVT 2013

Outlook

- Objectives
- Problem
- Motivation (Application to real peer-to-peer Skype and IPTV SopCast data)
- Survey
- Theoretical results
- Conclusions
- References

Objectives

- We investigate clusters of extremal events, i.e. conglomerates of exceedances of the process over a threshold.
- Clusters impact the risk for hazardous events like climate catastrophes, huge insurance claims, the loss and delay in telecommunication networks due to overloading
- Definitions of a cluster:
  - a cluster is a block of data with at least one exceedance over a threshold or,
  - clusters are data blocks separated by a fixed number of non-exceedances over the threshold (Beirlant et al. 2004)
- Objectives are to determine characteristics of such clusters and to derive their asymptotic distributions.

Problem

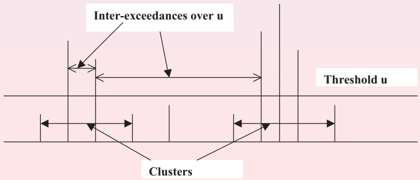
Let $\{R_n : n \geq 1\}$ be a stationary sequence of r.v.s with marginal cdf $F(x)$ and the extremal index θ, $M_n = \max\{R_1, ..., R_n\}$.

The number of inter-arrival times (IATs) between events arising between two consequent exceedances of the process $\{R_n\}_{n\geq 1}$ over **u**

$$T_1(u) = \min\{j \geq 1 : M_{1,j} \leq u, R_{j+1} > u | R_1 > u\}$$

The number of IATs between two consecutive non-exceedances

$$T_2(u) = \min\{j \geq 1 : L_{1,j} > u, R_{j+1} \leq u | R_1 \leq u\}$$



$M_{1,j} = \max\{R_2, ..., R_j\}$,
$M_{1,1} = -\infty$

$L_{1,j} = \min\{R_2, ..., R_j\}$,
$L_{1,1} = +\infty$

Aims

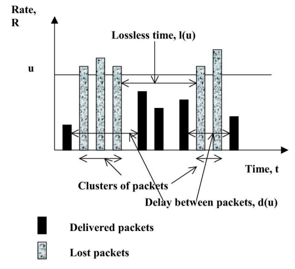
are to find limit distributions of

- the inter-cluster size $T_1(u)$;
- the cluster size $T_2(u)$

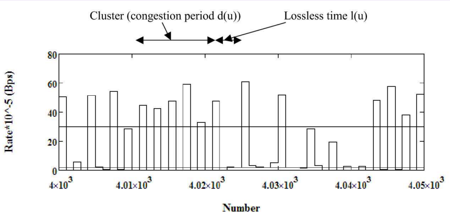
We shall show that

- asymptotically equal distributions of $T_1(u)$ and $T_2(u)$ are geometric-like and corrupted by the extremal index (Theorem 2).
- Asymptotically equal expectations $\mathbf{E}T_1(u)$ and $\mathbf{E}T_2(u)$ are obtained (Lemma 1).

The quality control of a packet flow transmission



P2P SopCast video stream



- Transmission rates $\{R_i = Y_i/X_i\}$, packet lengths $\{Y_i\}$, IATs between the packets $\{X_i\}$, physical capacity u of a channel
- Lossless time of the packet transmission $S_{T_1^*}(u) = \sum_{i=1}^{T_1^*(u)} X_i$;
- Delay between successfully transmitted packets $S_{T_2^*}(u) = \sum_{i=1}^{T_2^*(u)} X_i$

The mixing condition $\Delta^*(u_n)$ (**Ferro & Segers 2003**)

Definition

For real u and integer $1 \le k \le l$, let $\mathfrak{F}_{k,l}(u)$ be σ-filed generated by the events $\{X_i > u\}$, $k \le i \le l$. $\Delta^*(u_n)$ is fulfilled for the mixing coefficients

$$\alpha_{n,q}(u) = \max_{1 \le k \le n-q} \sup |\mathbf{P}(B|A) - \mathbf{P}(B)|,$$

where the supremum is taken over all $A \in \mathfrak{F}_{1,k}(u)$ with $\mathbf{P}(A) > 0$ and $B \in \mathfrak{F}_{k+q,n}(u)$, if there exist positive integers $r_n = o(n)$, $q_n = o(r_n)$ for which $\alpha_{cr_n,q_n}(u_n) \to 0$ as $n \to \infty$ for all $c > 0$.

The next theorem (**Ferro & Segers 2003**) states that $\overline{F}(u)T_1(u)$ is asymptotically exponentially distributed.
This agrees with the result (**Hsing et al 1988**) that the point process of exceedance times has a Poisson process limit.

Theory: the limit distribution of normalized inter-exceedance times, Ferro and Segers (2003)

Theorem

Let the positive integers $\{r_n\}$ and the thresholds $\{u_n\}$, $n \ge 1$, be such that $r_n \to \infty$, $r_n\overline{F}(u_n) \to \tau$ and $\mathbf{P}\{M_{r_n} \le u_n\} \to \exp(-\theta\tau)$ as $n \to \infty$ for some $\tau \in (0,\infty)$ and $\theta \in [0,1]$. If the condition $\Delta^*(u_n)$ is satisfied, then

$$\mathbf{P}\{\overline{F}(u_n)T_1(u_n) > t\} \to \theta\exp(-\theta t) \qquad (1)$$

for $t > 0$ as $n \to \infty$, where $\overline{F}(t) = 1 - F(t)$ is the tail function of $\{R_1\}$.

The result (1) implies that

$$\overline{F}(u_n)T_1(u_n) =^d T_\theta = \begin{cases} \eta, & \text{with probability} \quad \theta, \\ 0, & \text{with probability} \quad 1-\theta, \end{cases}$$

where η is exponentially distributed.

Survey: extremal index

Definition

(Leadbetter et al. 1983, p.53) The stationary sequence $\{R_n\}_{n\geq 1}$ is said to have extremal index $\theta \in [0,1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau, \qquad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta} \qquad \text{hold.} \qquad (2)$$

Interpretation

θ may be interpreted as the reciprocal of the limiting mean cluster size that is the mean number of exceedances per cluster

$$\theta \approx 1/\mathbf{E}T_2(u)$$

(Leadbetter et al. 1983).

Survey: interpretations of the extremal index

- **Leadbetter et al. 1983**: The reciprocal of the limiting mean cluster size that is the mean number of exceedances per cluster

$$\theta = \lim_{n\to\infty} \frac{P\{M_{p_n} > u_n\}}{p_n(1 - F(u_n))}$$

if Leadbetter's condition $D(u_n)$ holds, positive integers $\{p_n\}$ and $\{s_n\}$ such that $p_n = o(n)$ and $s_n = o(p_n)$ and $(n/p_n)\alpha(n, s_n) \to 0$ as $n \to \infty$ (**Beirlant et al. 2004, p. 377**)

$D(u_n)$-condition is satisfied if

for any $A \in \mathcal{I}_{1,l}(u_n)$ and $B \in \mathcal{I}_{l+s,n}(u_n)$, where $\mathcal{I}_{j,l}(u_n)$ is the set of all intersections of the events of the form $\{R_i \le u_n\}$ for $j \le i \le l$, and for some positive integer sequence $\{s_n\}$ such that $s_n = o(n)$, $|P\{(A \bigcap B)\} - P\{A\}P\{B\}| \le \alpha(n, s)$ holds and $\alpha(n, s_n) \to 0$ as $n \to \infty$.

Survey: interpretations of the extremal index

- **O'Brien 1987**:

$$\lim_{n\to\infty} P\{M_{1,p_n} \leq u_n | R_1 > u_n\} = \theta,$$

  where $\{p_n\}$ is an increasing sequence of positive integers, $p_n = o(n)$ as $n \to \infty$, $M_{1,p_n} = \max\{R_2, ..., R_{p_n}\}$.
- **Chernick et al. 1991**: If both $D(u_n)$ and $D'(u_n)$-conditions hold then

$$\lim_{n\to\infty} P\{M_{1,1} \leq u_n | R_1 > u_n\} = \theta = 1 \qquad (3)$$

  holds.
- **Robinson and Tawn 2000**: If $D(u_n)$ and $D''(u_n)$-conditions are satisfied then

$$\lim_{n\to\infty} P\{R_2 \leq u_n | R_1 > u_n\} = \theta$$

$D^{(k)}(u_n)$-condition for any positive integer k

Leadbetter and Nandagopalan 1989; Chernick et al. 1991

states that if the stationary sequence $\{R_t\}$ satisfies the $D(u_n)$-condition with $u_n = a_n x + b_n$ and normalizing sequences $a_n > 0$ and $b_n \in R$ such that for all x there exists $\mu \in R$, $\sigma > 0$ and $\xi \in R$, such that

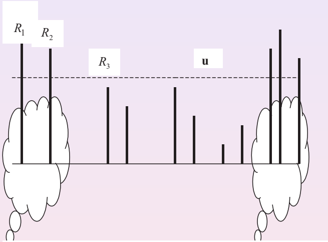
$$n\left(1 - F(a_n x + b_n)\right) \to (1 + \xi(x - \mu)/\sigma)_+^{-1/\xi}, \qquad \text{as} \qquad n \to \infty,$$

holds, where $(x)_+ = \max(x, 0)$, then

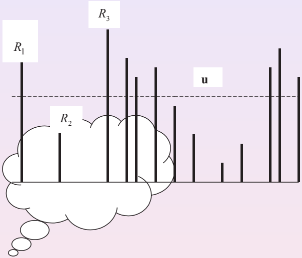
$$\lim_{n\to\infty} n \sum_{j=k+1}^{r_n} P\{R_1 > u_n \geq M_{1,k}, R_j > u_n\} = 0,$$

where $r_n = o(n)$, $s_n = o(n)$, $\alpha(n, s_n) \to 0$, $(n/r_n)\alpha(n, s_n) \to 0$ and $s_n/r_n \to 0$ as $n \to \infty$.
The $D'(u_n)$ and $D''(u_n)$-conditions correspond to $k = 1$ and $k = 2$, respectively.

$D''(u_n)$-condition: no internal drops down



Clusters with consecutive exceedances



Cluster with internal drop

The $D'(u_n)$-condition implies clusters with single exceedances only $\Leftrightarrow$ $\theta = 1$.

Survey: geometric distribution of $T_2(u)$

A geometric distribution has been used as a model of the limiting cluster size[1] distribution π, namely, $\pi(j) = \lim_{n\to\infty} \pi_n(j)$ for $j = 1, 2, ...,$ where

$$\pi_n(j) = P\{N_{r_n}(u_n) = j | N_{r_n}(u_n) > 0\} \qquad \text{for} \qquad j = 1, ..., r_n,$$

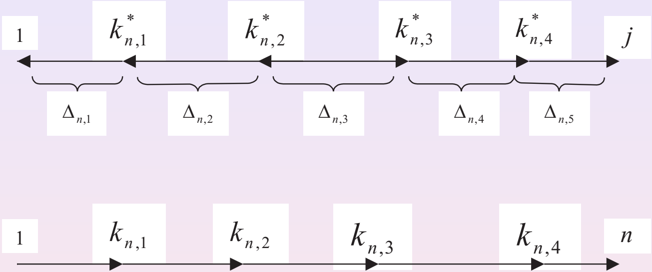
is the cluster size distribution, $r_n = o(n)$, $N_{r_n}(u_n)$ is the number of observations of $\{R_1, ..., R_{r_n}\}$ which exceed $u_n = a_n x + b_n$ that is required to satisfy the $D(u_n)$-condition (Hsing et al. 1988; Robinson and Tawn 2000). If the process R_t satisfies the $D''(u_n)$-condition then in our notations

$$\pi(j) = \lim_{n\to\infty} P\{T_2^*(u_n) - 1 = j\} = (1-\theta)^{j-1}\theta, \qquad j = 1, 2, ...$$

is proposed (Robinson and Tawn 2000, p. 126).

[1] The cluster size means the number of exceedances in a cluster.

Presented results



Partition of the interval $[1, j]$ for a fixed j

$k^*_{n,0} = 1, \quad k^*_{n,5} = j, \quad k^*_{n,i} = [jk_{n,i}/n] + 1, \quad i = \{1, 2\},$

$k^*_{n,3} = j - [jk_{n,4}/n], \quad k^*_{n,4} = j - [jk_{n,3}/n],$

$$\{k_{n,i-1} = o(k_{n,i}), i \in \{2, 3, 4\}\}, \qquad k_{n,4} = o(n). \tag{4}$$

Theorem 2

Let $\{x_{\rho_n}\}$ and $\{x_{\rho_n^*}\}$ be sequences of quantiles of R_1 of the levels $\{1-\rho_n\}$ and $\{1-\rho_n^*\}$,[2] those satisfy the conditions $\lim_{n\to\infty} n(1-F(x_{\rho_n})) = \tau$ and $\lim_{n\to\infty} P\{M_n \le x_{\rho_n}\} = e^{-\tau\theta}$ for each $0<\tau<\infty$ and, $q_n = 1-\rho_n$, $q_n^* = 1-\rho_n^*$, $\rho_n^* = (1-q_n^\theta)^{1/\theta}$. If there are positive integers $\{k_{n,i}\}$, $i=\overline{1,4}$, satisfying (4), $q_{n,i}^* = o(\Delta_{n,i})$, where $\Delta_{n,i} = k_{n,i}^* - k_{n,i-1}^*$, and $p_{n,i}^* = o(q_{n,i}^*)$, $i \in \{1,2,\ldots,5\}$, such that

$$\alpha_n^*(x_{\rho_n}) = \max\{\alpha_{\Delta_{n,i},p_{n,i}^*}(x_{\rho_n}), i \in \{1,2,...,5\}\} = o(1) \tag{5}$$

holds as $n\to\infty$, then it holds for $j\ge 1$

$$\lim_{n\to\infty} P\{T_1(x_{\rho_n}) = j\}/(\rho_n(1-\rho_n)^{(j-1)\theta}) \;=\; 1, \tag{6}$$

$$\lim_{n\to\infty} P\{T_2(x_{\rho_n^*}) = j\}/(q_n^*(1-q_n^*)^{(j-1)\theta}) \;=\; 1. \tag{7}$$

[2] $\overline{F}(x_{\rho_n}) = P\{R_1 > x_{\rho_n}\} = \rho_n$.

Theorem 2. Cont.

If additionally the sequence $\{R_n\}$ satisfies the $D''(x_{\rho_n})$-condition at $[1, k^*_{n,1}+2]$ and $[k^*_{n,4}-1, j+1]$, then it holds for $j \geq 1$

$$\lim_{n\to\infty} P\{T_1(x_{\rho_n}) = j\}/(\rho_n(1-\rho_n)^{(j-1)\theta}) \geq \theta^2, \quad (8)$$

$$\lim_{n\to\infty} P\{T_2(x_{\rho^*_n}) = j\}/(q^*_n(1-q^*_n)^{(j-1)\theta}) \geq \theta^2. \quad (9)$$

Presented results

Mixing condition $\alpha_n^*(x_{\rho_n}) = o(1)$ in Theorem 2 implies

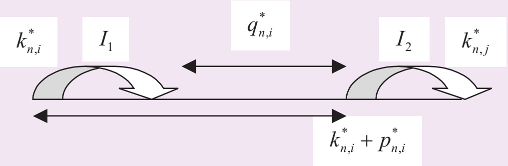
the independence of events $\{R_i > x_{\rho_n}, i \in I\}$ over

$I_1 = [k_{n,i}^*, k_{n,i}^* + p_{n,i}^* - q_{n,i}^*]$ and $I_2 = [k_{n,i}^* + q_{n,i}^*, k_{n,j}^*]$

in the sense that

$$\alpha_{k_{n,j}^* - k_{n,i}^*, q_{n,i}^*}(x_{\rho_n}) \to 0,$$

where $q_{n,i}^* = o(k_{n,j}^* - k_{n,i}^*)$ as $n \to \infty$



One has to check $|P(R_j > x_{\rho_n} | R_1 > x_{\rho_n}) - P(R_j > x_{\rho_n})| = o(1)$

Presented results

Extremal index θ shows the deviation of the asymptotic distribution from the geometric one.

One can rewrite results of Theorem 2 in a geometric form as

$$
\begin{aligned}
c_n P\{T_1(x_{\rho_n}) = j\} &\sim \eta_n(1-\eta_n)^{j-1}, \\
d_n P\{T_2(x_{\rho_n^*}) = j\} &\sim \chi_n(1-\chi_n)^{j-1}, \qquad \text{as} \qquad n \to \infty,
\end{aligned}
$$

using the replacements $(1-\rho_n)^\theta = 1-\eta_n$ and $(1-q_n^*)^\theta = 1-\chi_n$;

$$
c_n = \eta_n / \left(1 - (1-\eta_n)^{1/\theta}\right), \qquad 0 < \eta_n < 1,
$$

$$
d_n = \chi_n / \left(1 - (1-\chi_n)^{1/\theta}\right), \qquad 0 < \chi_n < 1
$$

Presented results

Lemma 1

Let the conditions of Theorem 2 be satisfied, and the sequence $\{R_n\}$ satisfy the mixing condition (5). If for some $\varepsilon > 0$

$$\sup_n E(T_1^{1+\varepsilon}(x_{\rho_n}))/\Lambda_{n,1} < \infty, \qquad \Lambda_{n,1} = 1/(1-(1-\rho_n)^{\theta})^2 \tag{10}$$

holds, then it follows

$$\lim_{n\to\infty} E(T_1(x_{\rho_n}))/(\Lambda_{n,1}\rho_n) = 1, \qquad \text{and if}$$

$$\sup_n E(T_2^{1+\varepsilon}(x_{\rho_n^*}))/\Lambda_{n,2} < \infty, \qquad \Lambda_{n,2} = q_n^*/(1-(1-q_n^*)^{\theta})^2, \tag{11}$$

holds, then it follows

$$\lim_{n\to\infty} E(T_2(x_{\rho_n^*}))/\Lambda_{n,2} = 1. \tag{12}$$

Presented results

Remark

The conditions (10) and (11) provide a uniform convergence of the ranges

$$\sum_{j=1}^{\infty} jP\{T_i(x_{\rho_n}) = j\}/\Lambda_{n,i}, \qquad i \in \{1,2\}$$

by n.

The conditions are fulfilled for geometrically distributed $T_i(x_{\rho_n})$, $i \in \{1,2\}$ with $\theta = 1$ when

$$P\{T_1(x_{\rho_n}) = j\} = \rho_n(1-\rho_n)^{(j-1)}$$

and

$$P\{T_2(x_{\rho_n}) = j\} = q_n(1-q_n)^{(j-1)}.$$

Example 1

ARMAX process $R_t = \max\{\alpha R_{t-1}, (1-\alpha)Z_t\}$, $0 \le \alpha < 1$

- Z_t is Frechet distributed with the df $F(x) = \exp\left(-(1-\alpha)/x\right)$, $x > 0$
- The extremal index of the process is $\theta = 1 - \alpha$

$$P\{T_1(x_\rho) = j\} = \left(q^{1-\theta} - q\right)^2 q^{\theta j}/(q(1-q)), \qquad j = 2, 3, \ldots,$$
$$P\{T_2(x_\rho) = 2\} = q^{\theta}(q^{\theta(1-\theta)} - q^{\theta})$$

$$P\{T_2(x_\rho) = j\} \ge q^{\theta}(q^{(1-\theta)\theta} - q^{\theta})^{j-1}, \qquad j \ge 3$$

where $q = 1 - \rho$, x_ρ is the $(1-\rho)$th quantile of R_1.

- The distributions $T_1(x_\rho)$ and $T_2(x_\rho)$ are geometric ones as $\alpha = 0$ (or $\theta = 1$)
- $E(T_1(x_\rho)) = \frac{q^{1-\theta}}{1-q}$, $\qquad E(T_1(x_\rho))^2 = \frac{q^{1-\theta}(1+q^{\theta})}{(1-q)(1-q^{\theta})}$

Example 2

Moving maxima process $R_t = \max_{i=0,\dots,m}\{\alpha_i \varepsilon_{t-i}\}$, $\alpha_i \geq 0$, $\sum_{i=0}^{m} \alpha_i = 1$

- ε_t are iid unit Frechet distributed with the df $F(x) = \exp(-1/x)$, $x > 0$
- The extremal index of the process is $\theta = \max_i\{\alpha_i\}$
  $\alpha_0 \geq \alpha_1 \geq \dots \geq \alpha_m$

$$P\{T_1(x_\rho) = j\} \geq q^{\theta(j-1)+1-\theta}\left(q^{\alpha_1} - q^{\theta}\right)\frac{1-q^{\theta}}{1-q}, \qquad j = 2, 3, \dots$$

$$P\{T_2(x_\rho) = 2\} = q^{\theta}\left(q^{\alpha_1} - q^{\theta}\right)$$

$$P\{T_2(x_\rho) = j\} \geq q(q^{\alpha_1} - q^{\theta})^{j-1}, \qquad j \geq 3,$$

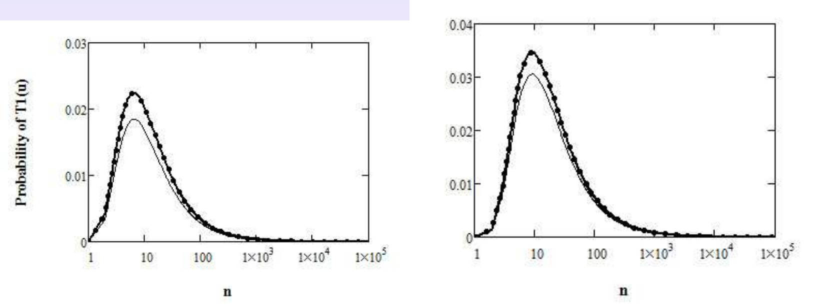
where $q = 1 - \rho$, x_ρ is the $(1-\rho)$th quantile of R_1.

- The distributions $T_1(x_\rho)$ and $T_2(x_\rho)$ are geometric ones as $\theta = \alpha_{(0)} = 1$
- $E(T_1(x_\rho)) \geq \frac{q(q^{\alpha_1-\theta}-1)}{(1-q^{\theta})(1-q)}$, $\qquad E(T_1(x_\rho))^2 \geq \frac{(1+q^{\theta})q(q^{\alpha_1-\theta}-1)}{(1-q^{\theta})^2(1-q)}$

Examples 1 and 2: checking the mixing conditions

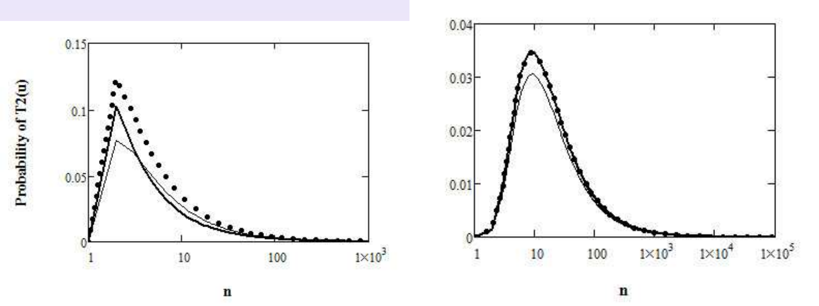
Process parameters	Mixing conditions		
	$\alpha_n^*(x_{\rho_n}) = o(1)$	$D'(u_n)$	$D''(u_n)$
ARMAX - process			
$\theta = 1$	$+$	$+$	$+$
$\theta < 1$	$-$	$-$	$+$
MM - process			
$\theta = 1$	$+$	$+$	$+$
$\alpha_0 \geq \alpha_1 \geq ... \geq \alpha_m$	$+$	$+$	$+$
$\alpha_0 = \theta$			

ARMAX process and geometric approximation



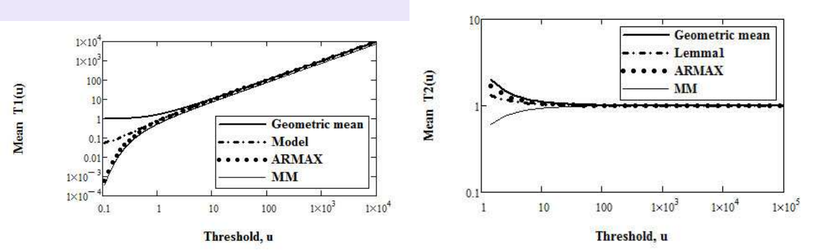
Distribution of $T_1(x_\rho)$ of ARMAX process (dotted line), of MM process (thin solid line) and geometric-like model $\theta^2\rho_n(1-\rho_n)^{(j-1)\theta}$ (solid line) for $j = 10$ against sample size n in logarithmic scale, when $\theta = 0.6$, $\alpha_1 = 0.4$ (a), and $\theta = 0.9$, $\alpha_1 = 0.1$ (b), and q_n is taken equal to $1 - 1/n$

Moving maximum process and geometric approximation



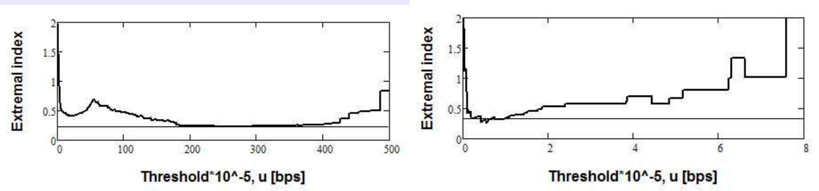
Lower bounds of the distribution $P\{T_2(x_\rho) = j\}$ of ARMAX process (dotted line) and of MM process (thin solid line) and geometric-like model $\theta^2 q_n^* (1 - q_n^*)^{(j-1)\theta}$ (solid line) for $j = 2$ against sample size n in logarithmic scale, when $\theta = 0.6$ and $\alpha_1 = 0.3$ (a), and $\theta = 0.9$ and $\alpha_1 = 0.1$ (b), and q_n is taken equal to $1 - 1/n$

$ET_1(x_\rho)$, $ET_2(x_\rho)$ and geometric approximation



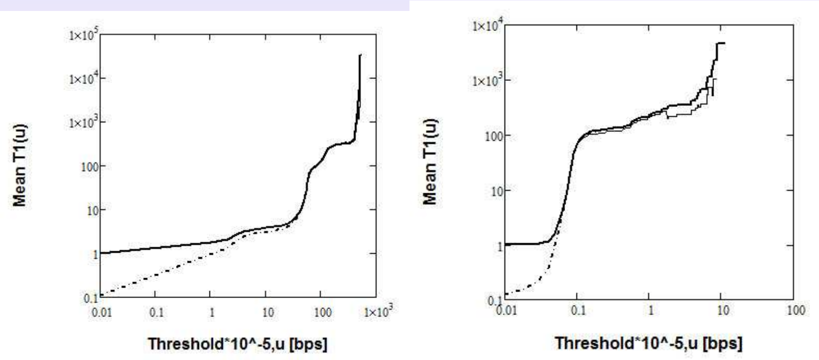
Geometric mean $1/\rho$, $ET_1(x_\rho)$ fitted by $\theta^2\rho/(1-(1-\rho)^\theta)^2$, $ET_1(x_\rho)$ of an ARMAX process and $ET_1(x_\rho)$ of a MM process with extremal index $\theta = 0.2$ and $\alpha_1 = 0.05$ (a); and the geometric mean $1/(1-\rho)$, $ET_2(x_\rho)$ obtained by Lemma 1 and $ET_2(x_\rho)$ of the same processes (b) against threshold $u = x_\rho = -1/ln(1-\rho)$ in logarithmic scale

Skype and SopCast IPTV traffic data



Intervals estimate (solid line) of the extremal index of the transmission rates of the SopCast IPTV (left) and Skype (right) data against the threshold u indicating the capacity of the channel measured in bits per second (bps) and recommended θ-values (thin solid lines) corresponding to the stability intervals of the plots

Skype and SopCast IPTV traffic data



Geometric mean (solid line) and $E(T_1(u))$ fitted by Lemma with $\theta(u) = 0.23$ (dashed line) and by the sample mean $\overline{T_1(u)}$ (thin solid line) obtained by SopCast IPTV data with a sample size equal to $6.553 \cdot 10^4$, against the threshold u (left), and similar curves shown for Skype data with a sample size equal to 4605 and with $\theta(u) = 0.388$ (right)

Reference

- Hsing T., Huesller J, Leadbetter M.R. 1988. On the exceedance point process for a stationary sequence. Probab. Theory Reltd Flds, 78, 97-112.
- Ferro, C.A.T. and Segers, J. 2003. Inference for Clusters of Extreme Values. Journal of the Royal Statistical Society, Series B, 65, 545-556.
- Leadbetter, M.R., Lingren, G., Rootzén, H.: Extremes and Related Properties of Random Sequence and Processes. ch.3, Springer, New York (1983)
- Leadbetter, M.R., Nandagopalan, L.: On exceedance point processes for stationary sequences under mild oscillation restrictions. Lect. Notes Statist., 51, 69-80 (1989)

Reference

- Markovich, N.M., Krieger, U.R. 2010. Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic. In: Performance Handbook Next Generation Internet: Performance Evaluation and Applications, Springer.
- Markovich N.M. 2013. Modeling clusters of extreme values. Extremes, Accepted.
- Markovich N. M. Quality Assessment of the Packet Transport of Peer-to-Peer Video Traffic in High-Speed Networks, Performance Evaluation, 70 (2013) 28-44
- Robinson, M.E., Tawn, J.A.: Extremal analysis of processes sampled at different frequences. Journal of the Royal Statistical Society Series B. 62(1), 117135 (2000)