**Modeling Clusters of Extreme Values in Time Series**

Natalia M. Markovich

Institute of Control Sciences of Russian Academy of Sciences, 117997 Moscow, Russia
nat.markovich@gmail.com, markovic@ipu.rssi.ru

**Abstract**

The study of clusters of extreme values of a time series (exceedances over a sufficiently high threshold) is of fundamental interest in many applied fields including climate research, insurance and telecommunications. Our main result states that limit distributions of cluster and inter-cluster sizes for a stationary sequence under specific mixing conditions have geometric forms. The cluster size implies the number of consecutive exceedances of the time series over a threshold and the inter-cluster size the number of consecutive observations running under the threshold. In Ferro and Segers (2003) the inter-cluster size normalized by the tail function is proved to be exponentially distributed. A geometric model of the limiting cluster size distribution was presented in Robinson and Tawn (2000) without rigorous proof. The inter-cluster size is also geometric distributed if clusters of exceedances are independent, Santos and Fraga Alves( 2012). Recent results of the author will be presented. It is shown that asymptotically equal distributions of both cluster and inter-cluster sizes are geometric like and can be represented by a level of a sufficiently high quantile of the underlying process used as the threshold and a so-called extremal index. The presented models are in a good agreement with cluster and inter-cluster distributions of autoregressive maximum and moving maxima processes and with real telecommunication data concerning packet traffic rates of Skype and Internet television peer-to-peer video applications.

Key Words: Cluster of exceedances over a threshold, distributions of cluster and inter-cluster sizes, extremal index

# 1 Introduction

The subject of our research concerns extremes in time series and more precisely, exceedances of an underlying process over sufficiently high threshold $u$. Due to dependence and heaviness of tails in time series such exceedances build clusters. The latter are separated by observations that run below $u$. Often clustering of exceedances corresponds to consecutive large losses occurring in a short period of time.

Our objectives are to find distributions of the cluster size (i.e. the number of consecutive exceedances over $u$ between two consecutive non-exceedances of the process) and the inter-cluster size (i.e. the number of observations running under $u$ between two consecutive exceedances). The necessity of these distributions arises in many applied fields like seismology, metereology, telecommunications and finance. An example of cluster structure in telecommunication data is shown in Fig. 1.

More precisely, let us consider a stationary sequence of random variables (rvs) $\{R_n\}_{n\geq 1}$ with marginal cumulative distribution function (cdf) $F(x)$ and the extremal index $\theta \in (0,1]$. The latter serves as a dependence measure. Its reciprocal has a simple interpretation as the limiting mean cluster size, Leadbetter et al. (1983). We consider two discrete rvs

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

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

[0]This work was supported in part by the Russian Foundation for Basic Research, grant 13-08-00744 A.

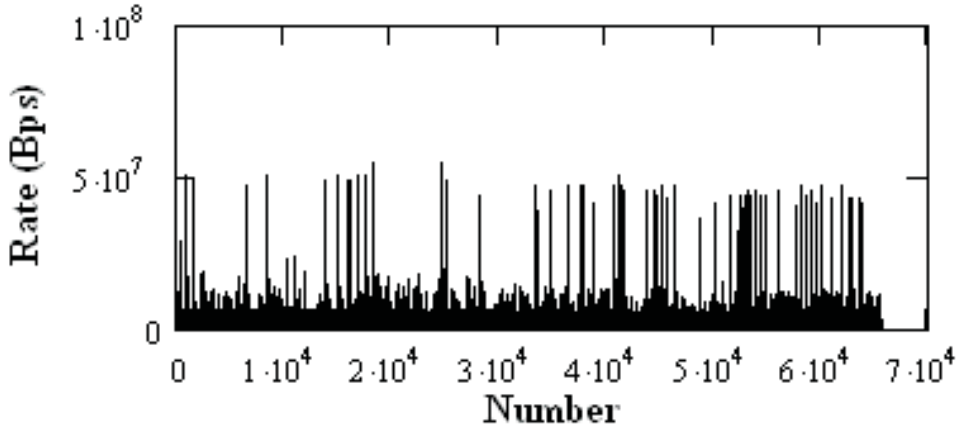


Figure 1: Clusters of exceedances of packet rates $\{R_t\}$ over a high threshold $u$ arising during the transmission of a video packet video stream in a peer-to-peer overlay network of SopCast, Markovich and Krieger (2011).

where $M_{1,j} = \max\{R_2, ..., R_j\}$, $M_{1,1} = -\infty$, $L_{1,j} = \min\{R_2, ..., R_j\}$, $L_{1,1} = +\infty$.
In many applications the cases $T_1(u) = 1$ and $T_2(u) = 1$ are excluded from consideration since they correspond to inter-arrival times between consecutive events $\{R_i\}$.
Distributions of sums $S_{T_i(u)} = \sum_{j=1}^{T_i(u)} X_j$, $i \in \{1,2\}$, where $\{X_j\}$ are inter-arrival times between observations of the process $\{R_n\}$, are probably of larger practical interest. $S_{T_1(u)}$ may be interpreted as the return interval between two consecutive clusters and $S_{T_2(u)}$ as the duration of cluster.
Intuitively it is clear that clusters of exceedances become rare and thus, independent as far as $u$ increases. In Hsing et al. (1988) it is proved that the limit process of the point process of exceedance times is a compound Poisson process. Hence the limit distribution of its return interval is exponential. The distribution of normalized r.v. $\overline{F(u)}T_1(u)$, where $\overline{F(u)} = 1 - F(u)$, is proved to be mixed exponential, Ferro and Segers (2003).
A geometric distribution has been used as a model of the limiting cluster size distribution $\pi$, 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 $u_n = a_n x + b_n$. The latter is required to satisfy Leadbetter's mixing condition[1] $D(u_n)$, Hsing et al. (1988), Robinson and Tawn (2000). If the process $R_t$ satisfies the $D''(u_n)$-condition[2] of Leadbetter and Nandagopalan (1989) then in our

[1] $D(u_n)$ is satisfied if for any $A \in \mathscr{I}_{1,l}(u_n)$ and $B \in \mathscr{I}_{l+s,n}(u_n)$, where $\mathscr{I}_{j,l}(u_n)$ 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 $n \to \infty$.

[2] The $D^{(k)}(u_n)$-condition for any positive integer $k$ 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

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

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

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

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.

notations

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

is proposed in Robinson and Tawn (2000), p. 126 without rigorous proof. Here, $\theta$ is an extremal index.

**Definition 1.** *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*

$$\lim_{n\to\infty} n(1-F(u_n)) = \tau \qquad \textit{and} \tag{2}$$

$$\lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta} \tag{3}$$

*hold, Leadbetter et al. (1983).*

The geometric nature of $T_1(u)$ without attraction of the extremal index is considered in Santos and Fraga Alves( 2012). In this paper $T_1(u)$ is called a 'duration between two consecutive violations' and it is used to test the independence hypothesis.
Recursive estimators of the limiting cluster size probabilities are considered in Robert (2009).
In Markovich (2013a) geometric-like asymptotically equivalent distributions of both $T_1(u)$ and $T_2(u)$ with a probability corrupted by the extremal index $\theta$ are derived. The latter allows us to take into account the dependence in the data.
The derived geometric models allow us to obtain the asymptotically equal means of $T_1(u)$ and $T_2(u)$, Markovich (2013a).

## 2 Results

In the next theorem the following mixing coefficient introduced in Weissman and Novak (1998) is used.

**Definition 2.** *For real $u$ and integers $1 \leq k \leq l$, let $\mathscr{F}_{k,l}(u)$ be the $\sigma$-field generated by the events $\{R_i > u\}$, $k \leq i \leq l$. Define the mixing coefficients $\alpha_{n,q}(u)$,*

$$\alpha_{n,q}(u) = \max_{1\leq k\leq n-q} \sup |P(B|A) - P(B)|, \tag{4}$$

*where the supremum is taken over all $A \in \mathscr{F}_{1,k}(u)$ with $P(A) > 0$ and $B \in \mathscr{F}_{k+q,n}(u)$ and $k$, $q$ are positive integers.*

In Markovich (2013a) quantiles of the underlying process $R_t$ are taken as thresholds $\{u_n\}$. The following result is derived. In order to deal with discrete rvs $T_1(u_n)$ and $T_2(u_n)$ we define the partition of the interval $[1,j]$ for a fixed $j$, namely,

$$\begin{aligned} k^*_{n,0} &= 1, \qquad k^*_{n,5} = j, \qquad k^*_{n,i} = [jk_{n,i}/n] + 1, \qquad i = \{1,2\}, \\ k^*_{n,3} &= j - [jk_{n,4}/n], \qquad k^*_{n,4} = j - [jk_{n,3}/n] \end{aligned} \tag{5}$$

that corresponds to the partition of the interval $[1,n]$

$$\{k_{n,i-1} = o(k_{n,i}), \qquad i \in \{2,3,4\}\}, \qquad k_{n,4} = o(n), \qquad n \to \infty, \tag{6}$$

where $n$ is the sample size.

**Theorem 1.** *(Markovich (2013a)) Let $\{R_n\}_{n\geq 1}$ be a stationary process with the extremal index $\theta$. Let $\{x_{\rho_n}\}$ and $\{x_{\rho_n^*}\}$ be sequences of quantiles of $R_1$ of the levels $\{1-\rho_n\}$ and $\{1-\rho_n^*\}$, respectively,[3] those satisfy the conditions (5) and (3) if $u_n$ is replaced by $x_{\rho_n}$ or by $x_{\rho_n^*}$ and, $q_n = 1-\rho_n$, $q_n^* = 1-\rho_n^*$, $\rho_n^* = (1-q_n^\theta)^{1/\theta}$. Let positive integers $\{k_{n,i}^*\}$, $i=\overline{0,5}$, and $\{k_{n,i}\}$, $i=\overline{1,4}$, be respectively as in (5) and (6), $p_{n,i}^* = o(\Delta_{n,i})$, $\Delta_{n,i} = k_{n,i}^* - k_{n,i-1}^*$, $q_{n,i}^* = o(p_{n,i}^*)$, $i\in\{1,2,...,5\}$ and $\{p_{n,3}^*\}$ be an increasing sequence, such that*

$$\begin{aligned}\alpha_n^*(x_{\rho_n}) &= \max\{\alpha_{k_{n,4}^*,q_{n,1}^*};\alpha_{k_{n,3}^*,q_{n,2}^*};\alpha_{\Delta_{n,3},q_{n,3}^*};\alpha_{j+1-k_{n,2}^*,q_{n,4}^*};\\ &\quad \alpha_{j+1-k_{n,1}^*,q_{n,5}^*};\alpha_{j+1,k_{n,4}^*-k_{n,1}^*}\} = o(1)\end{aligned} \tag{7}$$

*holds as $n\to\infty$, where $\alpha_{n,q} = \alpha_{n,q}(x_{\rho_n})$ is determined by (4), then it holds for $j\geq 2$*

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

$$\lim_{n\to\infty} P\{T_2(x_{\rho_n^*}) = j\}/(q_n^*(1-q_n^*)^{(j-1)\theta}) = 1, \tag{9}$$

*and 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 $j\geq 2$*

$$\lim_{n\to\infty} P\{T_1(x_{\rho_n}) = j\}/(\rho_n(1-\rho_n)^{(j-1)\theta}) \geq \theta^2, \tag{10}$$

$$\lim_{n\to\infty} P\{T_2(x_{\rho_n^*}) = j\}/(q_n^*(1-q_n^*)^{(j-1)\theta}) \geq \theta^2. \tag{11}$$

**Corollary 1.** *(Markovich (2013a)) Under the additional condition $D'(x_{\rho_n})$ (instead of $D''(x_{\rho_n})$-condition), $\theta = 1$ holds. Thus $P\{T_1(x_{\rho_n}) = j\} \sim \rho_n(1-\rho_n)^{(j-1)}$ and $P\{T_2(x_{\rho_n^*}) = j\} \sim q_n^*(1-q_n^*)^{(j-1)}$ as $n\to\infty$ follow.*[4]

**Lemma 1.** *(Markovich (2013a)) If the conditions of Theorem 1 are satisfied, and*

$$\sup_n E(T_2^{1+\varepsilon}(x_{\rho_n^*}))/\Lambda_{n,2} < \infty \tag{12}$$

*holds for some $\varepsilon > 0$, where $\Lambda_{n,2} = q_n^*/(1-(1-q_n^*)^\theta)^2$, and the sequence $\{R_n\}$ satisfies the mixing condition (7) then it holds*

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

**Remark 1.** *Condition (12) provides a uniform convergence the range $\sum_{j=1}^\infty jP\{T_2(x_{\rho_n^*}) = j\}/\Lambda_n$ by n. The condition is fulfilled for geometrically distributed $T_2(x_{\rho_n})$ when $P\{T_2(x_{\rho_n}) = j\} = q_n(1-q_n)^{(j-1)}$. For $0<\varepsilon<1$ we have $E(T_2^{1+\varepsilon}(x_{\rho_n}))/\Lambda_{n,2} < E(T_2^2(x_{\rho_n}))/\Lambda_{n,2} = (2-q_n)/q_n < \infty$ for such n that $q_n$ satisfies (2), i.e. $q_n \sim 1-\tau/n$.*
*Formally, one can prove that*

$$\lim_{n\to\infty} E(T_1(x_{\rho_n}))/\Lambda_{n,1} = 1, \quad \text{where} \quad \Lambda_{n,1} = \rho_n/(1-(1-\rho_n)^\theta)^2$$

*holds if the condition $\sup_n E(T_1^{1+\varepsilon}(x_{\rho_n}))/\Lambda_{n,1} < \infty$ is valid. However, the latter condition seems to be too hard and it is difficult to find an example when this is valid. At the same time, $E(T_1(x_{\rho_n}))/\Lambda_{n,1} = 1$ for any n if $P\{T_1(x_{\rho_n}) = j\} = \rho_n(1-\rho_n)^{(j-1)}$ holds.*

---

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

[4] The symbol $\sim$ means asymptotically equal to or $f(x)\sim g(x) \Leftrightarrow f(x)/g(x)\to 1$ as $x\to a$, $x\in M$ where the functions $f(x)$ and $g(x)$ are defined on some set $M$ and $a$ is a limit point of $M$.

Table 1: Checking the mixing conditions of Theorem 1

| Process parameters | Mixing conditions | | | Condition (12) |
|---|---|---|---|---|
| | $\alpha_n^*(x_{\rho_n}) = o(1)$ | $D'(u_n)$ | $D''(u_n)$ | |
| ARMAX - process | | | | |
| $\theta = 1$ | + | + | + | + |
| $\theta < 1$ | − | − | + | + |
| MM - process | | | | |
| $\theta = 1$ | + | + | + | + |
| $\alpha_0 \geq \alpha_1 \geq ... \geq \alpha_m$ $(\alpha_0 = \theta)$ | + | + | + | + |

The suggested geometric distributions give rise to further modeling of distributions of return intervals (the time intervals between clusters) and durations of clusters widely used in seismology and climatology. In Markovich (2013a) it is shown that the limit tail distribution of the duration of clusters $S_{T_2(u)} = \sum_{i=1}^{T_2(u)} X_i$ that is defined as a sum of a random number of weakly dependent regularly varying inter-arrival times $\{X_i\}$ with tail index $0 < \alpha < 2$ is bounded by the tail of stable distribution. In Markovich (2013b) it is found that the byte loss and the packet delay due to the loss in the clusters highly impacts on the quality and optimization of the packet transmission. The clusters are caused by packets whose transmission rates $\{R_i\}$ exceed the equivalent capacity $u$ of a bufferless channel.
Further development of the author's inferences concerns the score function of the inter-cluster size normalized by the tail function and its relation to the extremal index and the score function of the underlying process, Markovich and Stehlik (2013). The latter can be applied in numerous fields.

## 3 Examples

The ARMAX and MM processes satisfy mixing conditions (7), $D'(x_{\rho_n})$ and $D''(x_{\rho_n})$ (both are in conjunction with $D(x_{\rho_n})$-condition) and hence, the assumptions of Theorem 1 for specific values of their parameters $\alpha$ and $\{\alpha_i\}$.
Let us define the ARMAX process

$$R_t = \max\{\alpha R_{t-1}, (1-\alpha)Z_t\}, \qquad t \in \mathrm{Z}, \tag{13}$$

where $0 \leq \alpha < 1$, $\{Z_t\}$ are iid standard Fréchet distributed rvs with the cdf $F(x) = \exp(-1/x)$, for $x > 0$ and $P\{R_t \leq x\} = \exp(-1/x)$ holds assuming $R_0 = Z_0$. The extremal index $\theta$ of the process is equal to $1-\alpha$, Beirlant et al. (2004).
The $m$th order MM process is determined by formula

$$R_t = \max_{i=0,\ldots,m} \{\alpha_i Z_{t-i}\}, \qquad t \in \mathrm{Z}, \tag{14}$$

where $\{\alpha_i\}$ are constants with $\alpha_i \geq 0$, $\sum_{i=0}^m \alpha_i = 1$, and $Z_t$ are iid standard Fréchet distributed rvs. The distribution of $\{R_t\}_{t\geq 1}$ is also standard Fréchet. The extremal index of the process is determined by $\theta = \max_i\{\alpha_i\}$, Ancona-Navarrete and Tawn (2000). The checking the mixing conditions required in Theorem 1 and the condition (12) for both processes is shown in Table 1.
In Markovich (2013a) the analytical formulae of distributions of $T_1(x_\rho)$ and $T_2(x_\rho)$ and their moments are obtained. The latter were compared with models recommended in Theorem 1.

# 4 Conclusions

The clusters of exceedances of processes over high thresholds are investigated. The asymptoticaly equivalent geometric-like distributions of cluster and inter-cluster sizes are obtained. The expectation of the cluster size is presented. The models recommended in Theorem 1 and Lemma 1 are in good agreement with ARMAX and MM processes as well as with real telecommunication data, Markovich (2013a).

## References

[1] Ancona-Navarrete, M.A., Tawn, J. A. (2000) "A comparison of Methods for Estimating the Extremal Index," *Extremes*, 3:1, 5-38.

[2] Araújo Santos, P., Fraga Alves, M.I. (2012) "A new class of independence tests for interval forecasts evaluation," *Computational Statistics and Data Analysis* 56 (11), 3366-3380.

[3] Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004) *Statistics of Extremes: Theory and Applications*, Wiley, Chichester, West Sussex.

[4] Chernick, M.R., Hsing, T. and McCormick, W.P. (1991) "Calculating the extremal index for a class of stationary sequences," *Advances in Applied Probability*, 23, 835-850.

[5] Ferro, C.A.T., Segers, J. (2003) "Inference for Clusters of Extreme Values," *Journal of the Royal Statistical Society Series B,* 65, 545-556.

[6] Hsing, T., Huesler, J. and Leadbetter, M.R. (1988) "On the exceedance point process for a stationary sequence," *Probability Theory and Related Fields*, 78, 97-112

[7] Leadbetter, M.R., Lingren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequence and Processes*, ch.3, Springer, New York.

[8] Leadbetter, M.R., Nandagopalan, L. (1989) "On exceedance point processes for stationary sequences under mild oscillation restrictions," *Lecture Notes Statistics*, 51, 69-80.

[9] Markovich, N. (2013a) "Modeling clusters of extreme values," under revision.

[10] Markovich N.M. (2013b) "Quality Assessment of the Packet Transport of Peer-to-Peer Video Traffic in High-Speed Networks," *Performance Evaluation*, 70, 28-44.

[11] Markovich, N.M., Krieger, U. R. (2011) "Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic," *In: Kouvatsos, D. (ed.) Next Generation Internet: Performance Evaluation and Applications, LNCS 5233*, Springer, Heidelberg, 37-69.

[12] Markovich, N., Stehlik, M. (2013) "On relationship between score functions and extremal index," accepted for IFAC MIM 2013.

[13] Robert C.Y. (2009) "Inference for the limiting cluster size distribution of extreme values," *The Annals of Statistics*, 37, 1, 271-310.

[14] Robinson, M.E., Tawn, J.A. (2000) "Extremal analysis of processes sampled at different frequences," *Journal of the Royal Statistical Society Series B*, 62(1), 117135.

[15] Weissman, I., Novak S. Yu. (1998) "On blocks and runs estimators of the extremal index. Journal of Statistical Planning and Inference," 66, 281-288.