Disjoint and sliding blocks estimators for heavy tailed time series

Youssouph Cissokho

Joint work with Rafal Kulik

Department of Mathematics and Statistics (University of Ottawa)

2022 CMS Summer Meeting
(Memorial University in St. Johns)

August 12, 2022

uOttawa

CMS SMC

Outline

Motivation : What is an extreme event?

- Are rare by definition;
- High impact event:
  - Tornado outbreaks; large wildfires;
  - El Nino : a climate pattern that describes the unusual warming of surface waters (brings rains and extreme floods which destroys homes, hospitals, businesses, ...);

Motivation : What is an extreme event?

- Are rare by definition;
- High impact event:
  - Tornado outbreaks; large wildfires;
  - El Nino : a climate pattern that describes the unusual warming of surface waters (brings rains and extreme floods which destroys homes, hospitals, businesses, ...);

Hence, extremes remain a subject of active research and widely used in many other disciplines.

Motivation : What are extreme events?

Record heat under the dome : Lytton (northeast of Vancouver) set a record temperature of 50 °C on June 29, 2021, nearly 24 °C higher than normal. The next day, 90% of the small town of Lytton burned to the ground.

Figure: **Source: Environment and Climate Change Canada (600 people died in Vancouver, 650 000 farm animals perished).**

Motivation : What are extreme events?

British Columbia's flood of floods : between 200 and 300 mm in 2.5 days; 40 daily rainfall records were eclipsed with totals experienced only once every 100 years. One of the most destructive and expensive weather disasters in Canadian history.

Figure: **Source: Environment and Climate Change Canada; (where approximately 1.3 million animals died in flooded fields).**

Motivation

Consider a regularly varying sequence of i.i.d. nonnegative random variables $\{X_j^\dagger, j \in \mathbb{Z}\}$ with tail distribution $\overline{F}$. In particular:

- $\lim_{n\to\infty} \overline{F}(tx)/\overline{F}(x) = t^{-\alpha}$ for some $\alpha > 0$. (e.g. Pareto, Student).
- There exists a sequence $a_n \to \infty$ s.t.

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1} \max_{j=1,\ldots,n} \{X_j^\dagger\} \leq x) = \exp(-x^{-\alpha}) \;, \quad x > 0 \;.$$

Motivation

Consider a regularly varying sequence of i.i.d. nonnegative random variables $\{X_j^\dagger, j \in \mathbb{Z}\}$ with tail distribution $\overline{F}$. In particular:

- $\lim_{n\to\infty} \overline{F}(tx)/\overline{F}(x) = t^{-\alpha}$ for some $\alpha > 0$. (e.g. Pareto, Student).
- There exists a sequence $a_n \to \infty$ s.t.

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1} \max_{j=1,\dots,n} \{X_j^\dagger\} \le x) = \exp(-x^{-\alpha}) \ , \quad x > 0 \ .$$

If $\{X_j, j \in \mathbb{Z}\}$ is stationary, regularly varying with the same marginal tail df $\overline{F}$. Then

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1} \max_{j=1,\dots,n} \{X_j\} \le x) = \exp(-\theta x^{-\alpha}) \ , \quad x > 0 \ ,$$

where $\theta \in (0, 1]$ is called the *extremal index* (whenever exists).

Motivation

Goal: to estimate the quantity θ.

Motivation

Goal: to estimate the quantity θ.

The extremal index θ can be represented as

$$\lim_{x\to\infty} \mathbb{E}[H(X_j/x,\ j \in \mathbb{Z})]$$

for some $H : \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R} : \quad H(\boldsymbol{x}) = \mathbb{1}\left\{\max_{j\in\mathbb{Z}} x_j > 1\right\}$.

Questions:

- Can we consider different functionals $H : \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R}$?

Motivation

Goal: to estimate the quantity θ.

The extremal index θ can be represented as

$$\lim_{x\to\infty} \mathbb{E}[H(X_j/x,\ j \in \mathbb{Z})]$$

for some $H \ : \ \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R} \ : \ \ H(\boldsymbol{x}) = \mathbb{1}\left\{\max_{j\in\mathbb{Z}} x_j > 1\right\}.$

Questions:

- Can we consider different functionals $H \ : \ \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R}$?
- Yes, for specific choices of H we will define **H-cluster indices**.

Motivation

Goal: to estimate the quantity θ.

The extremal index θ can be represented as

$$\lim_{x\to\infty} \mathbb{E}[H(X_j/x,\ j\in\mathbb{Z})]$$

for some $H : \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R} : \quad H(\boldsymbol{x}) = \mathbb{1}\left\{\max_{j\in\mathbb{Z}} x_j > 1\right\}.$

Questions:

- Can we consider different functionals $H : \mathbb{R}_+^{\mathbb{Z}} \to \mathbb{R}$?
- Yes, for specific choices of H we will define **H-cluster indices**.
- How to estimate H-cluster indices? disjoint blocks, **sliding blocks** and runs estimators.

Tail process

Consider a stationary, regularly varying nonnegative time series $\boldsymbol{X} = \{X_j, j \in \mathbb{Z}\}$ with marginal distribution function F with tail index $\alpha > 0$.

[1] Basrak and Segers (2009)

Tail process

Consider a stationary, regularly varying nonnegative time series $\boldsymbol{X} = \{X_j, j \in \mathbb{Z}\}$ with marginal distribution function F with tail index $\alpha > 0$. Then, there exists $\boldsymbol{Y} = \{Y_j, j \in \mathbb{Z}\}$, called **tail process**[1], such that

$$\lim_{x\to\infty} \mathbb{P}(x^{-1}(X_i, \ldots, X_j) \in \cdot \,\big|\, |X_0| > x) = \mathbb{P}((Y_i, \ldots, Y_j) \in \cdot).$$

The process $\boldsymbol{Y}$ is <u>not stationary</u>. Explicit formulas do exist for some time series models.

[1] Basrak and Segers (2009)

Clusters of extremes, cluster functionals

Cluster functionals H

For $\boldsymbol{X} = \{X_j, j \in \mathbb{Z}\} \in (\mathbb{R})^{\mathbb{Z}}$. We denote $\boldsymbol{X}_{i,j} = (X_i, \ldots, X_j) \in (\mathbb{R})^{(j-i+1)}$ with $i \leq j \in \mathbb{Z}$. Then, we identify $H(\boldsymbol{X}_{i,j})$ with $H((\boldsymbol{0}, \boldsymbol{X}_{i,j}, \boldsymbol{0}))$, where $\boldsymbol{0} \in (\mathbb{R})^{\mathbb{Z}}$ is the zero sequence.

Clusters of extremes, cluster functionals

Cluster functionals H

For $\boldsymbol{X} = \{X_j, j \in \mathbb{Z}\} \in (\mathbb{R})^{\mathbb{Z}}$. We denote $\boldsymbol{X}_{i,j} = (X_i, \dots, X_j) \in (\mathbb{R})^{(j-i+1)}$ with $i \leq j \in \mathbb{Z}$. Then, we identify $H(\boldsymbol{X}_{i,j})$ with $H((\boldsymbol{0}, \boldsymbol{X}_{i,j}, \boldsymbol{0}))$, where $\boldsymbol{0} \in (\mathbb{R})^{\mathbb{Z}}$ is the zero sequence.

Given H on $(\mathbb{R})^{\mathbb{Z}}$, we want to estimate the limiting quantity (cluster index)

$$\nu^*(H) = \lim_{n\to\infty} \nu^*_{n,r_n}(H) = \lim_{n\to\infty} \frac{\mathbb{E}\Big[H(\boldsymbol{X}_{1,r_n}/u_n)\Big]}{r_n \mathbb{P}\Big(|X_0| > u_n\Big)},$$

with $r_n,\ u_n \to \infty$.

Question :

What are the conditions for the existence of such limit?

Assumptions

Assumptions on $r_n,\ u_n$ and the functional H are needed.

- $\lim_{n\to\infty} n\mathbb{P}(|X_0| > u_n) = \infty$ and $\lim_{n\to\infty} r_n\mathbb{P}(|X_0| > u_n) = 0$.

[2] Davis and Hsing (1995)

[3] Kulik, Soulier and Wintenberger (2019)

Assumptions

Assumptions on r_n, u_n and the functional H are needed.

- $\lim_{n\to\infty} n\mathbb{P}(|X_0| > u_n) = \infty$ and $\lim_{n\to\infty} r_n\mathbb{P}(|X_0| > u_n) = 0$.
- **Anticlustering condition** $\mathcal{AC}(r_n, u_n)$ Condition (extremes cannot persist for a infinite horizon time) holds if for all $x, y > 0$,[2]

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\max_{k\le|j|\le r_n} |X_j| > u_n x \;\middle|\; |X_0| > u_n y \right) = 0 \; .$$

It's valid e.g. geometrically ergodic Markov chains, short-memory linear or max-stable processes.[3]

[2] Davis and Hsing (1995)

[3] Kulik, Soulier and Wintenberger (2019)

Assumptions

Assumptions on $r_n,\ u_n$ and the functional H are needed.

- $\lim_{n\to\infty} n\mathbb{P}(|X_0| > u_n) = \infty$ and $\lim_{n\to\infty} r_n\mathbb{P}(|X_0| > u_n) = 0$.
- **Anticlustering condition** $\mathcal{AC}(r_n, u_n)$ Condition (extremes cannot persist for a infinite horizon time) holds if for all $x, y > 0$,[2]

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\max_{k\le|j|\le r_n} |X_j| > u_n x \,\middle|\, |X_0| > u_n y\right) = 0 .$$

  It's valid e.g. geometrically ergodic Markov chains, short-memory linear or max-stable processes.[3]

- However, H cannot be arbitrary. For e.g: of $H = 1$, then $\boldsymbol{\nu}^*(H) = \infty$, and if $K(\boldsymbol{x}) = \sum_{j\in\mathbb{Z}} \mathbb{1}\left\{x_j > 1\right\}$, then $\boldsymbol{\nu}^*(H) = 1$.

[2]Davis and Hsing (1995)

[3]Kulik, Soulier and Wintenberger (2019)

Example of H-cluster indices

Some cluster indices of interest are, among others:

- the **extremal index** obtained with $H_1(\boldsymbol{x}) = \mathbb{1}\left\{\sup_{j\in\mathbb{Z}} x_j > 1\right\}$.
- the **cluster size** distribution obtained with

$$H_{2,m}(\boldsymbol{x}) = \mathbb{1}\left\{\sum_{j\in\mathbb{Z}} \mathbb{1}\left\{\left|\boldsymbol{x}_j\right| > 1\right\} = m\right\}, \quad m \in \mathbb{N} \; ;$$

- the **large deviation index** of a univariate time series obtained with[4]

$$H_3(\boldsymbol{x}) = \mathbb{1}\{K(\boldsymbol{x}) > 1\} \,, \quad K(\boldsymbol{x}) = \left(\sum_{j\in\mathbb{Z}} x_j\right)_+ \quad .$$

[4] Mikosh and Wintenberger (2013, 2014)

Existence and representation

Theorem (1)

Let condition $\mathcal{AC}(r_n, u_n)$ hold. The sequence of measures converge vaguely $\boldsymbol{\nu}^*_{n,r_n} \to \boldsymbol{\nu}^*$, that is,

$$\lim_{n\to\infty} \boldsymbol{\nu}^*_{n,r_n}(H) = \lim_{n\to\infty} \frac{\mathbb{E}\Big[H(u_n^{-1}\boldsymbol{X}_{1,r_n})\Big]}{r_n\mathbb{P}\Big(|X_0| > u_n\Big)} = \boldsymbol{\nu}^*(H)\ .$$

for all bounded, continuous and shift invariant functions H with support separated from $\mathbf{0}$.

It has the following representation [5].

$$\boldsymbol{\nu}^*(H) = \mathbb{E}\Big[H(\boldsymbol{Y})\mathbb{1}\Big\{\boldsymbol{Y}^*_{-\infty,-1} \leq 1\Big\}\Big] = \mathbb{E}\left[\sup_{j\leq -1}|Y_j| < 1\right]\ .$$

[5] Kulik and Soulier (2020), Chapter VI

Disjoint blocks estimator

Consider the disjoint blocks statistics

$$\widetilde{\boldsymbol{DB}}_n(H) := \frac{1}{n\mathbb{P}(|X_0| > u_n)} \sum_{i=1}^{m_n} H\left(\boldsymbol{X}_{(i-1)r_n+1,ir_n}/u_n\right) ,$$

where $m_n = [n/r_n]$. Note that

$$\nu^*(H) = \lim_{n\to\infty} \mathbb{E}[\widetilde{\boldsymbol{DB}}_n(H)].$$

Disjoint blocks estimator

Consider the disjoint blocks statistics

$$\widetilde{\boldsymbol{DB}}_n(H) := \frac{1}{n\mathbb{P}(|X_0| > u_n)} \sum_{i=1}^{m_n} H\left(\boldsymbol{X}_{(i-1)r_n+1,ir_n}/u_n\right) ,$$

where $m_n = [n/r_n]$. Note that

$$\boldsymbol{\nu}^*(H) = \lim_{n\to\infty} \mathbb{E}[\widetilde{\boldsymbol{DB}}_n(H)].$$

For sequence of integers $k \to \infty$ such that $k/n \to \infty$, define $u_n = F^{\leftarrow}(1-k/n)$. Define the disjoint blocks estimator

$$\widehat{\boldsymbol{DB}}_n(H) = \frac{1}{k} \sum_{i=1}^{m_n} H\left(\boldsymbol{X}_{(i-1)r_n+1,ir_n}\Big/|X|_{(n:n-k)}\right) ,$$

where $|X|_{(n:1)} \leq \cdots \leq |X|_{(n:n)}$.

Sliding blocks estimator

Consider the sliding blocks statistics

$$\widetilde{\boldsymbol{S}B}_n(H) := \frac{1}{q_n r_n \mathbb{P}(|X_0| > u_n)} \sum_{i=0}^{q_n-1} H\left(\boldsymbol{X}_{i+1,i+r_n}/u_n\right) ,$$

where $q_n = n - r_n + 1$,

Sliding blocks estimator

Consider the sliding blocks statistics

$$\widetilde{SB}_n(H) := \frac{1}{q_n r_n \mathbb{P}(|X_0| > u_n)} \sum_{i=0}^{q_n-1} H\left(\boldsymbol{X}_{i+1,i+r_n}/u_n\right) ,$$

where $q_n = n - r_n + 1$,
and

$$\widehat{SB}_n(H) = \frac{1}{kr_n} \sum_{i=0}^{q_n-1} H\left(\boldsymbol{X}_{i+1,i+r_n}\Big/|X|_{(n:n-k)}\right) .$$

Sliding blocks estimator-CLT

Theorem (Cissokho and Kulik (2021), *Electronic Journal of Statistics*)

Let $\{X_j, j \in \mathbb{Z}\}$ be a stationary, regularly varying $\mathbb{R}$-valued and β-mixing time series and $s > 0$. Under the "appropriate" conditions

$$\sqrt{k}\left\{\widehat{SB}_n(H) - \boldsymbol{\nu}^*(H)\right\} \xrightarrow{\mathrm{d}} \mathbb{G}^*(H) ,$$

where $\mathbb{G}$ is a centered Gaussian process with covariance $\boldsymbol{\nu}^*(H\widetilde{H})$ and $\mathbb{G}^*(H) = \mathbb{G}(H - \boldsymbol{\nu}^*(H)\mathcal{E})$, $\mathcal{E}(\boldsymbol{x}) = \sum_{j\in\mathbb{Z}} \mathbb{1}\left\{\left|x_j\right| > 1\right\}$.

Sliding blocks estimator-CLT

Theorem (Cissokho and Kulik (2021), *Electronic Journal of Statistics*)

Let $\{X_j, j \in \mathbb{Z}\}$ *be a stationary, regularly varying* $\mathbb{R}$*-valued and* β*-mixing time series and* $s > 0$. *Under the "appropriate" conditions*

$$\sqrt{k}\left\{\widehat{SB}_n(H) - \boldsymbol{\nu}^*(H)\right\} \xrightarrow{\mathrm{d}} \mathbb{G}^*(H) \ ,$$

where $\mathbb{G}$ *is a centered Gaussian process with covariance* $\boldsymbol{\nu}^*(H\widetilde{H})$ *and*
$\mathbb{G}^*(H) = \mathbb{G}(H - \boldsymbol{\nu}^*(H)\mathcal{E})$, $\mathcal{E}(\boldsymbol{x}) = \sum_{j\in\mathbb{Z}} \mathbb{1}\left\{\left|x_j\right| > 1\right\}$.

The same asymptotics holds for disjoint blocks estimator as well.

Simulations-Stationary AR process

We start with a simple AR(1) process. For this process we have explicit formulas for all cluster indices. Samples of size $n = 1000$ are generated from AR(1) with $\alpha = 4$ and $\rho = 0.5, 0.9$.

Simulations-Stationary AR process

We start with a simple AR(1) process. For this process we have explicit formulas for all cluster indices. Samples of size $n = 1000$ are generated from AR(1) with $\alpha = 4$ and $\rho = 0.5, 0.9$.

Extremal index.

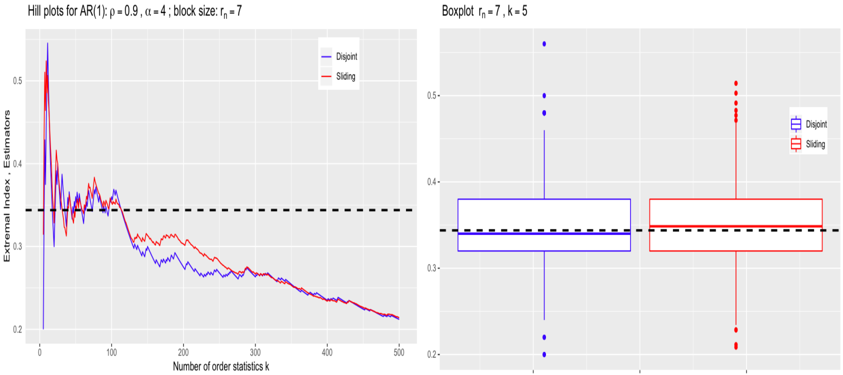
For AR(1) with $\rho > 0$ the extremal index is $\theta = 1 - \rho^{\alpha}$; (kulik and Soulier (2020)).

Simulations-Extremal index

	$\rho = 0.9$, Extremal Index=0.34				$\rho = 0.5$, Extremal Index= 0.94			
(k %)	$k = 5$		$k = 10$		$k = 5$		$k = 10$	
$r_n = 7$								
Disjoint bl	**0.34**	(0.05)	**0.31**	(0.03)	**0.68**	(0.05)	**0.58**	(0.03)
Sliding bl	**0.35**	(0.04)	**0.31**	(0.03)	**0.68**	(0.04)	**0.58**	(0.03)
$r_n = 8$								
Disjoint bl	**0.32**	(0.05)	**0.29**	(0.03)	**0.67**	(0.05)	**0.56**	(0.03)
Sliding bl	**0.33**	(0.04)	**0.29**	(0.03)	**0.67**	(0.04)	**0.56**	(0.03)
$r_n = 9$								
Disjoint bl	**0.32**	(0.05)	**0.28**	(0.03)	**0.66**	(0.05)	**0.53**	(0.03)
Sliding bl	**0.32**	(0.04)	**0.28**	(0.03)	**0.65**	(0.05)	**0.53**	(0.03)
$r_n = 10$								
Disjoint bl	**0.30**	(0.05)	**0.26**	(0.03)	**0.64**	(0.05)	**0.52**	(0.03)
Sliding bl	**0.30**	(0.04)	**0.26**	(0.03)	**0.63**	(0.05)	**0.52**	(0.03)

Figure: The median and the variance (in brackets) of disjoint and sliding blocks estimators for the extremal index. Data are simulated from AR(1) with $\alpha = 4$, $\rho = 0.5$ (thus, $\theta = 0.94$), and $\rho = 0.9$ (thus $\theta = 0.34$). Block size $r_n = 7,\ 8,\ 9,\ 10$. The number of order statistics is $k = 5\%$ and 10% for a sample $n = 1000$ based on $N = 1000$ Monte Carlo simulations.

Simulations-Extremal index



PoT vs. Block maxima

PoT method

- Drees and Neblung (2020) studied asymptotic normality of the sliding blocks estimators in general setting. they showed that it's limiting variance **does not exceed that of the disjoint blocks estimators**.
- For the extremal index, they **showed that the variances are equal**.

Note: we worked under PoT method.

Block maxima

- Robert, Segers, Ferro (2009) and bücher and Segers (2018a, 2918b), Zou, Volgusher and Bücher (2021): Sliding blocks estimators have smaller variance that the disjoint blocks.

Our contribution

To the best of our knowledge, this paper makes the following contribution:

- Central limit theorem for the data-based sliding blocks estimators under easy to verify assumptions.
- We give an explicit formula for the asymptotic variance. As such, we can conclude that **the sliding and the disjoint blocks estimators yield the same asymptotics**.
- This solves **the longstanding problem in the context of cluster functionals**.

Open questions

- Runs estimators (Cissokho and Kulik (2021)) (**accepted** for publication for EJS);
- Consistency of sliding blocks estimators under minimal conditions (Cissokho (2021)) ;
- Extend CLT for sliding blocks estimators (Theorem 1) to piecewise stationary processes. This line of research was proposed recently by Axel Bücher and his student. Piecewise stationary processes may be used in climate modeling.
- Obtain the results of (Theorem 1) under minimal conditions (that is, without relying on β-mixing and linear ordering of function classes). Do these results are valid under long range dependence?
- Can we extend the asymptotic results presented here to Gumbel domain of attraction? note that the probabilistic methods have to be completely different.
- Since the disjoint and sliding blocks statistics have the same asymptotic behaviour, is it possible to obtain an asymptotic expansion for the difference between these two statistics?
- Can we compare results between Peak-over-Threshold and Block Maxima methods?

Thank you and questions please. . .