Lomonosov Moscow State University
Faculty of mechanics and mathematics
Department of Probability Theory

Term paper

**Declustering parameter selection for semiparametric maxima estimators of the extremal index**

**Student:**
N.A. Akimov, groupe 433

**Advisors:**
E.M. Ryadnova, I.V. Rodionov

Moscow, 2021

# 1 Introduction

Let $X_1, X_2 \dots, X_n$ be a strictly stationary sequence of random variables (r.v.s) with cumulative distribution function (cdf) $F(x)$. According to Leadbetter et al. (1983) we say that such sequence has the extremal index $\theta \in (0, 1]$ if for each $0 < \tau < \infty$ there exists a sequence of real numbers $u_n = u_n(\tau)$ such that it holds

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau, \qquad \lim_{n\to\infty} P(M_{1,n} < u_n) = e^{-\tau\theta}, \tag{1}$$

where $M_{i,j} = \max\{X_{i+1}, X_{i+2} \dots, X_j\}$, $M_{1,1} = -\infty$. Denote tail function as $\overline{F}(x) = 1 - F(x)$. Now and further we assume that $F(x)$ and, therefore, $\overline{F}(x)$ are continuous.

Define a cluster as the number of consecutive observations exceeding the threshold $u$ between two consecutive non-exceedances. Ferro and Segers (2003) considered a random variable $T(u)$, equal in distribution to $\min\{j \geq 1 : X_{j+1} > u\}$ given $X_1 > u$.

**Defenition 1.1.** *( Ferro and Segers (2003)) For real $u$ and integers $1 \leq k \leq l$, let $\mathcal{F}_{k,l}(u)$ be the $\sigma$-field generated by events $\{X_i > u\}$, $k \leq i \leq l$. Define the mixing coefficients $\alpha_{n.q}(u)$,*

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

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

**Theorem 1.1.** *( Ferro and Segers (2003)) Let $\{X_n\}_{n\geq1}$ be a stationary process of r.v.s with tail function $\overline{F}(x)$. Let the positive integers $\{r_n\}$ and the thresholds $\{u_n\}$, $n \geq 1$ be such that $r_n \to \infty$, $r_n\overline{F}(u_n) \to \tau$ and $P\{M_{r_n} \leq u_n\} \to \exp(-\theta\tau)$ holds as $n \to \infty$ for some $\tau \in (0, \infty)$ and $\theta \in (0, 1]$. If there are positive integers $q_n = o(r_n)$ such that $\alpha_{cr_n,q_n}(u_n) = o(1)$ for any $c > 0$, then we get for $t > 0$*

$$P\{\overline{F}(u_n)T(u_n) > t\} \to \theta \exp(-\theta t).$$

For sample $X_1, X_2 \dots, X_n$ from strictly stationary sequence $\{X_i\}$ let $N = \sum\limits_{i=1}^{n} \mathbb{I}(X_i > u)$ be the number of exceedances over $u$ and $1 \leq S_i \leq n$ be the times of such exceedances. Denoting $T_i = S_{i+1} - S_i$, for $i = 1, \dots L$ where $L = N - 1$ and using Theorem 1.1 one can get the intervals estimator of the extremal index

$$\hat\theta_n(u) = \begin{cases} \min(1, \hat\theta^1_n(u)), \text{ if } \max\{T_i : 1 \leq i \leq L\} \leq 2, \\ \min(1, \hat\theta^2_n(u)), \text{ if } \max\{T_i : 1 \leq i \leq L\} > 2, \end{cases}$$

where

$$\hat\theta^1_n(u) = \frac{2(\sum_{i=1}^{L} T_i)^2}{L\sum_{i=1}^{L} T_i^2}, \quad \hat\theta^2_n(u) = \frac{2(\sum_{i=1}^{L}(T_i - 1))^2}{L\sum_{i=1}^{L}(T_i - 1)(T_i - 2)}.$$

Next we denote intervals estimator as $\hat\theta_0$.

## 2 Estimators and the discrepancy method

### 2.1 The discrepancy method and threshold selection for extremal index estimation

All known methods of extremal index estimation depends on some parameters such as threshold and/or declustering parameter. Markovich (2015) suggested a way to optimize this selection using the discrepancy method. In general, discrepancy methods are based on nonparametric statistics like Kolmogorov-Smirnov $D_n$ and Cramer-von Mises-Smirnov $\omega_n^2$.

Suppose we have a sample of r.v.s with cdf $F(x)$, which belongs to some parametric family of distributions with parameter $h$. We are looking for the best choice of $h$. The idea of the method is to solve the discrepancy equation

$$\rho(\hat{F}_h, F_n) = \delta$$

regarding $h$, where $\rho(\cdot,\cdot)$ is some metric on the space of density functions, $F_n$ is the empirical cdf and $\delta = \rho(F, F_n)$ is some value. Taking Cramer-von-Mises-Smirnov (C-M-S) statistic as $\rho(F, F_n)$ we get

$$\omega_n^2(h) = n \int_{-\infty}^{\infty} (F_n(x) - \hat{F}_h(x))^2 d\hat{F}_h(x).$$

Regarding practical applications Markovich (1989) propose to select the parameter $h$ as a solution of the equation

$$\hat{\omega}_n^2(h) = 0.05,$$

where

$$\hat{\omega}_n^2(h) = \sum_{i=1}^{n} \left( \hat{F}_h(X_{i,n}) - \frac{i-0.5}{n} \right)^2 + \frac{1}{12n} \tag{2}$$

was calculated by the order statistics $X_{1,n} \le \dots \le X_{n,n}$ corresponding to the sample $\{X_i\}_{i=1}^n$, and the value 0.05 corresponding to the mode of limit distribution of the C-M-S statistic $A_1$. A similar idea was explored in Markovich (2015) to estimate the extremal index.

From Theorem 1.1 we know that the cdf of $Y_i = (N_u/n)T_i$ is asymptotically equal to $G_\theta(t) = 1 - \theta\exp(-\theta t)$ for some $\theta \in (0,1]$. So, taking $G_{\hat{\theta}}(Y_{i,L})$ instead of $F_h(X_{i,n})$ where $\hat{\theta}(u)$ is some estimator of $\theta$ and using only $k$ largest inter-exceedance times we can rewrite (2) as

$$\omega_L^2(\hat{\theta}(u)) = \sum_{i=L-k+1}^{L} \left( \hat{G}_{\hat{\theta}}(Y_{i,L}) - \frac{i-0.5}{L} \right)^2 + \frac{1}{12L} = \delta. \tag{3}$$

The selection of $k$ and $\delta$ remains a problem. To overcome it, Markovich and Rodionov (2020) suggested a specific normalization of the discrepancy statistic $\omega_L^2(\hat{\theta}(u))$ in (3) such that its limit distribution coincides with the limit distribution $A_1$ of C-M-S statistic. Then its quantiles can be used as $\delta$. They introduced the statistic

$$\tilde{w}_L^2(\hat{\theta}(u)) = \sum_{i=L-k+1}^{k-1} \left( 1 - \frac{\hat{\theta}\exp(-Y_{L-i,L}\hat{\theta})}{1-\hat{t}_k} - \frac{k-i-0.5}{k} \right)^2 + \frac{1}{12k}, \tag{4}$$

where $\hat{t}_k = 1 - \hat{\theta}\exp(-Y_{L-k,L}\hat{\theta})$.

**Theorem 2.1.** *(Theorem 3.3, Markovich and Rodionov (2020)) Let the conditions of Theorem 1.1 and the condition (9) in Markovich and Rodionov (2020) be fulfilled and the estimator of extremal index $\hat{\theta} = \hat{\theta}_n$ be such that*

$$\sqrt{m_n}(\hat{\theta}_n - \theta) \xrightarrow{d} \zeta, \ n \to \infty,$$

*where the r.v. $\zeta$ has a nondegenerate distribution function $H$. Let us assume that the sequence $m_n$ is such that*

$$\frac{k}{m_n} = o(1) \ \textit{and} \ \frac{(\ln L)^2}{m_n} = o(1),$$

*as $n \to \infty$. Then*

$$\tilde{w}_L^2(\hat{\theta}) \xrightarrow{d} \xi \sim A_1$$

*holds, where $A_1$ is the limit distribution function of the C-M-S statistic.*

## 2.2 Declustering parameter selection for semiparametric maxima estimators of $\theta$

The first of semiparametric maxima estimators was suggested by Northrop (2015) and then some modification of them was considered in Berghaus, Bücher (2018). We are interested in both versions.

For $x \in (0,1)$ define sequences $u_n$ and $u'_n$ such that $u_n = F^{\leftarrow}(1 - x/n)$ and $u'_n = F^{\leftarrow}(e^{-x/n})$, where $F^{\leftarrow}(x)$ is the generalized inverse for $F(x)$. Then $n\overline{F}(u_n) = x$ and $n\overline{F}(u'_n) = n(1 - e^{-x/n}) \to x$, as $n \to \infty$. Also let $R_{1,n} = -n\log(F(M_{1,n}))$ and $Q_{1,n} = n(1 - F(M_{1,n}))$. We get by (1)

$$\begin{aligned} P\{Q_{1,n} \geq x\} &= P\{n(1 - F(M_{1,n})) \geq n(1 - F(u_n))\} = P\{M_{1,n} \leq u_n\} \to e^{-\theta x}, \\ P\{R_{1,n} \geq x\} &\sim P\{-n\log(F(M_{1,n})) \geq -n\log(F(u_n))\} = P\{M_{1,n} \leq u_n\} \to e^{-\theta x}, \end{aligned} \tag{5}$$

as $n \to \infty$.
We just showed that $Q_{1,n}$ and $R_{1,n}$ both asymptotically follow an exponential distribution with parameter $\theta$.

Divide $X_1, X_2 \ldots, X_n$ into $r_n$ blocks of length $b_n$ (we can omit last block if it has less then $b_n$ observations). We can use two approaches: disjoint blocks ($r_n = \lfloor n/b_n \rfloor$) or sliding blocks ($r_n = n - b_n + 1$). For block $B_i = \{X_{(i-1)b_n+1}, \ldots, X_{ib_n}\}$, $i = 1, \ldots, r_n$ define $R_{i,n}$ and $Q_{i,n}$ similarly to $R_{1,n}$ and $Q_{1,n}$ with $u_n$ replaced with $u(b_n)$ and $u'_n$ replaced with $u'(b_n)$. If $b_n$ is sufficiently large, then, according to (4), $R_{1,n}, \ldots, R_{r_n,n}$ and $Q_{1,n}, \ldots, Q_{r_n,n}$ have approximately the exponential distribution with mean $1/\theta$. Consider the maximum likehood estimators

$$\tilde{\theta}_n^N = \left(\frac{1}{r_n}\sum_{i=1}^{r_n} R_{i,n}\right)^{-1}, \qquad \tilde{\theta}_n^B = \left(\frac{1}{r_n}\sum_{i=1}^{r_n} Q_{i,n}\right)^{-1}.$$

.

Usually we do not know the cdf of $\{X_i\}$ and thus we need to estimate $F(x)$ and, therefore, $R_{i,n}$ and $Q_{i,n}$. For $y \in B_i$ Northrop (2015) used the empirical cdf based on $X_k \notin B_i$

$$\hat{F}_i(y) = \frac{1}{n - b_n + 1}\sum_{X_k \notin B_i} \mathbb{1}(X_k \leq y).$$

Taking $\hat{R}_{i,n} = -n\log(\hat{F}_i(M_{(i-1)b_n, ib_n}))$ and $\hat{Q}_{i,n} = n(1 - \hat{F}_i(M_{(i-1)b_n, ib_n}))$, we finally get

$$\hat{\theta}_n^N = \left(\frac{1}{r_n}\sum_{i=1}^{r_n}\hat{R}_{i,n}\right)^{-1}, \qquad \hat{\theta}_n^B = \left(\frac{1}{r_n}\sum_{i=1}^{r_n}\hat{Q}_{i,n}\right)^{-1}.$$

These estimators are asymptotically equivalent, consistent and asymptotically normal for both sliding and disjoint blocks approaches (see theorem 3.1, theorem 3.2 in Berghaus, Bücher (2018)), with $\sigma_{dj}^2 - \sigma_{sl}^2 = \theta^2(3 - 4\ln 2)$, where $\sigma_{dj}^2$ and $\sigma_{sl}^2$ are asymptotic variances.

Semiparametric maxima estimators highly depend on parameter $b_n$. We can use the discrepancy method for selection of $b_n$. The first way is use $\hat{\theta}_n^N$ or $\hat{\theta}_n^B$ instead of $\hat{\theta}$ with $b = b(u)$ in (4) and solve discrepancy equation regarding to $u$ and then get the value of $b$.

The second way is described below. We showed that $R_{i,n}$ and $Q_{i,n}$ both have approximately the exponential distribution with mean $1/\theta$. Then, replacing $G_{\hat{\theta}}(t)$ with $\hat{F}_{\hat{\theta}}(t) = 1 - \exp(-\hat{\theta}t)$ and $Y_{i,L}$ with $\hat{R}_{ni}$ or $\hat{Q}_{ni}$ in (4) we get

$$\tilde{\omega}_N^2(\hat{\theta}(b)) = \sum_{i=1}^{r_n}\left(\hat{F}_{\hat{\theta}}(R_{i,n}) - \frac{i-0.5}{r_n}\right)^2 + \frac{1}{12r_n}, \tag{6}$$

$$\tilde{\omega}_B^2(\hat{\theta}(b)) = \sum_{i=1}^{r_n}\left(\hat{F}_{\hat{\theta}}(Q_{i,n}) - \frac{i-0.5}{r_n}\right)^2 + \frac{1}{12r_n}. \tag{7}$$

Then we can choose $b_n$ as solution of the discrepancy equation

$$\tilde{\omega}_N^2 = \delta, \quad \tilde{\omega}_B^2 = \delta, \tag{8}$$

where $\delta$ is some quantile of $A_1$ distribution.

**Remark 2.1.** *The limit distributions of $\tilde{\omega}_N^2$ and $\tilde{\omega}_B^2$ are unknown. Its quantiles can be used as $\delta$ in (8) for better accuracy.*

# 3 Simulation study

In this section, we simulated some processes with known values of $\theta$ and calculated the interval and semiparametric maxima estimators for them using the discrepancy method for the selection of threshold $u$ for $\hat{\theta}_0$ and $b_n$ for $\hat{\theta}_n^N$ and $\hat{\theta}_n^B$. The simulation is repeated 1000 times with the sample size of $n = 10^5$ and $n = 5000$ of initial measurements.

## 3.1 Models

In this modelling were considered processes MM, ARMAX, AR(1), AR(2), MA(2) and GARCH(1,1) with known values of $\theta$.

The $mth$ order MM process is $X_t = \max_{0\le i\le m}\{\alpha_y Z_{t-i}\}$, $t \in \mathbb{Z}$, where $\{\alpha_i\}$ are constants with $\alpha_i \ge 0$, $\sum_{i=0}^m \alpha_i = 1$, and $Z_t$ are i.i.d standard Fréchet distributed r.v.s with cdf $F(x) = \exp(-1/x)$, for $x > 0$. The extremal index of this process is equal to $\theta = \max\{\alpha_i\}$, Ancona-Navarrete and Tawn (200). Values $m = 3$ and $\theta in\{0.5, 0.8\}$ corresponding to $\{\alpha_i\}_{i=0}^3 = \{0.5, 0.3, 0.15, 0.05\}$ and $\{\alpha_i\}_{i=0}^3 = \{0.8, 0.1, 0.008, 0.02\}$, respectively, are taken for our study.

The ARMAX process is determined as $X_t = \max\{\alpha X_{t-1}, (1-\alpha)Z_t\}$, $t \in \mathbb{Z}$, where $0 \le \alpha < 1$, $\{Z_t\}$ are i.i.d standard Fréchet distributed r.v.s and $P\{X_t \le x\} = \exp(-1/x)$ holds assuming $X_0 = Z_0$. The extremal index of the process was proven to be equal $\theta = 1 - \alpha$, Beirlant et al. (2004). We consider $\theta \in \{0.25, 0.75\}$.

The positively correlated AR(1) process with uniform noise (ARu+) is defined by $X_j = (1/r)X_{j-1} + \varepsilon_j$, $j \ge 1$ and $X_0 \sim U(0,1)$ with $X_0$ independent of $\{\varepsilon_j\}$. Then $X_j \sim U(0,1)$ holds for all $j \ge 1$. For a fixed integer $r \ge 2$ let $\varepsilon_n$, $n \ge 1$ be i.i.d. r.v.s with $P\{\varepsilon_1 = k/r\} = 1/r$, $k \in \{0, 1, \ldots, r-1\}$. The extremal index of ARu+ is $\theta = 1 - 1/r$ (Chernick et al. (1991)). $\theta \in \{0.5, 0.8\}$ corresponding to $r \in \{2, 5\}$ are taken.

The negatively correlated AR(1) process with uniform noise (ARu-) is defined by $X_j = -(1/r)X_{j-1} + \varepsilon_j$ with similarly distributed $\{\varepsilon_j\}$ but with support $k \in \{1, \ldots, r\}$. Its extremal index is $\theta = 1 - 1/r^2$ (Chernick et al. (1991)). The same $r'$s were taken corresponding to $\theta \in \{0.75, 0.96\}$.

We simulate the MA(2) process (Sun and Samorodnitsky (2018)) $X_i = pZ_{i-2} + qZ_{i-1} + Z_i$, $i \ge 1$, with $p > 0$, $q < 1$, and i.i.d. Pareto random variables $Z_{-1}, Z_0, Z_2, \ldots$ with $P\{Z_0 > x\} = 1$ if $x < 1$, and $P\{Z_0 > x\} = x^{-\alpha}$ if $x \ge 1$,for some $\alpha > 0$. The extremal index of the process is $\theta = (1 + p^\alpha + q^\alpha)^{-1}$. The cases $\alpha = 2$, $(p, q) = (1/\sqrt{2}, 1/\sqrt{2})$, $(1/\sqrt{3}, 1/\sqrt{6})$ with corresponding $\theta \in \{1/2, 2/3\}$ are considered.

We consider also processes studied in (Ferreira (2019); Northrop (2015); Süveges and Davison (2010)): the AR(1) process $X_j = 0.7X_{j-1} + \varepsilon_j$ , where $\{\varepsilon_j\}$ are standard Cauchy distributed and $\theta = 0.3$ (ARc); the AR(2) process $X_j = 0.95X_{j-1} - 0.89X_{j-2} + \varepsilon_j$, where $\{\varepsilon_j\}$ are Pareto distributed with tail index 2 and $\theta = 0.25$; and the GARCH(1,1) process $X_j = \sigma_j \varepsilon_j$ with $\sigma_j^2 = \alpha + \lambda X_{j-1}^2 + \beta\sigma_{j-1}^2$, $\alpha = 10^{-6}$, $\beta = 0.7$, $\lambda = 0.25$, the i.i.d. sequence of standard Gaussian r.v.s $\{\varepsilon_j\}_{j \ge 1}$ and $\theta = 0.447$ (see Laurini and Tawn (2012)).

## 3.2 Algorithms

For estimating we used the following algorithms.

**Algorithm 3.1. Threshold selection.** (Algorithm 4.1 from Markovich, Rodionov (2020)):

1. *Using* $X^n = \{X_i\}_{i=1}^n$ *and taking thresholds* $u$ *corresponding to quantile levels* $q \in \{0.90, 0.905, \ldots, 0.995\}$*, generate samples of the inter-exceedance times* $\{T_i(u)\}$ *and the normalized r.v.s*

$$\{Y_i\} = \{\overline{F}(u)T_i(u)\} = \{(N_u/n)T_i(u)\}, \quad i \in \{1, 2, \ldots, L\}, \quad L = L(u),$$

   *where* $N_u$ *is the number of exceedances over threshold* $u$*.*

2. *For each* $u$ *select* $k = \lfloor\hat{\theta}_0 L\rfloor$ *(in case* $\hat{\theta}_0 = 1$ *accept* $k = L - 1$*), where the intervals estimation was selected as a pilot estimator* $\hat{\theta}_0 = \hat{\theta}_0(u)$ *with the same* $u$ *as in Item 1.*

3. *Use the sorted sample* $Y_{L-k+1,L} \le \cdots \le Y_{L,L}$ *and find all solutions* $u_1, \ldots, u_l$ *among considered quantiles of the following discrepancy inequality*

$$\tilde{w}_L^2(\hat{\theta}_0) \le \delta,$$

   *where* $\delta = 1.49$ *is the 99.98 % quantile of the C-M-S statistic.*

4. *For each* $u_j$, $j \in \{1, \dots, l\}$ *calculate* $\hat{\theta}(u_j)$ *and find*

$$\hat{\theta}^{MR} = \frac{1}{l}\sum_{i=1}^{l} \hat{\theta}(u_i)$$

   *as the resulting estimator.*

**Remark 3.1.** *Let* $b = \frac{1}{1-q}$, *where* $q$ *is the level of quantile* $u$. *In our method we use* $\hat{\theta}_n^N$ *or* $\hat{\theta}_n^B$ *instead of* $\theta_0$ *in the latter algorithm. The estimates* $\hat{\theta}_1^N$ *and* $\hat{\theta}_1^B$ *appearing in item 4, are resulting.*

**Algorithm 3.2. Declustering parameter selection** :

1. *Using* $X^n = \{X_i\}_{i=1}^n$ *and taking* $b_1 = 2n \cdot 10^{-4}$, *and* $b_j = 5n(i-1) \cdot 10^{-4}$, *for all* $2 \le i \le 21$, *generate samples* $R_{i,n}(b_j)$, *for* $1 \le i \le r_n$ *and* $1 \le j \le 21$ *(see Fig.1 and Fig.2).*

2. *Use the sorted sample* $R_{(1,n)} \le \dots \le R_{(r_n,n)}$ *and find all solutions* $b_1, \dots, b_l$ *among considered declustering parameters of the following discrepancy inequality*

$$\tilde{\omega}_N^2(\hat{\theta}(b_j)) \le \delta, \tag{9}$$

   *where* $\delta = 1.49$ *is the 99.98 % quantile of the C-M-S statistic.*

3. *For each* $b_j$, $j \in \{1, \dots, l\}$ *calculate* $\hat{\theta}(b_j)$ *and find*

$$\hat{\theta}_2^N = \frac{1}{l}\sum_{i=1}^{l} \hat{\theta}(b_i)$$

   *as the resulting estimator.*

For $\hat{\theta}^B$ the algorithm is the same.

**Remark 3.2.** *Inequality (8) may not have solutions. In that case, we select* $b_0$ *as*

$$b_0 = \operatorname*{argmin}_B \left[\tilde{\omega}_N^2(\hat{\theta}(b_j))\right],$$

*where* $B$ *is the set from Item 1 of Algorithm 3.2., and then find*

$$\hat{\theta}_2^N = \hat{\theta}^N(b_0)$$

*as the resulting estimator.*

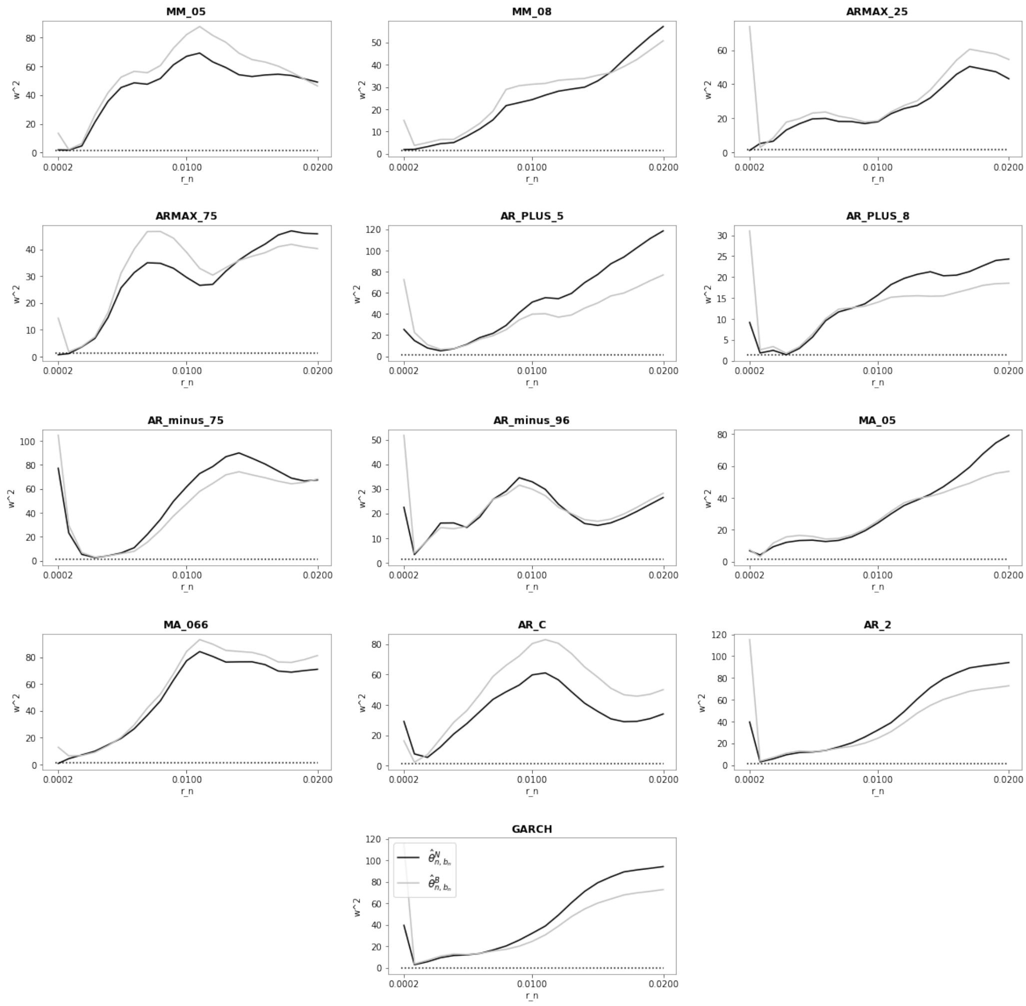


Figure 1: Dependence of $\tilde{\omega}^2_N(\hat{\theta}(b))$ and $\tilde{\omega}^2_B(\hat{\theta}(b))$ on $r_n$, where $b_n = r_n \cdot n$; $n = 10^5$. Dotted line shows level 1.49. We can see, that for large $n$, we should take $r_n$ less than $n \cdot 10^{-2}$, such that $\tilde{\omega}^2_N(\hat{\theta}(b))$ and $\tilde{\omega}^2_B(\hat{\theta}(b))$ reach 1.49 or less.

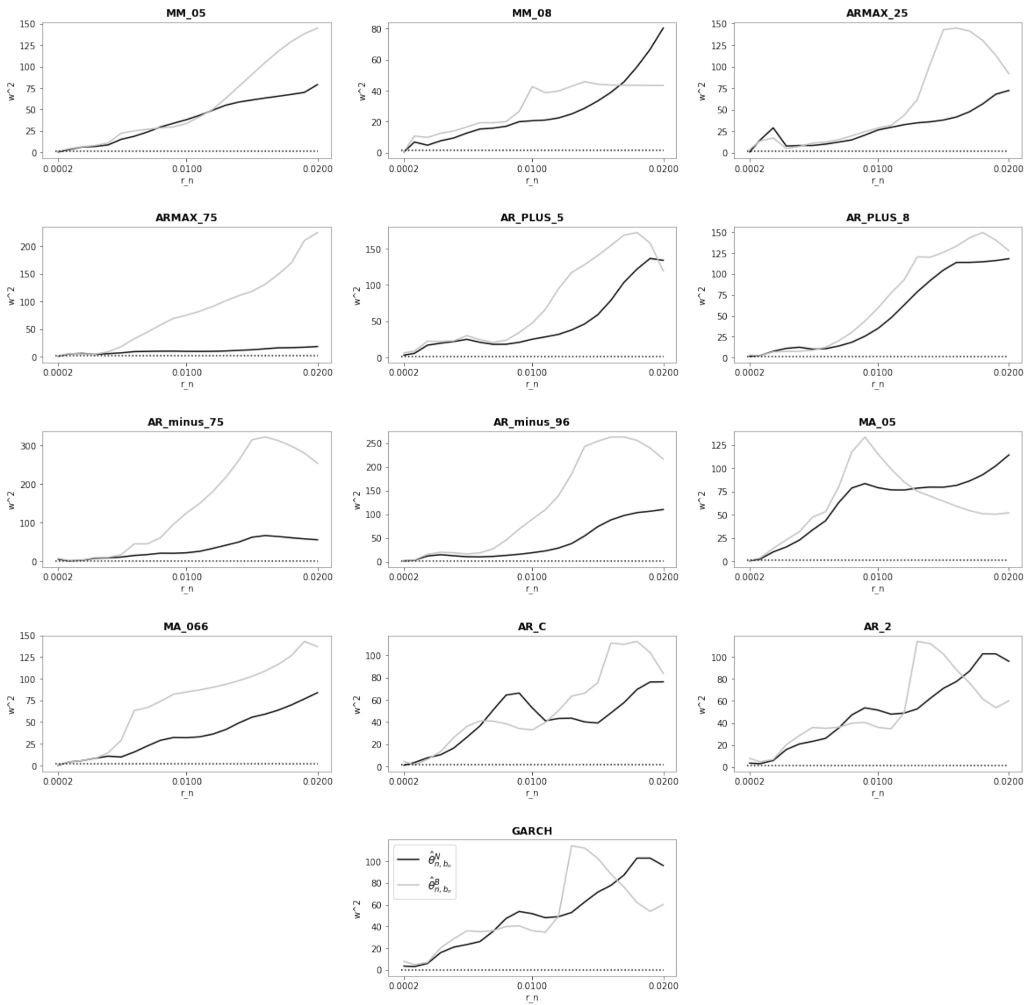


Figure 2: Dependence of $\tilde{\omega}_N^2(\hat{\theta}(b))$ and $\tilde{\omega}_B^2(\hat{\theta}(b))$ on $r_n$, where $b_n = r_n \cdot n$;$n = 5000$. Dotted line shows level 1.49. We can see, that for not so large $n$, we also should take $r_n$ less than $n \cdot 10^{-2}$, such that $\tilde{\omega}_N^2(\hat{\theta}(b))$ and $\tilde{\omega}_B^2(\hat{\theta}(b))$ reach 1.49 or less.

### 3.3 Results

Simulation results are presented below. We are interested in *Bias* and *RMSE* of our estimators.

Table 1: Bias$\cdot 10^4$ for $\hat{\theta}^{MR}$, $\hat{\theta}_1^N$, $\hat{\theta}_1^B$, $\hat{\theta}_2^N$ and $\hat{\theta}_2^B$ for $n = 10^5$ and $n = 5000$

| Bias$\cdot 10^4$ | MM | | ARMAX | | $ARu^+$ | | $ARu^-$ | | MA(2) | | ARc | AR(2) | GARCH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.5 | 0.8 | 0.25 | 0.75 | 0.5 | 0.8 | 0.75 | 0.96 | 0.5 | 2/3 | 0.3 | 0.25 | 0.447 |
| | $n = 10^5$ | | | | | | | | | | | | |
| $\hat{\theta}^{MR}$ | **-30** | **-15** | **-30** | **-17** | 270 | 316 | 428 | 266 | -200 | -509 | -79 | **310** | -431 |
| $\hat{\theta}_1^{N,SL}$ | 91 | 34 | 109 | 45 | 405 | 427 | 582 | 338 | -206 | -780 | -131 | 443 | -227 |
| $\hat{\theta}_1^{B,SL}$ | 184 | 168 | 171 | 172 | 432 | 467 | 611 | 362 | -41 | -527 | 77 | 489 | -179 |
| $\hat{\theta}_1^{N,DJ}$ | 88 | 28 | 107 | 39 | 402 | 426 | 564 | 326 | -207 | -780 | -97 | 443 | -224 |
| $\hat{\theta}_1^{B,DJ}$ | 174 | 153 | 167 | 158 | 434 | 463 | 574 | 348 | -48 | -537 | 73 | 484 | -181 |
| $\hat{\theta}_2^{N,SL}$ | 205 | 74 | 333 | 91 | 316 | 364 | 309 | 294 | -169 | -426 | -29 | 580 | **-1.8** |
| $\hat{\theta}_2^{B,SL}$ | 158 | 149 | 139 | 145 | 308 | 356 | 355 | 292 | **-21** | -439 | 127 | 504 | -38 |
| $\hat{\theta}_2^{N,DJ}$ | 42 | 19 | 55 | 24 | **224** | **206** | **268** | **69** | -115 | -354 | **-12** | 365 | -403 |
| $\hat{\theta}_2^{B,DJ}$ | 99 | 79 | 76 | 84 | 228 | 242 | 302 | 70 | -53 | **-302** | 48 | 393 | -376 |
| | $n = 5000$ | | | | | | | | | | | | |
| $\hat{\theta}^{MR}$ | **37** | **41** | **50** | **44** | 455 | 677 | 1054 | **226** | -224 | -712 | -60 | **552** | -1420 |
| $\hat{\theta}_1^{N,SL}$ | 226 | 58 | 288 | 88 | 800 | 740 | 1275 | 304 | -178 | -934 | -97 | 753 | -903 |
| $\hat{\theta}_1^{B,SL}$ | 509 | 456 | 471 | 469 | 931 | 1084 | 1408 | 354 | 195 | -445 | 294 | 896 | -888 |
| $\hat{\theta}_1^{N,DJ}$ | 230 | 68 | 291 | 94 | 807 | 758 | 1280 | 299 | -174 | -929 | -63 | 752 | **-875** |
| $\hat{\theta}_1^{B,DJ}$ | 465 | 386 | 443 | 404 | 880 | 1005 | 1333 | 323 | 151 | -495 | 267 | 841 | -890 |
| $\hat{\theta}_2^{N,SL}$ | 193 | 50 | 328 | 63 | 722 | 652 | 1014 | 286 | -188 | -794 | **-23** | 1230 | 1306 |
| $\hat{\theta}_2^{B,SL}$ | 509 | 427 | 473 | 436 | 726 | 906 | 1378 | 320 | 221 | -456 | 383 | 1051 | 1662 |
| $\hat{\theta}_2^{N,DJ}$ | 58 | -371 | 205 | -251 | **136** | **107** | **27** | -704 | -98 | -466 | 132 | 586 | -2223 |
| $\hat{\theta}_2^{B,DJ}$ | 116 | -315 | 265 | -194 | 192 | 161 | 72 | -685 | **-38** | **-404** | 196 | 643 | -2186 |

Table 2: RMSE$\cdot 10^4$ for $\hat{\theta}^{MR}$, $\hat{\theta}_1^N$, $\hat{\theta}_1^B$, $\hat{\theta}_2^N$ and $\hat{\theta}_2^B$ for $n = 10^5$ and $n = 5000$

| RMSE$\cdot 10^4$ | MM | | ARMAX | | $ARu^+$ | | $ARu^-$ | | MA(2) | | ARc | AR(2) | GARCH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.5 | 0.8 | 0.25 | 0.75 | 0.5 | 0.8 | 0.75 | 0.96 | 0.5 | 2/3 | 0.3 | 0.25 | 0.447 |
| | $n = 10^5$ | | | | | | | | | | | | |
| $\hat{\theta}^{MR}$ | **106** | 166 | **90** | 150 | **313** | 335 | 487 | 284 | 231 | 537 | 134 | **355** | 556 |
| $\hat{\theta}_1^{N,SL}$ | 115 | **117** | 129 | **112** | 427 | 437 | 622 | 341 | 213 | 783 | 241 | 455 | 301 |
| $\hat{\theta}_1^{B,SL}$ | 206 | 221 | 192 | 217 | 459 | 480 | 648 | 364 | 82 | 535 | 108 | 501 | 266 |
| $\hat{\theta}_1^{N,DJ}$ | 116 | 122 | 128 | 114 | 422 | 441 | 619 | 332 | 216 | 783 | 115 | 456 | 309 |
| $\hat{\theta}_1^{B,DJ}$ | 201 | 214 | 189 | 209 | 457 | 486 | 630 | 351 | 90 | 546 | **107** | 499 | **280** |
| $\hat{\theta}_2^{N,SL}$ | 216 | 121 | 344 | 128 | 399 | 408 | 467 | 324 | 190 | 499 | 115 | 617 | 347 |
| $\hat{\theta}_2^{B,SL}$ | 215 | 256 | 188 | 234 | 383 | 419 | 501 | 336 | **79** | 448 | 170 | 534 | 328 |
| $\hat{\theta}_2^{N,DJ}$ | 181 | 313 | 137 | 285 | 329 | **291** | **415** | **239** | 197 | 423 | 153 | 392 | 482 |
| $\hat{\theta}_2^{B,DJ}$ | 204 | 322 | 152 | 296 | 333 | 322 | 438 | 247 | 168 | **378** | 154 | 417 | 465 |
| | $n = 5000$ | | | | | | | | | | | | |
| $\hat{\theta}^{MR}$ | 342 | 467 | **335** | 441 | **-620** | 806 | 1166 | **283** | 382 | 820 | 349 | **745** | 1558 |
| $\hat{\theta}_1^{N,SL}$ | 294 | **311** | 349 | **295** | 848 | 785 | 1329 | 320 | **237** | 959 | 243 | 818 | 1345 |
| $\hat{\theta}_1^{B,SL}$ | 569 | 594 | 536 | 587 | 1006 | 1132 | 1474 | 359 | 284 | **519** | 376 | 963 | 1348 |
| $\hat{\theta}_1^{N,DJ}$ | 300 | 319 | 354 | 303 | 858 | 807 | 1337 | 313 | 240 | 956 | **170** | 823 | **1319** |
| $\hat{\theta}_1^{B,DJ}$ | 531 | 542 | 514 | 538 | 967 | 1071 | 1399 | 331 | 261 | 566 | 359 | 913 | 1351 |
| $\hat{\theta}_2^{N,SL}$ | **291** | 400 | 371 | 358 | 830 | **743** | 1378 | 393 | 276 | 898 | 279 | 1381 | 1397 |
| $\hat{\theta}_2^{B,SL}$ | 626 | 652 | 612 | 628 | 990 | 1048 | 1681 | 395 | 330 | 525 | 460 | 1193 | 1772 |
| $\hat{\theta}_2^{N,DJ}$ | 777 | 1057 | 675 | 1005 | 1110 | 876 | **1075** | 946 | 719 | 1010 | 732 | 894 | 2261 |
| $\hat{\theta}_2^{B,DJ}$ | 783 | 1037 | 697 | 991 | 1119 | 882 | 1074 | 928 | 713 | 983 | 747 | 933 | 2228 |

### 3.4 Notations

From tables 1 and 2, we can conclude that, in general, $\hat{\theta}_1^N$ gives us more satisfactory results than $\hat{\theta}_1^B$. The usage of sliding blocks for some processes has better accuracy than the usage of disjoint blocks, but for some it is worse, so we cannot make a common conclusion.

From the use of the discrepancy method for selection $b_n$ and $u_n$, we can infer that $\hat{\theta}^{MR}$ and $\hat{\theta}^{N,SL}$ has almost the same accuracy.

## References


[1] M.R. Leadbetter, G. Lindgren, H. Rootzén *Extremes and related properties of random sequence and processes*, ch.3. Springer-Verlag, New-York, 1983.

[2] C.A.T. Ferro, J. Segers *Inference for clusters of extreme values*. Journal of the Royal Statistical Society Series B. 65 545–556, 2003.

[3] P. J. Northrop *An efficient semiparametric maxima estimator of the extremal index*. Extremes 18:4 585–603., 2015.

[4] B. Berghaus, A. Bücher *Weak convergence of a pseudo maximum likelihood estimator for the extremal index*. The Annals of Statistics 46(5) 2307–2335, 2018.

[5] N.M. Markovich *Nonparametric estimation of extremal index using discrepancy method*. In: Proceedings of the X International conference "System identification and control problems" SICPRO–2015 Moscow January 26–29, V.A. Trapeznikov Institute of Control Sciences. 160–168. ISBN 978-5- 91450-162-1, 2015.

[6] N.M. Markovich *Experimental analysis of nonparametric distribution density estimates and methods of their smoothing*. Automation and Remote Control 50 941–948, 1989.

[7] L.N. Bolshev, N.V. Smirnov *Tables of Mathematical Statistics* Nauka, Moscow, 1965 (in Russian).

[8] N.M. Markovich, I.V.Rodionov *Threshold selection for extremal index estimation.* Electronic Journal of Statistics, 2020.

[9] M.A. Ancona–Navarrete, J. A Tawn *Comparison of Methods for Estimating the Extremal Index.* Extremes, 3:1 5–38, 2000.

[10] J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, *Statistics of Extremes: Theory and Applications.* Wiley, Chichester, West Sus- sex, 2004.

[11] M.R. Chernick, T. Hsing, W.P. McCormick *Calculating the extremal index for a class of stationary sequences.* Advances in Applied Probability,23 835–850, 1991.

[12] M. Ferreira. *Multiple thresholds in extremal parameter estimation.* Extremes, 22 317–341, 2019.

[13] J.Sun, G. Samorodnitsky. *Analysis of estimation methods for the extremal index.* Electronic Journal of Applied Statistical Analysis, 11:1 296–306, 2018.

[14] M. Süveges, A.C. Davison. *Model misspecification in peaks over threshold analysis.* The Annals of Applied Statistics, 4:1 203–221,2010.

[15] F. Laurini, J.A Tawn. *The extremal index for GARCH(1,1) processes.* Extremes 15, 511–529, 2012.