




On Threshold Selection Problem for Extremal Index Estimation

Igor Rodionov^(✉) 

Trapeznikov Institute of Control Sciences of RAS, Moscow, Russian Federation

Abstract. We study the properties of the new threshold selection method for non-parametric estimation of the extremal index of a stationary sequence proposed in [15]. The method is to apply the so-called discrepancy method based on the Cramér–von Mises–Smirnov’s statistic calculated by the largest order statistics of a sample. The limit distribution of this statistic is derived if the proportion of the largest order statistics used tends to some nonzero constant. We also use the non-standard modification of the Cramér–von Mises–Smirnov’s statistic to propose the goodness-of-fit test procedure of ω^2 type for distribution tails.

Keywords: Extremal index · Threshold selection · Discrepancy method · Stationary sequence · Cramér–von Mises–Smirnov’s statistic · Goodness-of-fit · Distribution tail

1 Introduction

Let (X_n) be a strictly stationary sequence with common cumulative distribution function (cdf) F . Introduce the following

Definition 1 [11]. *The sequence (X_n) is said to have the extremal index $\theta \in [0, 1]$, if for each $\tau > 0$ there exists a sequence of real numbers $u_n = u_n(\tau)$ such that*

$$\lim_{n \rightarrow \infty} n(1 - F(u_n)) = \tau, \quad \lim_{n \rightarrow \infty} P(M_n \leq u_n) = e^{-\theta\tau},$$

where $M_n = \max(X_1, \dots, X_n)$.

The extremal index exists for a wide class of stationary sequences and reflects a cluster structure of an underlying sequence and its local dependence properties. The extremal index of an independent sequence is equal to 1, and the opposite is not true. In particular, if the Berman’s condition holds, then a Gaussian stationary sequence has the extremal index equal to 1, [11].

We consider the problem of non-parametric estimation of the extremal index. It is important to note that absolutely all non-parametric extremal index estimators require the selection of the threshold parameter u and/or the block size

The author was partly supported by the Russian Foundation for Basic Research (grant No. 19-01-00090) and by Young Russian Mathematics Award.

b or other declustering parameter. The well-known blocks estimator [1] and its later modifications [19] depend of the choice of both b and u . Another classical one, the runs estimator [24], requires the selection of u and the number of consecutive observations r running below u separating two clusters. The intervals [8] and K-gaps [21] estimators require the choice of u only, whereas the sliding blocks estimators proposed in [16] and simplified in [2] depend on the block size only.

Less attention in the literature is devoted to methods of selection of the mentioned parameters, in particular, the threshold parameter u . Usually, the value of u is taken from high quantiles of an underlying sequence (X_n) or selected visually corresponding to a stability plot of values of some estimate $\hat{\theta}(u)$ with respect to u . Fukutome et al. [9] proposed the procedure of selection among pairs (u, K) for the K-gaps estimator based on Information Matrix Test.

Markovich and Rodionov [15] proposed the non-parametric tool to select one of the necessary parameters for extremal index estimation. Although the proposed method can be applied for selection of arbitrary aforementioned parameter, they focused on the selection of a threshold parameter u . The developed method is an automatic procedure of extremal index estimation in cases if it is based on estimators requiring the choice of only one parameter, in particular, the intervals and K-gaps estimators. But in [15] this procedure was established only if the proportion of the largest order statistics of a sample used vanishes as $n \rightarrow \infty$ (more precisely, see Theorem 3.3 [15]). The aim of this work is to investigate the opportunity of justification of the Markovich and Rodionov's method if the mentioned proportion tends to some positive constant c . The problem of goodness-of-fit testing of distribution tails is also studied.

2 Preliminaries

2.1 Inter-exceedance Times and Their Asymptotic Behavior

Let us discuss the properties of a stationary sequence (X_n) and its extremal index θ . Let L be the number of exceedances of level u by the sequence $(X_i)_{i=1}^n$ and $S_j(u)$ be the j -th exceedance time, that is,

$$S_j(u) = \min\{k > S_{j-1}(u) : X_k > u\}, \quad j = 1, \dots, L,$$

where $S_0 = 0$. Define the *inter-exceedance times* as

$$T_j(u) = S_{j+1}(u) - S_j(u), \quad j = 1, \dots, L - 1$$

and assume its number equal to L for convenience.

Introduce the following φ -mixing condition.

Definition 2 [8]. For real u and integers $1 \leq k \leq l$, let $\mathcal{F}_{k,l}(u)$ be the σ -field, generated by $\{X_i > u\}$, $k \leq i \leq l$. Introduce 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 the supremum is taken over all sets $A \in \mathcal{F}_{1,k}(u)$ with $P(A) > 0$ and $B \in \mathcal{F}_{k+l,n}(u)$.

The next theorem states that for some sequence of levels (u_n) it holds

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

where η is exponential with mean θ^{-1} .

Theorem 1 [8]. *Let the positive integers (r_n) and the thresholds (u_n) , $n \geq 1$, be such that $r_n \rightarrow \infty$, $r_n \overline{F}(u_n) \rightarrow \tau$ and $P\{M_{r_n} \leq u_n\} \rightarrow \exp(-\theta\tau)$ as $n \rightarrow \infty$ 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 all $c > 0$, then for $t > 0$*

$$P\{\overline{F}(u_n)T_1(u_n) > t\} \rightarrow \theta \exp(-\theta t) =: 1 - F_\theta(t), \quad n \rightarrow \infty. \quad (1)$$

The well-known intervals estimator of the extremal index is based on inter-exceedance times and is found via method of moments applied to the limit distribution (1). It is defined as ([8], see also [1], p. 391),

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

where

$$\hat{\theta}_n^1(u_n) = \frac{2 \left(\sum_{i=1}^L T_i(u_n) \right)^2}{L \sum_{i=1}^L (T_i(u_n))^2} \quad \text{and} \quad \hat{\theta}_n^2(u_n) = \frac{2 \left(\sum_{i=1}^L (T_i(u_n) - 1) \right)^2}{L \sum_{i=1}^L (T_i(u_n) - 1)(T_i(u_n) - 2)}.$$

It is known that

$$\sqrt{L}(\hat{\theta}_n(u_n) - \theta) \xrightarrow{d} N(0, \theta^3 v(\theta)), \quad (3)$$

where $v(\theta)$ is the second moment of the cluster size distribution $\{\pi(m)\}_{m \geq 1}$, [18]. Moreover, Theorem 2.4 [18] states that non-zero elements of the sequence

$$Z_i = \overline{F}(u_n)T_i(u_n), \quad i = 1, \dots, L, \quad (4)$$

are asymptotically independent under the assumptions of Theorem 1 and some assumptions on the cluster structure of the initial stationary sequence (X_n) .

To be able to use these properties, Markovich and Rodionov [15] assume that there exists a sequence $(E_i)_{i=1}^l$, $l = [\theta L]$, of independent exponentially distributed random variables with mean θ^{-1} such that

$$Z_{(L-k)} - E_{(l-k)} = o\left(\frac{1}{\sqrt{k}}\right) \quad (5)$$

uniformly for all $k \rightarrow \infty$ with $k/L \rightarrow 0$ as $L \rightarrow \infty$. This assumption is based on the following reasoning. It follows from Theorem 3.2 [18], that the limit distribution of the statistic

$$\sqrt{L} \left(\sum_{i=1}^L f(Z_i) - Ef(Z_1) \right)$$

for some class of continuous f does not depend on substitution of the set of r.v.s $\{Z_i^*\}_{i=1}^L$ with cdf F_θ instead of $\{Z_i\}_{i=1}^L$ under some regularity conditions. Moreover, for these r.v.s Theorem 2.2.1 and Lemma 2.2.3 [7] imply that if $k/L \rightarrow c$, $c \in [0, \theta)$ as $k \rightarrow \infty$, $L \rightarrow \infty$, then

$$\sqrt{k}(E_{(l-k)} - \ln(l/k)/\theta) = O_P(1).$$

2.2 Discrepancy Method

The method proposed in [15] is based on the so-called discrepancy method initially introduced in [12] and [22], see also [13], for optimal bandwidth selection in the problem of density estimation and applied at the first time for extremal index estimation in [14]. Let $\rho(\cdot, \cdot)$ be some distance on the space of probability measures, \hat{F}_n be the empirical cdf of the sequence $(X_i)_{i=1}^n$ and $\{F_u, u \in U\}$ be the family of cdfs parametrized by one-dimensional parameter u . Then the optimal value of u can be found as a solution of the discrepancy equation

$$\rho(\hat{F}_n, F_u) = \delta, \quad (6)$$

where δ , the so-called discrepancy value, is defined by the choice of ρ . The statistic of the Cramér-von Mises-Smirnov goodness-of-fit test (CMS statistic)

$$\omega_n^2 = n \int_{\mathbb{R}} (\hat{F}_n(x) - F_0(x))^2 dF(x)$$

was chosen as ρ in [15], though the statistics of other goodness-of-fit tests, e.g., the Kolmogorov and Anderson-Darling tests, can be applied in the discussing problem. Then quantiles of the limit distribution of the CMS statistic can be used as δ . The choice of the parameter u as $\hat{u} = \operatorname{argmin}_u \rho(\hat{F}_n, F_u)$ is usually not optimal.

3 Main Results

In this section we consider the problem of the discrepancy method application to choose the threshold/block-size parameter of the extremal index estimator. To simplify, let us assume that we choose the threshold parameter u . It seems that for this purpose one can take $F_u = T_{\hat{\theta}(u)}$, \hat{F}_n be equal to the empirical cdf of the sequence $\{Z_i\}_{i=1}^L$ and ρ be equal to the ω^2 distance in (6). However, we cannot directly apply the discrepancy method coupling with the ω^2 distance to this problem since T_θ is not a continuous distribution and thus the limit distribution of the CMS statistic would depend on θ . To overcome this difficulty, we introduce the modification of the CMS statistic based only on the largest order statistics corresponding to $\{Z_i\}_{i=1}^L$, since, as was mentioned in Sect. 2.1, the largest elements of this sequence are continuously distributed and asymptotically independent. Thus we face the problem of goodness-of-fit testing of left-censored

data and, in particular, distribution tails, see [20] for the principles of testing of distribution tails.

Let $(Y_i)_{i=1}^n$ be independent identically distributed random variables with common continuous cdf F_Y . Recall that if the hypothesis $H_0 : F_Y = F_0$ for continuous cdf F_0 holds, then the CMS statistic can be rewritten as

$$\omega_n^2 = \sum_{i=1}^n \left(F_0(Y_{(i)}) - \frac{i - 0.5}{n} \right)^2 + \frac{1}{12n},$$

where $Y_{(1)} \leq \dots \leq Y_{(n)}$ are the order statistics corresponding to $(Y_i)_{i=1}^n$. It is well-known that the limit distribution of ω_n^2 (denote its cdf as A_1) under H_0 does not depend on F_0 .

Goodness-of-fit procedures for various types of left- and right-censored data were proposed in a large number of works, we refer to the classical monograph [4] and recent monograph [23]. But to the best of author's knowledge, there are no works in the literature proposing the modifications of goodness-of-fit statistics for censored data having the same limit distribution as their full-sample analogues. Introduce the following modification of the CMS statistic

$$\hat{\omega}_k^2 = \sum_{i=0}^{k-1} \left(\frac{F_0(Y_{(n-i)}) - F_0(Y_{(n-k)})}{1 - F_0(Y_{(n-k)})} - \frac{k - i - 0.5}{k} \right)^2 + \frac{1}{12k}.$$

Theorem 2. *Let the hypothesis $H_0^t : \{F_Y(x) = F_0(x) \text{ for all large } x\}$ holds. Then there is c such that*

$$\hat{\omega}_k^2 \xrightarrow{d} \xi \sim A_1$$

as $k \rightarrow \infty$, $k/n < c$, $n \rightarrow \infty$.

Theorem 3.1, [15], is a particular case of the latter theorem for $F_0 = F_\theta$, where F_θ is defined in (1). It is worth noting that there is no necessity to require the continuity of F_0 for all real x ; we need this only for all sufficiently large x . Theorem 2 allows us to propose the goodness-of-fit test for continuous distribution tail of significance level α in the following way

$$\text{if } \hat{\omega}_k^2 > a_{1-\alpha}, \text{ then reject } H_0^t,$$

where $a_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of A_1 . Additionally, one can show that this test is consistent as $k \rightarrow \infty$, $k/n \rightarrow 0$, $n \rightarrow \infty$. Clear, only the largest order statistics of a sample can be used for testing the distribution tail hypotheses both if $k/n \rightarrow 0$ and if $k/n < c$. This problem is reasonable both if only the upper tail of the distribution is of interest and/or if only the largest order statistics of a sample are available.

Let us return to the problem of extremal index estimation. Consider

$$\tilde{\omega}_L^2(\theta) = \sum_{i=0}^{k-1} \left(\frac{F_\theta(Z_{(L-i)}) - F_\theta(Z_{(L-k)})}{1 - F_\theta(Z_{(L-k)})} - \frac{k - i - 0.5}{k} \right)^2 + \frac{1}{12k}.$$

The following theorem states that under some mild conditions the statistic $\tilde{\omega}_L^2(\theta)$ with some estimator $\hat{\theta}_n$ substituted for θ has the same limit distribution as the statistic $\hat{\omega}_k^2$ in Theorem 2 and the classical CMS statistic.

Theorem 3. [15]. *Let the assumptions of Theorem 1 and the condition (5) hold. Assume the extremal index estimator $\hat{\theta} = \hat{\theta}_n$ is such that*

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

where ζ has a non-degenerate cdf H . Assume the sequence of integers (m_n) is such that

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

as $n \rightarrow \infty$. Then

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

Remark 1. All extremal index estimators mentioned in Introduction satisfy the assumptions of Theorem 3 with H equal to the normal cdf with zero mean.

Theorem 3 guarantees the correctness of the discrepancy method

$$\tilde{\omega}_L^2(\hat{\theta}_n) = \delta, \tag{7}$$

where δ can be selected equal to 0.05, the mode of A_1 , and $k/L \rightarrow 0$ as $k \rightarrow \infty$ and $L \rightarrow \infty$. The simulation study provided in [15] shows that $u_{\max} = \max\{u_1, \dots, u_d\}$ is the best choice for threshold parameter both for the intervals and the K-gaps estimators of the extremal index, where $\{u_1, \dots, u_d\}$ are solutions of the discrepancy equation (7). The numerical comparison of the proposed method with other methods of threshold selection shows the significant advantage of the developed procedure on a wide class of model processes, see [15] for details. Although the limit distribution of the statistic $\tilde{\omega}_L^2(\hat{\theta}_n)$ does not depend on k , the selection of k for samples of moderate sizes remains a problem. The choice $k = \min(\hat{\theta}_0 L, L^\beta)$ with $\beta \in (0, 1)$, where $\hat{\theta}_0$ is some pilot estimate, has proven by simulation study to be the most suitable. But in case of $k/L \rightarrow c > 0$ as $k \rightarrow \infty$, $L \rightarrow \infty$ the distribution of $\hat{\theta}_n$ affects the limit distribution of the modified CMS statistic $\tilde{\omega}_L^2(\hat{\theta}_n)$ in contrast to the case $c = 0$, thus this limit distribution would differ from A_1 .

The asymptotic distributions of goodness-of-fit test statistics with parameters of an underlying distribution being estimated were intensively studied in the literature. The starting point for this classical theory was in works [5] and [10], whereas the common method to derive the limit distribution was proposed in [6]. However, this method based on a multivariate central limit theorem and convergence in the Skorokhod space cannot be directly applied to the problem of evaluating the limit distribution of the statistic $\tilde{\omega}_L^2(\hat{\theta}_n)$ when the assumption $k/m_n = o(1)$ does not hold ($m_n = O(L)$ for the intervals and K-gaps estimators, thus we can talk about the case $k/L = o(1)$). For this purpose we consider

another modification of the CMS statistic, the first analogue of which was introduced in [17],

$$\omega_{L,c}^2(\theta) = L \int_{x(c)}^{\infty} (F_L^*(x) - F_\theta(x))^2 dF_\theta(x), \quad (8)$$

where $F_L^*(x)$ is the empirical distribution function of the sequence $\{Z_i\}_{i \leq L}$ and $x(c) = \inf\{x : F_\theta(x) \geq 1 - c\} =: F_\theta^{*-}(1 - c)$, $c \in (0, 1)$. In the sequel, we will assume $0 < c < \theta$, therefore $F_\theta(x(c)) = 1 - c$. It follows from the results derived in [17] and the assumption (5) that the statistic $\omega_{L,c}^2$ converges in distribution to $\omega^2(c)$, where

$$\omega^2(c) = \int_{1-c}^1 B^2(t) dt$$

and $B(t)$ is the standard Brownian bridge, i.e. the Gaussian process on the interval $[0, 1]$ with mean zero and covariance function $\text{cov}(B_t, B_s) = \min(t, s) - ts$.

Denote $\ell_c(t) = \max(t - (1 - c), 0)$. Following the ideas of [6] we introduce the sample process

$$y_{L,c}(t, \theta) = \sqrt{L} \left(\hat{F}_{L,c}(t, \theta) - \ell_c(t) \right), \quad t \in [0, 1],$$

where

$$\hat{F}_{L,c}(t, \theta) = \frac{1}{L} \sum_{i=1}^L I(1 - c < F_\theta(Z_i) \leq t),$$

call it the truncated empirical distribution function of the sequence $\{Z_i\}_{i \leq L}$. Clear, since $\theta > c$ it holds

$$\int_0^1 y_{L,c}^2(t, \theta) dt = \omega_{L,c}^2(\theta).$$

Denote D , the Skorokhod space, i.e. the space of right-continuous functions with left-hand limits on $[0, 1]$ and metric $d(x, y)$ (see, e.g., [3], p. 111). The following theorem allows us to find the asymptotic distribution of the statistic $\omega_{L,c}^2(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the intervals estimator (2).

Theorem 4. *Let the sequence $\{Z_i\}$ defined by (4) satisfies the assumptions of Theorem 3.2 [18]. Assume $\theta > 0$. For every $c \in (0, \theta)$ the estimated sample process $\hat{y}_c(t) := y_{L,c}(t, \hat{\theta}_n)$ converges weakly in D as $n \rightarrow \infty$ to the Gaussian process $X(t)$, $t \in [1 - c, 1]$, with mean zero and covariance function*

$$C(t, s) = \ell_c(\min(t, s)) - (2 - 2/\theta)\ell_c(t)\ell_c(s) - \frac{1}{2\theta^2}h_c(t)(2h_c(s) + \tilde{h}_c(s)) - \frac{1}{2\theta^2}h_c(s)(2h_c(t) + \tilde{h}_c(t)) + \frac{v(\theta)}{\theta}h_c(t)h_c(s), \quad (9)$$

where

$$h_c(t) = (1-t) \log \left(\frac{1-t}{\theta} \right) - c \log \left(\frac{c}{\theta} \right), \quad \tilde{h}_c(t) = (1-t) \log^2 \left(\frac{1-t}{\theta} \right) - c \log^2 \left(\frac{c}{\theta} \right)$$

and $v(\theta)$ is defined in (3).

We see that the covariance function of the process $X(t)$ depends on θ and c . This fact makes the usage of quantiles of the limit distribution of the statistic $\omega_{L,c}^2(\hat{\theta}_n)$ (or some its appropriate normalization) as δ in the discrepancy method (7) quite inconvenient in practice. However, high efficiency of the discrepancy method apparently means that the values of the mentioned quantiles do not strongly depend on the values of θ and c .

4 Proofs

4.1 Proof of Theorem 2

To prove Theorem 2, we need the following

Lemma 1 (Lemma 3.4.1, [7]). *Let X, X_1, X_2, \dots, X_n be i.i.d. random variables with common cdf F , and let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the n th order statistics. The joint distribution of $\{X_{(i)}\}_{i=n-k+1}^n$ given $X_{(n-k)} = t$, for some $k = 1, \dots, n-1$, equals the joint distribution of the set of order statistics $\{X_{(i)}^*\}_{i=1}^k$ of i.i.d. r.v.s $\{X_i^*\}_{i=1}^k$ with cdf*

$$F_t(x) = P\{X \leq x | X > t\} = \frac{F(x) - F(t)}{1 - F(t)}, \quad x > t.$$

Assume $F_Y(x) = F_0(x)$ for all $x > x_0$ and set $c = 1 - F_0(x_0) - \varepsilon$ for some small $\varepsilon > 0$. Clear, since $k/n < c$ then $P(Y_{(n-k)} > x_0) \rightarrow 1$ under the assumptions.

Consider the conditional distribution of $\hat{\omega}_k^2$ given $F_0(Y_{(n-k)}) = t$, $t > 1 - c$. By Lemma 1 and the assumption $k/n < c$, the conditional joint distribution of the set of order statistics $\{F_0(Y_{(i)})\}_{i=n-k+1}^n$ coincides with the joint distribution of the set of order statistics $\{U_{(i)}^*\}_{i=1}^k$ of a sample $\{U_i^*\}_{i=1}^k$ from the uniform distribution on $[t, 1]$. Therefore, it holds

$$\hat{\omega}_k^2 \stackrel{d}{=} \frac{1}{(1-t)^2} \left(\sum_{i=1}^k \left(U_{(i)}^* - t - \frac{i-0.5}{k}(1-t) \right)^2 \right) + \frac{1}{12k}.$$

Next, $V_{(i)}^* = U_{(i)}^* - t$, $1 \leq i \leq k$, are the order statistics of a sample $\{V_i^*\}_{i=1}^k$ from the uniform distribution on $[0, 1-t]$. Hence, it follows

$$\hat{\omega}_k^2 \stackrel{d}{=} \frac{1}{(1-t)^2} \left(\sum_{i=1}^k \left(V_{(i)}^* - \frac{i-0.5}{k}(1-t) \right)^2 \right) + \frac{1}{12k}.$$

Finally, $W_{(i)}^* = V_{(i)}^*/(1-t)$, $1 \leq i \leq k$, are the order statistics of a sample $\{W_i^*\}_{i=1}^k$ from the uniform distribution on $[0, 1]$. Therefore, we get

$$\hat{\omega}_k^2 \stackrel{d}{=} \sum_{i=1}^k \left(W_{(i)}^* - \frac{i-0.5}{k} \right)^2 + \frac{1}{12k}.$$

It is easy to see, that the latter expression is exactly the CMS statistic and converges in distribution to a random variable ξ with cdf A_1 independently of the value of t .

4.2 Proof of Theorem 4

Assume for convenience that $\hat{\theta}_n = \hat{\theta}_n^2(u_n)$, where $\hat{\theta}_n^2(u_n)$ is defined by (2). For shortening, write $\hat{\theta}$ instead of $\hat{\theta}_n$. We need the following

Lemma 2. *Under the assumptions*

$$\sqrt{L}(\hat{\theta} - \theta) = \frac{\theta}{\sqrt{L}} \sum_{i=1}^L \left(2Z_i - \frac{\theta}{2} Z_i^2 - 1 \right) + o_P(1),$$

where $o_P(1)$ denotes a sequence of random variables vanishing in probability.

Proof. (of Lemma 2)

Observe that

$$n - \sum_{i=1}^L T_i(u_n) \leq T_{L+1}(u_n) \stackrel{d}{=} T_1(u_n),$$

where the last relation holds by stationarity and definition of inter-exceedance times. Denote $r_n = n\bar{F}(u_n)$. Theorem 3.2 [18] implies that

$$\sqrt{r_n}(L/r_n - 1) \xrightarrow{d} N(0, \theta v(\theta)), \quad (10)$$

thus for all $\varepsilon > 0$

$$\begin{aligned} P\left(\sqrt{r_n}\left(1 - \frac{1}{r_n} \sum_{i=1}^L Z_i\right) > \varepsilon\right) &= P\left(\sqrt{r_n} \frac{\bar{F}(u_n)}{r_n} \left(n - \sum_{i=1}^L T_i(u_n)\right) > \varepsilon\right) \\ &\leq P\left(\frac{\bar{F}(u_n)}{\sqrt{r_n}} T_{L+1}(u_n) > \varepsilon\right) = P\left(\frac{1}{\sqrt{r_n}} Z_1 > \varepsilon\right) \rightarrow 0, \end{aligned} \quad (11)$$

where the last relation holds by Theorem 1. Note that by (10) $L/r_n \xrightarrow{P} 1$, thus

$$\sqrt{L}(\hat{\theta} - \theta) = \sqrt{r_n}(\hat{\theta} - \theta) + o_P(1).$$

The latter relations imply that

$$\sqrt{L}(\hat{\theta} - \theta) = \sqrt{r_n} \left(\frac{2r_n/L}{\frac{1}{r_n} \sum_{i=1}^L Z_i^2} - \theta \right) + o_P(1) = \theta \sqrt{r_n} \left(\frac{2r_n/(\theta L) - \frac{1}{r_n} \sum_{i=1}^L Z_i^2}{\frac{1}{r_n} \sum_{i=1}^L Z_i^2} \right) + o_P(1).$$

Next, it follows from Lemma B.7 [18] that

$$\frac{1}{r_n} \sum_{i=1}^L Z_i^2 \xrightarrow{P} \frac{2}{\theta},$$

therefore

$$\sqrt{L}(\hat{\theta} - \theta) = \frac{\theta^2}{2} \sqrt{r_n} \left(\frac{2r_n}{\theta L} - \frac{1}{r_n} \sum_{i=1}^L Z_i^2 \right) + o_P(1).$$

It immediately follows from (10) and the delta method that

$$\sqrt{r_n}(r_n/L - 1) = \sqrt{r_n}(L/r_n - 1) + o_P(1).$$

Finally, we obtain from (11) and the latter

$$\begin{aligned}
\sqrt{r_n} \left(\frac{2r_n}{\theta L} - \frac{1}{r_n} \sum_{i=1}^L Z_i^2 \right) &= \sqrt{r_n} \left(\frac{4}{\theta} - \frac{2L}{\theta r_n} - \frac{1}{r_n} \sum_{i=1}^L Z_i^2 \right) + o_P(1) \\
&= \frac{4}{\theta} \sqrt{r_n} \left(1 - \frac{1}{r_n} \sum_{i=1}^L Z_i \right) \\
&\quad + \frac{2}{\theta} \sqrt{r_n} \left(2 \frac{1}{r_n} \sum_{i=1}^L Z_i - \frac{1}{r_n} \sum_{i=1}^L 1 - \frac{\theta}{2r_n} \sum_{i=1}^L Z_i^2 \right) + o_P(1) \\
&= \frac{2}{\sqrt{r_n} \theta} \sum_{i=1}^L \left(2Z_i - \frac{\theta}{2} Z_i^2 - 1 \right) + o_P(1).
\end{aligned}$$

Applying again $L/r_n \xrightarrow{P} 1$, we derive the result.

First of all, we observe that by (11)

$$\begin{aligned}
y_{L,c}(t, \theta) &= \sqrt{L} \left(\frac{1}{L} \sum_{i=1}^L I(1-c < F_\theta(Z_i) \leq t) - \ell_c(t) \right) \\
&= \frac{1}{\sqrt{L}} \sum_{i=1}^L (I(1-c < F_\theta(Z_i) \leq t) - \ell_c(t) Z_i) + \ell_c(t) \sqrt{L} \left(\frac{1}{L} \sum_{i=1}^L Z_i - 1 \right) \\
&= \frac{1}{\sqrt{L}} \sum_{i=1}^L (I(1-c < F_\theta(Z_i) \leq t) - \ell_c(t) Z_i) + o_P(1).
\end{aligned} \tag{12}$$

It is also worth noting that we can change all expressions of the form $\frac{1}{\sqrt{L}} \sum_{i=1}^L \xi_i$ appearing in our proof with $E\xi_i = 0$ for all i on expressions of the form $\frac{1}{\sqrt{r_n}} \sum_{i=1}^{r_n} \xi_i$ using the same argument as in the proof of Theorem 3.2 [18] based on the formula (10). It means that the randomness of L does not affect the asymptotic of $\hat{y}_c(t)$.

We follow the ideas of the proof of Theorem 2 in [6]. For shortening, write $\hat{\theta}$ instead of $\hat{\theta}_n$. Denote

$$\hat{t}(t) = F_{\hat{\theta}}(F_{\hat{\theta}}^{\leftarrow}(t)), \quad t \in [1-c, 1].$$

Since $\hat{\theta}$ is a consistent estimator of θ , see [18], we have

$$P(\hat{t}(t) \geq 1-c) \rightarrow 1 \tag{13}$$

for all $t \in (1-c, 1]$, and for $t = 1-c$ the latter probability tends to 1/2. First, we will show that

$$y_{L,c}(\hat{t}(t), \theta) - y_{L,c}(t, \theta) \xrightarrow{P} 0 \tag{14}$$

uniformly for $t \in [1-c, 1]$. We restrict ourselves to the study of the case $t \in (1-c, 1]$, the case $t = 1-c$ is similar. Denote $x_1(t) = F_{\theta_1}^{\leftarrow}(t)$ and $\tilde{t}(t) = F_{\theta_2}(x_1(t))$ for $\theta_1, \theta_2 \geq c$ and $t \in (1-c, 1]$. We have

$$\begin{aligned}
\sup_{t \in (1-c, 1]} |\tilde{t} - t| &\leq \sup_{t \in (1-c, 1]} |F_{\theta_2}(x_1(t)) - F_{\theta_1}(x_1(t))| \\
&= \sup_{x > F_{\theta_1}^{\leftarrow}(1-c)} \left| (\theta_2 - \theta_1) \frac{\partial F_\theta(x)}{\partial \theta} \Big|_{\theta=\theta^*} \right|,
\end{aligned}$$

where θ^* is between θ_1 and θ_2 . Since

$$\left| \frac{\partial F_\theta(x)}{\partial \theta} \right| = |(\theta^2 - 1)e^{-\theta x}| \leq 1$$

for all $\theta \in [0, 1]$ and $x > F_{\theta_1}^-(1 - c)$, we derive that \tilde{t} converges uniformly to t as $\theta_1 \rightarrow \theta$ and $\theta_2 \rightarrow \theta$. Therefore, since $\hat{\theta}$ converges to θ in probability,

$$\sup_{t \in (1-c, 1]} |\hat{t}(t) - t| \xrightarrow{P} 0.$$

An appeal to Lemma B.7 [18] gives us that $y_{L,c}(t, \theta) \xrightarrow{d} y(t)$, $t \in (1 - c, 1]$, in D where $y(t)$ is the Gaussian random process with mean zero and covariance function

$$\text{cov}(y(t), y(s)) = \ell_c(\min(t, s)) - (2 - 2/\theta)\ell_c(t)\ell_c(s).$$

The rest of the proof of (14) coincides with the corresponding steps in the proof of Lemma 1 [6].

Now let us show that

$$\hat{y}_c(t) = y_{L,c}(t, \theta) - \sqrt{L}(\hat{\theta} - \theta)(g(t, \theta) - g(1 - c, \theta)) + o_P(1), \quad t \in [1 - c, 1], \quad (15)$$

where

$$g(t, \theta) = \frac{1 - t}{\theta^2} \log \left(\frac{1 - t}{\theta} \right).$$

Note that by definition $\hat{y}_c(1 - c) = 0$ a.s., thus it remains to show (15) for $t \in (1 - c, 1]$. First we find the explicit form of the “estimated” empirical cdf $\hat{F}_{L,c}(\hat{t}(t), \theta)$. We have

$$\begin{aligned} \hat{F}_{L,c}(\hat{t}(t), \theta) &= \frac{1}{L} \sum_{i=1}^L I \left(1 - c \leq F_\theta(Z_i) < F_\theta(F_\theta^-(t)) \right) \\ &= \frac{1}{L} \sum_{i=1}^L I \left(F_\theta(F_\theta^-(1 - c)) < F_\theta(Z_i) \leq t \right) \\ &= F_{L,c}(t, \hat{\theta}) - \frac{1}{L} \sum_{i=1}^L I \left(1 - c \leq F_{\hat{\theta}}(Z_i) < F_{\hat{\theta}}(F_{\hat{\theta}}^-(1 - c)) \right) \\ &= F_{L,c}(t, \hat{\theta}) - \frac{1}{L} \sum_{i=1}^L I \left(\hat{t}(1 - c) < F_\theta(Z_i) \leq 1 - c \right). \end{aligned}$$

Denote

$$\tilde{F}_{L,c}(t, \theta) = \frac{1}{L} \sum_{i=1}^L I(t < F_\theta(Z_i) \leq 1 - c), \quad t \leq 1 - c.$$

Therefore we derive for the estimated sample process $\hat{y}_c(t)$

$$\begin{aligned}\hat{y}_c(t) &= \sqrt{L}(F_{L,c}(t, \hat{\theta}) - \ell_c(t)) \\ &= \sqrt{L}(F_{L,c}(\hat{t}(t), \theta) - \ell_c(\hat{t}(t))) + \sqrt{L}(\ell_c(\hat{t}(t)) - \ell_c(t)) + \sqrt{L}\tilde{F}_{L,c}(\hat{t}(1-c), \theta) \\ &= y_{L,c}(\hat{t}(t), \theta) + \sqrt{L}(\ell_c(\hat{t}(t)) - \ell_c(t)) + \sqrt{L}\tilde{F}_{L,c}(\hat{t}(1-c), \theta).\end{aligned}$$

Consider the third summand on the right-hand side. Fix $c_1 \in (c, \theta)$. Note that (14) remains true for all $c_1 < \theta$, thus we derive

$$y_{L,c_1}(\hat{t}(1-c), \theta) - y_{L,c_1}(1-c, \theta) \xrightarrow{P} 0.$$

On the other hand,

$$\begin{aligned}y_{L,c_1}(\hat{t}(1-c), \theta) - y_{L,c_1}(1-c, \theta) &= \sqrt{L}(F_{L,c_1}(\hat{t}(1-c), \hat{\theta}) - \ell_{c_1}(\hat{t}(1-c))) - \sqrt{L}(F_{L,c_1}(1-c, \hat{\theta}) - \ell_{c_1}(1-c)) \\ &= \sqrt{L}(1-c - \hat{t}(1-c)) - \sqrt{L}\tilde{F}_{L,c}(\hat{t}(1-c), \theta),\end{aligned}$$

therefore we derive

$$\sqrt{L}\tilde{F}_{L,c}(\hat{t}(1-c), \theta) = \sqrt{L}(1-c - \hat{t}(1-c)) + o_P(1).$$

Next, (13) implies that $\sqrt{L}(\ell_c(\hat{t}(t)) - \ell_c(t)) = \sqrt{L}(\hat{t}(t) - t) + o_P(1)$ in case of $t \in (1-c, 1]$. We have

$$\begin{aligned}\sqrt{L}(\ell_c(\hat{t}(t)) - \ell_c(t)) &= \sqrt{L}(F_\theta(F_\theta^-(t)) - F_{\hat{\theta}}(F_{\hat{\theta}}^-(t))) + o_P(1) \\ &= \sqrt{L}(\theta - \hat{\theta}) \left. \frac{\partial F_\gamma(x)}{\partial \gamma} \right|_{\substack{x=F_\theta^-(t) \\ \gamma=\theta^*}} + o_P(1),\end{aligned}$$

where θ^* is between θ and $\hat{\theta}$. Similarly to the corresponding steps in the proof of Lemma 2 [6] we can show that

$$\left. \frac{\partial F_\gamma(x)}{\partial \gamma} \right|_{\substack{x=F_{\hat{\theta}}^-(t) \\ \gamma=\theta^*}} = \left. \frac{\partial F_\gamma(x)}{\partial \gamma} \right|_{\substack{x=F_\theta^-(t) \\ \gamma=\theta}} + o_P(1),$$

since $\partial F_\gamma(x)/\partial \gamma$ is continuous with respect to (x, γ) for all $x > F_\theta^-(1-c)$ and $\gamma \in (0, 1]$. Clear,

$$\left. \frac{\partial F_\gamma(x)}{\partial \gamma} \right|_{\substack{x=F_\theta^-(t) \\ \gamma=\theta}} = (x-1)e^{-\gamma x} \Big|_{\substack{x=F_\theta^-(t) \\ \gamma=\theta}} = \frac{1-t}{\theta^2} \log\left(\frac{1-t}{\theta}\right) = g(t, \theta).$$

Note that the relation

$$\sqrt{L}(\hat{t}(t) - t) = \sqrt{L}(\theta - \hat{\theta})g(t, \theta) + o_P(1)$$

derived above for $t \in (1-c, 1]$ remains true also for $t = 1-c$. Finally, combining the previous relations and using (14), we derive (15).

Define the empirical process

$$z_L(t) = \frac{1}{\sqrt{L}} \sum_{i=1}^L \left(I(1-c < F_\theta(Z_i) \leq t) - \ell_c(t) Z_i - \theta(2Z_i - \frac{\theta}{2} Z_i^2 - 1)(g(t, \theta) - g(1-c, \theta)) \right)$$

and notice that $\hat{y}_c(t) = z_L(t) + o_P(1)$ by Lemma 2, (12) and (15). To complete the proof of Theorem 4 we need to prove that

$$(z_L(t_1), \dots, z_L(t_k)) \xrightarrow{P} (X(t_1), \dots, X(t_k)), \quad \text{for all } 1-c \leq t_1 < \dots < t_k \leq 1,$$

where $X(t)$ is the Gaussian process on $[1-c, 1]$ with mean zero and covariance function (9), and justify that the sequence of random elements (z_L) is tight. These parts of the proof are carried out similarly to the proofs of Lemma 3 and Lemma 4 [6], respectively.

5 Conclusion

The paper provides a study of properties of the new threshold selection method for non-parametric estimation of the extremal index of stationary sequences proposed in [15]. We consider a specific normalization of the discrepancy statistic based on some modifications of the Cramér–von Mises–Smirnov statistic ω^2 that is calculated by only k largest order statistics of a sample. We show that the asymptotic distribution of the truncated Cramér–von Mises–Smirnov statistic (8) as $k \rightarrow \infty, k/L \rightarrow c, L \rightarrow \infty$ depends both on c and the limit distribution of the extremal index estimator being substituted in the statistic. We also develop the goodness-of-fit test for distribution tails based on the ω^2 statistic modification, which limit distribution coincides with the limit distribution of the classical Cramér–von Mises–Smirnov statistic under null hypothesis.

References

1. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: *Statistics of Extremes: Theory and Applications*. Wiley, Chichester (2004)
2. Berghaus, B., Bücher, A.: Weak convergence of a pseudo maximum likelihood estimator for the extremal index. *Ann. Stat.* **46**(5), 2307–2335 (2018)
3. Billingsley, P.: *Convergence of Probability Measures*. Wiley, New York (1968)
4. D’Agostino, R.B., Stephens, M.A.: *Goodness of Fit Techniques*. Marcel Dekker, New York (1986)
5. Darling, D.A.: The Cramér–Smirnov test in the parametric case. *Ann. Math. Stat.* **26**, 1–20 (1955)
6. Durbin, J.: Weak convergence of the sample distribution function when parameters are estimated. *Ann. Stat.* **1**(2), 279–290 (1973)
7. de Haan, L., Ferreira, A.: *Extreme Value Theory: An Introduction*. Springer, Heidelberg (2006). <https://doi.org/10.1007/0-387-34471-3>
8. Ferro, C.A.T., Segers, J.: Inference for clusters of extreme values. *J. R. Stat. Soc. B.* **65**, 545–556 (2003)

9. Fukutome, S., Liniger, M.A., Süveges, M.: Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. *Theor. Appl. Climatol.* **120**, 403–416 (2015)
10. Kac, M., Kiefer, J., Wolfowitz, J.: On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Stat.* **26**, 189–211 (1955)
11. Leadbetter, M.R., Lindgren, G., Rootzén, H.: *Extremes and Related Properties of Random Sequence and Processes*. Springer, New York (1983). <https://doi.org/10.1007/978-1-4612-5449-2>
12. Markovich, N.M.: Experimental analysis of nonparametric probability density estimates and of methods for smoothing them. *Autom. Rem. Contr.* **50**, 941–948 (1989)
13. Markovich, N.M.: *Nonparametric Analysis of Univariate Heavy-Tailed data: Research and Practice*. Wiley, Hoboken (2007)
14. Markovich, N.M.: Nonparametric estimation of extremal index using discrepancy method. In: *Proceedings of the X International Conference “System Identification And Control Problems” SICPRO-2015*, pp. 160–168. V.A. Trapeznikov Institute of Control Sciences (2015)
15. Markovich, N.M., Rodionov, I.V.: Threshold selection for extremal index estimation. *Scand. J. Stat.* (2020). Under review. [arxiv:2009.02318](https://arxiv.org/abs/2009.02318)
16. Northrop, P.J.: An efficient semiparametric maxima estimator of the extremal index. *Extremes* **18**(4), 585–603 (2015)
17. Pettitt, A.N., Stephens, M.A.: Modified Cramer-von Mises statistics for censored data. *Biometrika* **63**(2), 291–298 (1976)
18. Robert, C.Y.: Asymptotic distributions for the intervals estimators of the extremal index and the cluster-size probabilities. *J. Stat. Plan. Infer.* **139**, 3288–3309 (2009)
19. Robert, C.Y., Segers, J., Ferro, C.A.T.: A sliding blocks estimator for the extremal index. *Electron. J. Stat.* **3**, 993–1020 (2009)
20. Rodionov, I.V.: On discrimination between classes of distribution tails. *Probl. Inform. Transm.* **54**(2), 124–138 (2018)
21. Süveges, M., Davison, A.C.: Model misspecification in peaks over threshold analysis. *Ann. Appl. Stat.* **4**(1), 203–221 (2010)
22. Vapnik, V.N., Markovich, N.M., Stefanyuk, A.R.: Rate of convergence in L_2 of the projection estimator of the distribution density. *Autom. Rem. Contr.* **53**, 677–686 (1992)
23. Voinov, V., Nikulin, M., Balakrishnan, N.: *Chi-squared Goodness-of-Fit Tests with Applications*. Academic Press, Boston (2013)
24. Weissman, I., Novak, S.Y.: On blocks and runs estimators of the extremal index. *J. Stat. Plan. Infer.* **66**, 281–288 (1998)