Threshold selection for extremal index estimation

Journal:	*Journal of Nonparametric Statistics*
Manuscript ID	GNST-2022-04-06.R2
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Markovich, Natalia; V.A.Trapeznikov Institute of Control Sciences Russian academy of sciences, Rodionov, Igor; Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences
Keywords:	Cram\'{e}r-von Mises-Smirnov statistic, Discrepancy method, Extremal index, Nonparametric estimation, Threshold selection
Classifications:	Nonparametric statistics, extreme value theory, statistical algorithms, Asymptotic Statistics, Extreme Value Methods
Maths:	62G32

Threshold selection for extremal index estimation

Natalia Markovich[a*] and Igor Rodionov [b]

[a] *V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, 117997 Russia*

[b] *Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russian Federation*

We propose a new threshold selection method for nonparametric estimation of the extremal index of stochastic processes. The discrepancy method was proposed as a data-driven smoothing tool for estimation of a probability density function. Now it is modified to select a threshold parameter of an extremal index estimator. A modification of the discrepancy statistic based on the Cramér-von Mises-Smirnov statistic ω^2 is calculated by k largest order statistics instead of an entire sample. Its asymptotic distribution as $k \to \infty$ is proved to coincide with the ω^2-distribution. Its quantiles are used as discrepancy values. The convergence rate of an extremal index estimate coupled with the discrepancy method is derived. The discrepancy method is used as an automatic threshold selection for the intervals and $K-$gaps estimators. It may be applied to other estimators of the extremal index. The performance of our method is evaluated by simulated and real data examples.

Keywords: Cramér-von Mises-Smirnov statistic; Discrepancy method; Extremal index; Nonparametric estimation; Threshold selection.

AMS Subject Classification: 62G32

1. Introduction

Let $\{X_i\}_{i=1,...,n}$ be a sample of random variables (r.v.s) from a strictly stationary time series with cumulative distribution function (cdf) $F(x)$. By Leadbetter et al. (1983) the stationary sequence $\{X_n\}_{n\geq 1}$ is said to have the 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 it holds

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

where $M_n = \max\{X_1, ..., X_n\}$. The extremal index reflects a cluster structure of an underlying sequence or its local dependence. For stationary sequences $\theta = 1$ when mixing

*Corresponding author. Email: nat.markovich@gmail.com

conditions $D(u_n)$ and $D'(u_n)$ hold (Leadbetter et al. 1983); in particular, it holds if $X_1, ..., X_n$ are independent.

Nonparametric estimators of θ require usually the choice of a threshold u and/or a declustering parameter. The well-known blocks and runs estimators of θ require u and the block size b or the number of consecutive observations r running below u to separate two consecutive clusters (Beirlant et al. 2004). A bias-corrected modification of the blocks estimator in Drees (2011) informs how to avoid the threshold selection by providing a rather stable plot of the extremal index estimates against u with some remaining uncertainty. In Sun and Samorodnitsky (2019) the multilevel blocks estimator is proposed where a sequence of increasing levels and a weight function have to be defined. The sliding blocks estimator has asymptotic variance smaller than the disjoint blocks estimator (Robert et al. 2009b) and both of them require the selection of a pair (u, b). The cycles estimator proposed by Ferreira and Ferreira (2018) needs both u and the cycle size s as parameters. The intervals estimator of θ by Ferro and Segers (2003) and the estimators introduced by Robert (2009b) require the choice of u. The K-gaps estimator is another threshold-based one (Süveges and Davison 2010).

One of the high quantiles of the sample $\{X_i\}_{i=1,...,n}$ is taken usually as u or u is selected visually corresponding to a stability interval of the plot of some estimate $\widehat{\theta}(u)$ against u. Following Süveges and Davison (2010), a list of pairs (u, K) is selected by the Information Matrix Test (IMT) in Fukutome et al. (2015). Then u is selected from such a pair that corresponds to the largest number of clusters of exceedances separated by more than K non-exceedances. The semiparametric maxima estimators depend on the block size only (Berghaus and Bücher 2018; Northrop 2015).

The objective of this paper is to propose a new nonparametric method based on a discrepancy statistic to find the threshold u. The latter statistic is built by the largest order statistics of normalized interexceedance times. We aim to find a limit distribution of the discrepancy statistic and to prove the consistency and the rate of convergence of extremal index estimates with u selected by the discrepancy method. The calculation algorithm and the comparison with other estimators of the extremal index will be presented.

The so-called discrepancy method was proposed in Markovich (1989) and Vapnik et al. (1992) as a data-driven smoothing tool for a probability density function (pdf) estimation by i.i.d. data. We aim to extend this method for extremal index estimation. The idea was to find an unknown parameter h of the pdf as a solution of the discrepancy equation

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

Here, $\widehat{F}_h(x) = \int_{-\infty}^{x} \widehat{f}_h(t)dt$ holds, $\hat{f}_h(t)$ is some pdf estimate, δ is a discrepancy value of the estimation of $F(x)$ by the empirical distribution function $F_n(x)$, i.e. $\delta = \rho(F, F_n)$. $\rho(\cdot, \cdot)$ is a nonnegative (loss) fuction in the space of cdf's. We will focus on the Cramér-von

Mises-Smirnov (C-M-S) statistic

$$\omega_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x) \tag{1}$$

as $\rho(F_n, F)$. Since δ is usually unknown, quantiles of its limit distribution may be proposed as δ. The latter limit distribution of the C-M-S statistic (denote it by A_1) is invariant regarding F and rather complicated (Bolshev and Smirnov 1965; Markovich 2007). For applications the tuning parameter h was proposed in Markovich (1989) as a solution of the equation

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

Here,

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

was calculated by the order statistics $X_{1,n} \leq ... \leq X_{n,n}$ of the sample $\{X_i\}_{i=1,...,n}$, the value 0.05 corresponds to the mode of A_1 and thus, the maximum likelihood value of A_1 was found from tables (see, e.g., Bolshev and Smirnov 1965) as δ.[1] To estimate the extremal index we replace $\widehat{F}_h(X_{i,n})$ in (2) by the exponential distribution model of normalized interexceedance times $\{Y_i\}_{i=1,...,L}$, $L = L(u)$ derived in Ferro and Segers (2003) and use in the sum only k largest order statistics of the latter sample.

The paper is organized as follows. In Section 2 related work is recalled. In Section 3 a modification of the C-M-S statistic denoted as $\widetilde{\omega}_L^2(\theta)$ is introduced. The limit distribution of $\widetilde{\omega}_L^2(\theta)$ built on independent observations is proved to coincide with the limit distribution of ω_n^2 (Theorem 3.1). The convergence of $\widetilde{\omega}_L^2(\widehat{\theta})$ distribution to A_1 is derived when the difference $\sqrt{m_n}(\widehat{\theta} - \theta)$ has a nondegenerate distribution (Theorem 3.2), where m_n is some sequence relating to k and L. In Theorem 3.3 the consistency and the inconsistency conditions for the normalized statistic $\widetilde{\omega}_L^2(\hat{\theta})$ are given. The rate of convergence of extremal index estimates with u selected by the discrepancy method is derived in Corollary 3.4. The choice of parameter k, the number of the largest order statistics, used for $\widetilde{\omega}_L^2$ calculation for samples of moderate sizes is discussed. An algorithm and a simulation study of the discrepancy method based on the normalized statistic $\widetilde{\omega}_L^2(\widehat{\theta})$ are given in Section 4. An application of the method to real data examples is provided in Section 5. Proofs can be found in Markovich and Rodionov (2022) and in a supplementary material file.

[1]The connection between (1) and (2) can be found in Markovich (2007), p.81.

2. Important mathematical results

Our results are based on Lemmas 2.2.3, 3.4.1 by de Haan and Ferreira (2006) concerning the limit distributions of the order statistics and Theorem 1 by Ferro and Segers (2003). Here and further, we denote for brevity a sequence of positive integers $\{k_n\}$ as k.

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

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

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

In Ferro and Segers (2003) a r.v. $T(u)$ equal in distribution to

$$\min\{j \ge 1 : X_{j+1} > u\} \quad \text{given} \quad X_1 > u$$

is considered. Theorem 1 by Ferro and Segers (2003) states that

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

where η is exponentially distributed with mean θ^{-1}. The zero asymptotic interexceedance times (the intracluster times) imply the times between the consecutive exceedances of the same cluster. The positive asymptotic interexceedance times are the inter-cluster times. $\to^d$ denotes convergence in distribution. Taking the exceedance times $1 \le S_1 < ... < S_{N_u} \le n$, the observed interexceedance times are $T_i = S_{i+1} - S_i$ for $i = 1, ..., N_u - 1$, where $N_u = \sum_{i=1}^n \mathbf{1}\{X_i > u\}$ is the number of observations exceeding a predetermined high threshold u.[2] Hereinafter we write $L \equiv L(u) = N_u - 1$.

The intuition for declustering of a sample is given in Ferro and Segers (2003). One can assume that the largest $C - 1 = \lfloor \theta L \rfloor$ interexceedance times are approximately independent inter-cluster times. The larger u corresponds to the larger interexceedance times whose number $L \equiv L(u)$ may be small. It leads to a larger variance of the estimates based on $\{T_i(u)\}$.

The intervals estimator is defined as (Ferro and Segers 2003),

$$\hat{\theta}_n(u) = \begin{cases} \min(1, \hat{\theta}_n^1(u)), \text{ if } \max\{T_i : 1 \le i \le L\} \le 2, \\ \min(1, \hat{\theta}_n^2(u)), \text{ if } \max\{T_i : 1 \le i \le L\} > 2, \end{cases} \tag{3}$$

[2]Theoretically, events $\{T_i = 1\}$ are allowed. In practice, such cases related to single inter-arrival times between consecutive exceedances are meaningless.

where

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

The K-gaps estimator proposed in Süveges and Davison (2010) is obtained by the maximum likelihood method using the model by Ferro and Segers (2003)

$$P\{\overline{F}(u_n)T(u_n) > t\} \to \theta\exp(-\theta t), \qquad n \to \infty. \tag{4}$$

and assuming that the K-gaps observations are independent. The K-gaps

$$S(u_n)^{(K)} = (\max(T(u_n) - K, 0)), \quad K = 0, 1, 2, ...$$

are obtained by truncation of the interexceedance times by the run parameter K. The normalized K-gaps $\overline{F}(u_n)S(u_n)^{(K)}$ have the same limiting mixture law (4) according to Theorem 2.1 in Süveges and Davison (2010). The K-gaps estimator has the following form

$$\widehat{\theta}^K = 0.5\left((a+b)/c + 1 - \sqrt{((a+b)/c+1)^2 - 4b/c}\right), \tag{5}$$

with $a = L - N_C$, $b = 2N_C$, $c = \sum_{i=1}^L \overline{F}(u_n)S(u_n)_i^{(K)}$. N_C is the number of non-zero K-gaps. The K-gaps estimator (5) is consistent and asymptotically normal as $n \to \infty$. Due to possible nonstationarity and violation of independence at extreme levels, K and u are to be selected by a misspecification test. The iterative weighted least squares estimator of Süveges (2007) explores the interexceedance times with $K = 1$. The automatic selection of an optimal pair (u, K) is proposed in Fukutome et al. (2015) by a choice of pairs for which values of the statistic of the information matrix test (the IMT) are less than 0.05. The test works satisfactorily when the number of exceedances is not less than 80.

The intervals estimator is derived to be consistent for m-dependent processes (Ferro and Segers 2003). Asymptotic normality property $\sqrt{m_n}(\hat{\theta}_n(u) - \theta) \to^d N(0, V)$ as $n \to \infty$ is derived for several extremal index estimators and different values of variance V. In most cases m_n is proportional or asymptotically proportional to n/r_n. Below r_n, u_n and τ have the same meaning as in Theorem 1 by Ferro and Segers (2003). First, $m_n = n\overline{F}(u_n)$ holds for the blocks and runs estimators in Weissman and Novak (1998), where $\overline{F}(u_n) = \tau/r_n(1+o(1))$; $m_n = L(u_n)$ is the number of interexceedance times $\{T_i(u_n)\}$ for the intervals estimator in Robert (2009a), where $L(u_n)$ is asymptotically equivalent to $\tau n/r_n$ in probability, see below; $m_n = n/r_n$ is taken for the multilevel blocks estimator in Sun and Samorodnitsky (2018), where $r_n\overline{F}(u_n^s) \to \tau_s$ and $s \in \{1, ..., m\}$ is the number of levels $\{u_n^s\}$; $m_n = n/r_n$ is used for the disjoint and sliding blocks estimators (Robert 2009a); $m_n = k_n$ is taken for the disjoint and sliding blocks estimators by Berghaus and Bücher (2018) and Northrop (2015), where k_n is a number of blocks of length b_n such

that $k_n = o(b_n^2)$ holds as $n \to \infty$. The latter results provide examples for the possible choice of m_n in assumption (10). It also relates to Remark 1.

3. Main results: ω^2-distribution of the modified Cramér-von Mises-Smirnov statistic

Using the largest order statistics of the sample $\{Y_i = (N_u/n)T_i\}$ let us rewrite (2) in the following form

$$\hat{\omega}_L^2(u) = \sum_{i=L-k+1}^{L} \left(1 - \theta \exp(-Y_{i,L}\theta) - \frac{i-0.5}{L}\right)^2 + \frac{1}{12L} \tag{6}$$

and investigate its limit distribution. Note that $L = L(u_n)$ is a sequence of r.v.s converging in probability to infinity such that $r_n L(u_n)/n \to \tau$ in probability, where the sequence of thresholds $u_n = u_n(\tau)$ and the sequence r_n were introduced in Theorem 1 by Ferro and Segers (2003).

According to Martynov (1978), Smirnov (1952) the limit distribution of the C-M-S statistic (1) (or, equivalently $\omega_n^2 = n\int_0^1 (F_n(t) - t)^2\, dt$) coincides with the distribution of

$$\Omega = \int_0^1 B^2(t)dt,$$

$B(t)$ is a Brownian bridge on $[0,1]$, i.e. the Gaussian random process with zero mean and the covariance function $R(s,t) = \min(s,t) - st$, $s,t \in [0,1]$. Thus, the statistic (6), built by k largest order statistics only, tends to 0 for $k = o(L)$ as $n \to \infty$, since the interval over which we integrate $B^2(t)$ tends to an empty set. Thus, (6) should be modified to have a non-degenerate limit distribution. Let us consider the modification of (6)

$$\widetilde{\omega}_L^2(\theta) = \frac{1}{(1-t_k)^2} \cdot \tag{7}$$
$$\cdot \sum_{i=L-k+1}^{L} \left(1 - \theta \exp(-Y_{i,L}\theta) - t_k - \frac{i-(L-k)-0.5}{k}(1-t_k)\right)^2 + \frac{1}{12k},$$

where $t_k = 1 - \theta \exp(-Y_{L-k,L}\theta)$. Let us explain in more detail why we need such modification. It follows from Theorem 1 by Ferro and Segers (2003) that there are a probability θ of asymptotic positive interexceedance times (the inter-cluster times) and a probability $1-\theta$ of zero asymptotic interexceedance times (the intra-cluster times). Moreover, the inter-cluster times are asymptotically independent exponential with mean $1/\theta$. Thus, one should built the statistic on only inter-cluster times to be able to employ the asymptotic independence property.

3.1. *Modified Cramér-von Mises-Smirnov statistic for known θ*

In this section we consider the following auxiliary statistic

$$\omega_k^2(\theta) = \sum_{i=L-k+1}^{L} \left(\frac{Z_{i,L} - Z_{L-k,L}}{1 - Z_{L-k,L}} - \frac{i - (L-k) - 0.5}{k} \right)^2 + \frac{1}{12k}, \tag{8}$$

where $Z_{i,L} = 1 - \theta \exp(-T^*_{i,L}\theta)$, $T^*_{1,L} \leq \ldots \leq T^*_{L,L}$ are order statistics of a sample $\{T_i^*\}$, $\{T_i^*\}$ are independent copies of T_θ. We assume in this section that L is a non-random sequence tending to infinity.

It follows from Theorem 1 by Ferro and Segers (2003) and Lemma 3.4.1 (de Haan and Ferreira 2006), that the conditional distribution of the set of order statistics $\{Z_{i,L}\}_{i=L-k+1}^{L}$ given $Z_{L-k,L} = s_k$ asymptotically agrees for $\limsup_{n\to\infty} k/L < \theta$ with the distribution of the set of order statistics $\{U^*_{j,k}\}$, $j = i - (L-k)$, of an i.i.d. sample $\{U_j^*\}$ from the uniform distribution on $[s_k, 1]$. The asymptotical distribution of $\omega_k^2(\theta)$ is given in the next theorem.

THEOREM 3.1 *It holds*

$$\omega_k^2(\theta) \xrightarrow{d} \xi$$

for $k \to \infty$, $\limsup_{n\to\infty} k/L < \theta$ as $n \to \infty$, where ξ obeys A_1 distribution, the limit distribution of the C-M-S statistic ω_n^2.

3.2. *Modified Cramér-von Mises-Smirnov statistic for unknown θ*

Here, we check whether one can substitute θ by its estimate $\widehat{\theta}$ in (7) and find the conditions that should be imposed on $\widehat{\theta}$ under which the limit distribution of $\widetilde{\omega}_L^2(\hat{\theta})$ will be the same as the limit distribution of $\omega_k^2(\theta)$. Recall again that the number of interexceedance times $L = L(u_n)$ is a sequence of r.v.s tending to $+\infty$ as $n \to \infty$ (Robert 2009a). In the spirit of Theorem 3.2, Robert (2009a), the limit distribution of the following statistic

$$\sqrt{L}\left(\sum_{i=1}^{L} f(Y_i) - Ef(Y_1) \right)$$

for some continuous f may not depend on a substitution of the set of r.v.s $\{T_i^*\}_{i=1}^{L}$ appearing in (8) instead of $\{Y_i\}_{i=1}^{L}$. Moreover, $T_i^* \stackrel{d}{=} T_\theta$, $i \in \{1, \ldots, L\}$ and there is a probability θ of the nonzero elements of this set that are independent exponentially distributed with parameter θ. For these r.v.s Theorem 2.2.1, de Haan and Ferreira (2006) implies that if $k/L \to 0$ and $k \to \infty$ as $n \to \infty$, then

$$\sqrt{k}(T^*_{L-k,L} - \ln(L\theta/k)/\theta) = O_P(1).$$

In light of these remarks let us assume that there exists a sample of independent exponentially distributed r.v.s $\{E_i^{(L)}\}_{i=1}^{l}$ with mean θ^{-1} for all large enough n such that

$$Y_{L-k,L} - E_{l-k,l}^{(L)} = o_P\left(\frac{1}{\sqrt{k}}\right) \tag{9}$$

if $k/L \to 0$ and $k \to \infty$ as $n \to \infty$ in probability, where we denote $l = \lfloor \theta L \rfloor$ and assume $k < l$. Theorem 3.1 remains valid if $T_{i,L}^*$, $i \in \{L-k, ..., L\}$ in $\omega_k^2(\theta)$ are substituted by $E_{i,l}^{(L)}$, $i \in \{l-k, ..., l\}$. This is possible due to condition $\limsup_{n\to\infty} k/L = 0 < \theta$. Thus, with probability tending to one $T_{L-k,L}^* > 0$ holds, and by Renyi's representation (for this argument see, e.g., the proof of Lemma 3.2.3, de Haan and Ferreira (2006)) we get $T_{L-i,L}^* - T_{L-k,L}^* =^d E_{l-i,l}^{(L)} - E_{l-k,l}^{(L)}$ given $T_{L-k,L}^* > 0$.

THEOREM 3.2 *Let the conditions of Theorem 1 by Ferro and Segers (2003) and the condition (9) be fulfilled and the estimator of the extremal index $\widehat{\theta} = \widehat{\theta}_n$ be such that*

$$\sqrt{m_n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \zeta, \quad n \to \infty, \tag{10}$$

where the r.v. ζ has a nondegenerate distribution function H. Let us assume that the sequence m_n is such that

$$\frac{k}{m_n} = o(1) \quad \textit{and} \quad \frac{(\ln L)^2}{m_n} = o(1) \tag{11}$$

in probability as $n \to \infty$. Then

$$\widetilde{\omega}_L^2(\widehat{\theta}_n) \xrightarrow{d} \xi \sim A_1$$

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

Remark 1 For instance, ζ is normally distributed with mean zero and $m_n = O(n/r_n)$ (hence, $m_n = O(L)$ in probability) for the intervals, blocks and sliding blocks estimators of the extremal index (Northrop (2015); Robert (2009a); Robert et al. (2009); Sun and Samorodnitsky (2018)), see also the last paragraph in Section 2.

Remark 2 The replacement of $o(1)$ by $O(1)$ in (11) violates Theorem 3.2. The assumption $k = O(m_n)$ may lead to the fact that the limit distribution of $\widetilde{\omega}_L^2(\widehat{\theta}_n)$ will differ from A_1, that is beyond the scope of our paper.

THEOREM 3.3 *Let the conditions of Theorem 1 by Ferro and Segers (2003) and (9) be fulfilled. Assume that the sequence of estimates $\{\hat{\theta}_n\}$ is such that for some $\alpha \in [0, 1/2]$*

$$k_{n_s}^{\alpha} |\hat{\theta}_{n_s} - \theta| \xrightarrow{P} +\infty, \qquad \textit{if} \qquad 0 < \alpha \le 1/2,$$
$$|\hat{\theta}_{n_s} - \theta| > \varepsilon \qquad \textit{for some} \qquad \varepsilon > 0, \qquad \textit{if} \qquad \alpha = 0$$

holds for some $n_s \to \infty$, $n_s \in \mathrm{N}$, $s \geq 1$, *where* $k = k_n = k(u_n)$, $k = o(L)$ *in probability. Then for corresponding subsequence* $\{L_{n_s}\}$ *of the sequence* $\{L\}$, $L = L(u_n)$

$$\widetilde{\omega}^2_{L_{n_s}}(\hat{\theta}_{n_s})/k^{1-2\alpha}_{n_s} \xrightarrow{P} +\infty$$

holds as $n \to \infty$.

Remark 3 Theorem 3.3 implies that the non-consistency of the estimator $\widehat{\theta}_n$ or the consistency with a sufficiently slow rate leads to the non-consistency of $\widetilde{\omega}^2_{L_{n_s}}(\widehat{\theta}_{n_s})$ in a sense that its limit distribution does not exist or the latter statistic tends to $+\infty$. In case that $\alpha \neq 0$ holds, the estimator $\widehat{\theta}_n$ may be consistent but with the rate of convergence slower than $k_n^{-\alpha}$. Hence, $\widetilde{\omega}^2_L(\widehat{\theta}_n)$ may be considered as a quality functional of $\widehat{\theta}_n$.

The consistency of the corresponding extremal index estimates follows from Theorem 3.3. The next corollary states, if the solutions of the discrepancy equation exist for each n, then the consistency is fulfilled.

COROLLARY 3.4 *Let the conditions of Theorem 1 by Ferro and Segers (2003) and (9) be fulfilled. Let* $\hat{\theta}_n(u_n)$ *be an estimator of* θ, $k = k(u_n) = f(L(u_n), \hat{\theta}_n(u_n))$, *where the function* $f(x, y)$ *is such that* $f(x, y) = o(x)$ *and* $f(x, y) \to \infty$ *as* $x \to \infty$ *uniformly for arbitrary* y. *Assume* $\{\tilde{u}_n\}$ *be some sequence of solutions of the discrepancy equation, such that every* n *corresponds to exactly one solution. Then* $\hat{\theta}_n(\tilde{u}_n) \xrightarrow{P} \theta$ *and for arbitrary* $\varepsilon > 0$

$$k(\tilde{u}_n)^{1/2-\varepsilon}|\hat{\theta}_n(\tilde{u}_n) - \theta| \xrightarrow{P} 0, \quad n \to \infty.$$

The proof of the corollary is based on a negation of the assertion of Theorem 3.3.

3.3. *The choice of k*

According to Theorem 3.2 the asymptotic distribution of $\widetilde{\omega}^2_L(\widehat{\theta}_n)$ does not depend on k. The k-selection gives another viewpoint that using only the largest interexceedance times screens out the smallest interexceedance times. It is helpful for the reasons discussed in Ferro and Segers (2003) and is the motivation for introducing the tuning parameter K in the $K-$gaps estimator of θ in Süveges and Davison (2010).

In practice, for each predetermined δ and u one may increase k until $k \leq \min\{\widehat{\theta}_0 L(u), L(u)^\beta\}$, $0 < \beta < 1$, holds ($\widehat{\theta}_0$ is some pilot estimate of θ) and the discrepancy equations have solutions, and select the largest one among such k's. This choice satisfies the assumptions of Theorem 3.2 but it is not unique. For instance, one can select $k = \lfloor (\ln L)^2 \rfloor$. In the following simulation study we examine three choices: $k = \lfloor \min\{\widehat{\theta}_0 L(u), \sqrt{L(u)}\} \rfloor$, $k = \lfloor (\ln L)^2 \rfloor$ and $k = \lfloor \widehat{\theta}_0 L(u) \rfloor$.

4. Simulation study

In our simulation study we focus on the threshold-based intervals and $K-$gaps estimators and aim to show the advantages of the discrepancy method used to find the threshold u for the latter estimators. The natural drawback of the intervals estimator is that it requires a large sample size n to obtain a moderate size $L(u)$ for a large u. The same concerns the K-gaps estimator.

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

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

where N_u is the number of exceedances over threshold u.

(2) For each u select $k = \lfloor \widehat{\theta}_0 L \rfloor$ (in case $\widehat{\theta}_0 = 1$, accept $k = L-1$), $k = \min\{\lfloor \widehat{\theta}_0 L \rfloor, \sqrt{L}\}$ or $k = \lfloor (\ln L)^2 \rfloor$, where the intervals estimator (3) may be selected as a pilot estimator $\widehat{\theta}_0 = \widehat{\theta}_0(u)$ with the same u as in Item 1.

(3) Use a sorted sample $Y_{L-k+1,L} \le ... \le Y_{L,L}$ and find all levels $u_1, ..., u_l$ among considered quantiles (here, l is a random number) such that

$$|\widetilde{\omega}_L^2(\widehat{\theta}) - \delta_1| < \varepsilon, \qquad \varepsilon = 0.01, \tag{12}$$

where

$$\widetilde{\omega}_L^2(\widehat{\theta}) = \sum_{i=0}^{k-1} \left(1 - \frac{\widehat{\theta}\exp(-Y_{L-i,L}\widehat{\theta})}{(1-\widehat{t}_k)} - \frac{k-i-0.5}{k}\right)^2 + \frac{1}{12k} = \delta_1, \tag{13}$$

$\widehat{t}_k = 1 - \widehat{\theta}\exp(-Y_{L-k,L}\widehat{\theta})$, $\widehat{\theta} = \widehat{\theta}(u)$ is calculated by (3), and $\delta_1 = 0.05$ is a mode of C-M-S statistic. If $L < 40$ we should replace $\widetilde{\omega}_L^2(\widehat{\theta})$ by

$$(\widetilde{\omega}_L^2(\widehat{\theta}))' = \left(\widetilde{\omega}_L^2(\widehat{\theta}) - \frac{0.4}{L} + \frac{0.6}{L^2}\right)\left(1 + \frac{1}{L}\right)$$

and use the same discrepancy value δ_1 (Kobzar 2006, p.217; Stephens 1974). [3]

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

$$\widehat{\theta}_1 = \frac{1}{l}\sum_{i=1}^{l} \widehat{\theta}(u_i), \quad \widehat{\theta}_2 = \widehat{\theta}(u_{min}), \qquad \widehat{\theta}_3 = \widehat{\theta}(u_{max}) \tag{14}$$

[3] The modification $(\hat{\omega}_n^2 - 0.4/n + 0.6/n^2)(1 + 1/n)$ of classical statistic (2) eliminates the dependence of the percentage points of the C-M-S statistic on the sample size (Stephens 1974). For $n > 40$ it changes the statistic on less than one percent. One can use the modification with regard to $\widetilde{\omega}_L^2(\widehat{\theta})$ for finite L due to the closeness of its distribution to the limit distribution of the C-M-S statistic by Theorem 3.2.

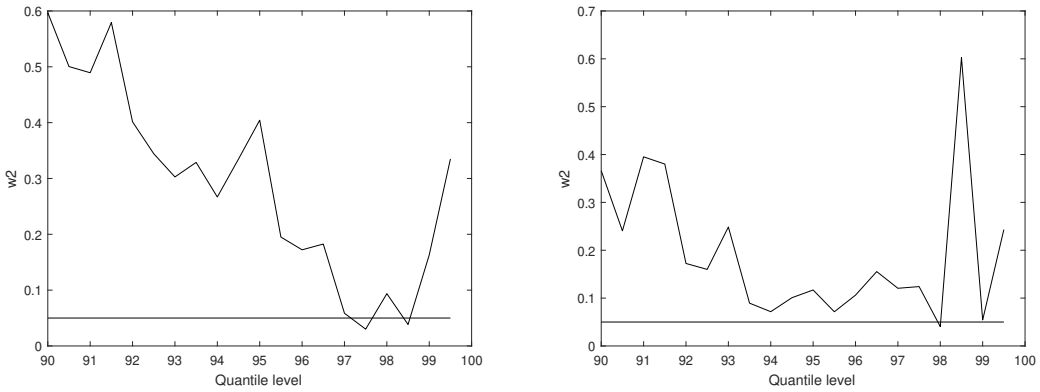

Figure 1. The left-hand side of (13) against quantile levels (cf. step 1) for the MM and $MA(2)$ processes both with extremal index 0.5 and sample size $n = 5000$.

as resulting estimates, where $u_{min} = \min\{u_1, ..., u_l\}$, $u_{max} = \max\{u_1, ..., u_l\}$.

The access to a GitHub project with computer Matlab and Python codes for our automatic threshold selection procedure is https://github.com/natmarkovich/discrepancy. We take the intervals estimator (3) as $\widehat{\theta}_0$ since it requires only u as parameter. Using the $K-$gaps estimator with the IMT method $\widehat{\theta}^{Kimt}$ which is computationally costly and the "plateau-finding" $\widehat{\theta}^{IA1}$ estimator as $\widehat{\theta}_0$ does not improve the results and therefore it is not shown. The solution of (12) can be improved by using intermediate quantiles apart of ones in Item 1 of the algorithm, but this may be computationally more costly, see Figure 1.

Remark 4 For the K-gaps estimator the algorithm is the same, but instead of $\{Y_i\}$ one should use the normalized K-gaps $\{\overline{F}(u)S(u)_i^{(K)}\}$ and K-gaps estimate as $\widehat{\theta}_0$ in (13). Using $\{Y_i\}$ in (13) was also considered but showed worse efficiency, hence we do not include these results to our simulation study. For each value of u one can examine different values of K, for instance, $K \in \{1, 2, ..., 20\}$ can be taken, as in Fukutome et al. (2015). We obtain a set of solutions $(u_i, K_i)_{i=1}^l$ of (12) by the algorithm. Then, $\hat{\theta}_1 = \frac{1}{l}\sum_{i=1}^l \widehat{\theta}(u_i, K_i)$. Next, among the solutions $(u_i, K_i)_{i=1}^l$ we select pairs with minimal value of u, denote it by u_{min}. Among the latter pairs $(u_{min}, K_{i_j})_{j=1}^d$ we select the pair with minimal value of K. This pair we denote as (u_{min}, K_{min}) and then $\hat{\theta}_2 = \widehat{\theta}(u_{min}, K_{min})$. Finally, $\hat{\theta}_3$ is equal to $\widehat{\theta}(u_{max}, K_{max})$, where the pair (u_{max}, K_{max}) is selected by similar way.

Remark 5 Since (12) may not be satisfied for considered quantiles and for given δ_1, k and K, we propose to use the inequality

$$\widetilde{\omega}_L^2(\widehat{\theta}) \leq \delta_2 \tag{15}$$

as an alternative to (12), where $\delta_2 = 1.49$ is the 99.98% quantile of the C-M-S statistic.

The discrepancy method is universal and any estimator depending on u can be cho-

sen. In case of the free-threshold estimators (see, e.g., Northrop (2015), Berghaus and Bücher (2018)) one can find a cluster identification parameter such as the block size by the discrepancy method. Another way is to express the latter parameter in terms of u and to apply the discrepancy method to determine u. For example, the block size can be selected as $b(u) = \lfloor n/L(u) \rfloor$. The simulation study examining free-threshold estimators is out of scope of our paper.

Based on the further simulation study we recommend to use the K-gaps estimator coupled with the discrepancy method (15) with $k = \lfloor \widehat{\theta}_0 L \rfloor$ and an accurate pilot estimate $\widehat{\theta}_0$, and the statistic $\widehat{\theta}_1$.

4.1. *Models*

In our simulation study we consider the processes MM, ARMAX, AR(1), AR(2), MA(2) and GARCH(1,1) with known values of θ. The simulation is repeated 1000 times with the sample size $n = 10^5$ of initial measurements $\{X_1, ..., X_n\}$. Big sample sizes may lead, however, to moderate sample sizes $L(u)$ of normalized inter-exceedance times $\{Y_1, ..., Y_{L(u)}\}$. We recall the definitions of the processes. The mth order MM process is $X_t = \max_{0 \le i \le m}\{\alpha_i Z_{t-i}\}$, $t \in \mathbf{Z}$, 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 $\theta = \max_i\{\alpha_i\}$, Ancona-Navarrete and Tawn (2000). Then $\{X_t\}_{t \ge 1}$ are standard Fréchet distributed. 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 $X_t = \max\{\alpha X_{t-1}, (1-\alpha)Z_t\}$, $t \in \mathbf{Z}$, 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). $P\{X_t \le x\} = \exp(-1/x)$ holds assuming $X_0 = Z_0$. We consider $\theta \in \{0.25, 0.75\}$.

The positively correlated AR(1) process with uniform noise (ARu^+) is defined by $X_j = (1/r)X_{j-1} + \epsilon_j$, $j \ge 1$ and $X_0 \sim U(0,1)$ with X_0 independent of $\{\epsilon_j\}$. Then $X_j \sim U(0,1)$ holds for all $j \ge 1$. For a fixed integer $r \ge 2$ let ϵ_n, $n \ge 1$ be i.i.d. r.v.s with $P\{\epsilon_1 = k/r\} = 1/r$, $k \in \{0, 1, \ldots, r-1\}$. 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 $X_j = -(1/r)X_{j-1} + \epsilon_j$ with similarly distributed $\{\epsilon_j\}$ 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 2019) $X_i = pZ_{i-2} + qZ_{i-1} + Z_i$, $i \ge 1$, with $p > 0$, $q < 1$, and i.i.d. Pareto random variables $Z_{-1}, Z_0, Z_1, ...$ 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. The distribution of the sum of weighted i.i.d. Pareto r.v.s behaves like a Pareto distribution at the tail; one can find its exact form in Ramsay (2008).

We consider also processes studied in (Ferreira 2018b; Northrop 2015; Süveges and Davison 2010): the AR(1) process $X_j = 0.7X_{j-1} + \epsilon_j$, where $\{\epsilon_j\}$ are standard Cauchy distributed and $\theta = 0.3$ (ARc); the $AR(2)$ process $X_j = 0.95X_{j-1} - 0.89X_{j-2} + \epsilon_j$, where $\{\epsilon_j\}$ are Pareto distributed with tail index 2 and $\theta = 0.25$; and the $GARCH(1, 1)$ process $X_j = \sigma_j \epsilon_j$ 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 $\{\epsilon_j\}_{j\geq 1}$ and $\theta = 0.447$ (see Laurini and Tawn 2012).

4.2. *Notations*

The sign $'-'$ in the tables means that there are no solutions of the discrepancy equation. In the tables we investigate different choices of k for the intervals and K-gaps estimators coupled with the discrepancy method. We study $k = \lfloor \widehat{\theta}_0 L \rfloor$ in Tables 1 and 2. Although $k = \lfloor \min(\widehat{\theta}_0 L, \sqrt{L}) \rfloor$ and $k = \lfloor (\ln L)^2 \rfloor$) were studied too, the results were generally worse. Therefore they are not represented. In Tables 1 and 2 the statistics (14) corresponding to the intervals estimates coupled with the discrepancy method (12) are denoted by $\{\hat{\theta}_i\}$, $i \in \{1, 2, 3\}$. The $K-$gaps estimates with pairs (u, K) selected by (12) are denoted by $\widehat{\theta}_i^{Kdis}$, $i \in \{1, 2, 3\}$, and with IMT-selected pairs (u, K) by $\widehat{\theta}^{Kimt}$. Statistics (14) relating to the intervals and $K-$gaps estimators and corresponding to solutions of the discrepancy inequality (15) are marked by asterisks in all tables. The intervals estimate with the threshold u selected by the "plateau-finding" algorithm A1 by Ferreira (2018a) is denoted by $\widehat{\theta}^{IA1}$. This algorithm seems to be the best one for the intervals estimator among other algorithms proposed in Ferreira (2018a) according to the provided simulation study. Applying this algorithm we use the bandwidth $d = [wn]$ with $w = 0.25$ and compute the moving average of $2d + 1$ "successive points" of $\widehat{\theta}$. The value $w = 0.005$ used in Ferreira (2018a) demonstrates slightly worse accuracy uniformly for all processes and we do not show it in Tables 1 and 2.

The values given in bold and italic bold correspond to the first and second best performances.

4.3. *Conclusions*

We propose to select a threshold of the threshold-based intervals and $K-$gaps estimators as a solution of the ω^2 discrepancy equation, where the discrepancy value is equal to the mode of the ω^2-distribution, i.e. to its most likelihood value.

On the first view, the intervals estimator does not require another parameter to be specified apart of the threshold. The intervals estimator coupled with the discrepancy method

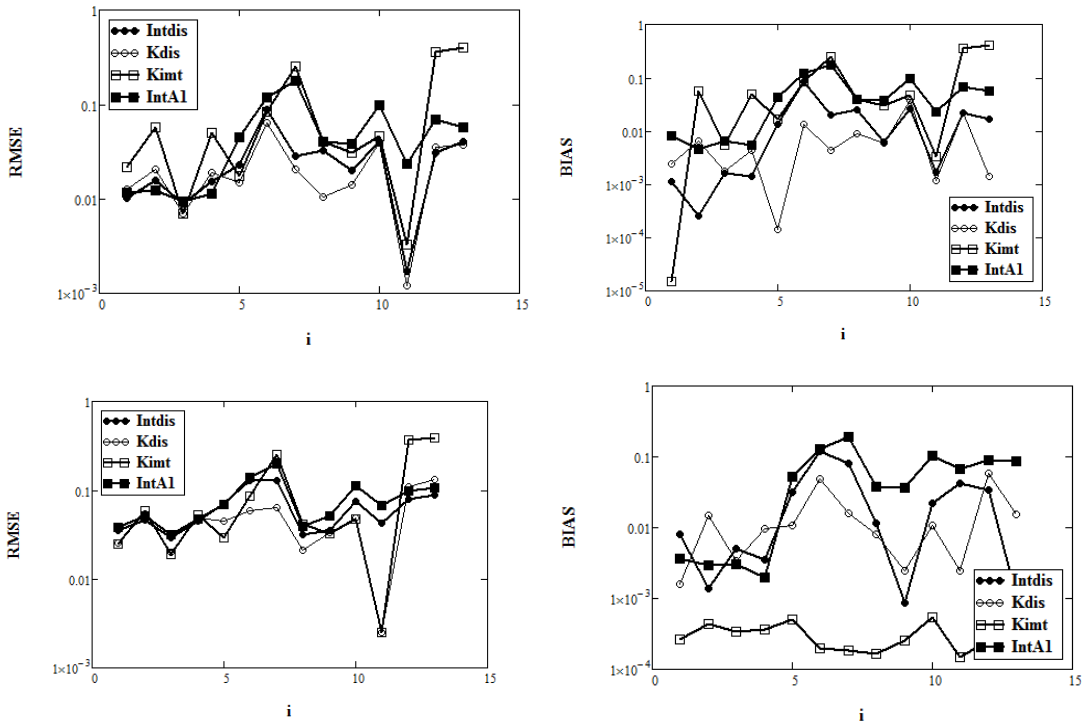

Figure 2. The best RMSE and Bias for the intervals estimator ('Intdis') and K-gaps ('Kdis') estimators with threshold u selected by (12) with the discrepancy equation (13) and the corresponding inequality (15), and for the K-gaps estimator with u selected by the test IMT ('Kimt') and the intervals estimator with the "plateau-finding" algorithm A1 to select u ('IntA1') against the number of processes related to the column labels in Tables 1 and 2, and enumerated from left to right as in the tables for sample size $n = 10^5$ (the upper row) and $n = 5000$ (the lower row).

works in the same way as the $K-$gaps estimator. An additional regularization parameter such as the moving window size for the "plateau-finding" algorithm A1 (Ferreira 2018a) or the number k of the largest order statistics is required to choose the threshold anyway. It is shown in our paper that there is a benefit in choosing k jointly with a threshold. This follows from the algorithm in Section 4 where k depends on $L(u)$.

It is proposed in Ferro and Segers (2003) to select the largest $C - 1 = \lfloor \theta L(u) \rfloor$ inter-exceedance times which are approximately independent, Robert (2009a). We follow a similar way, i.e. $k = \lfloor \widehat{\theta}_0 L \rfloor$ is used as one of the choices of k.

Using of $\widehat{\theta}^{Kimt}$ and $\widehat{\theta}^{IA1}$ estimators as $\widehat{\theta}_0$ does not improve the RMSE obtained by using the intervals estimator as $\widehat{\theta}_0$.

The discrepancy method is competitive with other threshold choices such as the IMT and "plateau-finding" algorithms and it improves substantially the existing intervals and $K-$gaps estimates coupled with the mentioned adjustment methods. Figure 2 corresponding to Tables 1 and 2 shows that the K-gaps estimator works better, if u is selected by the discrepancy method but not by the IMT method. According to our simulation study the K-gaps estimator coupled with the IMT method demonstrates a slow convergence as the sample size increases. The IMT method decreases the bias of the $K-$gaps

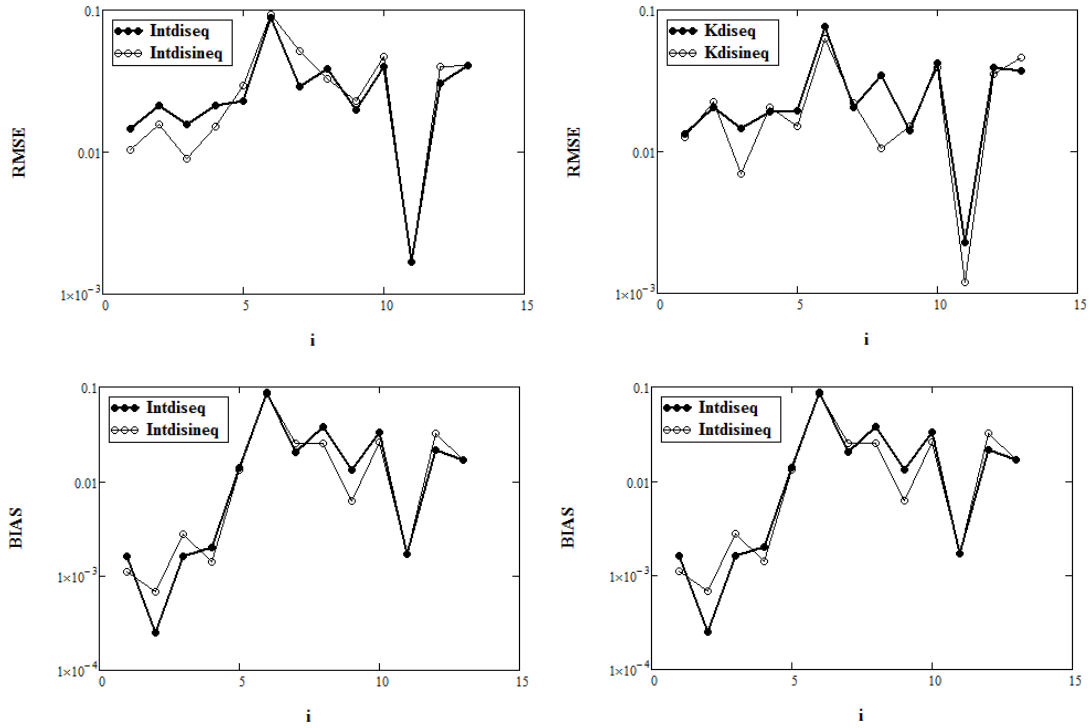

Figure 3. The best RMSE and Bias for the intervals estimator (left column) and K-gaps estimator (right column) obtained by the Algorithm with (12) and (13) notated as 'Intdiseq' and 'Kdiseq', and with the inequality (15) notated as 'Intdisineq' and 'Kdisineq' against the number of processes related to the column labels in Tables 1 and 2 for sample size $n = 10^5$.

estimates for smaller sample $n = 5000$, see Figure 2. However, the $K-$gaps estimate coupled with the discrepancy method provide mostly either better or comparable RMSEs than the IMT. The IMT method requires more computation time due to an exhaustive search among pairs (u, K). Generally, the K-gaps estimator works better than the intervals estimator both coupled with the discrepancy method. The intervals estimator coupled with the algorithm A1 provides the RMSE similar to the discrepancy method coupled with both intervals and K-gaps estimates only for MM and ARMAX processes, see Figure 2.

The discrepancy inequality (15) can be applied when the solutions of the (12) do not exist among the considered quantiles for given k and K. This may slightly improve the RMSE and the absolute bias of both intervals and K-gaps estimates in comparison with the usage of (12), see Figure 3. This property holds due to a larger number of solutions.

Figure 4 aims to compare the impact of the choice of k. It shows that the use of $k = \lfloor \min(\widehat{\theta}_0 L, \sqrt{L}) \rfloor$ and $k = \lfloor (\ln L)^2 \rfloor$ (both satisfying the assumptions of Theorem 3.2) provides similar values of the best RMSE, but $k = \lfloor \widehat{\theta}_0 L \rfloor$ provides the best accuracy among these three choices.

Figure 5 aims to find the best statistic among introduced in (14). Ratios $\{RMSE(\theta_j)/\min_{i\in\{1,2,3\}} RMSE(\theta_i)\}$, $j \in \{1, 2, 3\}$ are compared. One may conclude that $\widehat{\theta}_1$ provides consistently better accuracy than $\widehat{\theta}_2$ and $\widehat{\theta}_3$.

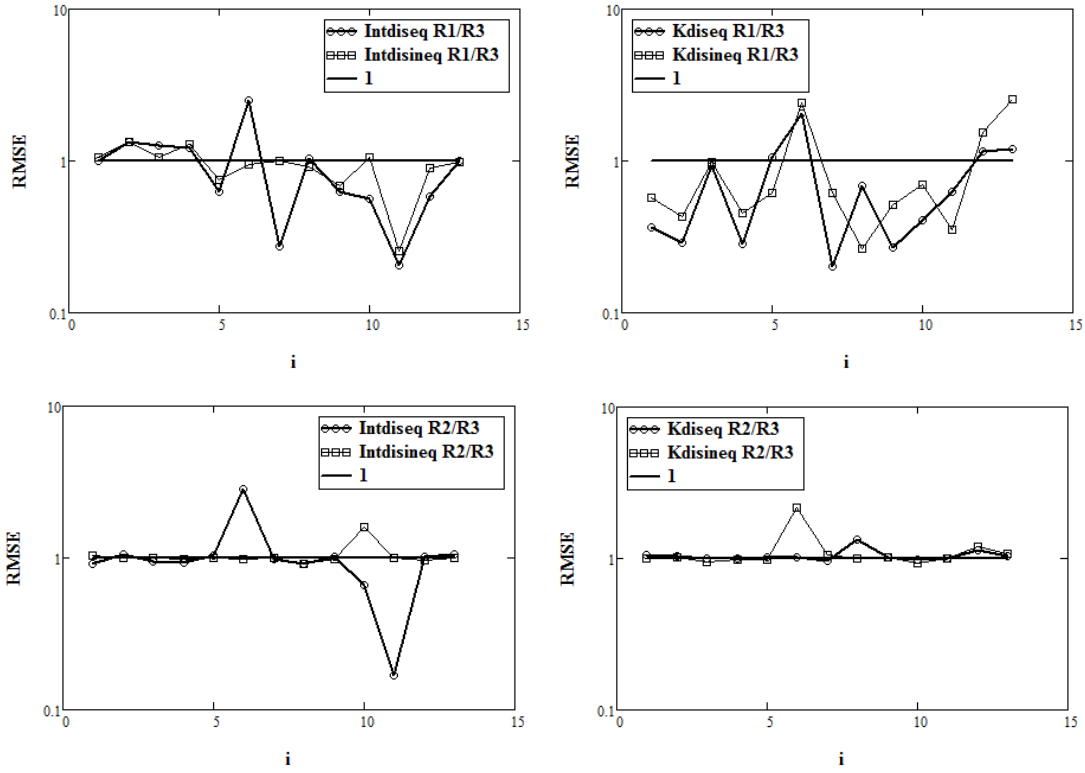

Figure 4. Ratios $R1/R3$ and $R2/R3$ of the best RMSE for the intervals estimator (left column) and K-gaps estimator (right column) obtained by the Algorithm with (12) and the equation (13) notated as 'Intdiseq' and 'Kdiseq', and with the inequality (15) notated as 'Intdisineq' and 'Kdisineq' against the number of processes related to the column labels in Table 1: The $R1/R3$ corresponds to the best results in Table 1 divided to those best with $k = \lfloor \min(\widehat{\theta}_0 L, \sqrt{L}) \rfloor$, and the $R2/R3$ - to those best RMSE with $k = \lfloor \min(\widehat{\theta}_0 L, \sqrt{L}) \rfloor$ divided to those best with $k = \lfloor (\ln L)^2 \rfloor$, respectively, for sample size $n = 10^5$.

The impact of the tail heaviness on the discrepancy method accuracy remains an open problem. Intuitively, the heaviness of distribution tail may impact on the rate of convergence of the exceedance point process to a compound Poisson process and hence, on the convergence of discrepancy statistic distribution to A_1, the limit distribution of C-M-S statistic.

5. Application to real data

5.1. *Daily maximum temperatures in Uccle, Belgium*

Following Ferreira (2018a) we consider two data sets of daily maximum temperatures (in 0.1 degrees Celsius) in July in Uccle (Belgium), from 1833 to 1999 and from 1900 to 1999 with sample sizes $n \in \{5177, 3100\}$, respectively, Figure 6. The data are available at "$http://lstat.kuleuven.be/Wiley/Data/ecad00045TX.txt$". The extremal index of the smaller sample was shown to be ranged between 0.49 and 0.56 in Beirlant et al. (2004); an application of bias-reduced version of the Nandagopalan's runs estimator in

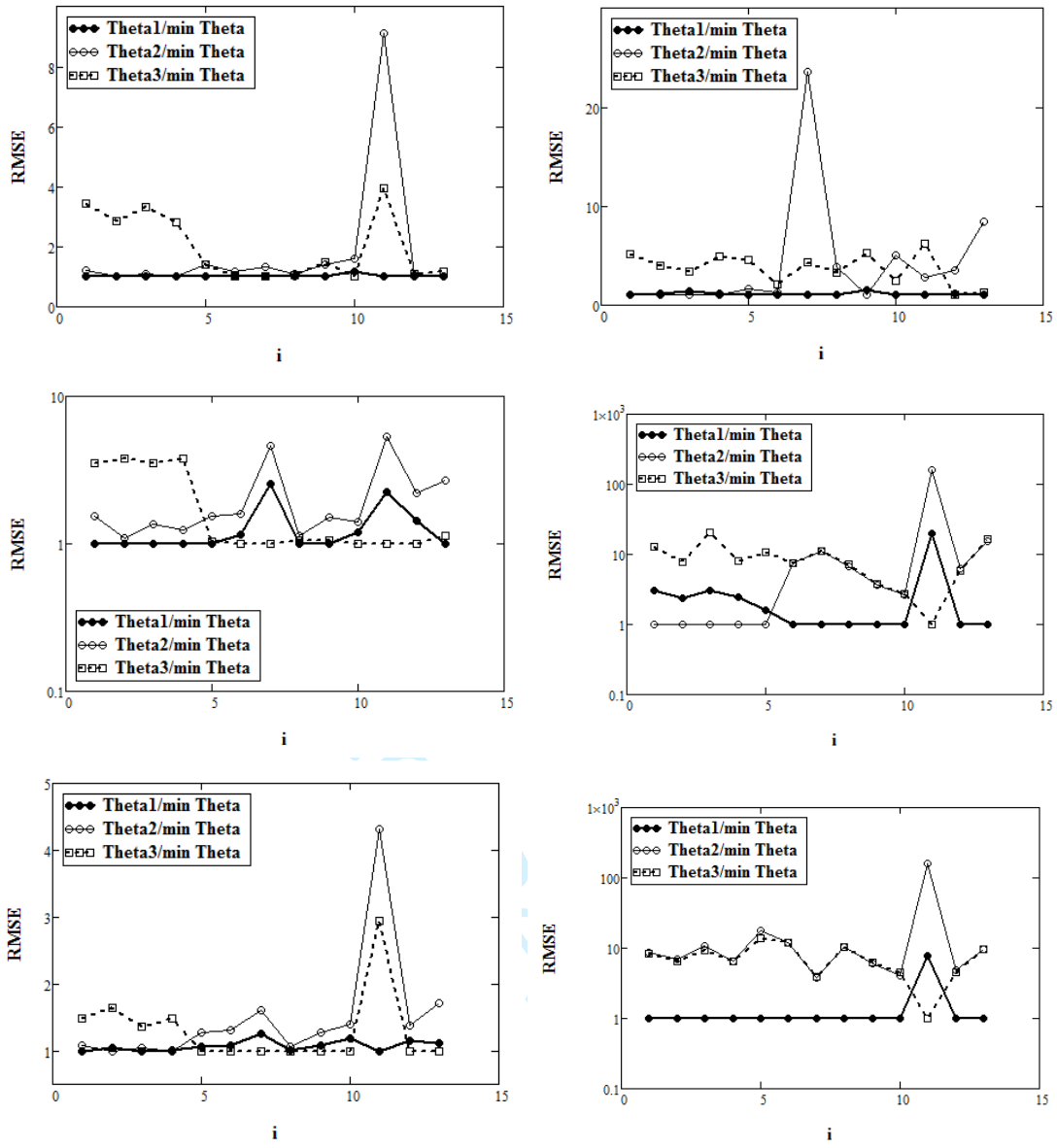

Figure 5. Ratios of the RMSEs $RMSE(\widehat{\theta}_i)/\min_i(RMSE(\widehat{\theta}_i))$ corresponding to estimates $\{\widehat{\theta}_i\}$, $i \in \{1,2,3\}$ in (14) for the intervals estimator (left column) and the K-gaps estimator (right column) obtained by the Algorithm with (15) against the number of processes related to the column labels in Table 1: The upper figures correspond to $k = \lfloor \widehat{\theta}_0 L \rfloor$ in Table 1, the middle figures to $k = \lfloor \min(\widehat{\theta}_0 L, \sqrt{L}) \rfloor$ and the lower figures to $k = \lfloor (\ln L)^2 \rfloor$ for sample size $n = 10^5$.

Ferreira (2018a) gave the values 0.41 and 0.57; and an application of the wide range of estimators gave the values from 0.10 to 0.57 ibid. We apply the intervals and $K-$gaps estimators coupled with the discrepancy method based on Algorithm 4.1. The $K-$gaps estimator coupling with the IMT method and the intervals estimator with "plateau-finding" algorithm A1 with $\omega = 0.3$ were also applied here and in the next example. 'Kdis' and 'Intdis' are calculated for $k = \lfloor sL \rfloor$, where s was taken equal to the pilot intervals estimate $\widehat{\theta}_0$ for each threshold value u or to values $\{0.51, 0.56\}$ for $n \in \{3100, 5177\}$, respectively, based on 'Kimt' estimates and previous estimation of θ provided in Beirlant

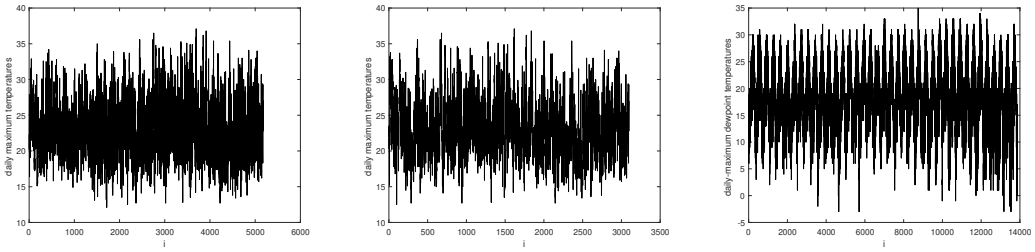

Figure 6. Daily maximum temperatures in July in Uccle, Belgium with sample sizes $n = 5177$ (left) and $n = 3100$ (middle); dewpoint temperatures at station Dhahran, Saudi Arabia with sample size $n = 13866$ (right).

et al. (2004). The discrepancy inequality method (15) is used. One may trust more $\widehat{\theta}_1^*$ as well as 'Kdis' estimate since they provide better results on the simulation. The results are shown in Table 3.

5.2. *Dewpoint temperatures at station Dhahran, Saudi Arabia*

We use the data corresponding to Figure S18 in Raymond et al. (2020) and kindly provided by the authors, which represent daily maximum dewpoint temperatures at station Dhahran, Saudi Arabia, Figure 6. This station is among several selected stations where a wet-bulb temperature (TW) has exceeded $TW = 33^oC$ at least 5 times. The dates span from 1 Jan 1979 to 31 Dec 2017. The sample size is equal to $n = 13866$ due to missing observations. As above it is suggested to trust more $\widehat{\theta}_1^*$ for 'Kdis' estimate. The estimated values of θ are shown in Table 4.

Acknowledgements

The work of N.M. Markovich in Sections 1, 2, 4 and 5 was supported by the Russian Science Foundation (grant No. 22-21-00177). The work of I. V. Rodionov in Section 3 and proofs in Markovich and Rodionov (2022) was performed at the Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences with the support of the Russian Science Foundation (grant No. 21-71-00035).

References

Ancona–Navarrete, M. A., and Tawn, J. A. (2000), 'A comparison of Methods for Estimating the Extremal Index', *Extremes*, 3(1), 5–38.

Balakrishnan, N., and Rao, C. R. (1998), Eds. Handbook of Statistics, Elsevier Science B.V.

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

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

Bolshev, L. N., and Smirnov, N. V. (1965), Tables of Mathematical Statistics, Nauka, Moscow (in Russian)

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.

Drees, H. (2011), 'Bias correction for estimators of the extremal index', *Preprint, arXiv: 1107.0935.*

de Haan, L., and Ferreira, A. (2006), Extreme Value Theory: An Introduction, Springer.

Ferreira, M. (2018a), Heuristic tools for the estimation of the extremal index: a comparison of methods, *REVSTAT – Statistical Journal*, 16(1), 115–136.

Ferreira, M. (2018b), 'Analysis of estimation methods for the extremal index', *Electronic Journal of Applied Statistical Analysis*, 11(1), 296–306.

Ferreira, H. and Ferreira, M. (2018), 'Estimating the extremal index through local dependence', *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 54(2), 587–605.

Ferro, C. A. T., and Segers, J. (2003), 'Inference for Clusters of Extreme Values', *Journal of the Royal Statistical Society Series B.*, 65, 545–556.

Fukutome, S., Liniger, M. A., and Süveges, M. (2015), 'Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland', *Theoretical and Applied Climatology*, 120, 403–416.

Hall, P. (1990), 'Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems', *Journal of Multivariate Analysis*, 32, 177–203.

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.

Kobzar, A. I. (2006), Applied mathematical statistics for engineers and scientists, Fizmatlit, Moscow (in Russian)

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

Leadbetter, M. R., Lingren, G., and Rootzén", H. (1983), Extremes and Related Properties of Random Sequence and Processes, Springer, New York.

Markovich, N. M. (1989), 'Experimental analysis of nonparametric probability density estimates and of methods for smoothing them', *Automation and Remote Control*, 50, 941–948.

Markovich, N. M. (2007), Nonparametric Analysis of Univariate Heavy–Tailed data: Research and Practice, Wiley, Chichester, West Sussex.

Markovich, N. M., Rodionov, I.V. (2022), 'Threshold selection for extremal index estimation', *arxiv.2009.02318.*

Martynov, G. V. (1978), The omega square tests, Nauka, Moscow (in Russian)

Northrop, P. J. (2015), 'An efficient semiparametric maxima estimator of the extremal index', *Extremes*, 18(4), 585–603.

Ramsay, C.M. (2008), 'The Distribution of Sums of I.I.D. Pareto Random Variables with Arbitrary Shape Parameter', *Communications in Statistics - Theory and Methods*, 37(14), 2177–2184.

Raymond, C., Matthews, T., and Horton, R. M. (2020), 'The emergence of heat and humidity too severe for human tolerance', *Science Advances*, 6(19), eaaw1838.

Robert, C. Y. (2009a), 'Asymptotic distributions for the intervals estimators of the extremal

index and the cluster–size probabilities', *Journal of Statistics Planning and Inference*, 139, 3288–3309.

Robert, C. Y. (2009b), 'Inference for the limiting cluster size distribution of extreme values', *The Annals of Statistics*, 37, 271–310.

Robert, C. Y., Segers, J., and Ferro, C. A. T. (2009), 'A sliding blocks estimator for the extremal index', *Electronic Journal of Statistics*, 3, 993–1020.

Smirnov, N. V. (1937), 'On the ω^2-distribution of von Mises', *Matematicheskij Sbornik*, 2(5), 973–993 (in Russian) (French abstract)

Stephens, M. A. (1974), 'EDF statistics for goodness-of-fit and some comparisons', *JASA*, 69, 730–737.

Sun, J., and Samorodnitsky, G. (2010), 'Estimating the extremal index, or, can one avoid the threshold-selection difficulty in extremal inference?', *Technical Report, Cornell University.*

Sun, J., and Samorodnitsky, G. (2019), 'Multiple thresholds in extremal parameter estimation', *Extremes*, 22, 317–341.

Süveges, M. (2007), 'Likelihood estimation of the extremal index', *Extremes*, 10, 41–55.

Süveges, M., and Davison, A. C. (2010), 'Model misspecification in peaks over threshold analysis', *The Annals of Applied Statistics*, 4(1), 203–221.

Vapnik, V. N., Markovich, N. M., and Stefanyuk, A. R. (1992), 'Rate of convergence in L_2 of the projection estimator of the distribution density', *Automation and Remote Control*, 53, 677–686.

Weissman, I., and Novak, S. Yu. (1998), 'On blocks and runs estimators of the extremal index', *Journal of Statistical Planning and Inference*, 66, 281–288.

Table 1. The root mean squared error ($k = \lfloor \widehat{\theta}_0 L \rfloor$), $\widehat{\theta}_0$ is a pilot intervals estimate.

$RMSE$	MM		ARMAX		ARu^+		ARu^-		$MA(2)$		ARc	AR(2)	GARCH
$\cdot 10^4/\theta$	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
$n = 10^5$													
$\widehat{\theta}_1$	147	215	159	211	230	887	287	383	199	***400***	-	**305**	-
$\widehat{\theta}_2$	146	213	158	211	230	889	287	384	201	402	-	**305**	-
$\widehat{\theta}_3$	154	222	164	215	235	886	287	383	203	***400***	-	**305**	-
$\widehat{\theta}_1^*$	**103**	160	89	***151***	292	928	516	***328***	229	542	***17***	400	413
$\widehat{\theta}_2^*$	123	***156***	99	***151***	410	1070	690	354	324	745	155	420	405
$\widehat{\theta}_3^*$	354	442	294	425	413	957	516	351	336	467	67	443	470
$\widehat{\theta}_1^{Kdis}$	133	204	145	190	195	772	209	349	**141**	423	-	390	**375**
$\widehat{\theta}_2^{Kdis}$	777	844	648	940	799	777	***208***	1396	799	420	-	519	***382***
$\widehat{\theta}_3^{Kdis}$	895	702	651	821	398	***763***	**207**	1115	927	637	-	477	389
$\widehat{\theta}_1^{Kdis*}$	136	264	100	237	**150**	**631**	220	**105**	229	**390**	**12**	394	464
$\widehat{\theta}_2^{Kdis*}$	127	226	***70***	207	249	788	5189	404	***152***	1944	34	1231	3929
$\widehat{\theta}_3^{Kdis*}$	652	897	238	1013	690	1300	963	345	798	958	75	***353***	573
$\widehat{\theta}^{Kimt}$	217	569	**69**	498	***173***	844	2501	401	309	466	33	3630	4028
$\widehat{\theta}^{IA1}$	***116***	**122**	95	**113**	447	1193	1756	399	387	977	233	693	580
$n = 5000$													
$\widehat{\theta}_1$	565	938	506	818	783	1431	1364	394	533	816	-	900	1497
$\widehat{\theta}_2$	557	913	476	787	748	1401	1337	396	537	810	-	897	1497
$\widehat{\theta}_3$	633	1013	620	903	879	1490	1416	394	593	850	-	910	1497
$\widehat{\theta}_1^*$	359	***496***	350	***464***	715	1294	1276	***315***	352	760	606	***835***	***955***
$\widehat{\theta}_2^*$	352	**466**	291	**450**	808	1587	1666	395	557	1179	422	**794**	**870**
$\widehat{\theta}_3^*$	1635	1564	1377	1652	1902	1644	1754	713	1505	1656	1455	1610	2036
$\widehat{\theta}_1^{Kdis}$	480	917	496	772	787	1186	1820	427	406	807	-	1690	1877
$\widehat{\theta}_2^{Kdis}$	1525	1880	982	1793	1836	1863	2672	2981	1218	2020	-	1855	2929
$\widehat{\theta}_3^{Kdis}$	1624	1993	1286	1815	1692	1686	2348	1775	1924	2349	-	1737	2337
$\widehat{\theta}_1^{Kdis*}$	320	605	299	507	***453***	**592**	**641**	**213**	404	754	***72***	1528	1491
$\widehat{\theta}_2^{Kdis*}$	***252***	548	***199***	487	535	***866***	2529	423	***335***	***488***	**25**	3684	3787
$\widehat{\theta}_3^{Kdis*}$	824	1007	931	871	1086	927	***916***	555	714	929	2321	1106	1324
$\widehat{\theta}^{Kimt}$	**247**	588	**188**	525	**293**	869	2518	418	**325**	**474**	**25**	3680	3900
$\widehat{\theta}^{IA1}$	385	513	319	478	694	1388	1985	394	514	1114	676	980	1077

Table 2. The absolute bias ($k = \lfloor \widehat{\theta}_0 L \rfloor$), $\widehat{\theta}_0$ is a pilot intervals estimate.

$\lvert Bias \rvert$	MM		ARMAX		ARu^+		ARu^-		$MA(2)$		ARc	AR(2) GARCH	
$\cdot 10^4/\theta$	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
						$n = 10^5$							
$\widehat{\theta}_1$	18	***2.515***	20	21	142	872	203	380	132	333	-	**217**	-
$\widehat{\theta}_2$	16	**2.498**	**16**	***20***	142	874	203	381	134	335	-	**217**	-
$\widehat{\theta}_3$	20	8.680	24	21	141	870	203	379	131	***331***	-	**217**	-
$\widehat{\theta}_1^*$	29	6.723	28	**14**	259	915	469	313	200	513	***17***	354	184
$\widehat{\theta}_2^*$	73	43	72	55	396	1064	669	336	300	719	155	386	198
$\widehat{\theta}_3^*$	***11***	42	66	63	132	858	254	253	***62***	**261**	67	320	170
$\widehat{\theta}_1^{Kdis}$	24	66	***18***	43	***29***	757	48	341	**60**	383	-	323	**14**
$\widehat{\theta}_2^{Kdis}$	133	138	172	145	93	763	***44***	***175***	161	380	-	284	***79***
$\widehat{\theta}_3^{Kdis}$	188	133	181	135	**1.395**	747	50	185	234	423	-	271	92
$\widehat{\theta}_1^{Kdis*}$	103	212	40	194	30	***620***	**43**	**90**	217	368	**12**	364	346
$\widehat{\theta}_2^{Kdis*}$	115	204	51	190	56	786	3497	404	137	448	34	1085	3020
$\widehat{\theta}_3^{Kdis*}$	225	488	34	485	141	**136**	306	321	359	636	75	***222***	249
$\widehat{\theta}^{Kimt}$	**0.148**	567	54	496	165	843	2501	401	306	462	33	3627	4027
$\widehat{\theta}^{IA1}$	82	45	64	54	436	1187	1752	399	378	972	233	687	563
						$n = 5000$							
$\widehat{\theta}_1$	144	336	100	255	339	1284	851	373	46	242	-	345	***11***
$\widehat{\theta}_2$	103	290	54	205	317	1245	811	376	***8.581***	265	-	345	***11***
$\widehat{\theta}_3$	185	391	148	305	360	1322	891	370	81	221	-	***344***	***11***
$\widehat{\theta}_1^*$	99	142	109	112	560	1236	1150	272	170	631	606	665	382
$\widehat{\theta}_2^*$	82	***14***	51	35	681	1534	1574	385	450	1094	422	636	378
$\widehat{\theta}_3^*$	812	779	719	754	992	1192	1021	114	735	658	1455	1047	842
$\widehat{\theta}_1^{Kdis}$	***16***	218	58	98	115	1048	1260	356	41	***109***	-	1186	1254
$\widehat{\theta}_2^{Kdis}$	437	158	290	269	454	727	926	421	305	332	-	853	379
$\widehat{\theta}_3^{Kdis}$	467	165	467	238	341	838	858	***80***	715	564	-	635	154
$\widehat{\theta}_1^{Kdis*}$	169	343	34	309	***108***	***480***	467	135	343	680	72	1458	1342
$\widehat{\theta}_2^{Kdis*}$	192	492	43	443	145	852	2529	423	290	429	***25***	3320	3781
$\widehat{\theta}_3^{Kdis*}$	129	146	295	101	304	534	***157***	110	25	306	2321	582	241
$\widehat{\theta}^{Kimt}$	**2.636**	**4.447**	**3.448**	**3.598**	**5.124**	**1.988**	**1.836**	**1.642**	**2.544**	**5.331**	**1.492**	**2.527**	**7.127**
$\widehat{\theta}^{IA1}$	36	29	***31***	***20***	523	1301	1919	385	361	1033	676	884	876

Table 3. Extremal index estimates for Uccle data.

n	Kimt	IntA1		Intdis		Kdis	
				$s=\widehat{\theta}_0$	$s=0.51$	$s=\widehat{\theta}_0$	$s=0.51$
3100	0.5133	0.4625	$\widehat{\theta}_1^*$	0.5329	0.5741	0.5670	0.5879
			$\widehat{\theta}_2^*$	0.4199	0.4637	0.5232	0.5232
			$\widehat{\theta}_3^*$	0.9575	0.9575	0.7244	0.7244
				$s=\widehat{\theta}_0$	$s=0.56$	$s=\widehat{\theta}_0$	$s=0.56$
5177	0.5695	0.4392	$\widehat{\theta}_1^*$	0.4655	0.4837	0.5632	0.6251
			$\widehat{\theta}_2^*$	0.4184	0.4919	0.5285	0.5662
			$\widehat{\theta}_3^*$	0.5618	0.5618	0.7024	0.7024

Table 4. Extremal index estimates for dewpoint temperatures data.

n	Kimt	IntA1		Intdis	Kdis
				$k=\lfloor\widehat{\theta}_0 L\rfloor$	
13866	0.4753	0.1541	$\widehat{\theta}_1^*$	0.2489	0.3178
			$\widehat{\theta}_2^*$	0.2003	0.1975
			$\widehat{\theta}_3^*$	0.4092	0.5607