A STUDY OF THE JACKKNIFE METHOD IN THE ESTIMATION OF THE EXTREMAL INDEX

Marta Ferreira
Center of Mathematics, University of Minho, Portugal
Center for Computational and Stochastic Mathematics, University of Lisbon, Portugal
Center of Statistics and Applications, University of Lisbon, Portugal
e-mail: *msferreira@math.uminho.pt*

Key words: Extreme value theory, Jackknife methodology, Local dependence conditions, Statistics of extremes.

Abstract: Clustering of high values occurs in many real situations and affects inference on extremal events. For stationary dependent sequences, under general local and asymptotic dependence conditions, the degree of clustering is measured through a parameter called the *extremal index*. The estimation of extreme events or parameters is usually based on a k number of top order statistics or on the exceedances of a high threshold u and is very sensitive to either of these choices. In particular, the bias increases with a growing k and a decreasing u. The use of the Jackknife methodology may help reduce bias. We analyse this method through a simulation study applied to several estimators of the extremal index. An application to real data sets illustrates the results.

1. Introduction

Let $\{X_n\}_{n\geq 1}$ be a stationary sequence with common distribution function (df) F, $M_{i,j} = \max(X_{i+1},\ldots,X_j)$, $M_{0,j} = M_j$ and $M_{i,j} = -\infty$ for $i > j$. We say that $\{X_n\}_{n\geq 1}$ has extremal index $\theta \in [0,1]$ if, for every real $\tau > 0$, there exists a sequence of thresholds $\{u_n \equiv u_n^{(\tau)}\}_{n\geq 1}$ such that

$$n(1-F(u_n)) \to \tau \tag{1}$$

and $P(M_n \leq u_n) \to \exp(-\theta\tau)$, as $n \to \infty$. A sequence satisfying (1) is usually indicative of normalised levels. The long range dependence condition $\mathrm{D}(u_n)$ of Leadbetter (1974), states that $\alpha_{n,l_n} \to 0$, as $n \to \infty$, for some sequence $l_n = o(n)$, where

$$\begin{aligned}\alpha_{n,l} = \sup\{&|P(M_{i_1,i_1+p} \leq u_n, M_{j_1,j_1+q} \leq u_n) - P(M_{i_1,i_1+p} \leq u_n)P(M_{j_1,j_1+q} \leq u_n)|:\\ &1 \leq i_1 < i_1+p+l \leq j_1 < j_1+q \leq n\}.\end{aligned}$$

If $\{X_n\}_{n\geq 1}$ satisfies $\mathrm{D}(u_n)$ for each positive τ within normalised levels (1) and $P(M_n \leq u_n)$ converges for some $\tau > 0$, then $P(M_n \leq u_n) \to \exp(-\theta\tau)$ for all $\tau > 0$ and $\{X_n\}_{n\geq 1}$ has extremal

AMS: 60G70, 62G32

index θ (Leadbetter and Rootzén, 1988). Condition D(u_n) establishes asymptotic independence as the extreme values become increasingly distant and is required for the local dependence conditions $\mathrm{D}^{(s)}(u_n)$ of Chernick, Hsing and McCormick (1991). Indeed, this latter holds for $\{X_n\}_{n\geq 1}$ satisfying D(u_n), if for some $\{b_n\}_{n\geq 1}$ such that,

$$b_n \to \infty,\, b_n\alpha_{n,l_n} \to 0,\, b_n l_n/n \to 0,$$

as $n \to \infty$, we have

$$nP(X_1 > u_n, M_{1,s} \leq u_n < M_{s,r_n}) \underset{n\to\infty}{\longrightarrow} 0,$$

with $\{r_n = [n/b_n]\}_{n\geq 1}$ ($[x]$ denotes the integer part of x). Condition $\mathrm{D}^{(s)}(u_n)$ implies $\mathrm{D}^{(t)}(u_n)$ for all $t > s$ and is implied by

$$n\sum_{j=s+1}^{r_n} P(X_1 > u_n, M_{1,s} \leq u_n < X_j) \underset{n\to\infty}{\longrightarrow} 0.$$

Condition $\mathrm{D}^{'}(u_n)$ of Leadbetter, Lindgren and Rootzén (1983) corresponds to $s = 1$ and restricts the occurrence of clusters of exceedances resembling an i.i.d. behaviour. Therefore we have a unit extremal index. The case $s = 2$ corresponds to condition $\mathrm{D}^{''}(u_n)$ of Leadbetter and Nandagopalan (1989). Under this condition, we have clustering of exceedances but a restriction on the occurrence of upcrossings.

If $\{X_n\}_{n\geq 1}$ satisfies $\mathrm{D}^{(s)}(u_n)$, we also conclude that the extremal index exists, given by

$$\theta = \lim_{n\to\infty} \theta(u_n, s) \equiv \lim_{n\to\infty} P(M_{1,s} \leq u_n | X_1 > u_n) \tag{2}$$

(see Chernick et al., 1991). This interpretation of the extremal index meets O'Brien (1987) characterisation where $\theta = \lim_{n\to\infty} \theta(u_n, r_n)$, with $r_n = o(n)$.

When extending the analysis of i.i.d. sequences to stationary ones the extremal index is a key parameter that influences the estimation of extremal properties. For instance, missing θ may lead us to underestimate high quantiles (see, Prata-Gomes and Neves, 2015). The characterisations above concerning existence and derivation of the extremal index allows the development of inference methods. The most common approach in statistics of extremes is conducted under a semi-parametric framework. The estimators are thus based on a number k of upper order statistics requiring a trade-off between variance and bias. More precisely, the variance decreases and the bias increases with increasing k. Contributions in the literature towards methods for bias reduction and stability along a substantial amount of thresholds (avoiding the increment of variance) are welcome. The Generalised Jackknife methodology revealed promising results in this context concerning the estimation of the extremal index (Gomes, Hall and Miranda, 2008; Prata-Gomes and Neves, 2015; Neves, Gomes, Figueiredo and Prata-Gomes, 2015). However, the method was only exploited for a simple estimator that holds under condition $\mathrm{D}^{''}(u_n)$. Financial time series, for instance, are commonly well modelled by GARCH processes where condition $\mathrm{D}^{''}(u_n)$ is quite implausible to hold (see Ferreira and Ferreira, 2015, and references therein). Here we analyse the application of the Jackknife method to other extremal index estimators which work under the more general condition $\mathrm{D}^{(s)}(u_n)$. The description of the methods is presented in Section 2. Our study is based on intensive simulation comprising several models and is conducted in Section 3. In Section 4 we illustrate our work through an application to real data sets within the areas of environment and finance. A small discussion concludes the paper in Section 5.

2. Estimators and Generalised Jackknife method

Classical estimators of θ correspond to the ratio between the number of independent clusters ($C_n(u_n)$) and the number of exceedances of a high threshold u_n ($N_n(u_n)$), that is,

$$\widehat{\theta} = \frac{C_n(u_n)}{N_n(u_n)}. \tag{3}$$

Different definitions of clusters lead to different estimators. Considering the well-known runs estimator, strongly motivated by O'Brien (1987) characterisation, two different groups of exceedances of u_n are identified as independent clusters if there are at least $r-1$ consecutive observations below the threshold between them. Thus, the runs estimator is defined by (3) where

$$C_n(u_n) \equiv C_n^R(u_n) = \sum_{i=1}^{n-r+1} \mathbb{1}_{\{X_i>u_n\}} \mathbb{1}_{\{X_{i+1}\leq u_n\}} \cdots \mathbb{1}_{\{X_{i+r-1}\leq u_n\}}.$$

Observe also that the runs estimator corresponds to the empirical counterpart of Chernick et al. (1991) formulation in (2), by taking $r = s$. In the sequel we denote it by $\widehat{\theta}^R$.

The blocks estimator (Leadbetter, 1983) is also defined by (3), where clusters correspond to blocks of length r_n ($r_n = o(n)$) where at least one exceedance of u_n occurs. Asymptotic properties of these estimators are derived in Hsing (1991, 1993) as well as in Smith and Weissman (1994) and Weissman and Novak (1998), where comparisons lead to the preference of the runs estimator. Other estimators were also proposed in the literature, e.g., maximum likelihood procedures (Ancona-Navarrete and Tawn, 2000; Süveges, 2007), a two-threshold estimator (Laurini and Tawn, 2003), and an intervals estimator (Ferro and Segers, 2003). More recently, "cycled"-type estimators were derived in Ferreira and Ferreira (2015). More precisely, if $\{X_n\}_{n\geq 1}$ satisfies condition $\mathrm{D}^{(s)}(u_n)$, we have that $\{Z_n\}_{n\geq 1}$ such that $Z_n = \bigvee_{j=(n-1)(s-1)+1}^{n(s-1)} X_j$, $n \geq 1$, is a sequence of cycles satisfying condition $\mathrm{D}^{(2)}(u_n)$ and we can estimate θ directly through

$$\widehat{\theta} = \frac{U_n^Z(u_n)}{N_n(u_n)}, \tag{4}$$

or indirectly through

$$\widehat{\theta} = \frac{\widehat{\theta}_Z N_n^Z(u_n)}{N_n(u_n)}, \tag{5}$$

where $U_n^Z(u_n)$ and $N_n^Z(u_n)$ are, respectively, the number of upcrossings of u_n and the number of exceedances of u_n within $\{Z_1,\ldots,Z_{[n/(s-1)]}\}$. The direct and indirect "cycled"-type estimators in (4) and (5) will be denoted $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$, respectively. Observe that $\widehat{\theta}^R$, $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$ work under condition $\mathrm{D}^{(s)}(u_n)$ which makes them natural competitors. This was corroborated in the simulation study conducted in Ferreira and Ferreira (2015). Moreover, estimators developed under $\mathrm{D}^{(2)}(u_n)$ can be used to calculate $\widehat{\theta}_Z$ in (5), since $\{Z_n\}_{n\geq 1}$ satisfies the condition of the upcrossings, namely the upcrossings estimator in Nandagopalan (1990), among others (for more details, see Ferreira and Ferreira, 2015, and references therein).

Resampling techniques like bootstrap and jackknife revealed promising results within extreme values inference, specially due to the scarce data provided by the tails. With respect to the estimation of θ, these methods have been applied to the Nandagopalan's estimator and, as far as we know, they are confined to this latter. The reason for this lies mainly with its simple calculation resulting in the ratio between the number of upcrossings and the number of exceedances of a large threshold. However, real data is not always likely to satisfy condition $\mathrm{D}^{(2)}(u_n)$. Think, for instance, in financial time series that usually present high volatility. The bootstrap methodology was analysed in Prata-Gomes and Neves (2015) and a generalised jackknife estimator was developed in Gomes et al. (2008). See also Neves et al. (2015) and Prata-Gomes and Neves (2015). Bootstrap is based on a computer-intensive resampling technique where extracting single observations within an independent scheme is replaced by block-resampling in a dependent context. Different ways of blocking mimics different features of the dependence structure of the data in the resampled one. The block length is an important parameter and is highly sensitive to the context. Different proposals are found in Hall, Horowitz and Jing (1995), Lahiri, Furukawa and Lee (2007), among others. A survey on this topic is presented in Prata-Gomes and Neves (2015).

If we consider u_n as a deterministic level in $[X_{n-k:n}, X_{n-k+1:n})$, where $X_{1:n} \leq \ldots \leq X_{n:n}$ are the order statistics of sample $(X_1, \ldots, X_n)$, we have the estimators as functions of k, i.e., $\widehat{\theta} \equiv \widehat{\theta}(k)$, $k = 1, \ldots, n-1$. We are thus reproducing a similar context of a semi-parametric estimation of other extremal parameters like the well-know tail index. In order to achieve consistency, $k \equiv k_n$ must be an intermediate sequence, that is, $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$. The given estimators present strong bias (see the left panel of Figures 2 – 7), particularly as k increases. The choice of an “optimal" k is difficult because it requires a trade-off between variance and bias (the variance is large in the beginning of the tail where few observations are used).

Consider three biased estimators, $\theta^{(1)}$, $\theta^{(2)}$ and $\theta^{(3)}$, for the parameter θ, each one with two dominant components within the bias, i.e.,

$$E(\widehat{\theta}^{(i)} - \theta) = g_1(\theta)\left(\frac{k}{n}\right) + g_2(\theta)\left(\frac{1}{k}\right) + o\left(\frac{k}{n}\right) + o\left(\frac{1}{k}\right), i = 1,2,3, \tag{6}$$

then the (second order) generalised jackknife (GJ) estimator is defined by

$$\widehat{\theta}_{GJ} := \frac{|M_1(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}, \widehat{\theta}^{(3)})|}{|M_1(1,1,1)|}, \tag{7}$$

where $|\cdot|$ denotes the determinant of a matrix and

$$M_1(\alpha_1, \alpha_2, \alpha_3) = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ g_1^{(1)}(n) & g_1^{(2)}(n) & g_1^{(3)}(n) \\ g_2^{(1)}(n) & g_2^{(2)}(n) & g_2^{(3)}(n) \end{pmatrix}.$$

This method is developed in Gray and Schucany (1972). It is not difficult to conclude that $\widehat{\theta}_{GJ}$ is an unbiased estimator for the parameter θ. In Gomes et al. (2008) we can see an illustration of the validity of (6) for some dependent models in the case of the Nandagopalan's estimator, i.e., the runs estimator $\widehat{\theta}^R$ with parameter $r = 2$. Thus, in accordance with (7), the authors propose a GJ estimator for the extremal index based on three levels, k, $[\delta k]+1$ and $[\delta^2 k]+1$, where $\delta \in (0,1)$ is a tuning parameter. More precisely,

$$\widehat{\theta}_{GJ}(k) = \frac{|M_1(\widehat{\theta}^R([\delta^2 k]+1), \widehat{\theta}^R([\delta k]+1), \widehat{\theta}^R(k))|}{|M_1(1,1,1)|},$$

with $g_1^{(i)}(n) = \delta^{3-i}$ and $g_2^{(i)}(n) = 1/g_1^{(i)}(n)$, $i = 1,2,3$, leading to

$$\widehat{\theta}_{GJ}(k,\delta) = \frac{(\delta^2+1)\widehat{\theta}^R([\delta k]+1) - \delta(\widehat{\theta}^R([\delta^2 k]+1) + \widehat{\theta}^R(k))}{(1-\delta)^2}. \tag{8}$$

Here we apply the GJ second order methodology to the runs estimator for any value r, as well as to the cycled-type estimators $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$ mentioned above. Our analysis consists of an intensive simulation study applied to several models with diverse dependence structures, namely:

- Max-autoregressive process (MAR): $X_i = \phi X_{i-1} \vee \varepsilon_i$, with $0 < \phi < 1$, $\{\varepsilon_i\}_{i\geq 1}$ an i.i.d. sequence of r.v.'s with d.f. $F_\varepsilon(x) = \exp(-(1-\phi)/x)$, $x > 0$ and $\theta = 1-\phi$; we consider $\phi = 1/2$ ($\theta = 1/2$).

- Moving maxima (MM), $X_i = \bigvee_{j=0,\dots,m} \alpha_j \varepsilon_{i-j}$, with $\sum_{j=0}^m \alpha_j = 1$, $\alpha_j \geq 0$, $\{\varepsilon_i\}_{i\geq 1}$ an i.i.d. sequence of unit Fréchet distributed r.v.'s and $\theta = \vee_{j=0,\dots,m}\alpha_j$; we consider $m = 3$, and two cases:

  MM(I): $\alpha_0 = 1/6$, $\alpha_1 = 1/2$, $\alpha_2 = 1/3$ ($\theta = 1/2$);

  MM(II): $\alpha_0 = 1/3$, $\alpha_1 = 1/6$, $\alpha_2 = 1/2$ ($\theta = 1/2$).

- Autoregressive Gaussian (AR): $X_i = \beta X_{i-1} + \varepsilon_i$, with $|\beta| < 1$, $\{\varepsilon_i\}_{i\geq 1}$ an i.i.d. sequence of $N(0, 1-\alpha^2)$ distributed r.v.'s ($\theta = 1$).

- Autoregressive Cauchy (ARCauchy): $X_i = \beta X_{i-1} + \varepsilon_i$, $|\beta| < 1$ and $\theta = 1-\beta^2$; we consider $\beta = -3/5$ ($\theta = 0.64$).

- Uniform autoregressive (ARUnif): $X_i = -(1/m)X_{i-1} + \varepsilon_i$, with $\{\varepsilon_i\}_{i\geq 1}$ an i.i.d. sequence, $P(\varepsilon_1 = j/m) = 1/m$ for $j = 1,\dots,m$ and $\theta = 1-1/m^2$; we consider $m = 2$ ($\theta = 3/4$).

- Bivariate extreme value Markov (MCBEV): $P(X_i \leq x, X_{i+1} \leq y) = \exp(-(x^{1/\gamma} + y^{1/\gamma})^\gamma)$; we consider $\gamma = 0.5$ ($\theta = 0.328$) .

- GARCH(1,1): $X_i = \sigma_i \varepsilon_i$, with $\sigma_i^2 = \alpha + \lambda X_{i-1}^2 + \beta\sigma_{i-1}^2$, $\alpha, \lambda, \beta > 0$, with $\{\varepsilon_i\}_{i\geq 1}$ an i.i.d. sequence of standard Gaussian r.v.'s; we consider $\alpha = 10^{-6}$, $\lambda = 1/4$ and $\beta = 7/10$ ($\theta = 0.447$).

Figure 1 illustrates a sample path of each model. It is proved in the literature that condition $D'(u_n)$ holds for AR (Leadbetter et al., 1983), condition $D^{(2)}(u_n)$ holds for MAR (Hall, 1996) and MM(I) (Ferreira and Ferreira, 2015) and condition $D^{(3)}(u_n)$ holds for models MM(II) (Ferreira and Ferreira, 2015), ARCauchy and ARUnif (Chernick et al., 1991). In Ferreira and Ferreira (2015) conditions $D^{(4)}(u_n)$ and $D^{(5)}(u_n)$ were (empirically) validated for models MCBEV and GARCH(1,1), respectively.

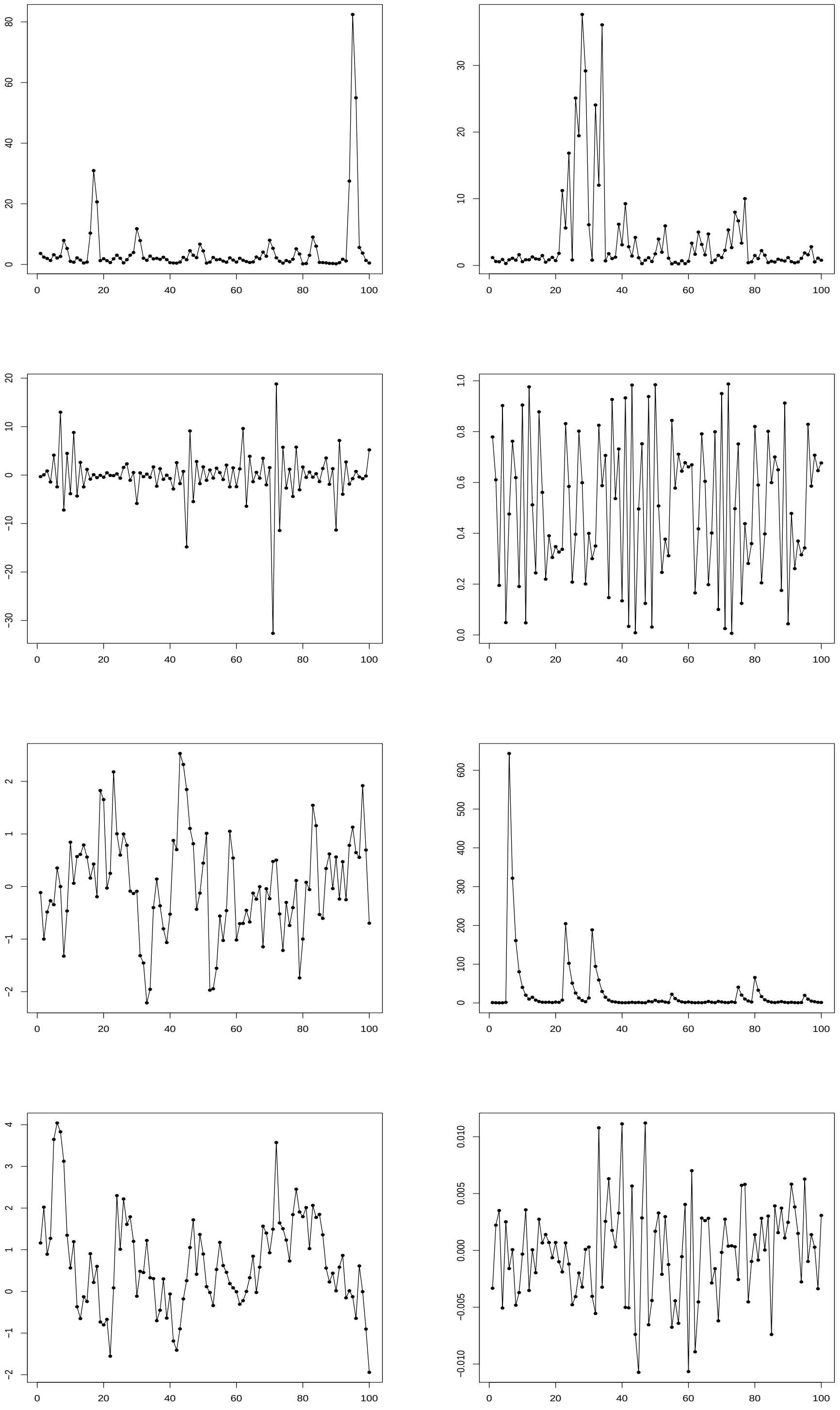

Figure 1: Sample paths of models (left-to-right and top-to-bottom): MM(I), MM(II), ARCauchy, ARUnif, AR, MAR, MCBEV, GARCH.

We now apply the runs estimator $\widehat{\theta}^R \equiv \widehat{\theta}^R(k)$, $k = 1, \ldots, n-1$, by taking $r = s$ of the $\mathrm{D}^{(s)}(u_n)$ condition validated in the respective model, and the same value of s is considered within the cycle-type estimators $\widehat{\theta}^{CD} \equiv \widehat{\theta}^{CD}(k)$ and $\widehat{\theta}^{CI} \equiv \widehat{\theta}^{CI}(k)$, $k = 1, \ldots, n-1$, given in, respectively, (4) and (5). In each case, we also compute the GJ associated estimator as in (8), by replacing $\widehat{\theta}^R$ by $\widehat{\theta}^{CD}$ or $\widehat{\theta}^{CI}$, accordingly. Observe that $\widehat{\theta}^{CI}$ is based on the estimation of the extremal index θ_Z of cycles $\{Z_n\}_{n\geq 1}$ where condition $\mathrm{D}^{(2)}(u_n)$ holds and thus we derive $\widehat{\theta}_Z$ through the Nandagopalan's estimator. In addition, a second GJ variation of $\widehat{\theta}^{CI}$ is implemented by applying the GJ estimator only to $\widehat{\theta}_Z$. We will use the notation $\widehat{\theta}^R_{GJ}$, $\widehat{\theta}^{CD}_{GJ}$, $\widehat{\theta}^{CI}_{GJ}$ and $\widehat{\theta}^{CI_2}_{GJ}$ for the respective GJ versions of $\widehat{\theta}^R$, $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$ (the CI_2 version corresponds to the second GJ variation referred to above).

It will be found that, in this new context, the GJ methodology continues to fulfil its main goal of improving the inference by combining information coming from the observed data. In particular, we will see that the bias assumption in (6) is quite general and reasonable, since the subsequent second order GJ estimation seems to work for diverse estimators and models.

3. Simulation study

Our study is based on 1000 replicas generated from the models above concerning samples of sizes $n = 500, 1000, 5000$ (we do not consider smaller samples, since they may compromise the performance of the cycled-type estimators; see Ferreira and Ferreira, 2015). Previous simulation analysis corroborate the choice $\delta = 1/4$ in (8) suggested in Gomes et al. (2008), and thus we assume $\widehat{\theta}_{GJ}(k) \equiv \widehat{\theta}_{GJ}(k, 1/4)$.

The sample paths of the estimated absolute bias (abias) and root mean squared error (rmse) for samples of size $n = 1000$ are plotted in Figures 2 – 7. In order to evaluate the GJ methodology when compared to the usual estimation, we also compute indicators of the eventual reduction within the bias and the rmse, as well as an indicator of the increase in the sample path stability considered in Gomes et al. (2008). More precisely, we estimate the optimal number of top order statistics to consider, in the sense of $k_o = \arg\min_k mse(\widehat{\theta}(k))$. The bias reduction, the relative efficiency and the sample paths stability indicators are thus given by, respectively,

$$BR = \frac{abias_o}{abias_o^{GJ}}, RE = \sqrt{\frac{mse_o}{mse_o^{GJ}}} \text{ and } SPS = \frac{\sum_{k=1}^{n-1} \mathbb{1}_{\{abias(\widehat{\theta}_{GJ}(k)) \leq 0.01\}}}{\sum_{k=1}^{n-1} \mathbb{1}_{\{abias(\widehat{\theta}(k)) \leq 0.01\}}},$$

where $abias_o \equiv abias(\widehat{\theta}(k_o))$, $abias_o^{GJ} \equiv abias(\widehat{\theta}_{GJ}(k_o^{GJ}))$, and similarly for the mse. Observe that larger values indicate that GJ is the better estimator.

The simulation results are reported in Tables 1 – 3.

Table 1: Simulation results for $n = 500$.

R	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	33	30	30	123	1	56	43	24
k_o^{GJ}	416	375	375	124	1	416	165	124
$abias_o$	0.0146	0.0269	0.0365	0.0082	0.0000	0.0306	0.0587	0.0883
$abias_o^{GJ}$	0.0233	0.0294	0.0150	0.0055	0.0000	0.0261	0.0442	0.0813
$rmse_o$	0.0517	0.0642	0.0899	0.0349	0.0000	0.0669	0.0875	0.1200
$rmse_o^{GJ}$	0.0659	0.0706	0.0807	0.1259	0.0000	0.0784	0.1193	0.1478
BR	0.6266	0.9150	2.4333	1.4909	1.0000	1.1724	1.3281	1.0861
RE	0.7845	0.9093	1.1140	0.2772	1.0000	0.8533	0.7334	0.8119
SPS	22.7692	21.6250	45.2000	0.7857	1.0000	13.0000	1.6000	1.0000
CD	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	28	23	30	55	1	38	170	72
k_o^{GJ}	243	249	244	112	1	246	1662	1249
$abias_o$	0.0308	0.0232	0.0512	0.0188	0.0090	0.0417	0.0400	0.0780
$abias_o^{GJ}$	0.0216	0.0279	0.0054	0.0295	0.0090	0.0149	0.0308	0.0755
$rmse_o$	0.0654	0.0687	0.0967	0.0565	0.0949	0.0811	0.0548	0.0947
$rmse_o^{GJ}$	0.0823	0.0832	0.0976	0.1345	0.0949	0.0888	0.0481	0.0850
BR	1.4259	0.8315	9.4815	0.6373	1.0000	2.7987	1.2987	1.0331
RE	0.7947	0.8257	0.9908	0.4201	1.0000	0.9133	1.1393	1.1141
SPS	18.8000	18.8300	54.2500	0.5161	1.0000	26.3333	1.0000	2.6667
CI	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	23	23	30	55	1	27	25	16
k_o^{GJ}	248	244	232	112	1	240	165	123
$abias_o$	0.0343	0.0284	0.0518	0.0188	0.0090	0.0334	0.0551	0.1003
$abias_o^{GJ}$	0.0075	0.0046	0.0279	0.0295	0.0090	0.0067	0.0143	0.0416
$rmse_o$	0.0672	0.0676	0.0971	0.0565	0.0949	0.0847	0.0879	0.1338
$rmse_o^{GJ}$	0.0749	0.0751	0.1031	0.1345	0.0949	0.0848	0.0820	0.1110
BR	4.5733	6.1739	1.8566	0.6373	1.0000	4.9851	3.8531	2.4111
RE	0.8972	0.9001	0.9418	0.4201	1.0000	0.9988	1.0720	1.2054
SPS	44.2000	45.0000	43.7500	0.5161	1.0000	44.4000	1.6667	2.0000
CI_2	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o^{GJ}	247	247	247	248	1	247	165	123
$abias_o^{GJ}$	0.0434	0.0587	0.0174	0.0033	0.0090	0.0405	0.0304	0.0639
$rmse_o^{GJ}$	0.0732	0.0797	0.0897	0.0695	0.0949	0.0754	0.0693	0.0917
BR	0.7903	0.4838	2.9770	5.6970	1.0000	0.8247	1.8125	1.5696
RE	0.9180	0.8482	1.0825	0.8129	1.0000	1.1233	1.2684	1.4591
SPS	3.0000	3.2000	17.7500	0.7097	1.0000	7.0000	2.6667	4.0000

Table 2: Simulation results for $n = 1000$.

R	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	73	42	54	246	1	99	121	139
k_o^{GJ}	732	631	708	240	1	804	880	752
$abias_o$	0.0225	0.0183	0.0417	0.0046	0.0000	0.0309	0.0240	0.0081
$abias_o^{GJ}$	0.0165	0.0230	0.0039	0.0037	0.0000	0.0197	0.0077	0.0031
$rmse_o$	0.0414	0.0501	0.0766	0.0259	0.0000	0.0578	0.0496	0.0413
$rmse_o^{GJ}$	0.0496	0.0550	0.0562	0.0883	0.0000	0.0576	0.0480	0.0559
BR	1.3636	0.7957	10.6923	1.2432	1.0000	1.5685	3.1169	2.6129
RE	0.8347	0.9109	1.3630	0.2933	1.0000	1.0035	1.0333	0.7388
SPS	19.7778	23.2222	46.5000	0.9772	1.0000	16.6129	6.5161	2.4667
CD	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	46	38	47	108	1	55	70	35
k_o^{GJ}	483	471	496	212	1	499	332	247
$abias_o$	0.0278	0.0218	0.0448	0.0104	0.0080	0.0310	0.0502	0.0822
$abias_o^{GJ}$	0.0160	0.0229	0.0040	0.0281	0.0080	0.0190	0.0376	0.0764
$rmse_o$	0.0538	0.0534	0.0814	0.0402	0.0894	0.0705	0.0728	0.1087
$rmse_o^{GJ}$	0.0600	0.0608	0.0688	0.0981	0.0894	0.0691	0.0884	0.1193
BR	1.7375	0.9520	11.2000	0.3701	1.0000	1.6316	1.3351	1.0759
RE	0.8967	0.8783	1.1831	0.4098	1.0000	1.0203	0.8235	0.9111
SPS	23.4286	25.9091	57.2500	0.9642	2.0000	25.2143	1.2500	4.0000
CI	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	42	33	48	108	1	45	37	24
k_o^{GJ}	496	492	416	212	1	496	332	243
$abias_o$	0.0344	0.0206	0.0457	0.0104	0.0080	0.0315	0.0430	0.0913
$abias_o^{GJ}$	0.0078	0.0049	0.0159	0.0281	0.0080	0.0060	0.0133	0.0349
$rmse_o$	0.0548	0.0517	0.0814	0.0402	0.0894	0.0734	0.0708	0.1210
$rmse_o^{GJ}$	0.0533	0.0511	0.0761	0.0981	0.0894	0.0583	0.0589	0.0812
BR	4.4103	4.2041	2.8742	0.3701	1.0000	5.2500	3.2331	2.6160
RE	1.0281	1.0107	1.0696	0.4098	1.0000	1.2590	1.2020	1.4901
SPS	43.2727	46.8000	43.7500	0.9643	2.0000	42.5455	2.3333	1.0000
CI_2	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o^{GJ}	499	307	499	499	1	499	331	243
$abias_o^{GJ}$	0.0413	0.0438	0.0171	0.0004	0.0080	0.0427	0.0302	0.0594
$rmse_o^{GJ}$	0.0577	0.0665	0.0639	0.0492	0.0894	0.0604	0.0529	0.0760
BR	0.8329	0.4703	2.6725	26.0000	1.0000	0.7377	1.4238	1.5370
RE	0.9497	0.7774	1.2739	0.8171	1.0000	1.2152	1.3384	1.5921
SPS	7.2727	5.9000	11.7500	1.1071	1.0000	6.9091	2.3333	4.0000

Table 3: Simulation results for $n = 5000$.

R	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	211	106	137	1235	1	281	420	646
k_o^{GJ}	3260	2148	3396	1232	1	3048	4275	3760
$abias_o$	0.0137	0.0116	0.0239	0.0012	0.0000	0.0181	0.0106	0.0008
$abias_o^{GJ}$	0.0126	0.0093	0.0024	0.0020	0.0000	0.0117	0.0043	0.0001
$rmse_o$	0.0244	0.0306	0.0466	0.0112	0.0000	0.0344	0.0289	0.0198
$rmse_o^{GJ}$	0.0248	0.0271	0.0261	0.0388	0.0000	0.0290	0.0222	0.0241
BR	1.0873	1.2473	9.9583	0.6000	1.0000	1.5470	2.4651	8.0000
RE	0.9839	1.1292	1.7854	0.2887	1.0000	1.1862	1.3018	0.8216
SPS	19.1157	29.7808	56.8478	1.0016	2.0000	20.2101	5.2062	2.3590

CD	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	141	106	116	546	1	195	170	72
k_o^{GJ}	2097	1919	2476	1028	1	2368	1662	1249
$abias_o$	0.0180	0.0147	0.0238	0.0035	0.0010	0.0255	0.0400	0.0780
$abias_o^{GJ}$	0.0124	0.0142	0.0065	0.0276	0.0010	0.0149	0.0308	0.0755
$rmse_o$	0.0299	0.0320	0.0493	0.0177	0.0316	0.0427	0.0548	0.0947
$rmse_o^{GJ}$	0.0296	0.0313	0.0317	0.0500	0.0316	0.0323	0.0481	0.0850
BR	1.4516	1.0352	3.6615	0.1268	1.0000	1.7114	1.2987	1.0331
RE	1.0101	1.0224	1.5552	0.3540	1.0000	1.3220	1.1393	1.1141
SPS	26.5942	29.5636	60.1707	1.0037	2.0000	25.4744	1.0000	2.6667

CI	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o	110	87	116	546	1	146	105	54
k_o^{GJ}	2456	2336	1896	1028	1	2468	1664	1247
$abias_o$	0.0185	0.0131	0.0238	0.0035	0.0010	0.0235	0.0402	0.0829
$abias_o^{GJ}$	0.0073	0.0033	0.0094	0.0276	0.0010	0.0083	0.0082	0.0377
$rmse_o$	0.0308	0.0314	0.0493	0.0177	0.0316	0.0435	0.0552	0.0983
$rmse_o^{GJ}$	0.0235	0.0240	0.0371	0.0500	0.0316	0.0270	0.0267	0.0501
BR	2.5342	3.9697	2.5319	0.1268	1.0000	2.8313	4.9024	2.1989
RE	1.3106	1.3083	1.3288	0.3540	1.0000	1.6111	2.0674	1.9621
SPS	47.6154	49.4000	46.9500	1.0037	2.0000	41.6949	4.3636	4.0000

CI_2	MM(I)	MM(II)	ARCauchy	ARUnif	AR	MAR	MCBEV	GARCH
k_o^{GJ}	947	672	2451	2464	1	1207	1665	1247
$abias_o^{GJ}$	0.0219	0.0206	0.0148	0.0056	0.0010	0.0249	0.0301	0.0606
$rmse_o^{GJ}$	0.0369	0.0374	0.0319	0.0257	0.0316	0.0382	0.0357	0.0641
BR	0.8447	0.6359	1.6081	0.6250	1.0000	0.9438	1.3355	1.3680
RE	0.8347	0.8396	1.5455	0.6887	1.0000	1.1387	1.5462	1.5335
SPS	8.9808	5.8400	11.8000	1.1284	3.0000	7.4068	1.8182	3.0000

The AR model has a dependent structure with $\theta = 1$. In these cases, all estimators are biased since they actually compute $\theta(k^*) < \theta = 1$ for some k^* (see Ancona-Navarrete and Tawn, 2000). Although the GJ methodology allows for a reduction of the bias and the rmse, it still underestimates the extremal index.

In the following comments we always exclude this case. Observe that the least rmse produced by the estimated optimal level $X_{n-k_o:n}$ tends to relapse on the runs estimator, both in its simple form and based on the GJ procedure. When comparing the usual approach with the respective GJ procedure, the rmse at the “optimal" level always decreases in the ARCauchy whilst the opposite occurs with model ARUnif. The improvement is more evident for large sample sizes and models MAR and MCBEV. In the GARCH model, the runs estimator always presents smaller rmse than the GJ version. The largest BR, for $n \leq 1000$, is registered in estimator $\widehat{\theta}^{CI}$ (except in models ARCauchy and ARUnif, where this is observed, respectively, in $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI_2}$). For $n = 5000$ the conclusions change for models ARCauchy and GARCH where $\widehat{\theta}^{R}$ is better. With regards to efficiency, the largest RE seems mostly associated to the CI and CI$_2$ estimators. The stability indicator is low for model ARUnif. Indeed, the last line panels of Figures 2, 4 and 6 show no improvement of the GJ method within a large range of levels in the ARUnif model, suggesting that the bias assumption (6) may not hold in this case. The largest SPS is observed for estimator CI in the majority of the cases. Although the GJ method does not necessarily lead to improvements in all values of the indicators nor smaller bias and rmse at optimal levels, the stability of the trajectories around the true value over a wider range of levels, observed in Figures 2 – 7, is of vital importance with regard to practical applications.

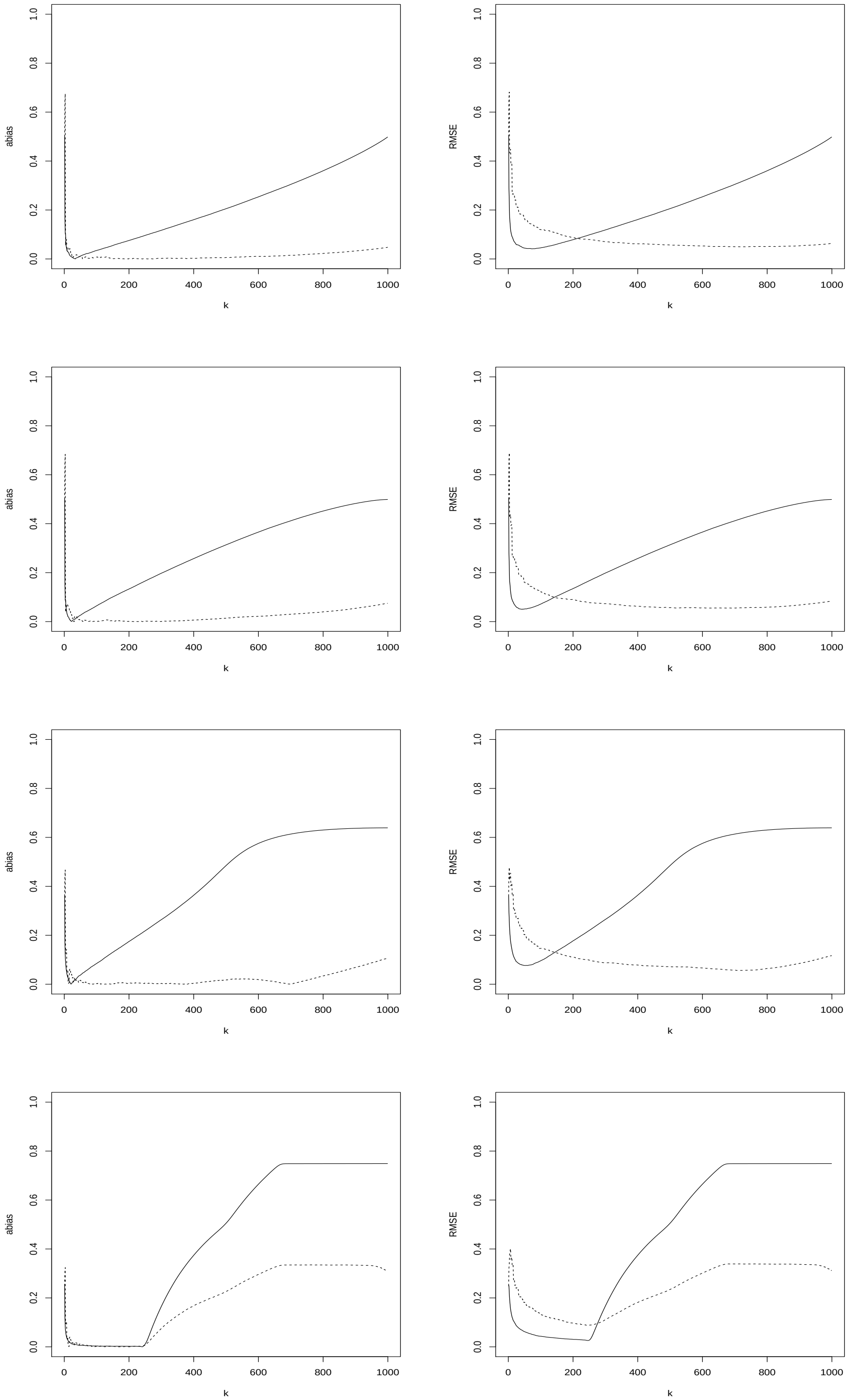

Figure 2: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^R$ (full) and $\widehat{\theta}^R_{GJ}$ (dashed) of models (top-to-bottom): MM(I), MM(II), ARCauchy, ARUnif.

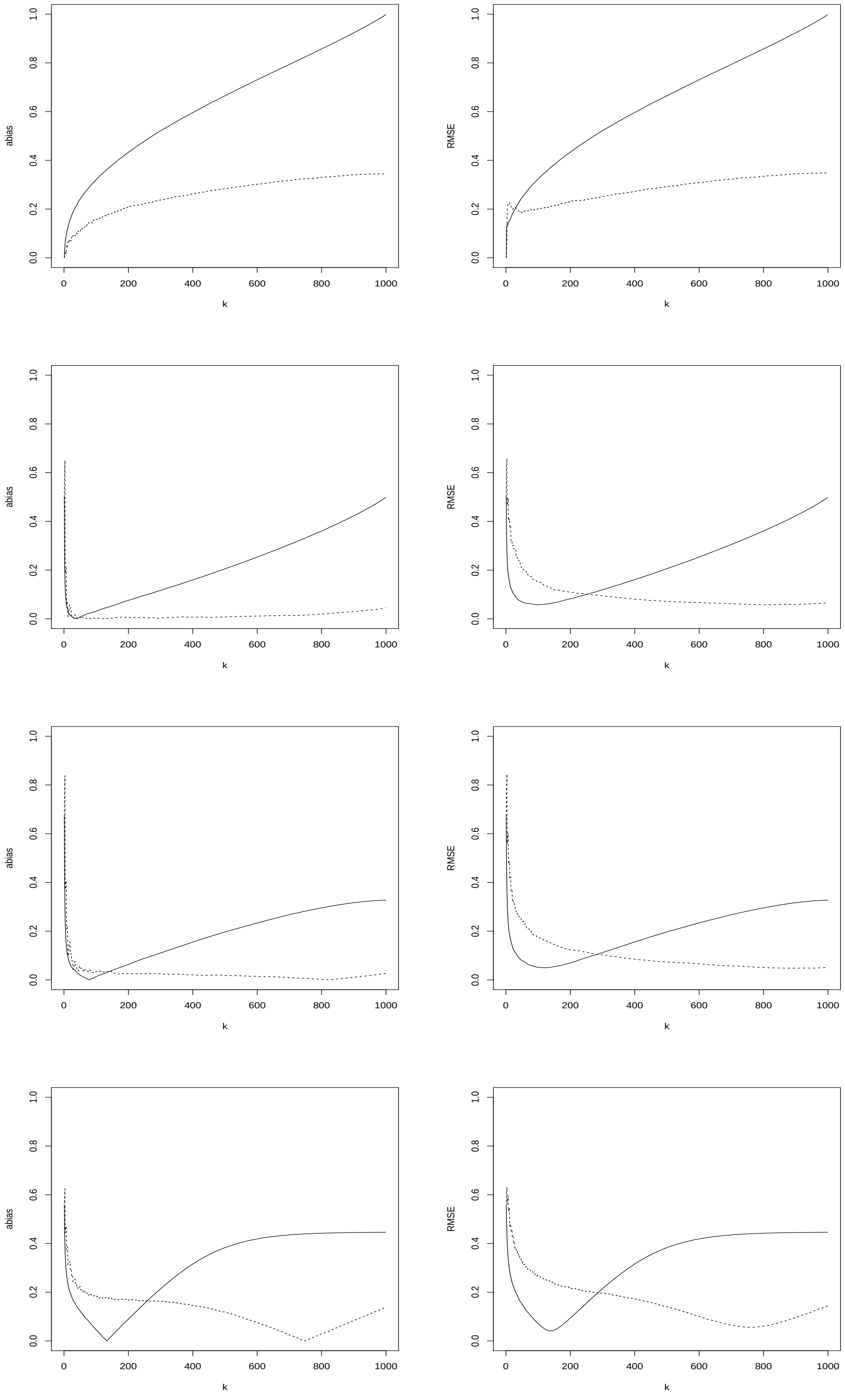

Figure 3: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^R$ (full) and $\widehat{\theta}^R_{GJ}$ (dashed) of models (top-to-bottom): AR, MAR, MCBEV, GARCH.

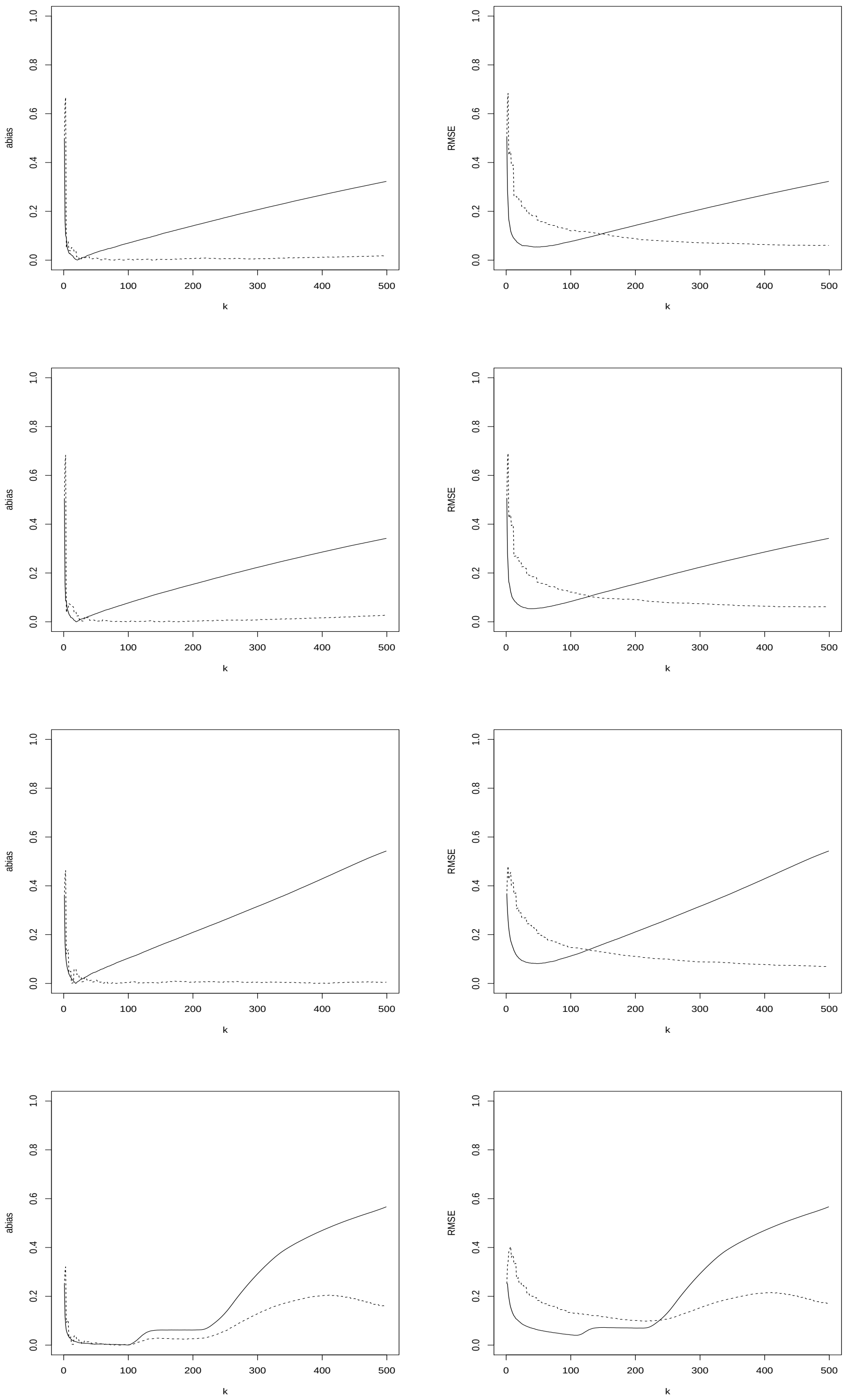

Figure 4: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^{CD}$ (full) and $\widehat{\theta}^{CD}_{GJ}$ (dashed) of models (top-to-bottom): MM(I), MM(II), ARCauchy, ARUnif.

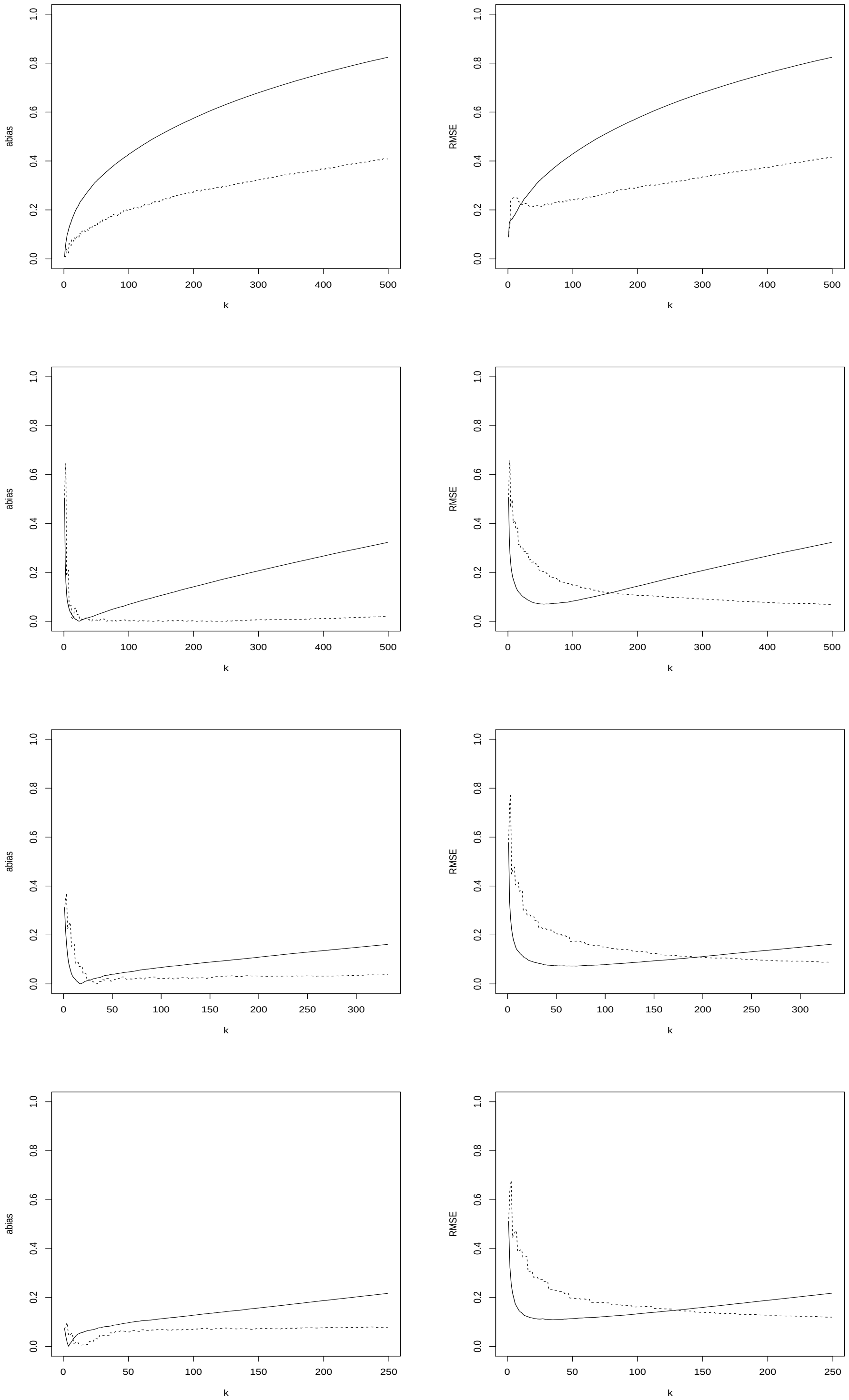

Figure 5: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^{CD}$ (full) and $\widehat{\theta}^{CD}_{GJ}$ (dashed) of models (top-to-bottom): AR, MAR, MCBEV, GARCH.

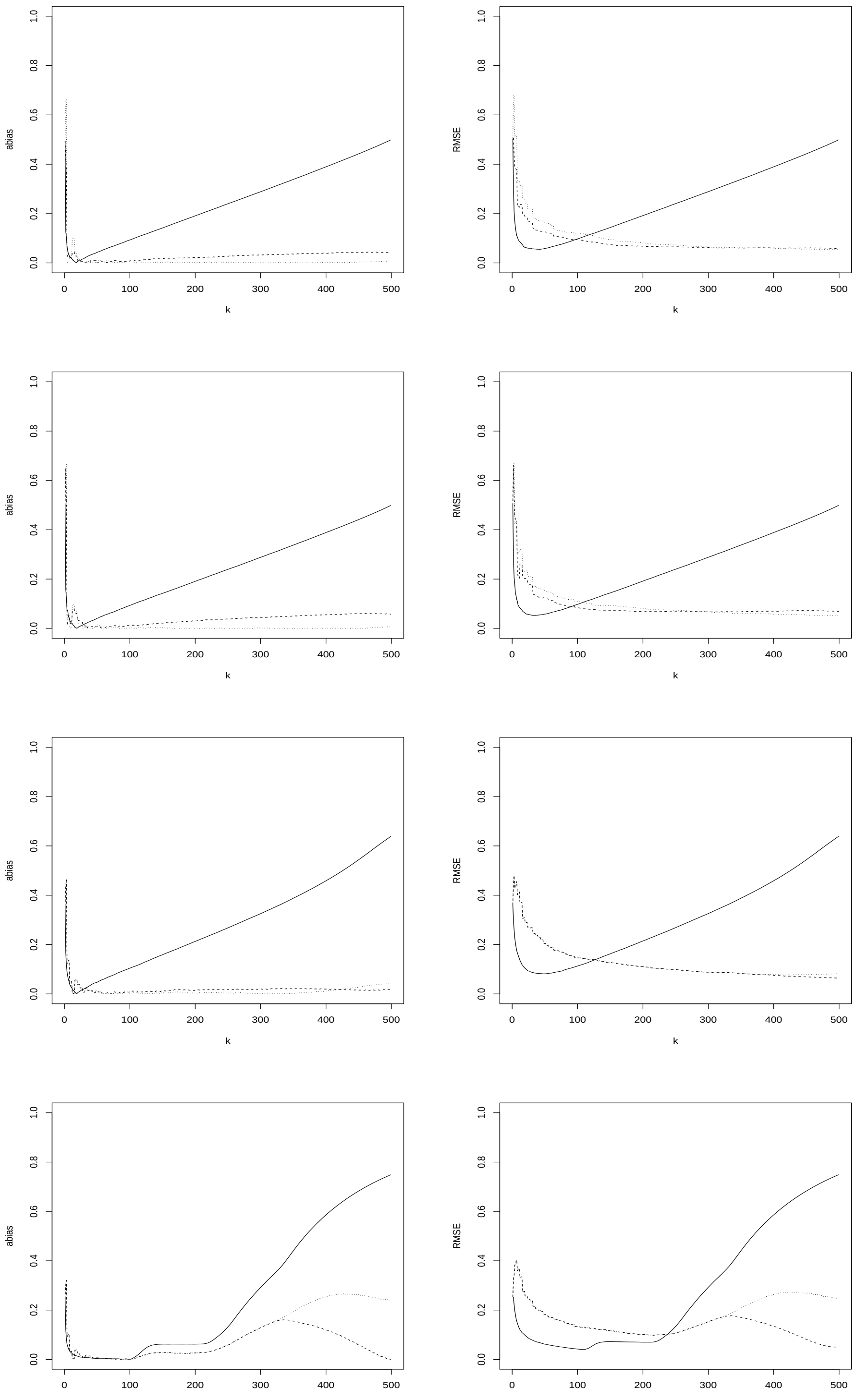

Figure 6: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^{CI}$ (full), $\widehat{\theta}^{CI}_{GJ}$ (dashed) and $\widehat{\theta}^{CI_2}_{GJ}$ (dotted) of models (top-to-bottom): MM(I), MM(II), ARCauchy, ARUnif.

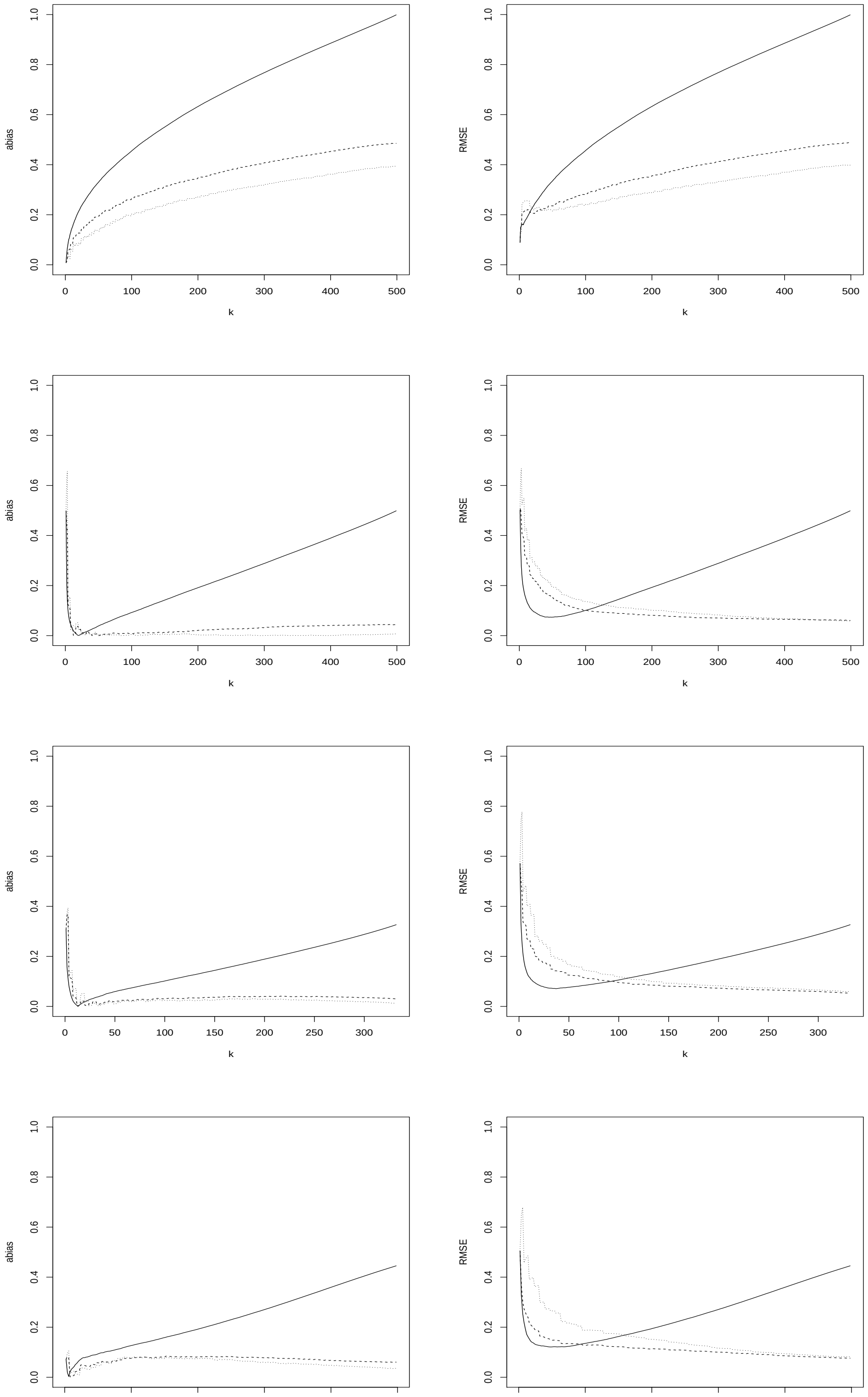

Figure 7: Absolute bias (left) and rmse (right), for $n = 1000$, of $\widehat{\theta}^{CI}$ (full), $\widehat{\theta}^{CI}_{GJ}$ (dashed) and $\widehat{\theta}^{CI_2}_{GJ}$ (dotted) of models (top-to-bottom): AR, MAR, MCBEV, GARCH.

4. Applications

4.1. Environmental data

We consider the daily maximum temperatures registered at Uccle (Belgium) in the period 1901-1999 (`http://lstat.kuleuven.be/Wiley/Data/ecad00045TX.txt`). In order to keep the stationarity assumption, we consider the July observations, corresponding typically to the warmest month. See the plotted data in Figure 8. An empirical evaluation of conditions $D^{(s)}(u_n)$ was conducted in Ferreira (2015) leading to the choice $s = 3$. Beirlant, Goegebeur, Segers and Teugels (2004) suggested the run length 4. Both values were considered and the option $s = 3$ led to closer estimates of $\theta \approx 0.55$ derived in the previous reference under parametric modelling. The sample path estimators plotted in Figure 9 thus correspond to this case. We can observe, within the GJ estimators, a clear decrease of the bias and more stability around the horizontal line corresponding to the referred estimate $\theta \approx 0.55$.

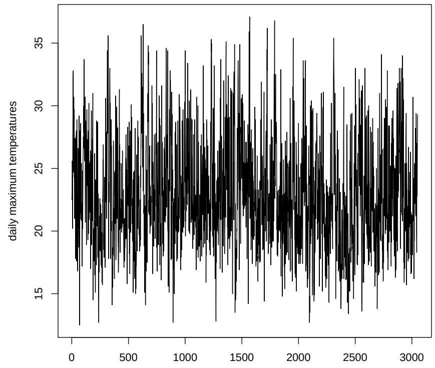

Figure 8: Uccle temperatures in July during 1901 – 1999.

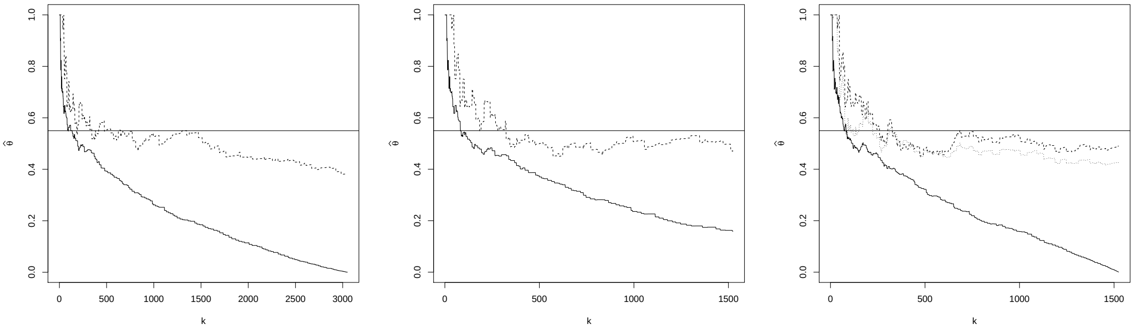

Figure 9: Extremal index estimation within Uccle data (from left-to-right): sample paths of $\widehat{\theta}^R$, $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$ (full) and respective GJ versions (dashed); the dotted line in the last panel refers to $\widehat{\theta}^{CI_2}_{GJ}$. The horizontal line corresponds to the estimate 0.55.

4.2. Financial data

Financial analysts are often focused on large gains or losses where the clustering phenomena is of particular concern. We consider the daily closing prices of the Dow Jones index over the period 1996 – 2000 (Figure 10). More precisely, we take a reasonable stationary series by deriving the log-returns (logarithms of the ratios of successive prices). The analyses performed in Coles (2001) suggests $r = 4$ and leads to $\theta \approx 0.865$. Based on this, we compute the sample paths of the proposed estimators, plotted in Figure 11. We can see that the GJ estimates oscillate closer around the horizontal line ($\theta \approx 0.865$), particularly in the case of the runs estimator.

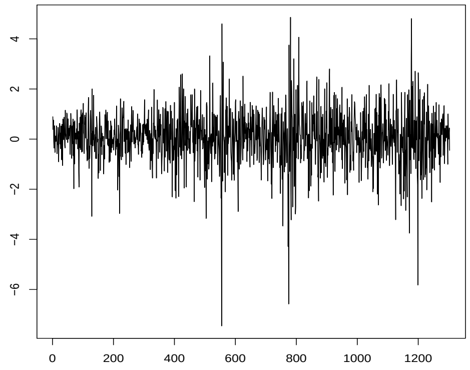

Figure 10: Dow Jones daily log-returns during 1996 – 2000.

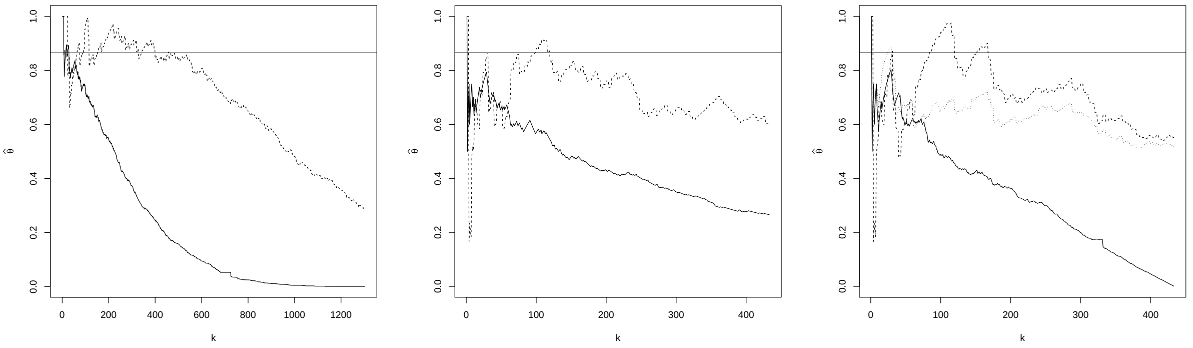

Figure 11: Extremal index estimation within Dow Jones index data (from left-to-right): sample paths of $\widehat{\theta}^{R}$, $\widehat{\theta}^{CD}$ and $\widehat{\theta}^{CI}$ (full) and respective GJ versions (dashed); the dotted line in the last panel refers to $\widehat{\theta}^{CI_2}_{GJ}$. The horizontal line corresponds to the estimate 0.865.

5. Discussion

The extremal index is a crucial parameter whenever clustering of high values takes place, since it is implicated in the estimation of rare events such as exceptional high quantiles, return levels or return periods. Semi-parametric estimators usually bear a large bias, a common feature when we are limited to the tail. In this paper we analyse the performance of second order GJ methods in the estimation of the extremal index. We can see that, in several estimators and diverse models, it accomplishes the expected task of decreasing the bias for a wider range of the trajectory estimates, a useful feature from a practical point of view and thus a motivation for applications. However, in dependent sequences with tails resembling i.i.d. structures ($\theta = 1$), the bias reduction is still not enough. Also, there are dependent structures where the two-dominant components form assumed for the bias in (6) is not the most convenient. These topics will be the aim of future work.

Acknowledgements

The author wishes to thank the reviewers for the helpful corrections.

References

ANCONA-NAVARRETE, M. A. AND TAWN, J. A. (2000). A comparison of methods for estimating the extremal index. *Extremes*, **3**, 5 – 38.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., AND TEUGELS, J. (2004). *Statistics of Extremes: Theory and Application.* John Wiley and Sons: New York.

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.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag: London.

FERREIRA, H. AND FERREIRA, M. (2015). Estimating the extremal index through local dependence. ArXiv:1505.02077v1 (submitted).

FERREIRA, M. (2015). Heuristic tools for the estimation of the extremal index: a comparison of methods. (submitted).

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.

GOMES, M. I., HALL, A., AND MIRANDA, M. C. (2008). Subsampling techniques and the jackknife methodology in the estimation of the extremal index. *Computational Statistics & Data Analysis*, **52** (4), 2022 – 2041.

GRAY, H. L. AND SCHUCANY, W. R. (1972). *The Generalized Jackknife Statistic.* Marcel Dekker: New York.

HALL, A. (1996). Maximum term of a particular autoregressive sequence with discrete margins. *Communications in Statistics – Theory and Methods*, **25** (4), 721 – 736.

HALL, P., HOROWITZ, J. L., AND JING, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, **50**, 561 – 574.

HSING, T. (1991). Estimating the parameters of rare events. *Stochastic Processes and Their Applications*, **37**, 117 – 139.

HSING, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. *Annals of Statistics*, **21**, 2043 – 2071.

LAHIRI, S., FURUKAWA, K., AND LEE, Y. (2007). Nonparametric plug-in method for selecting the optimal block lengths. *Statistical Methodology*, **4**, 292 – 321.

LAURINI, F. AND TAWN, J. (2003). New estimators for the extremal index and other cluster characteristics. *Extremes*, **6**, 189 – 211.

LEADBETTER, M. R. (1974). On extreme values in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **28** (4), 289 – 303.

LEADBETTER, M. R. (1983). Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **65** (2), 291 – 306.

LEADBETTER, M. R., LINDGREN, G., AND ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer: New York.

LEADBETTER, M. R. AND NANDAGOPALAN, S. (1989). *On Exceedance Point Processes for Stationary Sequences under Mild Oscillation Restrictions*, volume 51 of *Lecture Notes in Statistics*. Springer: New York, pp. 69 – 80.

LEADBETTER, M. R. AND ROOTZÉN, H. (1988). Extremal theory for stochastic processes. *Annals of Probability*, **16** (2), 431 – 478.

NANDAGOPALAN, S. (1990). *Multivariate Extremes and Estimation of the Extremal Index*. Ph.D. thesis, University of North Carolina: Chapel Hill, U.S.A.

NEVES, M., GOMES, M. I., FIGUEIREDO, F., AND PRATA-GOMES, D. (2015). Modeling extreme events: Sample fraction adaptive choice in parameter estimation. *Journal of Statistical Theory and Practice*, **9** (1), 184 – 199.

O'BRIEN, G. L. (1987). Extreme values for stationary and Markov sequences. *Annals of Probability*, **15**, 281 – 291.

PRATA-GOMES, D. AND NEVES, M. (2015). Bootstrap and other resampling methodologies in statistics of extremes. *Communications in Statistics – Simulation and Computation*, **44** (10), 2592 – 2607.

SMITH, R. AND WEISSMAN, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society: Series B*, **56**, 515 – 528.

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

WEISSMAN, I. AND NOVAK, S. (1998). On blocks and runs estimators of the extremal index. *Journal of Statistical Planning and Inference*, **66**, 281 – 288.

Manuscript received, 2015-11-30, revised, 2016-01-19, accepted, 2016-01-22.