Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Vergelijken van schatters van de extreme index
Bachelor eindproject Technische Wiskunde

Comparison on estimations for the extremal index

Author:
Josephine Clercx (4568095)

Supervisor:
Dr. J. -J. Cai

Other committee member:
Dr. Ir. L. E. Meester

June 25, 2020

Abstract

The clustering of events can have a large impact on society. The extremal index θ tells how much extreme events cluster. We will compare different types of estimators in this project. First, we review the extremes of different sequences which have different values of θ. We have found significant differences between the extremes. Then, 2 different types of estimators are introduced which both use different ways to divide the data, using disjoint blocks and using sliding blocks. The optimal block lengths are simulated for all those estimators. Using those block lengths, θ is simulated with all the estimators. From the simulations we conclude that the estimators using sliding blocks perform better. The best-performed estimator that we found from the simulations is used to estimate θ on data from the KNMI, comparing wind gusts and precipitation at weather stations De Bilt and Vlissingen.

Contents

1 Introduction

Clustering of extreme events in a short period can have a large impact on society. For example, high temperatures can lead to forest fires, or multiple days of heavy rain can results in floods. Clustering of extremes is due to sequentially extremal dependence in a time series. The extremal index θ is a parameter that measures the clustering of extreme values. Its value ranges between 0 and 1. When $\theta = 1$, the extreme events are independent, they do not cluster. When $\theta < 1$, the extreme events are dependent, so they cluster together. The lower the extremal index, the more dependent the extreme events are, the more they cluster together.

The extremal index is defined by (Leadbetter (1983) [6]):

Definition 1. *Let $\{\xi_n\}_{n\geq 1}$ be a strictly stationary sequence of random variables with marginal distribution function F, finite or infinite right end point $\omega = sup\{x : F(x) < 1\}$ and tail function $\bar{F} = 1 - F$. For integers $0 \leq k < l$ and $n \geq 1$ put $M_{k,l} = \max\{\xi_i : i = k+1, ..., l\}$ and $M_n = M_{0,n}$. The process $\{\xi_n\}_{n\geq 1}$ is said to have extremal index $\theta \in [0,1]$ if for each $\tau > 0$ there is a sequence $\{u_n\}_{n\geq 1}$ such that, as $n \to \infty$,*

(a) $n\bar{F}(u_n) \to \tau$ and
(b) $P(M_n \leq u_n) \to e^{-\theta\tau}$.

Combining a and b gives another interpretation of the definition of the extremal index. Let

$$u_n = \bar{F}^{-1}(\frac{\tau}{n})$$

Resulting in,

$$P(M_n \leq \bar{F}^{-1}(\frac{\tau}{n})) = P(\bar{F}(M_n) \geq \frac{\tau}{n}) = P(n(1 - F(M_n)) \geq \tau) \to e^{-\theta\tau}$$

Therefore, $P(n(1 - F(M_n)) \geq \tau) \to e^{-\theta\tau}$ defines the extremal index. This implies that $n(1 - F(M_n))$ follows an $Exp(\theta)$-distribution asymptotically. We will use this later on the project to build up the estimator.

The project is structured around the following 3 research questions:

- *What is the behaviour of extremes of different types of sequences?*
- *Which type of estimator is the best for estimating θ, using the simulated optimal block length b_n for every sequence?*
- *How does the extremal index behave for data from the KNMI?*

The first research question will be handled in Sections 2 and 3. In Section 2 the extremes are presented of different sequences: independent sequences with $\theta = 1$ and dependent sequences with different values of θ. Continuing in Section 3, a comparison study will be made on the maximum value of different sequences. The second research question will be discussed in Sections 4, 5 and 6. Different types of estimators are introduced in Section 4, where we distinguish between the use of disjoint blocks and sliding blocks. In Section 5 the optimal block length is simulated for every estimator, using those block lengths. In Section 6, is θ estimated for different sequences with independent and dependent extreme values. Finally, in Section 7 data from the KNMI (Koninklijk Nederlands Meteorologisch Instituut) will be introduced, with this data θ is estimated for different block lengths.

1.1 Codes

All the simulations of the project have been done in the program 'RStudio', the codes can be found on `www.github.com/JosephineClercx/BEP`.

2 Extremes of different sequences

The extremes of sequences can be either dependent ($\theta < 1$) or independent ($\theta = 1$). Both cases are explained with some examples of different sequences. These sequences will be used in all the simulations and calculations of the project.

2.1 Independent extremes of sequences

When $\theta = 1$ for a sequence, the extreme values of that sequence are independent, they will not cluster together. Below you find some examples, the first ones are based on an independent and identical distribution (iid) of random variables, the last one is a moving average sequence.

IID

First, we will prove that for every sequence that is iid, $\theta = 1$ holds.

Theorem 1. *Let $X_1, X_2, ..., X_n$ be a sequence which is Independent and Identical Distributed, then the extreme values of $X_1, X_2, ...$ are independent. Hence $\theta = 1$ according to Definition 1.*

Proof. Let $X_1, X_2, ..., X_n$ be iid with continuous cumulative distribution function (cdf) F and $M_n = \max_{1 \le i \le n} X_i$. $F(M_n)$ has the same distribution as the distribution of $\max_{1 \le i \le n} U_i$ with U_i iid $U(0,1)$ for every i. This is because, $F(X_i)$ is equal in distribution to $U(0,1)$ for every i, which follows from the probability integral transform (page 353 in Rice (2007) [8]). Then, as $n \to \infty$, for each $\tau > 0$,

$$P(\max_{1 \le i \le n} U_i \le 1 - \tau/n) \overset{(1)}{=} (1 - \tau/n)^n \overset{(2)}{\to} e^{-\tau} \tag{1}$$

(1) derives from the fact that U_i is iid for every i, because the following holds (page 109 in Dekking et al. (2005) [4]): Let $Z = \max_{1 \le i \le n} U_i$, then

$F_Z(1 - \tau/n) = P(Z \le 1 - \tau/n) = P(\max_{1 \le i \le n} U_i \le 1 - \tau/n) = P(U_1 \le 1 - \tau/n, U_2 \le 1 - \tau/n, ..., U_n \le 1 - \tau/n) = P(U_1 \le 1 - \tau/n) P(U_2 \le 1 - \tau/n) \cdots P(U_n \le 1 - \tau/n) = (F(1 - \tau/n))^n$.

The last step, (2), derives from the standard limit $\lim_{n \to \infty} (1 + \frac{x}{n})^n \to e^x$, here the short proof:

$$\lim_{n \to \infty} (1 + \frac{x}{n})^n = \lim_{n \to \infty} e^{\ln n (1 + \frac{x}{n})} = e^{\lim_{n \to \infty} \ln n (1 + \frac{x}{n})} = e^{\lim_{n \to \infty} \frac{1 + \frac{x}{n}}{\frac{1}{n}}} \overset{(3)}{=} e^{\lim_{n \to \infty} \frac{\frac{-x}{n^2} \frac{1}{1 + \frac{x}{n}}}{-\frac{1}{n^2}}} = e^{\lim_{n \to \infty} \frac{x}{1 + \frac{x}{n}}} = e^x$$

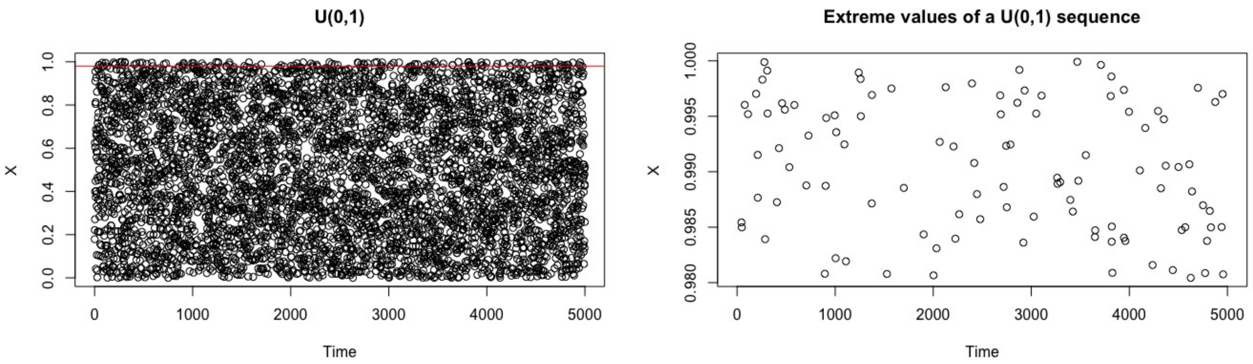
In step (3) is L'Hopital's Rule used. Rewriting Equation 1 gives,

$$P(\max_{1 \le i \le n} U_i \le 1 - \tau/n) = P(n(1 - \max_{1 \le i \le n} U_i) \ge \tau) \to e^{-\tau} \tag{2}$$

Therefore, $P(n(1 - F(M_n)) \ge \tau) \to e^{-\tau}$. Hence, an iid sequence with continuous cdf has independent extreme values and $\theta = 1$. □

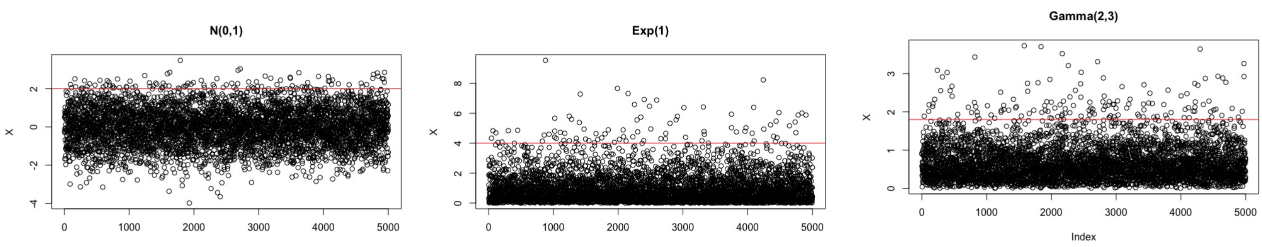
Here follow some examples of different iid sequences. In Figure 1 we have simulated 5000 random variables from a $U(0,1)$ distribution. We also looked at the extreme values in Figure 1b. The results give a homogeneous distribution without obvious clusters, therefore we can say that the extreme values do not cluster and the extreme values of this model are independent.

In Figure 2 we did the same, we simulated 5000 random variables of the $N(0,1)$-, $Exp(1)$- and $Gamma(2,3)$-distribution. Figures 2d, 2e and 2f show the extreme values of the different sequences, respectively. All the sequences do not show clustering at the extreme values and therefore the extreme values of these models are also independent.

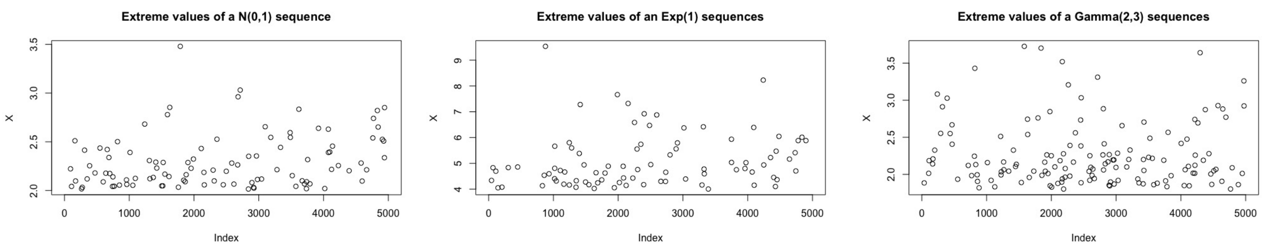


(a) $U(0,1)$ sequence with red line at 0.98 (b) Values above 0.98

Figure 1: Extreme values of a $U(0,1)$ sequence



(a) $N(0,1)$ sequence with red line at 2 (b) $Exp(1)$ sequence with red line at 4 (c) $Gamma(2,3)$ sequence with red line at 1.8



(d) Values above 2 (e) Values above 4 (f) Values above 1.8

Figure 2: Extreme values of 3 different iid sequences

Moving average

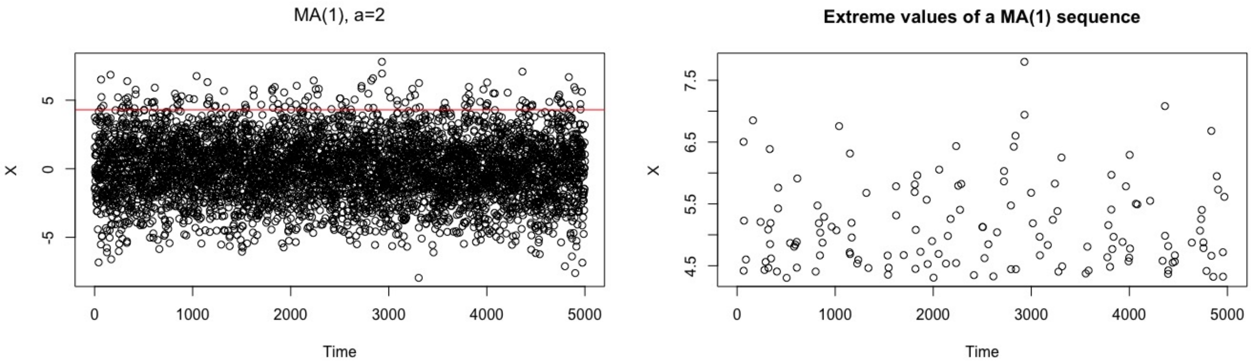
A moving average sequence is an example of a sequence that has an independent extremal index but it is not iid. Below is the definition given (page 11 in VanderMeulen (2019) [10]) :

Definition 2. *Let* $Z_i \sim WN(0,\sigma^2)$ *and* $a \in \mathbb{R}$,

$$X_i = Z_i + aZ_{i-1}, \qquad i = 0, \pm 1, \pm 2, \ldots \tag{3}$$

X_i *is a first order moving average process, it can be written as* $X_i \sim MA(1)$.

In Figure 3 we simulated 5000 MA(1) random variables with $a = 2$ and $\sigma^2 = 1$. We zoomed in at the extreme values of the MA(1) sequence in Figure 3b, we don't see any obvious clusters. Therefore we can conclude that the extreme values are independent.



(a) MA(1) sequence with red line at 4.3

(b) Values above 4.3

Figure 3: Extreme values of a MA(1) sequence with $a = 2$

2.2 Dependent extremes of sequences

In this section, we present different sequences where $\theta < 1$ for.

ARCH

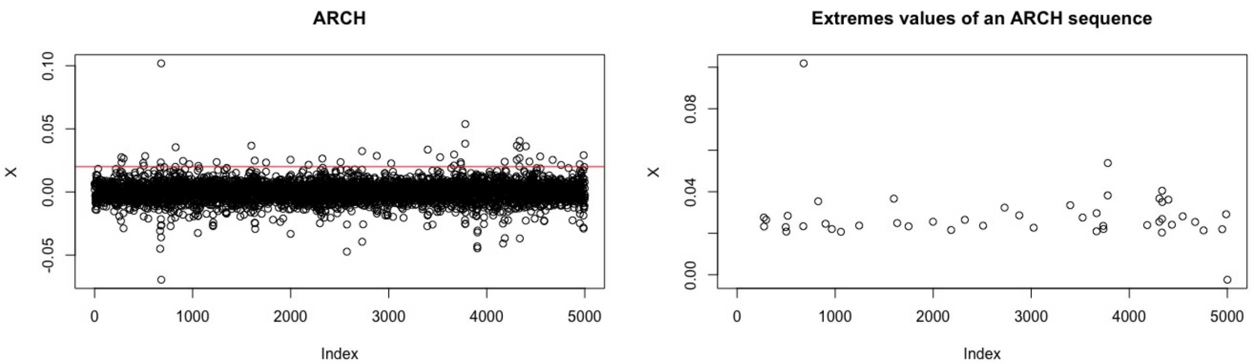
This is an ARCH model given by:

Definition 3. *Let ϵ_i be iid N(0,1) for every i and*

$$X_i = (2 \times 10^{-5} + 0.7 X_{i-1}^2)^{1/2} \epsilon_i, \quad i \geq 1$$

X_i is an ARCH model.

For this model, $\theta = 0.721$ (Table 3.2 in de Haan et al. (1989) [3]). This means that $\theta < 1$, therefore the extreme values are not independent of this model. In Figure 4b we see some extreme values that tend to cluster together, however there are not clear peaks.



(a) ARCH sequence with red line at 0.02

(b) Values above 0.02

Figure 4: Extreme values of an ARCH sequence

sARCH

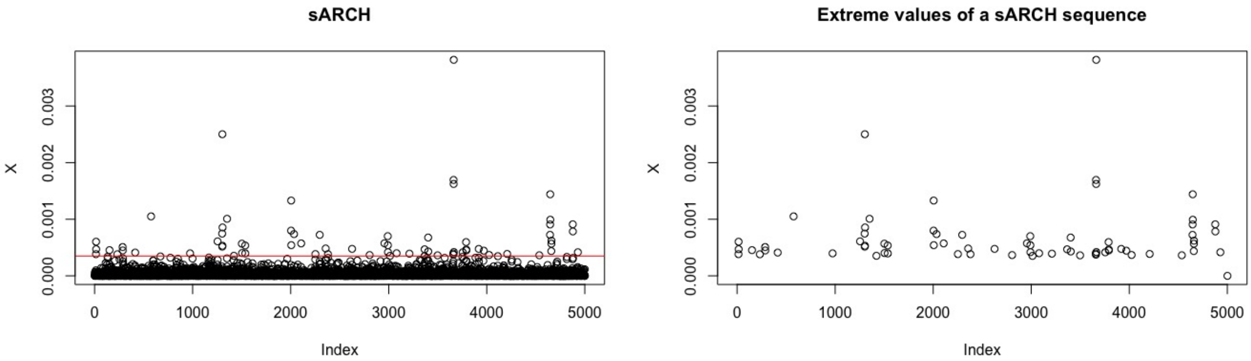
sARCH is a squared ARCH model, given by:

Definition 4. *Let ϵ_i be iid N(0,1) for every i and*

$$X_i = (2 \times 10^{-5} + \frac{1}{2} X_{i-1}) \epsilon_i^2, \quad i \geq 1$$

X_i is sARCH model.

For this model, $\theta = 0.727$ (Table 3.2 in de Haan et al. (1989) [3]). So $\theta < 1$, therefore the extreme values are not independent. In figure 5b we also see some extreme values that tend to cluster together.



(a) sARCH sequence with red line at 0.00035 (b) Values above 0.00035

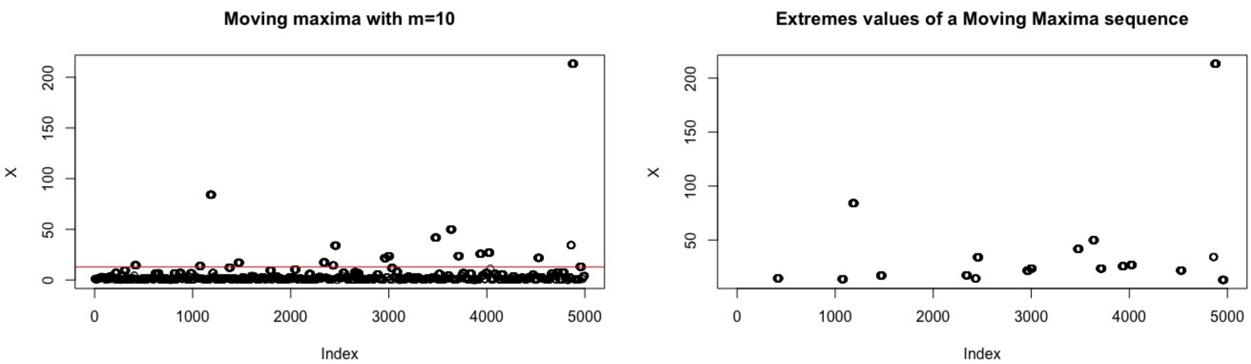
Figure 5: Extreme values of a sARCH sequence

Moving Maxima

A Moving Maxima sequence (Cai (2019) [2]) is given by:

Definition 5. *Let $m \geq 2$ a fixed constant, let ϵ_i be an iid sequence with $P(\epsilon_i \leq x) = \exp(-\frac{1}{mx})$ and $X_i = \max_{0 \leq j \leq m} \epsilon_{j+i}$ for $i \geq 1$. Then, X_i is a Moving Maxima model.*

For this model $\theta = \frac{1}{m}$. In Figure 6b the extreme values are plotted of a Moving Maxima sequence with $m = 25$, so $\theta = 0.04$, which is a very small value. The dependence of the extreme values are clearly visible.



(a) Moving Maxima sequence with red line at 13 (b) Values above 13

Figure 6: Extreme values of a Moving Maxima sequence

AR-C

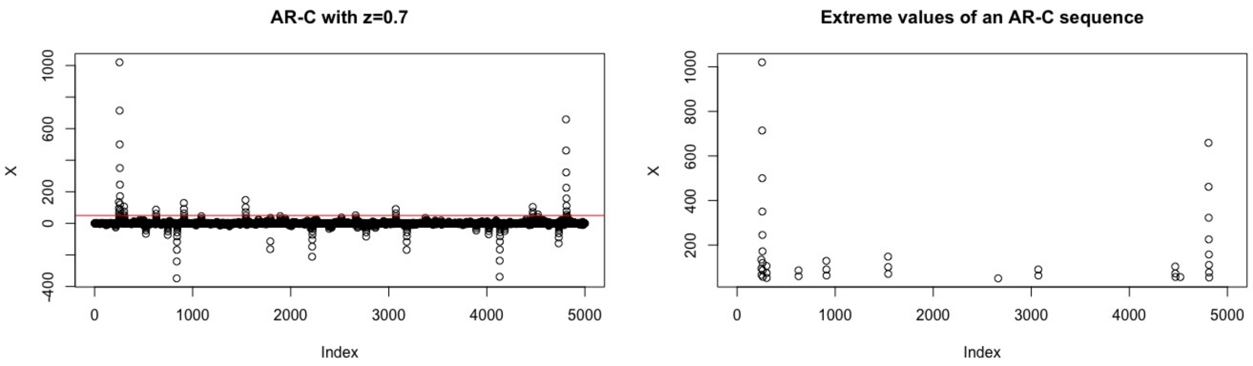
This is an AR(1) model with Cauchy margin (Cai (2019) [2]):

Definition 6. *Let $z \in (-1, 1)$, let ϵ_i be iid with density $\frac{1-|z|}{\pi(x^2+(1-|z|)^2)}$, let X_0 have a standard Cauchy density $\frac{1}{\pi(1+x^2)}$ and let*

$$X_i = zX_{i-1} + \epsilon_i, \quad i \geq 1 \tag{4}$$

Then, X_i is an AR-C model.

For this model, if $z \geq 0$ then $\theta = 1 - z$ and if $z \geq 0$ then $\theta = 1 - |z|^2$. In figure 7b the extreme values are plotted of an AR-C sequence with $z = 0.7$, so $\theta = 0.3$. We see that almost every extreme value is dependent on another.



(a) AR-C sequence with red line at 50

(b) Values above 50

Figure 7: Extreme values of an AR-C sequence with z=0.7

We have discussed and demonstrated different sequences above, there is a clear difference between sequences with $\theta = 1$ and $\theta < 1$. We can conclude that the lower the value of θ, the more dependent the extremes are. This is clearly visible when you compare figures 5b and 6b, with $\theta = 0.727$ and $\theta = 0.04$ respectively.

3 Compare sequences

In this section, we will investigate the maximum values of different sequences. We compare 3 pairs of sequences, each one has the same distribution but differs in dependence. MA(1) and N(0,5), Moving Maxima and Std. Fréchet, AR-C and Std. Cauchy will be compared. Below are the definitions used for the comparison, given in pairs.

$$\begin{cases} X_i \sim MA(1), \quad X_i = Z_i + aZ_{i-1}, \quad Z_i \sim WN(0,\sigma^2), a \in R \\ \tilde{X}_i \sim N(0,(1+a^2)\sigma^2) \end{cases}$$

$$\begin{cases} Y_i = \max_{0\le j\le m} \epsilon_{j+i}, \quad i \ge 1, m \ge 2, \quad \epsilon_i \sim P(\epsilon_i \le x) = \exp(-\frac{1}{mx})(= \text{Fréchet}(1,\frac{1}{m},0)) \quad (\text{MovingMaxima, Def. } 5) \\ \tilde{Y}_i \sim \text{Fréchet}(1,\frac{1}{m},0) \end{cases}$$

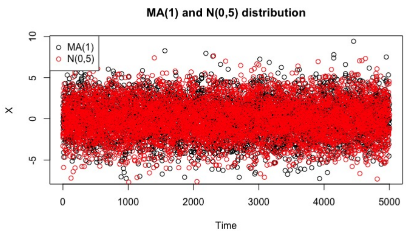
$$\begin{cases} Z_i = zZ_{i-1} + \epsilon_i, \quad z \in (-1,1), \quad \epsilon_i \sim \frac{1-|z|}{\pi(x^2+(1-|z|)^2)}(= \text{Cauchy}(0,1-|z|)) \quad (\text{AR}-\text{C, Def. } 6) \\ \tilde{Z}_i \sim \text{Cauchy}(0,1-|z|) \end{cases}$$

For the further simulations and calculations the following parameter values are chosen: $a = 2, \sigma^2 = 1, m = 10$ and $z = 0.7$

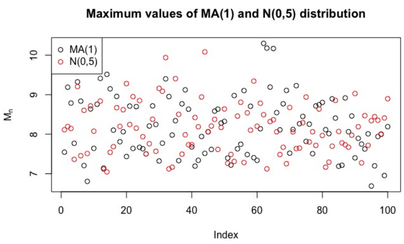
In Figures 8a, 8d and 8g the different sequences are represented, in every figure it seems that the pair of sequences behave the same. We took 100 times the maximum of 5000 random variables from the sequences, see Figures 8b, 8e and 8h. There are no differences between the behaviour of the maximum values of the sequences. Note that in figures 8e and 8h the logarithmic values are presented to see the results better.

In Figures 8c, 8f and 8i we have simulated a density plot of $F(M_n)$ for every pair. The $F(M_n)$ of MA(1) and N(0,5) have the same density behaviour, this is because for both distributions $\theta = 1$. For the other 2 pairs this is different, see Figures 8f and 8i. The density for the standard independent sequences is much higher at the values close to 1, compared to the dependent sequences. In Appendix A, Figures 8f and 8i are bigger presented. The explanation for the differences is: Assume Y_i has cdf F and $\tilde{Y}_i$ has cdf $\tilde{F}$, then F and $\tilde{F}$ are the same because Y_i and $\tilde{Y}_i$ have the same distribution, however $F(M_n)$ and $\tilde{F}(\tilde{M}_n)$ (with $M_n = \max(X_i)$ and $\tilde{M}_n = \max(\tilde{X}_i)$) are not the same. This is because Y_i is not an independent distribution, while $\tilde{Y}_i$ is. The same applies to Z_i and $\tilde{Z}_i$

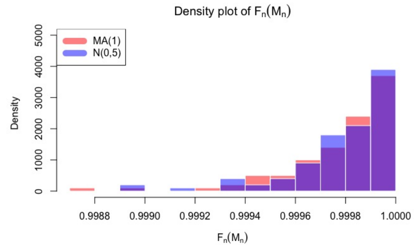
For Y_i is $\theta = 1/m = 0.1$ and for Z_i is $\theta = 1 - z = 0.3$. Accordingly, the smaller θ is, the larger the variation in density is.



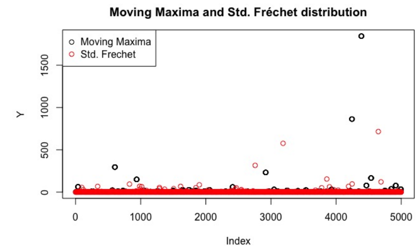
(a) MA(1) and $N(0,5)$ sequence



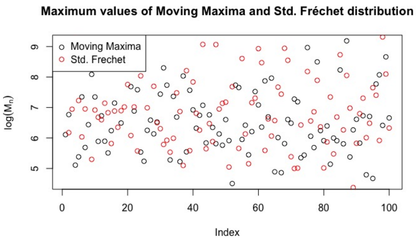
(b) Maximum values of MA(1) and $N(0,5)$ sequence



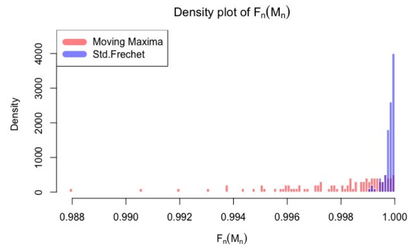
(c) Density plot of $F(M_n)$



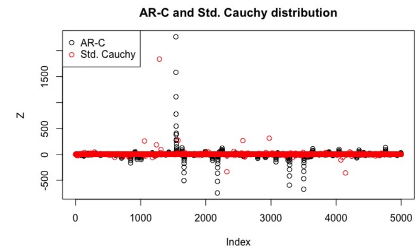
(d) Moving Maxima and Std. Fréchet sequence



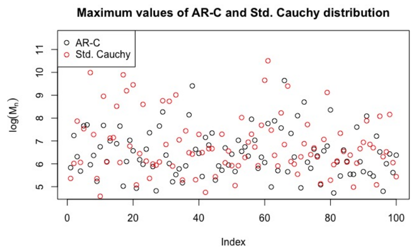
(e) Maximum values of Moving Maxima and Std. Fréchet sequence



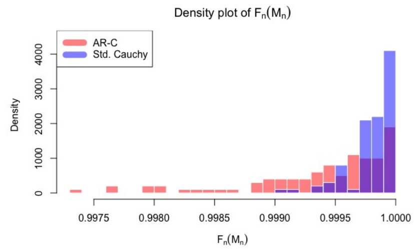
(f) Density plot of $F(M_n)$



(g) AR-C and Std. Cauchy sequence



(h) Maximum values of AR-C and Std. Cauchy sequence



(i) Density plot of $F(M_n)$

Figure 8: Comparing sequences

4 Maximum likelihood estimation of the extremal index

A way to estimate the extremal index θ is by using the maximum likelihood estimator for the exponential distribution, based on a sample of estimated block maxima (Berghaus (2013) [1]). We use the maximum likelihood estimator for the exponential distributions because, in Section 1 we explained that the definition of the extremal index follows an $Exp(\theta)$-distribution asymptotically. The maximum likelihood estimator for the exponential distribution is defined in the following steps. The probability density function of the exponential distribution is defined as

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{5}$$

Its likelihood function is

$$\mathcal{L}(\lambda; x_1, ...x_n) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \tag{6}$$

To calculate the maximum likelihood solve the following for λ:

$$\frac{d\ln(\mathcal{L}(\lambda; x_1, ..., x_n))}{d\lambda} = 0 \tag{7}$$

$$\frac{d\ln(\mathcal{L}(\lambda; x_1, ..., x_n))}{d\lambda} = \frac{d\ln(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i})}{d\lambda} = \frac{dn\ln(\lambda) - \lambda \sum_{i=1}^{n} x_i}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \tag{8}$$

Finally we get (page 318 in Dekking et al. (2015) [4]),

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i} \tag{9}$$

Combining these to estimate θ, the formula becomes:

$$\tilde{\theta}_n = (\frac{1}{n} \sum_{i=1}^{n} X_i)^{-1} \tag{10}$$

With $X_1, X_2, ..., X_n$ a stationary sequence of real-valued random variables with stationary cdf F. The data can be separated in 2 ways, with disjoint blocks and with sliding blocks.

4.1 Disjoint blocks

Looking at the sample block maxima, suppose n observations from time series $(X_i)_{i\geq 1}$. Divide the sample into k_n blocks of length b_n and assume $n = k_n b_n$. For $i = 1, ..., k_n$ let $M_{ni} = M_{((i-1)b_n+1):(ib_n)} = \max\{X_{(i-1)b_n+1}, ..., X_{ib_n}\}$, the maximum over all the X_i in the ith block. Let $N_{ni} = F(M_{ni}) = \max\{U_{(i-1)b_n+1}, ..., U_{ib_n}\}$, the last equality follows from the fact that $F(M_{ni})$ is a probability integral transform, see the proof of Theorem 1. Let $Y_{ni} = -b_n \log(N_{ni})$. With b_n sufficiently large the random variables $Y_{ni}, ..., Y_{nk}$ follow an approximate sample from the $Exp(\theta)$-distribution. Using the maximum likelihood estimator in Equation 10, we replace X_i with Y_{ni} and n with k_n, we get:

$$\tilde{\theta}_n = (\frac{1}{k_n} \sum_{i=1}^{k_n} Y_{ni})^{-1} \tag{11}$$

Note that, $\tilde{\theta}_n$ is not an estimator for θ, because it is based on the unknown cdf F, therefore we call $\tilde{\theta}_n$ an oracle for θ. For this reason, we have the following definitions: $\hat{N}_{ni} = \hat{F}_n(M_{ni})$ with $\hat{F}_n(x) = n^{-1} \sum_{s=1}^{n} \mathbf{1}(X_s \leq x)$ denotes the emperical cdf of $X_1, ..., X_n$. Then, $\hat{Y}_{ni} = -b_n \log(\hat{N}_{ni})$. We use Equation 10 again, but now with $\hat{Y}_{ni}$ for X_i and k_n for n:

$$\hat{\theta}_n^{\mathrm{N,dj}} = (\frac{1}{k_n} \sum_{i=1}^{k_n} \hat{Y}_{ni})^{-1} \tag{12}$$

$\hat{\theta}_n^{\mathrm{N,dj}}$ is an estimator for θ (Northrop (2015) [7]).
There is another variant of the estimator, using the following formulas:

$$\hat{\theta}_n^{\mathrm{B,dj}} = (\frac{1}{k_n} \sum_{i=1}^{k_n} \hat{Z}_{ni})^{-1}, \qquad \hat{Z}_{ni} = b_n(1 - \hat{N}_{ni}) \tag{13}$$

The 2 variants of the estimator are almost the same, because $\hat{Y}_{ni}$ and $\hat{Z}_{ni}$ are almost the same. Using $\log(1-x) \approx -x$ for $|x| < 1$, with $x = \hat{N}_{ni} = \hat{F}_n(M_{ni})$, $\hat{Y}_{ni}$ can be rewritten:

$$\hat{Y}_{ni} = -b_n \log(\hat{N}_{ni}) = -b_n \log(1 - (1 - \hat{N}_{ni})) \approx b_n(1 - \hat{N}_{ni}) = \hat{Z}_{ni}$$

Therefore, $\hat{Y}_{ni} \approx \hat{Z}_{ni}$, the next step is showing that $\hat{Z}_{ni}$ is connected to the extremal index with Definition 1. Rewriting $\hat{Z}_{ni}$ gives, $\hat{Z}_{ni} = b_n(1 - \hat{N}_{ni}) = b_n(1 - \hat{F}(M_{ni}))$, then using Definition 1 with $F = \hat{F}$, $u_n = \bar{\hat{F}}(\frac{\tau}{b_n})$ and $M_n = M_{ni}$, we can conclude that $b_n(1 - \hat{F}(M_{ni}) \to e^{-\theta\tau}$. This confirms the choice of $\hat{Y}_{ni}$ and $\hat{Z}_{ni}$.

4.2 Sliding blocks

Another way to estimate θ, is a variant based on sliding blocks (Northrop (2015) [7]) instead of disjoint blocks. Divide the sample into $n - b_n + 1$ blocks of length b_n, for $t = 1, ..., n - b_n + 1$, let $M_{nt}^{\mathrm{sl}} = M_{t:(t+b_n-1)} = \max\{X_t, ..., X_{t+b_n-1}\}$. Using the same formulas as for the estimators with the disjoint blocks, $N_{nt}^{\mathrm{sl}} = F(M_{nt}^{\mathrm{sl}})$, $Z_{nt}^{\mathrm{sl}} = b_n(1 - N_{nt}^{\mathrm{sl}})$ and $Y_{nt}^{\mathrm{sl}} = -b_n \log N_{nt}^{\mathrm{sl}}$, and the empirical counterparts, $\hat{N}_{nt}^{\mathrm{sl}} = \hat{F}(M_{nt}^{\mathrm{sl}})$, $\hat{Z}_{nt}^{\mathrm{sl}} = b_n(1 - \hat{N}_{nt}^{\mathrm{sl}})$ and $\hat{Y}_{nt}^{\mathrm{sl}} = -b_n \log \hat{N}_{nt}^{\mathrm{sl}}$. Then the 2 variants of estimators are defined as

$$\hat{\theta}_n^{\mathrm{N,sl}} = (\frac{1}{n - b_n + 1} \sum_{t=1}^{n-b_n+1} \hat{Y}_{nt}^{\mathrm{sl}})^{-1}, \qquad \hat{\theta}_n^{\mathrm{B,sl}} = (\frac{1}{n - b_n + 1} \sum_{t=1}^{n-b_n+1} \hat{Z}_{nt}^{\mathrm{sl}})^{-1} \tag{14}$$

Note that, for both estimators above, the data should not be discarded if b_n is not a divisor of the sample size n. This is different for the 2 estimators which use disjoint blocks, there the data will be discarded if b_n is not a divisor of n.

5 Optimal block length

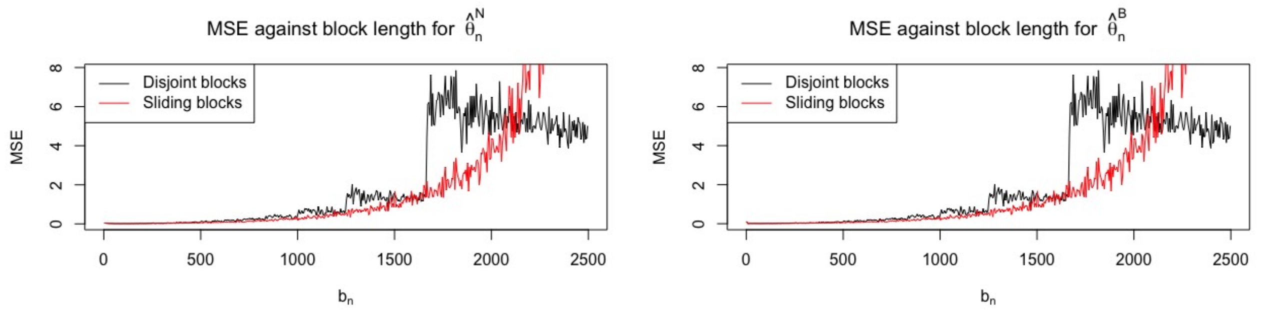
For different sequences and different estimators the size of the block length and the amount of blocks can affect the result of the estimator. In this section the optimal block length will be investigated by running a simulation based on the Mean Squared Error (MSE) of the estimation.

To compare all the different relevant block lengths with sample size $n = 5000$, we will compare block lengths $b_n = 2, 7, 12, ..., 2497$. Note that block length $b_n = 1$ is not used, because then $M_{ni} = X_i$ and the concept of using blocks is gone. Also, the maximum compared block length is $b_n = 2497$ because, for larger block lengths the MSE goes to infinity. This is because, if the amount of blocks is $k_n = 1$, there is a chance that $\hat{N}_{ni} = \hat{F}(M_{ni}) = 1$ and then the estimator will not work.

With all those block lengths, the estimator is simulated 100 times. Then the MSE is calculated using the original value of θ for that sequence. We did this for the 4 estimators from Section 4 and for all the sequences in Section 2.

We will discuss the results of the MA(1) and the Moving Maxima sequence, the other distribution results are in Table 2 and Appendix B. This is because some results, like the sequences who are all iid, are very much the same.

We start with the results of the MA(1) sequence. In Figure 9a are the results of the MSE with different block lengths for estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$, in Figure 9b are the results for estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$. Both figures look the same, however in Figure 9 at the beginning, there is a small difference compared to Figure 9a. We will focus on this difference later.



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for MA(1) for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for MA(1) for different block lengths

Figure 9: MSE of MA(1) for different block lengths

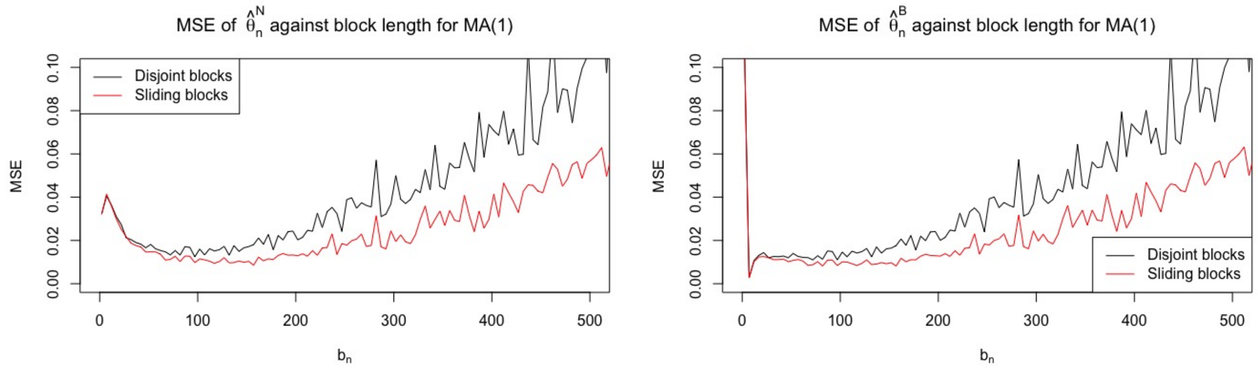
In both the figures, there are some remarkable observations. First, for both the disjoint blocks estimators, there are some jumps around $b_n = 1000, 1250, 1600$. This is because for estimations with disjoint blocks there are few values for which $k_n = n/b_n$ is an integer. Using the `floor()` argument in RStudio, there are simulations with the same amount of blocks k_n. However, those blocks aren't the same, see an example in Table 1. When the value of `floor(` k_n`)` decreases, a large amount of data is not used, see the last column in Table 1. This results in a higher MSE. Therefore, the jumps are due to the changing amounts of blocks k_n.

b_n	k_n	`floor(` k_n`)`	**First block**	**Second block**	**Third block**
1665	3,003	3	$X_1, .., X_{1665}$	$X_{1666}, ..., X_{3331}$	$X_{3332}, ..., X_{4997}$
1666	3,001	3	$X_1, .., X_{1666}$	$X_{1667}, ..., X_{3333}$	$X_{3334}, ..., X_{4999}$
1667	2,999	2	$X_1, .., X_{1667}$	$X_{1668}, ..., X_{3335}$	Out of range
1668	2,998	2	$X_1, .., X_{1668}$	$X_{1669}, ..., X_{3337}$	Out of range

Table 1: Explanation of the jumps in figures 9a and 9b

Another remarkable thing that we observe, is that for the sliding blocks estimators, when the block length increases, the MSE increases as well. This is because if the block length b_n increases, then $\hat{N}_{ni} = \hat{F}(M_{ni})$ increases as well, this results in a higher $\hat{Y}_{ni}$ and $\hat{Z}_{ni}$, which results into a higher MSE.

Now, we zoom in on Figures 9a and 9b and look at Figures 10a and 10b. For the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ the block length $b_n = 7$ is the optimal block length, however for estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ the optimal block length is larger. For $\hat{\theta}_n^{\mathrm{N,dj}}$ it is $b_n = 97$ and for $\hat{\theta}_n^{\mathrm{N,sl}}$ it is $b_n = 157$. It can be concluded that the MSE for the sliding block estimators are lower than the disjoint block estimators for different block lengths.

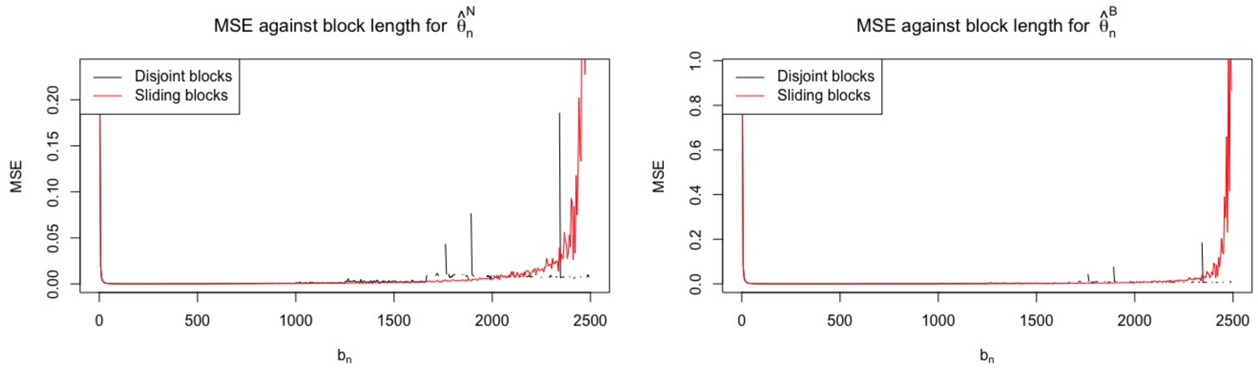


(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for MA(1) for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for MA(1) for different block lengths

Figure 10: MSE of MA(1) for different block lengths

Also, for the Moving Maxima sequence we discuss the results in detail. In Figure 11a the results of the MSE with different block lengths for estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ are presented, in Figure 11b the results for estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ are presented. At the beginning of both figures, the MSE decreases until its optimal point, and from $b_n = 2000$ the MSE grows exponential. Note that there are some gaps in both figures, that is because the MSE value is infinity. Let's zoom in at the optimal point.

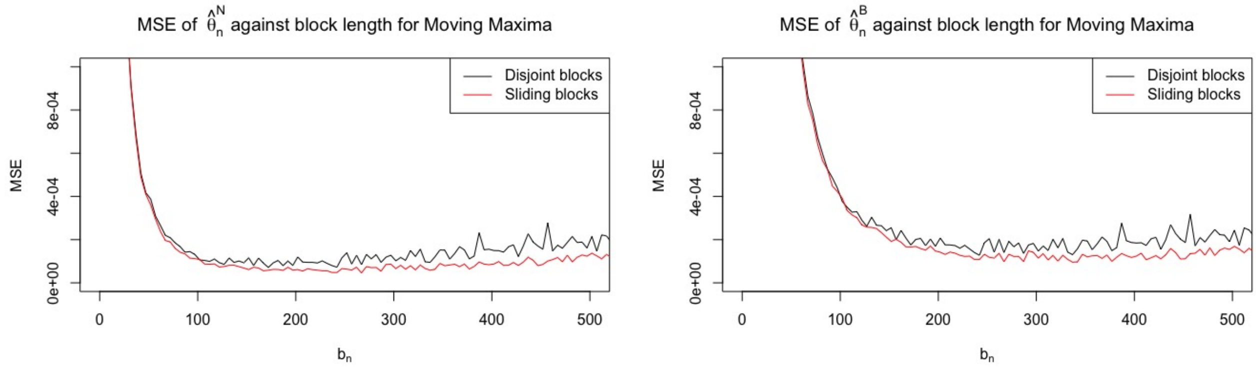


(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for Moving Maxima for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for Moving Maxima for different block lengths

Figure 11: MSE of Moving Maxima for different block lengths

Zooming in on Figures 11a and 11b results in Figure 12. The MSE decreases exponentially until its optimal point, then it increases. All the optimal points are at different block lengths, for $\hat{\theta}_n^{\mathrm{N,dj}}$ it is at $b_n = 172$ and for $\hat{\theta}_n^{\mathrm{N,sl}}$, $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ it is at $b_n = 267, 242, 337$ respectively. Also in these cases the MSE for the sliding blocks estimators are smaller than the disjoint blocks estimators for different block lengths.



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for Moving Maxima for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for Moving Maxima for different block lengths

Figure 12: MSE of Moving Maxima for different block lengths

For all the other sequences in this paper we have made the same simulations, in the table below the results are represented. It seems that all the iid sequences work optimally for the same block lengths approximately, while the other sequences have different results. Also, for almost all the sequences, except the ARCH and Moving Maxima sequences, there is a clear difference in block length between the 2 variants of the estimator.

	Optimal b_n for estimator			
Model	$\hat{\theta}_n^{\mathrm{N,dj}}$	$\hat{\theta}_n^{\mathrm{N,sl}}$	$\hat{\theta}_n^{\mathrm{B,dj}}$	$\hat{\theta}_n^{\mathrm{B,sl}}$
U(0,1)	2	2	32	57
N(0,1)	2	2	27	42
Exp(1)	2	2	32	32
Gamma(2,3)	2	2	27	27
MA(1)	97	157	7	7
ARCH	37	92	92	92
sARCH	17	17	37	47
Moving Maxima	172	267	242	337
AR-C	12	12	47	47

Table 2: Optimal b_n for every estimator

We will use the results of Table 2 for every simulation to estimate θ in further sections of this paper. In Appendix B the zoomed figures for all the other models are presented, that we did not discussed in detail.

6 Estimating the extremal index

For different sequences with sample size $n = 5000$, θ has been estimated 100 times using 2 different estimators with 2 different types of dividing blocks, disjoint blocks and sliding blocks. Also for every distribution, the optimal block length has been computed in the previous section. The results of the optimal block lengths computations are given in Table 2. We will discuss the results of 2 models in detail, the $U(0,1)$ and Moving Maxima model.

In Figure 13 the histograms for a $U(0,1)$ sequence are presented. The first variant of the estimator ($\hat{\theta}_n^{\mathrm{N,dj}}$, $\hat{\theta}_n^{\mathrm{N,sl}}$) is more accurate than the second variant ($\hat{\theta}_n^{\mathrm{B,dj}}$, $\hat{\theta}_n^{\mathrm{B,sl}}$), the second variant estimators are slightly too high. Comparing the disjoint with the sliding estimators doesn't give a clear, significant difference. The MSE and absolute bias of the sliding estimators are smaller than the disjoint estimators, see Table 3. Therefore the $\hat{\theta}_n^{\mathrm{N,dj}}$ estimator is the best for the $U(0,1)$ model.

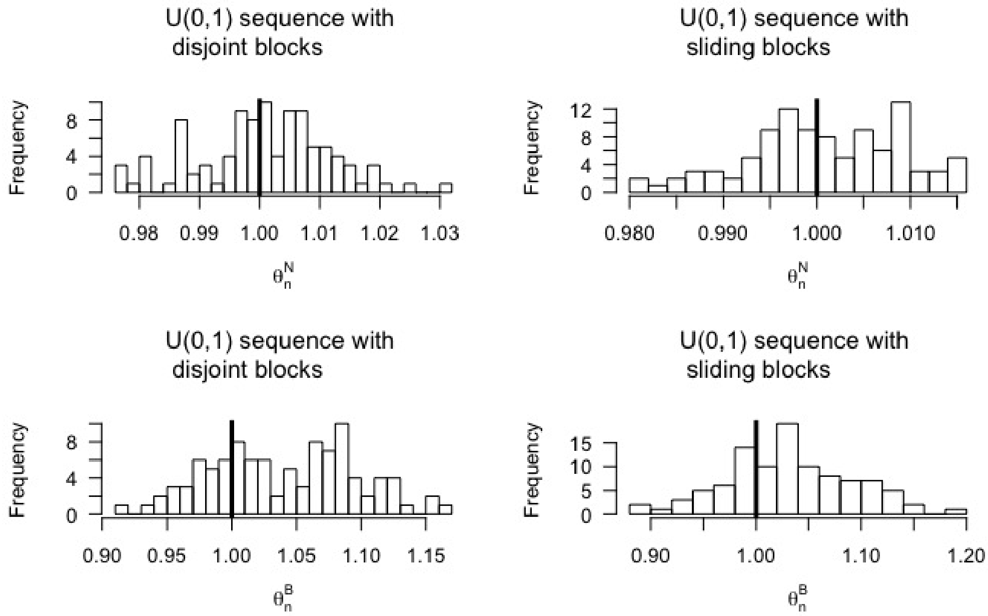


Figure 13: Histograms of estimators of θ for $U(0,1)$ sequence

In Figure 14 the histograms for a sARCH sequence are presented. In the upper 2 histograms, the bold line (the real value of $\theta = 0.727$) is closer to the middle than the two lower histograms. The MSE and absolute bias for the lower 2 histograms are also higher, see Table 3. Therefore, first variant of the estimator ($\hat{\theta}_n^{\mathrm{N,dj}}$, $\hat{\theta}_n^{\mathrm{N,sl}}$) is better than the second variant ($\hat{\theta}_n^{\mathrm{B,dj}}$, $\hat{\theta}_n^{\mathrm{B,sl}}$) for the sARCH model. Also with this model, comparing disjoint and sliding estimators to each other does not give useful results. The MSE of the sliding blocks estimators are smaller than the MSE of the disjoint estimators, but there is not a clear obvious difference between them. It can be concluded that also the $\hat{\theta}_n^{\mathrm{N,dj}}$ estimator is the best estimator for the sARCH model.

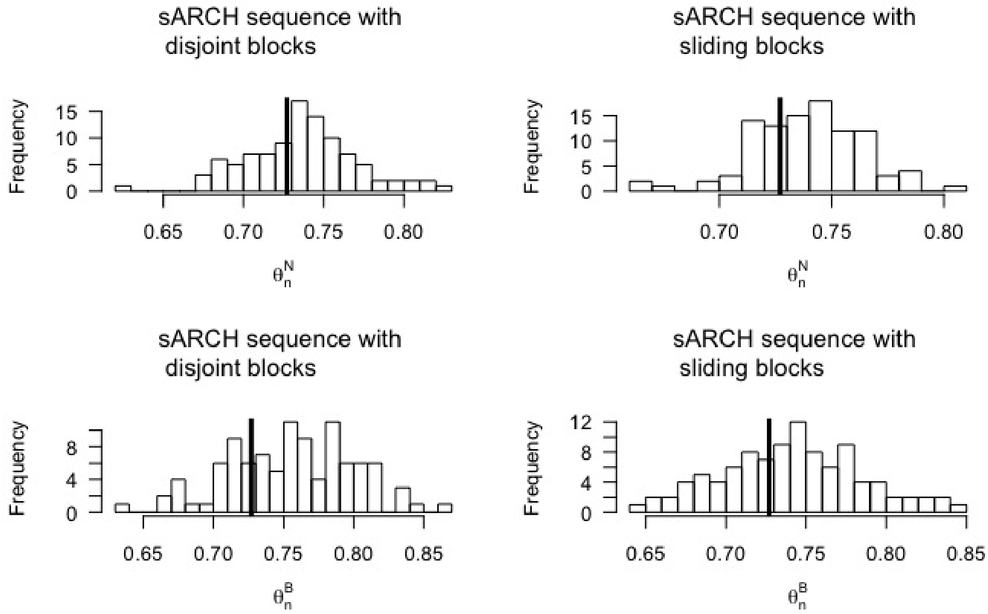


Figure 14: Histograms of estimators of θ for sARCH sequence

For the other distributions the results are in Table 3. All the histograms can be found in Appendix C. Comparing all the estimators, the sliding estimators have a smaller MSE than the disjoint estimators for almost all the distributions. We can conclude that the sliding block estimators are better for estimating θ.

Comparing the 4 estimators, the $\hat{\theta}_n^{\mathrm{N,sl}}$ estimator is the best estimator for estimating θ for almost all the models.

		Absolute Bias				**MSE**			
Model	θ	$\hat{\theta}_n^{\mathrm{N,dj}}$	$\hat{\theta}_n^{\mathrm{N,sl}}$	$\hat{\theta}_n^{\mathrm{B,dj}}$	$\hat{\theta}_n^{\mathrm{B,sl}}$	$\hat{\theta}_n^{\mathrm{N,dj}}$	$\hat{\theta}_n^{\mathrm{N,sl}}$	$\hat{\theta}_n^{\mathrm{B,dj}}$	$\hat{\theta}_n^{\mathrm{B,sl}}$
U(0,1)	1	0.001855	0.0005824	0.04920	0.03360	0.0001414	7.6118e-05	0.004120	0.002646
N(0,1)	1	0.001571	0.001511	0.05152	0.04641	0.0001271	7.2317e-05	0.004551	0.003158
Exp(1)	1	0.002701	0.001421	0.03646	0.03907	0.0001400	6.9881e-05	0.003691	0.003102
Gamma(2,3)	1	0.002727	0.002043	0.05299	0.04831	0.0001419	6.0257e-05	0.005392	0.01271
MA(1)	1	0.05562	0.02999	0.05239	0.05295	0.01190	0.008995	0.002957	0.002984
ARCH	0.721	0.05867	0.04238	0.06414	0.05301	0.004782	0.005080	0.007498	0.006158
sARCH	0.727	0.006896	0.007003	0.01165	0.006339	0.001308	0.0008225	0.002903	0.002773
Moving Maxima	0.04	0.006888	0.006081	0.01244	0.009341	8.1320e-05	6.4232e-05	0.0001431	0.0001156
AR-C	0.3	0.001702	0.0004053	0.01950	0.01760	0.0001599	0.0001148	0.001255	0.0008936

Table 3: Results of estimating θ

7 Application to wind gust data and precipitation data

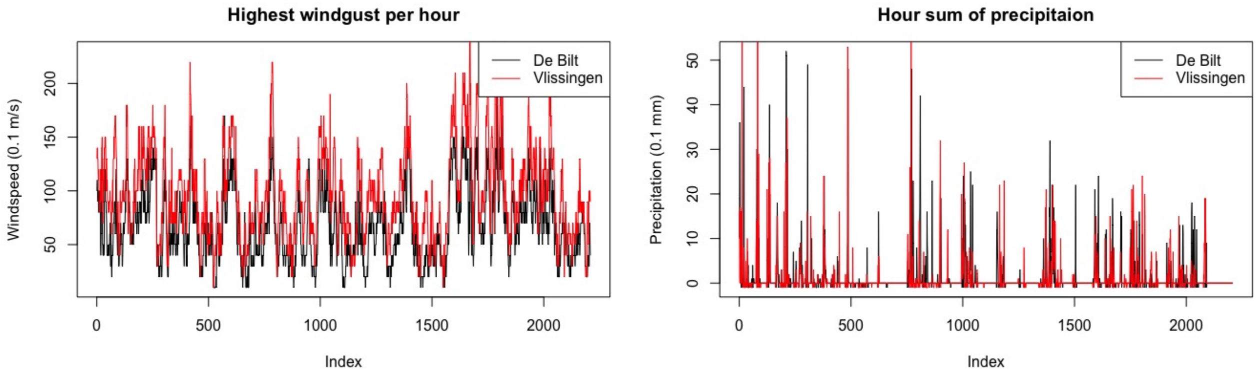
In this section, the value of θ will be compared for different types of weather conditions (wind gusts and precipitation) and at different weather stations (De Bilt and Vlissingen). Estimator $\hat{\theta}_n^{\mathrm{N,sl}}$ will be used to estimate θ, because in Section 6 we concluded that it is the best estimator.

7.1 Data

The data which is used, is taken from the KNMI (Koninklijk Nederlands Meteorologisch Instituut) database [1]. It is the national data- and knowledge-center for the weather, the climate and the seismology for The Netherlands. The data set contains the highest wind gust per hour in 0.1 m/s and the hour sum of precipitation in 0.1 mm starting at 2019-10-01 00:00, till 2019-12-31 23:59, at weather stations De Bilt and Vlissingen in The Netherlands. Note that when the precipitation is less than 0.05 mm, the given value is -1. The codes on the website for the highest wind gust per hour and hour sum of precipitation are FX and RH, respectively.

7.2 Plotted data

Vlissingen is located at the coast of The Netherlands, while De Bilt is in the middle of The Netherlands. At the coast, wind and wind gusts can behave differently (Endean (2010) [5]) due to the wind blowing over and around solid obstructions, which influences the overall weather pattern. In Figures 15a and 15b the wind gust and precipitation data are plotted at stations De Bilt and Vlissingen. It seems that the wind gusts are stronger in Vlissingen than in De Bilt. The precipitation seems to be slightly different in the 2 cities.



(a) Highest wind gust per hour in October, November and December 2019

(b) Hour sum of the precipitation per hour in October, November and December 2019

Figure 15: Data from KNMI at stations De Bilt and Vlissingen

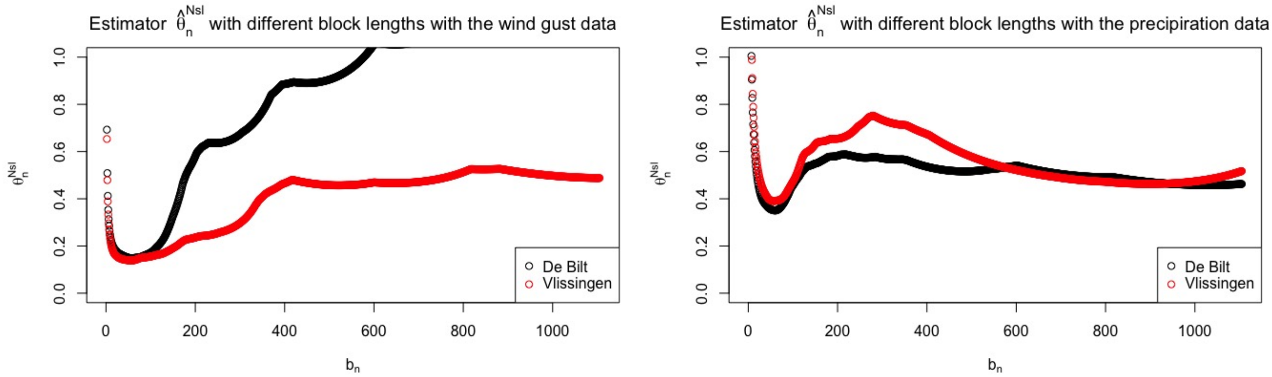
7.3 Estimating the extremal index

Estimator $\hat{\theta}_n^{\mathrm{N,sl}}$ has been used to estimate θ in Figures 16a and 16b for the wind gust and precipitation data at stations The Bilt and Vlissingen. A clear result is that θ is lower for the wind gusts for certain block lengths compared to the precipitation, this means that the wind gusts extremes are more dependent on each other than the precipitation extremes. This makes sense, it is more likely to stop raining immediately than that it stops to being windy.

Another result, looking at Figure 16a, $\hat{\theta}_n^{\mathrm{N,sl}}$ seems to be higher at The Bilt than at Vlissingen, this means that the wind gusts at Vlissingen are more dependent on each other compared to the wind gusts at De Bilt.

Looking at Figure 16b, $\hat{\theta}_n^{\mathrm{N,sl}}$ fluctuates for both stations with different block lengths. It is hard to conclude the difference at 2 the weather stations. Apparently, the two locations do not influence the way the extreme values of precipitation behave.

[1] https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi



(a) Estimation of θ for the wind gusts at De Bilt and Vlissingen for different block lengths

(b) Estimation of θ for the precipitation at De Bilt and Vlissingen for different block lengths

Figure 16: Estimations of θ for different block lengths

8 Conclusion

In this project, the main goal was investigating the 3 research questions:

- *What is the behaviour of extremes of different types of sequences?*
- *Which type of estimator is the best for estimating θ, using the simulated optimal block length b_n for every sequence?*
- *How does the extremal index behave for data from the KNMI?*

In the first part of the project, we investigated the behaviour of different sequences. We saw an obvious difference between the extreme values of sequences where $\theta = 1$ for and where $\theta < 1$. The extremes of sequence with $\theta = 1$ were independent and randomly distributed. Compared to sequences where θ was very low, there were clear peaks with multiple extreme values. Also, we made a comparison study on the maximum values of sequences which are the same but different in dependence. The results showed that for sequences where $\theta < 1$, the $F(M_n)$-values behaved different.

Continuing, we introduced different estimators, 2 different variants of estimators with disjoint blocks or sliding blocks. First, the optimal block length was determined with the use of the MSE. For almost all the models there was a clear difference in the 2 variants of estimators. The iid sequences all had approximately the same optimal block length. For the other distributions, the optimal block lengths were very different from each other. Another result, was that the MSE of the sliding block estimators were lower than the disjoint block estimators for all the block lengths. Then, θ was estimated using the optimal block lengths. The estimated value was compared to the original value. Again, there was a clear pattern in the MSE of all the iid distributions. Comparing the disjoint block estimators with the sliding block estimator, the sliding block estimators performed better, the MSE were smaller for almost all the models. Comparing all the estimators, the $\hat{\theta}_n^{\mathrm{N,sl}}$ estimator is the best for almost all the models.

Finally, the best estimator, $\hat{\theta}_n^{\mathrm{N,sl}}$, was used to estimate θ for real data from the KNMI. The wind gusts and precipitation were compared for two weather stations, De Bilt and Vlissingen. It resulted in that the wind gusts have a lower extremal index than the precipitation. Also, the extremes of the wind gusts at Vlissingen are more dependent on each other than at De Bilt. For the precipitation at the 2 stations, no large, significant difference was observed.

8.1 Further research

Different topics have been studied in this project. When interested in the optimal block length, it is possible to simulate more precisely. We did not simulated all the block lengths, but block lengths $b_n = 2, 7, 12, ...,$ because it took a long running time for the program 'RStudio' and my laptop, to simulate them all. So with better and faster hardware these simulations can be made.

Another interesting thing could be comparing more different types of estimators. Moreover, other variants could be used or the data could be divided in different ways. Also, in this project the estimators have been compared by looking at simulated results. Another way to compare these, is by looking more theoretical at them.

Last, the peaks of the distributions could be looked at more closely. When is a peak a cluster? How many extremes are needed to become a cluster. Identifying those clusters is called declustering. To use the declustering methods, Segers presents them in his paper (Segers (2014)[9]).

Appendices

A Compare sequences

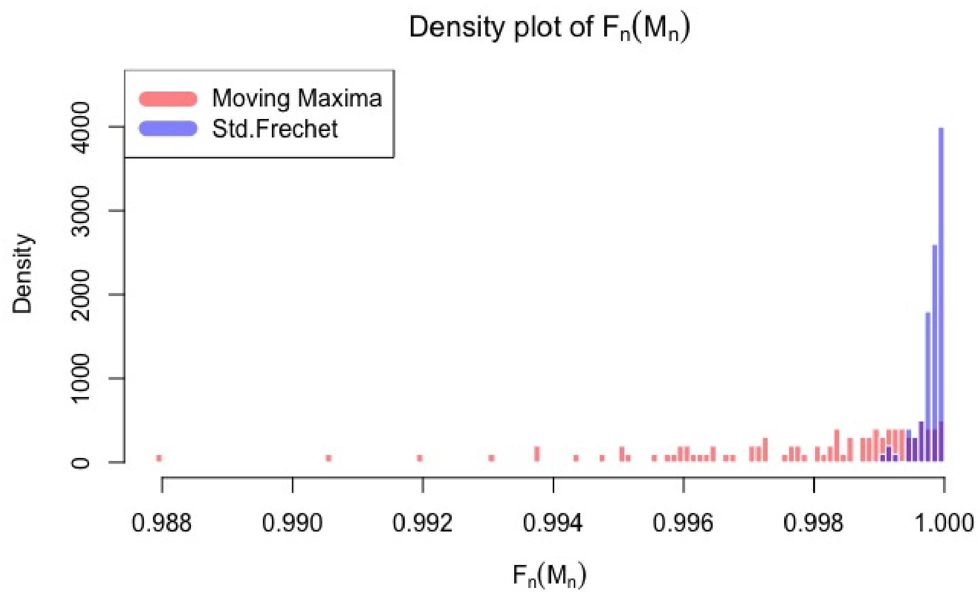


Figure 17: Density plot of $F(M_n)$

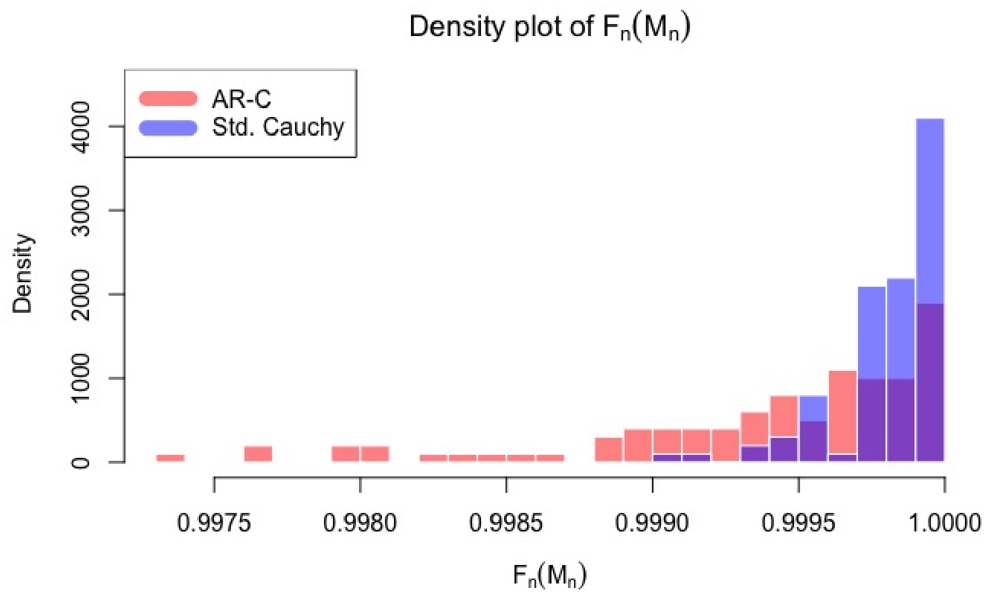
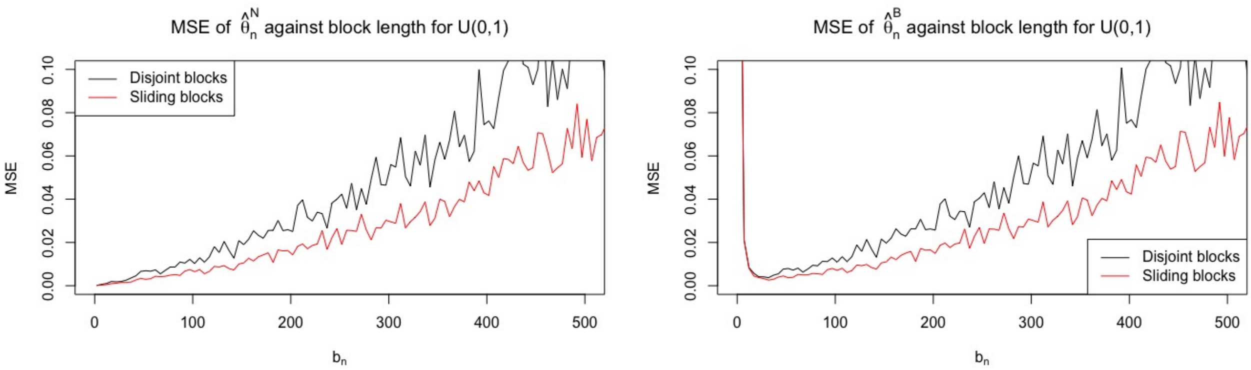


Figure 18: Density plot of $F(M_n)$

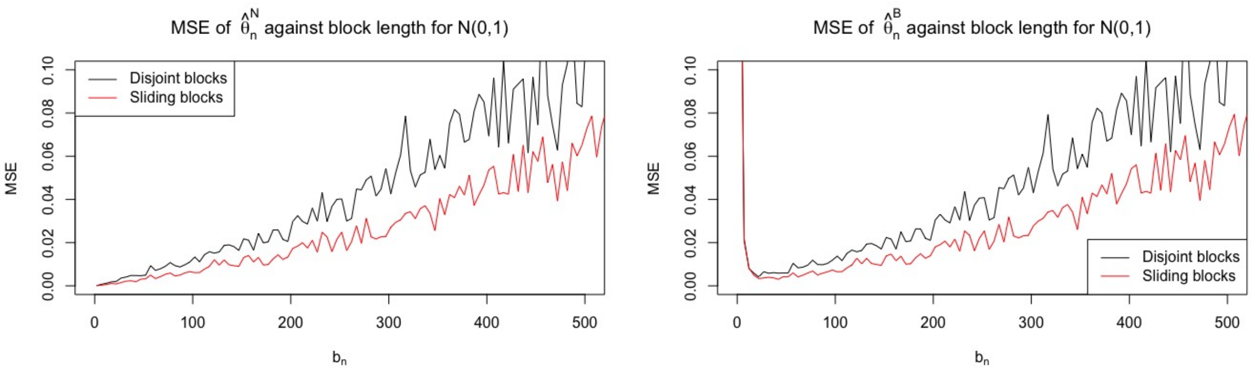
B Optimal block length results



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for $U(0,1)$ for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for $U(0,1)$ for different block lengths

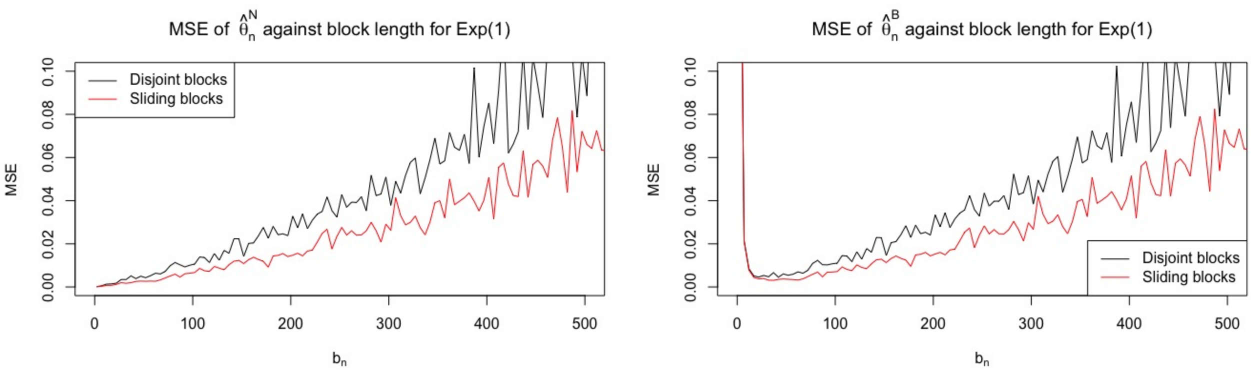
Figure 19: MSE of $U(0,1)$ for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for $N(0,1)$ for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for $N(0,1)$ for different block lengths

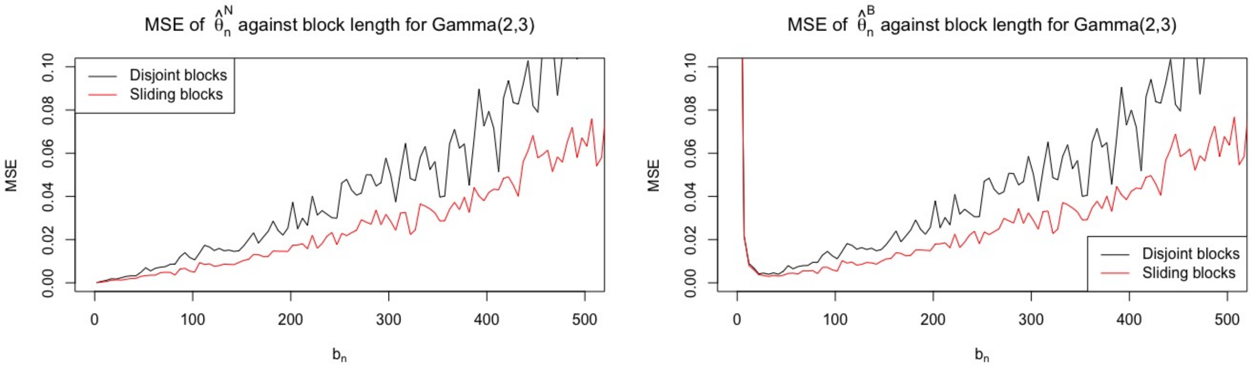
Figure 20: MSE of $N(0,1)$ for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for $Exp(1)$ for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for $Exp(1)$ for different block lengths

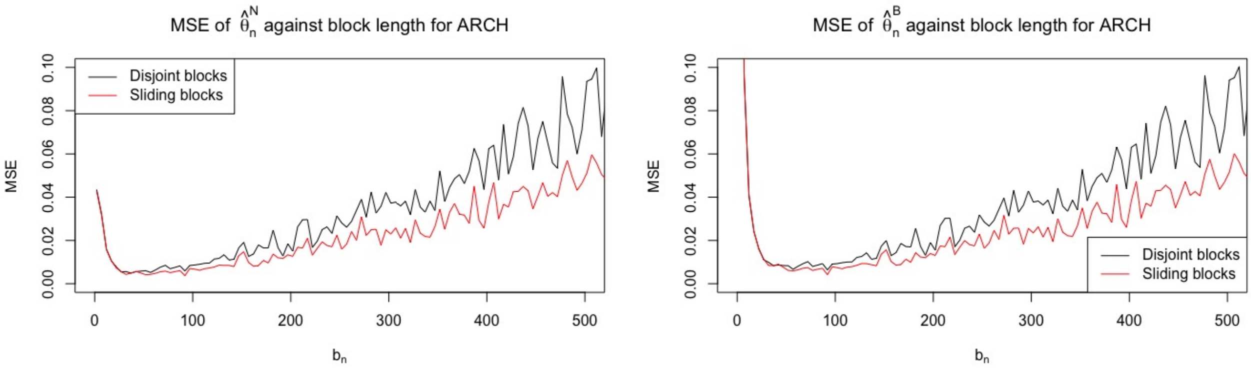
Figure 21: MSE of $Exp(1)$ for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for $Gamma(2,3)$ for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for $Gamma(2,3)$ for different block lengths

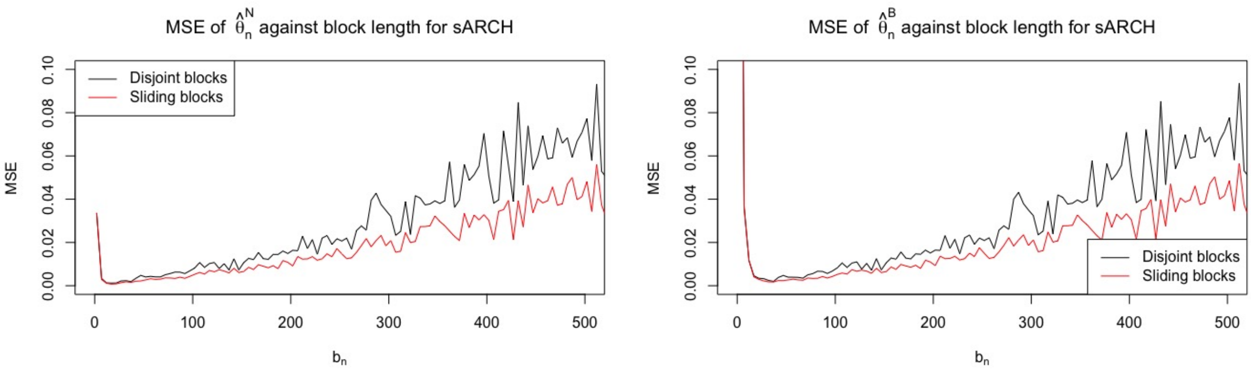
Figure 22: MSE of $Gamma(2,3)$ for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for ARCH for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for ARCH for different block lengths

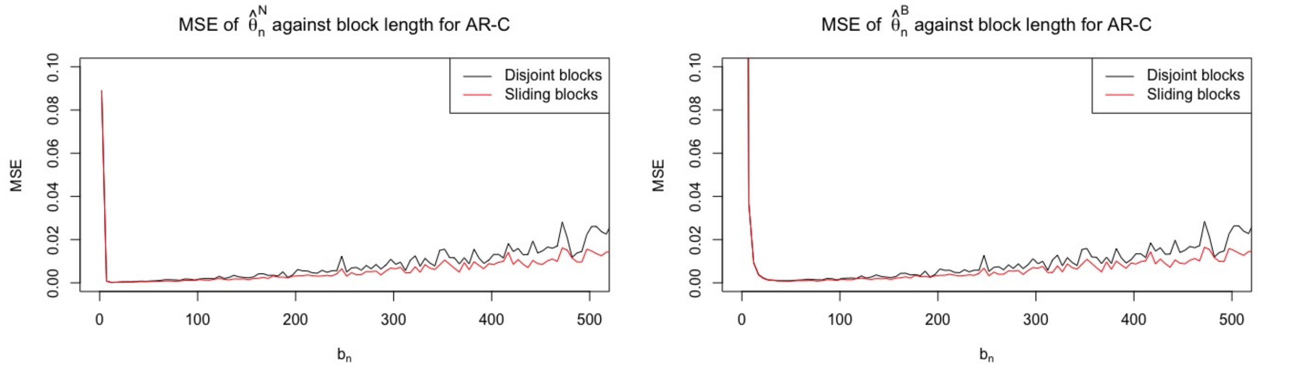
Figure 23: MSE of ARCH for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for sARCH for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for sARCH for different block lengths

Figure 24: MSE of sARCH for different block lengths



(a) MSE of the estimators $\hat{\theta}_n^{\mathrm{N,dj}}$ and $\hat{\theta}_n^{\mathrm{N,sl}}$ for AR-C for different block lengths

(b) MSE of the estimators $\hat{\theta}_n^{\mathrm{B,dj}}$ and $\hat{\theta}_n^{\mathrm{B,sl}}$ for AR-C for different block lengths

Figure 25: MSE of AR-C for different block lengths

C Estimating the extremal index histograms

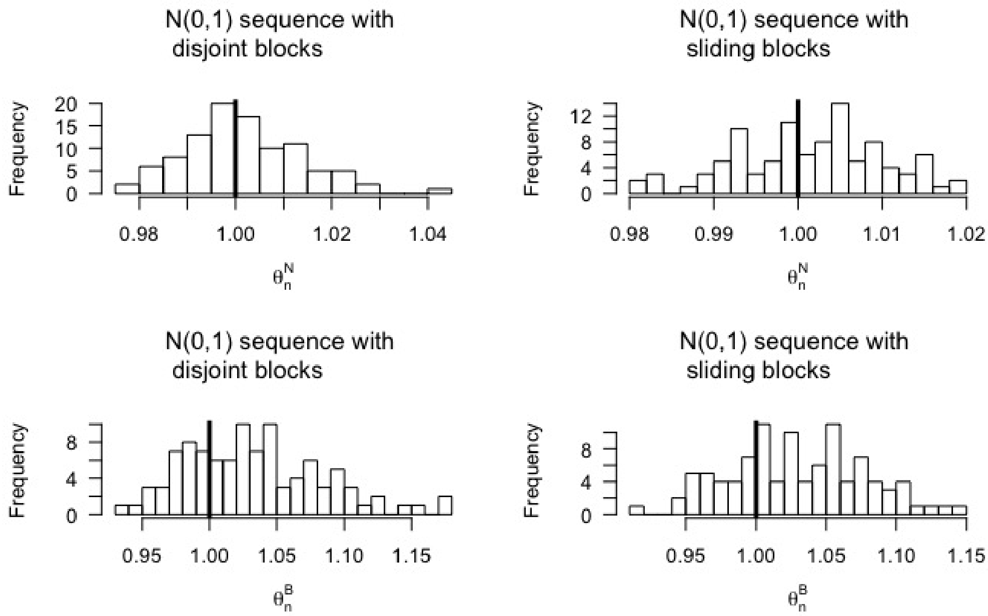


Figure 26: Histograms of estimators of θ for $N(0,1)$ model

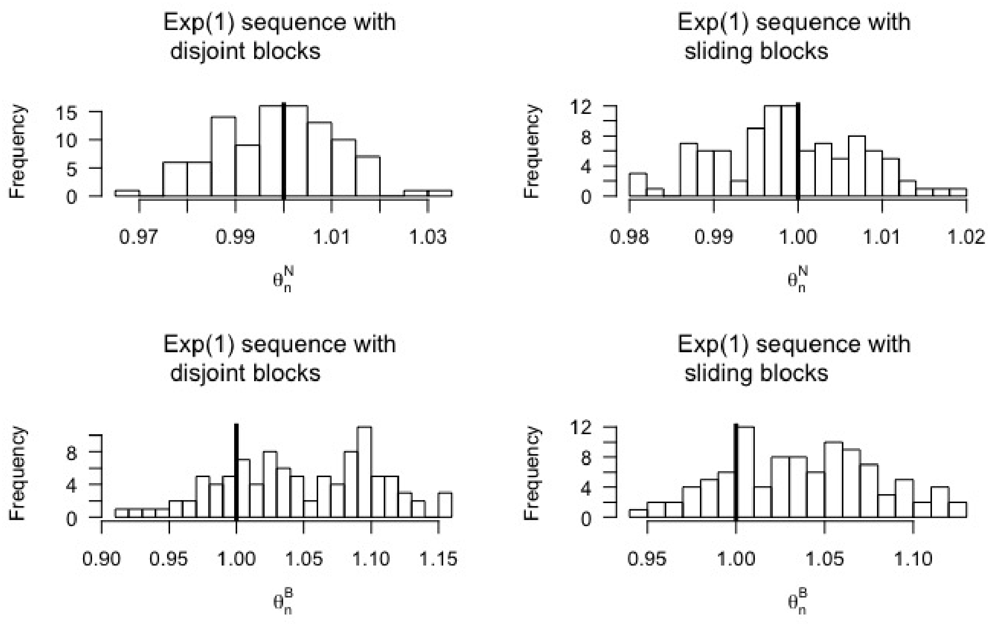


Figure 27: Histograms of estimators of θ for $Exp(1)$ model

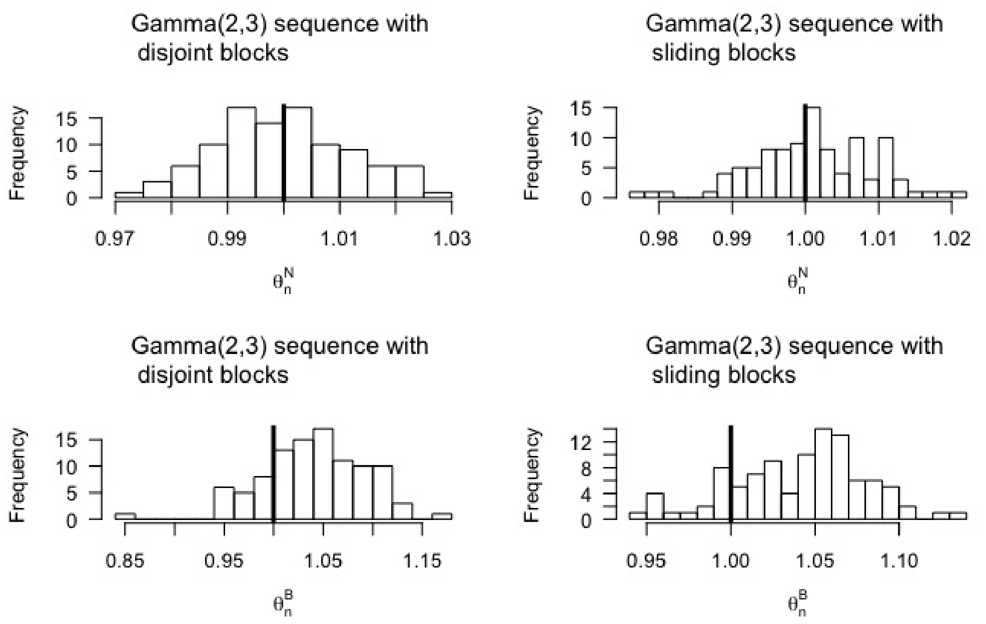


Figure 28: Histograms of estimators of θ for $Gamma(2,3)$ model

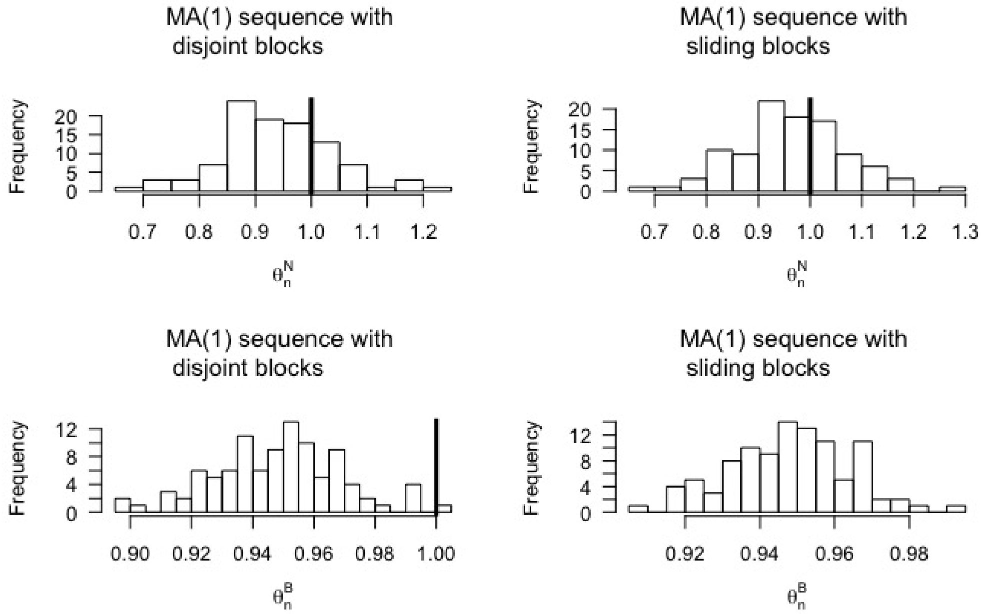


Figure 29: Histograms of estimators of θ for MA(1) model

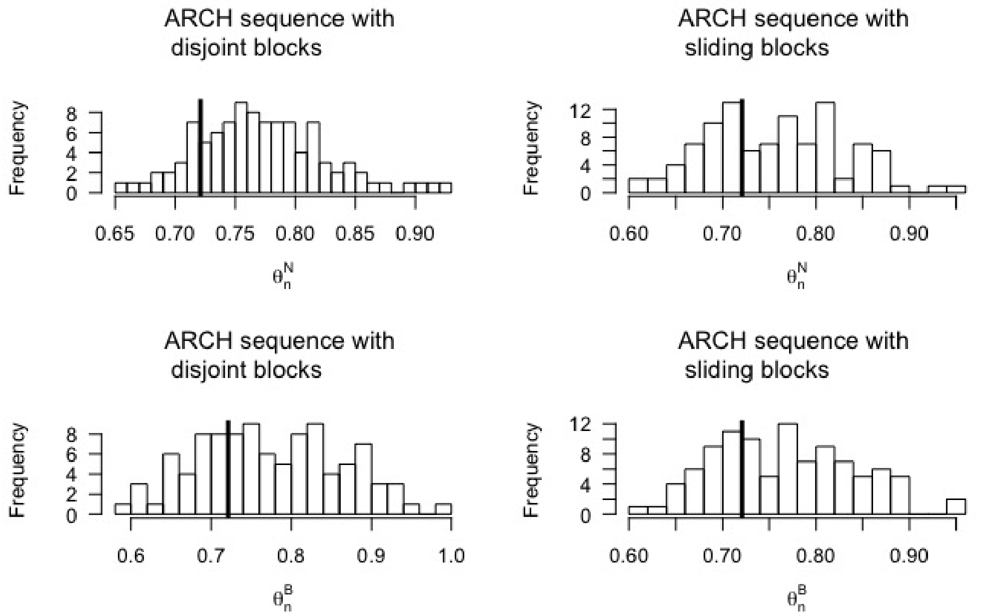


Figure 30: Histograms of estimators of θ for ARCH model

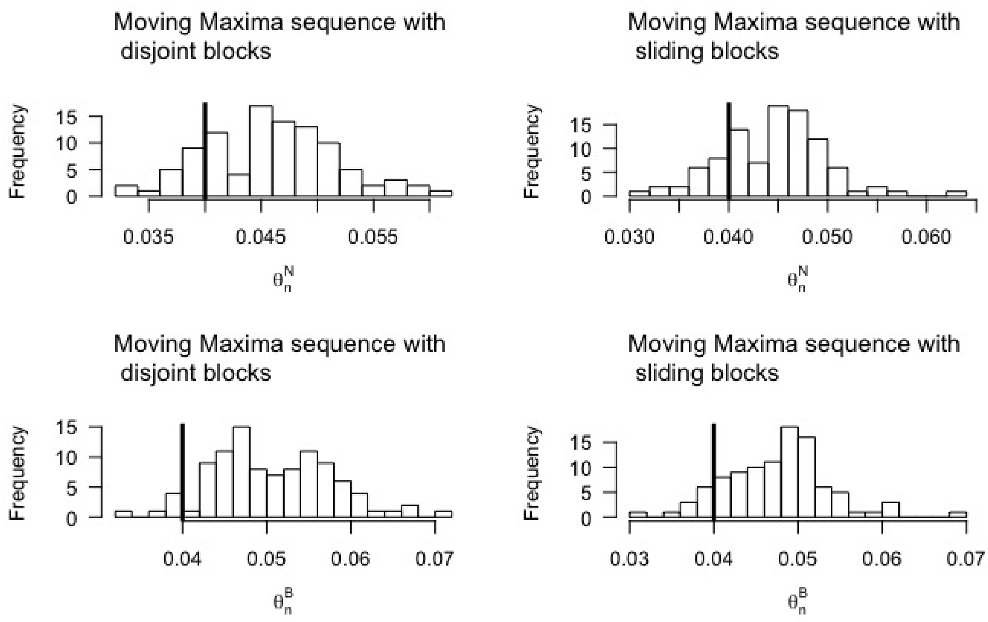


Figure 31: Histograms of estimators of θ for Moving Maxima model

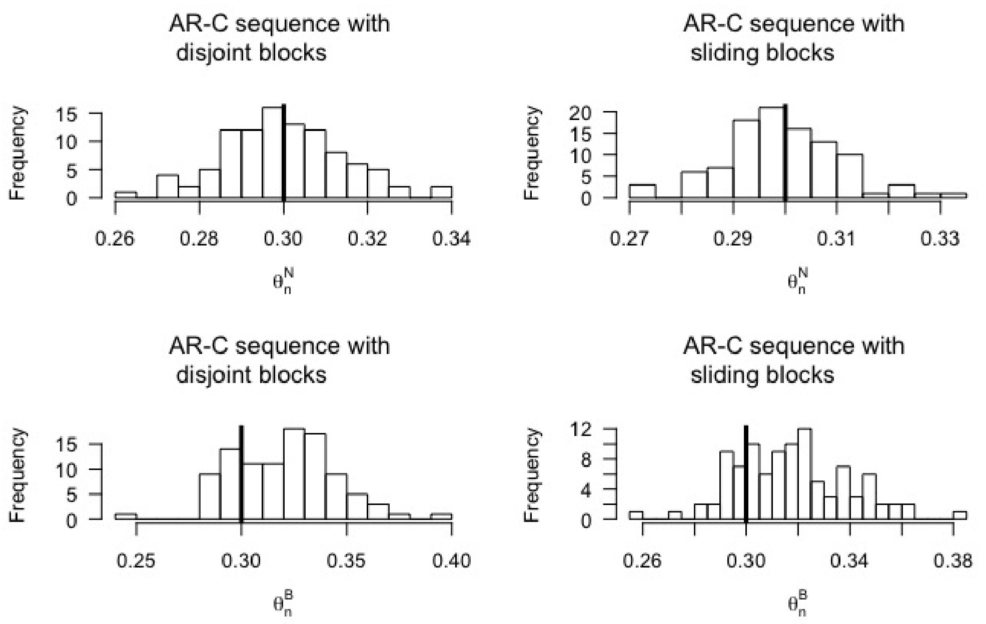


Figure 32: Histograms of estimators of θ for AR-C model

References

Betina Berghaus and Axel Bücher. Weak convergence of a psuedo maximum likelihood estimator for the extremal index. *Journal of Chemical Information and Modeling*, 53(9):1689–1699, 2013.

Juan Juan Cai. A nonparametric estimator of the extremal index. pages 1–26, 2019.

Laurens de Haan, Sidney I. Resnick, Holger Rootzén, and Casper G. de Vries. Extremal behaviour of solutions to a stochastic difference equation with applications to arch processes. *Stochastic Processes and their Applications*, 32(2):213–224, 1989.

F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester. A Modern Introduction to Probability and Statistics: Understanding Why and How. *Journal of the American Statistical Association*, mar 2005.

Ken Endean. *Coastal Turmoil: Winds, Waves and Tidal Races*. A&C Black, 1 edition, 2010.

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

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

John A. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole, Duxbury, 3rd edition, 2007.

Johan Segers. Automatic declustering of extreme values via an estimator for the extremal index Automatic declustering of extreme values via an estimator for the extremal index. (April 2002), 2014.

Frank Van der Meulen. *Lecture Notes Financial Time Series*. Delft, 1.04 edition, 2019.