UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 05-18
Economics Series 09
April 2005

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624 98 75

TESTING THE EXISTENCE OF CLUSTERING IN THE EXTREME VALUES *

José Olmo [1]

Abstract

This paper introduces an estimator for the extremal index as the ratio of the number of elements of two point processes defined by threshold sequences u_n, v_n and a partition of the sequence in different blocks of the same size. The first point process is defined by the sequence of the block maxima that exceed u_n. This paper introduces a thinning of this point process, defined by a threshold v_n with $v_n > u_n$, and with the appealing property that under some mild conditions the ratio of the number of elements of both point processes is a consistent estimator of the extremal index. The method supports a hypothesis test for the extremal index, and hence for testing the existence of clustering in the extreme values. Other advantages are that it allows some freedom to choose u_n, and it is not very sensitive to the choice of the partition. Finally, the stylized facts found in financial returns (clustering, skewness, heavy tails) are tested via the extremal index, in this case for the DaX returns

Keywords: Extremal Index, Extreme Value Theory, GARCH processes, Poisson Processes.

JEL code: C12, C13, C15, C51.

* **Acknowledgements:**.. Financial support DGCYT Grant (SEC01-0890) is gratefully acknowledged. The author is deeply grateful to the Department of Statistics and Operations Research of UNC at Chapel Hill, specially to Ross Leadbetter and Francisco Chamu. The author also thanks the seminar participants of the 3rd Symposium on Extreme Value Analysis: Theory and Practice celebrated in Aveiro (Portugal) as well as its scientific committee for the C.E.A.U.L. prize for young researchers.

[1] Departamento de Economía ,Universidad Carlos III de Madrid, C/ Madrid, 126 28903 Getafe (Madrid), Spain. E-mail: jose.olmo@uc3m.es

Testing the Existence of Clustering in the Extreme Values

Jose Olmo *
Dept. of Economics, Universidad Carlos III de Madrid

This version, January 2005

Abstract

This paper introduces an estimator for the extremal index as the ratio of the number of elements of two point processes defined by threshold sequences $\{u_n\}$, $\{v_n\}$ and a partition of the sequence in different blocks of the same size. The first point process is defined by the sequence of the block maxima that exceed $\{u_n\}$. This paper introduces a thinning of this point process, defined by a threshold $\{v_n\}$ with $\{v_n\} > \{u_n\}$, and with the appealing property that under some mild conditions the ratio of the number of elements of both point processes is a consistent estimator of the extremal index. The method supports a hypothesis test for the extremal index, and hence for testing the existence of clustering in the extreme values. Other advantages are that it allows some freedom to choose $\{u_n\}$, and it is not very sensitive to the choice of the partition. Finally, the stylized facts found in financial returns (clustering, skewness, heavy tails) are tested via the extremal index, in this case for the DaX returns.

Keywords: Extremal Index, Extreme Value Theory, GARCH processes, Poisson Processes.

1 Background

Suppose a random sample from an unknown distribution function F, and let G be the limiting distribution of the sample maximum M_n. Classical Extreme Value Theory shows that under some regularity conditions on the tail of F and for some suitable constants $a_n > 0$, b_n,

$$P\{a_n^{-1}(M_n - b_n) \leq x\} \to G(x), \tag{1}$$

where G must be of the following types (see de Haan (1976)),

Type I: (Gumbel) $\quad G(x) = e^{-e^{-x}}, \quad -\infty < x < \infty.$

Type II: (Fréchet) $\quad G(x) = \begin{cases} 0 & x \leq 0, \\ e^{-x^{-\frac{1}{\xi}}} & x > 0, \xi > 0. \end{cases}$

Type III: (Weibull) $\quad G(x) = \begin{cases} 1 & x \geq 0, \\ e^{-(-x)^{-\frac{1}{\xi}}} & x < 0, \xi < 0. \end{cases}$

*Address for correspondence: Universidad Carlos III de Madrid, C/ Madrid 126, 28903, Getafe (Madrid). E-mail: jose.olmo@uc3m.es. Financial support DGCYT Grant (SEC01-0890) is gratefully acknowledged. The author is deeply grateful to the Department of Statistics and Operations Research of UNC at Chapel Hill, especially to Ross Leadbetter and Francisco Chamu. The author also thanks the seminar participants of the 3^{rd} Symposium on Extreme Value Analysis: Theory and Practice celebrated in Aveiro (Portugal) as well as its scientific committee for the C.E.A.U.L. prize for young researchers. The software developed for this paper is implemented in Matlab 6.1. and may be downloaded from the author homepage wwww.unc.edu/home/olmo.

This important result may be extended to study the maximum of a wide class of dependent processes. We concentrate here on stationary sequences where the dependence is restricted by different distributional *mixing* conditions. We distinguish two types of dependence: long range and short range dependence. To limit the first type of dependence we assume a variation of the distributional mixing condition $D(u_n)$ of Leadbetter et al. (1983). Leadbetter's mixing condition is said to hold for a sequence $\{u_n\}$ if for any integers $1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_{p'} \leq n$ for which $j_1 - i_p \geq l$, we have

$$D(u_n): \quad |F_{i_1,\ldots,i_p,j_1,\ldots,j_{p'}}(u_n) - F_{i_1,\ldots,i_p}(u_n)F_{j_1,\ldots,j_{p'}}(u_n)| \leq \alpha_{n,l},$$

where $\alpha_{n,l_n} \to 0$ as $n \to \infty$ for some $l_n = o(n)$, and $F_{i_1,\ldots,i_p}(u_n)$ denotes $P\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\}$. Let $D'(u_n)$ be the alternative mixing condition that will be used throughout the paper. This condition is as follows,

$$D'(u_n): |P\left\{X_{i_1} > u_n \, or \ldots or \, X_{i_p} > u_n \, or X_{j_1} > u_n \, or \ldots or \, X_{j_{p'}} > u_n\right\} -$$

$$-P\left\{X_{i_1} > u_n \, or \ldots or \, X_{i_p} > u_n\right\} P\left\{X_{j_1} > u_n \, or \ldots or \, X_{j_{p'}} > u_n\right\}| \leq \alpha_{n,l}. \tag{2}$$

Note that these conditions only concern events of the form $\{X_i > u_n\}$ in contrast to more restrictive mixing conditions, for example the strong mixing condition introduced in Rosenblatt (1956).

These mixing conditions alone are sufficient to extend the central result given in (1) to stationary sequences for some suitable constants not necessarily the ones obtained from the *iid* context. In particular these constants $a_n > 0$, b_n and the extreme value distribution G are the same of the *iid* case under a condition $D''(u_n)$ restricting short range dependence, Leadbetter (1983), that avoids the presence of clusters,

$$D''(u_n): \quad \underset{n\to\infty}{limsup} \quad n \sum_{j=2}^{[n/k_n]} P\{X_1 > u_n, X_j > u_n\} \to 0 \quad as \quad k_n \to \infty, \tag{3}$$

with k_n a sequence that defines a partition of the sample.

Otherwise, for a stationary sequence $\{X_n\}$ satisfying only $D'(u_n)$ with $u_n = a_n x + b_n$, we typically have

$$P\{a_n^{-1}(M_n - b_n) \leq x\} \to G^{\theta}(x), \tag{4}$$

where θ is the key parameter for extending extreme value theory for *iid* random variables to stationary sequences. This concept, originated in papers by Loynes (1965), O'Brien (1974) and developed in detail by Leadbetter (1983), reflects the effect of the clustering of the observations exceeding u_n on the limiting distribution of the maximum.

There are different interpretations of the extremal index θ, concerning diverse features of the clustering of the largest observations. Loynes (1965) under different mixing conditions found that

$$P\{M_n \leq u_n\} = F^{n\theta}(u_n). \tag{5}$$

O'Brien (1987) showed that

$$P\{M_{2,r_n} \leq u_n | X_1 > u_n\} \to \theta, \tag{6}$$

where M_{2,r_n} is the maximum of $\{X_2, \ldots, X_{r_n}\}$, and $r_n = o(n)$ satisfies certain growth conditions. Note that from this definition of the extremal index it is straightforward to see that $0 \leq \theta \leq 1$. Alternatively Leadbetter (1983) showed that the inverse of the extremal index is

the limiting mean number of exceedances of u_n in an interval of length r_n, *i.e.*

$$E\left[\sum_{j=1}^{r_n} I(X_j > u_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1\right] \to \theta^{-1}, \tag{7}$$

with $I(X > 0)$ the indicator function. By stationarity this is called the limiting mean cluster size of the process. Finally, Hsing (1993) and Ferro and Segers (2003) take advantage of the limiting probability

$$P\{M_n \leq u_n\} \to e^{-\theta\tau}, \tag{8}$$

with $0 < \tau < \infty$, in two different ways. Hsing approximates the distribution of $n(1 - F(M_n))$ by an exponential distribution with mean θ^{-1}, and Ferro and Segers model the process of the interexceedance times defined by u_n by the same limiting exponential distribution.

Expression (8) is a transformation of (4) where τ is the exponent of an extreme value distribution and $u_n = a_n x + b_n$. In the same way the limiting probability (1) may be written as $P\{M_n \leq u_n\} \to e^{-\tau}$. Taking logs in this expression, it is immediate to derive that $n(1 - F(u_n)) \to \tau$ for u_n sufficiently high. Then for *iid* sequences, $B_n^{(u_n)} = \sum_{j=1}^{n} I(X_j > u_n)$ converges in distribution to a Poisson random variable with mean τ.

However for dependent stationary sequences where $D''(u_n)$ is not satisfied $B_n^{(u_n)}$ does not converge to a Poisson random variable (the exceedances of u_n are not mutually independent), nevertheless we can define a point process as the result of thinning $B_n^{(u_n)}$. This thinning defines the process $N_{k_n}^{(u_n)}$ formed by the maxima over k_n blocks of length r_n and exceeding u_n, and converges to a Poisson process N with mean $\theta\tau$, see Leadbetter (1983) or Leadbetter et al. (1983). This paper presents an alternative derivation of the extremal index as the result of thinning twice $B_n^{(u_n)}$. The second thinning of $B_n^{(u_n)}$, and hence thinning of $N_{k_n}^{(u_n)}$, defines another point process $N_{k_n}^{(v_n)}$ that converges in distribution to a Poisson process with intensity $\theta^2\tau$. The sequence $\{v_n\}$ satisfies $n(1 - F(v_n)) \to \theta\tau$ and is defined by $E[\sum_{j=1}^{r_n} I(X_j > v_n) | \sum_{j=1}^{r_n} I(X_j > u_n) \geq 1] \to 1$. Under some mild conditions on the threshold sequence, this method provides a consistent estimator of the extremal index that outperforms most of the popular estimators and such that it is not very sensitive to the choice of the block size r_n nor the choice of the sequence $\{u_n\}$ in contrast to the rest of the candidates that estimate θ.

The paper is structured as follows. Section 2 introduces a definition of the extremal index as the ratio of two point processes derived from the asymptotic distribution of the maximum. A natural estimator for this parameter based on these techniques is introduced in Section 3. This section also reviews some of the most popular estimators found in the literature and their statistical properties, in particular bias and variance. The corresponding properties of our estimator are also studied with special emphasis in the analysis of the mean square error of the different methods. The optimal block size selection is also considered and the section concludes with a hypothesis test for the extremal index that is sufficient to test the existence of clustering in the extremes. A simulation experiment for different examples presented in the literature is conducted in Section 4 stressing a Monte-Carlo experiment for the mean square error. Section 5 presents an application to DaX Index returns in order to gain some understanding about the clustering in the extremes and in the volatility of the process. Finally the conclusions are found in Section 6.

2 Definition of the extremal index

Suppose throughout that we have n observations from a stationary sequence $\{X_i, i \geq 1\}$ with marginal distribution function F satisfying $[1 - F(x)]/[1 - F(x^-)] \to 1$ as $x \to \infty$. This

condition is sufficient to define a sequence $\{u_n\}$ for each $0 < \tau < \infty$ such that

$$n(1 - F(u_n)) \to \tau. \tag{9}$$

Consider from now on that $\{X_n\}$ satisfies $D'(u_n)$, as defined in (2), for each $\tau > 0$. Intuitively this condition gives a measure of the degree of dependence in the process and permits the construction of *almost* independent blocks by the definition of sequences $\{k_n\}$, $\{r_n\}$ with $k_n \to \infty$, $k_n = o(n)$ and $k_n l_n = o(n)$, while r_n is the integer part of n/k_n. The interpretation of these sequences is: k_n is the number of blocks of the sequence of length n, and r_n the size of each block.

Under these assumptions, if $P\{M_n \leq u_n\}$ converges for some $\tau > 0$ then

$$P\{M_n \leq u_n\} \to e^{-\theta\tau}, \tag{10}$$

for all $\tau > 0$, with $0 \leq \theta \leq 1$ (see theorem 3.7.1. of Leadbetter et al. (1983) for a detailed proof). The parameter θ is called the extremal index of the sequence $\{X_n\}$ and is the key parameter for extending extreme value theory from *iid* random variables to stationary processes.

Consider $\{k_n\}$, $\{r_n\}$ that define a suitable partition of the sequence $\{X_n\}$, then a sufficient condition for the existence of the extremal index is

$$k_n(1 - F_{1,\ldots,r_n}(u_n)) \to \theta\tau. \tag{11}$$

This result is immediate by the approximation of $P\{M_n \leq u_n\}$ by $P^{k_n}\{M_{r_n} \leq u_n\}$ for suitable choices of k_n and r_n, and (10) and the linear polynomial expansion of the exponential function. The converse of this result is also true, *i.e.* a stationary sequence with extremal index θ satisfies (11) for each $\tau > 0$. The proof is obtained by taking logs in the expression $P^{k_n}\{M_{r_n} \leq u_n\}$ that approximates $e^{-\theta\tau}$.

Consider the number of exceedances of u_n within a block of size r_n. This event defines a sequence of random variables $B_{r_n}^{(u_n)} = \sum_{j=1}^{r_n} I(X_j > u_n)$ for $r_n \to \infty$, and $r_n = o(n)$ whose expected value, by the stationarity of the process, converges to the mean cluster size of the exceedances of u_n in the sequence $\{X_n\}$, that is, the inverse of the extremal index,

$$E\left[B_{r_n}^{(u_n)} | B_{r_n}^{(u_n)} \geq 1\right] \to \theta^{-1}.$$

This is readily seen since $E\left[B_{r_n}^{(u_n)} | B_{r_n}^{(u_n)} \geq 1\right] = \sum_{j=1}^{\infty} jP\left\{B_{r_n}^{(u_n)} = j | B_{r_n}^{(u_n)} \geq 1\right\} = \frac{r_n P\{X_j > u_n\}}{P\{\bigcup_{j=1}^{r_n}(X_j > u_n)\}}$,

and therefore $E\left[B_{r_n}^{(u_n)} | B_{r_n}^{(u_n)} \geq 1\right] = \frac{r_n(1-F(u_n))}{1-F_{1,\ldots,r_n}(u_n)} \to \theta^{-1}$, if (11) holds.

The same argument may be applied to define a process $B_{r_n}^{(v_n)}$ with $v_n \geq u_n$ satisfying

$$E\left[B_{r_n}^{(v_n)} | B_{r_n}^{(u_n)} \geq 1\right] \to 1. \tag{12}$$

It is of interest to note that the sequence $\{v_n\}$ satisfies condition $D'(v_n)$ since $v_n \geq u_n$ and

$$n(1 - F(v_n)) \to \theta\tau \quad as \quad n \to \infty. \tag{13}$$

In addition, by the structure of dependence (see (9) and (10)) we have $P\{M_n \leq v_n\} \to e^{-\theta^2\tau}$. It is immediate now to see that (11) holds for the sequence $\{v_n\}$ by

$$k_n(1 - F_{1,\ldots,r_n}(v_n)) \to \theta^2\tau. \tag{14}$$

The event $\{X_i > u_n\}$ and the sequences $\{k_n\}$, $\{r_n\}$ divide the sequence $\{X_n\}$, with extremal index θ, in approximately independent groups of exceedances of u_n where $M_{(j-1)r_n+1,jr_n}$

is the block maxima for $j = 1, \ldots, k_n$. It is clear that the sequence $\{M_{(j-1)r_n+1,jr_n}\}$ is approximately serially independent as n increases if $D'(u_n)$ holds for $\{X_n\}$.

Consider the points j as points in time and define for each n, and k_n, a process $\eta_{k_n}(j/k_n) = M_{(j-1)r_n+1,jr_n}$. The time scale is normalized $t = j/k_n$ on the unit interval $(0, 1]$. Then the exceedances of u_n by the process $\eta_{k_n}(t)$ define a point process $N_{k_n}^{(u_n)}$ on the unit interval (see Kallenberg (1976) for the theory of point processes). Moreover, the point process $N_{k_n}^{(u_n)}$ converges in distribution to a Poisson process N on $(0, 1]$ with intensity parameter $\theta\tau$. To prove this result it is only necessary to show that $E[N_{k_n}^{(u_n)}(a, b]] \to E[N(a, b]]$ for $0 < a < b \leq 1$ and $P\{N_{k_n}^{(u_n)}(A) = 0\} \to P\{N(A) = 0\}$ for each finite disjoint union A of sets $(a_i, b_i] \subset (0, 1]$. The proof is analog to the corresponding one found in theorem 4.1. in Leadbetter (1983).

It is interesting to see that the same argument may be applied to construct a thinning of $N_{k_n}^{(u_n)}$ by a sequence $\{v_n\}$ satisfying (12). This sequence defines the point process $N_{k_n}^{(v_n)}$ on the unit interval that converges to a Poisson process with intensity measure $\theta^2\tau$. The proof is identical to the case $N_{k_n}^{(u_n)}$ since (14) and $D'(v_n)$ hold with $v_n \geq u_n$.

These results provide the setting to define the extremal index as the ratio of the limiting expected value of the point processes $N_{k_n}^{(u_n)}$ and $N_{k_n}^{(v_n)}$,

$$\theta = \lim_{n\to\infty} \frac{E[N_{k_n}^{(v_n)}]}{E[N_{k_n}^{(u_n)}]}. \tag{15}$$

The extremal index can also be interpreted as the conditional excess probability of u_n. From the results given in (9) and (13),

$$\theta = 1 - \lim_{n\to\infty} F_{u_n}(v_n), \tag{16}$$

with $F_{u_n}(v_n) = \frac{F(v_n)-F(u_n)}{1-F(u_n)}$. It is clear that as the dependence in the extremes (exceedances of u_n) of the stationary sequence decreases, v_n approaches u_n and θ gets closer to one as for the *iid* case or for weak dependence ($D'(u_n)$ and $D''(u_n)$ hold).

These definitions of the extremal index are also valid for threshold sequences where (9) does not hold but the mixing condition in (2) still does. Consider $\widetilde{u}_n$ such that $n(1 - F(\widetilde{u}_n)) = \tau_n$, with $\tau_n \to \infty$, and $\tau_n = o(n)$. This condition implies that $P\{M_n \leq \widetilde{u}_n\} \to 0$.

A necessary condition for $\widetilde{u}_n$ in order to define the extremal index in the same way as in (15) is that the ratio $\frac{-logP\{M_n\leq\widetilde{u}_n\}}{n(1-F(\widetilde{u}_n))}$ converges to a constant in $(0, 1)$. If the sequence $\{X_n\}$ has extremal index θ conditions (9) and (11) are satisfied for certain sequence u_n. Then, a sufficient condition for $\widetilde{u}_n$ is that

$$\frac{(1 - F(\widetilde{u}_n))(1 - F_{1,\ldots,r_n}(u_n))}{(1 - F(u_n))(1 - F_{1,\ldots,r_n}(\widetilde{u}_n))} \to 1. \tag{17}$$

This condition entails this, $k_n(1 - F_{1,\ldots,r_n}(\widetilde{u}_n)) = \tau'_n$ with $\tau'_n \to \infty$ and $\tau'_n/\tau_n \to \theta$. The same results that for u_n and τ constant are achieved now for $\widetilde{u}_n$ and τ_n. Therefore, the sequence $B_{r_n}(\widetilde{u}_n)$ satisfies that

$$E\left[B_{r_n}^{(\widetilde{u}_n)} | B_{r_n}^{(\widetilde{u}_n)} \geq 1\right] \to \theta^{-1},$$

and there exists a sequence $\widetilde{v}_n$ such that $n(1 - F(\widetilde{v}_n)) = \tau'_n$. Under condition (17) for $\{\widetilde{v}_n\}$ instead of $\{\widetilde{u}_n\}$ we obtain that $k_n(1 - F_{1,\ldots,r_n}(\widetilde{v}_n)) = \tau''_n$, with $\tau''_n \to \infty$ and $\tau''_n/\tau'_n \to \theta$, and the extremal index may be defined as in (15) for the corresponding $\widetilde{u}_n$ and $\widetilde{v}_n$ given that $D'(\widetilde{u}_n)$ holds.

For estimation purposes we will refer to the number of elements of the processes $N_{k_n}^{(\widetilde{u}_n)}$ and $N_{k_n}^{(\widetilde{v}_n)}$ as $Z^*_{\widetilde{u}_n}$ and $Z^*_{\widetilde{v}_n}$ respectively, and $Z_{\widetilde{u}_n}$ and $Z_{\widetilde{v}_n}$ will be used to denote the number of exceedances of $\widetilde{u}_n$ and $\widetilde{v}_n$ by the sequence $\{X_n\}$. Analog notation will be used for the corresponding exceedances of u_n and v_n. Note the variables $Z^*_{\widetilde{u}_n}$ and $Z^*_{u_n}$ can be interpreted

as the number of blocks of the partition defined by $\{k_n\}$, $\{r_n\}$ where there is at least one exceedance of $\widetilde{u}_n$ and u_n respectively.

3 Estimation of the extremal index

The extremal index represents the clustering of the largest observations determined by a sequence $\{u_n\}$ sufficiently high to satisfy a condition of type (9). The serial dependence in these observations has an effect on the distribution of the maximum of the stationary sequence, that is, $P\{M_n \leq u_n\}$ is $F^{n\theta}(u_n)$ instead of $F^n(u_n)$ for n and u_n sufficiently large.

This result leads to the first estimator of the extremal index for appropriate sequences k_n, r_n satisfying that $P^{k_n}\{M_{r_n} \leq u_n\}$ approximates $P\{M_n \leq u_n\}$. Then, by taking logs in both expressions, $\theta = \frac{logP\{M_{r_n} \leq u_n\}}{r_n log F(u_n)}$. A natural estimator for the extremal index is in this case,

$$\hat{\theta}_n^{(1)} = \frac{log(1 - Z^*_{u_n}/k_n)}{r_n log(1 - Z_{u_n}/n)}, \tag{18}$$

with the notation introduced in the last section. The ratio Z_{u_n}/n is an estimator of $1-F(u_n)$, and $Z^*_{u_n}/k_n$ an estimator of $1 - F_{1,\ldots,r_n}(u_n)$.

On the other hand the concept of extremal index introduced by Leadbetter (1983), θ^{-1} the limiting mean cluster size of the exceedances, yields this estimator

$$\hat{\theta}_n^{(2)} = \frac{Z^*_{u_n}}{Z_{u_n}}. \tag{19}$$

This method is called the blocks method and may be considered a simplified version of $\hat{\theta}_n^{(1)}$. Another popular method is the runs estimator, that may be seen as the estimator of the extremal index for the definitions introduced in O'Brien (1987) or in Hsing (1993),

$$\overline{\theta}_n = \frac{W_{u_n}}{Z_{u_n}}, \tag{20}$$

where $W_{u_n} = \sum_{i=1}^{n-r_n} I(X_i > u_n)(1 - I(X_{i+1} > u_n)) \cdot \cdot (1 - I(X_{i+r_n} > u_n))$.

Our definition of the extremal index yields an appealing estimator of θ given by the ratio of $Z^*_{v_n}$ and $Z^*_{u_n}$ or alternatively $Z^*_{\widetilde{v}_n}$ and $Z^*_{\widetilde{u}_n}$. For u_n and v_n sequences satisfying (9) and (13) our estimator $\widetilde{\theta}_n$ is given by

$$\widetilde{\theta}_n = \frac{Z^*_{v_n}}{Z^*_{u_n}}, \tag{21}$$

representing the corresponding thinnings defined by the sequence k_n and the thresholds u_n and v_n. The estimator, however, is not fully specified since these sequences are not determined. By (9) an appropriate candidate for this threshold sequence is given by extreme order statistics (see section 2.5. in Leadbetter et al. (1983)). In turn an adequate choice of v_n is given by the order statistic of the stationary sequence $\{X_n\}$ satisfying the empirical counterpart of (12), *i.e.*

$$v_n = \max_{1 \leq i \leq n} \left\{ x_i, i = 1, \ldots, n | \frac{1}{Z^*_{u_n}} \sum_{j=1}^{k_n} B^{(x_i)}_{r_n,j} = 1 \right\}, \tag{22}$$

with $B^{(u_n)}_{r_n,j} = \sum_{k=(j-1)r_n+1}^{jr_n} I(X_k > u_n)$. This expression boils down to $v_n = x_{(n-Z^*_{u_n})}$, extreme order statistic, with $x_{(1)} \leq \ldots \leq x_{(n)}$ the sequence of order statistics. By (17) the corresponding expressions apply to $\widetilde{u}_n$ and $\widetilde{v}_n$ being intermediate order statistics.

If the threshold u_n is estimated by an extreme order statistic the point process $N^{(u_n)}_{k_n}$

converges to a Poisson process, and its variance in consequence converges to a constant. This is a serious inconvenient for the consistency of the majority of the estimators of θ that is overcome in our setup by using $\widetilde{u}_n$ (intermediate order statistic).

3.1 Statistical properties of the different estimators

Consider first the case of $\widetilde{\theta}_n$ as the quotient of the random variables $Z^*_{v_n}$ and $Z^*_{u_n}$ where v_n and u_n satisfy (13) and (9) respectively, that is,

$$\widetilde{\theta}_n = \tfrac{Z^*_{v_n}}{Z^*_{u_n}}.$$

By the second order Taylor expansion of $E[Z^*_{v_n}/Z^*_{u_n}]$ about the respective expected values (delta method) we have that

$$E[\widetilde{\theta}_n] = \frac{E[Z^*_{v_n}]}{E[Z^*_{u_n}]}\left(1 + \frac{V[Z^*_{u_n}]}{E[Z^*_{u_n}]^2} - \frac{Cov[Z^*_{v_n}, Z^*_{u_n}]}{E[Z^*_{u_n}]E[Z^*_{v_n}]}\right) + O(\frac{1}{\tau^2}). \tag{23}$$

The different contributions to $Z^*_{u_n}$ are not mutually independent. In particular, $E[Z^{*2}_{u_n}] = k_n P\{M_1 > u_n\} + \sum_{i=1}^{k_n}\sum_{i\neq j}^{k_n} P\{M_i > u_n, M_j > u_n\}$, where M_i is used to denote the maximum of $\{X_{(i-1)r_n+1}, \ldots, X_{ir_n}\}$. By stationarity the variance can be expressed as $V[Z^*_{u_n}] = E[Z^*_{u_n}] + k_n^2 P\{M_1 > u_n, M_2 > u_n\} - E^2[Z^*_{u_n}] - k_n P\{M_1 > u_n, M_2 > u_n\}$. Under $D'(u_n)$ the difference between $k_n^2 P\{M_1 > u_n, M_2 > u_n\}$ and $E^2[Z^*_{u_n}]$ converges to 0 as n increases, and $k_n P\{M_1 > u_n, M_2 > u_n\}$ is well approximated by $E[Z^*_{u_n}]P\{M_1 > u_n\}$ that in turn also converges to 0. The covariance takes a similar expression, $Cov[Z^*_{u_n}, Z^*_{v_n}] = E[Z^*_{v_n}] + k_n^2 P\{M_1 > u_n, M_2 > v_n\} - E[Z^*_{u_n}]E[Z^*_{v_n}] - k_n P\{M_1 > u_n, M_2 > v_n\}$ that boils down to $Cov[Z^*_{u_n}, Z^*_{v_n}] = E[Z^*_{v_n}]$.

Therefore expression (23) for n sufficiently high is as follows

$$E[\widetilde{\theta}_n] = \frac{E[Z^*_{v_n}]}{E[Z^*_{u_n}]}\left(1 + \frac{E[Z^*_{u_n}]}{E[Z^*_{u_n}]^2} - \frac{E[Z^*_{v_n}]}{E[Z^*_{u_n}]E[Z^*_{v_n}]}\right) + O(\frac{1}{\tau^2}),$$

and it is immediate to see that the expected value of our estimator takes this expression,

$$E[\widetilde{\theta}_n] = \theta + O(\frac{1}{\tau^2}). \tag{24}$$

For the analysis of the variance it is useful to derive the conditional moments. Consider $Z^*_{u_n} = z^*_{u_n}$ known, and note that the sequences u_n and v_n are related by this expression,

$$1 - F_{1,\ldots,r_n}(v_n) = (1 - F_{1,\ldots,r_n}(u_n))\left(1 - \frac{F_{1,\ldots,r_n}(v_n) - F_{1,\ldots,r_n}(u_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right). \tag{25}$$

Then by (11), $E[\widetilde{\theta}_n | Z^*_{u_n} = z^*_{u_n}] = 1 - \frac{F_{1,\ldots,r_n}(v_n) - F_{1,\ldots,r_n}(u_n)}{1 - F_{1,\ldots,r_n}(u_n)}$, and the conditional variance takes this form

$$V[\widetilde{\theta}_n | Z^*_{u_n} = z^*_{u_n}] = \frac{1}{z^*_{u_n}}\left(1 - \frac{F_{1,\ldots,r_n}(v_n) - F_{1,\ldots,r_n}(u_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right). \tag{26}$$

By the law of iterated expectations the unconditional variance can be decomposed in two different terms, $V[\widetilde{\theta}_n] = V[E[\widetilde{\theta}_n | Z^*_{u_n}]] + E[V[\widetilde{\theta}_n | Z^*_{u_n}]]$. It is clear the first term is 0, and by the Taylor expansion of $E[1/Z^*_{u_n}]$ about $E[Z^*_{u_n}]$ we obtain that

$$E[V[\widetilde{\theta}_n | Z^*_{u_n} = z^*_{u_n}]] = \left(1 - \frac{F_{1,\ldots,r_n}(v_n) - F_{1,\ldots,r_n}(u_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right)\left(\frac{1}{E[Z^*_{u_n}]} + \frac{V[Z^*_{u_n}]}{E^3[Z^*_{u_n}]}\right). \tag{27}$$

In consequence,

$$V[\widetilde{\theta}_n] = \left(1 - \frac{F_{1,\ldots,r_n}(v_n) - F_{1,\ldots,r_n}(u_n)}{1 - F_{1,\ldots,r_n}(u_n)}\right)\left(\frac{1}{\theta\tau} + O\left(\frac{1}{\tau^2}\right)\right) = O\left(\frac{1}{\tau}\right). \tag{28}$$

Therefore the mean square error (MSE) of our estimator is of order $O(\frac{1}{\tau})$ with τ constant. This result implies that this estimator is not consistent for u_n and v_n defined by extreme order statistics. The consistency, however, will be achieved when these sequences are replaced by $\widetilde{u}_n$ and $\widetilde{v}_n$ intermediate order statistics as it is shown in the following section.

Our estimator may be interpreted as a refinement of the standard blocks method $\hat{\theta}_n^{(2)}$ by writing $\widetilde{\theta}_n = \frac{Z^*_{v_n}/Z_{u_n}}{\hat{\theta}_n^{(2)}}$. The asymptotic properties of the latter estimator $\hat{\theta}_n^{(2)}$ are derived in Hsing (1991) or in Smith and Weissman (1994). By means of the delta method they find that $E[\hat{\theta}_n^{(2)}] = \theta + O(\frac{1}{\tau})$, and the variance is $V[\hat{\theta}_n^{(2)}] = O(\frac{1}{\tau})$. Therefore the bias of this estimator is higher than the bias of $\widetilde{\theta}_n$, but the mean square error (MSE) of both estimators is $O(1/\tau)$.

For the logs method,

$$E[\hat{\theta}_n^{(1)}] = \frac{E[Z^*_{u_n}]}{E[Z_{u_n}]}\left(1 + \frac{E[Z^*_{u_n}]}{2k_n} + \frac{E^2[Z^*_{u_n}]}{6k_n^2}\right) = \theta + O\left(\frac{\tau}{k_n}\right), \text{ and } V[\hat{\theta}_n^{(1)}] = O(\tfrac{1}{\tau}).$$

This estimator is asymptotically unbiased, but it is not consistent either for τ constant.

3.2 Inference for the Extremal Index

Consider now the sequences $\widetilde{u}_n$ and $\widetilde{v}_n$ defined by the conditions $\tau'_n = k_n(1 - F_{1,\ldots,r_n}(\widetilde{u}_n))$, $\tau''_n = k_n(1 - F_{1,\ldots,r_n}(\widetilde{v}_n))$, with $\tau'_n \to \infty$, $\tau''_n \to \infty$ and $\tau''_n/\tau'_n \to \theta$. In this case the first two moments of the random variables $Z^*_{\widetilde{u}_n}$ and $Z^*_{\widetilde{v}_n}$ diverge to infinity. By stationarity the variance is given by this expression,

$$V[Z^*_{\widetilde{u}_n}] = E[Z^*_{\widetilde{u}_n}] + \left(k_n^2 P\{M_1 > \widetilde{u}_n, M_2 > \widetilde{u}_n\} - E^2[Z^*_{\widetilde{u}_n}]\right) - k_n P\{M_1 > \widetilde{u}_n, M_2 > \widetilde{u}_n\}.$$

Note that in this case, under $D'(u_n)$ for n sufficiently high, the variance is

$$V[Z^*_{\widetilde{u}_n}] = E[Z^*_{\widetilde{u}_n}] - E[Z^*_{\widetilde{u}_n}]P\{M_1 > \widetilde{u}_n\}. \tag{29}$$

The covariance in turn takes this expression,

$$Cov[Z^*_{\widetilde{u}_n}, Z^*_{\widetilde{v}_n}] = E[Z^*_{\widetilde{v}_n}] - E[Z^*_{\widetilde{v}_n}]P\{M_1 > \widetilde{u}_n\}.$$

Therefore expression (23) is as follows

$$E[\widetilde{\theta}_n] = \frac{\tau''_n}{\tau'_n}\left(1 + \frac{\tau'_n P\{M_1 \le \widetilde{u}_n\}}{(\tau'_n)^2} - \frac{\tau''_n P\{M_1 \le \widetilde{u}_n\}}{\tau'_n \tau''_n}\right) + o\left(\frac{1}{\tau_n}\right), \tag{30}$$

that boils down to $E[\widetilde{\theta}_n] = \theta + o\left(\frac{1}{\tau_n}\right)$, by the definition of τ'_n and τ''_n.

This estimator of θ is now asymptotically unbiased, and for $\tau_n < k_n$, $\tau_n^2 > k_n$, $\widetilde{\theta}_n$ outperforms $\hat{\theta}_n^{(1)}$ in this sense. For $\tau_n > k_n$ this result is trivial.

In order to find the unconditional variance in this case, we calculate first the conditional moment.

$$V[\widetilde{\theta}_n | Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}] = \frac{1}{z^{*2}_{\widetilde{u}_n}} V[Z^*_{\widetilde{v}_n} | Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}]. \tag{31}$$

Applying (29) to the random variable $Z^*_{\widetilde{v}_n}|Z^*_{\widetilde{u}_n}$,

$$V[Z^*_{\widetilde{v}_n} | Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}] = E[Z^*_{\widetilde{v}_n} | Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}]P\{M_1 \le \widetilde{v}_n | M_1 > \widetilde{u}_n\},$$

that amounts to

$$V[Z^*_{\widetilde{v}_n} | Z^*_{\widetilde{u}_n} = z^*_{\widetilde{u}_n}] = z^*_{\widetilde{u}_n} P\{M_1 > \widetilde{v}_n | M_1 > \widetilde{u}_n\} P\{M_1 \le \widetilde{v}_n | M_1 > \widetilde{u}_n\}. \tag{32}$$

Then, in the same way as in (27),

$$V[\widetilde{\theta}_n] = \left(1 - \tfrac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)}\right)\left(\tfrac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)}\right)\left(\tfrac{1}{E[Z^*_{\widetilde{u}_n}]} + \tfrac{V[Z^*_{\widetilde{u}_n}]}{E^3[Z^*_{\widetilde{u}_n}]}\right),$$

that in turn is

$$V[\widetilde{\theta}_n] = \left(1 - \frac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)}\right)\left(\frac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)}\right)\frac{1}{\tau'_n} + o\left(\frac{1}{\tau_n}\right). \quad (33)$$

Under $D'(\widetilde{u}_n)$ the distribution of $Z^*_{\widetilde{v}_n}|Z^*_{\widetilde{u}_n}$ is well approximated ($\sim$) by a binomial distribution with parameters $bin\left(Z^*_{\widetilde{u}_n}, 1 - \frac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)}\right)$, and $Z^*_{\widetilde{u}_n}$ by a $bin\,(k_n, 1 - F_{1,\ldots,r_n}(\widetilde{u}_n))$. Then, the distribution of $\widetilde{\theta}_n$ can be approximated by a normal distribution with parameters given in (30) and (33).

On the other hand the relation between the tails introduced in (25) holds for $\widetilde{u}_n$ and $\widetilde{v}_n$, and by assumption $\tau''_n/\tau'_n \to \theta$, yielding that $1 - \frac{F_{1,\ldots,r_n}(\widetilde{v}_n) - F_{1,\ldots,r_n}(\widetilde{u}_n)}{1 - F_{1,\ldots,r_n}(\widetilde{u}_n)} \to \theta$. In turn, the distribution of $\widetilde{\theta}_n$ is approximated by

$$\widetilde{\theta}_n \overset{w}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{\tau'_n}\right). \quad (34)$$

By the structure of dependence $\frac{\tau'_n}{\tau_n} \to \theta$ as n goes to infinity, and hence $\widetilde{\theta}_n \overset{w}{\sim} N\left(\theta, \frac{1-\theta}{\tau_n}\right)$ results a valid approximation for the distribution of $\widetilde{\theta}_n$. More formally, we can obtain a test statistic that is asymptotically parameter free,

$$T_n = \frac{\widetilde{\theta}_n - \theta}{\sqrt{1-\theta}}\sqrt{\tau_n} \overset{w}{\to} N(0,1). \quad (35)$$

The asymptotic confidence intervals for θ are easily calculated from the former expression.

$$\theta \in \left[\widetilde{\theta}_n \pm z_{1-\alpha/2}\sqrt{\frac{1-\widetilde{\theta}_n}{\tau_n}}\right], \quad (36)$$

with $z_{1-\alpha/2}$ the quantile of the standard normal distribution. This interval is an approximation of the true confidence interval for finite samples. The exact confidence region for small sample sizes may be better approximated by resampling techniques. The confidence interval takes this expression

$$\theta \in \left[\widetilde{\theta}_n - \sqrt{\frac{1-\widetilde{\theta}_n}{\tau_n}}J_n^{-1}(1-\frac{\alpha}{2}, F), \widetilde{\theta}_n - \sqrt{\frac{1-\widetilde{\theta}_n}{\tau_n}}J_n^{-1}(\frac{\alpha}{2}, F)\right], \quad (37)$$

where $J_n^{-1}(1-\alpha, F)$ is the $1-\alpha$ quantile of the sampling distribution $J_n(F)$ of the statistic T_n. In practice this quantile is approximated by the order statistic $T_{n,((1-\alpha)B)}$ of the sample $T_{n,1}, \ldots, T_{n,B}$ with B the number of iterations. The notation F in the distribution $J_n(F)$ refers to Monte Carlo simulation, that is, the generating process of the data is known. Otherwise $J_n(F)$ must be approximated by $J_b(F)$ with $b < n$, $b/n \to 0$ (subsampling), or by $J_n(F^*)$ with F^* representing blocks bootstrap methods. The naïve bootstrap does not work in this context due to the serial dependence in the data. Nevertheless, as it is seen in the simulations, the gaussian asymptotic intervals give reliable approximations of the exact confidence regions for moderate sample sizes.

Our interest however lies on testing hypotheses of the type $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$. This one-sided hypothesis test permits to assess the mixing condition $D''(u_n)$ introduced in (3) by imposing $\theta_0 = 1$. In other words, if $\theta = 1$ there are no clusters of extreme values (exceedances of some threshold) in the stationary sequence. The null hypothesis amounts

to see if θ_0 is contained in the interval $\left(-\infty, \widetilde{\theta}_n + z_{1-\alpha}\sqrt{\frac{1-\widetilde{\theta}_n}{\tau_n}}\right]$ or alternatively in $\left(-\infty, \widetilde{\theta}_n - \sqrt{\frac{1-\widetilde{\theta}_n}{\tau_n}} J_n^{-1}(\alpha, F)\right]$, the bootstrap approximation.

Consider the example due to Chernick (1981) for $\{X_n\}$ a strictly stationary first order autoregressive sequence driven by the model $X_i = \frac{1}{r}X_{i-1} + \varepsilon_i$, where $r \geq 2$ is an integer, ε_i are discrete uniforms on $\{0, 1/r, \ldots, (r-1)/r\}$, being independent of X_{i-1}, and X_i having a uniform distribution on $[0, 1]$. The extremal index is given by $\theta = \frac{r-1}{r}$. The plot given in figure 6.1 describes the curve of the estimates of θ for different partitions and different sample sizes. The upper panel is for $n = 200$ and the lower panel considers $n = 1000$.

(INSERT FIGURE 6.1)

Two conclusions stem from these plots. First, it is clear that condition $D''(\hat{u}_n)$ is rejected with $\alpha = 0.05$, and second, the confidence intervals for θ are smaller as n increases. This is caused by the choice of an increasing order statistic, $\hat{u}_n = x_{(n-k)}$ with $k = \sqrt{2n}$ as threshold.

3.3 Some comments on the block size selection

The partition of the sequence $\{X_n\}$ in k_n blocks of size r_n has two main features: First, it defines a point process $N_{k_n}^{(u_n)}$ that converges to a serially independent process, and second, the distribution of this sequence converges to a Poisson process as k_n goes to infinity. The majority of the estimators for the extremal index found in the literature are tied to that partition. This dependence turns explicit for example for the logs method where k_n appears in the expression for $\hat{\theta}_n^{(1)}$, as well as in its expected value.

If the observations are independent or weak dependent, $N_{k_n}^{(u_n)}$ with $k_n = n$ defines itself an *iid* point process and $\theta = 1$. Otherwise the extremal index is less than 1 and the partition $k_n < n$ plays a central role in the estimation of the extremal index.

Provided n, an adequate choice of the sequence k_n along with a suitable threshold u_n define a sequence given by the maxima over the corresponding blocks of observations. Testing condition $D'(u_n)$ in practice is replaced by testing for serial independence in this sequence of maxima. Hypothesis tests for the latter condition require large sample sizes and most of them rely on gaussian assumptions. A naïve alternative to these methods is dropping the first partitions defined by $k_n = n, n-1, \ldots$, that have a high likelihood of entailing serial dependence in $N_{k_n}^{(u_n)}$, and analyzing the performance of the estimators by the stability of the corresponding estimates along the different partitions.

The influence of the choice of k_n on the estimates of the extremal index also depends on the choice of the threshold u_n. For example in the blocks method the estimates of θ are driven by the corresponding partition for low u_n. In this case $Z_{u_n}^*$ and k_n take the same values resulting in a sequence of estimates that approaches 0 when k_n decreases.

A similar situation occurs for the logs method with u_n a low threshold. The numerator in this case collapses to $-\infty$ and the estimator is not defined for the corresponding partition. These effects are not present in $\widetilde{\theta}_n$ because a larger ratio $Z_{u_n}^*/k_n$ given by a low threshold is compensated by a large value of $Z_{v_n}^*/k_n$, that is, the sequence v_n varies according to u_n.

4 Simulations: Some examples

We now consider some examples from the literature showing short range dependence ($0 < \theta < 1$) reflected by a distribution of the maximum satisfying (5).

The following example is the doubly stochastic model studied by Smith and Weissman (1994). Let $\{\xi_i, i \geq 1\}$ be *iid* with distribution function F, and suppose that $Y_1 = \xi_1$, and for

$i > 1$, $Y_i = Y_{i-1}$ with probability ψ, and $Y_i = \xi_i$ with probability $1-\psi$. The doubly stochastic sequence $\{X_i\}$ is defined by $X_i = Y_i$ with probability η, and $X_i = 0$ with probability $1-\eta$, with these different events mutually independent. The extremal index is

$$\theta = \tfrac{1-\psi}{1-\psi+\psi\eta}.$$

The following pictures represent the paths of the different estimators for the extremal index. The threshold sequence is estimated by an order statistic: $\hat{u}_n = x_{(n-k)}$. We have implemented two different types of order statistics for samples of size $n = 200$ and $n = 1000$ observations. An extreme order statistic ($k = 20$ fixed), and an intermediate order statistic ($k = \sqrt{2n}$). We only present the estimates of θ for the intermediate order statistic since the other threshold sequence provides similar results for these sample sizes in this example.

The curves describe the sample means of the different estimates of the extremal index and for different partitions of the sample for $m = 100$ simulated sequences generated from the model introduced in Smith and Weissman with $\psi = 0.9$ and $\eta = 0.7$. Suppose also F, a Fréchet distribution $F(x) = \exp(-x^{-\alpha})$ with $\alpha = 1$ and $x \in (0, \infty)$.

(INSERT FIGURE 6.2)

The confidence intervals for $\widetilde{\theta}_n$ derived in the last section are not plotted. Instead we have represented the simulated standard deviation of the different estimators for $m = 100$ in order to present a fair comparison between the three competitors. The standard deviation for the different partitions is estimated via Monte Carlo simulation by $\hat{\sigma}_{k_n}$ with

$$\hat{\sigma}^2_{k_n} = \tfrac{1}{m-1} \sum_{i=1}^{m} (\theta_{i,est} - \overline{\theta}_{est})^2,$$

and $\overline{\theta}_{est}$ the sample mean of the different estimates.

Apparently the blocks method is the best method. After the first partitions of the sample the curve of the estimates of θ remains stable very close to the target line for the three methods. Nevertheless, the blocks method estimator has smaller variance. In addition, focusing on figure 6.3 it is clear that the different estimators of θ analyzed in this example are consistent and the blocks method is more efficient. The mean square error is estimated from the simulated sequences generated for figure 6.2, and takes this expression,

$$MSE(\theta_{est}) = \tfrac{1}{m} \sum_{i=1}^{m} (\theta_{i,est} - \theta)^2.$$

These results agree with the conclusions found in Smith and Weissman (1994).

(INSERT FIGURE 6.3)

However the impressive performance of the blocks method may be due to the low value of the extremal index ($\theta = 0.137$) and the choice of a low threshold estimate. Under these circumstances, the curve of estimates of θ by the blocks method is decaying as k_n decreases (r_n increases) and approaches the true parameter. To get an insight into this, we also study a doubly stochastic process where the extremal index is significantly higher. Suppose $\psi = 0.5$ and $\eta = 0.5$, *i.e.* $\theta = 0.66$. The following plots are the analogs of figures 6.2 and 6.3.

(INSERT FIGURE 6.4)

(INSERT FIGURE 6.5)

The blocks method in this example does not work. The number of blocks with an exceedance of $\hat{u}_n$ ($Z^*_{\hat{u}_n}$) is similar to k_n for each partition. Therefore the estimator decreases as k_n decreases since $Z_{\hat{u}_n}$ remains constant. On the other hand the logs method improves as n increases and the mean square error of $\widetilde{\theta}_n$ and $\hat{\theta}^{(1)}_n$ are negligible for $n = 1000$.

Finally the exact and asymptotic confidence intervals for θ are displayed to assess the estimates given by $\widetilde{\theta}_n$.

(INSERT FIGURE 6.6)

5 Clustering in Financial Series: The Case of DaX Index

Financial returns are characterized by a series of stylized facts: leverage effect (after periods of high volatility the likelihood of losses is higher than in calm periods), heavy tails, clustering of the largest observations and some skewness towards the losses tail. The seminal paper of Engle and Bollerslev (1986) proposed the popular GARCH models, Generalized Auto-Regressive Conditional Heteroscedastic volatility models, to explain these features of the data. In general, the Garch(1,1) is sufficient to model most of the financial returns. It takes this expression,

$$X_i = \epsilon_i \sigma_i, \qquad \sigma_i^2 = \omega + \alpha X_{i-1}^2 + \beta \sigma_{i-1}^2,$$

with ω, α, $\beta > 0$, and $\alpha + \beta < 1$, that can be interpreted as an ARMA(1,1) model for the squares,

$$X_i^2 = \omega + (\alpha + \beta) X_{i-1}^2 + \nu_i - \beta \nu_{i-1},$$

with $\nu_i = \sigma_i^2(\varepsilon_i^2 - 1)$.

According to this model, the dependence found in the financial returns is driven by the second moments. The literature concerning this topic is enormous; up to the extent that there exist different GARCH type models to explain particular characteristics of the financial series.

We propose to analyze some of these stylized facts, in particular the clustering of the largest observations, by means of the extremal index. A value of θ significatively less than 1 shows certain short range dependence reflected in the clustering of the largest observations. This may be interpreted as a pattern in the occurrence of the extreme values, that is, once a large loss in the asset return has occurred we can expect a period of large losses (values exceeding some threshold). The average length of this period is the inverse of the extremal index.

The data we use to illustrate this methodology consists on the analysis of the Frankfurt financial market (DaX Index) over the period 19/12/1994 − 20/04/2001. These data have been collected from *www.freelunch.com*. The observations considered for the analysis are the logarithmic returns measured in percentage terms and denoted as r_t:

$$r_t = 100\ (logP_t - logP_{t-1}),$$

with P_t the original prices at time t.

(INSERT TABLE 6.7)

The analysis of the extremal index for both tails shows certain clustering in the occurrence of the positive and negative extreme values. The confidence intervals derived from $\widetilde{\theta}_n$ do not contain $\theta = 1$ for $\alpha = 0.05$ (figure 6.8).

(INSERT FIGURE 6.8)

These pictures also depict a higher level of clustering for the largest negative returns than for the positive values. This fact can be statistically tested by means of a confidence interval for the difference of the extremal indexes corresponding to the positive and negative tail. This confidence interval takes this expression

$$\theta_{pos} - \theta_{neg} \in \left[\widetilde{\theta}_{n,pos} - \widetilde{\theta}_{n,neg} \pm z_{1-\alpha/2} \sqrt{\frac{1-\widetilde{\theta}_{n,pos}}{\tau_{n,pos}} + \frac{1-\widetilde{\theta}_{n,neg}}{\tau_{n,neg}}} \right]. \tag{38}$$

It is important to mention that $\widetilde{\theta}_{n,pos}$ and $\widetilde{\theta}_{n,neg}$ are considered independent. This can lead to obtain smaller confidence intervals given α compared to considering dependent estimators with positive correlation. For some partitions of the sample it is statistically significant that the clustering for the positive extreme values is smaller than for the largest negative returns (figure 6.9).

(INSERT FIGURE 6.9)

The analysis of the clustering of the largest values for the sequence of the volatility of the returns deserves some interesting comments. The confidence interval introduced in (38) may be applied to test the difference between the extremal index of the volatility sequence θ_{vol} and θ_{pos} or θ_{neg} (figure 6.10). The results derived from both tests, $\theta_{pos} - \theta_{vol}$ and $\theta_{neg} - \theta_{vol}$, point out that the extreme values of the volatility sequence are driven by the negative extreme values. Therefore these observations are bigger in absolute value than the largest positive returns. This fact explains the negative skewness of the returns sequence.

Finally it is worth mentioning the stylized fact of heavy tails. By Berman's condition (Berman, 1964), if $\{r_t\}$ is a standard normal sequence and $Cov(r_t, r_{t-j}) log\ j \to 0$ as $j \to \infty$, the extremal index of the sequence is $\theta = 1$. In practice, the autocorrelation function of the returns of a financial series is usually close to zero, also in this case and then the second part of Berman's condition holds. Therefore, if $\theta < 1$ the sequence of the returns of the DaX Index is not normally distributed but heavy tailed. This suggests that the existence of clustering of the extreme values in a financial series implies that the distribution of the observations is heavy tailed. Hence it is not sufficient with the second moments of $\{r_t\}$ to know the structure of dependence of the sequence. Moreover, the dependence in the extremes plays an important role and this dependence stems from the heavy tails.

(INSERT FIGURE 6.10)

6 Conclusion

The aim of this paper has been to propose an estimator for the extremal index defined by the ratio of the number of exceedances of two threshold sequences. This estimator possesses two appealing properties: First, it is not necessary to choose a sequence $\{u_n\}$ satisfying the Poisson condition in the limit, and second it is not very sensitive to the block size selection.

Regarding the asymptotic properties of our estimator, we can conclude that are estimator has the same order of convergence than the standard methods (the respective variances are of the same order). However, under very general conditions our estimator is asymptotically unbiased outperforming the other two methods that are not free from a residual term. Our

estimator also works better than these methods in two manners: it is not so dependent of the corresponding partition of the sequence, and it relaxes the selection of the threshold sequence.

In addition, the absence of dependence on the Poisson condition permits to propose a hypothesis test for the extremal index. We find this test useful in different ways: it formally assesses the estimates of the extremal index, it introduces an innovative procedure for testing the existence of clustering in the occurrence of extreme events, and it may be useful to determine the skewness and kurtosis of the distribution of the data by testing the difference of extremal indexes between both tails.

Finally, the application of these methodologies to financial series (DaX Index) confirms the existence of short range dependence in the extreme observations; that is, some clustering of the extreme values of the positive and negative returns. The clustering is higher for the negative tail. By Berman's condition, the distribution of the observations is heavy tailed since θ is statistically less than 1. These results agree with the stylized facts found in most of the financial series.

References

Berman, S.M., (1964): Limit theorems for the maximum term in stationary sequences. Annals of Mathematical Statistics 35, 502-516.

Chernick, M.R., (1981): A limit theorem for the maximum of autorregresive processes with uniform marginal distribution, Annals of Probability 9, $145 - 149$.

Engle, R.F., Bollerslev, T., (1986): Modelling the persistence of conditional variances. Econometric Reviews 5, $1 - 50$ (with discussion).

Ferro, C.A., and Segers, J., (2003): Inference for clusters of extremes. Journal of the Royal Statistical Society B, 65, $545 - 556$.

Haan, L. de, (1976): Sample extremes: an elementary introduction. Statist. Neerlandica, 30, $161 - 172$.

Hsing, T., (1991): Estimating the parameters of rare events. Stochastic Processes and Applications, 37, $117 - 139$.

Hsing, T., (1993): Extremal Index Estimation for a Weakly Dependent Stationary Sequence. Annals of Statistics, 21, $2043 - 2071$.

Kallenberg, O., (1976): *Random Measures.* Akademie Verlag, Berlin and Academic Press, New York.

Leadbetter, M. R., (1983): Extremes and Local Dependence in Stationary Sequences, Z. Wahrscheinlichkeitstheorie verw. Gebiete 65, $291 - 306$.

Leadbetter, M. R., Lindgren, G., and Rootzén, H., (1983): *Extremes and Related Properties of Random Sequences and Processes.* Ed. Springer-Verlag, New York.

Loynes, R.M., (1965): Extreme Values in Uniformly Mixing Stationary Stochastic Processes. Annals of Mathematical Statistics, 36, $993 - 999$.

O'Brien, G.L., (1974):The maximum term of uniformly mixing stationary sequences.

Z. Wahrscheinlichkeitstheorie verw. Gebiete 30, 57 − 63.

O'Brien, G.L., (1987): Extreme Values for Stationary and Markov Sequences. Annals of Probability, 15, 281 − 291.

Rosenblatt, M., (1956): A central limit theorem and a strong mixing condition. Proc. Nat. Acad. Sci. USA, 42, 43 − 47.

Smith, R.L., Weissman, I., (1994): Estimating the Extremal index. Journal of the Royal Statistical Society B, 56, 515 − 528.

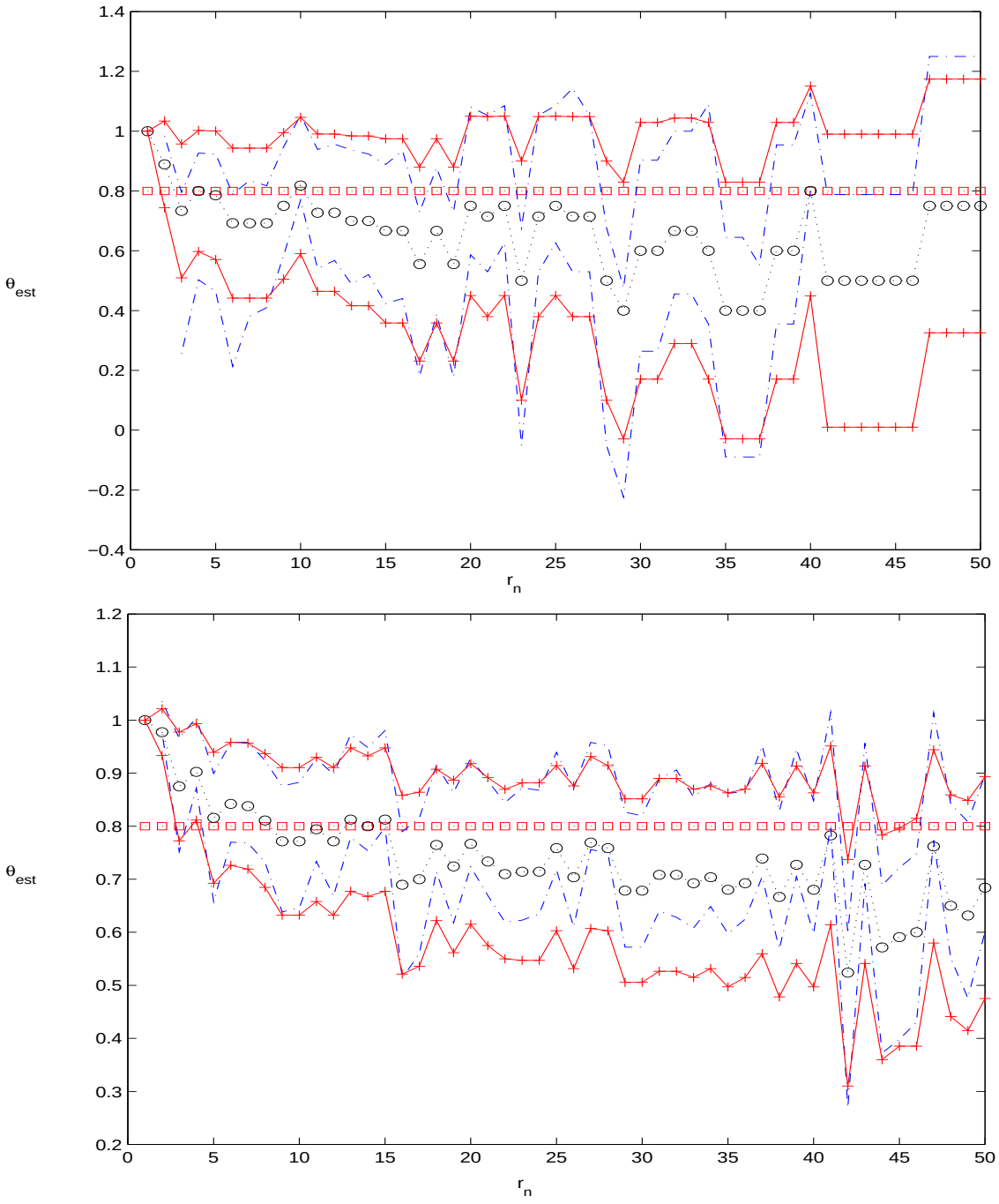

Figure 6.1. *Estimated values of the extremal index for the Chernick model with* $r = 5$. *The extremal index is* $\theta = 0.8$ *plotted by* $\square$ *line. The partitions* r_n *considered are in the range* $[1, 50]$. $\widetilde{\theta}_n$ *is represented by* $(\cdots)$ *and* o*; the dash line describes the bootstrap confidence interval with* $B = 1000$ *and* $(+-)$ *is employed for the asymptotic intervals. The significance level is* $\alpha = 0.05$. *The sample sizes are* $n = 200$ *and* $n = 1000$ *respectively. The threshold sequence is* $\hat{u}_n = x_{(n-k)}$ *with* $k = \sqrt{2n}$.

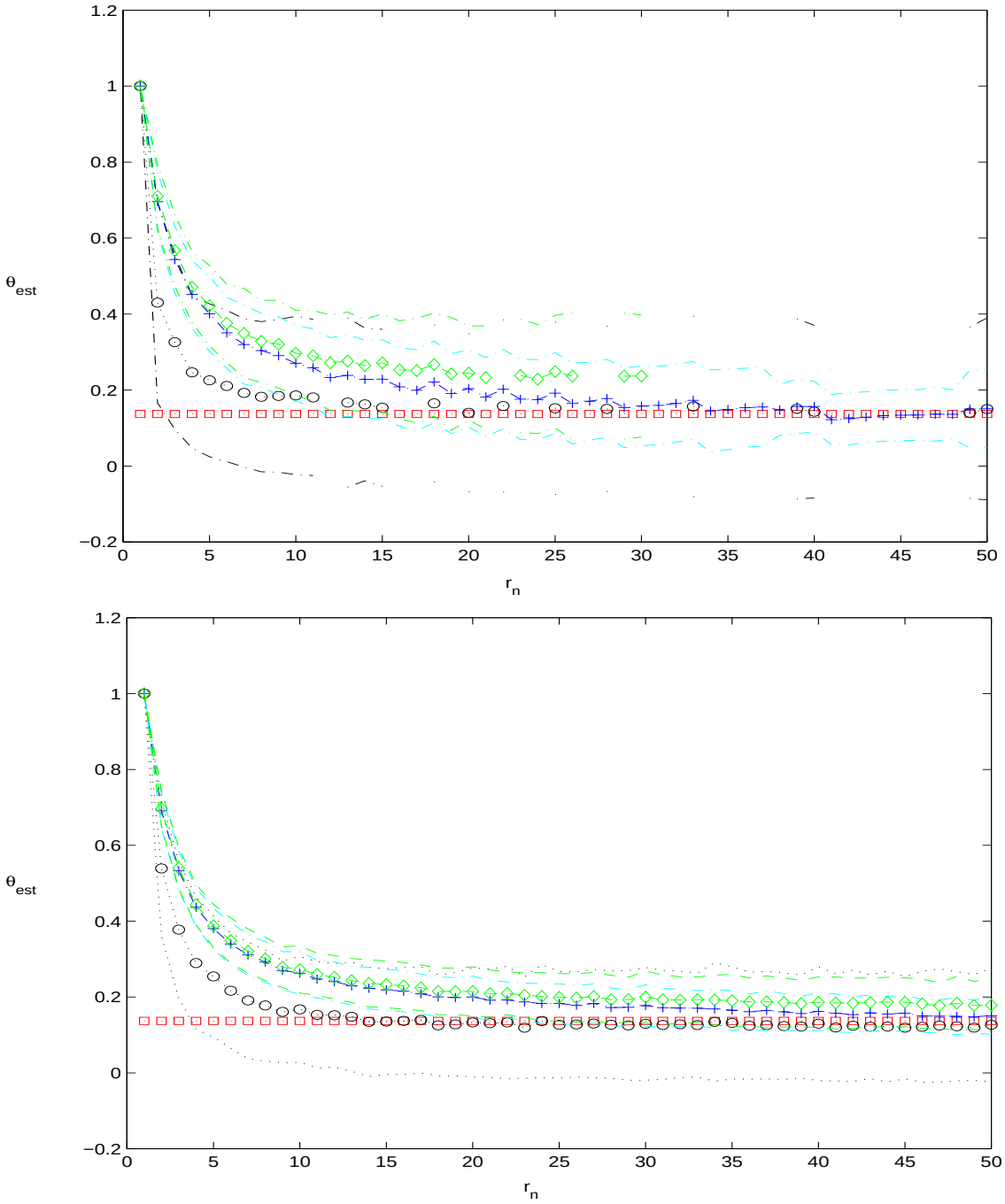

Figure 6.2. *Estimated values of the extremal index for the doubly stochastic model with* $\psi = 0.9$ *and* $\eta = 0.7$. *The extremal index is* $\theta = 0.137$ *plotted by* $\square$ *line. The partitions* r_n *considered are in the range* $[1, 50]$. $\widetilde{\theta}_n$ *is represented by* $(\cdots)$ *and* o*; the corresponding standard deviation is plotted with* $(\cdots)$. *The logs method* $\hat{\theta}_n^{(1)}$ *is represented with* $(- - -\!-)$ *and* $\diamond$. *The standard deviation with* $(- - -\!-)$. *The blocks method* $\hat{\theta}_n^{(2)}$ *with* $(\cdot - \cdot\!-)$ *and* $+$, *and* $(\cdot - \cdot\!-)$ *for the standard deviation. The sample sizes are* $n = 200$ *and* $n = 1000$ *respectively.* $m = 100$ *simulations are used. The threshold sequence is* $\hat{u}_n = x_{(n-k)}$ *with* $k = \sqrt{2n}$.

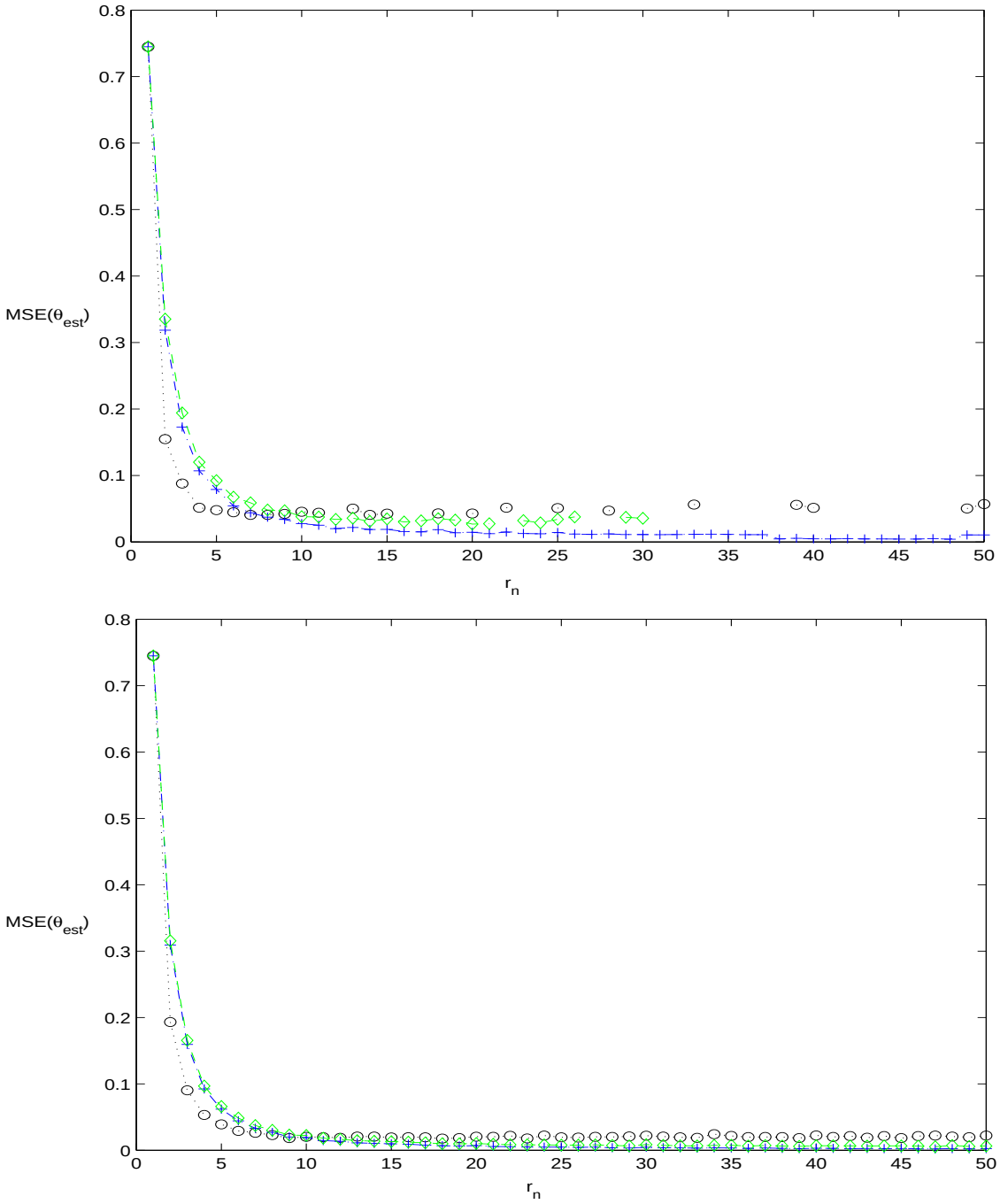

Figure 6.3. *Simulated mean square error (MSE) of the estimators of θ for the doubly stochastic model with $\psi = 0.9$ and $\eta = 0.7$. The partitions r_n considered are in the range $[1, 50]$. $m = 100$ simulations of the model are used. $\widetilde{\theta}_n$ is represented by $(\cdot\,\cdot)$ and o, $\hat{\theta}_n^{(1)}$ with $(- - -\!-)$ and $\diamond$, and $(\cdot - \cdot\!-)$ and $+$ for $\hat{\theta}_n^{(2)}$. The sample sizes are $n = 200$ and $n = 1000$ respectively. The threshold sequence is $\hat{u}_n = x_{(n-k)}$ with $k = \sqrt{2n}$.*

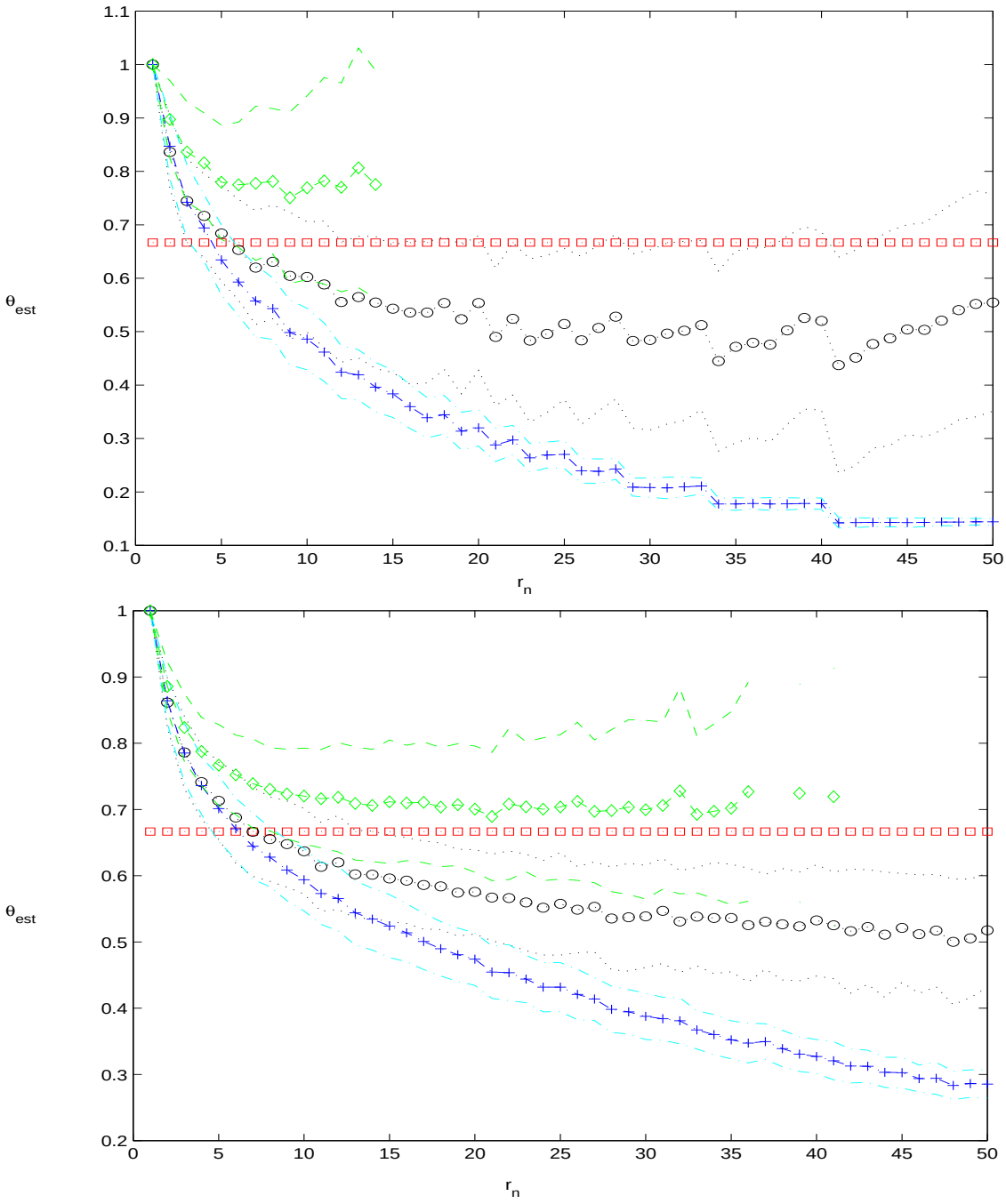

Figure 6.4. *Estimated values of the extremal index for the doubly stochastic model with* $\psi = 0.5$ *and* $\eta = 0.5$. *The extremal index is* $\theta = 0.66$ *plotted by* $\square$ *line. The partitions* r_n *considered are in the range* $[1, 50]$. $\widetilde{\theta}_n$ *is represented by* $(\cdots)$ *and* o*; the corresponding standard deviation is plotted with* $(\cdots)$. *The logs method* $\hat{\theta}_n^{(1)}$ *is represented with* $(- - -\!-)$ *and* $\diamond$. *The standard deviation with* $(- - -\!-)$. *The blocks method* $\hat{\theta}_n^{(2)}$ *with* $(\cdot - \cdot\!-)$ *and* $+$, *and* $(\cdot - \cdot\!-)$ *for the standard deviation. The sample sizes are* $n = 200$ *and* $n = 1000$ *respectively.* $m = 100$ *simulations are used. The threshold sequence is* $\hat{u}_n = x_{(n-k)}$ *with* $k = \sqrt{2n}$.

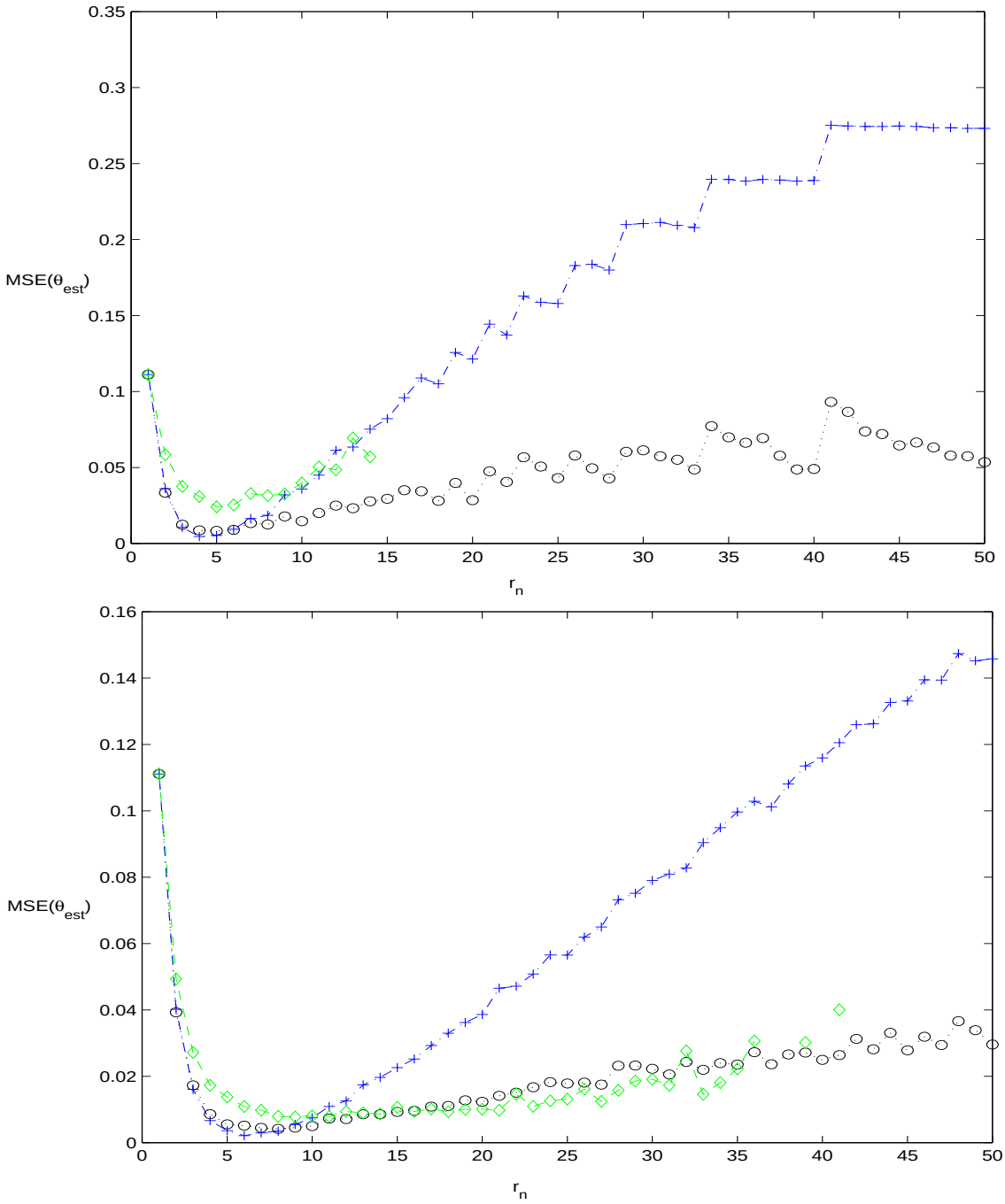

Figure 6.5. *Simulated mean square error (MSE) of the estimators of θ for the doubly stochastic model with $\psi = 0.5$ and $\eta = 0.5$. The partitions r_n considered are in the range $[1, 50]$. $m = 100$ simulations of the model are used. $\widetilde{\theta}_n$ is represented by $(\cdots)$ and o, $\hat{\theta}_n^{(1)}$ with $(---)$ and $\diamond$, and $(-\cdot-\cdot)$ and $+$ for $\hat{\theta}_n^{(2)}$. The sample sizes are $n = 200$ and $n = 1000$ respectively. The threshold sequence is $\hat{u}_n = x_{(n-k)}$ with $k = \sqrt{2n}$.*

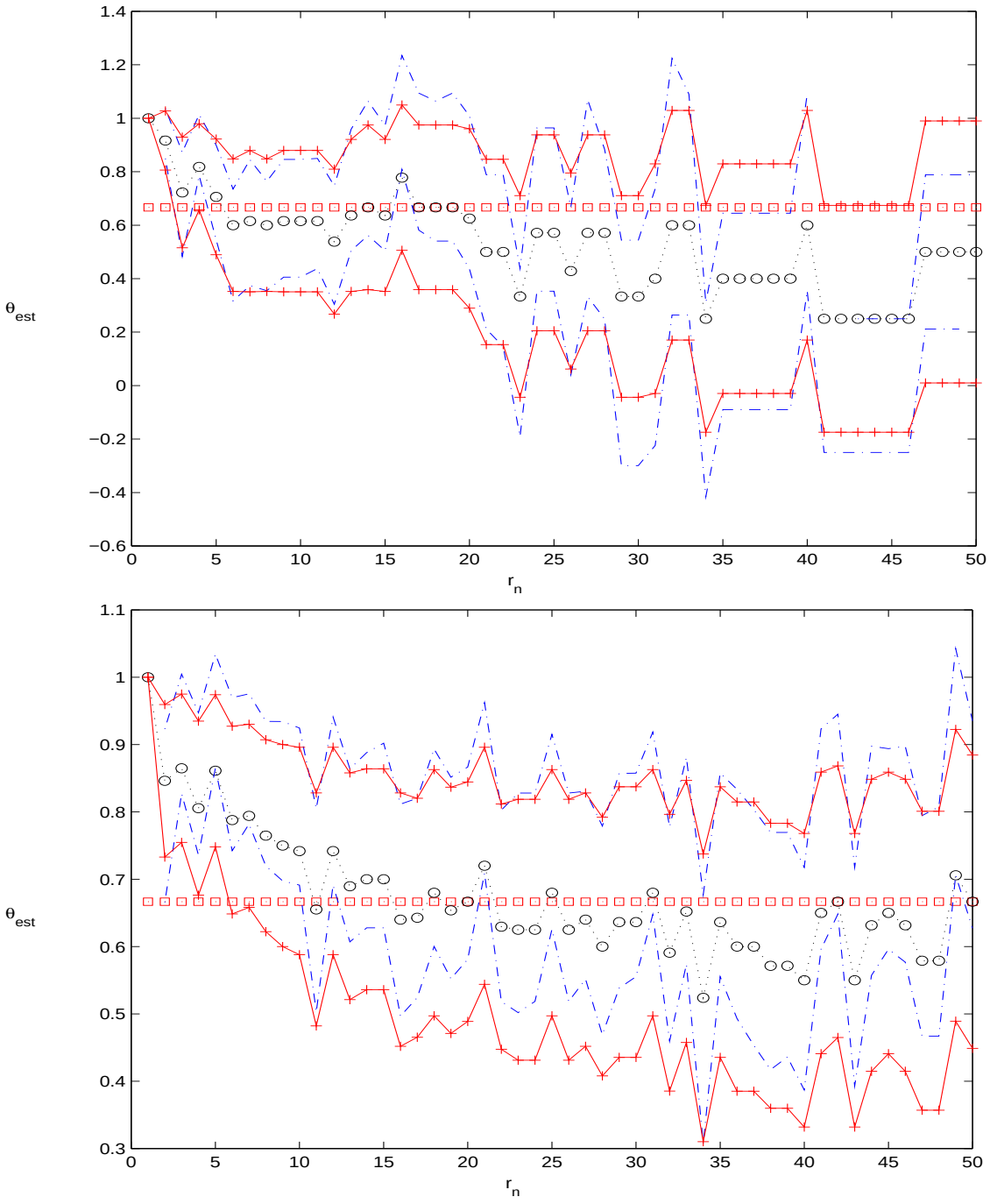

Figure 6.6. *Estimated values of the extremal index for the doubly stochastic model with* $\psi = 0.5$ *and* $\eta = 0.5$. *The extremal index is* $\theta = 0.66$ *plotted by* $\square$ *line. The partitions* r_n *considered are in the range* $[1, 50]$. $\widetilde{\theta}_n$ *is represented by* $(\cdots)$ *and* o*; the dash line describes the bootstrap confidence interval with* $B = 1000$ *and* $(+-)$ *is employed for the asymptotic intervals. The significance level is* $\alpha = 0.05$. *The sample sizes are* $n = 200$ *and* $n = 1000$ *respectively. The threshold sequence is* $\hat{u}_n = x_{(n-k)}$ *with* $k = \sqrt{2n}$.

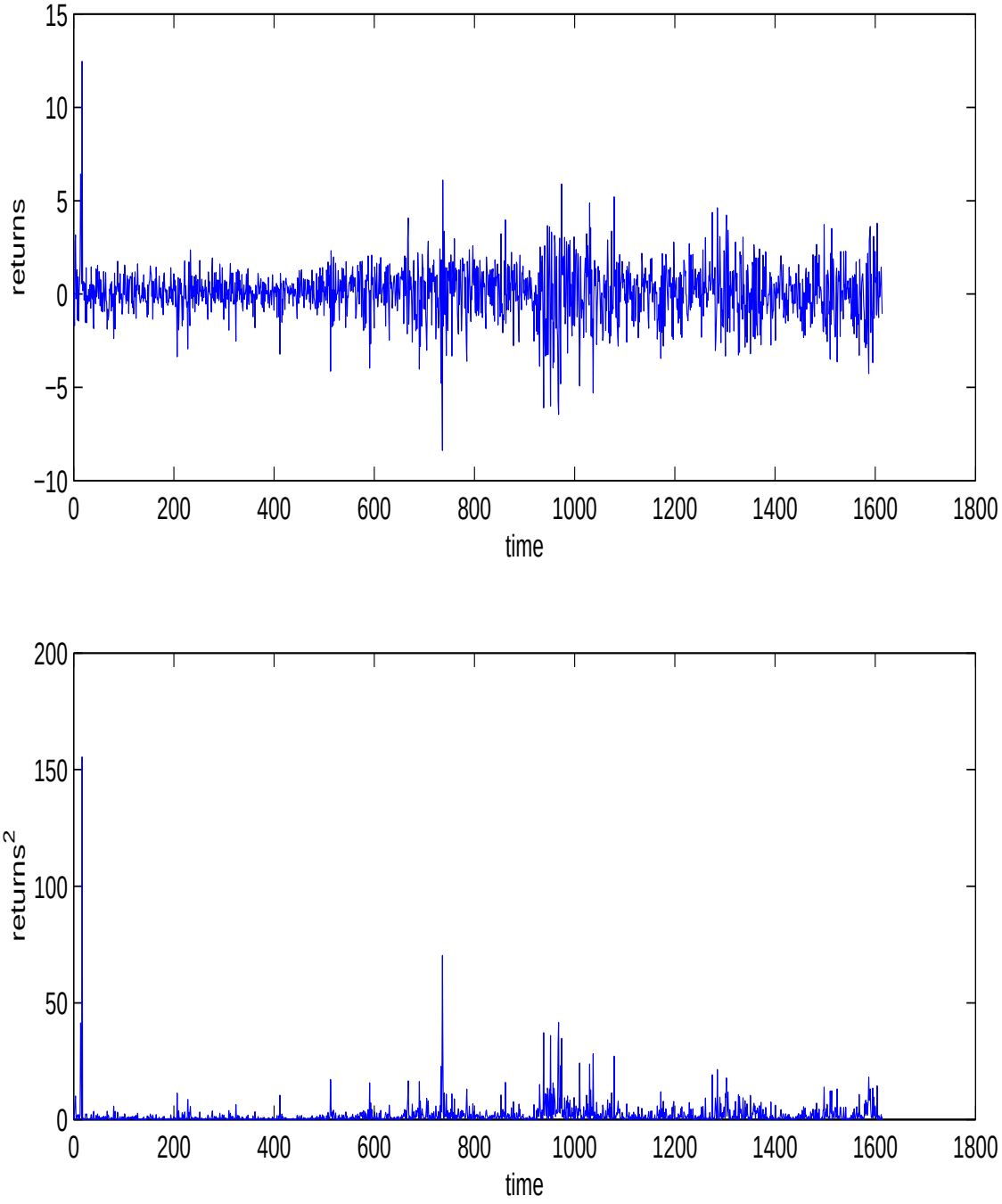

Figure 6.7. *DaX Index returns are represented in the upper panel. Squared returns showing the patterns of volatility are plotted in the lower panel. The sample period is* 19/12/1994 − 20/04/2001 *(*$n = 1614$ *observations).*

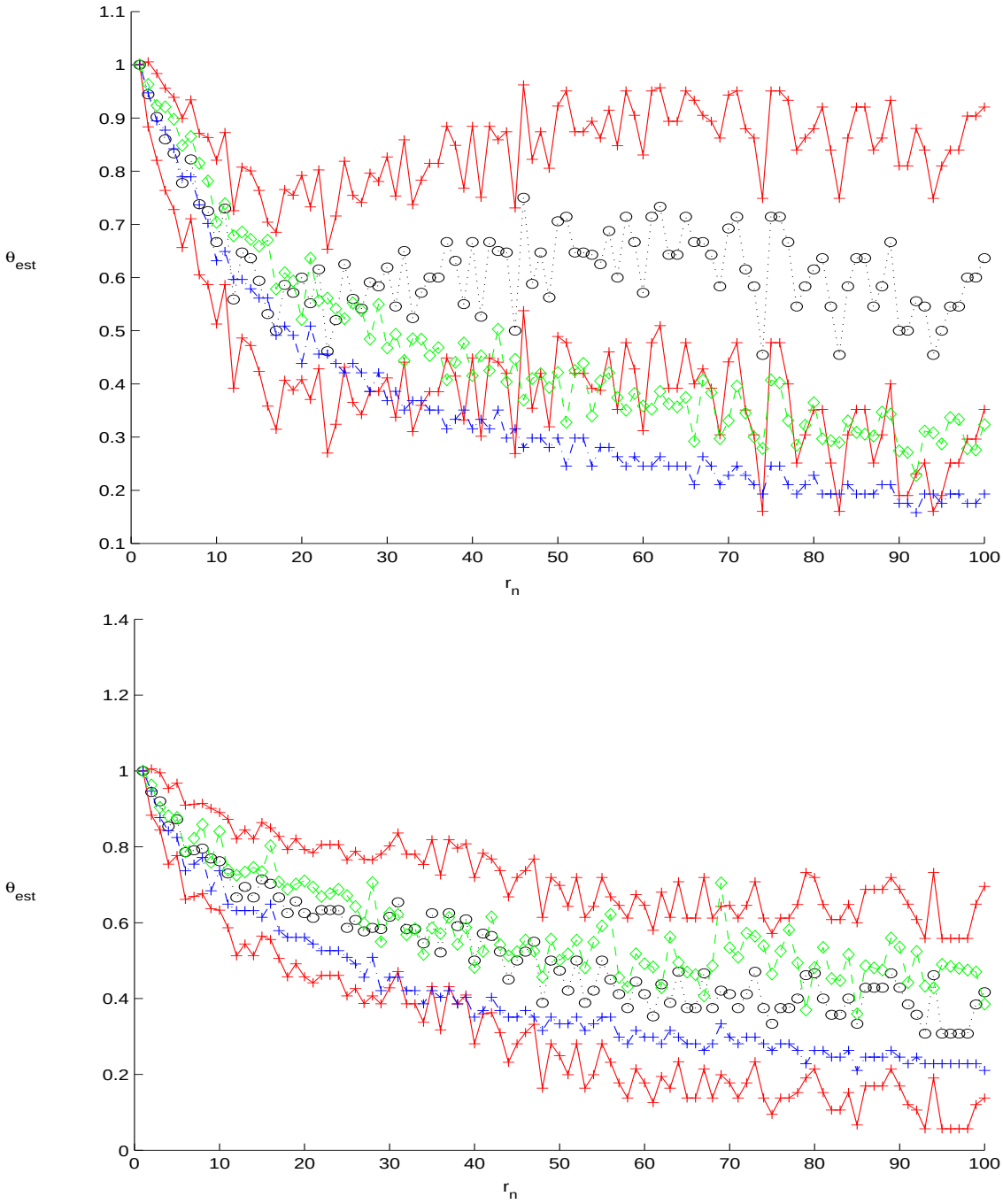

Figure 6.8. *Estimated values of θ_{pos} and θ_{neg} for the DaX Index returns over the period* $19/12/1994 - 20/04/2001$ *($n = 1614$). The upper panel estimates θ_{pos} and the lower panel θ_{neg}. $r_n \in [1, 100]$. $\widetilde{\theta}_n$ is represented by $(\cdots)$ and o; $(+-)$ describes the asymptotic confidence intervals with $\alpha = 0.05$. $\hat{\theta}_n^{(1)}$ with $(----)$ and $\diamond$, and $(-\cdot-\cdot)$ and $+$ for $\hat{\theta}_n^{(2)}$. $\hat{u}_{n,pos} = x_{(n-k)}$ and $\hat{u}_{n,neg} = x_{(k)}$ with $k = \sqrt{2n}$ are the corresponding thresholds.*

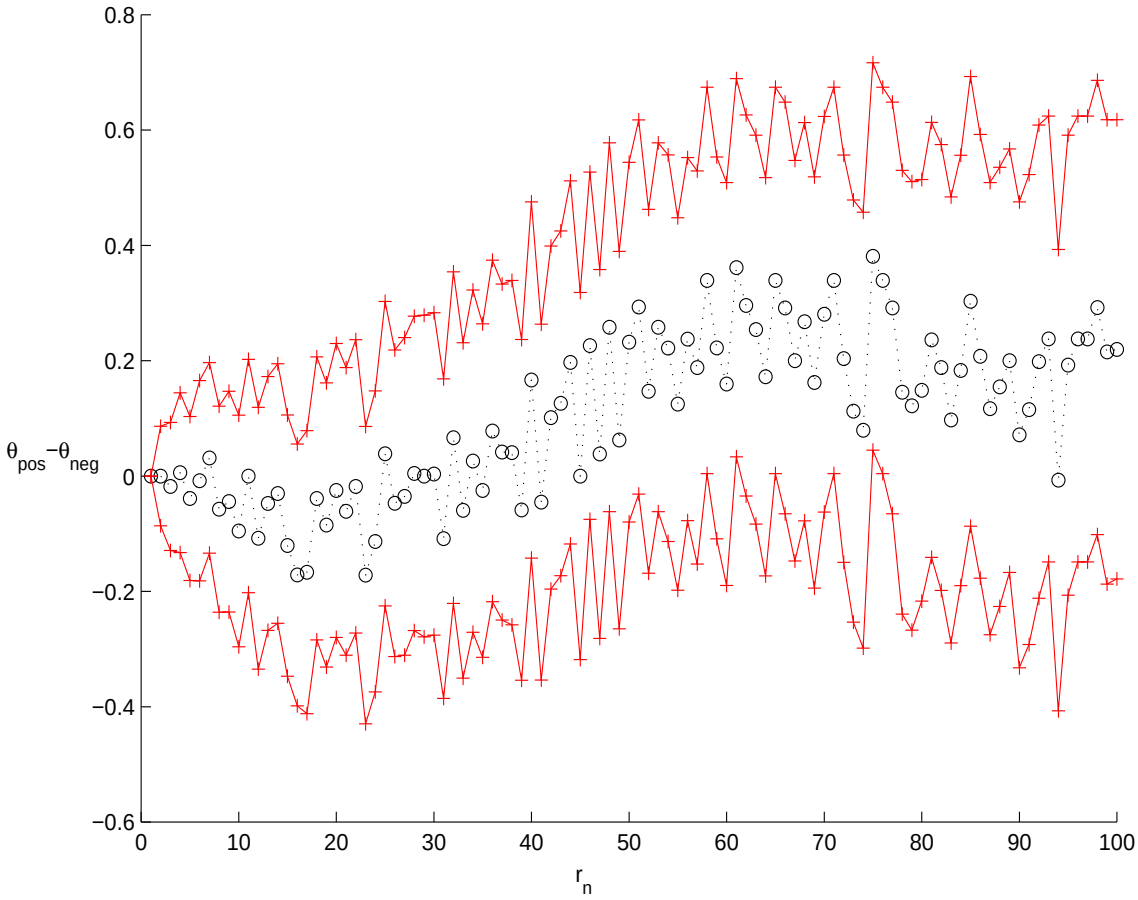

Figure 6.9. *Estimated values of* $\theta_{pos} - \theta_{neg}$ *for the DaX Index returns over the period* 19/12/1994 − 20/04/2001 *(n=1614).* $r_n \in [1, 100]$. $\widetilde{\theta}_{n,pos} - \widetilde{\theta}_{n,neg}$ *is represented by* $(\cdots)$ *and o;* $(+-)$ *describes the asymptotic confidence intervals with* $\alpha = 0.05$. $\hat{u}_{n,pos} = x_{(n-k)}$ *and* $\hat{u}_{n,neg} = x_{(k)}$ *with* $k = \sqrt{2n}$ *are the corresponding thresholds.*

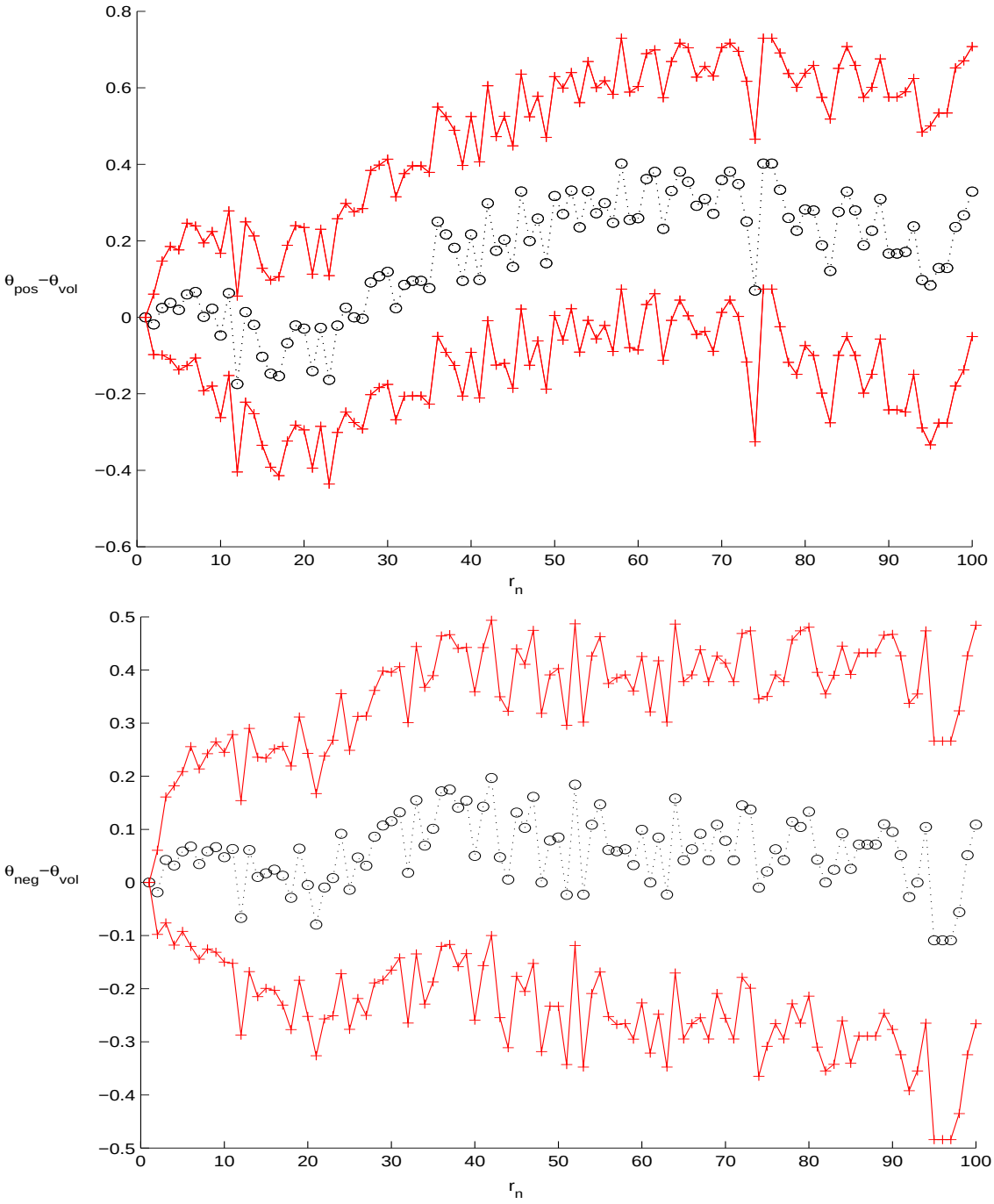

Figure 6.10. *Estimated values of $\theta_{pos} - \theta_{vol}$ (upper panel) and $\theta_{neg} - \theta_{vol}$ (lower panel) for the DaX Index returns over the period 19/12/1994 − 20/04/2001 ($n = 1614$). $r_n \in [1, 100]$. $\widetilde{\theta}_{n,pos} - \widetilde{\theta}_{n,vol}$ and $\widetilde{\theta}_{n,neg} - \widetilde{\theta}_{n,vol}$ are represented by $(\cdots)$ and o; $(+-)$ describes the asymptotic confidence intervals with $\alpha = 0.05$. $u_{n,pos} = x_{(n-k)}$ is the threshold for the positive exceedances and $u_{n,neg} = x_{(k)}$ for the negative exceedances, with $k = \sqrt{2n}$.*