On the Estimation of the Heavy–Tail Exponent in Time Series using the Max–Spectrum

Stilian A. Stoev†

Department of Statistics, The University of Michigan, Ann Arbor, U.S.A.

George Michailidis

Department of Statistics, The University of Michigan, Ann Arbor, U.S.A.

Summary. This paper addresses the problem of estimating the tail index α of distributions with heavy, Pareto–type tails for dependent data, that is of interest in the areas of finance, insurance, environmental monitoring and teletraffic analysis. A novel approach based on the max self–similarity scaling behavior of block maxima is introduced. The method exploits the increasing lack of dependence of maxima over large size blocks, which proves useful for time series data.
We establish the consistency of the proposed max–spectrum estimator for certain classes of dependent time series and demonstrate its robustness to short–lived contaminations in the data. The max–spectrum estimator exhibits *linear* computational time and memory complexity and can be calculated in a sequential manner, that makes it particularly well suited both for massive, as well as streaming data sets. Further, it provides a natural time–scale perspective in the analysis of heavy–tailed time series, not available in Hill–type techniques. The performance of the max–spectrum estimator is illustrated on synthetic and real data.

Keywords: heavy–tail exponent, max–spectrum, block–maxima, heavy tailed time series, moving maxima, max–stable, Fréchet distribution

1. Introduction

The problem of estimating the exponent in heavy tailed data has a long history in statistics, due to its practical importance and the technical challenges it poses. Heavy tailed distributions are characterized by the slow, hyperbolic decay of their tail. Formally, a real valued random variable X with cumulative distribution function (c.d.f.) $F(x) = \mathbb{P}\{X \leq x\},\ x \in \mathbb{R}$ is (right) heavy–tailed with index $\alpha > 0$, if

$$\mathbb{P}\{X > x\} = 1 - F(x) \sim \sigma_0^\alpha x^{-\alpha}, \text{ as } x \to \infty, \tag{1}$$

with $\sigma_0 > 0$, and where $\sim$ denotes that the ratio of the left–hand side to the right–hand side in (1) tends to 1, as $x \to \infty$. For simplicity purposes, we assume here that X is almost surely positive i.e. $F(0) = 0$. The *tail index* (*exponent*) α controls the rate of decay of the tail of F.

The presence of heavy tails in data was originally noted in the work of Zipf on word frequencies in languages (Zipf (1932)), who also introduced a graphical device for their detection (de Sousa and Michailidis (2004)). Subsequently, Mandelbrot (1960) noted their presence in financial data. Since the early 1970s heavy tailed behavior has been noted in many other scientific fields, such as hydrology, insurance claims and social and biological networks (see, e.g. Finkenstädt and Rootzén (2004) and Barabasi (2002)). In particular, the emergence of the Internet and the World Wide Web gave a new impetus to the study of heavy tailed distributions, due to their omnipresence in Internet packet and flow data, the topological structure of the Web, the size of computer files, etc. (see e.g. Adler et al. (1998), Resnick (1997), Faloutsos et al. (1999), Adamic and Huberman (2000,

†*Address for correspondence:* Stilian A. Stoev, Department of Statistics, The University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109-1107, U.S.A. *E-mail:* sstoev@umich.edu.

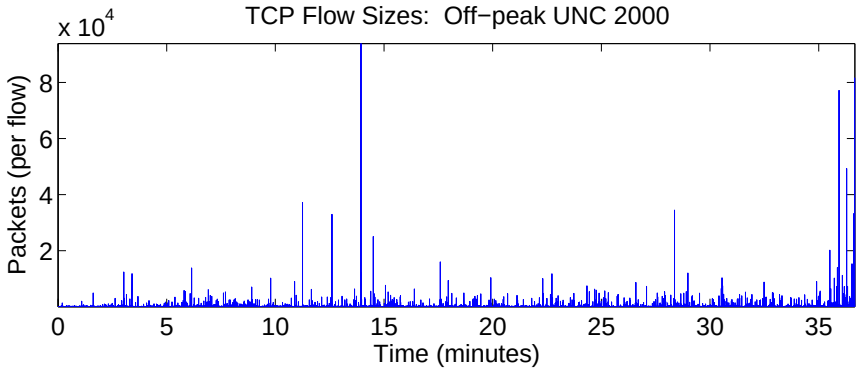

Fig. 1. TCP connection flow sizes (in number of packets) as a function of time.

2002), Park and Willinger (2000)). In fact, heavy tailed behavior is a characteristic of highly optimized physical systems, as argued in Carlson and Doyle (1999).

However, the availability of Internet and other high frequency data pose a set of new challenges for the problem of estimating the tail index discussed next in the context of the following data set that contains the size (in packets) of $142,170$ IP (Internet Protocol) flows, collected over a period of about 36 minutes on a link at the UNC Chapel Hill campus network starting at $8:30$ a.m. in 2000. A time plot of the data, shown in Figure 1, reveals several spikes indicating the potential presence of heavy tails. Some important features of such data are: (i) their large size due to the fine time scale resolution at which they are collected, resulting in hundreds of thousands or even millions of observations, (ii) their temporal structure that introduces dependence amongst observations, and (iii) their sequential nature, since observations are added to the data set over time. Traditional methods for estimating the tail index are not well suited for addressing these issues, as discussed below.

The majority of the approaches proposed in the literature focus on the scaling behavior of the largest order statistics $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(N)}$ obtained from an i.i.d. sample $X(1), \ldots, X(N)$ from F; typical examples include Hill's estimator Hill (1975) and its numerous variations (Kratz and Resnick (1996), Resnick and Stărică (1997)), kernel based estimators (Csörgő et al. (1985) and Feuerverger and Hall (1999)). A review of these methods and their applications is given in de Haan et al. (2000) and de Sousa and Michailidis (2004)). The most widely used in practice is the Hill estimator $\widehat{\alpha}_H(k)$ defined as:

$$\widehat{\alpha}_H(k) := \Big(\frac{1}{k}\sum_{i=1}^{k} \ln X_{(i)} - \ln X_{(k+1)}\Big)^{-1}, \tag{2}$$

with $k, 1 \leq k \leq N-1$ being the number of included order statistics. The parameter k is typically selected by examining the plot of the $\widehat{\alpha}_H(k)$'s versus k, known as the *Hill plot*. In practice, one chooses a value of k where the Hill plot exhibits a fairly constant behavior (see e.g. de Haan et al. (2000)). However, the use of order statistics requires *sorting* the data that: (i) is computationally demanding (`quicksort` requires on average $\mathcal{O}(N\log(N))$ steps, although its worst case performance is $\mathcal{O}(N^2)$), (ii) destroys the time ordering of the data and hence their temporal structure and (iii) summary statistics can not be recursively updated in the presence of new data, since a new sorting is required. Further, as can be seen from the brief review above, most of the emphasis has been placed on point estimation of the tail index and little on constructing confidence intervals. Exceptions can be found in the work of Cheng and Peng (2001) and Lu and Peng (2002) for the construction of confidence intervals and of Resnick and Stărică (1995) on the estimation of α for dependent data.

The purpose of this study is to introduce a method for estimating the tail index that overcomes the above listed shortcomings of other techniques. It is based on the asymptotic *max self–similarity* properties of heavy–tailed maxima. Specifically, the maximum values of data calculated over blocks of size m, scale at a rate of $m^{1/\alpha}$.

Therefore, by examining a sequence of growing, dyadic block sizes $m = 2^j,\ 1 \leq j \leq \log_2 N,\ j \in \mathbb{N}$, and subsequently estimating the mean of logarithms of block–maxima (log–block–maxima) one obtains an estimate of the tail index α. Notice that by using blocks of data, the temporal structure of the data is preserved. As shown in Section 2.2, this procedure requires $\mathcal{O}(N)$ operations, making it particularly useful for large data sets; further, the estimates for α can be updated recursively as new data become available, by using only $\mathcal{O}(\log_2 N)$ memory and without the knowledge of the entire data set, thus making the proposed estimator particularly suitable for streaming data (see e.g. Henzinger et al. (1998), Gilbert et al. (2001), and Muthukrishnan (2003)). Estimators based on max–self similarity for the tail index for i.i.d. data were introduced in Stoev et al. (2006), where their consistency and asymptotic normality was established. In this paper, we extend them to dependent data, prove their consistency, examine and illustrate their performance using synthetic and real data sets and discuss a number of implementation issues.

The remainder of the paper is structured as follows: in Section 2 the estimators are introduced and their computational complexity examined. Their consistency is established in Section 3.1 and the construction of confidence intervals is addressed in Section 3.2. Several implementation issues are examined in Section 4. Applications to financial time series are discussed in Section 5, while some concluding remarks are drawn in Section 6.

2. Max self–similarity and tail exponent estimators

In this section, we introduce the max self–similarity estimators for the tail exponent and demonstrate several of their characteristics.

2.1. Definition and basic properties

We briefly review the basic ideas behind the proposed estimators for independent and identically distributed (i.i.d.) data. A detailed exposition is given in Stoev et al. (2006). We discuss the case of *dependent* $X(i)$*'s* in Section 3.1 below.

Consider the sequence of block–maxima

$$X_m(k) := \max_{1 \leq i \leq m} X(m(k-1)+i) = \bigvee_{i=1}^{m} X(m(k-1)+i),\ \ k = 1, 2, \ldots,\ \ m \in \mathbb{N},$$

where $X_m(k)$ denotes the largest observation in the $k-$th block. By the Fisher–Tippett–Gnedenko Theorem we have that

$$\Big\{ \frac{1}{m^{1/\alpha}} X_m(k) \Big\}_{k \in \mathbb{N}} \xrightarrow{d} \Big\{ Z(k) \Big\}_{k \in \mathbb{N}}, \tag{3}$$

where $\xrightarrow{d}$ denotes convergence of the finite–dimensional distributions, with the $Z(k)$'s being independent copies of an $\alpha-$Fréchet random variable. For large values of m, the normalized block–maxima behave like a sequence of i.i.d. $\alpha-$Fréchet variables. Further, a sequence of i.i.d. variables is called *max self–similar* in the following sense:

Definition 2.1. A sequence of random variables $X = \{X(k)\}_{k \in \mathbb{N}}$ (defined on the same probability space) is said to be max self–similar with self–similarity parameter $H > 0$, if for any $m > 0$,

$$\Big\{ \bigvee_{i=1}^{m} X(m(k-1)+i) \Big\}_{k \in \mathbb{N}} \stackrel{d}{=} \Big\{ m^H X(k) \Big\}_{k \in \mathbb{N}}, \tag{4}$$

with $=^d$ denoting equality of the finite–dimensional distributions.

Relationship (4) holds asymptotically for i.i.d. data and exactly for Fréchet distributed data. Hence, any sequence of i.i.d. heavy–tailed variables can be regarded as *asymptotically max self–similar* with self–similarity

parameter $H = 1/\alpha$. This feature suggests that an estimator of H and consequently α can be obtained by focusing on the scaling of the maximum values in blocks of growing size. A similar idea applied to block–wise sums was used in Crovella and Taqqu (1999) for estimating α, in the case $0 < \alpha < 2$.

For an i.i.d. sample $X(1), \ldots, X(N)$ from F, define

$$D(j,k) := \max_{1 \le i \le 2^j} X(2^j(k-1)+i) = \bigvee_{i=1}^{2^j} X(2^j(k-1)+i), \;\; k = 1, 2, \ldots, N_j, \tag{5}$$

for all $j = 1, 2, \ldots, [\log_2 N]$, where $N_j := [N/2^j]$ and where $[x]$ denotes the largest integer not greater than $x \in \mathbb{R}$. By analogy to the discrete wavelet transform, we refer to the parameter j as the *scale* and to k as the *location* parameter. We consider dyadic block–sizes for algorithmic and computational convenience (see Section 2.2 for more details). Introduce the statistics

$$Y_j := \frac{1}{N_j} \sum_{k=1}^{N_j} \log_2 D(j,k), \;\; j = 1, 2, \ldots, [\log_2(N)]. \tag{6}$$

Then, a regression–based estimator of $H = 1/\alpha$ (and hence α) for a range of scales $1 \le j_1 \le j \le j_2 \le [\log_2(N)]$ is given by:

$$\widehat{H}_w(j_1, j_2) := \sum_{j=j_1}^{j_2} w_j Y_j, \quad \text{and} \quad \widehat{\alpha}_w(j_1, j_2) := 1/\widehat{H}_w(j_1, j_2), \tag{7}$$

where the weights w_j are chosen so that

$$\sum_{j=j_1}^{j_2} w_j = 0 \quad \text{and} \quad \sum_{j=j_1}^{j_2} j w_j = 1. \tag{8}$$

The optimal weights w_j can be calculated through *generalized least squares* (GLS) regression using the asymptotic covariance matrix of the Y_j's. In practice, it is important to at least use *weighted least squares* (WLS) regression which accounts for the magnitude of the variances of the Y_j's (see, Stoev et al. (2006)).

We propose to use the estimator defined in (7) in dependent time series data. We first illustrate its usage through a simulated data example. A data set of size $N = 2^{15} = 32,768$ was generated from an auto–regressive time series of order one with Pareto innovations. Specifically,

$$X(k) = \phi X(k-1) + Z(k) = \sum_{i=0}^{\infty} \phi^i Z(k-i), \;\; k = 1, \ldots, N,$$

where $\phi = 0.9$ and $\mathbb{P}\{Z(k) > x\} = x^{-\alpha}, \; x > 1$, with $\alpha = 1.5$. The data together with its Hill plot are shown in Figure 2. Notice that even though the Hill estimator work best for Pareto data, the dependence structure in the model makes the Hill plot quite misleading (see the bottom left panel). The zoomed–in version of the Hill plot (bottom right panel) however indicates that the tail exponent lies in the range 1.5 to 2 and the plot stabilizes very briefly around 1.75. Resnick and Stărică (1997) have shown that the Hill estimator is consistent for such dependent data sets. Nevertheless, as this example indicates, the Hill plot can be difficult to assess in practice.

In Figure 3, the *max–spectrum* plot is shown; i.e. the plot of the statistics Y_j versus the available dyadic scales $j, \; 1 \le j \le [\log_2 N] (= 15)$. The estimated tail exponent over the range of scales $(10, 15)$ is 1.4844, which is very close to the nominal value of $\alpha = 1.5$. Moreover, the max–spectrum is easy to assess and interpret. One clearly sees a "knee" in the plot near scale $j = 10$, where the max–spectrum curves upwards and thus it is natural to choose the range of scales $(10, 15)$ to estimate α. The choice of the scales (j_1, j_2) can be also automated, as briefly discussed in Section 4.1 below.

We examine next some important characteristics of the max–spectrum tail index estimators.

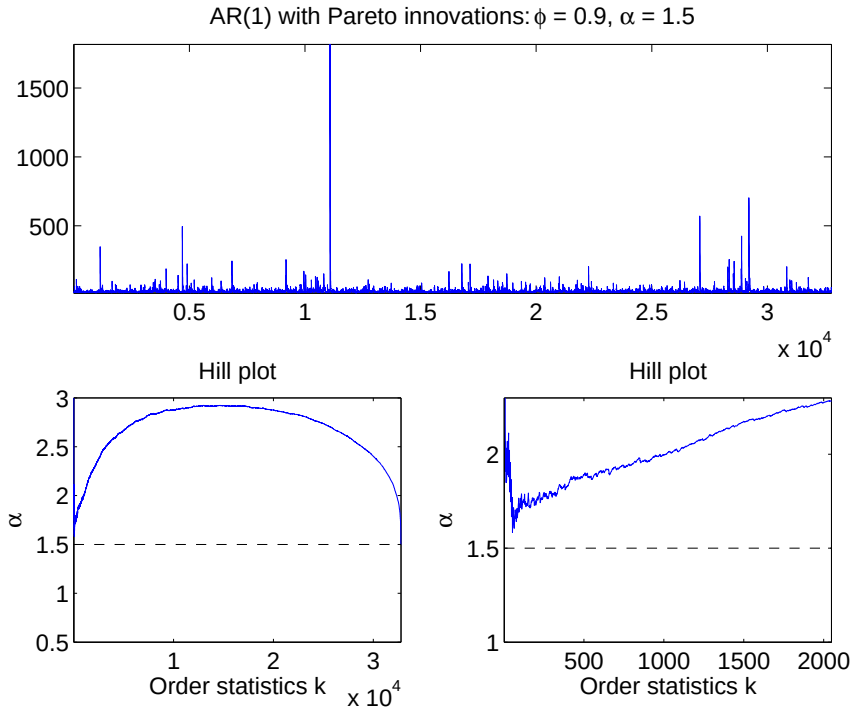

Fig. 2. *Top panel*: auto–regressive time series of order 1 with Pareto innovations of tail exponent $\alpha = 1.5$. *Bottom left and right panels*: the Hill plot for this data set and its zoomed–in version, respectively. The dashed horizontal line indicates the value of $\alpha = 1.5$.

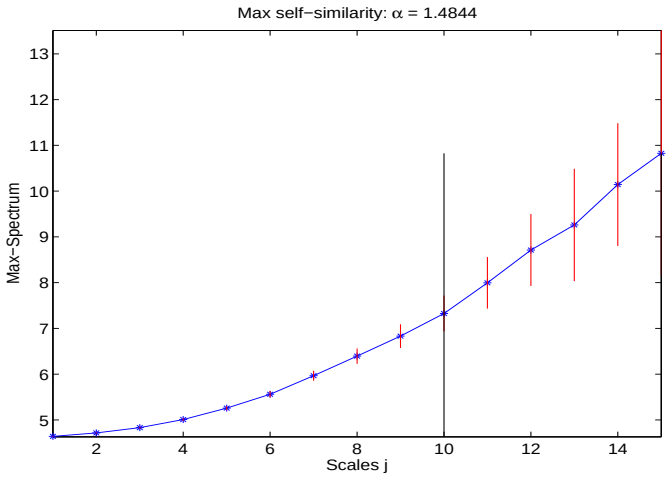

Fig. 3. The max–spectrum plot for the data set in Figure 2. The max self–similarity estimator of the tail exponent, obtained from the range of scales $(j_1, j_2) = (10, 15)$, is $\widehat{\alpha}(10, 15) = 1.4844$.

2.2. Algorithmic and computational features

The nature of the max-spectrum estimators offers significant computational advantages over the existing Hill–type and kernel–based estimators. Given a sample of size N, one can compute the max–spectrum Y_j, $1 \leq j \leq$

$[\log_2 N]$ in $\mathcal{O}(N)$ operations, since it involves $\mathcal{O}(N/2^j)$ pair–wise maxima and sums, for $j = 1, \ldots, [\log_2 N]$ for a total of

$$\mathcal{O}\Big(\sum_{j=1}^{[\log_2 N]} [N/2^j] \Big) = \mathcal{O}(N)$$

operations.

In contrast, methods based on the order statistics involve sorting the sample and hence require at least $\mathcal{O}(N \log_2(N))$ operations. Moreover, one can compute the max–spectrum of the data by using only $\mathcal{O}(N)$ memory, which is essentially the memory required to store the data. The *sequential* algorithm presented below uses $\mathcal{O}(\log_2 N)$ memory and operations per update. The computational advantages of our estimators become particularly important when dealing with massive data sets that have become ubiquitous in many applied areas, such as Internet traffic monitoring, sensor networks, finance and insurance. Further, the ability to compute the tail index sequentially proves useful when monitoring high frequency processes in real time, such as TCP flows of Internet traffic, traded volumes of stocks, etc. On the other hand, in order to use a Hill–type estimator, the whole sample of historical data needs to be stored for updating the order statistics. This is a tall requirement in areas where gigabytes of data are generated in short time (e.g. computer network traces, volume and price of financial data) or in situations where storage or computational resources are limited (e.g. sensor networks). An algorithm for sequentially updating the Y_j's is given next.

Algorithm

Variables: Keep N and a list of variables $\{Y_j,\ N_j,\ M_j$ and $R_j\}_{j=1}^{J}$.

Initialization: Given a preliminary sample $X(1), \ldots, X(N)$, one initializes the above variables as follows. First, set $J := [\log_2 N]$ and $N_j := [N/2^j],\ j = 1, \ldots, J$. Then, let $Y_j,\ j = 1, \ldots, J$ be as in (6) and define:

$$M_j := N - 2^j N_j,\ j = 1, \ldots, J \quad \text{and} \quad R_j := \max_{1 \le i \le M_j} X(2^j N_j + i).$$

Update: When a new sample point $X(N+1)$ is observed:

Step 1: For all $j = 1, \ldots, J$, set $M_j := M_j + 1$ and $R_j := \max\{R_j, X(N+1)\}$. Set $N := N + 1$.
Step 2: For all $j = 1, \ldots, J-1$: If $M_j = 2^j$, then update $Y_j,\ N_j, M_j$ and R_j as follows:

$$Y_j := (N_j Y_j + \log_2(R_j))/(N_j + 1) \text{ and set } N_j := N_j + 1,\ M_j := 0 \text{ and } R_j := 0.$$

Step 3: For $j = J$: if $M_j = 2^j$, then create $Y_{J+1} := \max\{Y_J, \log_2(R_J)\}$ and update $Y_J,\ N_J,\ M_J$ and R_J as in *Step 2*. Create $N_{J+1} := 1$, $M_{J+1} := 0$ and $R_{J+1} := 0$, and set $J := J + 1$.

Remark 1: Since $N_j = [N/2^j]$, for each scale j, there may be data points (left–over at the end of the data set) that are not enough to fit in a block of size 2^j. The variables R_j in the above algorithm contain the maxima of these data points, namely, $X(2^j N_j + 1), \ldots, X(N)$ and $M_j = N - 2^j N_j \le 2^j - 1$, denotes their number. When a new data point $X(N+1)$ arrives, one updates R_j as $R_j \vee X(N+1)$ and $M_j = M_j + 1$ (*Step 1*). If now the left–over data points are enough to complete a block, one updates Y_j and sets R_j and M_j to zero (*Step 2*). *Step 3* concerns the situation where there is enough new data to directly increase the number of available scales J to $J + 1$.

Remark 2: The algorithm operates with a small amount of memory of order $\mathcal{O}(J) = \mathcal{O}(\log_2 N)$ and when a new data point becomes available, the max–spectrum $Y_j,\ j = 1, \ldots, J$ statistics are updated by using only $\mathcal{O}(J) = \mathcal{O}(\log_2 N)$ operations. Further, the algorithm becomes operational even with a sample of 2 observations. In this case, to compute the max–spectrum of a sample of size N, one performs

$$\mathcal{O}(\log_2(2) + \log_2(3) + \cdots + \log_2(N)) = \mathcal{O}(N \log_2(N) - N) = \mathcal{O}(N \log_2(N))$$

operations.

Remark 3: In order to estimate α using the Hill estimator, one should maintain a *heap data structure* of size at least $\mathcal{O}(N)$ (see, e.g. Ch. II in Cormen et al. (2002)), as opposed to the $\mathcal{O}(\log N)$ memory requirement of the max–spectrum estimator.

Remark 4: In applications, when monitoring streaming data, one may also wish to discount the effect of observations in the distant past. This can be done by modifying the sequential algorithm above to include suitable exponential moving average versions of the statistics Y_j. For simplicity, in Figure 4 we merely computed the max–spectrum over 36 non–overlapping windows of the data.

2.3. TCP connection sizes – a sequential approach

We illustrate next the max self–similarity estimator and their sequential nature using the Internet data set of TCP connection flow sizes, shown in Figure 4. The top panel of Figure 4 diplays sizes (in packets) of the

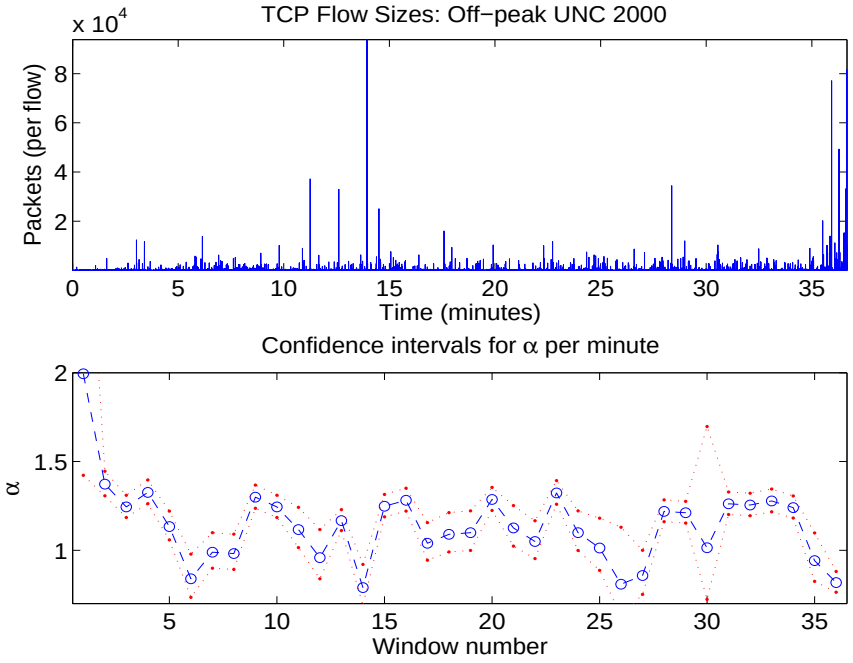

Fig. 4. *Top panel:* TCP connection (flow) sizes $X(k)$ (in number of packets) as a function of time. Every data point $X(k)$ has a time–stamp $t(k)$ of the time when the TCP connection ended. *Bottom panel:* max self–similarity estimates $\widehat{\alpha}(j_1, j_2)$ of the heavy tail exponent α, computed over one minute long windows of data. The scale j_2 equals 11 and the scale j_1 was chosen automatically with tuning parameters $p = 0.01$ and $b = 4$ (see Section 4.1). The dotted envelope indicates 95% confidence intervals for α obtained by using an asymptotic approximation.

$142,170$ TCP connections active over the main UNC Chapel Hill campus link during an off–peak period of about 36 minutes in 2000. The x–axis indicates the time the connection ended. The observed heavy tailed distribution is typical for such data and it can be attributed to the distribution of file sizes and/or durations of user activities (see e.g. Crovella et al. (1998) and the collection of papers in Park and Willinger (2000)). The heavy tailed nature of TCP connection sizes and durations is closely related to the observed self–similarity and long–range dependence of network traffic traces (see e.g. Park et al. (1996), Crovella and Bestavros (1996), Taqqu et al. (1997)). Thus, the value of the heavy–tail index α has important implications on the dimensioning

and management of the network. In fact, simply monitoring the tail exponent sequentially in time, as new traffic arrives, provides valuable feedback on the current state of the network. The bottom panel of Figure 4 displays max self–similarity estimates, calculated from one minute long windows of the data for each of the 36 available minutes. The dotted lines indicate 95% asymptotic confidence intervals for the corresponding windows (for their calculation see Section 3.2).

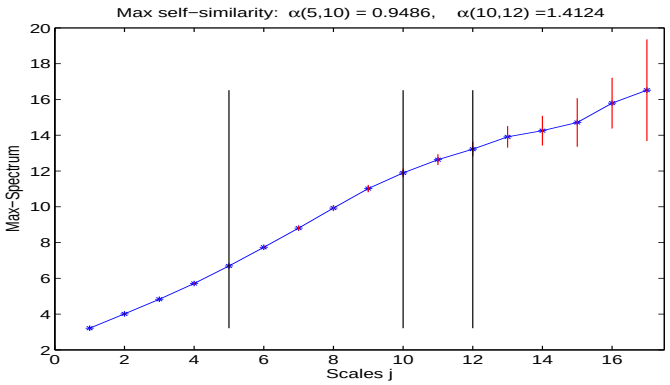

Fig. 5. The max–spectrum of the TCP connection sizes data set. The long vertical lines indicate two ranges of scales where the max–spectrum has different slopes. The range of smaller scales $(5, 10)$ yelds $\widehat{\alpha}(5, 10) = 0.9486$, which corresponds to the tail behavior of the data over time scales of about 2^{10} observations or 15 seconds. The second range $(10, 12)$ yields $\widehat{\alpha}(10, 12) = 1.4124$, which corresponds to the tails of the maxima over durations of about 15 seconds to 1 minute. Depending on the goals and time horizont of interest, the practitioner may choose to work with different parts of the max–spectrum.

Observe that the estimators track well the periods at which different magnitude heavy–tailed TCP connections occur. For example, the large spikes in the top plot around the 14-th and 36-th minutes correspond to tail exponents less than 1. In these cases, the confidence intervals for α do not contain 1. On the other hand, the confidence interval of the tail exponent around the 30-th minute is wide, which indicates that this fluctuation may not be significant. Further, the estimated tail exponents cluster around two values: 1.4 and 0.9. This is also seen on the max–spectrum plot of the entire data set in Figure 5. The block–maxima scale over small time periods of 2^{10} observations (about 15 seconds) with $\widehat{\alpha}(5, 10) = 0.9486$ and over intermediate ones (2^{10} to 2^{12} observations (15 seconds to 1 minute) with $\widehat{\alpha}(10, 12) = 1.4124$. This is largely due to the fact that over larger periods there are relatively few fluctuations, notably the spikes around the 14-th and 36-th minutes, resulting in a flatter slope for the max–spectrum.

For comparison purposes, we present in Figure 6 the Hill plot of the same data set, which proves challenging to assess. An inspection around the largest ($1 \leq k \leq 1,000$) and intermediate ($5,000 \leq k \leq 20,000$) order statistics, yields values similar to those obtained from the max–spectrum plot. For example, $\widehat{\alpha}_H(300) = 1.4114$ and $\widehat{\alpha}_H(12,000) = 0.9296$ match the max self–similarity estimators $\widehat{\alpha}(10, 12)$ and $\widehat{\alpha}(5, 10)$, respectively. Nevertheless, the Hill plot remains difficult to interpret objectively and can be quite misleading in practice. On the other hand, the max–spectrum is robust and easy to interpret. Finally, contrary to the max–spectrum estimator, it is not practical to compute and interpret the Hill plots sequentially, as new Internet traffic data become available. To do so, one needs to store either a sufficiently large window, or the entire past data.

3. Asymptotic properties

3.1. Consistency

The estimators $\widehat{H}_w$ and $\widehat{\alpha}_w = 1/\widehat{H}_w$ in (7) utilize the scaling properties of the max–spectrum statistics Y_j in (6). Therefore, their asymptotic behavior is closely linked to the asymptotic behavior of the max–spectrum as both

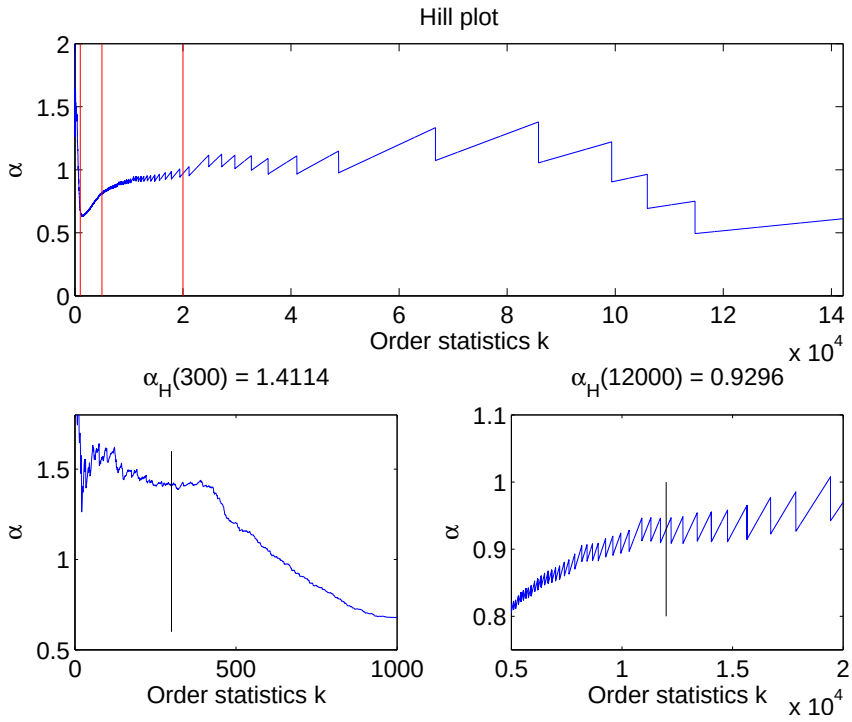

Fig. 6. *Top panel:* the Hill plot ($\widehat{\alpha}_H(k)$ versus k, see (2)) for the TCP connection sizes data set in Figure 4. Observe the volatile, saw–tooth shape of the graph, which does not appear to stabilize. Two zoomed–in regions of the Hill plot are displayed in the bottom (indicated by the vertical lines on the top plot).

the scale j and the sample size N tend to infinity. In the case of i.i.d. data, the Law of Large Numbers implies that for fixed j, as $N_j \to \infty$, the Y_j's are consistent and unbiased estimators of $\mathbb{E}Y_j = \mathbb{E}\log_2 D(j,1)$, if finite (see Corollary 3.1 in Stoev et al. (2006)). On the other hand, the asymptotic max self–similarity (3) of X and (5) imply that, as $j \to \infty$,

$$\mathbb{E}Y_j = \mathbb{E}\log_2 D(j,1) \simeq j/\alpha + C, \tag{9}$$

where $C = C(\sigma_0, \alpha) = \mathbb{E}\log_2 \sigma_0 Z$, and where $\simeq$ means that the difference between the left– and the right–hand side tends to zero, with Z being an $\alpha-$Fréchet variable with unit scale coefficient. A random variable Z is said to be $\alpha-$Fréchet, $\alpha > 0$, with scale coefficient $\sigma > 0$, if

$$\mathbb{P}\{Z \le x\} = \begin{cases} \exp\{-\sigma^\alpha x^{-\alpha}\} & , \ x > 0 \\ 0 & , \ x \le 0 \end{cases} \tag{10}$$

The previous discussion suggests that the max self–similarity estimators in (7) will be consistent as both the scale j and N_j tend to infinity. The consistency and asymptotic normality of these estimators was established in Stoev et al. (2006) for i.i.d. data. This was accomplished by assessing the rate of convergence of moment type functionals of block–maxima, such as $\mathbb{E}\log_2 D(j,1)$, under mild conditions on the rate of the tail decay in (1). Here, we focus on the case of dependent data and obtain sufficient conditions for the consistency of the proposed max self–similarity estimators. Deriving their asymptotic distribution for dependent data appears to be a challenging problem, which we plan to address in future work.

Consider a strictly stationary process (time series) $X = \{X(i)\}_{i\in\mathbb{N}}$ with heavy–tailed marginal c.d.f. F as in (1). Further, assume that the $X(i)$'s are positive, almost surely, that is, $F(0) = 0$. It is expected that the max

self–similarity estimators will work for dependent, time series data, provided that the block–maxima of X scale at a rate $m^{1/\alpha}$ as the block size m grows. The latter holds for processes with a positive *extremal index*.

Definition 3.1. A strictly stationary process X with marginal c.d.f. F as in (1) is said to have an extremal index $\theta_X > 0$, if

$$\frac{1}{n^{1/\alpha}}\Big(X(1) \vee \cdots \vee X(n)\Big) \xrightarrow{d} \theta_X^{1/\alpha} Z, \quad \text{as } n \to \infty, \tag{11}$$

where Z is an $\alpha-$Fréchet varible with scale coefficient σ_0.

The above definition is equivalent to the usual definition of the extremal index in the special case when the marginal c.d.f. F of X is as in (1) (see Ch. 3.7, p. 67 in Leadbetter et al. (1983)). As shown in Theorem 3.7.1 of Leadbetter et al. (1983), under the general *asymptotic independence* condition "$D(u_n)$" (see, p. 53 in Leadbetter et al. (1983)), the heavy–tailed time series X has an extremal index θ_X, where $0 \le \theta_X \le 1$. Observe that when the $X(i)$'s are i.i.d. then θ_X equals 1. Therefore, the extremal index θ_X can be viewed as a measure of the degree of extremal dependence among the $X(i)$'s – the closer θ_X to zero, the stronger the dependence and vice versa.

In order to gain some insight into the proposed approach, suppose for the time being, that X has a positive extremal index $\theta_X > 0$. In view of (11), relationship (9) holds with $C = C(\theta_X^{1/\alpha}\sigma_0, \alpha)$. Therefore, to obtain consistent estimates for the parameter $H = 1/\alpha$ one should focus on a range of scales which grow as the sample size increases. We thus *fix* a range $j_1 \le j \le j_2,\ j_1, j_2 \in \mathbb{N}$ and focus on the vectors $Y_r := \{Y_{j+r}\}_{j=j_1}^{j_2}$, with $r \in \mathbb{N},\ r + j_2 \le [\log_2 N]$ and where the parameter $r = r(N)$ grows with the sample size.

As in (7), define

$$\widehat{H} := \sum_{j=j_1}^{j_2} w_j Y_{j+r}, \tag{12}$$

where the w_j's are as in (8).

We present next sufficient conditions on the dependence structure of the process X that help us establish the consistency of the estimator $\widehat{H}$. Subsequently, we show that a large class of max–stable processes satisfies these conditions.

Condition E: *Let* $M_n := n^{-1/\alpha} \vee_{1\le i\le n} X(i),\ n \in \mathbb{N}$. *Suppose that* $\mathbb{E}|\log_2(M_n)|^p < \infty,\ p = 1, 2$, *and that,*

$$\mathbb{E}\Big(\log_2(M_n)\Big)^p \longrightarrow c_p, \quad \textit{as } n \to \infty, \tag{13}$$

where $c_p \neq 0,\ p = 1, 2$.

Condition C: *Let* $D(j, k)$ *be as in* (5). *Suppose that, for some* $j_0 \in \mathbb{N}$,

$$K(n) := \sup_{j\ge j_0, j\in\mathbb{N}} \Big|\mathrm{Cov}\Big(\log_2 D(j, n+1),\, \log_2 D(j, 1)\Big)\Big| \longrightarrow 0, \quad \textit{as } n \to \infty. \tag{14}$$

Observe that these conditions *do not* involve the extremal index θ_X of X, *nor* do they assume its existence. In practice, the extremal index of a time series X may be hard to evaluate and in fact may be difficult to relate to the asymptotic properties of the max–spectrum of X. The above conditions are more directly related to the asymptotic nature of the max self–similarity estimators. We discuss further Conditions E and C in the Remarks below. The following result shows that these conditions imply the consistency of $\widehat{H}$, with its proof given in Section 7.

Theorem 3.1. *Let* $X = \{X(i)\}_{i\in\mathbb{N}}$ *be a strictly stationary process with heavy–tailed marginal c.d.f.* F *as in* (1) *with* $F(0) = 0$. *Let* $\widehat{H},\ j_1 < j_2$ *and* $r = r(N)$ *be as in* (12) *with* w_j*'s as in* (8). *If* $r = r(N) \to \infty$ *and* $N/2^r \to \infty$, *as* $N \to \infty$, *then Conditions E and C imply*

$$\mathbb{E}(\widehat{H} - H)^2 \longrightarrow 0, \quad \textit{as } N \to \infty, \tag{15}$$

where $H := 1/\alpha$. *In particular,* $\widehat{H} \to^P H,\ N \to \infty$.

Condition C is readily satisfied if the sequence $X(i),\ i \in \mathbb{N}$ is $m-$dependent. We next show that the *moving maxima* Fréchet processes always satisfy Conditions E and C. Let $\alpha > 0$ and $a(k) \geq 0,\ k \in \mathbb{Z}$ be such that

$$\sum_{k\in\mathbb{Z}} a(k)^\alpha < \infty \tag{16}$$

and let $Z(k),\ k \in \mathbb{Z}$ be independent and identically distributed $\alpha-$Fréchet variables with unit scale coefficients. Then, the process $X = \{X(k)\}_{k\in\mathbb{Z}}$,

$$X(k) := \bigvee_{i\in\mathbb{Z}} a(k-i)Z(i) = \bigvee_{i\in\mathbb{Z}} a(i)Z(k-i),\ \ k \in \mathbb{Z}, \tag{17}$$

is said to be a moving maxima with kernel $a = \{a(k)\}_{k\in\mathbb{Z}}$ and innovations $Z(k),\ k \in \mathbb{Z}$. The analogous moving minima processes were first introduced by Deheuvels (1983). Davis and Resnick (1993) study the context of maxima, in a general setting. The maxima of moving maxima, or M3 processes, of Smith and Weissman (1996) provide flexible extensions to the moving maxima models (see also Zhang and Smith (2004)). General max–stable processes with Fréchet marginal distributions can be handled by using de Haan's spectral representation (de Haan (1984)), or by using the extremal stochastic integrals in Stoev and Taqqu (2006). Here, we focus on the simple, but useful moving maxima model although we expect that Conditions E and C continue to hold for M3 processes, as well.

The condition (16) guarantees that the maxima in (17) converge in probability and, in fact, almost surely, by monotonicity. Moreover, $X = \{X(k)\}_{k\in\mathbb{Z}}$ is a strictly stationary max–stable process; i.e. its finite–dimensional distributions are multivariate max–stable (see, e.g. Ch. 5 in Resnick (1987)). Specifically, for all $k_j \in \mathbb{Z},\ x_j > 0,\ j = 1,\ldots,n,\ n \in \mathbb{N}$, one has

$$\mathbb{P}\{X(k_j) \leq x_j,\ j = 1,\ldots,n\} = \exp\Big\{-\sum_{i\in\mathbb{N}}\Big(\bigvee_{1\leq j\leq n} a(k_j - i)^\alpha x_j^{-\alpha}\Big)\Big\}, \tag{18}$$

(see, e.g. Stoev and Taqqu (2006)).

For convenience, the scale coefficient of an $\alpha-$Fréchet variable ξ, will be often denoted by $\|\xi\|_\alpha$. Note that $\|\xi\|_\alpha$ does *not* equal $(\mathbb{E}\xi^\alpha)^{1/\alpha}$, which is infinite. Observe that, by (18), for any $k_j \in \mathbb{Z}$ and $b_j > 0,\ j = 1,\ldots,n,\ n \in \mathbb{N}$, the max–linear combinations $\bigvee_{j=1}^n b_j X(k_j)$ are $\alpha-$Fréchet variables with scale coefficients

$$\Big\|\bigvee_{j=1}^n b_j X(k_j)\Big\|_\alpha = \Big(\sum_{i\in\mathbb{Z}}\Big(\bigvee_{1\leq j\leq n} b_j a(k_j - i)\Big)^\alpha\Big)^{1/\alpha}. \tag{19}$$

This follows from (18) by setting $x_j := b_j^{-1},\ j = 1,\ldots,n$.

Note that by picking various sequences $a = \{a(k)\}_{k\in\mathbb{Z}}$, one can obtain moving maxima processes X with very different dependence structures. The slower the $a(k)$'s tend to zero, as $k \to \infty$, the stronger the dependence of the $X(k)$'s. In fact, by (19),

$$\Big\|\frac{1}{n^{1/\alpha}} X(1) \vee \cdots \vee X(n)\Big\|_\alpha^\alpha = \frac{1}{n}\sum_{i\in\mathbb{Z}}\bigvee_{1\leq j\leq n} a(j-i)^\alpha.$$

The following lemma of Smith and Weissmann implies that the extremal index θ_X of the moving maxima process X equals

$$\theta_X = \bigvee_{i\in\mathbb{Z}} a(i)^\alpha \Big/ \sum_{i\in\mathbb{Z}} a(i)^\alpha. \tag{20}$$

Lemma 3.1. *(Lemma 3.2 in Smith and Weissman (1996)) Let $b_k \geq 0,\ k \in \mathbb{Z}$ and suppose $\sum_{k\in\mathbb{Z}} b_k < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i\in\mathbb{Z}}\bigvee_{i+1\leq k\leq i+n} b_k = \bigvee_{k\in\mathbb{Z}} b_k. \tag{21}$$

We show next that the max self–similarity estimators are consistent for an arbitrary moving maxima $\alpha-$Fréchet process X.

Theorem 3.2. *The $\alpha-$Fréchet moving maxima process X defined in* (17) *satisfies Conditions E and C, where* $c_p = \mathbb{E}(\log_2 a^* + \log_2 Z(1))^p,\ p = 1, 2,$ *with* $a^* := \bigvee_{k\in\mathbb{Z}} a(k)$.

Proof: Given in the Appendix.

Remark 1: To establish the asymptotic distribution of the estimator $\widehat{H}$ in (12), one needs to study the covariance structure of the vector $\{Y_{j+r}\}_{j=j_1}^{j_2}$, as $r \to \infty$. Such a study would depend on the concrete model for the dependence of the $X(i)$'s and it falls beyond the scope of the present work.

Remark 2: Let X be a positive and strictly stationary process with heavy–tailed marginal distributions. Suppose that X has a positive extremal index; i.e. relation (11) holds. Then, the *uniform integrability* of the random variables $|\log M_n|^p,\ p = 1, 2,\ n \in \mathbb{N}$ implies Condition E. This follows from the definition of uniform integrability by using truncation and applying the continuous mapping theorem. Thus, if for some $\epsilon > 0$, we have $\sup_{n\geq 1} \mathbb{E}|\log M_n|^{2+\epsilon} < \infty$, then Condition E holds.

Remark 3: As indicated above, strictly stationary, $m-$dependent sequences X satisfy Condition C. The previous remark also shows that many such sequences satisfy Condition E and hence the max self–similarity estimator $\widehat{H}$ is consistent (Theorem 3.1).

3.2. On the construction of confidence intervals

In many applications, an uncertainty assessment about the estimated tail exponent is important and therefore one has to construct confidence intervals. In Stoev et al. (2006), confidence intervals for H in (7) were proposed, in the case of *i.i.d. data*, based on the asymptotic distribution of the max–spectrum. Theorem 4.1 therein implies, under certain conditions on the rate of convergence in (1), that

$$\sqrt{N_{j_2+r}}\Big(Y_{r+j} - \mu_r(j)\Big)_{j=j_1}^{j_2} \xrightarrow{d} \mathcal{N}(\vec{0}, \Sigma_\alpha(j_1, j_2)), \tag{22}$$

as $N \to \infty$, so that $r = r(N) \to \infty$ and $r(N)/\log_2(N) \to 0$. Here

$$\mu_r(j) := (j + r)/\alpha + C(\sigma_0, \alpha) \quad \text{and} \quad \Sigma_\alpha(j_1, j_2) := \alpha^{-2}\Big(\psi_{ij}(j_1, j_2)\Big)_{m\times m},$$

where $C = C(\sigma_0, \alpha)$ is as in (9), and where

$$\psi_{ij}(j_1, j_2) := 2^{\max\{i,j\}-j_2}\mathrm{Cov}(\log_2(Z_1),\, \log_2(Z_1 \vee (2^{|i-j|} - 1)Z_2)), \tag{23}$$

for independent, standard $1-$Fréchet variables Z_1 and Z_2.

In view of (7), the result (22) readily yields the asymptotic distribution of the estimator $\widehat{\alpha} = 1/\widehat{H}$. Thus, one has the following *asymptotic confidence interval* for α of level $\gamma,\ 0 < \gamma < 1$:

$$\Big((\widehat{H} - \widehat{H} z_{(1-\gamma)/2}\sqrt{c_w}/\sqrt{N_{r+j2}})^{-1}, (\widehat{H} + \widehat{H} z_{(1-\gamma)/2}\sqrt{c_w}/\sqrt{N_{r+j2}})^{-1}\Big), \tag{24}$$

where $z_{(1-\gamma)/2}$ is $(1+\gamma)/2-$quantile of the standard normal distribution, and where $c_w := \sum_{i,j=j_1}^{j_2} w_i w_j \psi_{ij}(j_1, j_2)$ with w_j as in (8). Here, as recommended in Stoev et al. (2006), we use the reciprocal of a symmetric confidence interval for H to obtain a confidence interval for $\alpha = 1/H$.

Observe that the asymptotic covariance matrix in (22) does not involve the scale parameter σ_0 (see (1)). Therefore, one does not need to estimate it to obtain a confidence interval for α. Numerical experiments with *independent data* show that the asymptotic confidence intervals in (24) work very well in practice.

Now, let us turn to the dependent case. There are very few results on confidence intervals for the heavy tail exponent even in the case of independent data. We are not aware of any general results on the asymptotic

j_1	3	4	5	6	7	8	9	10	11
90% c.i. $\phi = 0.1$	0.891	0.894	0.912	0.919	0.897	0.903	0.889	0.895	0.875
$\phi = 0.3$	0.759	0.888	0.914	0.915	0.899	0.901	0.889	0.895	0.875
$\phi = 0.5$	0.229	0.772	0.889	0.915	0.892	0.899	0.888	0.895	0.875
$\phi = 0.7$	0.000	0.299	0.801	0.895	0.895	0.899	0.887	0.895	0.875
$\phi = 0.9$	0.000	0.000	0.070	0.641	0.843	0.890	0.877	0.890	0.875
95% c.i. $\phi = 0.1$	0.943	0.952	0.954	0.953	0.949	0.950	0.931	0.931	0.904
$\phi = 0.3$	0.844	0.940	0.952	0.953	0.949	0.950	0.931	0.931	0.904
$\phi = 0.5$	0.321	0.854	0.950	0.954	0.948	0.950	0.931	0.931	0.904
$\phi = 0.7$	0.000	0.395	0.872	0.946	0.944	0.950	0.931	0.931	0.904
$\phi = 0.9$	0.000	0.000	0.123	0.738	0.911	0.941	0.927	0.930	0.904
99% c.i. $\phi = 0.1$	0.990	0.990	0.989	0.991	0.987	0.993	0.975	0.972	0.947
$\phi = 0.3$	0.946	0.985	0.990	0.991	0.987	0.992	0.975	0.972	0.947
$\phi = 0.5$	0.552	0.953	0.984	0.990	0.987	0.991	0.975	0.972	0.947
$\phi = 0.7$	0.000	0.642	0.959	0.981	0.988	0.990	0.974	0.972	0.947
$\phi = 0.9$	0.000	0.000	0.276	0.897	0.968	0.984	0.973	0.972	0.947

Table 1. Coverage probabilities of the asymptotic confidence intervals (24) for α for max–AR(1) time series as in (25) of length 2^{15}. Max self–similarity estimators $\widehat{H} = \widehat{H}(j_1, j_2)$ were used with $1 \leq j_1 \leq j_2$ and $j_2 = 15$. Results for three confidence levels: 90%, 95% and 99% are shown for different values of j_1.

distribution of the Hill or the moment estimator of α for dependent data. No theory for the asymptotic distribution of the max–spectrum is available either, in the dependent case. Developing such a theory is beyond the scope of the present work. Nevertheless, we will present next a heuristic asymptotic argument justifying the use of the confidence intervals (24) even when the data are dependent.

Let $X = \{X(i)\}_{i \in \mathbb{N}}$ be a strictly stationary time series with heavy–tailed marginal c.d.f. as in (1) with $F(0) = 0$. Suppose that X has a positive extremal index θ_X (see (11)). Under many dependence scenarios, the block maxima $X_m(k) := \max_{1 \leq i \leq m} X(m(k-1)+i)$ become "weakly dependent" as m grows. In fact, suppose that as in the independent case, (3) holds, where the $Z(k)$'s are i.i.d. $\alpha-$Fréchet. In this case, relation (11) implies that the scale coefficients of the $Z(k)$'s involve the extremal index of X. This suggests that, regardless of the dependence structure of X, relation (22) will continue to hold, under certain conditions on the rate of $r(N)$ as a function of N. We expect only the constant C in the centering quantity $\mu_r(j)$ to be affected by the dependence. Due to the asymptotic independence of the block–maxima, we expect the max–spectrum to have the *same* asymptotic covariance structure, as in the independent case.

If the limit covariance structure of the max–spectrum is the same as in the independent case, then the same confidence intervals as in (24) would apply. The fact that the extremal index is not likely to appear in the asymptotic distribution of $\widehat{\alpha}$ is important since its estimation is a challenging problem in its own right (see e.g. Ancona-Navarrete and Tawn (2000)).

Table 1 illustrates the coverage probabilities of the confidence intervals in (24) for dependent data. Specifically, $1\,000$ independent replications of max–AR(1) time series $X = \{X(k)\}_{k \in \mathbb{Z}}$:

$$X(k) := \phi X(k-1) + Z(k) = \bigvee_{i=0}^{\infty} \phi^i Z(k-i), \;\; k = 1, \ldots, N, \tag{25}$$

of size $N = 2^{15} = 32\,768$ for different values of ϕ were generated. The coverage probabilities for 90%, 95% and 99% levels, for confidence intervals based on $\widehat{H} = \widehat{H}(j_1, j_2)$ are reported in each row, as a function of j_1. Observe that when the data are closer to independent ($\phi = 0.1$), the coverage probabilities match the nominal values even for small j_1's. As the degree of dependence grows, larger values for j_1 are required to achieve accurate coverage probabilities. Nevertheless, even in the most dependent setting ($\phi = 0.9$) the value of $j_1 = 8$ yields very good results. The fact that coverage probabilities deteriorate for very large scales j_1 is due to the inadequacy of the normal approximation in (22) in the presence of a limited number of block–maxima.

These brief numerical experiments suggest that, provided the scales j_1 and j_2 are well–chosen, the confidence intervals in (24) work well in practice, even for dependent data. We plan to study further the problem of constructing confidence intervals for α in future work, both theoretically and through more extensive simulations.

4. Parameter selection, time scales and robustness

In this Section, several issues that relate to the performance and tuning of the max-spectrum estimator are examined.

4.1. On the automatic selection of the cut–off scale j_1

In the ideal case of $\alpha-$Frechet i.i.d. data, the max–spectrum plot of Y_j is linear in j. When the distribution of the data is not Fréchet, or when the data are dependent, then the max–spectrum is asymptotically linear, as the scales j tend to infinity. It is therefore important to select appropriately the range of large scales j for estimation purposes. In view of (9), one can always choose $j_2 = [\log_2 N]$ to be the largest available scale and hence, the problem is reduced to choosing the scale $j_1, 1 \le j_1 < j_2$. The estimator of α is then obtained by performing a WLS or GLS linear regression of Y_j versus $j,\ j_1 \le j \le j_2$ (see (7)).

The "cut-off" parameter j_1 can be selected either by visually inspecting the max–spectrum or through a data driven procedure. In Stoev et al. (2006) an automatic procedure for selecting the cut–off parameter was proposed, in the case of independent data, whose main steps are briefly summarized next. We also demonstrate that it performs satisfactorily for dependent data. The algorithm sets $j_2 := [\log_2 N]$ and $j_1 := \max\{1, j_2 - b\}$, with $b = 3$ or 4 in practice for moderate sample sizes. Next, j_1 is iteratively decreased until statistically significant deviations from linearity of $Y_j,\ j_1 \le j \le j_2$ are detected. Namely, as $j_1 > 1$, at each iteration over the scale j_1 the following two quantities are calculated $\widehat{H}_{\rm new} = \widehat{H}(j_1 - 1, j_2)$ and $\widehat{H}_{\rm old} = \widehat{H}(j_1, j_2)$. Whenever the value of zero is *not* contained a confidence interval centered at $(\widehat{H}_{\rm new} - \widehat{H}_{\rm old})$, the algorithm stops and returns the selected j_1 and $\hat{\alpha} = 1/\widehat{H}_{\rm old}$; otherwise, it sets $j_1 := j_1 - 1$ and proceeds accordingly. The construction of the confidence interval about $(\widehat{H}_{\rm new} - \widehat{H}_{\rm old})$ utilizes the covariance matrix Σ_1 (see (22)). Obviously, due to the fact that the exact asymptotic distribution is unknown for dependent data, the above procedure is heuristic in nature, but nevertheless exhibits a good performance in practice.

Figure 7 demonstrates the performance of the automatic selection procedure in the case of dependent data. Even though the marginal distributions of X are Fréchet, the dependence causes a knee in the max–spectrum plot (see, e.g. Figure 3). The automatic selection procedure picks up this "knee" and yields reasonably unbiased and precise *automatic estimates of* α (see the top–right panel in Figure 7). Comparing the MSE plot and the histogram of the selected j_1 values, we see that over 70% of the times the value $j_1 = 5$ was chosen, which is close to the optimal value of $j_1 = 6$. The histogram of the resulting automatic estimates of α (top–right panel) is not "very different" from the histogram of the estimators corresponding to the MSE–optimal $j_1 = 6$ (bottom–right panel).

Recall Table 1, and observe that the case $\phi = 0.9$ corresponds to the time series analyzed in Figure 7. The coverage probabilities of the confidence intervals for α essentially match the nominal levels, for $j_1 \ge 8$. On the other hand the MSE–optimal value is $j_2 = 6$ (Figure 7) which is only slightly smaller than $j_1 = 8$. This can be contributed to the fact that the bias involved in the estimators at $j_1 = 6$, although comparable to their standard errors is significant and noticeably *shifts* the confidence interval. As the scale j_1 grows, the bias quickly becomes negligible and the resulting confidence intervals become accurate.

These brief experiments suggest that the automatic procedure is practical and works reasonably well in the case of dependent moving maxima time series. Similar experiments for independent heavy–tailed data (not shown here) indicate that the automatic selection procedure continues to perform well and chooses values of j_1 close to the MSE–optimal ones, thus making it appropriate for use in empirical work. Nevertheless, a detailed study of its performance under a combination of heavy–tailed distributions and dependence structures, as well as its sensitivity to the choice of the back–start parameter b and the level of significance p, is necessary and the subject of future work.

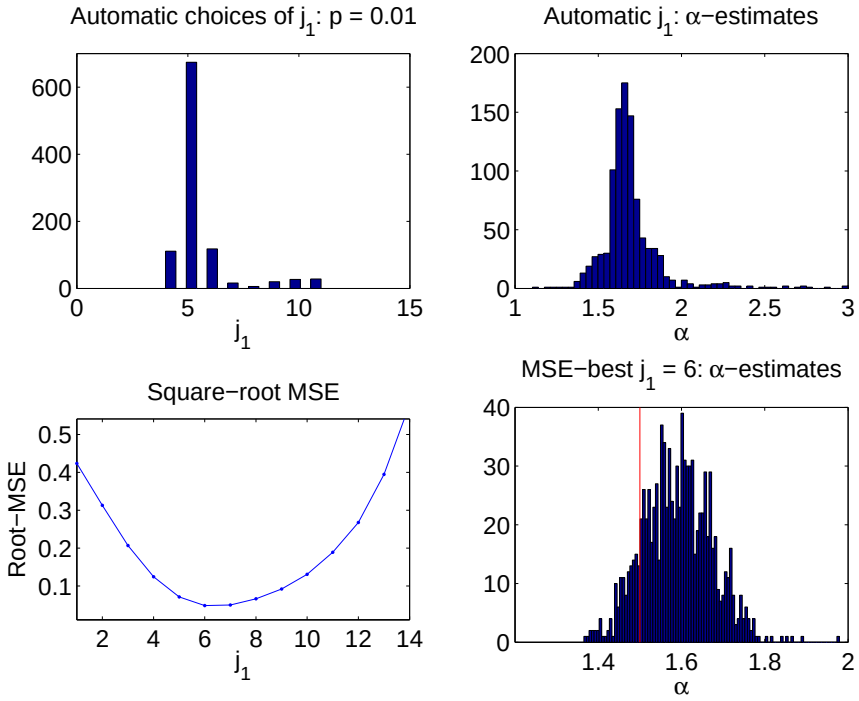

Fig. 7. The top–plot shows the histogram of automatically selected j_1 values for $1,000$ independent samples of size $N = 2^{15}$ from an exponential moving maxima $\alpha-$Fréchet process, $X = \{X(k)\}_{k\in\mathbb{Z}}$, defined as in (25) with $\phi = 0.9$ and with i.i.d. $1.5-$Fréchet innovations. We used significance level and back–start parameters are $p = 0.01$ and $b = 4$, respectively. The top–right plot show the histogram of the resulting $\widehat{\alpha} = 1/\widehat{H}$ estimates. The bottom–left plot shows estimates of the square root of the mean squared error (MSE) $\mathbb{E}(\widehat{H} - H)^2$ as a function of j_1. The bottom–right plot contains a histogram of $\widehat{\alpha}$ estimates obtained with the MSE–optimal choice of $j_1 = 11$.

4.2. Robustness and the emergence of extreme value time scales: numerical illustrations

In this section, we focus on two useful features for applications of the max self–similarity estimators: their robustness to short–lived non–stationary contaminants in the data and their ability to identify the time scales where heavy–tailed behavior emerges. The objective is to illustrate such features and *not* an exhaustive quantitative assessment.

In Figure 8, we demonstrate the max self–similarity estimators for a GARCH(1,1) time series $X = \{X_t\}_{t\in\mathbb{Z}}$:

$$X_t = \sigma_t(Z_t - \mathbb{E}Z_t), \quad \text{where} \quad \sigma_t^2 = \phi_0 + \phi_1 X_{t-1}^2 + \theta_1\sigma_{t-1}^2, \ \ t \in \mathbb{Z}, \tag{26}$$

with $\phi_0,\ \phi_1,\ \theta_1 > 0$. The Z_t's are i.i.d. Pareto distributed random variables ($\mathbb{P}\{Z_t > x\} = x^{-\alpha},\ \alpha = 5$, for $x \geq 1$). The parameters of the model were set to $\phi_0 = 1,\ \phi_1 = 0.1$ and $\theta = 0.8$. The fact that $\mathbb{E}Z_t^4 < \infty$, $\phi_0 > 0$ and $\phi_1 + \theta_1 < 1$ imply the existence of a stationary solution to the equation (26), called a GARCH(1,1) time series (see Theorem 1 in Bollerslev (1986)). The GARCH models and their numerous generalizations provide flexible classes of time series, widely used for modeling financial data (see e.g. Tsay (2005)).

To limit the effect of negative values, we analyzed a simple modification of the GARCH(1,1) time series, namely $X_t + \mu$, where $\mu = 15$. The classical Hill plot of the data, shown on the bottom left panel of Figure 8 is quite volatile, although it provides reasonably good estimates of $\alpha = 5$ for small and moderate values of k. The max–spectrum (in the bottom right panel) is more robust (than the Hill plot) and yields an estimate $\widehat{\alpha} \approx 5.3$ close to the nominal value. This was achieved by selecting the parameter $j_1 = 6$ automatically (with $p = 0.01$

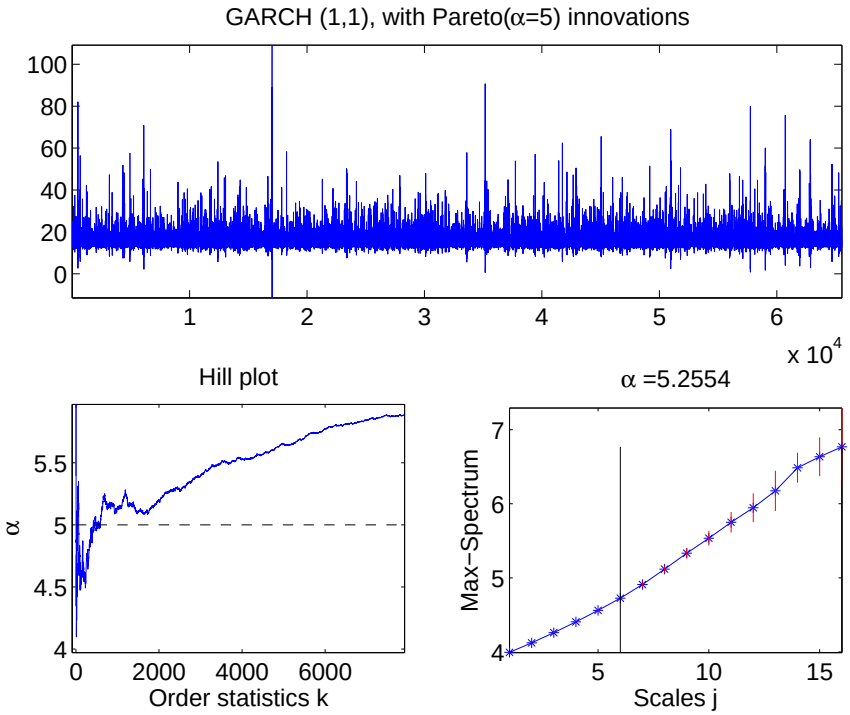

Fig. 8. The *top panel* displays a sample of a GARCH(1,1) time series with Pareto ($\alpha = 5$) innovations. *Bottom–left:* a plain Hill plot, zoomed to a range where the $\widehat{\alpha}_H(k)$'s are relatively constant in k. The resulting estimates are relatively close to the nominal value of $\alpha = 5$ (indicated by the horizontal dotted line) *only* for small values of k. The Hill plot exhibits a sizable bias for moderate and large k's. *Bottom–right:* the max–spectrum i.e. the log–scale statistics Y_j versus the scale j. The reciprocal of the slope yields an estimate of $\widehat{\alpha} = 5.2554$ for the nominal value of $\alpha = 5$.

and $b = 4$, see Section 4.1). Even in the difficult GARCH dependence context the max self–similarity estimator works well, even in the presence of Pareto data.

A well known feature of the Hill plot is its non–robustness to data coming from mixture distributions. The top plot of Figure 9, shows the GARCH time series from Figure 8, where about 1.5% of the data are replaced by i.i.d. Pareto variables with tail exponent $\alpha' = 1$. Notice that the data set in Figure 9 is not an i.i.d. sample from a mixture distribution, since the heavier tailed component is located at the beginning of the series. Nevertheless, the Hill plot involves the order statistics, and it cannot distinguish between mixtures and short–lived contamination in the data. The Hill plot shown, briefly stabilizes for small values of k taking values close to 1. It grows rapidly for larger order statistics and stabilizes briefly near the nominal level 5. Once again, it proves difficult to assess and interpret and in particular pick out the special features underlying this data set. On the other hand, the max–spectrum is relatively robust and essentially unaffected by the contamination. It yields an estimate of $\widehat{\alpha} \approx 4.89$, close to the nominal value, obtained by choosing the parameter $j_1 = 6$ automatically (with $p = 0.01$ and $b = 4$). The presence of the contaminant affects the max–spectrum only at the largest scales by making is slightly more steep.

We next demonstrate another interesting feature of the max self–similarity estimation framework; namely, that it provides insight into the time scales of the data, where certain heavy–tailed behavior becomes relevant. This aspect is particularly important when studying time series data, where the index of the observations has a physical meaning. In Figure 10, we show the heavy tailed GARCH(1,1) data set studied above, with a heavy–

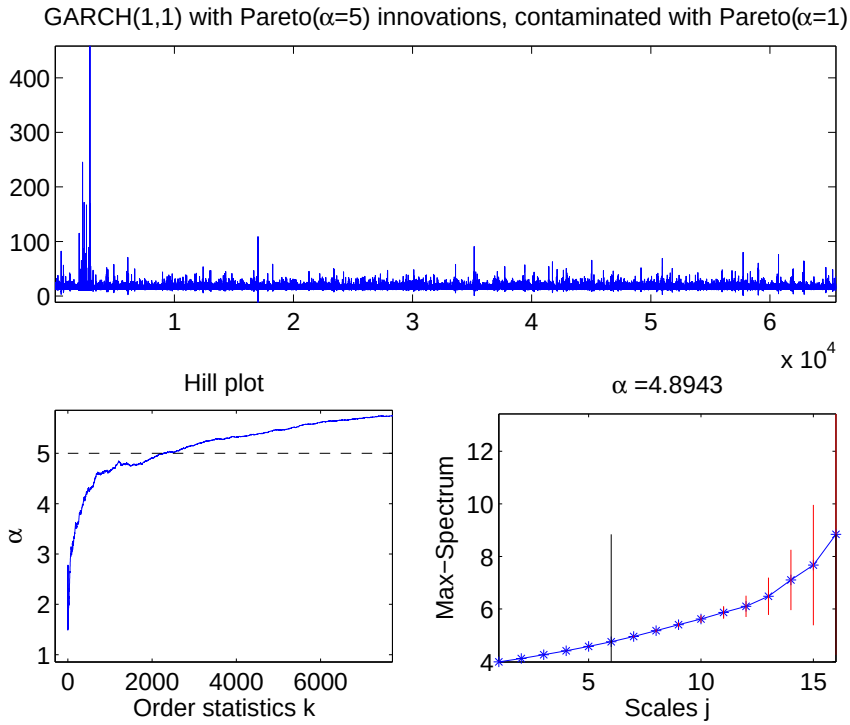

Fig. 9. The *top panel* displays the GARCH(1,1) time series from Figure 8, where about 1.5% of the sample was contaminated (replaced) by a Pareto distribution with tail exponent $\alpha' = 1$. Observe the region of larger values in the beginning of the data. The *bottom–left* and *bottom–right* panels show the Hill plot and the max–spectrum of the data, respectively.

tailed Pareto component inserted at every $256-$th sample value. Observe that over a very short segment of k values the Hill estimators $\widehat{\alpha}_H(k)$ happen to approximately identify the tail exponent $\alpha = 1$ of the Pareto component. The max self–similarity estimator also identifies the presence of the heavy–tailed component and estimates its index as $\widehat{\alpha} = 1.6994$. In addition, the procedure of automatic selection of the range of scales j chooses $j_1 = 10$ (solid vertical line), which can be linked to the time scale $256 = 2^8$ where the Pareto component starts to dominate the block–maxima. This is an important advantage of the max self–similarity estimators over the classical tail exponent estimators. The latter ones are blind to the temporal structure in the data since they involve sorted observations.

Observe also that the max–spectrum yields an estimate $\widehat{\alpha} \approx 5.25$, over the range of scales $[j_1, j_2] = [4, 8]$. This corresponds to the tail exponent of the underlying GARCH(1,1) time series. This region of scales also reflects the time scale where the heavier–tailed component in the data is irrelevant. One cannot obtain such information from the Hill plot, which does not provide time scale information and does not appear to stabilize near the level $\alpha = 5$ (horizontal dashed line).

5. Applications to Financial Data

We analyze market transactions for two stocks -Intel (symbol INTC) and Google (GOOG)- using the max–spectrum. The data sets were obtained from the *Trades and Quotes* (TAQ) data base of *consolidated transactions* of the *New York Stock Exchange* (NYSE) and NASDAQ (see Wharton Research Data Service (url)) and include the following information about every single trade of the underlying stock: *time of transaction* (up to seconds),

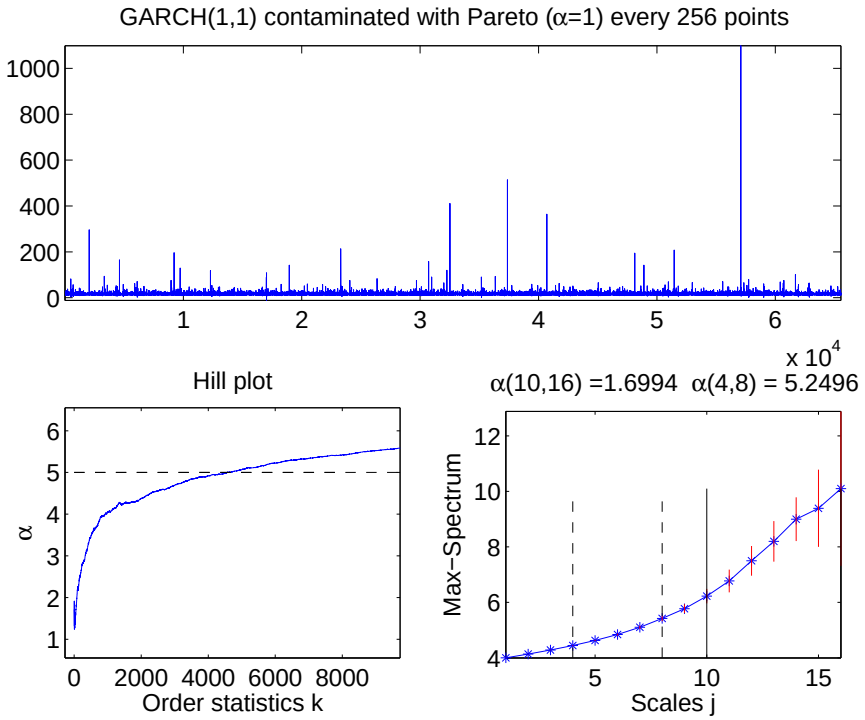

Fig. 10. The *top panel* displays the GARCH(1,1) time series from Figure 8. Every $256-$th data point is replaced by a sample from a Pareto distribution with tail exponent $\alpha' = 1$. The *bottom panels* display Hill plots and the max self–similarity spectrum of the data, respectively.

price (of the share) and *volume* (in number of shares). In our analysis, we focus on the traded volumes of the two stocks for November 2005, that could provide information about the respective sector's, as well as the market's economic conditions (Lo and Wang (2000)).

A ubiquitous feature of the volume data sets is the presence of heavy, Pareto type tails, as can be seen in Figure 12. Specifically, the top panel shows transaction volumes for the Google stock on November 7, 2005, while the bottom panels show the Hill and the max–spectrum plots, respectively. The tail exponent, estimated from the max–spectrum over the range of scales $(11, 15)$ is $\widehat{\alpha} = 1.0729$. The Hill plot indicates heavy–tail exponent estimates between 1.5 and 2, which correspond to the slope of the max–spectrum over the range of scales $(1, 10)$. The small dip in the Hill plot for very large order statistics (small k's) can be related to the behavior of the max–spectrum for scales $(11, 15)$ (see also Section 4.2, above). Such behavior is typical for almost all liquid stocks, as well as the presence of non–stationarity and dependence. In order to minimize the intricate non–stationarity effects, we focus here on traded volumes within a day. The max–spectrum yields consistent tail exponent estimates even in the presence of dependence. This fact and the demonstrated robustness of the max–spectrum (see Section 4.2) suggest that it may be safely used in various practical scenarios involving heavy–tailed data. In Figure 11, we show the max self–similarity estimates of the tail exponents, for each of the 21 trading days in November, 2005. The max–spectra of these 21 time series (not shown here) of trading volumes are essentially linear. This confirms the validity of a heavy–tailed model for the data, valid over a wide range of time scales – from seconds up to hours and days. Further, at the beginning and end of the trading day, several large volume transactions are observed, as documented in Hong and Wang (2000). Nevertheless, the trading activity of Google, remains essentially linear over the period under study, with a few bumps at the largest scales due to diurnal effects and other non–stationarities.

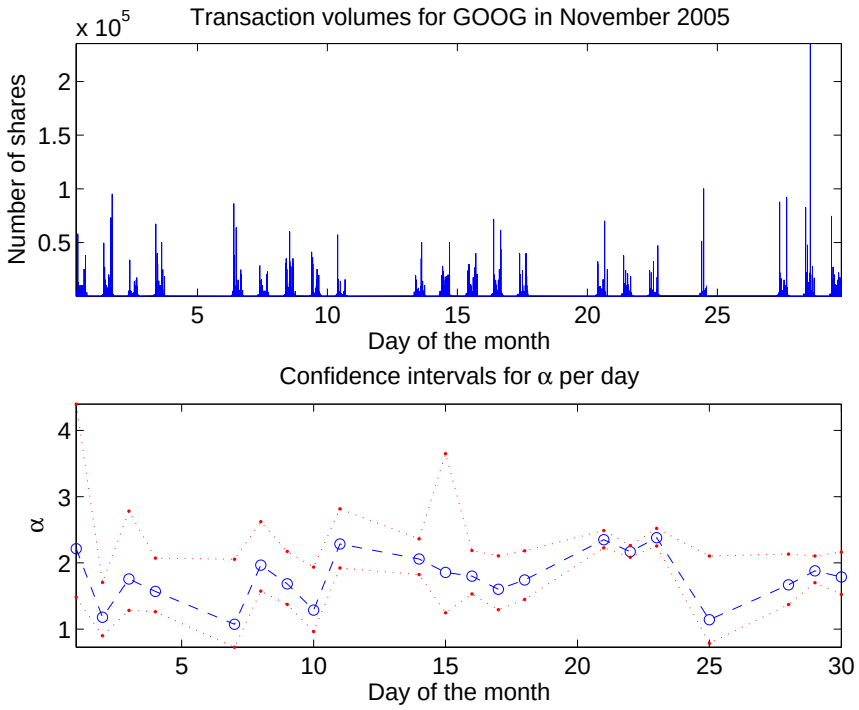

Fig. 11. *Top panel:* traded volumes for the Google stock from the TAQ data base of consolidated trades of NYSE and NASDAQ for the month of November, 2005. The x– axis and y–axis correspond to time and number of traded shares, respectively. This is a high– frequency data set, where each data point corresponds to the volume of a single transaction and no temporal aggregation is performed. The gaps of zeros in the data correspond to hours of the day with no trading and/or weekends. *Bottom panel:* estimated tail exponents (indicated by circles) from the max–spectrum and their corresponding 95% confidence intervals (indicated by broken lines), based on the asymptotic approximation for independent and identically distributed data. Automatic selection of the cut-off scale j_1 was done with $p = 0.1$ and $b = 3$ (see Section 4.1). Every estimate was computed from a day worth of transaction volumes.

In Figure 11, the daily tail exponent estimates are shown for the Google stock, which fluctuate between 1 and 2, along with pointwise confidence intervals (broken lines). These estimates indicate that the tail exponent exhibits a significant degree of variability over the period of a month, and that an infinite variance model may be most appropriate for modeling trading volumes. For example, on November 7 (see Figure 12), the estimate of α is nearly 1, which may be due to the several extremely large peaks in the volume data. The upward knee in the max–spectrum of this data set is likely caused by these peaks. The max–spectra on most other days are much closer to linear than the one in Figure 12. Such correspondence between the presence of large peaks in the data and the behavior of the max–spectrum can be used to identify statistically significant fluctuations in the volume data. Hence, the max–spectrum plot can be used not only to estimate α, but also to detect changes in the market. We illustrate this last point next, by examining an unusual trading pattern in the Intel stock towards the end of November, 2005.

Figure 13 shows the max–spectrum estimates of the tail exponents for the traded volumes of the Intel stock for 21 trading days in November 2005. Notice that up to November 21, the tail exponent is fairly constant, fluctuating between 1.2 and 2. On November 22 (Tue) and 23 (Wed), before the Thanksgiving holiday on November 24 (Thur), the tail exponent takes values larger than 3 and 5, respectively. This change is quite surprising and it is deemed significant by the corresponding confidence intervals. A closer look at the data from November 23

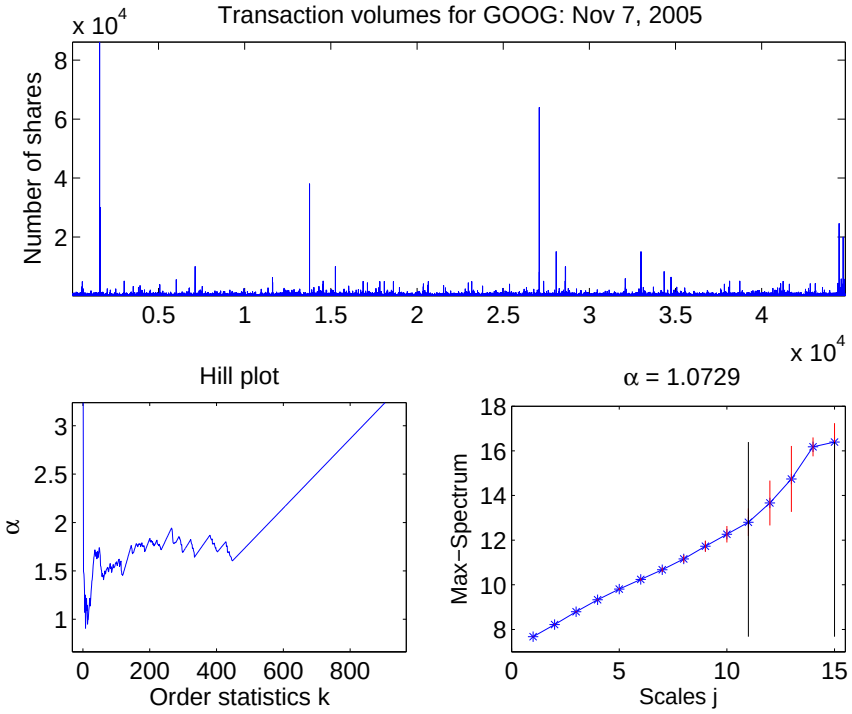

Fig. 12. *Top panel:* the transaction volumes during the trading hours of November 7, 2005. The x–axis corresponds to the number of the transaction and the y–axis to number of shares. Note that about $50,000$ transactions occurred on this day, which is typical for the Google stock. Observe also the fairly classical heavy–tailed nature of the volume data. *Bottom panels:* the Hill plot (left) and the max–spectrum (right) of the data. The Hill plot is zoomed–in to a range where it is fairly constant and a tail exponent between 1.5 and 2 can be identified. The max–spectrum reveals more: on large scales the plot is steeper than on small scales with the tail exponent about 1 on the range of scales $(11, 15)$ and exponent about 1.7 on scales $(1, 10)$. The presence of a knee in the max– spectrum plot suggests different behavior of the largest volumes on large time scales than on small time scales and can be contributed to the several very large spikes of over $20,000$ traded shares (about 5 million US dollars) the top plot.

(Figure 14) shows a changing but persistent pattern of trading as compared to November 21; see for example Figure 15).

This behavior proves persistent and continues on November 25, after the Thanksgiving holiday. Moreover, no such behavior was observed for the Google data on any of the 21 trading days in November, 2005. Although trading of extremely large volumes occurs on November 23, as seen in Figure 14, these trades are very regular and hence inconsistent with a heavy–tailed model. Although regular in time, these large transactions occur on a time scale of several minutes, and hence the small scales of the max–spectrum are not affected by these peaks and behave as on a normal trading day (see Figure 15). However, the large peaks dominate the larger scales j and their regularity makes the max–spectrum essentially horizontal. The Hill plot, shown on the bottom–left panel of Figure 14, fails to pick up the unusual behavior, since it suggests values of $\alpha \approx 1$, which corresponds only to the smallest portion of the max–spectrum, where $\widehat{\alpha}(7, 11) = 1.0578 \approx 1$.

Our best guess is that this change in activity is related to the approval by the board of directors of the Intel Corp. on November 10 of a program for a stock buy–back worth of up to 25 billion US dollars; (see, e.g. the Financial Times, London, on Thursday November 11, page 27); hence, some of the delayed effects of the announcement of the program and market reaction to it are demonstrated in the volume activity discussed above.

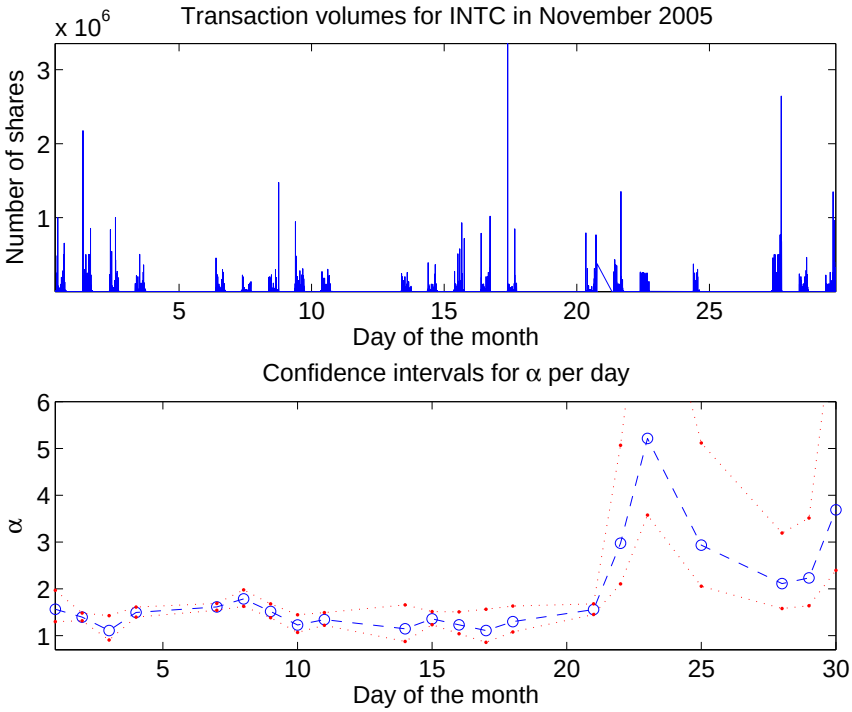

Fig. 13. This figure has the same format as Figure 11. On the top panel, the traded volumes of the Intel stock for the month of November, 2005 are shown. Observe that the tail exponent estimates on the bottom plot fluctuate between 1.5 and 2 up to November 21. On and after November 22, unusually high values of α appear (compare with the case of the Google stock in Figure 11). This is further analyzed in Figures 14 and 15, below.

6. Conclusion

In this paper, the problem of estimating the tail index of heavy tailed dependent data is studied and an estimator based on the max self–similarity scaling behavior of block maxima is introduced. Its consistency is established and several of its features discussed. Analysis of Internet traffic and stock volume data demonstrate its usefulness in applications. As indicated on several occasions, the problem of deriving the asymptotic distribution of the proposed estimator is of interest for further study.

7. Proofs

PROOF OF THEOREM 3.1: In view of (12) and (8), we have that

$$\widehat{H} - H = \sum_{j=j_1}^{j_2} w_j (Y_{j+r} - \mathbb{E}Y_{j+r}) + \sum_{j=j_1}^{j_2} w_j (\mathbb{E}Y_{j+r} - (j+r)/\alpha - c_1), \tag{27}$$

where c_1 is as in (13). Now, by applying the Cauchy–Schwartz inequality and the inequality

$$\Big| \sum_{i=1}^{m} x_i \Big|^p \leq m^{0 \vee (p-1)} \sum_{i=1}^{m} |x_i|^p, \ m \in \mathbb{N}, \quad \text{valid for all } p, \ x_i \in \mathbb{R}, \ i = 1, \ldots, m, \tag{28}$$

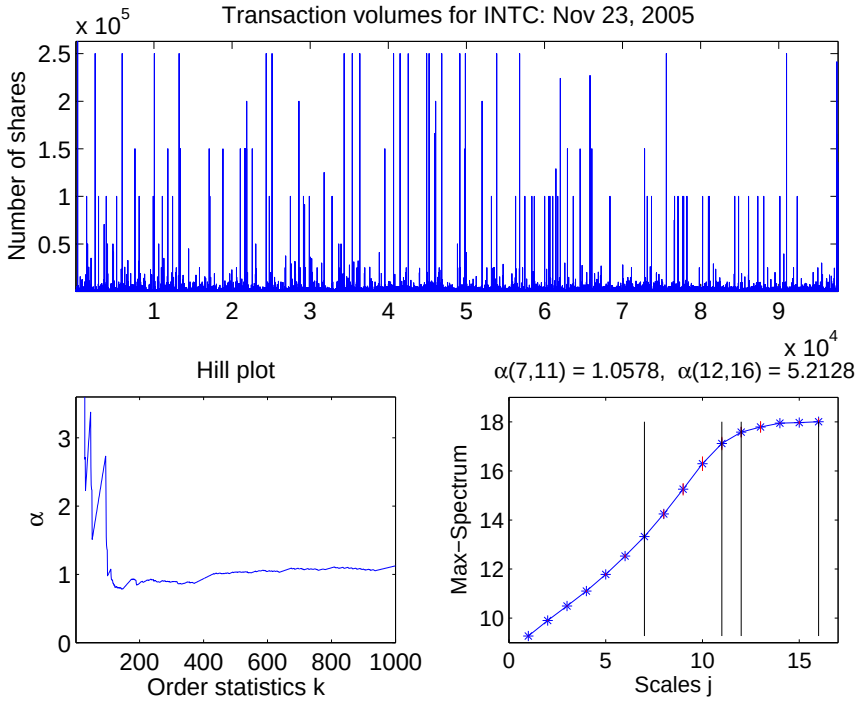

Fig. 14. *Top panel:* traded volumes of the Intel stock for November 23, 2005. Observe the regular occurrence of many very large trades of approximately the same sizes: 10, 000, 15, 000, 25, 000 and a few of 20, 000 shares. This is a very unusual behavior of the volume data, as compared to a typical trading day (see, e.g. Figure 15). *Bottom panels:* the Hill plot and the max–spectrum of the data. Notice that the Hill plot fails to identify the unusual behavior of the data, whereas the max–spectrum flattens out, on large scales due to the regular non–heavy tailed behavior of the largest traded volumes. Once identified on the max–spectrum plot, one can perhaps read–off these details from the volatile Hill plot for very small k's. On small scales, where the regular large transactions are not frequent and do not play a role, the max–spectrum yields tail exponents about 1. This is in line with the Hill plot.

we obtain

$$\mathbb{E}(\widehat{H}-H)^2 \le \text{const}\Big(\sum_{j=j_1}^{j_2}\mathbb{E}(Y_{j+r}-\mathbb{E}Y_{j+r})^2+\sum_{j=j_1}^{j_2}(\mathbb{E}Y_{j+r}-(j+r)/\alpha-c_1)^2\Big). \tag{29}$$

Fix $j,\ j_1\le j\le j_2$ and observe that

$$\begin{aligned}\mathbb{E}(Y_{j+r}-\mathbb{E}Y_{j+r})^2 &= \frac{1}{N_{j+r}}\mathrm{Var}(\log_2 D(j+r,1)) \\ &\quad+\frac{2}{N_{j+r}}\sum_{k=1}^{N_{j+r}-1}\frac{(N_{j+r}-k)}{N_{j+r}}\mathrm{Cov}(\log_2 D(j+r,k+1),\ \log_2 D(j+r,1)).\end{aligned} \tag{30}$$

Note that

$$\mathrm{Var}(\log_2 D(j+r,1))=\mathrm{Var}\Big(\log_2(D(j+r,1)/2^{(j+r)/\alpha})\Big)=\mathrm{Var}\Big(\log_2 M_{2^{j+r}}\Big),$$

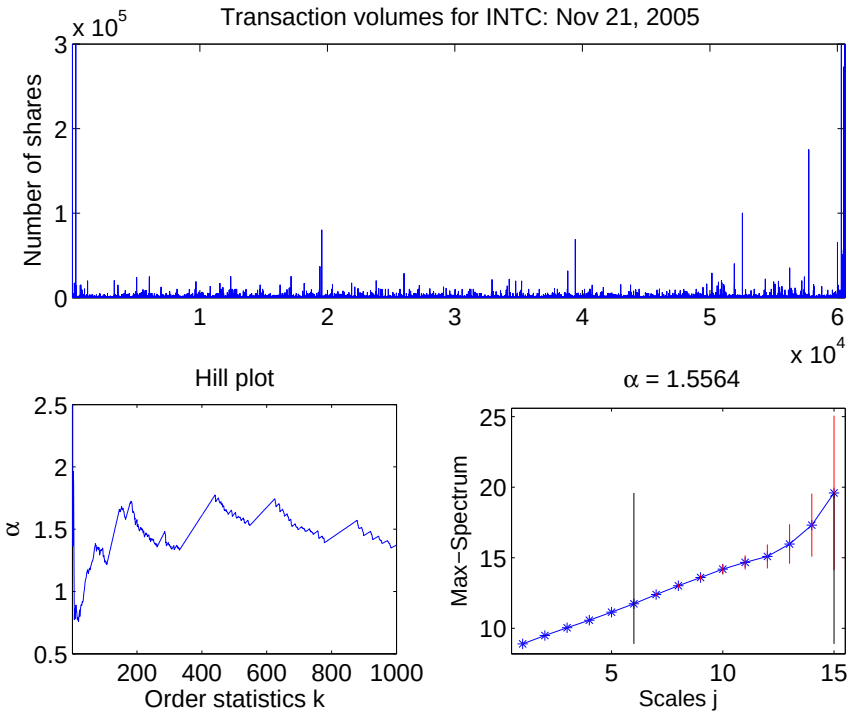

Fig. 15. This figure has the same format as Figure 14. The top plot shows the volumes of INTC during November 21, 2005, which as the volumes of GOOG in Figure 12, behave like a classical heavy–tailed sample. The Hill plot and the max–spectrum (bottom left and right panels, respectively) identify tail exponents around 1.5. The cut off scale in the max–spectrum plot was selected automatically with $p = 0.1$ and $b = 3$ (as in Figure 13. Notice the volatile, saw–tooth shape of the Hill plot which is due to its non–robustness to deviations from the Pareto model. The max–spectrum is more robust and fairly linear with a small knee on scale $j = 12$, which may be due to a few clusters of large volumes in the beginning and at the end of the trading day.

where $M_{2^{j+r}}$ is as in Condition E. Thus, by (13), the first term in the right–hand side (r.h.s.) of (30) vanishes, as $N_r = N/2^r \to \infty$. On the other hand, Condition C implies that the second term in the r.h.s. of (30) also vanishes, as $N/2^r \to \infty$.

We have thus shown that the first term (variance) in the r.h.s. of (29) tends to zero, as $N \to \infty$. Now, focus on the second term in the r.h.s. of (29). In view of (6), for all $j,\ j_1 \le j \le j_2$,

$$\mathbb{E}Y_{j+r} - (j+r)/\alpha - c_1 = \mathbb{E}\log_2\left(D(j+r,1)/2^{(j+r)/\alpha}\right) - c_1 = \mathbb{E}\log_2 M_{2^{(j+r)/\alpha}} - c_1,$$

which vanishes, as $r = r(N) \to \infty$, by Condition E. Therefore, the r.h.s. of (29) converges to zero, as $r \to \infty$ and $N/2^r \to \infty$ and the proof of the theorem is complete. □

The next three elementary results are used in the proof of Theorem 3.2.

Lemma 7.1. *Let $b_k \ge 0,\ k \in \mathbb{Z}$ and suppose $\sum_{k\in\mathbb{Z}} b_k < \infty$. Then, for any $n \in \mathbb{N}$,*

$$\bigvee_{k\in\mathbb{Z}} b_k \le \frac{1}{n}\sum_{i\in\mathbb{Z}} \bigvee_{i+1\le k\le i+n} b_k \le \sum_{k\in\mathbb{Z}} b_k. \tag{31}$$

PROOF. We have $b_{k^*} = \bigvee_{k\in\mathbb{Z}} b_k$, for some $k^* \in \mathbb{Z}$, since $\sum_{k\in\mathbb{Z}} b_k < \infty$. Thus, $\bigvee_{i+1\le k\le i+n} b_k = b_{k^*}$, for all $i = k^* - n, \ldots, k^* - 1$. Since at least n of the terms $\bigvee_{i+1\le k\le i+n} b_k$ are equal to the maximum $b_{k^*} = \bigvee_{k\in\mathbb{Z}} b_k$, we obtain that the first inequality in (31) holds. The second bound in (31) follows from the fact that $\bigvee_{i+1\le k\le i+n} b_k \le \sum_{i+1\le k\le i+n} b_k$. □

Lemma 7.2. *Let Z_1 and Z_2 be independent $\alpha-$Fréchet variables, $\alpha > 0$, with positive scale coefficients $\|Z_1\|_\alpha$ and $\|Z_2\|_\alpha$, respectively. Then, for all $q > 0$ and $\gamma \in (0, \alpha/q)$, there exists a constant $C_{q,\gamma}$, independent of $Z_i,\ i = 1, 2$, such that*

$$\mathbb{E}|\log_2(Z_1 \vee Z_2) - \log_2 Z_1|^q \le C_{q,\gamma}\Big(\|Z_2\|_\alpha/\|Z_1\|_\alpha\Big)^{q\gamma}. \tag{32}$$

PROOF. The expectation in (32) equals

$$\mathbb{E}|\log_2(1 \vee (Z_2/Z_1))|^q \le C_\gamma^q \mathbb{E}(Z_2^{q\gamma}/Z_1^{q\gamma}) = C_\gamma^q(\mathbb{E}Z_2^{q\gamma})(\mathbb{E}Z_1^{-q\gamma}), \tag{33}$$

for some constant C_γ, which is independent of Z_1 and Z_2. In the last relation, we used the inequality $0 \le \log_2(1 \vee x) \le C_\gamma x^\gamma$, valid for all $x,\ \gamma > 0$ and some $C_\gamma > 0$, and the independence of Z_1 and Z_2.

Now, observe that $\mathbb{E}(Z_2^{q\gamma}) = \|Z_2\|^{q\gamma}\mathbb{E}(Z^{q\gamma})$, and $\mathbb{E}(Z_1^{-q\alpha}) = \|Z_1\|^{-q\gamma}\mathbb{E}(Z^{-q\gamma})$, where Z is an $\alpha-$Fréchet random variable with unit scale coefficient. The expectation $\mathbb{E}(Z^{-q\alpha})$ is finite, for any $\alpha q > 0$, since $q\alpha \in (0, \alpha)$, the expectation $\mathbb{E}(Z^{q\alpha})$ is also finite. This, in view of (33), implies (32). □

Lemma 7.3. *For any two non–empty index sets $\mathcal{A}$ and $\mathcal{B}$, and real numbers $x_{m,n}$, $m \in \mathcal{A}$ and $n \in \mathcal{B}$, we have* $\sup_{m\in\mathcal{A}}\Big(\sup_{n\in\mathcal{B}} x_{m,n}\Big) = \sup_{(m,n)\in\mathcal{A}\times\mathcal{B}} x_{m,n}$.

This result follows directly from the definition of the supremum.

PROOF OF THEOREM 3.2: We first show Condition E. In view of (19), we have that M_n equals in distribution to $\|M_n\|_\alpha Z(1)$, where

$$\|M_n\|_\alpha^\alpha = \frac{1}{n}\sum_{i\in\mathbb{Z}} \bigvee_{1\le k\le n} a(k-i)^\alpha,$$

denotes the scale coefficient of the $\alpha-$Fréchet variable M_n.

Lemma 3.1, applied to $b_k := a(k)^\alpha,\ k \in \mathbb{Z}$, implies that $\|M_n\|_\alpha \to a^* := \bigvee_{k\in\mathbb{Z}} a(k)$, as $n \to \infty$. We therefore obtain

$$\mathbb{E}(\log_2 M_n)^p = \mathbb{E}(\log_2\|M_n\|_\alpha + \log_2 Z(1))^p \longrightarrow c_p := \mathbb{E}(\log_2 a^* + \log_2 Z(1))^p,\ \ p = 1, 2,$$

which implies Condition E.

The validation of Condition C is more involved. Observe that the sequences $D(j, n+1),\ n \in \mathbb{Z}$ in (14) are in fact a type of moving maxima processes. Indeed, in view of (5) and (17), we obtain

$$\begin{aligned} D(j, n+1) &= \bigvee_{i=1}^{2^j}\bigvee_{k\in\mathbb{Z}} a(k)Z(2^j n + i - k) \\ &= \bigvee_{k\in\mathbb{Z}}\Big(\bigvee_{i=1}^{2^j} a(k+i)\Big)Z(2^j n - k) =: \bigvee_{k\in\mathbb{Z}} b_j(k)Z(2^j n - k), \end{aligned} \tag{34}$$

where the second equality in (34) follows by making the change of variables $k := k - i$ and interchanging the two maxima (Lemma 7.3). Thus, we have that $D(j, n+1) = \widetilde{X}_j(2^j n)$, where $\widetilde{X}_j(k) := \bigvee_{i\in\mathbb{Z}} b_j(i)Z(k-i),\ k \in \mathbb{Z}$ is a moving maxima process as in (17). That is, the process $\{D(j,k)\}_{k\in\mathbb{Z}}$ is a *down–sampled version* of a moving maxima process.

Let now $m \in \mathbb{N}$ be an arbitrary parameter and consider the decomposition:

$$\frac{1}{2^j/\alpha} D(j, n+1) =: D_m(j, n+1) \vee \widetilde{D}_m(j, n+1), \;\; n \in \mathbb{Z}, \tag{35}$$

where

$$D_m(j, n+1) := \frac{1}{2^{j/\alpha}} \bigvee_{k \in \mathbb{Z}} \Big(\bigvee_{j=1}^{2^j} a(k+i) \Big) 1_{\{|k| \le m2^j\}} Z(2^j n - k) \tag{36}$$

and

$$\widetilde{D}_m(j, n+1) := \frac{1}{2^{j/\alpha}} \bigvee_{k \in \mathbb{Z}} \Big(\bigvee_{j=1}^{2^j} a(k+i) \Big) 1_{\{|k| > m2^j\}} Z(2^j n - k). \tag{37}$$

Note the presence of the normalization factor $2^{j/\alpha}$ in (35) as well as in (36) and (37).

We now give some intuition behind the decomposition (35). Observe that the indicator functions $1_{\{|k| \le m2^j\}}$ and $1_{\{|k| > m2^j\}}$ restrict the outer maxima in (36) and (37) to the sets of indices $\{k \in \mathbb{Z} : \; |k| \le m2^j\}$ and $\{k \in \mathbb{Z} : \; |k| > m2^j\}$, respectively. Also, as argued above, the sequences $D_m(j, n+1), \; n \in \mathbb{Z}$ and $\widetilde{D}_m(j, n+1), \; n \in \mathbb{Z}$ are down–sampled moving maxima. Note, however that, for all $j \in \mathbb{N}$, $\{D_m(j, n+1)\}_{n \in \mathbb{Z}}$ is a $2m-$dependent sequence, and therefore its covariances at lags greater than $2m$ vanish. We therefore proceed by showing that the covariances in (14) can be bounded in terms of the $D_m(j, k)$'s plus an asymptotically negligible contribution due to the $\widetilde{D}_m(j, k)$'s.

We start by deriving several inequalities for the scale coefficients $\|D(j,1)/2^{j/\alpha}\|_\alpha$, $\|D_m(j,1)\|_\alpha$ and $\|\widetilde{D}_m(j,1)\|_\alpha$. Observe that by (34),

$$\|D(j,1)/2^{j/\alpha}\|_\alpha^\alpha = \frac{1}{2^j} \Big\| \bigvee_{k \in \mathbb{Z}} \Big(\bigvee_{i=1}^{2^j} a(k+i) \Big) Z(-k) \Big\|_\alpha^\alpha = \sum_{k \in \mathbb{Z}} \frac{1}{2^j} \bigvee_{i=1}^{2^j} a(k+i)^\alpha, \tag{38}$$

where in the last equality we used the fact that the $Z(k)$'s are i.i.d. standard $\alpha-$Fréchet variables (see (18)).

Now, by applying Lemma 7.1 to the r.h.s. of (38), we obtain

$$\bigvee_{k \in \mathbb{Z}} a(k)^\alpha \le \|D(j,1)/2^{j/\alpha}\|_\alpha^\alpha \le \sum_{k \in \mathbb{Z}} a(k)^\alpha. \tag{39}$$

As in (38), we also have that

$$\|D_m(j,1)\|_\alpha^\alpha = \sum_{k \in \mathbb{Z}} \frac{1}{2^j} \bigvee_{i=1}^{2^j} a(k+i)^\alpha 1_{\{|k| \le m2^j\}} \;\text{ and }\; \|\widetilde{D}_m(j,1)\|_\alpha^\alpha = \sum_{k \in \mathbb{Z}} \frac{1}{2^j} \bigvee_{i=1}^{2^j} a(k+i)^\alpha 1_{\{|k| > m2^j\}}.$$

By using that $|k+i| \le (m-1)2^j \le (m-1)$ implies $|k| \le m2^j$, for all $i = 1, \ldots, 2^j, \; j \in \mathbb{N}$, we obtain

$$\|D_m(j,1)\|_\alpha^\alpha \ge \sum_{k \in \mathbb{Z}} \frac{1}{2^j} \bigvee_{i=1}^{2^j} a(k+i)^\alpha 1_{\{|k+i| \le (m-1)2^j\}} \ge \bigvee_{k \in \mathbb{Z}} a(k)^\alpha 1_{\{|k| \le m-1\}}, \tag{40}$$

where the last inequality follows from Lemma 7.1. By Lemma 7.1, as in (39), we also have

$$\|D_m(j,1)\|_\alpha^\alpha \le \sum_{k \in \mathbb{Z}} a(k)^\alpha. \tag{41}$$

Since $|k| > m2^j$ implies $|k+i| > (m-1)2^j \ge (m-1)$, for all $i = 1, \ldots, 2^j$, by Lemma 7.1,

$$\|\widetilde{D}_m(j,1)\|_\alpha^\alpha \le \sum_{k \in \mathbb{Z}} \frac{1}{2^j} \bigvee_{i=1}^{2^j} a(k+i)^\alpha 1_{\{|k+i| > (m-1)2^j\}} \le \sum_{k \in \mathbb{Z}} a(k)^\alpha 1_{\{|k| > m-1\}}. \tag{42}$$

Consider now the covariances in (14) and observe that they are equal to

$$\mathrm{Cov}\Big(\log_2(D(j,n+1)/2^{j/\alpha}), \log_2(D(j,1)/2^{j/\alpha})\Big) =: \mathbb{E}\xi(j)\eta(j) - \mathbb{E}\xi(j)\mathbb{E}\eta(j),$$

where, for brevity, we let $\xi(j) := \log_2(D(j,n+1)/2^{j/\alpha})$, and $\eta(j) := \log_2(D(j,1)/2^{j/\alpha})$. Let also $\xi_m(j) := \log_2 D_m(j,n+1)$ and $\eta_m(j) := \log_2 D_m(j,1)$ and observe that the triangle and the Cauchy–Schwartz inequalities, imply

$$\begin{aligned} |\mathbb{E}\xi(j)\eta(j) - \mathbb{E}\xi_m(j)\eta_m(j)| &\le \mathbb{E}|\xi_m(j)(\eta(j)-\eta_m(j))| + \mathbb{E}|\eta(j)(\xi(j)-\xi_m(j))| \\ &\le (\mathbb{E}\xi_m(j)^2)^{1/2}\Big(\mathbb{E}(\eta(j)-\eta_m(j))^2\Big)^{1/2} + (\mathbb{E}\eta(j)^2)^{1/2}\Big(\mathbb{E}(\xi(j)-\xi_m(j))^2\Big)^{1/2}. \end{aligned} \tag{43}$$

Note that, by stationarity, $D_m(j,1) =^d D_m(j,n+1) =^d \|D_m(j,1)\|_\alpha Z(1)$, and hence

$$\mathbb{E}\xi_m(j)^2 = \mathbb{E}\eta_m(j)^2 = \mathbb{E}\Big(\log_2 \|D_m(j,1)\|_\alpha + \log_2 Z(1)\Big)^2.$$

We similarly have that

$$\mathbb{E}\xi(j)^2 = \mathbb{E}\eta(j)^2 = \mathbb{E}\Big(\log_2 \|D(j,1)/2^{j/\alpha}\|_\alpha + \log_2 Z(1)\Big)^2.$$

Therefore, the inequalities (39), (40) and (41) imply that $\sup_{j\in\mathbb{N}} \mathbb{E}\xi_m(j)^2 = \sup_{j\in\mathbb{N}} \mathbb{E}\eta_m(j)^2$ and $\sup_{j\in\mathbb{N}} \mathbb{E}\xi(j)^2 = \sup_{j\in\mathbb{N}} \mathbb{E}\eta(j)^2$ are finite, for all sufficiently large m. Indeed, this follows from the inequality $|\log_2(x)| \le \log_2(e)(x+1/x)$, valid for all $x>0$. Thus, by (43), for all sufficiently large m, we have

$$\begin{aligned} \sup_{j\in\mathbb{N}} |\mathbb{E}\xi(j)\eta(j) - \mathbb{E}\xi_m(j)\eta_m(j)| &\le \text{const} \sup_{j\in\mathbb{N}} \Big(\mathbb{E}(\xi(j)-\xi_m(j))^2\Big)^{1/2} \\ &= \text{const} \sup_{j\in\mathbb{N}} \Big(\mathbb{E}(\log_2(D(j,1)/2^{j/\alpha}) - \log_2 D_m(j,1))\Big)^{1/2}. \end{aligned} \tag{44}$$

Note that $D_m(j,1)$ and $\widetilde{D}_m(j,1)$ are independent since they involve maxima of non–overlapping ranges of $Z(k)$'s (see (36) and (37)). Thus, in view of (35), Lemma 7.2, applied to the r.h.s. of (44) implies that

$$\begin{aligned} \sup_{j\in\mathbb{N}} |\mathbb{E}\xi(j)\eta(j) - \mathbb{E}\xi_m(j)\eta_m(j)| &\le \text{const} \sup_{j\in\mathbb{N}} \Big(\frac{\|\widetilde{D}_m(j,1)\|_\alpha}{\|D_m(j,1)\|_\alpha}\Big)^{\alpha/4} \\ &\le \text{const} \Big(\frac{\sum_{k\in\mathbb{Z}} a(k)^\alpha 1_{\{|k|>(m-1)\}}}{\bigvee_{k\in\mathbb{Z}} a(k)^\alpha 1_{\{|k|\le(m-1)\}}}\Big)^{\alpha/4} =: \text{const}\, R(m), \end{aligned} \tag{45}$$

where the last inequality follows from (40).

Similarly, by using the triangle inequality and the fact that $\mathbb{E}|X| \le (\mathbb{E}X^2)^{1/2}$, we get

$$\sup_{j\in\mathbb{N}} |\mathbb{E}\xi(j)\mathbb{E}\eta(j) - \mathbb{E}\xi_m(j)\mathbb{E}\eta_m(j)| \le \text{const}\, R(m), \tag{46}$$

for all sufficiently large $m\in\mathbb{N}$, where $R(m)$ is as in (45). Now, by combining relations (45) and (46), we obtain that

$$\begin{aligned} \sup_{j\in\mathbb{N}} |\mathrm{Cov}(\xi(j),\eta(j))| &\le \sup_{j\in\mathbb{N}} |\mathrm{Cov}(\xi_m(j),\eta_m(j))| + \text{const}\, R(m) \\ &= \sup_{j\in\mathbb{N}} |\mathrm{Cov}(\log_2 D_m(j,n+1), \log_2 D_m(j,1))| + \text{const}\, R(m). \end{aligned} \tag{47}$$

Since $R(m) = \Big((\sum_{|k|>(m-1)} a(k)^\alpha)/(\bigvee_{|k|\le(m-1)} a(k)^\alpha)\Big)^{\alpha/4} \to 0$, as $m\to\infty$, one can make the second term in the r.h.s. of (47) arbitrarily small, for all sufficiently large m's. However, since $D_m(j,1)$ and $D_m(j,n+1)$ are independent, for all $n>2m$, and for any $j\in\mathbb{N}$, the first term in the r.h.s. of (47) vanishes for all $n>2m$. This implies that (14) holds and completes the proof of the theorem. □

References

Adamic, L. and B. Huberman (2000). The nature of markets in the world wide web. *Quarterly Journal of Electronic Commerce 1*, 5–12.

Adamic, L. and B. Huberman (2002). Zipf's power law and the Internet. *Glottometrics 3*, 143–150.

Adler, R., R. Feldman, and M. S. Taqqu (Eds.) (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston: Birkhäuser.

Ancona-Navarrete, M. and J. Tawn (2000). A comparison of methods for estimating the extremal index. *Extremes 3*(1), 5–38 (2001).

Barabasi, A.-L. (2002). *Linked: The New Science of Networks*. Cambridge, MA, USA: Perseus Publishing.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Carlson, J. M. and J. Doyle (1999). Highly optimized tolerance: a mechanism for power laws in designed systems. *Physical Review E 60*(2), 1412–1427.

Cheng, S. and L. Peng (2001). Confidence intervals for the tail index. *Bernoulli 7*(5), 751–760.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2002). *Introduction to Algorithms*. McGraw–Hill.

Crovella, M. E. and A. Bestavros (1996, May). Self-similarity in World Wide Web traffic: evidence and possible causes. In *Proceedings of the 1996 ACM SIGMETRICS. International Conference on Measurement and Modeling of Complex Systems*, pp. 160–169.

Crovella, M. E. and M. S. Taqqu (1999). Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability 1*, 55–79.

Crovella, M. E., M. S. Taqqu, and A. Bestavros (1998). Heavy-tailed probability distributions in the World Wide Web. In R. Adler, R. Feldman, and M. S. Taqqu (Eds.), *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Boston, pp. 3–25. Birkhäuser.

Csörgő, S., P. Deheuvels, and D. Mason (1985). Kernel estimates of the tail index of a distribution. *Annals of Statistics 13*(3), 1050–1077.

Davis, R. A. and S. I. Resnick (1993). Prediction of stationary max–stable processes. *The Annals of Applied Probability 3*(2), 497–525.

de Haan, L. (1984). A spectral representation for max–stable processes. *Annals of Probability 12*(4), 1194–1204.

de Haan, L., H. Drees, and S. Resnick (2000). How to make a Hill plot. *Annals of Statistics 28*(1), 254–274.

de Sousa, B. and G. Michailidis (2004). A diagnostic plot for estimating the tail index of a distribution. *Journal of Computational and Graphical Statistics 13*(4), 974–995.

Deheuvels, P. (1983). Point processes and multivariate extreme values. *Journal of Multivariate Analysis 13*(2), 257–272.

Faloutsos, M., P. Faloutsos, and C. Faloutsos (1999). On power–law relationships of the Internet topology. In *SIGCOMM*, pp. 251–262.

Feuerverger, A. and P. Hall (1999). Estimating a tail exponent by modeling departure from a Pareto distribution. *Ann. Statist. 27*(2), 760–781.

Finkenstädt, B. and H. Rootzén (Eds.) (2004). *Extreme Values in Finance, Telecommunications, and the Environment*, Volume 99 of *Monographs on Statistics and Applied Probability*. New York: Chapman and Hall / CRC.

Gilbert, A., Y. Kotidis, S. Muthukrishnan, and M. Strauss (2001). Surfing wavelets on streams: one-pass summaries for approximate aggregate queries. In *Procedings of VLDB, Rome, Italy*.

Henzinger, M., P. Raghavan, and S. Rajagopalan (1998). Computing on data streams.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics 3*, 1163–1174.

Hong, H. and J. Wang (2000). Trading and returns under periodic market closures. *Journal of Finance 55*, 297–354.

Kratz, M. and S. I. Resnick (1996). The qq-estimator and heavy tails. *Stochastic Models 12*, 699–724.

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

Lo, A. and J. Wang (2000). Trading volume: Definitions, data analysis, and implications of portfolio theory. *Review of Financial Studies 13*, 257–300.

Lu, J.-C. and L. Peng (2002). Likelihood based confidence intervals for the tail index. *Extremes 5*(4), 337–352 (2003).

Mandelbrot, B. B. (1960). The Pareto-Lévy law and the distribution of income. *International Economic Review 1*, 79–106.

Muthukrishnan, S. (2003). Data streams: algorithms and applications. Preprint. `http://www.cs.rutgers.edu/~muthu/`.

Park, K., G. Kim, and M. E. Crovella (1996, October). On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proceedings of the Fourth International Conference on Networks Protocols (ICNP'96)*. To appear.

Park, K. and W. Willinger (Eds.) (2000). *Self-Similar Network Traffic and Performance Evaluation*. New York: J. Wiley & Sons, Inc.

Resnick, S. and C. Stărică (1995). Consistency of Hill's estimator for dependent data. *Journal of Applied Probability 32*, 139–167.

Resnick, S. and C. Stărică (1997). Smoothing the Hill estimator. *Adv. in Appl. Probab. 29*(1), 271–293.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. New York: Springer-Verlag.

Resnick, S. I. (1997). Heavy tail modeling and teletraffic data. *The Annals of Statistics 25*, 1805–1869. With discussions and rejoinder.

Smith, R. L. and I. Weissman (1996). Characterization and estimation of the multivariate extremal index. Preprint, `http://www.unc.edu/depts/statistics/postscript/rs/extremal.pdf`.

Stoev, S., G. Michailidis, and M. Taqqu (2006). Estimating heavy–tail exponents through max self–similarity. Technical Report 445, Department of Statistics, University of Michigan, Ann Arbor.

Stoev, S. and M. S. Taqqu (2006). Extremal stochastic integrals: a parallel between max–stable processes and $\alpha-$stable processes. To appear in *Extremes*.

Taqqu, M. S., W. Willinger, and R. Sherman (1997). Proof of a fundamental result in self-similar traffic modeling. *Computer Communications Review 27*(2), 5–23.

Tsay, R. (2005). *Analysis of Financial Time Series* (second ed.). Hoboken, New Jersey, U.S.A.: John Wiley & Sons.

Wharton Research Data Service (url). `https://wrds.wharton.upenn.edu/`. Wharton School of Management, Universty of Pennsylvania.

Zhang, Z. and R. Smith (2004). The behavior of multivariate maxima of moving maxima processes. *Journal of Applied Probability 41*(4), 1113–1123.

Zipf, G. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.