Aust. N. Z. J. Stat. **50**(1), 2008, 1–11 doi: 10.1111/j.1467-842X.2007.00496.x

A FINE-TUNED ESTIMATOR OF A GENERAL CONVERGENCE RATE

TUCKER MCELROY[1] AND DIMITRIS N. POLITIS[2*]

US Census Bureau and University of California at San Diego

Summary

A general rate estimation method based on the in-sample evolution of appropriately chosen diverging/converging statistics has recently been proposed by D.N. Politis [*C. R. Acad. Sci. Paris, Ser. I*, vol. 335, pp. 279–282, 2002] and T. McElroy & D.N. Politis [*Ann. Statist.*, vol. 35, pp. 1827–1848, 2007]. In this paper, we show how a modification of the original estimators achieves a competitive rate of convergence. The modified estimators require the choice of a tuning parameter; an optimal such choice is generally a non-trivial problem in practice. Some discussion to that effect is given, as well as a small simulation study in a heavy-tailed setting.

Key words: heavy-tail index; rate of convergence; stationary sequence; time series.

1. Introduction

Let $X_1, \ldots, X_n$ be an observed stretch from a strictly stationary, weakly dependent time series $\{X_t\}$. A number of converging and/or diverging statistics can be computed from a dataset of this type. In many instances, however, the rate of convergence/divergence of some statistic of interest may be unknown; that is, it may depend on some unknown feature of the underlying probability law P.

For each given context, i.e. choice of statistic and assumptions on the time series $\{X_t\}$, a context-specific rate estimator can be devised and its properties analysed. By contrast, a general approach for rate estimation that is not context-dependent has been recently proposed by Politis (2002) and McElroy & Politis (2007). In this paper, we propose a modification of the original estimators that achieves a competitive rate of convergence at the expense of having to choose a tuning parameter; an optimal such choice is, however, a nontrivial problem in practice.

In the next section, the rate estimation methodology of Politis (2002) and McElroy & Politis (2007) is briefly reviewed, the new fine-tuned estimator is introduced, and its properties are analysed; some discussion on the choice of tuning parameter is also given. In Section 3, a specific example of interest is described in detail, namely the estimation of the heavy-tail index with data from a linear time series model. In that setting, the results of a small simulation study are also presented.

*Author to whom correspondence should be addressed.

[1]Statistical Research Division, US Census Bureau, 4700 Silver Hill Road, Washington, DC 20233-9100, USA.

[2]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA.
e-mail: politis@euclid.ucsd.edu

2. Rate estimation based on diverging/converging statistics

2.1. Scanning a sequence

The notion of 'scanning' a sequence $X_1, \ldots, X_n$ was introduced in McElroy & Politis (2007), and is reviewed below.

Definition 2.1 *A scan is a collection of n block-subsamples of the sequence $X_1, \ldots, X_n$ with the following two properties: (a) within each scan there is a single block of each size $k = 1, \ldots, n$; and (b) those n blocks are nested, i.e. the block of size k_1 can be found as a stretch within the block of size k_2 when $k_1 \leq k_2$.*

As usual, a block-subsample of the sequence $X_1, \ldots, X_n$ is a block of consecutive observations, that is, a set of the type $X_j, X_{j+1}, \ldots, X_{j+m}$; see, for example, Politis, Romano & Wolf (1999).

We will therefore say that the sequence $X_1, \ldots, X_n$ has been *scanned* if a block corresponding to each block size $k = 1, \ldots, n$ has been extracted, and if those blocks are nested; that is, the block of size k_1 can be found as a stretch within the block of size k_2 when $k_1 \leq k_2$. For example, consider the '*direct*' scan

$$(X_1), (X_1, X_2), (X_1, X_2, X_3), \ldots, (X_1, \ldots, X_{n-1}), (X_1, \ldots, X_n),$$

and the '*reverse*' scan

$$(X_n), (X_{n-1}, X_n), (X_{n-2}, X_{n-1}, X_n), \ldots, (X_2, \ldots, X_n), (X_1, \ldots, X_n).$$

In general, a scan will start at time-point j (say) and then the blocks will proceed growing/expanding to the left and/or to the right. For example, a valid scan is

$$(X_5), (X_4, X_5), (X_3, X_4, X_5), (X_3, X_4, X_5, X_6), \ldots, (X_1, \ldots, X_n);$$

note how within each block the natural time order is preserved, and how all scans end with the block containing the full data set. The number of possible scans is large. In fact, there are 2^{n-1} different scans of the sequence $X_1, \ldots, X_n$ when no ties are present; see McElroy & Politis (2007) and www.math.ucsd.edu/∼politis/PAPER/scansAlgorithms.pdf, where a number of algorithms for generating randomly chosen scans are presented.

2.2. Basic rate estimation methodology based on scans

We outline below the basic rate estimation method of Politis (2002) and McElroy & Politis (2007), and review its key properties. To do this, let $T_n = T_n(X_1, \ldots, X_n)$ be some positive statistic whose rate of convergence/divergence depends on some unknown real-valued parameter λ. Similarly, let $T_k = T_k(X_1, \ldots, X_k)$ for $k = 1, \ldots, n$.

Assumption A1. *Assume that, for some slowly varying function $L(n)$ and for some known invertible function $g(\cdot)$ that is continuous over an interval that contains λ, we have $U_n = O_P(1)$ as $n \to \infty$, where*

$$U_k = \log\left(k^{-g(\lambda)} L(k) T_k\right) \text{ for } k = 1, \ldots, n. \tag{1}$$

The following two conditions will also be found useful:

$$U_n \xrightarrow{\mathcal{L}} \text{some r.v. } U, \text{ with } \mathrm{E}\big(U_n^2\big) \to \mathrm{E}(U^2), \text{ as } n \to \infty \tag{2}$$

and

$$\mathrm{E}(U_n) - \mathrm{E}(U) = O(n^{-p}), \text{ for some } p > 0. \tag{3}$$

Consider all the 2^{n-1} different scans of the sequence $X_1, \ldots, X_n$; order the scans in some arbitrary fashion, focus on the Ith such scan, and let $T_k^{(I)}$ denote the value of the statistic T_k as computed from the block of size k of the Ith scan of the sequence $X_1, \ldots, X_n$. Politis (2002) and McElroy & Politis (2007) suggest estimating λ by $\hat{\lambda}^{(I)} = g^{-1}(\hat{g}^{(I)})$, where

$$\hat{g}^{(I)} = \frac{\sum_{k=1}^{n}(Y_k - \bar{Y})(\log k - \overline{\log n})}{\sum_{k=1}^{n}(\log k - \overline{\log n})^2},$$

and $Y_k = \log T_k^{(I)}$ for $k = 1, \ldots, n$; $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k$; and $\overline{\log n} = \frac{1}{n}\sum_{k=1}^{n} \log k$.

The following theorem now ensues on the properties of the general rate estimator $\hat{\lambda}^{(I)}$.

Theorem 2.1 [McElroy & Politis, 2007] *Assume Assumption A1.*
(i) If condition (2) holds, then $\mathrm{E}(\hat{g}^{(I)}) \to g(\lambda)$ *and* $\mathrm{var}(\hat{g}^{(I)}) = O(1)$ *as* $n \to \infty$.
(ii) If condition (3) holds, then $\mathrm{E}(\hat{g}^{(I)}) = g(\lambda) + A_2 + O(n^{-p}\log n)$, *where*

$$A_2 = -\frac{\sum_{k=1}^{n}(\log L(k) - \overline{\log L})(\log k - \overline{\log n})}{\sum_{k=1}^{n}(\log k - \overline{\log n})^2}.$$

Theorem 2.1 suggests an approach for potentially improving the estimators $\hat{\lambda}^{(I)}$ by combining/averaging the estimators based on scans. Consider the estimators $\hat{\lambda}^{(1)}, \ldots, \hat{\lambda}^{(N)}$ for some integer N, and define $\hat{\lambda}^{*} = N^{-1}\sum_{i=1}^{N} \hat{\lambda}^{(i)}$. A different way of combining estimators is given by the median; so, we also define $\hat{\lambda}^{\sharp} = \mathit{median}\{\hat{\lambda}^{(I)}, \ldots, \hat{\lambda}^{(N)}\}$. The median estimator $\hat{\lambda}^{\sharp}$ will exhibit a variance reduction behaviour similar to that of the mean estimator $\hat{\lambda}^{*}$. However, the median may be preferable in practice because of its robustness.

The idea of combining estimators that are computed over subsets of the data is in the spirit of the notion of 'bagging' of Breiman (1996). It is easy to see that averaging over scans 'does not hurt' in the sense that both $\hat{\lambda}^{*}$ and $\hat{\lambda}^{\sharp}$ will be at least as accurate as any given $\hat{\lambda}^{(I)}$. Although it is very difficult to generally quantify the variance reduction effect of scanning estimators, the simulations in McElroy & Politis (2007) show a very spectacular effect even with a small value of N. Note that N is really tied to the practitioner's computational facilities, and not so much to the sample size n or the number of scans, 2^{n-1}. The recommendation is to take N as big as computationally feasible; in practice, however, even taking N as small as 100 should give a significant benefit, especially if the N scans under consideration are very different from one another. A way to ensure this is to use N randomly selected scans from an algorithm that gives (close to) equal weight to each scan. A collection of algorithms to generate randomly selected scans can be found at www.math.ucsd.edu/∼politis/PAPER/scansAlgorithms.pdf, where some properties of those algorithms are also discussed.

2.3. A fine-tuned estimator

From the discussion in Section 2.2, it is apparent that the estimator $\hat{\lambda}^{(I)}$ is based on a log–log least squares (LS) regression in a plot of the evolution of $T_k^{(I)}$ vs. the index k; here, $T_k^{(I)}$ is our statistic computed from the block of size k of the Ith scan of the sequence $X_1, \ldots, X_n$. It follows that each scan of the sequence $X_1, \ldots, X_n$ corresponds to a different scatterplot, and therefore to different slope estimators. Although all scans have the same final point, namely $T_n = T_n(X_1, \ldots, X_n)$, the paths leading to that point can be very different.

As suggested by Theorem 2.1, the estimator $\hat{\lambda}^{(I)}$ has very small bias under some conditions, but may benefit from a variance reduction technique. To elaborate, let us assume that L is constant, which implies that $A_2 = 0$ in Theorem 2.1. Under conditions (2) and (3), Theorem 2.1 implies that

$$\mathrm{E}\big(\hat{g}^{(I)}\big) = g(\lambda) + O(n^{-p} \log n) \text{ and } \mathrm{var}\big(\hat{g}^{(I)}\big) = O(1). \tag{4}$$

In other words, the bias of $\hat{g}^{(I)}$ decays polynomially but the variance is only bounded. The methodology of scanning has as its goal to reduce this variance. The improvement resulting from the effect of scanning, namely the improvement that $\hat{\lambda}^*$ and $\hat{\lambda}^\sharp$ have to offer over each of the $\hat{\lambda}^{(I)}$, would be quantifiable if we could identify some independent (or approximately independent) scans.

To make this variance reduction phenomenon more explicit, we have to resort to a sample-size trick; its purpose is to free the scanned scatterplots from the obligation to have the same final point, which would result in the scans being dependent. Thus let $b = [c_b n^\zeta]$, where ζ is some fixed constant in (0, 1), c_b is some positive constant, and $[\cdot]$ denotes the integer part. Let $\hat{g}_b^{(I)}$ be the LS regression (with intercept) slope estimator that uses only block sizes $k = 1, \ldots, b$ from the Ith scan. In other words, let

$$\hat{g}_b^{(I)} = \frac{\sum_{k=1}^b (Y_k - \bar{Y}_b)(\log k - \overline{\log b})}{\sum_{k=1}^b (\log k - \overline{\log b})^2};$$

here, $Y_k = \log T_k^{(I)}$ as in Section 2.2, but now $\bar{Y}_b = \frac{1}{b}\sum_{k=1}^b Y_k$, and $\overline{\log b} = \frac{1}{b}\sum_{k=1}^b \log k$. As $b = [c_b n^\zeta]$, from (4) it follows that

$$\mathrm{E}\big(\hat{g}_b^{(I)}\big) = g(\lambda) + O(n^{-p\zeta} \log n) \text{ and } \mathrm{var}\big(\hat{g}_b^{(I)}\big) = O(1). \tag{5}$$

Now consider the case for which the stationary time series $\{X_t\}$ is m-dependent for some non-negative integer m. For example, $\{X_t\}$ may be based on an underlying i.i.d. sequence $\{Z_t\}$ via a function of the type $X_t = \phi(Z_t, Z_{t-1}, \ldots, Z_{t-m})$; the case $m = 0$ is the special case of the sequence $\{X_t\}$ being independant and identically distributed (i.i.d.). Let $scan_1$, $scan_2, \ldots, scan_Q$ with $Q = [n/(b+m)]$ be scans satisfying the following: for each $i = 1, \ldots, Q$, the block of size b in $scan_i$ is the block $(X_{(i-1)(b+m)+1}, \ldots X_{(i-1)(b+m)+b})$. As a matter of fact, as only blocks up to size b from each scan will be used by estimator $\hat{g}_b$, generating $scan_i$ is equivalent to treating the block $(X_{(i-1)(b+m)+1}, \ldots X_{(i-1)(b+m)+b})$ as if it were the whole dataset, and using any of the algorithms of McElroy & Politis (2007) to generate scans; see www.math.ucsd.edu/∼politis/PAPER/scansAlgorithms.pdf.

It is apparent that, as a result of the m-dependence assumption, all blocks of size smaller than or equal to b found in $scan_i$ are independent of all blocks of size smaller than or equal to b found in $scan_j$ for $i \neq j$. Therefore, the statistics $\hat{g}_b^{(1)}, \ldots, \hat{g}_b^{(Q)}$ are independent. Furthermore,

because the *same* algorithm (random or deterministic) is used to generate each scan, then $\hat{g}_b^{(1)}, \ldots, \hat{g}_b^{(Q)}$ are i.i.d. As in the previous subsection, to obtain an improved estimator by means of scanning we define

$$\hat{g}_b^* = Q^{-1} \sum_{i=1}^{Q} \hat{g}_b^{(i)} \text{ and } \hat{g}_b^\sharp = \text{median}\{\hat{g}_b^{(1)}, \ldots, \hat{g}_b^{(Q)}\}.$$

The corresponding estimates of λ are defined by $\hat{\lambda}_b^* = g^{-1}(\hat{g}_b^*)$ and $\hat{\lambda}_b^\sharp = g^{-1}(\hat{g}_b^\sharp)$.

From the above discussion, we have essentially proved the following.

Theorem 2.2 *Let the time series $\{X_t\}$ be strictly stationary and m-dependent for some non-negative integer m. Assume assumption A1, conditions (2) and (3), and that the slowly varying function $L(k)$ is constant. Then*

$$\text{bias}(\hat{\lambda}_b^*) = O\left(\frac{\log n}{n^{p\zeta}}\right) \text{ and } \text{var}(\hat{\lambda}_b^*) = O(1/Q),$$

where $Q = n/(b+m)$ and $b = [c_b n^\zeta]$.

In particular, Theorem 2.2 shows that scanning *works* in reducing the variance, and has as a result a polynomial rate of convergence of the scanned mean estimator $\hat{\lambda}_b^*$ under the aforementioned assumptions. As $\hat{g}_b^\sharp$ is the sample median of the i.i.d. (triangular array-type) random variables $\hat{g}_b^{(1)}, \ldots, \hat{g}_b^{(Q)}$, it is expected that a similar result will hold for the scanned median estimator $\hat{\lambda}_b^\sharp$; that is, that

$$\lambda_b^\sharp = \lambda + O\left(\frac{\log n}{n^{p\zeta}}\right) + O_P(1/\sqrt{Q}), \tag{6}$$

under a few extra assumptions controlling the behaviour of the common distribution of $\hat{g}_b^{(1)}, \ldots, \hat{g}_b^{(Q)}$ in a neighbourhood of $g(\lambda)$. See, for example, the results of Mizera & Wellner (1998). Although these extra assumptions are typically very weak, they are unfortunately very difficult to verify in any practical set-up; thus, we will not pursue this theoretical approach any further here.

Remark 2.1 Theorem 2.2 implies that the mean squared error (MSE) of estimator $\hat{\lambda}_b^*$ satisfies

$$\text{MSE}(\hat{\lambda}_b^*) = O\left(\frac{\log^2 n}{n^{2p\zeta}}\right) + O(b/n).$$

As $b = [c_b n^\zeta]$, it follows that letting $\zeta = 1/(2p+1)$ minimizes – up to a logarithmic term – the asymptotic order of $\text{MSE}(\hat{\lambda}_b^*)$, and thus optimizes the large-sample accuracy of the scanned estimators. In this sense, ζ is a tuning parameter whose choice affects and fine-tunes the accuracy of $\hat{\lambda}_b^*$ (and $\hat{\lambda}_b^\sharp$). Using the aforementioned optimal value of ζ yields

$$\hat{\lambda}_b^* = \lambda + O_P\left(n^{-p/(2p+1)} \log n\right).$$

In other words, there is a polynomial rate of convergence for $\hat{\lambda}_b^*$, with a similar result expected for $\hat{\lambda}_b^\sharp$.

Remark 2.2 There seems no reason – other than for convenience – to limit ourselves to the particular independent scans $scan_1^*, scan_2^*, \ldots, scan_Q^*$; taking a large number of scans

will ensure that most of the independent scans will be included in the average (and in the right proportions). Similarly, the m-dependence assumption is just for convenience: an appropriate weak dependence, for example strong mixing, condition can take its place, and still the variance reduction effect of scanning will be of the same order of magnitude. This is manifested in our empirical results in the next section.

Remark 2.3 It is apparent that b here plays the role of a 'bandwidth' parameter influencing the rate of convergence of $\hat{\lambda}_b^*$ and $\hat{\lambda}_b^\sharp$. Optimally choosing such a 'bandwidth' parameter is always a challenging problem. We now give some very general guidelines in that direction, bearing in mind that context-specific estimators may also exist and might be preferable. To start with, recall that (5) implies that $\mathrm{E}(\hat{g}_b^{(i)}) = g(\lambda) + O(b^{-p}\log b)$ and $\mathrm{var}(\hat{g}_b^{(i)}) = O(1)$ for all $i = 1, 2, \ldots, Q$ corresponding to any given set of scans such as $scan_1, scan_2, \ldots, scan_Q$. We can then loosely re-write the above as

$$\hat{g}_b^{(i)} \simeq g(\lambda) + C\frac{\log b}{b^p} + \text{error}, \tag{7}$$

where C is some unknown constant (not necessarily positive). Equation (7) can be interpreted as a (non-linear) regression of $\hat{g}_b^{(i)}$ on b with unknown parameters: $g(\lambda)$, C and p. Because the 'responses' $\hat{g}_b^{(i)}$ are available for $i = 1, 2, \ldots, Q$ and for b taking a value in any chosen range, say among some values $\{b_1, b_2, \ldots, b_K\}$ for some finite K, the non-linear regression (7) can be fitted for $i = 1, 2, \ldots, Q$ and $b \in \{b_1, b_2, \ldots, b_K\}$. The result would yield a (rough) estimate of p to be used in connection with the optimal choice of rate of b, i.e. $b = [c_b n^\zeta]$ with $\zeta = 1/(2p + 1)$. Note that the above does not constitute a complete solution to the problem of choosing b, as the optimal proportionality constant c_b appears intractable in this general setting; nevertheless, it may provide some rough guidance in this difficult problem.

Remark 2.4 To ensure that the bias of $\hat{\lambda}$ decays polynomially fast in Theorem 2.2, we had to assume that the slowly varying function $L(k)$ is constant, which is not always the case. Actually, Theorem 2.2 would hold true even if L were not constant, as long as $A_2 = O(n^{-p}\log n)$, where the quantity A_2 is as defined in Theorem 2.1. If even that fails, our estimator must be modified. To outline such a modification, consider the less restrictive assumption that

$$L(k) = \delta_1(\log k)^{\delta_2} \quad \text{for some numbers } \delta_1 > 0,\ \delta_2 \in \mathbb{R}. \tag{8}$$

Let $Y_k = \log T_k^{(I)}$ as computed from the Ith scan, and consider the identity

$$Y_k = g(\lambda)\log k + U_k - \log L(k) \tag{9}$$

for $k = 1, \ldots, n$. Using assumption (8) we have

$$\begin{aligned} Y_k &= g(\lambda)\log k + U_k - \log L(k) \\ &= g(\lambda)\log k + (U_k - \mathrm{E}(U)) + (\mathrm{E}(U) - \log\delta_1) - \delta_2 \log\log k \end{aligned}$$

for $k = 1, \ldots, b$. Here, U denotes the weak limit of U_k as in condition (2). The term $U_k - \mathrm{E}(U)$ in the above serves the role of the error in a regression of $Y_k = \log T_k^{(I)}$ on three factors: intercept, $\log k$ and $\log\log k$.

Thus, let $\tilde{g}_b^{(I)}$ be the LS estimator of the coefficient of the $\log k$ term in the above regression of Y_k on the three factors intercept, $\log k$ and $\log\log k$; that is, $\tilde{g}_b$ estimates $g(\lambda)$.

As before, only block sizes $k = 1, \ldots, b$ are used from the Ith scan, where $b = [c_b n^{\zeta}]$ for some $\zeta \in (0, 1)$ and $c_b > 0$. Finally, let $\tilde{\lambda}_b^{(I)} = g^{-1}(\tilde{g}_b^{(I)})$, and

$$\tilde{\lambda}_b^{*} = Q^{-1} \sum_{i=1}^{Q} \tilde{\lambda}_b^{(i)} \text{ and } \tilde{\lambda}_b^{\sharp} = \text{median}\{\tilde{\lambda}_b^{(1)}, \ldots, \tilde{\lambda}_b^{(Q)}\},$$

where $\tilde{\lambda}_b^{(i)}$ is computed from $scan_i$ picked from the aforementioned independent scans $scan_1$, $scan_2, \ldots, scan_Q$, or a large number of randomly chosen scans. It is then expected that an analogue of Theorem 2.2 will still hold true with $\tilde{\lambda}_b^{*}$ instead of $\hat{\lambda}_b^{*}$, with a similar result expected to hold for $\tilde{\lambda}_b^{\sharp}$ as well.

3. Example: estimation of the heavy-tail index

3.1. An illustration of the methodology at work

Throughout this section, we will assume that the data $X_1, \ldots, X_n$ are an observed stretch of a *linear* time series satisfying $X_t = \sum_{j \in \mathbb{Z}} \psi_j Z_{t-j}$, for all $t \in \mathbb{Z}$, where $\{Z_t\}$ is i.i.d. from some (possibly heavy-tailed) distribution $F \in D(\lambda)$; the (generally unknown) filter coefficients $\{\psi_j\}$ are assumed to be absolutely summable. Here, $D(\lambda)$ denotes the domain of attraction of a λ-stable law with $\lambda \in (0, 2]$; see, for example, Embrechts, Klüppelberg & Mikosch (1997, Chap. 2).

In this context, it is well known that there exist sequences a_n and b_n such that $a_n^{-1}(\sum_{t=1}^{n} Z_t - b_n) \stackrel{\mathcal{L}}{\Longrightarrow} S_\lambda$, where S_λ denotes a generic λ-stable law with unspecified scale, location and skewness; recall that $a_n = n^{1/\lambda} \tilde{L}(n)$ for some slowly varying function $\tilde{L}(\cdot)$. The centring sequence b_n can be taken to be zero if either $\lambda < 1$ or $\lambda > 1$ and Z_t has mean zero. When $\lambda = 1$, we can only let $b_n = 0$ if Z_t is symmetric about zero. Our goal is the estimation of λ, which is tantamount to the estimation of the main part of the rate a_n.

Tail index estimators typically are based upon a number q of extreme order statistics; see Csörg, Deheuvels & Mason (1985) and Csörg & Viharos (1998) for a general categorization of such tail index estimators, including the well-known Hill estimator. A practical problem for the Hill estimator, but also for the other above-mentioned estimators, lies in choosing the number of order statistics q to be used. Although it is known that we must have $q \to \infty$ and $q/n \to 0$ as $n \to \infty$ to ensure consistency, optimally choosing q in a given finite sample situation is a challenging problem; see, for example, Danielsson *et al.* (2001) and references therein.

To address the tail index estimation problem using the methodology of Section 2, let $T_n = \frac{1}{n} \sum_{t=1}^{n} X_t^2$, i.e. the sample second moment. As it turns out, T_n diverges with rate $n^{-1+2/\lambda}$ (modulo a slowly varying function). Thus, in the language of Section 2, we have $g(\lambda) = -1 + 2/\lambda$. More formally, under the assumed conditions, we have that

$$n^{-2/\lambda} L(n) \sum_{t=1}^{n} X_t^2 \stackrel{\mathcal{L}}{\Longrightarrow} J, \tag{10}$$

where $L(\cdot)$ is a slowly varying function and J has a positively skewed $S_{\lambda/2}$ distribution when $\lambda \in (0, 2)$; see, for example, McElroy & Politis (2002). When $\lambda = 2$, the expression (10) is valid if Z_t has finite variance, in which case L can be taken to be a constant by the Law of Large

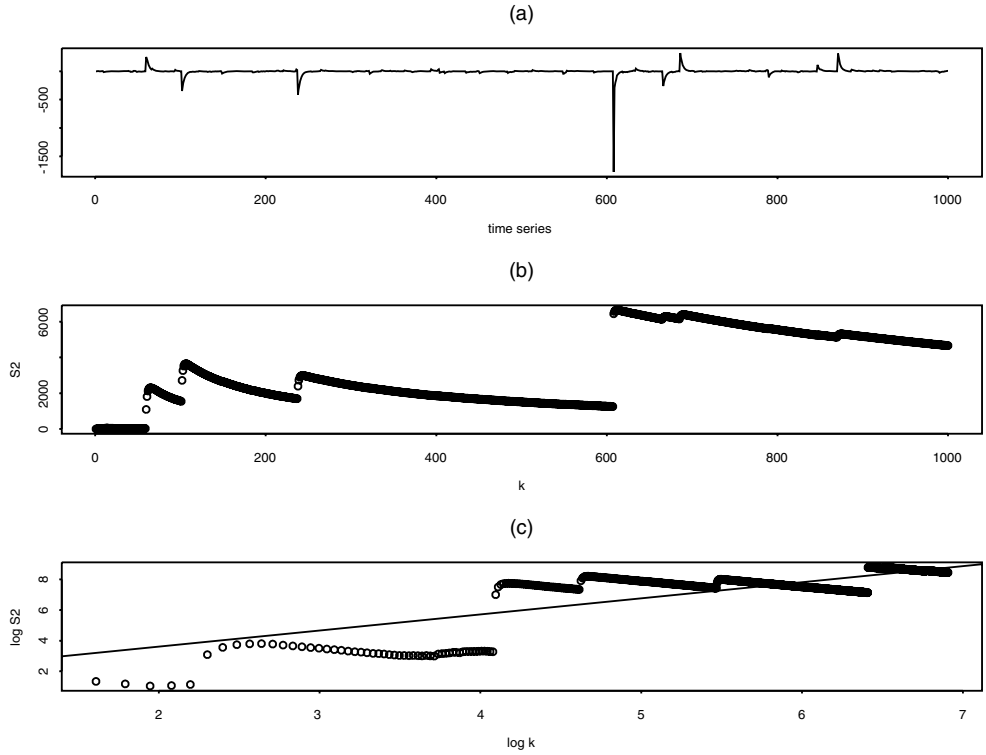

Figure 1. (a) Sample-path of a Cauchy time series generated from model (11) with $n = 1000$ and $\rho = 0.7$. (b) Plot of $T_k = \frac{1}{k}\sum_{t=1}^{k} X_t^2$ vs. k. (c) Plot of $\log T_k$ vs. $\log k$ with superimposed least squares regression line (with intercept term) (s2 denotes the sample second moment).

Numbers (see, for example, Corollary 2.2.17 of Embrechts *et al.* 1997). Otherwise, when the variance is infinite, in many cases it is necessary to centre the sample second moment.

Figure 1 gives an illustration of the proposed methodology of estimating the heavy-tail index, λ. Figure 1(a) shows a typical sample-path of a Cauchy time series satisfying the AR(1) model (11) with $n = 1000$ and $\rho = 0.7$. Note that, in the Cauchy case, the slowly varying function L is a constant, and $\lambda = 1$. Figure 1(b) shows a plot of T_k vs. k, where the divergence of T_k for increasing values of k is apparent; interestingly, this divergence occurs with steps/jolts induced by each outlying value the Cauchy distribution generates. Finally, Figure 1(c) depicts the core of the basic estimation procedure: a plot of $\log T_k$ vs. $\log k$ with a superimposed LS regression line. The $\hat{\lambda}$ corresponding to this plot was found to be equal to 0.96.

3.2. A small simulation study

To perform a small simulation study, the simple AR(1) model

$$X_t = \rho X_{t-1} + Z_t \tag{11}$$

was employed with $\rho = -0.5$, 0.1 or 0.7. As before, $\{Z_t\}$ is i.i.d. from some distribution $F \in D(\lambda)$. The distributions under consideration in this simulation were the following: (i)

$\{Z_t\} \sim$ i.i.d. Cauchy; (ii) $\{Z_t\} \sim$ i.i.d. 1.5–Stable (symmetric); (iii) $\{Z_t\} \sim$ i.i.d. 1.9–Stable (symmetric); (iv) $\{Z_t\} \sim$ i.i.d. N(0,1); (v) $\{Z_t\} \sim$ i.i.d. Pareto(2, 1); (vi) $\{Z_t\} \sim$ i.i.d. Burr (2, 1, 0.5); (vii) $Z_t = \tilde{Z}_t \cdot \max(1, \log_{10}|\tilde{Z}_t|)$, where $\{\tilde{Z}_t\} \sim$ i.i.d. Burr (2, 1, 0.5).

Cases (i) to (vi) fall under a 'normal' domain of attraction; that is, the slowly varying function L is constant, as Theorem 2.2 assumes. The variation (vii) on the Burr distribution has as its purpose the construction of a non-normal domain of attraction. For definitions of the Stable, Pareto, and Burr distributions, see Embrechts *et al.* (1997).

For each combination of the value of ρ and the distribution F, 100 time series stretches were generated, each of length $n = 1000$. From each series, Algorithm B′ of McElroy & Politis (2007) – as described in www.math.ucsd.edu/∼politis/PAPER/scansAlgorithms.pdf – was employed to generate $N = 200$ random scans on the basis of which the estimators $\hat{\lambda}_b^*$ and $\hat{\lambda}_b^\sharp$ were computed for three values of b: 100, 300, and 700. (The information that $0 < \lambda \leq 2$ was explicitly used here, in that values of $\hat{\lambda}$ greater than 2 were truncated to the value 2; interestingly, no occurrences of a negative $\hat{\lambda}$ were observed. This truncation is necessary for the good performance of $\hat{\lambda}^*$, but is superfluous/unnecessary for $\hat{\lambda}^\sharp$, as the latter is based on a median that 'clips' outlying values.) The results of our simulation

TABLE 1

Empirical MSEs of the proposed estimators with $N = 200$; data from model (11) with $n = 1000$ and (a) $\rho = -0.5$, (b) $\rho = 0.1$, and (c) $\rho = 0.7$. The best choice of b is indicated by the MSE shown in boldface

(a)

	$\hat{\lambda}^*_{100}$	$\hat{\lambda}^*_{300}$	$\hat{\lambda}^*_{700}$	$\hat{\lambda}^\sharp_{100}$	$\hat{\lambda}^\sharp_{300}$	$\hat{\lambda}^\sharp_{700}$
(i)	0.067	**0.052**	0.063	0.038	**0.037**	0.063
(ii)	**0.006**	0.017	0.038	**0.020**	0.027	0.055
(iii)	0.019	**0.010**	0.011	**0.004**	0.006	0.014
(iv)	0.030	0.013	**0.006**	0.006	0.003	**0.001**
(v)	0.271	0.230	**0.202**	0.273	0.253	**0.241**
(vi)	**0.025**	0.026	0.049	**0.016**	0.026	0.058
(vii)	**0.006**	0.015	0.032	**0.041**	0.064	0.079

(b)

	$\hat{\lambda}^*_{100}$	$\hat{\lambda}^*_{300}$	$\hat{\lambda}^*_{700}$	$\hat{\lambda}^\sharp_{100}$	$\hat{\lambda}^\sharp_{300}$	$\hat{\lambda}^\sharp_{700}$
(i)	0.060	**0.049**	0.062	**0.031**	0.034	0.063
(ii)	**0.007**	0.018	0.038	**0.019**	0.027	0.053
(iii)	0.014	**0.009**	0.010	**0.003**	0.005	0.014
(iv)	0.019	0.008	**0.004**	0.002	0.001	<**0.00**
(v)	0.234	0.194	**0.169**	0.216	0.205	**0.199**
(vi)	**0.022**	0.025	0.049	**0.016**	0.025	0.057
(vii)	**0.006**	0.016	0.034	**0.049**	0.067	0.080

(c)

	$\hat{\lambda}^*_{100}$	$\hat{\lambda}^*_{300}$	$\hat{\lambda}^*_{700}$	$\hat{\lambda}^\sharp_{100}$	$\hat{\lambda}^\sharp_{300}$	$\hat{\lambda}^\sharp_{700}$
(i)	0.076	**0.057**	0.064	0.048	**0.038**	0.064
(ii)	**0.005**	0.016	0.036	**0.016**	0.026	0.051
(iii)	0.029	0.014	**0.012**	0.007	**0.006**	0.015
(iv)	0.046	0.020	**0.010**	0.013	0.006	**0.003**
(v)	0.147	0.118	**0.102**	0.103	**0.102**	0.110
(vi)	**0.060**	0.048	0.067	0.033	**0.031**	0.065
(vii)	0.024	**0.017**	0.033	**0.017**	0.044	0.069

TABLE 2

Empirical MSE of the optimal Hill estimator $H_{q_{\text{opt}}}$ based on q_{opt} order statistics; data from model (11) with $n = 1000$ and $\rho = -0.5$, $\rho = 0.1$, and $\rho = 0.7$

	$\rho = -0.5$		$\rho = 0.1$		$\rho = 0.7$	
	$H_{q_{\text{opt}}}$	q_{opt}	$H_{q_{\text{opt}}}$	q_{opt}	$H_{q_{\text{opt}}}$	q_{opt}
(i)	0.011	180	0.007	140	0.019	200
(ii)	0.013	220	0.013	200	0.031	220
(iii)	0.015	220	0.017	220	0.035	220
(iv)	n/a	n/a	n/a	n/a	n/a	n/a
(v)	0.118	40	0.086	40	0.026	400
(vi)	0.034	60	0.027	60	0.017	300
(vii)	0.052	20	0.059	20	0.048	100

are summarized in Table 1, where the empirical MSE of each estimator is given. The best choice of b is indicated by the MSE shown in boldface.

Note that, in this set-up, the benchmark for comparison among estimators of λ is given by the Hill estimator H_q based on q extreme order statistics. Table 2 shows the (empirically found) *true* optimal values of q, denoted by q_{opt}; in other words, $H_{q_{\text{opt}}}$ had the smallest MSE empirically computed from the model in question over a wide range of q values. In this way, the performance of $H_{q_{\text{opt}}}$ constitutes a practically unattainable benchmark, because the value of q_{opt} is *not* known by the practitioner. As mentioned earlier, estimation of q_{opt} is not a trivial matter and is further complicated when the data are dependent; see Embrechts *et al.* (1997) or Danielsson *et al.* (2001) and references therein. This phenomenon is manifested in our simulations, especially in cases (v)–(vii), i.e. the Pareto and Burr distributions, for which the value of the true q_{opt} seems to be quite unstable as a function of the dependence factor ρ.

The conclusions of our simulation are quite striking.

(1) Comparing the best estimator from Table 1 with the best Hill estimator from Table 2, i.e. allowing for optimal 'bandwidth' choice for all estimators in question, we see that the estimators $\hat{\lambda}_b^*$ and/or $\hat{\lambda}_b^\sharp$ outperformed the Hill estimator uniformly in all cases except the Cauchy (i) and the Pareto (v).

(2) The estimators $\hat{\lambda}_b^*$ and/or $\hat{\lambda}_b^\sharp$ performed excellently in the Gaussian case (iv); note that the Hill estimator is inapplicable/inconsistent in this case as it diverges to infinity – hence the 'n/a's in Table 2. (Indeed, in the Gaussian case the Hill estimator diverges to infinity, albeit very slowly. As it turns out, for a sample size of $n = 1000$ the distribution of Hill's estimator misleadingly appears to be centred around the value 2. It takes sample sizes of 5000 or more to see the Hill estimator actually diverging from the neighborhood of 2.) The reason is that Hill estimates the heavy-tail index directly. By contrast, our method estimates the index of the domain of attraction, which appears in the rate of convergence of the statistic of interest.

(3) The excellent performance of $\hat{\lambda}_b^*$ and/or $\hat{\lambda}_b^\sharp$ in case (vii) suggests that the assumption of Theorem 2.2 that the slowly varying function L is (asymptotic to) a constant is not crucial for the validity of the method.

(4) The comparison between $\hat{\lambda}_b^*$ and $\hat{\lambda}_b^\sharp$ is inconclusive based on our simulations. Note, however, that $\hat{\lambda}_b^*$ is aided by the explicit truncation of the original estimator to the value 2. On the other hand, $\hat{\lambda}_b^\sharp$ is more robust, and thus may be recommendable in a general set-up when outside information – such as the restriction $\lambda \in (0, 2]$ – might not be available.

(5) Comparing the results of Table 1 with the entries of Table 1 of McElroy & Politis (2007) that are associated with $N = 200$, it is apparent that using b smaller than n is advantageous in all cases except the Normal (iv) and Pareto (v).

3.3. Conclusions

The simulation confirms that our proposed methodology leads to reasonable estimates of the heavy-tail index under (linear) dependence and possibly non-normal domains of attraction, i.e. a non-constant slowly varying function L. Nevertheless, it should be stressed that our methodology has general applicability, and is not specific to the particular context as Hill's estimator is. Of course, it is expected that context-specific, carefully optimized estimators may give a better performance than this general 'off-the-shelf' tool. The fact that, in most of the cases considered, for example (ii), (iii) and (vii), our general methodology seems to outperform the optimally fine-tuned Hill estimator (using the true q_{opt}) can be considered remarkable.

As previously mentioned, optimally choosing a 'bandwidth' parameter such as b in $\hat{\lambda}_b^*$ and $\hat{\lambda}_b^\sharp$ is a very challenging problem. Although even with a suboptimal choice of b the consistency and the polynomial rate of convergence of $\hat{\lambda}_b^*$ and $\hat{\lambda}_b^\sharp$ are maintained, it is hoped that the guidelines of Remark 2.3 will prove useful in this general quest; more work is definitely needed in that direction.

References

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

CSÖRG, S., DEHEUVELS, P. & MASON, D. M. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.* **13**, 1050–1077.

CSÖRG, S. & VIHAROS, L. (1998). Estimating the tail index. In *Asymptotic Methods in Probability and Statistics*, ed. B. Szyszkowicz, pp. 833–881, Amsterdam: North Holland.

DANIELSSON, J., DE HAAN, L., PENG, L. & DE VRIES, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivar. Analysis* **76**, 226–248.

EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Berlin: Springer-Verlag.

MCELROY, T. & POLITIS, D. N. (2002). Robust inference for the mean in the presence of serial correlation and heavy tailed distributions. *Econometric Theory* **18**, 1019–1039.

MCELROY, T. & POLITIS, D. N. (2007). Computer-intensive rate estimation, diverging statistics, and scanning. *Ann. Statist.* **35**, 1827–1848. [Also available from: www.math.ucsd.edu/∼politis/PAPER/scans.pdf]

MIZERA, I. & WELLNER, J. A. (1998). Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. *Ann. Statist.* **26**, 672–691.

POLITIS, D. N. (2002). A new approach on estimation of the tail index. *C. R. Acad. Sci. Paris, Ser. I* **335**, 279–282.

POLITIS, D. N., ROMANO, J. P. & WOLF, M. (1999). *Subsampling*. New York: Springer.