Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 1389–1406
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Moment-based tail index estimation

Tucker McElroy[a], Dimitris N. Politis[b,*]

[a] US Bureau of the Census, University of California, San Diego, USA
[b] University of California, San Diego, USA

Received 20 June 2003; accepted 28 April 2006
Available online 2 June 2006

Abstract

A general method of tail index estimation for heavy-tailed time series, based on examining the growth rate of the logged sample second moment of the data was proposed and studied in Meerschaert and Scheffler (1998. A simple robust estimator for the thickness of heavy tails. J. Statist. Plann. Inference 71, 19–34) as well as Politis (2002. A new approach on estimation of the tail index. C. R. Acad. Sci. Paris, Ser. I 335, 279–282). To improve upon the basic estimator, we introduce a scale-invariant estimator that is computed over subsets of the whole data set. We show that the new estimator, under some stronger conditions on the data, has a polynomial rate of consistency for the tail index. Empirical studies explore how the new method compares with the Hill, Pickands, and DEdH estimators.

Keywords: Heavy tails; Domains of attraction; Subsampling

1. Introduction

Let $X_1, \dots, X_n$ be an observed stretch of a linear-dependent time series satisfying

$$X_t = \sum_{j \in \mathbb{Z}} \psi_j Z_{t-j} \tag{1}$$

for all $t \in \mathbb{Z}$, where $\{Z_t\}$ is *iid* (independent and identically distributed) from some continuous distribution F; the case where $\psi_j = 0$ for $j \neq 0$ is the special case of $\{X_t\}$ being *iid*, and is considered in detail throughout Section 2. We assume that F belongs to $D(\alpha)$, the domain of attraction of an α-stable law; however, the heavy-tail index α is unknown and must be estimated from the data. In this context, there exist sequences a_n and b_n such that $a_n^{-1}(\sum_{t=1}^n Z_t - b_n) \stackrel{\mathscr{L}}{\Longrightarrow} S_\alpha$, where S_α denotes a generic α-stable law with unspecified scale, location and skewness, and $\alpha \in (0, 2)$; it is always true that we can write $a_n = n^{1/\alpha} L(n)$ for some slowly varying function L. If L is either constant or asymptotically tends to a nonzero constant, we say that F is in the normal domain of attraction, denoted by $\mathrm{ND}(\alpha)$. We restrict to the case that $\alpha < 2$ to ensure that the variance of the data is always infinite.

Estimators of the tail index are often constructed from extreme order statistics—see Csörgő et al. (1985) for a general class of such estimators. The well-known Hill estimator H_q falls into this class. A challenging problem lies in

* Corresponding author.
E-mail address: politis@math.ucsd.edu (D.N. Politis).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2006.04.002

choosing the number of order statistics q to be used in practice; see Embrechts et al. (1997) and the references therein for a discussion of this topic. The work of Politis (2002) presents an alternative estimation approach that is based on empirically examining the growth of appropriately chosen diverging statistics. A prime example of such a statistic is given by the sample variance that diverges to infinity in the absence of a finite second moment. As a matter of fact, a consistent—albeit with logarithmic rate—tail index estimator can be constructed by simply taking the ratio of the logarithm of the sample variance to the logarithm of the sample size; see Meerschaert and Scheffler (1998). For a survey on nonparametric methods for heavy-tailed data, see Meerschaert and Scheffler (2003); these authors also extend their methodology to heavy-tailed random vectors.

In this paper, a new class of tail index estimators are presented, which are in the same philosophy as the estimators of Meerschaert and Scheffler (1998) and Politis (2002); in particular, they do not rely on extreme order statistics. The main estimator is an effort to improve on the convergence rate of the tail index estimator of Meerschaert and Scheffler (1998), hereafter referred to as the MS statistic. Our approach is simple and intuitive, and can also be generalized to rate estimation settings other than the heavy-tail problem. We also propose some ways to improve upon the basic form of the new estimators, and give some finite-sample simulation results. Section 2 covers the theoretical results that establish the asymptotics of the tail index estimator, while Section 3 deals with the more practical issues of how to conduct inference for α, and presents the result of several simulation studies. All proofs are placed in the Appendix.

2. Theory

2.1. Motivation for the new approach

The theory developed in this section motivates the construction of the tail index estimators that we explore empirically in Section 3. In order to facilitate simple proofs, the results in this section are presented for *iid* data, i.e., the case $\psi_j = 0$ for $j \neq 0$ in (1). Some notations that are consistently used in this paper: $\mathbb{E}$ denotes the expectation operator, whereas $\mathbb{D}$ is the variance operator (D for dispersion). By the notation $\mathbb{D}[A, B]$, we denote the covariance between variables A and B. Also, when we write a random variable without a subscript, we indicate a common version.

Let us define the sum of squares process $S_n(X^2) = \sum_{t=1}^{n} X_t^2$. It is well-known (see, for example Theorem 4.2 of Davis and Resnick (1985) for the $MA(\infty)$ case) that $S_n(X^2)$ diverges (when $\alpha < 2$) at rate a_n^2, i.e., the normalized partial sums of squares $U_n = a_n^{-2} S_n(X^2)$ converge in distribution to a nondegenerate random variable. Since $a_n = n^{1/\alpha} L(n)$, the rate of divergence of $S_n(X^2)$ may give crucial information about α. So define $\zeta = 1/\alpha$ and $M = L^2$, and consider the identity

$$\log S_n(X^2) = 2\zeta \log n + \log M(n) + \varepsilon_n, \tag{2}$$

where the random variables

$$\varepsilon_n = \log(a_n^{-2} S_n(X^2))$$

can be thought of as "residuals" in a regression of $\log S_n(X^2)$ on $\log n$. This is the basic motivation behind the regression estimator of Politis (2002), as well as the MS statistic:

$$\hat{\zeta}_n^{\mathrm{MS}} = \log^+ S_n(X^2)/(2 \log n).$$

Here $\log^+ x = \max\{\log x, 0\}$. This differs from the statistic defined by Meerschaert and Scheffler (1998) in that the sample second moments rather than the sample variance is computed; the centering makes no difference to asymptotics when $\alpha < 2$. In this paper we shall use log rather than $\log^+$; the latter function has the advantage of disregarding the behavior of X^2 near zero, whose negative values of $\log X^2$ we essentially explore. However, the logarithm allows for the intuitive decomposition given by (2). As suggested in Meerschaert and Scheffler (1998), these estimators can be easily adapted to handle the case that $\alpha \geqslant 2$ by computing fourth or six sample moments; thus, we define

$$\hat{\zeta}_n^{\mathrm{BAS}r} = \log S_n(X^{2r})/(2r \log n),$$

where the integer r is taken sufficiently large such that the $2r$th moment of the data's distribution does not exist. The notation denotes a "basic" estimator, dependent on a user-defined integer r. The following result is similar to