*The Annals of Statistics*
1997, Vol. 25, No. 5, 1805–1869

# SPECIAL INVITED PAPER

## HEAVY TAIL MODELING AND TELETRAFFIC DATA[1]

BY SIDNEY I. RESNICK

*Cornell University*

Huge data sets from the teletraffic industry exhibit many nonstandard characteristics such as heavy tails and long range dependence. Various estimation methods for heavy tailed time series with positive innovations are reviewed. These include parameter estimation and model identification methods for autoregressions and moving averages. Parameter estimation methods include those of Yule–Walker and the linear programming estimators of Feigin and Resnick as well estimators for tail heaviness such as the Hill estimator and the qq-estimator. Examples are given using call holding data and interarrivals between packet transmissions on a computer network. The limit theory makes heavy use of point process techniques and random set theory.

**1. Introduction.** Classical queuing and network stochastic models contain simplifying assumptions guaranteeing the Markov property and insuring analytical tractability. Frequently, interarrivals and service times are assumed to be iid and typically underlying distributions are derived from operations on exponential distributions. At a minimum, underlying distributions are usually assumed nice enough that moments are finite.

Increasing instrumentation of teletraffic networks has made possible the acquisition of large amounts of data. Analysis of this data is disturbing since there is strong evidence that the classical queuing assumptions of thin tails and independence are inappropriate for this data. Video conference data and packet counts per unit time in Ethernet traffic appear to exhibit long range dependence and self-similarity [Beran, Sherman, Taqqu and Willinger (1995); Beran (1994); Willinger, Taqqu, Leland and Wilson (1995)] while such phenomena as file lengths, cpu time to complete a job, call holding times, interarrival times between packets in a network and lengths of on/off cycles appear to be generated by distributions which have heavy tails [Duffy, McIntosh, Rosenstein and Willinger (1993, 1994); Meier–Hellstern, Wirth, Yan and Hoeflin (1991); Willinger, Taqqu, Sherman and Wilson (1997).

Although we will not dwell heavily on long range dependence in this survey, it is instructive to quickly view in Figure 1 a video conferencing data set in excess of 48,000 data. The top graph is the time series plot and the bottom plot

Received September 1995; revised June 1996.

[1]Research partially supported by NSF Grant DMS-94-00535 at Cornell University and NSA Grant MDA904-95-H-1036.

*AMS* 1991 *subject classifications*. 62M10, 62M09.

*Key words and phrases*. Heavy tails, regular variation, Hill estimator, Poisson processes, linear programming, autoregressive processes, parameter estimation, weak convergence, consistency, time series analysis, estimation, independence.

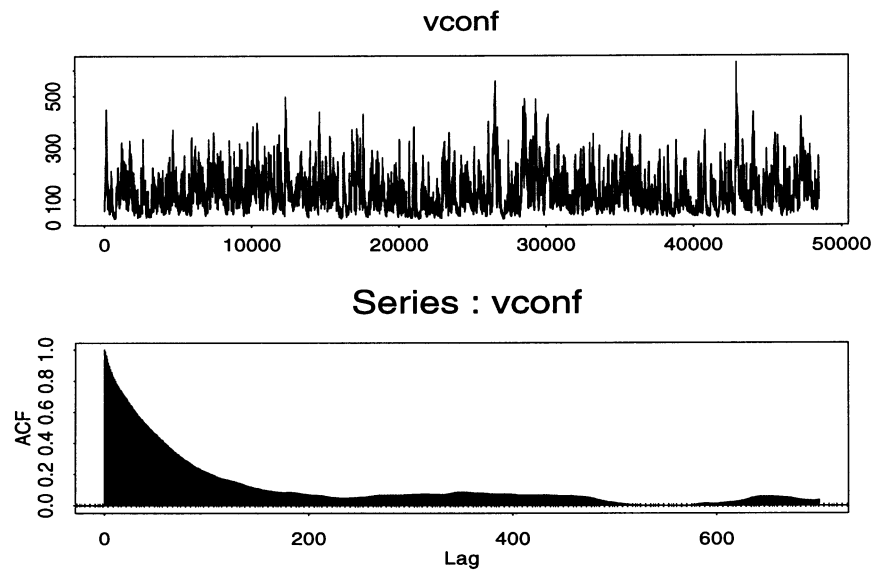


FIG. 1. *Autocorrelations of video conferencing data.*

is the sample autocorrelation function (acf), graphed out to 700 lags. Usual short range dependent data sets would show a sample correlation function dying after only a few lags and then persisting within the magic 95% confidence window determined by Bartlett's formula [see, e.g., Brockwell and Davis (1991)]. However the video conference data has an acf showing significant correlations for hundreds of lags. This data also exhibits self-similarity; see Beran, Sherman, Taqqu and Willinger (1995), Garrett and Willinger (1994).

It is logical and natural to question whether departures from classical queuing assumptions matter that much. There is persistent skepticism about the importance of correctly detecting heavy tails or long range dependence which is founded on unverified faith in the robustness of classical assumptions and perhaps, understandably, a bit of inertia. Also, mathematical analysts have been slow to embrace these features since queues and networks with such complex inputs are difficult to analyze mathematically. However there is simulation evidence [e.g., see Livny, Melamed and Tsiolis (1993)] that correctly accounting for dependence and correctly modeling tail heaviness is crucial. Resnick and Samorodnitsky (1997) verify that for a simple $G/M/1$ queue, a stationary input with long range dependence can induce heavy tails for the waiting time distributions and for the distribution of the number in the system. The inputs in the Resnick and Samorodnitsky model are of the form of interarrivals $\{T_n\}$ of customers where $\{T_n\}$ is stationary with a form of long range dependence. The dramatic degradation of system performance exemplified by heavy tailed outputs compared with what would be predicted by a model satisfying classical assumptions indicates that the phenomena of long range dependence and heavy tails cannot be ignored if estimation of network

capacity is the goal. This point will undoubtedly receive more systematic study in the near future and it is expected that long range dependence and heavy tails will be of critical importance for even crude first-order solutions of and guidelines for a wide range of network engineering problems.

It is becoming increasingly clear that there are strong connections between long range dependence and heavy tails, but these connections have not yet been systematically explored. The Resnick and Samorodnitsky (1997) paper suggests that long range dependent inputs to a queuing system can induce heavy tailed outputs. It is also clear that heavy tails can induce long range dependence. Consider the following example of an on/off model for packet transmission of a source and destination pair fashioned after one described in Willinger, Taqqu, Sherman and Wilson (1997) and discussed in Heath, Resnick and Samorodnitsky (1997): we have a stationary alternating renewal process [Resnick (1992)] with counting function $\{N(t),\ t \geq 0\}$ and renewal times $\{S_n,\ n \geq 0\}$. A renewal interval consists of an *on* period which has distribution $F_{\text{on}}$ and an *off* period with distribution $F_{\text{off}}$. Let the means of $F_{\text{on}}$ and $F_{\text{off}}$ be $\mu_{\text{on}}$ and $\mu_{\text{off}}$, respectively. Set $\mu = \mu_{\text{on}} + \mu_{\text{off}}$. The process $\{N(t)\}$ can be constructed to be stationary by flipping a coin to decide if the process starts in an on or off period, and then choosing the right residual distribution. The coin chooses on with probability $\mu_{\text{on}}/(\mu_{\text{on}} + \mu_{\text{off}})$. If the coin indicates for example on, then after the initial residual period, we alternate independent off periods with independent on periods. $N(t)$ is the number of points corresponding to termination of off periods in $[0, t]$. Next, we may define the stationary process $\{Z_t,\ t \geq 0\}$ so that $Z_t = 1$ iff $t$ is in an on period. If

$$1 - F_{\text{on}}(t) = t^{-\alpha}L(t), \qquad t \to \infty,\ 1 < \alpha < 2$$

$$1 - F_{\text{off}}(t) = o\big(1 - F_{\text{on}}(t)\big), \qquad t \to \infty,$$

where $L(t)$ is slowly varying, then

$$\text{Cov}(Z_0, Z_t) \sim \frac{\mu_{\text{off}}^2}{(\alpha - 1)\mu^3} t^{-(\alpha-1)} L(t), \qquad t \to \infty.$$

The slow rate of decay of the covariance function is characteristic of long range dependence.

The effects of heavy tails in the inputs to an associated reservoir model are quite dramatic. Suppose the alternating renewal process controls inputs to the reservoir; water flows into the reservoir at unit rate in on periods and evaporates at rate $r > \mu_{\text{on}}/\mu$ during off periods. Let $X(t)$ be the contents of the reservoir at time $t$. Then because of the regenerative nature of $\{X(t)\}$ we have

$$X(t) \Rightarrow X(\infty)$$

and it follows that the limit distribution of $X(\infty)$ is heavy tailed

$$P[X(\infty) > x] \sim (\text{const}) x^{-(\alpha-1)} L(x).$$

These results follow by standard results in the theory of regularly varying functions; see Resnick (1987), de Haan (1970), Bingham, Goldie and Teugels (1987), Geluk and de Haan (1987).

This model is a simplified idealization but effectively illustrates the simple way in which heavy tails can induce long range dependence. Traffic on an Ethernet system can be considered as a superposition of traffic from many source–destination pairs. Willinger, Taqqu, Sherman and Wilson (1997) show how superimposing independent copies of the $Z_t$ process and taking limits in a suitable manner yields fractional Brownian motion [Samorodnitsky and Taqqu (1994)] as a limit. This offers one plausible explanation for the observed self-similar nature of Ethernet traffic.

This example serves as a reminder that for modeling the data the following choices must be faced.

1. Is it better to use structural models such as on/off models where specific physical features are accounted for? Structural models which successfully capture relevant features have immediate connections to applications. Here the engineering comfort level may be high and there may be some possibility of analytical work succeeding if such physical models are used as model inputs.
2. Or should we use black box models from time series analysis? Such models ignore physical structure but have a long tradition in data analysis. They may be easier for analysts who do not have the close ties to engineers required to intelligently construct structural models. Furthermore, such models may be applicable across a broad range of disciplines.

If we use structural models, there is no guarantee we have correctly summarized relevant features. Statistical verification of goodness of fit can be challenging. The probability analysis may be easier but the statistics may be harder. Time series models have an advantage of being applicable across a broad range of disciplines. However, if we use time series models, we have to decide whether to use linear or nonlinear models. Linear models are the simplest, but there is no guarantee that in the heavy tailed world they form a large flexible class. In traditional finite variance time series modeling based on Hilbert space methodology, the Yule–Walker estimators guarantee that any correlation structure may be mimicked out to any fixed number of lags by autoregressions of a suitable order [see Brockwell and Davis (1991), page 240] and in this very limited sense linear models are sufficient for data analysis. For infinite variance models, we have no such confidence that linear models are adequately flexible and in fact the theoretical perspective offered by Rosinski's (1995) work as well as some data experience [Resnick (1997); Feigin and Resnick (1996)] make this unlikely.

This is an important point because evidence is emerging that for heavy tailed modeling there are significant differences in the behavior of key summary statistics depending on whether the model is linear or nonlinear. For example, consider the sample autocorrelation function (acf). In the heavy tailed case, it is best to define it as follows: if the stationary time series is

$\{X_t,\ t = 0, \pm 1, \pm 2, \ldots\}$ then define for $h = 0, 1, \ldots$

$$\hat{\rho}_H(h) = \frac{\sum_{i=1}^{n-h} X_t X_{t+h}}{\sum_{i=1}^{n} X_t^2}.$$

This substitutes for the usual correlation function in finite variance time series modeling. If the model is linear, the sample acf at lag $h$ converges in probability to a constant depending on $h$ [Davis and Resnick (1985a,b, 1986)] but if the model is nonlinear, Davis and Resnick (1996) show that the sample acf at lag $h$ may converge in distribution to a nondegenerate random variable depending on $h$.

Sets of data displaying characteristics of heavy tails are encountered in diverse fields other than teletraffic engineering, for example in hydrology [Gumbel (1958); Castillo (1988)], economics and finance [Koedijk, Schafgans and de Vries (1990); Janson and de Vries (1991)] reliability and structural engineering [Grigoriu (1995)]. Because of the diversity of applications and because heavy tails are one of the causes of long range dependence, we expect interest in the detection and modeling of heavy tailed phenomena to grow and hope the survey which follows will contribute to understanding the uses and limitations of the growing body of techniques.

Section 2 discusses some mathematical background of regular variation and specifies the type of heavy tailed models we will study. Section 3 focuses on the obvious steps in heavy tailed modeling: detecting heavy tails and detecting dependencies. We describe various graphical techniques of an exploratory nature which can be helpful but point out some limitations. Heavy tailed autoregressive models have been successfully studied and relevant model selection and estimation methods are summarized in Section 4, which also contains some brief remarks on the bootstrap. An example is provided in Section 5. Section 6 contains some closing remarks.

## 2. Background for heavy tailed models.

2.1. *Regular variation.* What is a heavy tail? The bulk of statistical work deals with light tails, which decay exponentially fast as typified by the normal distribution tail, since by Mills' ratio, as $x \to \infty$,

$$P[N > x] \sim \frac{n(x)}{x} = \frac{1}{\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x} \to 0.$$

Contrast this with a typical heavy tail such as possessed by the Pareto distribution: $X$ has a Pareto tail with index $\alpha > 0$ if, for $x > 0$,

$$P[X > x] = x^{-\alpha}, \qquad x > 1.$$

More generally, we say $X$ has a heavy tailed distribution $F$ if

$$P[X > x] = x^{-\alpha} L(x), \tag{2.1}$$

where $L$ is slowly varying; that is, for $x > 0$,

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1. \tag{2.2}$$

Note that when $X \geq 0$,

$$E(X^\beta) < \infty, \qquad \beta < \alpha,$$
$$E(X^\beta) = \infty, \qquad \beta > \alpha.$$

Typical slowly varying functions include the following examples:

$$\begin{aligned} L(x) &= c + o(1), && x > 0, \\ &= \log x, && x > 1, \\ &= \log(\log x), && x \text{ large}, \\ &= 1/\log x, && x > 1. \end{aligned}$$

In the first example of $L$ where $L(x) = c + o(1)$, the term $o(1)$ may look innocent and harmless but can wreak havoc and there can be a big difference between detecting Pareto tails and detecting, say, tails of stable distributions [Samorodnitsky and Taqqu (1994)] where

$$1 - F(x) \sim x^{-\alpha}, \qquad x \to \infty.$$

This point is illustrated by examples in Section 3.1.

Another way to express (2.1) is to say that $1 - F$ is regularly varying with index $-\alpha$: a distribution $F$ concentrating on $[0, \infty)$ has a tail $1 - F(x)$ which is regularly varying with index $-\alpha$, $\alpha > 0$ (written $1 - F \in RV_{-\alpha}$) if

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}, \qquad x > 0. \tag{2.3}$$

The distribution tail $1 - F$ is second order regularly varying with first-order parameter $-\alpha$ and second-order parameter $\rho$ (written $1 - F \in 2RV(-\alpha, \rho)$) if there exists a function $A(t) \to 0$, $t \to \infty$ which ultimately has constant sign such that the following refinement of (2.3) holds:

$$\begin{aligned} &\lim_{t\to\infty} \frac{(1 - F(tx)/(1 - F(t))) - x^{-\alpha}}{A(t)} \\ &\qquad = H(x) := cx^{-\alpha} \int_1^x u^{\rho-1}\, du, \qquad x > 0, \end{aligned} \tag{2.4}$$

for $c \neq 0$. Note that for $x > 0$,

$$H(x) = \begin{cases} cx^{-\alpha} \log x, & \text{if } \rho = 0, \\ cx^{-\alpha} \dfrac{x^\rho - 1}{\rho}, & \text{if } \rho < 0. \end{cases}$$

It follows that $|A| \in RV_\rho$ and no other choices of $\rho$ are consistent with $A(t) \to 0$. See de Haan and Stadtmüller (1996), Geluk and de Haan (1987).

Regular variation is the basic analytic theory underlying extreme value theory and stable processes [de Haan (1970), Geluk and de Haan (1987), Bingham, Goldie and Teugals (1987); Resnick (1987), Samorodnitsky and Taqqu (1994)]. Second-order regular variation has proven very useful and natural for establishing asymptotic normality of extreme value statistics (see the discussion in Section 3 on the Hill estimator) and also for the study of rates of convergence to extreme value and stable distributions [de Haan and Resnick (1996), de Haan and Peng (1995a,b,c), Smith (1982)].

There is a well-developed technique for integrating regularly varying functions called Karamata's theorem which roughly says that when one integrates a regularly varying function, one may treat the slowly varying function as a constant.

2.2. *Point processes and random measures.* Mathematical and asymptotic properties of heavy tailed models are analyzed with heavy reliance on weak convergence theory and point process techniques. Occasionally mildly exotic items like weak convergence of random closed sets are necessary [Feigin and Resnick (1994)]. The central role of point processes for the mathematical analysis of heavy tailed phenomena is well documented [Resnick (1986, 1987, 1991)] and will not be emphasized here beyond some reminders about notation and one central result. In what follows $M_+(E)$ is the set of positive Radon measures on a nice locally compact space $E$; $M_p(E)$ is the set of point measures. $M_+(E)$ is metrized by the vague metric [cf. Kallenberg (1983), Resnick (1987), Neveu (1976)]. We denote weak convergence of random elements or probability measures by $\Rightarrow$ and $\to_v$ denotes vague convergence of measures in $M_+(E)$. For $x \in E$ and $A \subset E$, define

$$\varepsilon_x(A) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

A Radon point measure with points in $E$ is denoted $\sum_i \varepsilon_{x_i}$. The collection of all such point measures is $M_p(E)$. The following two results giving weak convergence results for special random measures provides the necessary theoretical background.

PROPOSITION 2.1. *Let $m = m(n)$ be a sequence satisfying $m \to \infty$, $m/n \to 0$ as $n \to \infty$.*

(i) *Suppose for each $n$ that $\{Z_j(n),\ j \geq 1\}$ are iid random elements of the space $E$. The sequence of point processes*

$$\sum_{j=1}^{n} \varepsilon_{Z_j(n)}, \qquad n = 1, 2, \ldots$$

*converges weakly to a limiting Poisson process on $E$ with mean measure $\nu$ iff*

$$E\left( \sum_{j=1}^{n} \varepsilon_{Z_j(n)}(\cdot) \right) = nP[Z_1(n) \in \cdot] \overset{v}{\to} \nu(\cdot).$$

(ii) *Suppose for each $n$ that $\{Y_j(n),\ j \geq 1\}$ are iid random elements of the space $E$. The sequence of random measures*

$$\frac{m}{n}\sum_{j=1}^{n} \varepsilon_{Y_j(n)}, \qquad n = 1, 2, \ldots$$

*converges weakly to a limiting nonrandom measure $\nu \in M_+(E)$, iff*

$$E\left(\frac{m}{n}\sum_{j=1}^{n} \varepsilon_{Y_j(n)}(\cdot)\right) = mP[Y_1(n) \in \cdot] \xrightarrow{v} \nu.$$

Part (i) of the proof is Proposition 3.21, page 154 in Resnick (1987) and part (ii) is 3.57, page 161, Resnick (1987) and is also given in Resnick (1986).

An application of Proposition 2.1 of immediate interest to analysis of heavy tailed phenomena is as follows: let $\{Z_j,\ j \geq 1\}$ be iid and nonnegative with common distribution $F$. Set $E = (0, \infty]$, $b(t) = F^{\leftarrow}(1 - 1/t)$ where $F^{\leftarrow}$ is the left continuous inverse of the monotone function $F$, $Z_j(n) = Z_j/b(n)$, $Y_j(n) = Z_j/b(m)$ and $m = m(n)$ is a sequence satisfying $m \to \infty$, $m/n \to 0$ as $n \to \infty$. Also, the measure $\nu$ is given by $\nu((x, \infty]) = x^{-\alpha}$. We then have the following equivalences to the regular variation condition given in (2.3).

1. We have in $M_+(E)$,

$$nP\left[\frac{Z_1}{b(n)} \in \cdot\right] \xrightarrow{v} \nu.$$

2. In $M_+(E)$,

$$N_n := \sum_{j=1}^{n} \varepsilon_{Z_j/b(n)} \Rightarrow N_\infty,$$

where $N_\infty$ is a Poisson process on $E$ with mean measure $\nu$.

3. In $M_+(E)$,

$$\nu_n := \frac{m}{n}\sum_{j=1}^{n} \varepsilon_{Z_j/b(m)} \Rightarrow \nu.$$

2.3. *Heavy tailed time series models.* We will try to model data with linear time series models, but keep in mind that although this class of models is relatively simple, there is no guarantee that it is adequate. In particular, this choice of class excludes nonlinear models and such random coefficient models as ARCH and GARCH, which are dearly beloved in economics. Linear time series models is familiar territory but one often encounters the situation that many statistical techniques are tried in the heavy tailed domain because they work in the finite variance arena, and this is fairly weak justification.

We will deal with moving average processes of order infinity, written MA($\infty$). These are specified as follows: let $\{Z_t,\ -\infty < t < \infty\}$ be iid and nonnegative with common distribution $F$ satisfying (2.3). For suitable constants

$\{c_j\}$, define

$$X_t = \sum_{j=0}^{\infty} c_j Z_{t-j}, \qquad t = 0, \pm 1, \ldots . \tag{2.5}$$

Summability conditions must be assumed for the $\{c_j\}$ to guarantee that the random series in (2.5) converges; the following is typically assumed:

$$\sum_{j=0}^{\infty} |c_j|^{\delta} < \infty \quad \text{for some } 0 < \delta < \alpha \wedge 1.$$

Note we assume $Z_j \geq 0$ but only assume $c_j \in \mathbb{R}$. It may be mildly puzzling and even controversial to assume positive $Z$'s but the sort of data that we have in mind to model is inherently positive, and if the support of $Z_i$ included negative values, it would be difficult to preclude an $X_t$ from taking negative values with positive probability. Furthermore, there is an important practical reason for the assumption: without this assumption limit distributions of important statistics are a mess and the only practical alternative to assuming positivity is to assume the distribution of the $Z$'s is symmetric, which seems more harsh than assuming positivity.

Of course, the usual classical models are special cases of MA($\infty$) [cf. Brockwell and Davis (1991)]. These include:

1. Autoregressions of order $p$, abbreviated AR($p$):

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + Z_t, \qquad t = 1, \ldots, n.$$

2. Finite order moving averages of order $q$, written MA($q$):

$$X_t = \sum_{j=0}^{q} \theta_j Z_{t-j}, \qquad t = 1, \ldots, n.$$

3. Autoregressive moving average processes with orders $p, q$ written ARMA($p, q$):

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=0}^{q} \theta_j Z_{t-j}, \qquad t = 1, \ldots, n.$$

In each case, the process can be written as an MA($\infty$).

Issues in a classical context when fitting time series models to data must also be dealt with in the heavy tailed case. These include:

1. Model selection: Is ARMA a good choice of a parametric family and if so, which ARMA should we choose; that is, how do we choose $p$ and $q$?
2. Parameter estimation: After $p$ and $q$ are chosen, how do you estimate parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, $\alpha$, and so on?
3. Confirmation: After model selection and estimation have been accomplished, how does one confirm that the fitted model is an adequate description of what produced the data?

4. Given a well-selected, estimated and confirmed model, what can you do with it? Predict? There does not seem to be any practical prediction theory for heavy tailed phenomena, but if a good fit to the data has been achieved, synthetic data can be fed into a network model to try to estimate things like the distribution of the time to buffer overflow.

**3. First steps for heavy tailed modeling.** Suppose one is inclined to try fitting a heavy tailed model to a data set. The following seems like a rational set of exploratory first steps:

1. Decide if the data could plausibly be explained by a heavy tailed model.
2. Try to assess if there is dependency in the data.

We take up these two steps in turn.

3.1. *Are heavy tails present?* To help with assessing whether heavy tails are present and to estimate the index $\alpha$ in (2.3), various exploratory plotting techniques are available. These are based on the Hill estimator and the qq–plot. We begin with the Hill estimator, which is widely used.

3.1.1. *The Hill estimator.* Suppose $X_1, \ldots, X_n$ are iid from a distribution $F$. Let

$$X_{(1)} > X_{(2)} > \cdots > X_{(n)}$$

be the order statistics. If $F$ has an exact Pareto distribution,

$$1 - F(x) = x^{-\alpha}, \qquad x > 1,$$

then taking logarithms $\log X_1, \ldots, \log X_n$ yields a sample from an exponential density with parameter $\alpha$. Since the mean is $\alpha^{-1}$, the maximum likelihood estimator (mle) of $\alpha^{-1}$ is the sample mean and thus

$$H_n = \frac{1}{n} \sum_{i=1}^{n} \log X_{(i)}$$

is the mle of $\alpha^{-1}$. If instead of assuming a Pareto distribution, we only assume

$$1 - F(x) = x^{-\alpha} L(x), \qquad x \to \infty,$$

then we may pick $k < n$ and define the Hill estimator [Hill (1975)] to be

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}}.$$

Note that $k$ is the number of upper-order statistics used in the estimation.

The rough idea behind using only $k$ upper-order statistics is that you should only sample from that part of the distribution which looks most Pareto-like. A more precise explanation is that conditional on $X_{(k+1)}$, the sample

$$\frac{X_{(1)}}{X_{(k+1)}}, \ldots, \frac{X_{(k)}}{X_{(k+1)}}$$

is distributed like the order statistics from a sample of size $k$ from the distribution with tail

$$\frac{1-F(xX_{(k+1)})}{1-F(X_{(k+1)})}, \qquad x \geq 1.$$

Because of (2.3), if $X_{(k+1)}$ is large, the regular variation condition implies that

$$\frac{1-F(xX_{(k+1)})}{1-F(X_{(k+1)})} \approx x^{-\alpha}$$

and thus it seems sensible to do what we did in the Pareto case.

Here are the theoretical properties of the Hill estimator.

PROPOSITION 3.1. *Suppose* $\{X_t\}$ *is stationary and the marginal distribution satisfies*

$$P[X_1 > x] = x^{-\alpha}L(x), \qquad x \to \infty.$$

*If either*

(i) $\{X_t\}$ *is iid* [*Mason* (1982)] *or*

(ii) $\{X_t\}$ *is weakly dependent* [*Rootzen, Leadbetter and de Haan* (1990); *Hsing* (1991)] *or*

(iii) $\{X_t\}$ *is an MA*$(\infty)$ *process* [*Resnick and Stărică* (1995, 1996b)]

*then if* $n \to \infty$ *and* $k \to \infty$ *but* $k/n \to 0$, *we have*

$$H_{k,n} \xrightarrow{P} \alpha^{-1} \tag{3.1}$$

*and usually* (*one needs an extra unverifiable assumption on the distribution such as second-order regular variation and a further restriction on* $k$) *Hill's estimator is asymptotically normal. For the iid case we have*

$$\sqrt{k}(H_{k,n} - \alpha^{-1}) \Rightarrow N(0, \alpha^{-2}). \tag{3.2}$$

Our purposes would not be served by worrying about the precision of the statement of Proposition 3.1 but note the form of the asymptotics: we need to let $k \to \infty$ but $k/n \to 0$. So $k$, the number of upper order statistics used in the estimation, is considered a function of $n$, the sample size. We note that regular variation is equivalent to consistency of the Hill estimator in a manner made precise in Mason (1982) and second-order regular variation is equivalent to asymptotic normality of Hill's estimator in a manner made precise in Geluk, de Haan, Resnick and Stărică (1997). The condition of second-order regular variation controls the bias $EH_{k,n} - \alpha^{-1}$. See also Csörgő and Mason (1985), Davis and Resnick (1984), Dekkers and de Haan (1989), Hall (1982), Mason (1988), Mason and Turova (1994), Resnick and Stărică (1997).

Proofs of the facts given in Proposition 3.1 can be based on the following observation. Suppose the Hill estimator is based on the stationary observations $\{X_1, \ldots, X_n\}$, which have marginal distribution $G(x) = P[X_1 \leq x]$, and

quantile function

$$b(t) = G^{\leftarrow}(1 - t^{-1}).$$

For this sequence of $X$'s, define the tail empirical process as in Section 2.2 by

$$\nu_{X,n}(\cdot) = \frac{1}{k}\sum_{i=1}^{n} \varepsilon_{X_i/b(n/k)}(\cdot).$$

Suppose

$$\nu_{X,n}(\cdot) \xrightarrow{P} \nu(\cdot)$$

where $\nu(x, \infty] = x^{-\alpha}, \ x > 0$. Simple inversion and scaling arguments show that it then follows that the same limit relation follows with $b(n/k)$ replaced by $X_{(k)}$, the $k$th largest order statistic. Since

$$H_{k,n} = \int_1^{\infty} \log y \ \frac{1}{k}\sum_{i=1}^{n} \varepsilon_{X_i/X_{(k)}}(dy),$$

it appears that the convergence of the random measures can drag with it the convergence of the integral functional $H_{k,n}$. This argument is the one given in Resnick and Stărică (1995). The philosophy can also be adapted to verify asymptotic normality at least when the sequence $\{X_n\}$ is iid.

In practice, the Hill estimator is used as follows: we graph

$$\{(k, H_{k,n}^{-1}), \ 1 \le k \le n\}$$

and hope the graph looks stable so you can pick out a value of $\alpha$.

Sometimes this works beautifully but sometimes there are problems and it pays to be on good terms with a higher power. Consider Figure 2, which shows two cases where the procedure is heart-warming. The top row are time series plots. The top left plot is 4045 simulated observations from a Pareto distribution with $\alpha = 1$ and the top right plot is 4045 telephone call holding times indexed according to the time of initiation of the call. Both plots are scaled by division by 1000. The range of the Pareto data is (1.0001, 10206.477) and the range of the call holding data is (2288,11714735). Readers with teenagers living at home or who use a modem to dial into remote computers will not be surprised that call holding times can be heavy tailed. The bottom two plots are Hill plots $\{(k, H_{k,n}^{-1}), \ 1 \le k \le 4045\}$, the bottom left plot being for the Pareto sample and the bottom right plot for the call holding times. Both Hill plots are gratifyingly stable after settling down and are in a tight neighborhood. The Hill plot for the Pareto seems to nail $\alpha = 1$ correctly and the estimate in the call holding example seems to be between 0.9 and 1. (So in this case, not only does the variance not exist but the mean appears to be infinite as well.) The Hill plots could be modified to include a confidence interval based on the asymptotic normality of the Hill estimator, but we have not done this.

The Hill plot is not always so revealing. Consider Figure 3, which has come to be known as the Hill Horror Plot. The left plot is for a simulation of size 10,000 from a symmetric $\alpha$-stable distribution with $\alpha = 1.7$. One would be

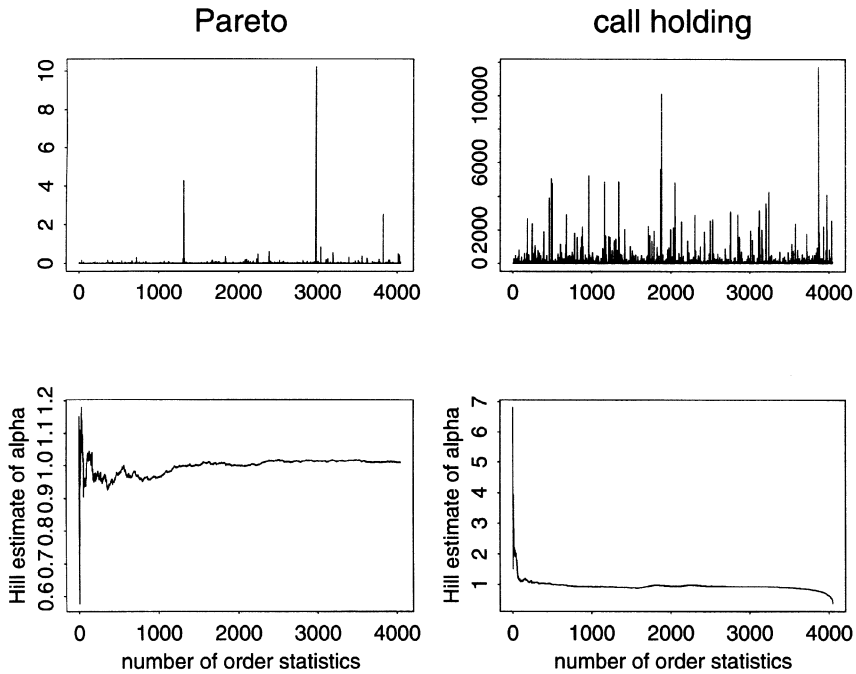


FIG. 2. *Time series and Hill plots for Pareto (left) and call holding (right) data.*

hard pressed to discern the correct answer of 1.7 from the plot. The middle plot is for a sample of size 10,000 called *perturb* from the distribution tail

$$1 - F(x) \sim x^{-1}(\log x)^{10}, \qquad x \to \infty,$$

so that $\alpha = 1$. The plot exhibits extreme bias and comes nowhere close to indicating the correct answer of 1. The problem of course is that the Hill estimator is designed for the Pareto distribution and thus does not know how to interpret information correctly from the factor $(\log x)^{10}$ and merely readjusts its estimate of $\alpha$ based on this factor rather than identifying the logarithmic perturbation. The third plot is 783 real data called *packet* representing interarrival times of packets to a server in a network. The problem here is that the graph is volatile and it is not easy to decide what the estimate should be.

Here is a summary of difficulties when using the Hill estimator.

1. How do you get a point estimate from a graph? What value of $k$ do you use?
2. The graph may exhibit considerable volatility and/or the true answer may be hidden in the graph.
3. The Hill estimate has optimality properties only when the underlying distribution is close to Pareto. If the distribution is far from Pareto, there may be outrageous bias even for sample sizes such as 1,000,000.

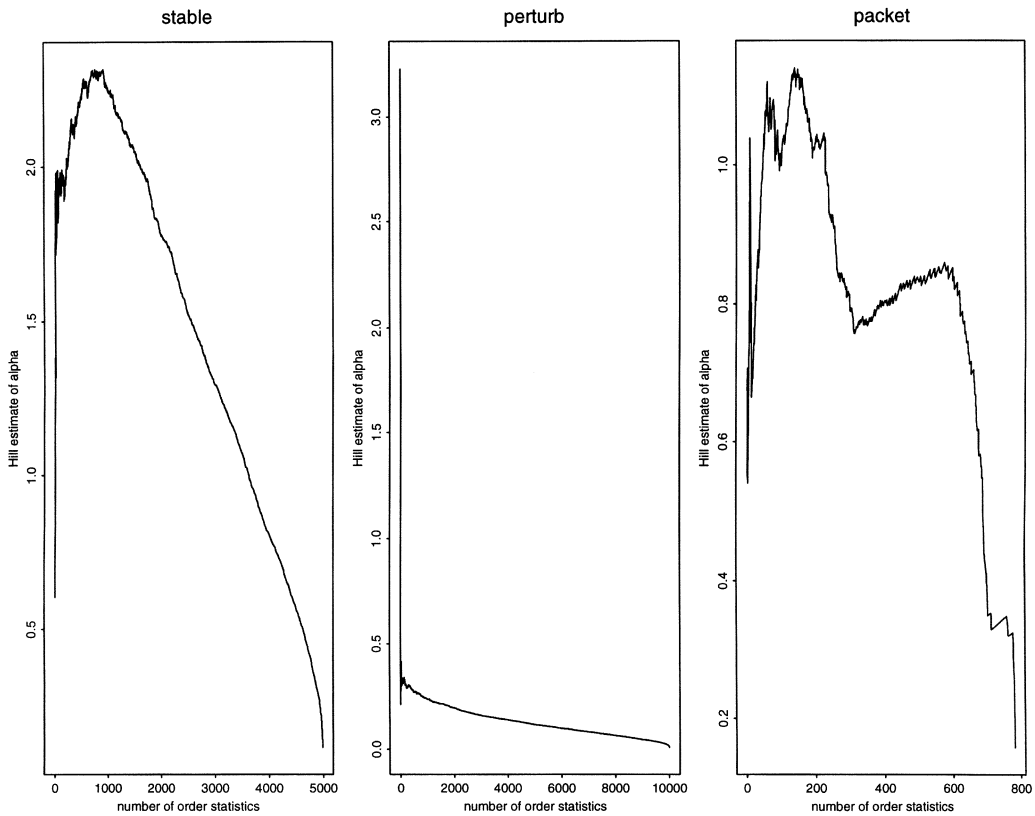


FIG. 3. *The Hill Horror Plot.*

For point 1, several previous studies advocate choosing $k$ to minimize the asymptotic mean squared error of Hill's estimator [Hall (1982), Dekkers and de Haan (1991), de Haan and Peng (1995d)]. There are several problems with such formulas. First, they require one to know the distribution rather explicitly and thus, although interesting and welcome, are not a practical solution although there is a possibility of adaptively modifying the procedure which would improve on practicality. Second, the formulas are frequently only asymptotic formulas and asymptotic equivalence is often not helpful for finite samples. If $k^* = k^*(n)$ is the choice of $k$ which minimizes the asymptotic mse, then an equally acceptable asymptotic solution is

$$k_1^* = \left(1 + \frac{10^{96}}{n}\right) k^*.$$

Even if one accepts a value of $k^*$ for finite $n$ from a displayed formula, this does not always work well in practice.

For point 2, there are simple smoothing techniques which always help to overcome the volatility of the plot, and plotting on a different scale frequently

overcomes the difficulty associated with the stable example. These techniques are discussed in the next paragraph. For the bias problem, there is no completely satisfactory resolution yet. Two possibilities under investigation are the bootstrap and fitting within a smaller parametric family.

3.1.2. *SmooHill: smoothing the Hill estimator.* A simple smoothing technique [Resnick and Stărică (1997)], although powerless to correct bias, reduces the volatility of the plot and the uncertainty about how to pick out an estimate of $\alpha$. Pick integer $u$ (usually 2 or 3) and define the SmooHill estimator:

$$\text{Smoo}H_{k,n} := \frac{1}{(u-1)k} \sum_{j=k+1}^{uk} H_{j,n}.$$

The asymptotic variance of the Hill estimator $H_{k,n}$ is $1/\alpha^2$. The asymptotic variance of $\text{Smoo}H_{k,n}$ is less, namely:

$$\frac{1}{\alpha^2}\frac{2}{u}\left(1 - \frac{\log u}{u}\right).$$

The bigger the $u$, the more the asymptotic variance is reduced. However, there is a tradeoff between variance reduction and the fact that for large $u$ fewer points are plotted in SmooHill. Usually we pick $u$ between $n^{0.1}$ and $n^{0.2}$ where $n$ is the sample size.

3.1.3. *Alt plotting: changing the scale.* As an alternative to the Hill plot, it is sometimes useful to display the information provided by the Hill estimation as

$$\left\{\left(\theta, H^{-1}_{\lceil n^\theta \rceil, n}\right),\ 0 \le \theta \le 1\right\},$$

where we write $\lceil y \rceil$ for the smallest integer greater or equal to $y \ge 0$. We call such a plot the *alternative Hill plot*, abbreviated AltHill. The alternative display is sometimes revealing since the initial order statistics get shown more clearly and cover a bigger portion of the displayed space. This method was suggested by C. Stărică and efforts are currently underway to quantify the improvement.

We return now to the examples given in Figure 3. Figure 4 gives four views of the Hill plot of the stable data where $\alpha = 1.7$. The traditional Hill plot offers little hope of correctly discerning the answer, but the alternate scale of AltHill in the upper right seems to reveal the answer fairly clearly. The bottom left plot is the SmooHill plot in alt scale and the bottom right presents both Hill and SmooHill in alt scale together.

Figure 5 analyzes the packet interarrival data introduced for Figure 3 and displays the Hill analysis in a manner parallel to the previous Figure 4. The Hill plot might lead one to guess a value of $\alpha$ of about 0.8 until one notices with unease that this occurs around $k = 400$ which means $\theta = \log(400)/\log(783) = 0.899$. Picking $k$ midway between 1 and 783 or picking $\theta$ so close to 1 seem unwise. Examining AltSmooHill makes $\alpha = 1.1$ a more likely choice. The sample size of 783 seems too small to provide much assurance of a correct estimate.

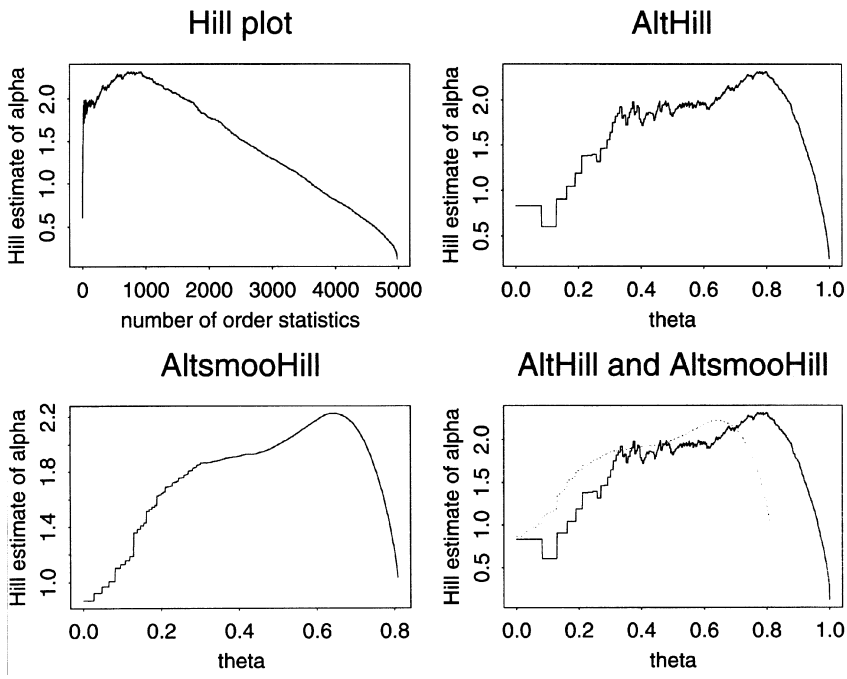


FIG. 4. *Stable*, $\alpha = 1.7$.

3.1.4. *Alternative estimators: qq-plotting.* The following statement, while not precise, is suggestive: the closer the underlying distribution $F$ is to Pareto, the better the Hill estimator seems to do. The qq-plot can help assess this. This graphical technique is a commonly used method of visually assessing goodness of fit and of estimating location and scale parameters. See for example Rice (1988) and Castillo (1988). It can be adapted to the problem of detecting heavy tails and for estimating the $\alpha$. It rests on the simple observation that for a sample of size $n$ uniformly distributed on $(0, 1)$, since the spacings are identically distributed, plotting $i/(n+1)$ vs. the $i$th largest in the sample should yield approximately a straight line of slope 1.

Suppose $\{X_1, \ldots, X_n\}$ are iid with distribution $F$. Pick $k$ upper order statistics

$$X_{(1)} > X_{(2)} > \cdots X_{(k)}$$

and neglect the rest. Plot

$$\left\{\left(-\log\left(1 - \frac{j}{k+1}\right), \log X_{(j)}\right),\ 1 \le j \le k\right\}. \tag{3.3}$$

If the data is approximately Pareto or even if $1 - F$ is only regularly varying and satisfies (2.3), this should be approximately a straight line with slope $= 1/\alpha$. The slope of the least squares line through the points is an estimator called the qq-estimator [Kratz and Resnick (1996)]. Computing the slope we

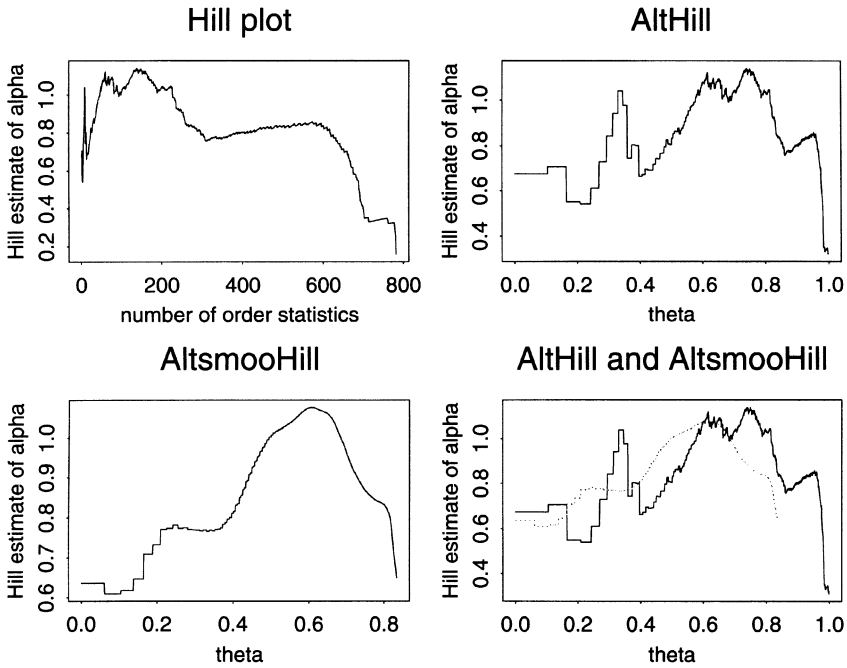


FIG. 5. *Hill plots for packet interarrivals.*

find that the qq-estimator is given by

$$\widehat{\alpha^{-1}}_{k,n} = \left[\frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{(k+1)}\right)\right)\log\left(\frac{X_{(i)}}{X_{(k+1)}}\right) - \frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)H_{k,n}\right] \times \left[\frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{(k+1)}\right)\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{(k+1)}\right)\right)^2\right]^{-1} \tag{3.4}$$

There are two different plots one can make based on the qq-estimator. There is the dynamic qq-plot obtained from plotting $\{(k, 1/\widehat{\alpha^{-1}}_{k,n}),\ 1 \le k \le n\}$ which is similar to the Hill plot. Another plot, the static qq-plot, is obtained by choosing and fixing $k$, plotting the points in (3.3) and putting the least squares line through the points while computing the slope as the estimate of $\alpha^{-1}$.

The qq-estimator is consistent if $k \to \infty$ and $k/n \to 0$ and under a second-order regular variation condition and further restriction on $k(n)$, it is asymptotically normal with asymptotic variance $2/\alpha^2$. This is larger than the asymptotic variance of the Hill estimator, but bias and volatility of the plot seem to be more of an issue than asymptotic variance. The volatility of the qq-plot always

seems to be less than that of the Hill estimator. As with the Hill estimator, sensitivity to choice of $k$ is an important issue.

Figure 6 compares the Hill plot with the dynamic qq-plot for the call holding data and Figure 7 does the same thing for the packet interarrival data. Figure 8 gives two static qq-plots for the call holding data, one using $k = 3500$ and the other using $k = 1500$, yielding estimates of $\alpha$ of 0.95 and 0.977, respectively. For this data set, the estimators are unusually insensitive to the choice of $k$; the Hill plots and the dynamic qq-plots are quite stable and the static qq-plots do not change much as $k$ varies. Figure 9 gives two static qq-plots for the packet interarrival data. The data set is only 783 in length and now there is some sensitivity to $k$.

3.1.5. *De Haan's moment estimator.* The extreme value distributions [Resnick (1987), de Haan (1970), Leadbetter, Lindgren and Rootzen (1983), Castillo (1988)] can be parameterized as a one-parameter family

$$G_\gamma(x) = \exp\{-(1+\gamma x)^{-\gamma^{-1}}\}, \qquad \gamma \in \mathbb{R},\ 1+\gamma x > 0.$$

When $\gamma = 0$, we interpret $G_0$ as the Gumbel distribution

$$G_0(x) = \exp\{-e^{-x}\}, \qquad x \in \mathbb{R}.$$

A distribution whose sample maxima when properly centered and scaled converges in distribution to $G_\gamma$ is said to be in the *domain of attraction* of $G_\gamma$,

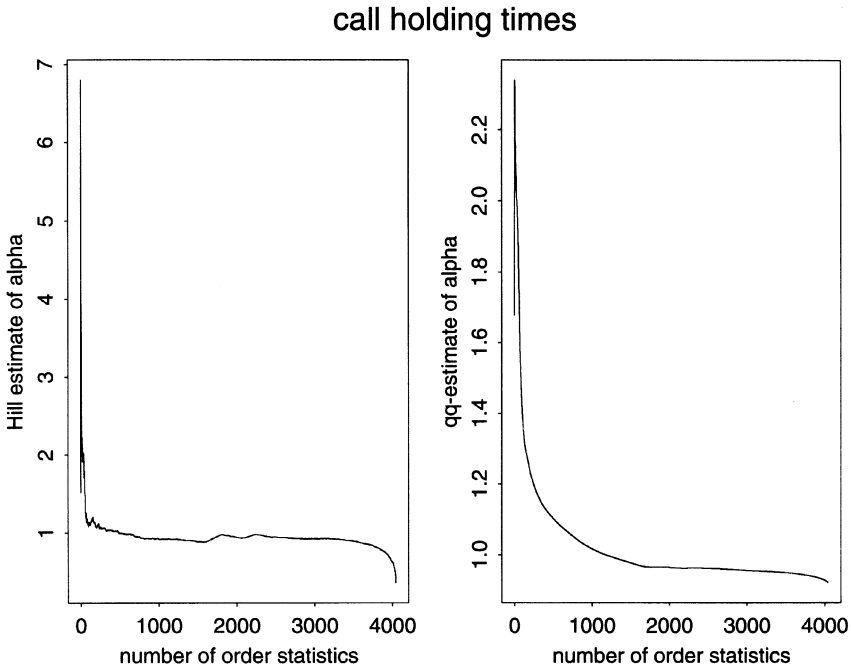


FIG. 6. *Hill and* qq*-plots for call holding times.*

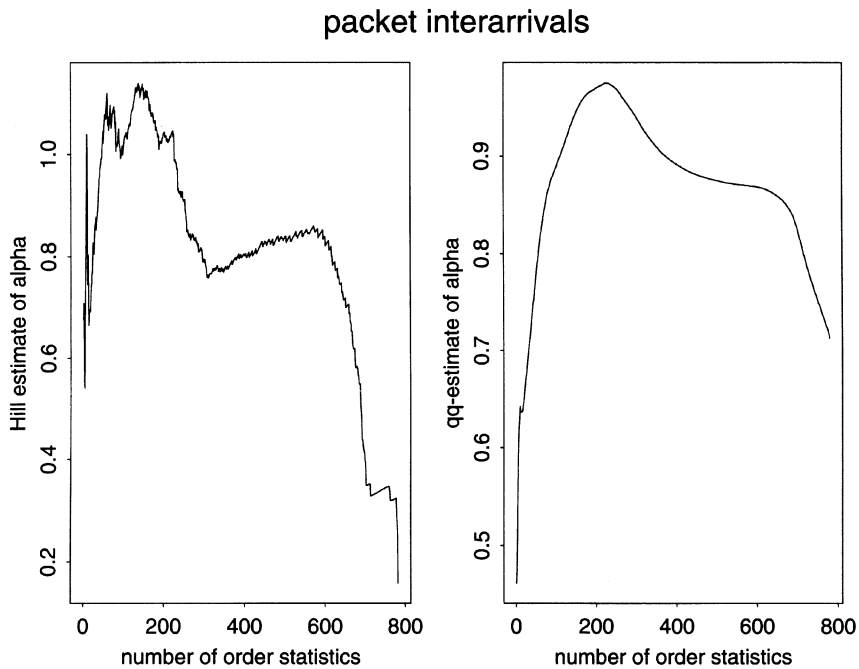


FIG. 7. *Hill and* qq*-plots for packet interarrivals.*

which is written $F \in D(G_\gamma)$. If $\gamma > 0$ and $F \in D(G_\gamma)$, then $1 - F \in RV_{-1/\gamma}$. De Haan's moment estimator $\hat{\gamma}$ [Dekkers, Einmahl and de Haan (1989), de Haan (1991), Dekkers and de Haan (1991), Resnick and Stărică (1996a)] is designed to estimate $\gamma$ from a random sample in the domain of attraction of $G_\gamma$. When $\gamma > 0$, this is the same thing as estimating $\gamma = 1/\alpha$. Since the exponential, normal, gamma densities and many others are in the $D(G_0)$, the domain of attraction of the Gumbel distribution, this provides another method of deciding when a distribution is heavy tailed or not. If $\hat{\gamma}$ is negative or very close to zero, there is considerable doubt that heavy tailed analysis should be applied.

The moment estimator is defined as follows: let $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ be the order statistics from a random sample of size $n$. Define for $r = 1, 2$,

$$H_{k,n}^{(r)} = \frac{1}{k} \sum_{i=1}^{k} \left( \log \frac{X_{(i)}}{X_{(k+1)}} \right)^r$$

so that $H_{k,n}^{(1)}$ is the Hill estimator. Define

$$\hat{\gamma}_n = H_{k,n}^{(1)} + 1 - \frac{1/2}{1 - (H_{k,n}^{(1)})^2 / H_{k,n}^{(2)}}. \tag{3.5}$$

Then, assuming $F \in D(G_\gamma)$, we have consistency

$$\hat{\gamma}_n \xrightarrow{P} \gamma,$$

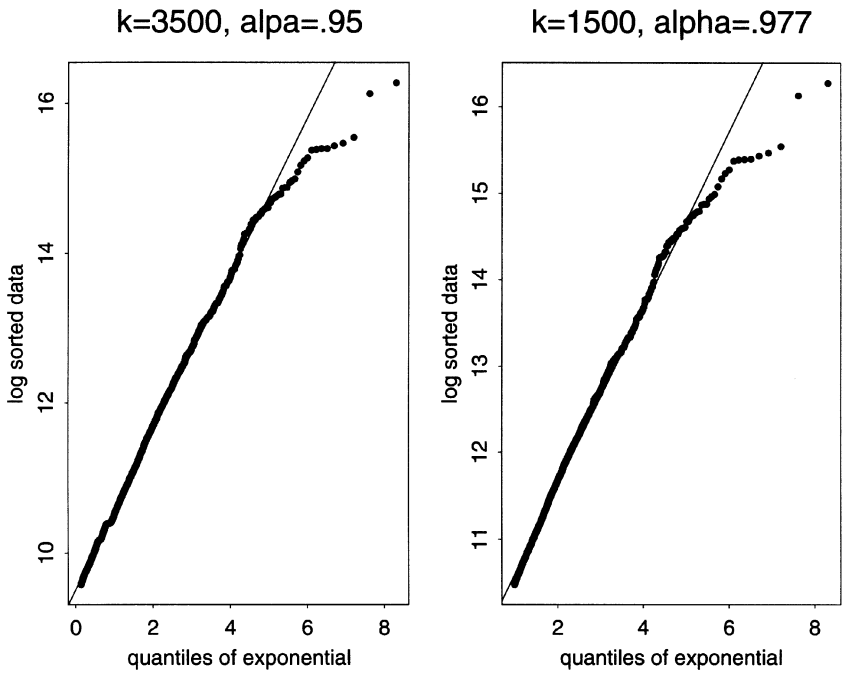


FIG. 8. *Static* qq*-plots for call holding times.*

as $n \to \infty$ and $k/n \to 0$. Furthermore under a second-order condition and a further restriction on $k$,

$$\sqrt{k}(\hat{\gamma} - \gamma) \Rightarrow N,$$

where $N$ is a normal random variable with 0 mean and variance

$$\sigma(\gamma) = \begin{cases} 1 + \gamma^2, & \text{if } \gamma \geq 0, \\ (1-\gamma)^2(1-2\gamma)\left(4 - 8\dfrac{1-2\gamma}{1-3\gamma} + \dfrac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right), & \text{if } \gamma < 0. \end{cases}$$

The asymptotic variance of the moment estimator exceeds that of the Hill estimator when $\gamma > 0$, so from the point of view of asymptotic variance, there is no reason to prefer it. However, the moment estimator discerns a light tail more effectively than the Hill estimator, and thus it is often useful to apply the moment estimator to see if $\gamma \leq 0$, which would rule out heavy tail analysis.

Figure 10 compares the effectiveness of the moment estimator for discerning a light tail ($\gamma = 0$) with that of of the Hill estimator $\alpha = \infty$. The Hill estimator does not seem very reliable for this purpose. Both estimators are applied to the same random sample of size 1000 taken from an exponential distribution with unit mean.

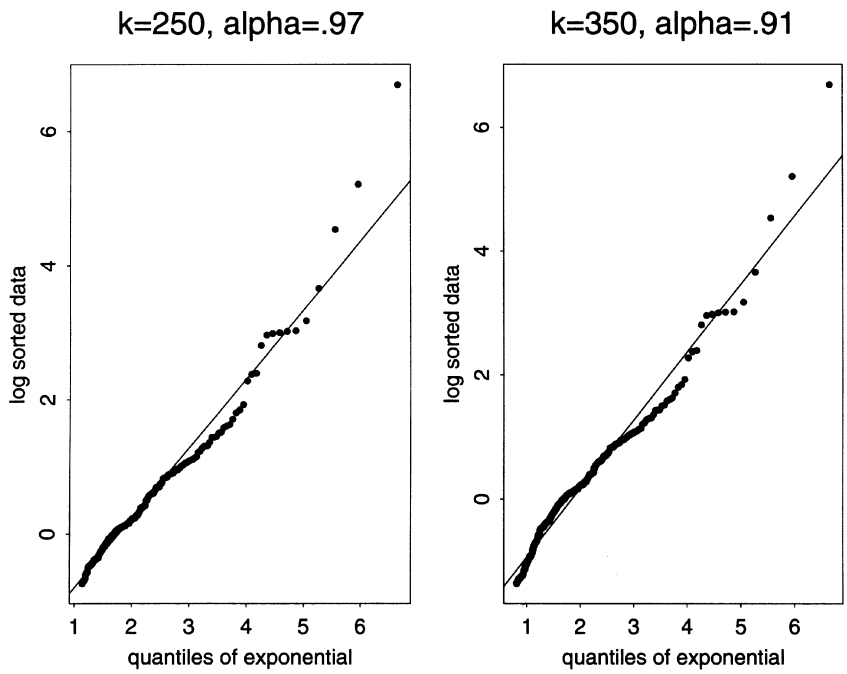


FIG. 9. *Static* qq*-plots for packet inter-arrivals.*

Figure 11 shows the moment estimator applied to the call holding data on the left and the packet interarrival data on the right. Keep in mind when comparing these graphs with previous graphs that $\gamma = 1/\alpha$.

3.2. *Are dependencies present?* Mature statistical computer packages have built-in routines to graph the classical sample autocorrelation function (acf) of the data given by

$$\hat{\rho}(h) = \frac{\sum_{i=1}^{n-|h|}(X_i - \bar{X})(X_{i+h} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{3.6}$$

In the classical $L_2$ case where the variance of the marginal distribution is finite and correlations exist, the sample correlation $\hat{\rho}(h)$ estimates the mathematical correlation $\rho(h)$ and in fact

$$\hat{\rho}(h) \xrightarrow{P} \rho(h).$$

In the heavy tailed case where variances and even means may be infinite, there is no point to the centering by $\bar{X}$ and the following heavy tailed modification is more appropriate:

$$\hat{\rho}_H(h) = \frac{\sum_{i=1}^{n-|h|} X_i X_{i+h}}{\sum_{i=1}^{n} X_i^2}. \tag{3.7}$$

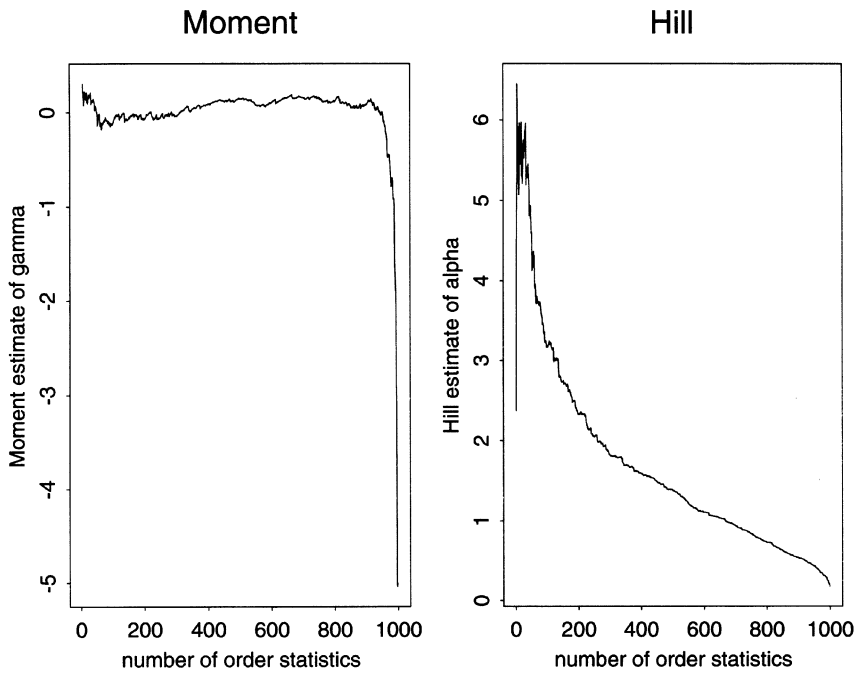


FIG. 10. *Unit exponential data.*

If we have the heavy tailed model given by (2.5),

$$X_t = \sum_{j=0}^{\infty} c_j Z_{t-j},$$

where $\{Z_t\}$ are nonnegative, iid and heavy tailed, the mathematical correlations do not exist if $\alpha < 2$. However $\hat{\rho}_H(h)$ still converges to a limiting constant [Davis and Resnick (1985a,b; 1986)]

$$\hat{\rho}_H(h) \xrightarrow{P} \frac{\sum_{j=0}^{\infty} c_j c_{j+h}}{\sum_{j=1}^{\infty} c_j^2} := \rho(h). \tag{3.8}$$

The mean corrected function given in (3.6) also converges in probability to the same limit. The limit law for $\hat{\rho}_H(h)$ is complex and is established in Davis and Resnick (1986, 1985b). The most tractable case is for $\alpha < 1$ and is given as follows.

PROPOSITION 3.2. *Suppose* $\alpha < 1$ *and for some* $\delta < \alpha$,

$$\sum_{j=0}^{\infty} j|c_j|^{\delta} < \infty.$$

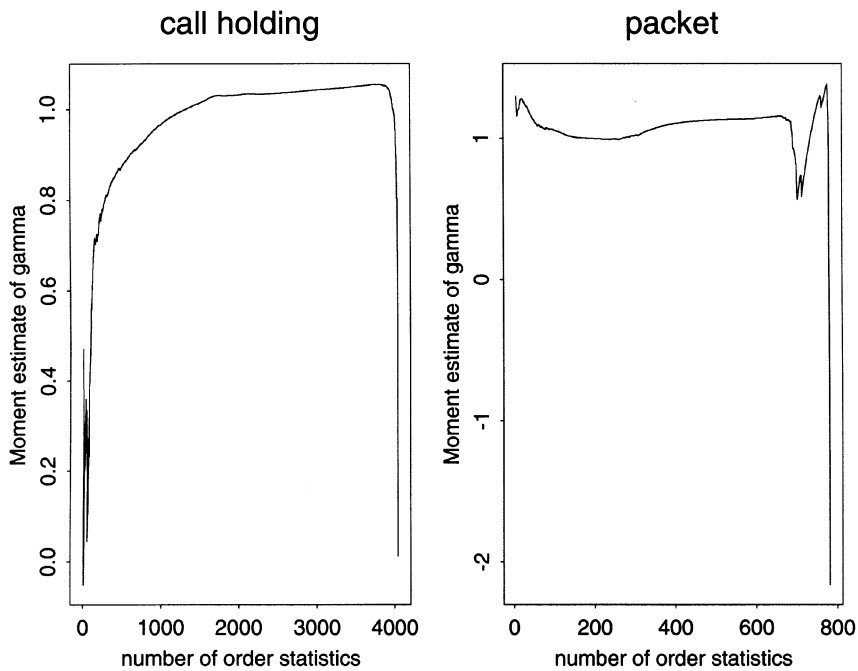


FIG. 11. *Moment estimator applied to call holding and packet data.*

*Define the quantile functions*

$$b(n) = \left(\frac{1}{1-F}\right)^{\leftarrow}(n), \qquad \tilde{b}(n) = \left(\frac{1}{P[Z_1 Z_2 > \cdot]}\right)^{\leftarrow}(n).$$

*Then for any* $l \geq 1$,

$$\left(\tilde{b}(n)^{-1} b(n)^2 (\hat{\rho}_H(h) - \rho(h)),\ 1 \leq h \leq l\right) \Rightarrow (Y_1, \ldots, Y_l),$$

*in* $\mathbb{R}^l$ *where*

$$Y_h = \sum_{j=1}^{\infty} (\rho(h+j) + \rho(h-j) - 2\rho(j)\rho(h)) \frac{S_j}{S_0}.$$

*Here* $S_0, S_1, \ldots$ *are independent and* $S_0$ *is one-sided stable of index* $\alpha/2$ *and* $S_1, S_2, \ldots$ *are iid, one-sided stable of index* $\alpha$.

Even in this simplest case, the limit distribution is quite complex. [Compare this with the classical Bartlett formula where the sample acf has a limiting normal distribution. See Brockwell and Davis (1991), page 221, ff, page 538, ff.] In distribution, $Y_h$ can be reexpressed as

$$\left(\sum_{j=1}^{\infty} |\rho(h+j) + \rho(h-j) - 2\rho(j)\rho(h)|^{\alpha}\right)^{1/\alpha} \frac{U}{V},$$

where $U$ and $V$ are independent, nonnegative stable random variables and the index of $V$ is $\alpha/2$ and the index of $U$ is $\alpha$. So even if the value of $\alpha$ were known, the percentiles of the distribution of $Y_h$ are not easy to obtain and certainly impossible to obtain analytically.

Nonetheless, $\hat{\rho}_H$ can be used as an exploratory tool to make preliminary investigations of dependence. Note that if the MA($\infty$) process $\{X_t\}$ is iid, so that $X_t = Z_t$, then $c_j = 0$ for $j \geq 1$ and for $h \geq 1$,

$$\hat{\rho}_H(h) \xrightarrow{P} 0.$$

This give an exploratory indication of independence: if on graphing the sample heavy tailed acf, one finds only small values, then it may be possible to model the data as iid. Similarly, if the sample acf is small beyond lag $q$, then there is some evidence that MA($q$) may be an appropriate model. Of course, without firm knowledge of the quantiles of the limit distribution of $\hat{\rho}_H(h)$, it is impossible to say with precision what *small* means.

Figure 12 shows the classical acf for the call holding data side by side with the heavy tailed modification. The right graph of the heavy tailed sample acf has a dotted line drawn at height $h = 0.0035$ and the interval $[0, h]$ is a 95% confidence window analogous to the one given by Bartlett's formula [Brockwell and Davis (1991)] in classical time series. The confidence window

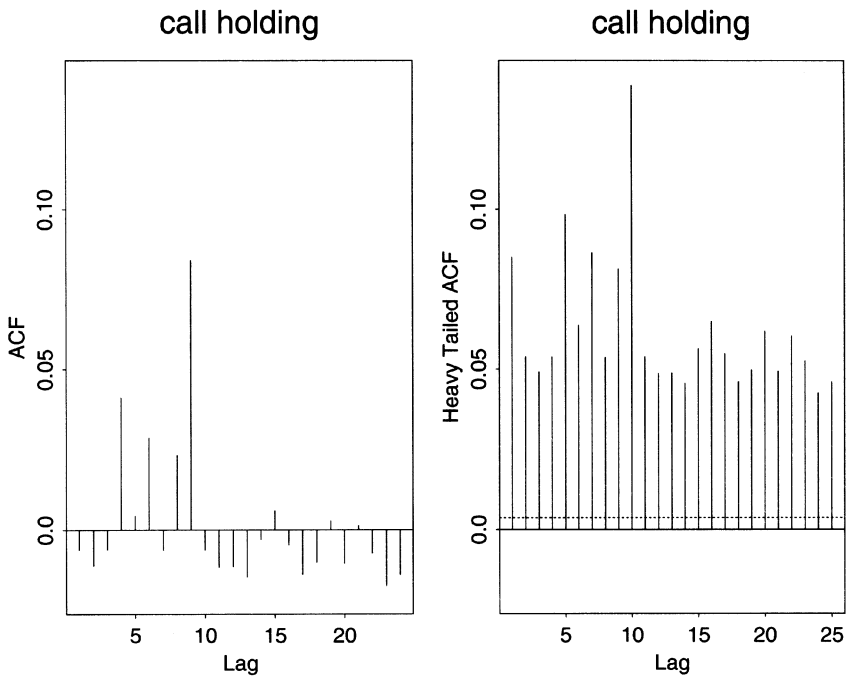


FIG. 12. *Call holding times.*

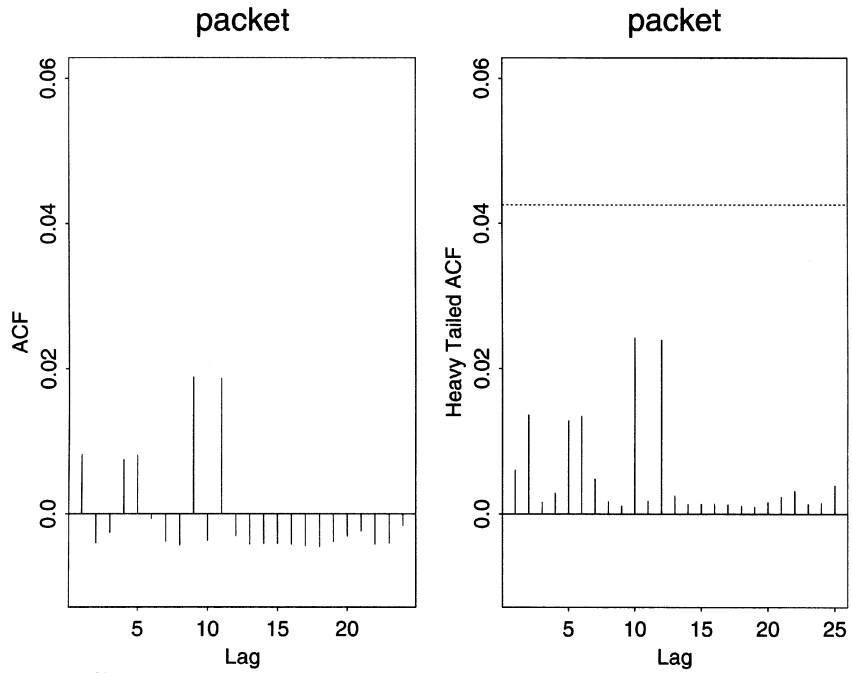


FIG. 13. *Packet inter-arrivals.*

is drawn based on the assumption that the data is independent and has Pareto tails. According to Theorem 3.3 of Davis and Resnick (1986), $h$ is given by

$$h = l\alpha^{1/\alpha}\frac{n^{-1/\alpha}}{\log n},$$

where $l$ is the quantile satisfying

$$P[U/V \leq l] = 0.95$$

for $U$, $V$ independent positive stable random variables with indices $\alpha$ and $\alpha/2$. In the case of the call holding data, we used the estimated value of $\alpha = 0.97$. The quantile $l$ was estimated by simulation. The position of $h$ relative to the heights of the sample heavy tailed acf values casts serious doubts on the assumption of independence. Figure 13 exhibits comparable graphs for the packet interarrival data. The right heavy tailed graph does not offer evidence against the hypothesis of independence.

**4. Modeling dependent data.** Modeling data which is not a realization of an iid model presents great challenges. There is no guarantee that the simple linear ARMA models form a large enough subclass of the class of heavy tailed models that hunting within this class will yield a model with an acceptable fit to the data.

From the theoretical point of view, there has been considerable success in analyzing autoregressive models of order $p$, abbreviated AR($p$). These are models of the form

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + Z_t, \qquad t = 0, \pm 1, \pm 2, \ldots, \tag{4.1}$$

where $\{Z_t\}$ is iid and heavy tailed.

Several methods have been proposed for estimating the coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. These methods are all consistent and asymptotic distributions have been worked out. They include:

1. Yule–Walker estimators; after slight modification, these work for heavy tailed autoregressions [Davis and Resnick (1985a,b; 1986)],
2. Spectral density estimators [Mikosch, Gadrich, Klüppelberg and Adler (1995)],
3. Least gamma deviation estimators [Davis, Knight and Liu (1991)],
4. Linear programming (LP) estimators [Feigin and Resnick (1992, 1994, 1997); Feigin, Resnick and Stărică (1995); Feigin, Kratz and Resnick (1996)].

The LP and least gamma deviation estimators have the best rate of convergence. We will only deal with the Yule–Walker and LP methods. For both methods we review model fitting and estimation.

4.1. *The Yule–Walker method.* The classical Yule–Walker method is based on sample correlations. In the heavy tailed context, the heavy tailed sample correlation function is the basis for fitting and estimation.

4.1.1. *Yule–Walker estimation.* Recall the definition of $\hat{\rho}_H(h)$ and $\rho(h)$ from (3.7) and (3.8). Suppose $0 < \alpha < 2$ and that $\{X_t\}$ is a stationary, invertible autoregressive process of order $p$ of the form given by (4.1) which can be inverted and written as the MA($\infty$) process

$$X_t = \sum_{j=0}^{\infty} c_j Z_{t-j}. \tag{4.2}$$

Write $c_j = 0$ if $j < 0$ so that

$$\rho(-i) = \frac{\sum_{k=i}^{\infty} c_k c_{k-i}}{\sum_{k=0}^{\infty} c_k^2} = \rho(i).$$

Set

$$\mathbf{R} = (R_{ij})_{i,\,j=1}^{p} = (\rho(i-j))_{i,\,j=1}^{p}, \qquad \boldsymbol{\rho} = (\rho(1), \ldots, \rho(p))'$$

and we have the Yule–Walker equation

$$\mathbf{R}\boldsymbol{\phi} = \boldsymbol{\rho}, \tag{4.3}$$

where recall $\boldsymbol{\phi}$ is the $p$-vector of autoregressive coefficients in (4.1). Furthermore, for every $m$, the matrix

$$\mathbf{R}_m = (\rho(i-j))_{i,j=1}^{m} \tag{4.4}$$

is nonsingular, provided $\sum_k c_k^2 > 0$. The heavy tailed Yule–Walker estimator $\hat{\boldsymbol{\phi}}^{YW}$ of $\boldsymbol{\phi}$ satisfies

$$\hat{\mathbf{R}}\hat{\boldsymbol{\phi}}^{YW} = \hat{\boldsymbol{\rho}}_H, \tag{4.5}$$

where

$$\hat{\mathbf{R}} = (\hat{\rho}_H(i-j))_{i,j=1}^{p}, \qquad \hat{\boldsymbol{\rho}}_H = (\hat{\rho}_H(1), \ldots, \hat{\rho}_H(p))'.$$

Since $\hat{\boldsymbol{\rho}}_H \to_P \boldsymbol{\rho}$ [Davis and Resnick (1985a)] and $\hat{\mathbf{R}} \to_P \mathbf{R}$ as $n \to \infty$, the consistency of the heavy tailed Yule–Walker estimators follows.

Furthermore, in nice cases (e.g., when $\alpha < 1$),

$$\tilde{b}(n)^{-1} b(n)^2 (\hat{\boldsymbol{\phi}}^{YW} - \boldsymbol{\phi})$$

has a limit distribution which is a function of the limit distribution obtained for the sample correlation function [Davis and Resnick (1986)]. The rate of convergence of this limit distribution as measured by $\tilde{b}(n)^{-1} b(n)^2$ is inferior to that of the linear programming estimators.

4.1.2. *Model selection based on sample correlations: pacf and AIC.* Standard techniques for model selection based on sample correlations are available for heavy tailed analysis after minor modification. Define

$$\boldsymbol{\phi}_m^* = \begin{cases} (\phi_1, \ldots, \phi_m)', & \text{if } m \le p, \\ (\phi_1, \ldots, \phi_p, 0, \ldots, 0)', & \text{if } m \ge p \end{cases}$$

and

$$\boldsymbol{\rho}_m = (\rho(1), \ldots, \rho(m))'.$$

For $m > p$ we have

$$\mathbf{R}_m \boldsymbol{\phi}_m^* = \boldsymbol{\rho}_m$$

and so

$$\boldsymbol{\phi}_m^* = \mathbf{R}_m^{-1} \boldsymbol{\rho}_m.$$

Recall that in classical time series analysis, the $m$th component on the right would be the partial autocorrelation at lag $m$ [Brockwell and Davis (1991), page 102] and we call the $m$th component of the $m$-vector

$$\hat{\boldsymbol{\phi}}_m^* = \hat{\mathbf{R}}_m^{-1} \hat{\boldsymbol{\rho}}_m \tag{4.6}$$

the sample heavy tailed partial autocorrelation function (pacf) at lag $m$. For $m > p$ we have

$$\hat{\boldsymbol{\phi}}_m^* \xrightarrow{P} \boldsymbol{\phi}_m^*$$

in $\mathbb{R}^m$ so that for the $m$th component we have, when $m > p$,

$$\hat{\phi}^*_{m,m} \xrightarrow{P} 0.$$

Again, in simple cases such as when $\alpha < 1$, we have that for $m > p$,

$$\tilde{b}_n b_n^2(\hat{\boldsymbol{\Phi}}^*_m - \boldsymbol{\Phi}^*_m)$$

has a limit distribution depending on the limit achieved for the sample acf. This of course means that the $m$th component $\tilde{b}_n b_n^2 \hat{\phi}^*_{m,m}$ has a limit distribution.

This yields an exploratory technique for diagnosing when an autoregression might be a suitable candidate model for a data set. Graph the heavy tailed sample pacf and if it dies after $p$ lags, try fitting an AR($p$). Again, because of the complexity of the limit distribution, there is difficulty in deciding at what lag the graph has died.

The classical AIC criterion for order selection is not consistent for selecting finite variance models since it tends to overestimate the order. Brockwell and Davis (1991) have a nice example where they simulate an AR(1) process and then apply the AIC criterion, which insists that the simulated process is an AR(2). However, the AIC criterion is consistent for heavy tails [Knight (1989a), Bhansali (1988)]. Define

$$\hat{\sigma}^2(0) = \frac{1}{n}\sum_{i=1}^n X_i^2,$$

$$\hat{\sigma}^2(m) = \hat{\sigma}^2(0)\prod_{j=1}^m (1 - \hat{\phi}^*_{j,j})^2, \qquad m \geq 1.$$

The heavy tailed AIC function is defined by

$$AIC(k) = n \log \hat{\sigma}^2(k) + 2k,$$

and the estimate of $p$ is obtained by minimizing this function:

$$\hat{p} = \operatorname*{argmin}_{k \leq K} AIC(k),$$

where $K$ is an upper bound which is assumed to exist. As $n \to \infty$, $\hat{p} \to_P p$, the true order.

Graphing $\{\text{AIC}(k),\ k \geq 1\}$ helps in determining the order. In Figure 14 we display the heavy tailed pacf plot and the AIC plot $\{(k, \text{AIC}(k)),\ 1 \leq k \leq 15\}$ for 500 simulated data from the heavy tailed AR(2) process

$$X_t = 1.3X_{t-1} - 0.7X_{t-2} + Z_t, \tag{4.7}$$

where $\{Z_t\}$ are iid with standard Pareto distribution $P[Z_1 > x] = 1/x,\ x \geq 1$. Both techniques are seen to work quite well. However, for real data, it may

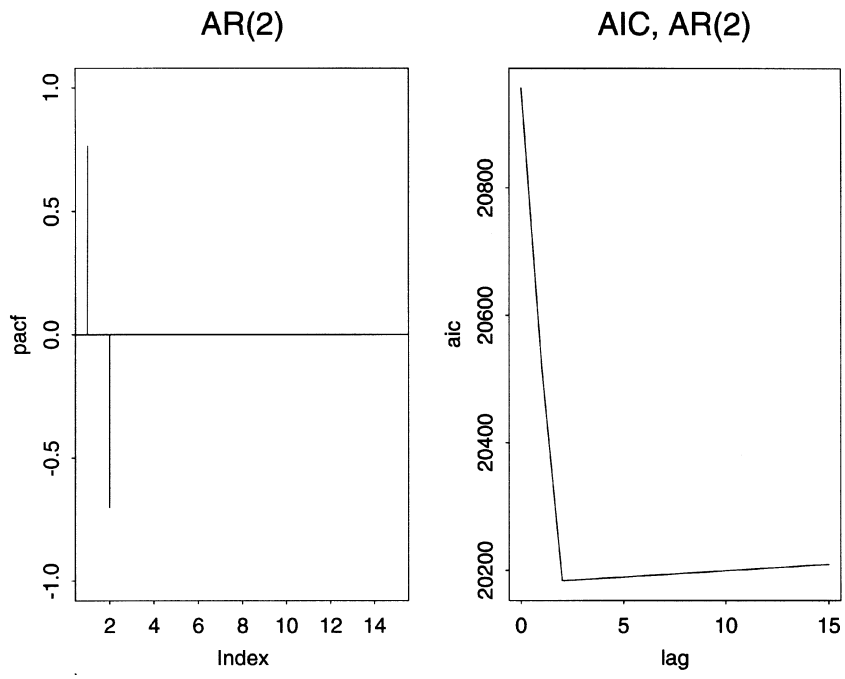


FIG. 14. *Heavy tailed* pacf *and* AIC *for simulated* AR(2).

be the case that even though AIC($k$) has a nice minimum, residual analysis of the model fitted by the AIC criterion reveals a lack of iid structure in the estimated residuals, throwing into doubt the goodness of the fit.

4.2. *Linear programming estimators.* The linear programming (LP) estimators were devised by Feigin and Resnick (1992, 1994) to capitalize on the assumption that the residuals $\{Z_t\}$ in (4.1) are nonnegative. This work built on ideas of Davis and McCormick (1989) and Andel (1989).

4.2.1. *Linear programming estimation.* Assume the innovations $\{Z_t\}$ in (4.1) are nonnegative with a common distribution whose left endpoint is 0. The LP estimators of the AR($p$) coefficients $\boldsymbol{\phi}$ are

$$\hat{\boldsymbol{\phi}}(n) = \arg\max_{\boldsymbol{\delta} \in D_n} \boldsymbol{\delta}' \mathbf{1} \tag{4.8}$$

where $\mathbf{1}' = (1, \ldots, 1)$ and where the feasible region $D_n$ is defined as

$$D_n = \bigcap_{t=1}^{n} \left\{ \boldsymbol{\delta} \in R^p \colon X_t - \sum_{i=1}^{p} \delta_i X_{t-i} \geq 0 \right\}. \tag{4.9}$$

Here is a rapid review of one derivation of this estimator: suppose temporarily that the common distribution of the $Z$'s is unit exponential so that

$$P[Z_1 > x] = e^{-x}, \qquad x > 0.$$

In this case, conditionally on $\{X_0 = x_0,\ X_{-1} = x_{-1}, \ldots, X_{-p+1} = x_{-p+1}\}$, the likelihood is proportional to [using $I(\cdot)$ for the indicator function]:

$$I\Big(\bigwedge_{t=1}^{n}\Big(X_t - \sum_{i=1}^{p}\phi_i X_{t-i}\Big) \geq 0\Big)\exp\Big\{\phi_1\sum_{1}^{n}X_{t-1} + \phi_2\sum_{1}^{n}X_{t-2} + \cdots + \phi_n\sum_{1}^{n}X_{t-p}\Big\}$$
$$\approx I\Big(\bigwedge_{t=1}^{n}\Big(X_t - \sum_{i=1}^{p}\phi_i X_{t-i}\Big) \geq 0\Big)\exp\Big\{\sum_{i=1}^{p}\phi_i\sum_{t=1}^{n}X_t\Big\}.$$

Assuming that $\sum_{t=1}^{n} X_t$ is ultimately positive, the corresponding maximum likelihood estimator will thus be approximately determined by solving the linear program (LP)

$$\max\Big(\sum_{i=1}^{p}\phi_i\Big)$$

subject to

$$X_t \geq \sum_{i=1}^{p}\phi_i X_{t-i}; \qquad t = 1, \ldots, n.$$

The simplified approximate form of the objective function is justified by the fact that $\sum_1^n X_{t-1}/\sum_1^n X_{t-i} \approx 1$.

Now drop the assumption that the density of the $Z$'s is exponential. The LP estimator still gives an estimation procedure with good properties which is applicable to the heavy tailed case.

The consistency and asymptotic distribution for the LP estimator was established in Feigin and Resnick (1992, 1994) and Feigin, Resnick and Stărică (1995).

THEOREM 4.1. *Suppose $\{X_t\}$ is a stationary autoregression given by* (4.1) *where $\{Z_t\}$ are iid nonnegative innovation variables with common distribution $F$, which has a regularly varying tail of index $-\alpha$. Suppose also that for some $\beta > \alpha$,*

$$EZ_1^{-\beta} = \int_0^\infty u^{-\beta}F(du) < \infty.$$

*Let $\hat{\boldsymbol{\phi}}(n)$ be the linear programming estimator based on $X_1, \ldots, X_n$ given in* (4.8) *and* (4.9). *Let $\{E_j,\ j \geq 1\}$ be iid unit exponential random variables and define*

$$\Gamma_k = E_1 + \ldots + E_k, \qquad k \geq 1,$$

*so that $\{\Gamma_k\}$ are the points of a homogeneous Poisson process. Define $b_n$ by*

$$b_n = \Big(\frac{1}{1-F}\Big)^{\leftarrow}(n) = F^{\leftarrow}\Big(1 - \frac{1}{n}\Big)$$

*and for* $|z| \le 1$ *set*

$$C(z) = \sum_{j=0}^{\infty} c_j z^j = \frac{1}{\Phi(z)}, \qquad \Phi(z) = 1 - \sum_{i=1}^{p} \phi_i(0) z^i,$$

*where* $\{\phi_i(0),\ i = 1, \ldots, p\}$ *are the true autoregressive coefficients. Then*

$$b_n(\hat{\boldsymbol{\phi}}(n) - \boldsymbol{\phi}(0)) = O_p(1)$$

*so the rate of convergence of* $\hat{\boldsymbol{\phi}}(n)$ *to* $\boldsymbol{\phi}(0)$ *is* $b_n$. *Furthermore, if for any* $p-1$ *distinct indices* $\{l_1, \ldots, l_{p-1}\}$, *for which the set of* $p$ *vectors*

$$\{\mathbf{1}, (c_{l_j}, c_{l_j - 1}, \ldots, c_{l_j - p + 1});\ 1 \le j \le p-1\}$$

*does not contain the zero vector, the set is also linearly independent, then*

$$b_n(\hat{\boldsymbol{\phi}}(n) - \boldsymbol{\phi}(0)) \Rightarrow \mathbf{L},$$

*where* $\mathbf{L}$ *is nondegenerate,*

$$\mathbf{L} \stackrel{d}{=} \arg\max_{\boldsymbol{\delta} \in \Lambda} \boldsymbol{\delta}' \mathbf{1} \tag{4.10}$$

*and*

$$\Lambda = \{\boldsymbol{\delta} \in \mathbb{R}^p \colon \boldsymbol{\delta}' \mathbf{1} \ge -1, \boldsymbol{\delta}' \mathbf{v}_k \le 1,\ k \ge 1\}. \tag{4.11}$$

*The points* $\{\mathbf{v}_k\}$ *are specified as follows. Let* $\{Y_{kl},\ k \ge 1,\ l \ge 0\}$ *be a doubly infinite array of iid random variables which is independent of* $\{\Gamma_k\}$ *with the distribution* $F$. *Then*

$$\mathbf{v}_l = \left( \bigvee_{k=1}^{\infty} \Gamma_k^{-1/\alpha} Y_{kl}^{-1} \right) (c_{l-1}, \ldots, c_{l-p})' = V_l(c_{l-1}, \ldots, c_{l-p}); \qquad l = 1, 2, \ldots .$$

Note that the asymptotic distribution is complicated and depends on a limiting Poisson process and the unknown distribution of the autoregression. In one case, however, namely if $\boldsymbol{\phi}(0) = \mathbf{0}$ making $X_t = Z_t$, the limit distribution considerably simplifies and this is the basis of a test for independence discussed in the next subsection.

We applied the LP estimator and the Yule–Walker estimator to a sample of size 100 from the AR(2) described in (4.7). The LP estimator yielded $(\phi_1, \phi_2)$ values of (1.30202, −0.7004143), which is quite good performance in view of the small sample size. The corresponding Yule–Walker were not nearly so accurate and were (0.9084571, −0.4122331).

4.2.2. *Model selection and confirmation.* The LP estimator can be used to fashion a test for independence against autoregressive alternatives. Test if

$$\phi_1 = \cdots \phi_p = 0$$

by rejecting when

$$\bigvee_{i=1}^{p} |\hat{\phi}_i(n)|$$

is large [Feigin, Resnick and Stărică (1995)]. This also provides a model selection tool since a well-fitted model should have the property that the estimated residuals

$$\hat{Z}_t(n) := X_t - \sum_{i=1}^{p} \hat{\phi}_i(n) X_{t-i}, \qquad t = p+1, \ldots, n$$

are approximately iid and we may test this using the independence test.

It would not be possible to fix the size of the test if the limit distribution of the LP estimator did not considerably simplify. Fortunately it does and under the null hypothesis of $\boldsymbol{\phi}(0) = \mathbf{0}$,

$$b_n \hat{\boldsymbol{\phi}}(n) \Rightarrow \mathbf{L} \equiv (V_1^{-1}, \ldots, V_p^{-1}),$$

where for $x_i \geq 0;\ i = 1, \ldots, p$ we have that

$$\begin{aligned} &P[V_i \leq x_i,\ i = 1, \ldots, p] \\ &\quad = \exp\left\{ - \int_{(y_1, \ldots, y_p) \in [0,\infty)^p} \left( \bigwedge_{l=1}^{p} y_l x_l \right)^{-\alpha} F(dy_1) \cdots F(dy_p) \right\}. \end{aligned} \tag{4.12}$$

This means that if we want a 0.05 level rejection region, we should reject when

$$P\left[ \bigvee_{i=1}^{p} |\hat{\phi}_i(n)| > K(0.05) \right] = 0.05$$

and to find an approximate value of $K(0.05)$ we write

$$\begin{aligned} P\left[ \bigvee_{i=1}^{p} |\hat{\phi}_i(n)| > K(0.05) \right] &\approx P\left[ \bigvee_{i=1}^{p} L_i > b_n K(0.05) \right] \\ &\leq p P[L_1 > b_n K(0.05)] \\ &= p \exp(-c(b_n K(0.05))^{\alpha}), \end{aligned} \tag{4.13}$$

where $c = E(Z_1^{-\alpha})$. This yields

$$K(0.05) \approx \frac{(-\log(0.05/p)/c)^{1/\alpha}}{b_n} = \frac{(\log(20p)/c)^{1/\alpha}}{b_n}.$$

We need to estimate $\alpha, c$ and $b_n$. The qq-plot yields both $\hat{b}_n$ and $\hat{\alpha}$ and then we can get

$$\hat{c} = n^{-1} \sum_{i=1}^{n} X_i^{-\hat{\alpha}}.$$

Figure 15 shows the implementation of the independence test for the data set of packet interarrivals. The earlier discussion of the estimate of $\alpha$ was not entirely conclusive so the test was conducted three times with various values of $\alpha$ which are shown in the figure. We tested against the alternative for five autoregressive coefficients. Because the subadditive approximation does more damage for larger values of $p$, we limited ourselves to $p = 5$. The test fails

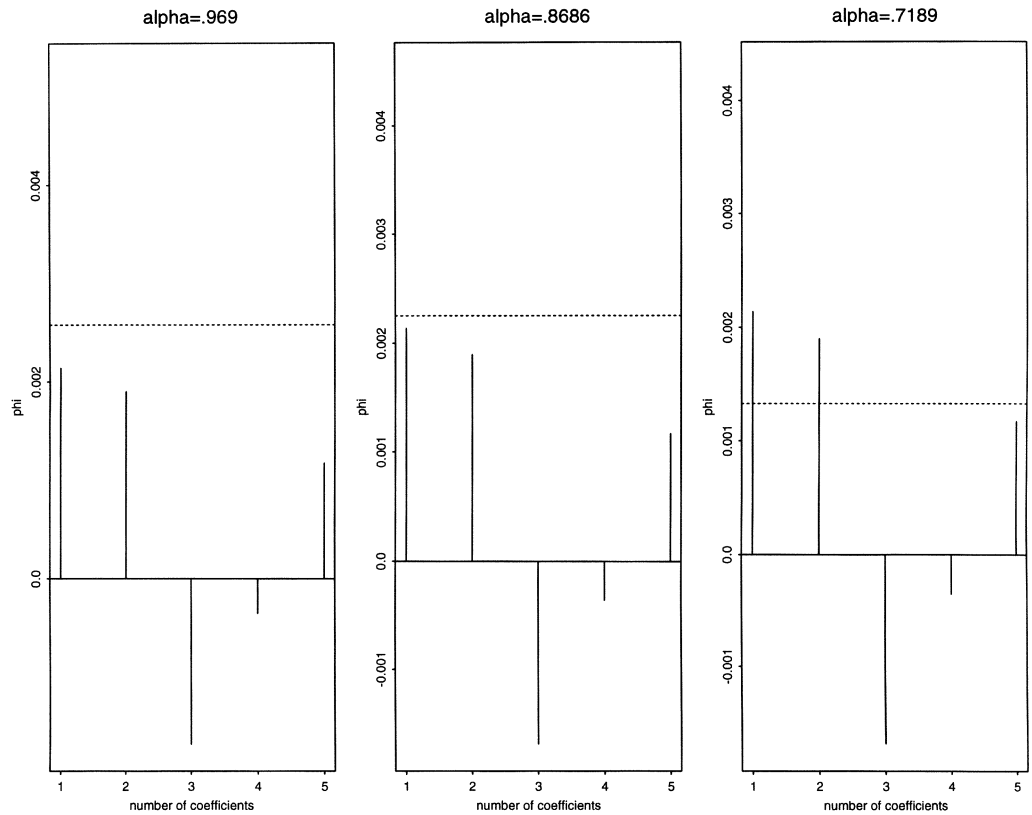


FIG. 15. *Tests for independence for packet interarrivals.*

only for the improbably small estimate of $\alpha \approx 0.7$ and this coupled with the heavy tailed acf and pacf plots indicates strong evidence for independence. In the graphs, the dotted horizontal line represents $K(0.05)$ and the vertical lines represent $(\hat{\phi}_i^{(n)},\ i = 1, \ldots, 5\}$.

4.2.3. *Left tail analysis.* The linear programming estimators and test for independence are designed to work under the assumption that either the right tail is regularly varying as in (2.3) or that the left tail is regularly varying [Feigin and Resnick (1994a)]. For the left tail case, the precise assumptions that ensure consistency and an asymptotic distribution for the LP estimator are as follows.

CONDITION M (Model specification). The process $\{X_t\colon t = 0, \pm 1, \pm 2, \ldots\}$ satisfies the equations

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t, \qquad t = 0, \pm 1, \pm 2, \ldots,$$

where $\{Z_t\}$ is an independent and identically distributed sequence of random variables with essential infimum (left endpoint) equal to 0 and common distribution function $F$.

CONDITION S (Stationarity). The coefficients $\phi_1, \ldots, \phi_p$ satisfy the stationarity condition that the autoregressive polynomial $\Phi(z) \equiv 1 - \sum_1^p \phi_i z^i$ has no roots in the unit disk $\{z : |z| \leq 1\}$. Furthermore we assume $\Phi(1) > 0$; that is, we require

$$\sum_{i=1}^{p} \phi_i < 1.$$

CONDITION L (Left tail). The distribution $F$ of the innovations $Z_t$ satisfies, for some $\alpha > 0$:

$$\lim_{s \downarrow 0} \frac{F(sx)}{F(s)} = x^{\alpha} \quad \text{for all } x > 0; \tag{4.14a}$$

$$E(Z_t^{\beta}) = \int_0^{\infty} u^{\beta} F(du) < \infty \quad \text{for some } \beta > \alpha. \tag{4.14b}$$

A notable success in the left tail case was given in Feigin, Resnick and Stărică (1995) where the LP estimator was used to fit an autoregression to the lynx data and the test for independence was used to fine tune the fit and perform model confirmation.

Under regular variation of either the left or right tail of $Z_1$, the rate of convergence of the LP estimator is of the order of $n^q$ where $q$ is the reciprocal of the index of variation. For some phenomena, the distribution of $Z_1$ may have both tails regularly varying and in this case best results are obtained by working with the heavier tail, that is, the tail with the smallest index. This achieves the best rate of convergence.

Figure 16 displays four views of the Hill plot for 1/callholding, giving strong indication that the *left* tail of the call holding data is regularly varying with an index in the neighborhood of $\alpha = 4$. Figure 17 gives a qq-plot yielding an estimate of 3.52. Since the left tail parameter is so much bigger than the right tail index, there is little temptation to switch to left tail analysis.

4.2.4. *The bootstrap and LP estimation.* A review of Theorem 4.1 affirms that the limit distribution for the LP estimators cannot be calculated explicitly except in the case of independence. To overcome this difficulty, a bootstrap procedure can be devised [Feigin and Resnick (1997), Datta and McCormick (1995)]. Caution: it remains to be seen how practical such procedures will be, and right now there is a sizable gap between theory and practice.

The proof [Feigin and Resnick (1997)] of the validity of the bootstrap depends heavily on a stochastic version of Karamata's theorem where $1 - F$ is replaced by $\nu_n(x, \infty]$ where

$$\nu_n(x, \infty] = \frac{m}{n} \sum_{t=1}^{n} \varepsilon_{Z_t/b(m)}$$

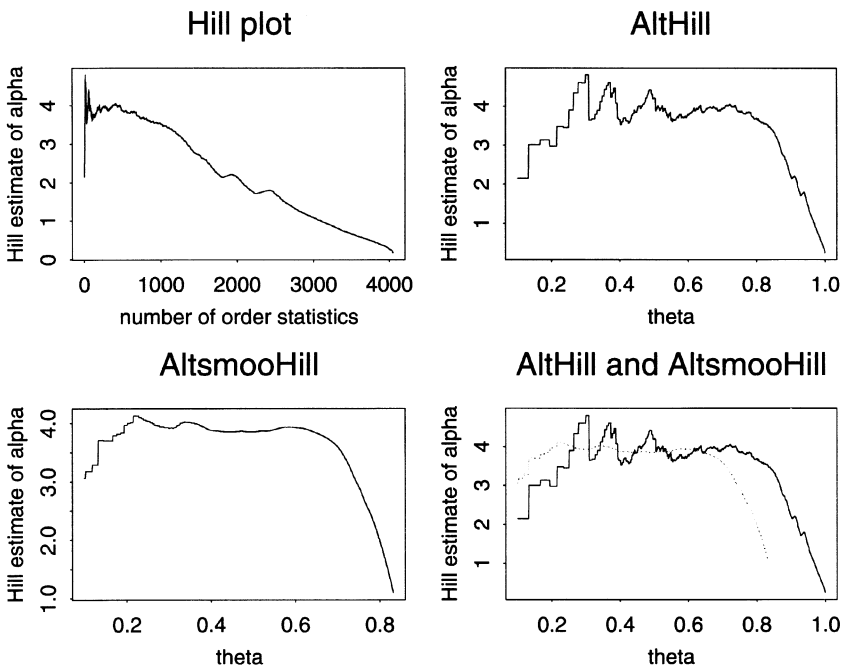


FIG. 16. *Hill plots for left tail of call holding data.*

and also by a version of the Karamata theorem where $1 - F$ is replaced by $\hat{\nu}_n(x, \infty]$ where

$$\hat{\nu}_n = \frac{m}{n} \sum_{t=1}^{n} \varepsilon_{\hat{Z}_t(n)/b(m)}$$

and $\hat{Z}_t(n),\ t = 1, 2, \ldots$ are the estimated residuals.

The necessity for the bootstrap sample size to be $m = o(n)$ makes use of the bootstrap difficult in practice. Just as we had difficulty picking $k$ when using the Hill estimator, for the bootstrap we must pick $m$ without reliable guidelines. In connection with bootstrapping extremes and heavy tailed phenomena, many authors have noticed that if the original sample is of size $n$, in order for the bootstrap asymptotics to work as desired the bootstrap sample should be of size $m$ where $m$ is a function of $n$ and $m/n \to 0$ as $n \to \infty$. See for example Athreya (1987), Giné and Zinn (1989), Hall (1990), Kinateder (1992), Knight (1989b), LePage (1992) and Deheuvels, Mason and Shorack (1993). Some perspective on this necessity to reduce the bootstrap sample size to something of smaller order than the observed sample size is provided in the discussion of the behavior of random measures in Proposition 2.1. See Feigin and Resnick (1997) for a discussion.

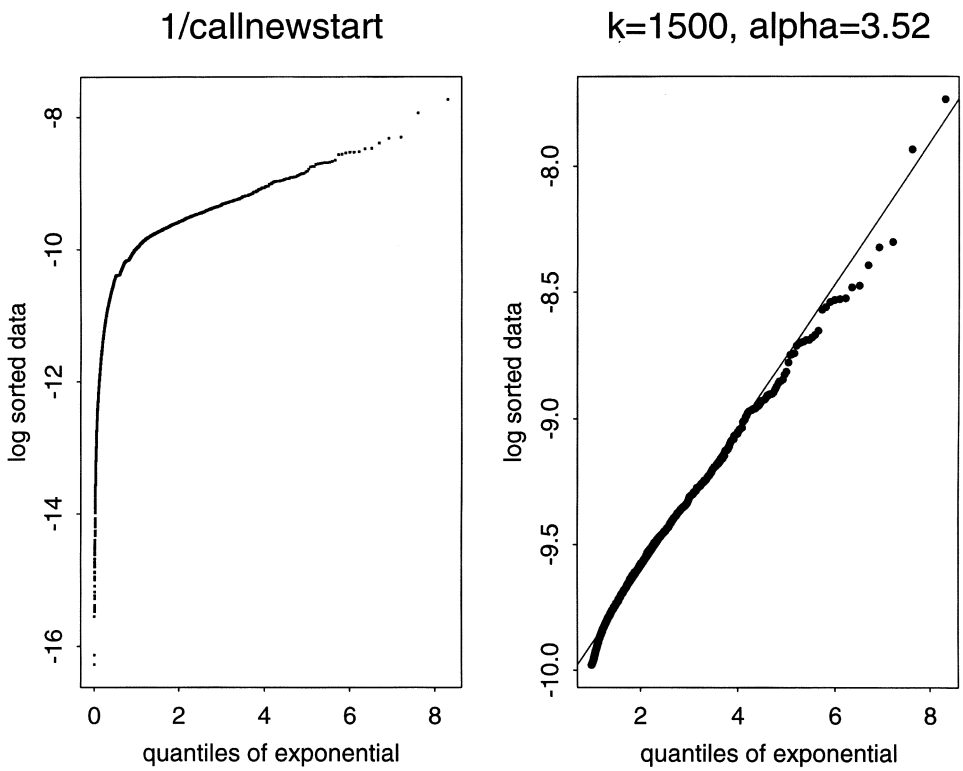


FIG. 17. qq-*plots for left tail of call holding data.*

4.2.5. *Estimating $\alpha$ for autoregressions.* Suppose you observe $X_1, \ldots, X_n$ from a heavy tailed autoregression

$$X_t = \sum_{i=1}^{p} \phi_j X_{t-j} + Z_t, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{4.15}$$

where $Z_t \geq 0$ and

$$P[Z_1 > x] = x^{-\alpha} L(x). \tag{4.16}$$

There are two possible ways to estimate $\alpha$.

1. Apply the Hill estimator directly to $X_1, \ldots, X_n$, that is,

$$H_{k,n}(X) = \frac{1}{k} \sum_{i=1}^{k} (\log X_{(i)} - \log X_{(k+1)}).$$

This seems a sensible thing to do since the tail of $X_1$ contains the same information as the tail of $Z_1$ by a result of Cline (1983), which says that

$$P[X_1 > x] \sim (\text{const}) P[Z_1 > x].$$

We know this works from Proposition 3.1.

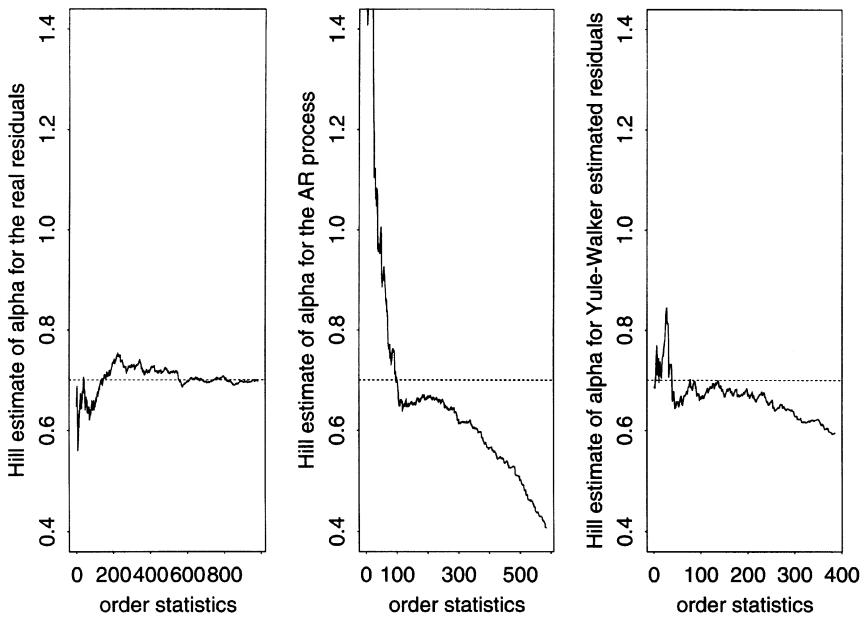


FIG. 18. *Hill plots for an autoregression.*

2. Alternatively we could estimate autoregressive coefficients $\boldsymbol{\phi}$ with a consistent estimator $\hat{\boldsymbol{\phi}}^{(n)}$ and then estimate the residuals

$$\hat{Z}_t(n) = X_t - \sum_{i=1}^{p} \hat{\phi}_i^{(n)} X_{t-i}, \ t = 1, \ldots, p.$$

These estimate $Z_1, \ldots, Z_n$. If we apply Hill's estimator to the estimated residuals we should get a sensible procedure:

$$H_{k,n}(\hat{Z}) = \frac{1}{k} \sum_{i=1}^{k} \left(\log \hat{Z}_{(i)}^{(n)} - \log Z_{(k+1)}^{(n)}\right).$$

Both are consistent estimators of $\alpha^{-1}$ [Resnick and Stărică (1995)]. Experience and asymptotic variance calculations [Resnick and Stărică (1996b)] indicate that the second is surely a better procedure.

To compare these procedures we present three Hill plots in Figure 18. An autoregression of length 1000 was simulated with $\phi_1 = 1.3$, $\phi_2 = -0.7$, using iid Pareto random variables with $\alpha = 0.7$. The left-hand Hill plot is for the actual Pareto residuals used. The middle graph is a Hill plot for $\{X_t\}$ and the right-hand Hill plot is based on the estimated residuals. The middle plot seems to be considerably worse than the right plot based on estimated residuals. The dotted horizontal line in each case represents the true value of $\alpha$.

**5. A data example.** We consider a data set consisting of 3802 interarrival times of isdn D-channel packets. Figure 19 gives the time series plot and Figure 20 gives the acf and pacf plots in both classical and heavy tailed form. Independence does not seem likely based on these graphs.

We next checked the heavy tailed nature of the data. The Hill estimator was exceptionally stable and so were the qq-plots. Based on these diagnostics, a value of $\alpha = 1.06$ was estimated. The plots are given in Figure 21.

Based on the acf/pacf plots, it is unlikely that the data can be modeled by independence, but to confirm this we applied the independence test with $p = 6$. The independence hypothesis is rejected at the 5% level as shown in Figure 22.

If we wish to try modeling the data using a heavy tailed autoregression, we have to decide on the order. The pacf plot indicates $p = 6$ is likely and this is confirmed by the AIC plot given in Figure 23.

We then estimated autoregressive coefficients using the LP estimator and obtained coefficients

$$\begin{aligned}&\big(\hat{\phi}_1^{(n)}, \ldots, \hat{\phi}_6^{(n)}\big)\\ &\qquad = (0.00135, 0.00186, -0.00033, -0.00003, 0.00056, -0.00003).\end{aligned}$$

The estimated coefficients are rather small and it does not appear that the autoregressive structure changes the data very much. However, applying the independence test to the residuals as before with $p = 6$ (Figure 24) yields better results and the independence hypothesis cannot now be rejected.

The satisfaction one feels at apparently finding a suitable fit for the data is tempered by the acf plots in Figure 25 which splits the data into three successive parts of length 1000 each and computes an acf plot for each. The plots do not look very similar which could be an indication of lack of stationarity or, more disturbing, could be an indication of nonlinearity in the data. Tools need to be developed to cope with this possibility. This is discussed in greater detail in Feigin and Resnick (1996), Resnick (1996) and Davis and Resnick (1997). In any event, tests on estimated residuals should always be interpreted cautiously since the estimation method used usually tries to whiten residuals.

**6. Closing remarks.** At this point in the development of the subject, it is clear that there is an abundance of data which seems to need heavy tailed modeling but there is not an abundance of heavy tailed models with a rich set of accompanying data fitting techniques. It is difficult to get autoregressions to fit given data sets, and although this class offers attractive features for analysis, it is probably too small a class within the heavy tailed universe. There is a crying need to develop tools which will work with a larger class. The class of nonlinear models and the class of hidden Markov models are two possible classes worthy of investigation. See Meier-Hellstern, Wirth, Yan and Hoeflin (1991) for a closely related example.

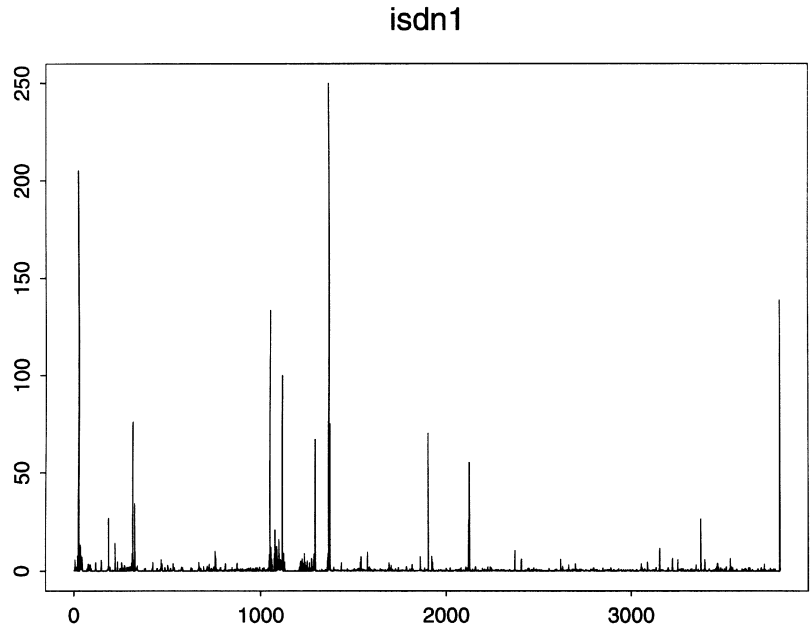


FIG. 19. *Time series plot of* isdn1.

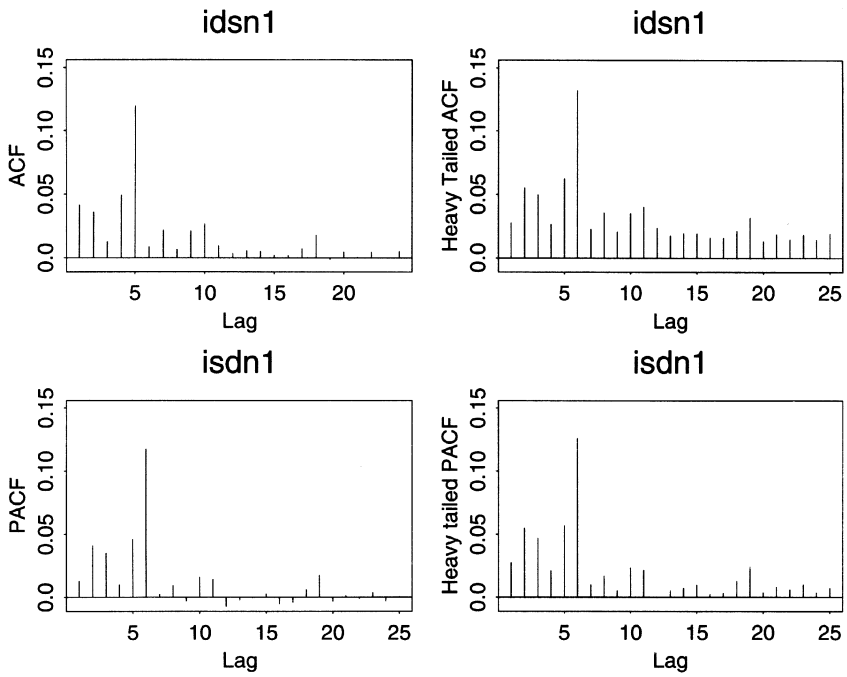


FIG. 20. Acf *and* pacf, *heavy and classical, for* isdn1.

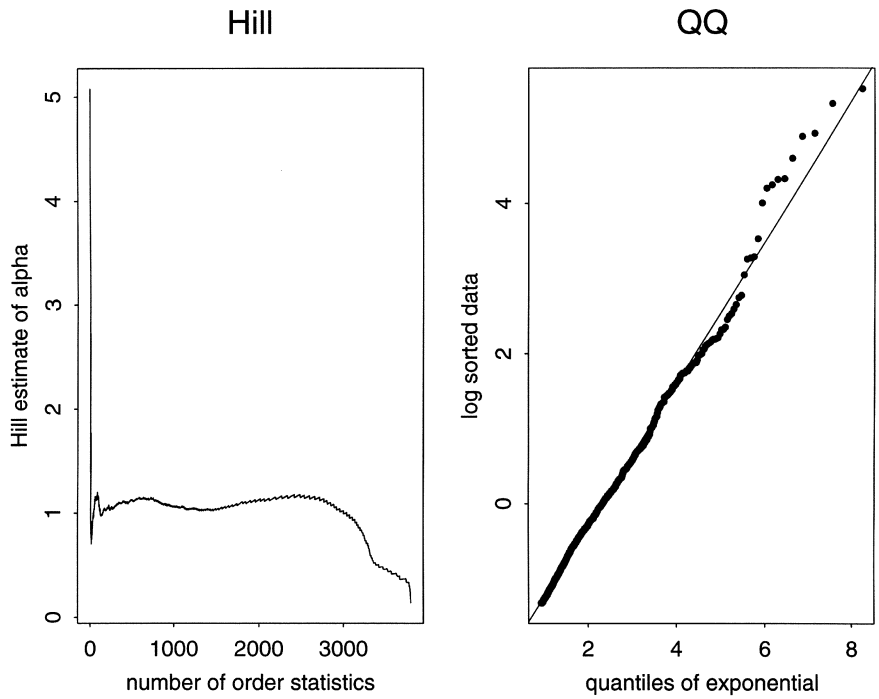


FIG. 21. *Hill and* qq*-plots for* isdn1.

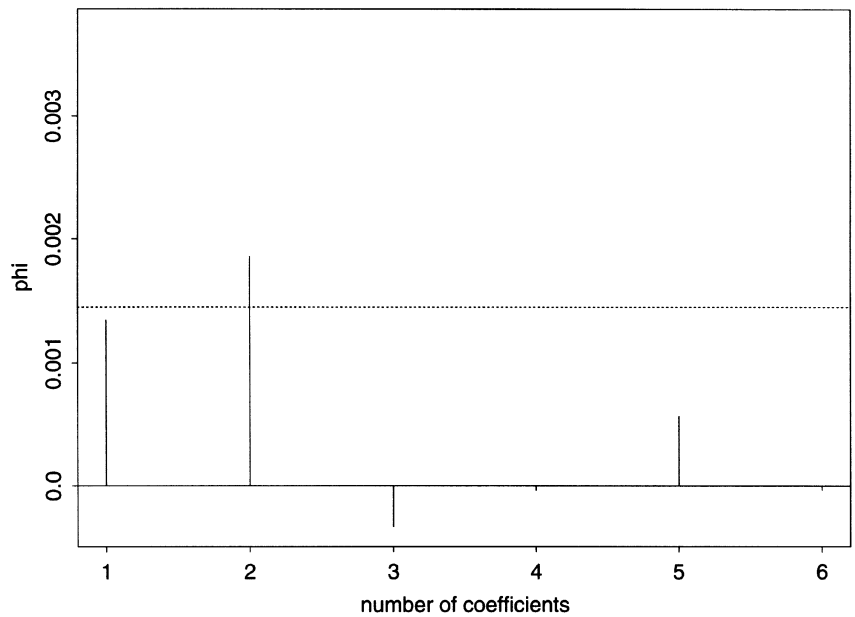


FIG. 22. *Independence test on* isdn1.

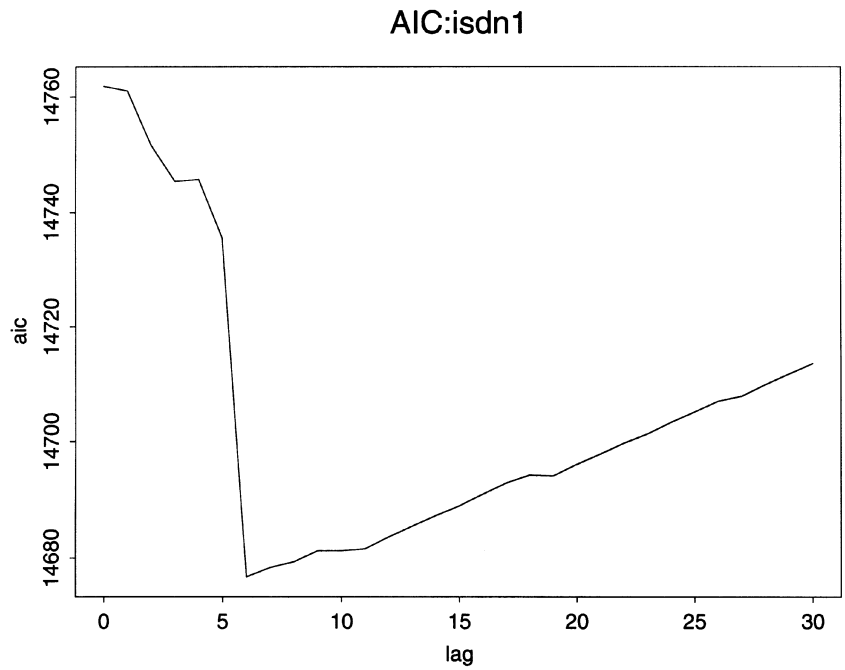


FIG. 23. AIC *plot for* isdn1.

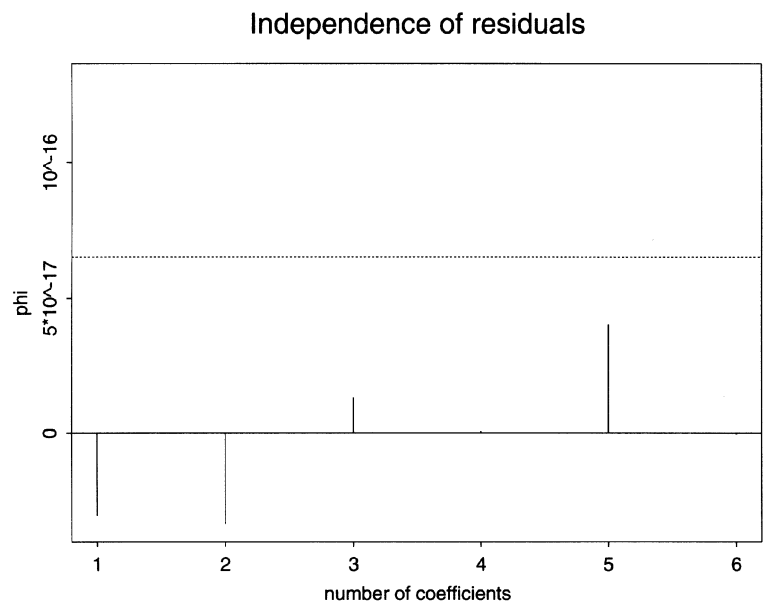


FIG. 24. *Independence test for residuals of* isdn1.

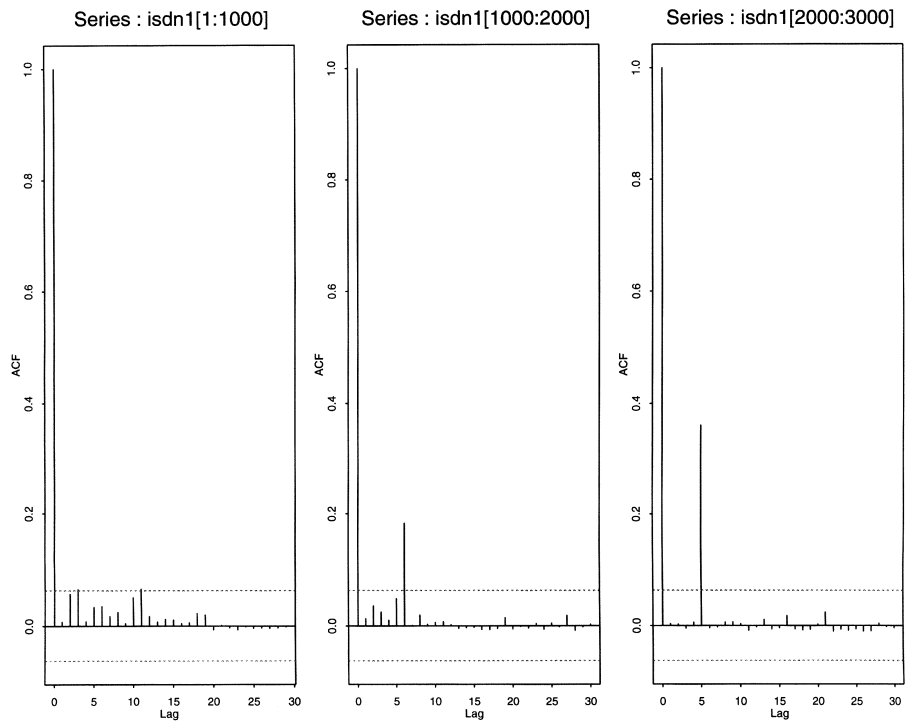


FIG. 25. Acf *of partitioned data.*

**Acknowledgments.** This paper is partly based on the Statistics Society of Canada Presidential address given in Montreal, Canada in July of 1995. Special thanks to President R. James Tomkins and the SSC for the invitation to speak and for the encouragement to organize my thoughts on the subject of heavy tailed modeling.

## REFERENCES

ANDEL, J. (1989). Nonnegative autoregressive processes. *J. Time Ser. Anal.* **10** 1–11.

ATHREYA, K. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.* **15** 724–731.

BERAN, J. (1994). *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

BERAN, J., SHERMAN, R., TAQQU, M. and WILLINGER, W. (1995). Long-range dependence in Variable-bit rate video traffic. *IEEE Trans. Comm.* **43** 1566–1579.

BHANSALI, R. (1988). Consistent order determination for processes with infinite variance. *J. Roy. Statist. Soc. Ser. B* **50** 46–60.

BINGHAM, N., GOLDIE, C. and TEUGELS, J. (1987). *Regular variation*. Cambridge Univ. Press, Cambridge.

BROCKWELL, P. and DAVIS, R. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York.

CASTILLO, E. (1988). *Extreme Value Theory in Engineering.* Academic Press, San Diego.

CLINE, D. (1983). Estimation and linear prediction for regression, autoregression and ARMA with infinite variance data. Ph.D. dissertation, Dept. Statistics, Colorado State Univ.

CSÖRGO, S. and MASON, D. (1985). Central limit theorems for sums of extreme values. *Math. Proc. Cambridge Philos. Soc.* **98** 547–558.

DATTA, S. and MCCORMICK, W. (1995). Bootstrap inference for a first order autoregression with positive innovations. *J. Amer. Statist. Soc.* **90** 1289–1301.

DAVIS, R., KNIGHT, K. and LIU, J. (1991). $M$-estimation for autoregressions with infinite variance. *Stochastic Process. Appl.* **40** 145–180.

DAVIS, R. and MCCORMICK, W. (1989). Estimation for first-order autoregressive processes with positive or bounded innovations. *Stochastic Process. Appl.* **31** 237–250.

DAVIS, R. and RESNICK, S. (1984). Tail estimates motivated by extreme value theory. *Ann. Statist.* **12** 1467–1487.

DAVIS, R. and RESNICK, S. (1985a). Limit theory for moving averages of random variables with regularly varying tail probabilities. *Ann. Probab.* **13** 179–195.

DAVIS, R. and RESNICK, S. (1985b). More limit theory for the sample correlation function of moving averages. *Stochastic Process. Appl.* **20** 257–279.

DAVIS, R. and RESNICK, S. (1986). Limit theory for the sample covariance and correlation functions of moving averages. *Ann. Statist.* **14** 533–558.

DAVIS, R. and RESNICK, S. (1996). Limit theory for bilinear processes with heavy tailed noise. Unpublished manuscript. *Ann. Appl. Probab.* **6** 1191–1210.

DE HAAN, L. (1970). On regular variation and its application to the weak convergence of sample extremes. *Math. Centre Tract* **32**. Math. Centre, Amsterdam.

DE HAAN, L. (1991). Extreme value statistics. Lecture notes, Econometric Institute, Erasmus Univ., Rotterdam.

DE HAAN, L. and PENG, L. (1995a). Rate of convergence for bivariate extremes (total variation metric). Report 9465/A, Econometric Institute, Erasmus Univ., Rotterdam.

DE HAAN, L. and PENG, L. (1995b). Rate of convergence for bivariate extremes (uniform metric). Report 9466/A, Econometric Institute, Erasmus Univ., Rotterdam.

DE HAAN, L. and PENG, L. (1995c). Exact rates of convergence to a symmetric stable law. Report 9467/A, Econometric Institute, Erasmus Univ., Rotterdam.

DE HAAN, L. and PENG, L. (1995d). Comparison of tail index estimators. Report 9464/A, Econometric Institute, Erasmus Univ., Rotterdam.

DE HAAN, L. and RESNICK, S. (1996). Second-order regular variation and rates of convergence in extreme-value theory. *Ann. Probab.* **24** 97–124.

DE HAAN, L. and STADTMÜLLER, U. (1996). Generalized regular variation of second order. *J. Austral. Math. Soc. Ser. A* **61** 381–395.

DEHEUVELS, P., MASON, D. and SHORACK, G. (1993). Some results on the influence of extremes on the bootstrap. *Ann. Inst. H. Poincaré* **29** 83–103.

DEKKERS, A. and DE HAAN, L. (1989). On the estimation of the extreme value index and large quantile estimation. *Ann. Statist.* **17** 1795–1832.

DEKKERS, A. and DE HAAN, L. (1991). Optimal choice of sample fraction in extreme value estimation. Ph.D. dissertation, Erasmus Univ., Rotterdam.

DEKKERS, A., EINMAHL, J. and DE HAAN, L. (1989). A moment estimator for the index of an extreme value distribution. *Ann. Statist.* **17** 1833–1855.

DUFFY, D., MCINTOSH, A., ROSENSTEIN, M. and WILLINGER, W. (1993). Analyzing telecommunications traffic data from working common channel signaling subnetworks. In *Proceedings of the 25th Interface, San Diego*. Interface Foundation of North Americ̀a.

DUFFY, D., MCINTOSH, A., ROSENSTEIN, M. and WILLINGER, W. (1994). Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. *IEEE Journal on Selected Areas in Communications* **12** 544–551.

FEIGIN, P., KRATZ, M. and RESNICK, S. (1996). Parameter estimation for moving averages with positive innovations. *Ann. Appl. Probab.* **6** 1157–1190.

FEIGIN, P. and RESNICK, S. (1992). Estimation for autoregressive processes with positive innovations. *Stochastic Model.* **8** 479–498.

FEIGIN, P. and RESNICK, S. (1994). Limit distributions for linear programming time series estimators. *Stochastic Process. Appl.* **51** 135–166.

FEIGIN, P. and RESNICK, S. (1996). Pitfalls of fitting autoregressive models for heavy-tailed time series. Unpublished manuscript.

FEIGIN, P. and RESNICK, S. (1997). Linear programming estimators and bootstrapping for heavy tailed phenomena. *Adv. Appl. Probab.* To appear. Available as TR1124.ps.Z at http://www.orie.cornell.edu/trlist/trlist.html.

FEIGIN, P. RESNICK, S. and STĂRICĂ, C. (1995). Testing for independence in heavy tailed and positive innovation time series. *Comm. Statist. Stochastic Models* **11** 587–612.

GARRETT, M. and WILLINGER, W. (1994). Analysis, modeling and generation of self-similar VBR video traffic. *Proceedings of ACM SigComm, London*.

GELUK, J. and DE HAAN, L. (1987). Regular variation, extensions and Tauberian theorems. *CWI Tract* **40**, CWI, Amsterdam.

GELUK, J., DE HAAN, L., RESNICK, S. and STĂRICĂ, C. (1997). Second order regular variation, convolution and the central limit theorem. *Stochastic Process. Appl.* To appear. Available as TR1133.ps.Z at http://www.orie.cornell.edu/trlist/trlist.html.

GINÉ, E. and ZINN, J. (1989). Necessary conditions for the bootstrap of the mean. *Ann. Statist.* **17** 684–691.

GRIGORIU, M. (1995). *Applied Non-Gaussian Processes*. Prentice Hall, Englewood Cliffs, NJ.

GUMBEL, E. (1958). *Statistics of Extremes*. Columbia Univ. Press, New York.

HALL, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B* **44** 37–42.

HALL, P. (1990). Asymptotic properties of the bootstrap for heavy-tailed distributions. *Ann. Probab.* **18** 1342–1360.

HEATH, D., RESNICK, S. and SAMORODNITSKY, G. (1997). Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.* To appear. Available as TR1144.ps.Z at http://www.orie.cornell.edu/trlist/trlist.html.

HILL, B. (1975). A simple approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174.

HSING, T. (1991). On tail estimation using dependent data. *Ann. Statist.* **19** 1547–1569.

JANSEN, D. and DE VRIES, C. (1991). On the frequency of large stock returns: putting booms and busts into perspective. *Review of Economics and Statistics* **73** 18–24.

KALLENBERG, O. (1983). *Random Measures*, 3rd ed. Akademie, Berlin.

KINATEDER, J. (1992). An invariance principle applicable to the bootstrap. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.). Wiley, New York.

KNIGHT, K. (1989a). Order selection for autoregressions. *Ann. Statist.* **17** 824–840.

KNIGHT, K. (1989b). On the bootstrap of the sample mean in the infinite variance case. *Ann. Statist.* **17** 1168–1175.

KOEDIJK, K., SCHAFGANS, M. and DE VRIES, C. (1990). The tail index of exchange rate returns. *Journal of International Economics* **29** 93–108.

KRATZ, M. and RESNICK, S. (1966). The qq estimator and heavy tails. *Stochastic Models* **12** 699–724.

LEADBETTER, M., LINDGREN, G. and ROOTZEN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

LEPAGE, R. (1992). Bootstrapping signs. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 215–224. Wiley, New York.

LIVNY, M., MELAMED, B. and TSIOLIS, A. K. (1993). The impact of autocorrelations on queueing systems. *Management Sciences* **39** 322–339.

MASON, D. (1982). Laws of large numbers for sums of extreme values. *Ann. Probab.* **10** 754–764.

MASON, D. (1988). A strong invariance theorem for the tail empirical process. *Ann. Inst. H. Poincaré* **24** 491–506.

MASON, D. and TUROVA, T. (1994). Weak convergence of the Hill estimator process. In *Extreme Value Theory and Applications* (J. Galambos, J. Lechner and E. Simiu, eds.) 419–432. Kluwer, Dordrecht.

MEIER-HELLSTERN, K., WIRTH, P., YAN, Y. and HOEFLIN, D. (1991). Traffic models for ISDN data users: office automation application. In *Teletraffic and Datatraffic in a Period of Change. Proceedings of the 13th ITC* (A. Jensen and V. B. Iversen, eds.) 167–192. North Holland, Amsterdam.

MIKOSCH, T., GADRICH, T., KLÜPPELBERG, C. and ADLER, R. (1995). Parameter estimation for ARMA models with infinite variance innovations. *Ann. Statist.* **23** 305–326.

NEVEU, J. (1976). Processus ponctuels. *Ecole d'Eté de Probabilités de Saint-Flour VI. Lecture Notes in Math.* **598**. Springer, Berlin.

RESNICK, S. (1986). Point processes, regular variation and weak convergence. *Adv. in Appl. Probab.* **18** 66–138.
RESNICK, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.
RESNICK, S. (1991). Point processes and Tauberian theory. *Math. Sci.* **16** 83–106.
RESNICK, S. (1992). *Adventures in Stochastic Processes.* Birkhäuser, Boston.
RESNICK, S. (1997). Why non-linearities can ruin the heavy tailed modeler's day. In *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions* (R. Adler, R. Feldman and M. Taqqu, eds.). Birkhäuser, Boston.
RESNICK, S. and SAMORODNITSKY, G. (1997). Performance decay in a single server exponential queuing model with long range dependence. *Oper. Res.* To appear.
RESNICK, S. and STĂRICĂ, C. (1995). Consistency of Hill's estimator for dependent data. *J. Appl. Probab.* **32** 139–167.
RESNICK, S. and STĂRICĂ, C. (1996a). Smoothing the moment estimator of the extreme value parameter. Preprint. Available by ftp from ftp.orie.cornell.edu as TR1158.ps.Z in directory /ftp/pub/techreps or at http://www.orie.cornell.edu/trlist/trlist.html.
RESNICK, S. and STĂRICĂ, C. (1996b). On the asymptotic behavior of Hill's estimator for dependent data. Unpublished manuscript.
RESNICK, S. and STĂRICĂ, C. (1997). Smoothing the Hill estimator. *J. Appl. Proabab.* **29** 271–293.
RICE, J. (1988). *Mathematical Statistics and Data Analysis*. Brooks/Cole, Pacific Grove, CA.
ROOTZEN, H., LEADBETTER, M. and DE HAAN, L. (1990). Tail and quantile estimation for strongly mixing stationary sequences. Technical Report 292, Center for Stochastic Processes, Dept. Statistics, Univ. North Carolina, Chapel Hill.
ROSIŃSKI, J. (1995). On the structure of stationary stable processes. *Ann. Probab.* **23** 1163–1187.
SAMORODNITSKY, G. and TAQQU, M. (1994). *Stable Non-Gaussian Random Processes*. Chapman and Hall, New York.
SMITH, R. (1982). Uniform rates of convergence in extreme value theory. *Adv. in Appl. Probab.* **l4** 543–565.
WILLINGER, W., TAQQU, M., LELAND, W. and WILSON, D. (1995). Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. *Statist. Sci.* **10** 67–85.
WILLINGER, W., TAQQU, M., SHERMAN, R. and WILSON, D. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking* **5** 71–96.

CORNELL UNIVERSITY
SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
ETC BUILDING
ITHACA, NEW YORK 14853
E-MAIL: sid@orie.cornell.edu

## DISCUSSION

ROBERT J. ADLER

*Technion, Israel Institute of Technology and*
*University of North Carolina, Chapel Hill*

I shall break my comments into three sections: one on the general philosophy of the paper, one of specific details and a final "public service announcement."

**General philosophy.** Resnick is to be commended on having written a paper that has not only forced him, in his own words, to organize his thoughts, but is going to force a number of other people to do the same.

Unfortunately, many of these thoughts seem to be very negative. In particular, the Hill estimator, about which so much has been written over the last decade, performs far less well in practice (to put it mildly) than it does in theory; linear, ARMA time series models are of questionable versatility; and, even if assumed, they seem to be difficult to identify and estimate. I will turn to some of these in a moment, but, along with other difficulties, they lead one to ask the following (only partly tongue-in-check) question:

Long before we had a theory for handling heavy tailed processes, we used to use transformations, so that $\{X_t\}$ was the data, $\{Z_t\} := \{\phi(X_t)\}$ was, *marginally*, Gaussian. "Marginal" is important here, since, although any stationary series can be made marginally Gaussian, a point transformation of this kind will not ensure full multivariate normality, which is what makes the standard techniques so powerful.

One of the main objections to this approach was, of course, that the class of processes that could be represented as point transformations of Gaussian processes is of somewhat limited generality. However, as Resnick points out, we really have very little idea how wide, for example, is the class of ARMA processes among all stable processes, so that at this point it seems that we have done little but to exchange one level of uncertainty for another.

Hence, one begins to wonder if the kind of effort directed into heavy tailed processes over the past decade or so should not be redirected back towards a more conventional approach of finding "better" transformations that will bring data into a framework in which we have tools that we know we can trust.

As an example of where one might look, transformations of the form

$$Z_t = \phi(X_t, \ldots, X_{t-k})$$

would be able to transform any heavy tailed (or other) time series into one whose $k$-dimensional distributions were Gaussian, and perhaps one would do better attempting to find optimal estimators of $\phi$ and then working with precision technology on such an approximate Gaussian process rather than in attacking $\{X_t\}$ directly. (The neural networks community have, in fact, wholeheartedly adopted this approach, although their language and motivation may at times seem unfamiliar, and theoretical justification of their techniques is generally lacking: See [7] for a very nice collection of papers in this vein and [8] for a different approach to nonlinear (and so non-Gaussian) time series modelling motivated by neural nets.)

**On specific details.** Many of the problems faced by the Hill and related estimators of the tail decay parameter $\alpha$ can be overcome if one is prepared to adopt a more parametric model, and assume, for example, stable innovations. While parametric models clearly have their own problems, these may not be as bas as one might initially expect. For example, Calder and Davis [3] describe a time series analysis of some models with Pareto innovations, using both

Pareto and stable techniques. Although the Pareto and stable distributions are quite different everywhere other than in their tails it turns out that the analyses (parameter estimates, etc.) were virtually identical, indicating that most of the estimation was somehow being done "in the tail."

One of the big advantages of assuming stable innovations is that in this case there is a very accurate estimator of $\alpha$ (and the other stable parameters) due to McCulloch [6]. (Interestingly, and in contrast to what Resnick reports for Hill estimators, the McCulloch estimator seems to be marginally more efficient when applied to the original data, rather than the fitted residuals. Details of some simulation studies to this effect appear in [1].)

Overall, it seems that the time may have come to relegate Hill-like estimators to the *Annals of Not-Terribly-Useful Ideas*.

Another issue that needs some understanding is the use of confidence windows for testing ACF's, whether they be "heavy tail modified" or not. Although heavy tails seem at first to be pleasant to work with, since the sample correlations approach their limit at a faster rate than in the Gaussian case, it is also unfortunately true that the limiting distributions are approached much slower than in the $L_2$ case. This is well documented in a number of situations (cf. [4] for references) although there does not seem to be much theory for the time series setting.

The following table, abstracted from Table 4 of [1], shows the results of 10,000 simulations of symmetric, stable, white noise with $\alpha = 1.8$ and differing sample sizes $n$. It records the percentage of times when the ACF at lag one lay outside the 95% confidence interval described in Resnick's paper. Note how very large the sample size $n$ has to be before one gets close to the nominal 5%.

Percentage outside 95% confidence interval for correlation

| $\alpha$ | $n = 10^3$ | $n = 10^4$ | $n = 10^6$ |
|---|---|---|---|
| 1.8 | 10.55 | 8.65 | 6.56 |

It is not clear how to overcome this problem. The obvious approach, of dropping asymptotic distributions and resorting to bootstrapping, is also beset with slow convergence problems (e.g., [5]).

In any case, the punch line here is that asymptotic distributions have to be taken with a heavy pinch of salt in the heavy tailed situation, and so the negative feeling one gets from Resnick's ACF/PACF analyses may not be totally justified. It is clear, however, that this is an area in need of some theoretical study.

**A public service announcement.** The reader who enjoyed Resnick's paper should also find the 23 papers in [2] illuminating. They cover a wide variety of problems associated with heavy tailed processes, in a number of different areas, from a practical viewpoint.

## REFERENCES

[1] ADLER, R. J., FELDMAN, R. and GALLAGHER, C. (1997). Analyzing stable time series. In *A User's Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*

*and Processes* (R. J. Adler, R. Feldman and M. Taqqu, eds.). Birkhäuser, Boston. To appear.

[2] ADLER, R. J., FELDMAN, R. and TAQQU, M. (eds.) (1997). *A User's Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions and Processes*. Birkhäuser, Boston.

[3] CALDER, M. and DAVIS, R. A. (1997). Inference for linear processes with stable noise. In *A User's Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions and Processes* (R. J. Adler, R. Feldman and M. Taqqu, eds.). Birkhäuser, Boston. To appear.

[4] CHRISTOPH, G. (1991). On some differences in limit theorems with a normal or a nonnormal stable limit law. *Math. Nachr.* **153** 247–256.

[5] LEPAGE, R., PÓDGÓRSKI, K. and RYZNAR, M. (1997). Bootstrapping signs and permutations for regression with heavy tailed errors: a seamless resampling. In *A User's Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions and Processes* (R. J. Adler, R. Feldman and M. Taqqu, eds.). Birkhäuser, Boston. To appear.

[6] MCCULLOCH, J. H. (1986). Simple consistent estimators of stable distribution parameters. *Comm. Statist. Simulation Comput.* **15** 1109–1136.

[7] WEIGEND, A. S. and GERSHENFELD, N. A. (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, Reading, MA.

[8] ZEEVI, A., MEIR, R. and ADLER, R. J. (19xx). Time series modeling using mixtures of experts. Preprint.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
AND
FACULTY OF INDUSTRIAL ENGINEERING
TECHNION, HAIFA, 32000
ISRAEL
E-MAIL: robert@adler.stat.unc.edu
robert@ieadler.technion.ac.il

## DISCUSSION

JAN BERAN

*University of Konstanz*

I would like to congratulate Sidney Resnick on the interesting review of heavy tail modeling. The paper demonstrates convincingly that there is an urgent need for developing statistical methods for heavy tailed stochastic processes, including the possibility of long-range dependence. One reason why heavy tail modeling may not be popular in practice is the lack of flexible yet sufficiently simple time series methods. Even the question of how to estimate the tail of a distribution optimally is not solved completely. Historically, the main method is Hill's estimator. However, as is shown in this paper, various modifications have to be made in order to make it sufficiently reliable. For dependent data the situation is even worse, since linear models appear to be less general than for processes with finite moments. It is, however, not easy to find sufficiently general models that could be, at the same time, useful

for data analysis. Statistical models are often purely theoretical constructions that approximate the stochastic behavior of a system, but have otherwise nothing to do with the real data generating process. (Exceptions are models that are based on specific facts from a subject science.) Simple statistical methods are therefore usually more successful in practice. In the following, a few possibilities, regarding heavy tail and long memory estimation are discussed briefly.

**1. (Nonstandard) robust estimation.** The problem of estimating the tail of a distribution resembles semiparametric estimation of the pole of the spectral density $f$ for long-memory processes [see, e.g., Geweke and Porter-Hudak (1993), Robinson (1996)]. The pole is assumed to be of the form $f(\lambda) \sim L(\lambda)|\lambda|^{1-2H}$. The long-memory parameter $H$ is estimated from the lowest $m$ periodogram ordinates with $m \to \infty$ but $m/n \to 0$. For the derivation of the limiting distribution of $\hat{H}$, assumptions on the slowly varying function $L$ are needed that are difficult to check in practice. The rate of convergence of $\hat{H}$ is $\sqrt{m}$ and thus slower than the parametric rate $\sqrt{n}$. Also, how to choose $m$ has been resolved only partially so far [see, e.g., Giraitis and Samarov (1996)]. The reason for these problems is that the aim is to have an omnibus method that is consistent although the unknown spectral density $f$ is essentially completely arbitrary, except for the behavior in an (unknown) neighborhood of the origin. For many applications, this aim may be unnecessarily ambitious and perhaps too pessimistic. Approximate consistency may often be sufficient for practical purposes, in particular since statistical models are only approximations anyway. Moreover, one might have a priori knowledge about the approximate qualitative shape of the spectrum. One may thus start with a reasonable but simple parametric model and estimate $H$ in a way that is robust against a sufficiently large class of deviations from the assumed spectral shape. Graf (1983) and Graf, Hampel and Tacier (1984) propose, for instance, to estimate $H$ by a huberized frequency-domain maximum likelihood method with fractional Gaussian noise as the central model. For details see, for example, Beran (1994), Chapter 7.3. The idea is to bound the score function by frequency dependent upper and lower limit curves $u(\lambda)$ and $v(\lambda)$. The functions $u$ and $v$ are chosen to be equal to infinity and minus infinity, respectively, in a small neighborhood of zero. Outside of this neighborhood, $u$ and $v$ are finite and monotonically decreasing and increasing, respectively, with increasing frequency. Boundedness for "high" frequencies makes sure that even under deviations from the central parametric model, $\hat{H}$ has a relatively small bias. For fractional Gaussian noise, the asymptotic bias is zero. Moreover, since all periodogram ordinates are used, a $\sqrt{n}$ rate of convergence is achieved. Thus, this method yields estimates that achieve a parametric rate of convergence while remaining almost consistent in a (hopefully) realistic neighborhood around the assumed model. Note that fractional Gaussian noise can be replaced by other parametric models.

This idea can be carried over to the estimation of the tail parameter $\alpha$ of the distribution of a positive random variable. Starting with a central parametric

model, an estimate of $\alpha$ can be defined by a huberized maximum likelihood method. Here, huberizing has to be strong for small values and no huberizing should be applied for extreme values. This is rather nonstandard, since usually robust methods aim at bounding the influence of extreme observations. As an example, take the Pareto distribution as the central parametric model. Under this ideal model, $\log X_i (i = 1, \ldots, n)$ are iid exponential. Thus, the MLE of $\theta = \alpha^{-1}$ can be written as the solution of

$$\sum_{i=1}^{n} \left( \frac{\log X_i}{\theta} - 1 \right) = 0. \tag{1}$$

To obtain a huberized estimate of $\theta$, define for $p \in [0, 1]$ functions $u(p) \geq 0$ and $v(p) \leq 0$ that are bounded for $p \leq p_o < 1$ where $p_o$ is a fixed constant, and such that $u(p)$ and $-v(p)$ are monotonically increasing in $p$. For $p > p_o$, one may choose $u(p) = -v(p) = \infty$. Define now

$$g(x, p; \theta) = \left[ \frac{\log x}{\theta} - 1 \right]_{v(p)}^{u(p)}, \tag{2}$$

where $[y]_v^u = \min\{\max(y, v), u\}$ and $\mu(p) = E[g(Z, p; 1)]$ where $\log Z$ is standard exponential. Finally, let $p_i = F_n(X_i)$ be the empirical distribution function at observations $X_i$ and

$$\psi(X_i, p_i; \theta) = g(X_i, p_i; \theta) - \mu(p_i). \tag{3}$$

Then $\hat{\theta}$ is defined by

$$\sum_{i=1}^{n} \psi(X_i, p_i; \hat{\theta}) = 0. \tag{4}$$

How (4) works in practice and which huberizing functions $u$ and $v$ exactly are suitable would need to be investigated in detail. Versions of (4) based on the characteristic function may also be of interest. For time series, (4) can be applied to estimate $\alpha$ of the residual process.

**2. Hierarchical models and model choice.** An alternative to robust estimation is to use a hierarchichal class of parametric models with an arbitrary number of parameters and choose the number of parameters by a suitable model choice criterion. For example, a well-known method for estimating the spectrum of a short-memory process is "autoregressive spectral estimation" (see, e.g., the discussion and references in Priestley, chapter 7.8). The true process is assumed to have an AR($\infty$)-representation. Finite order AR($p$)-spectra are used to approximate the true spectral density. The order of the AR($p$) process is chosen by criteria such as the AIC, the CAT and so on. How to adapt this method to long-memory models (using, for instance, fractional AR-models) is an open problem. Ideally, with suitable model selection, it should be possible to obtain a consistent estimate of $H$. Note that, in contrast

to semiparametric or robust estimation, this would allow also for modelling the whole spectral density and not just the long-memory parameter. Also note that models that are defined directly via the spectral density may be particularily useful in this context. For instance, for fractional exponential (FEXP) models, generalized linear regression can be applied [Beran (1993)].

Similarily, the tails of a distribution may be estimated by approximating the whole distribution or its characteristic function by parametric densities or characteristic functions, respectively, with the number of parameters estimated by a model choice criterion. In contrast to semiparametric or robust estimation, this would allow for modelling the whole distribution function and not just the tails.

**3. Heteroscedastic models.** Simulated series of nonlinear models with random heteroscedasticity, such as GARCH models [Bollerslev (1986), Engle (1982)], often resemble sample paths of stable processes, at least visually. In particular, the occurence of occasional (single or patches of) "outliers" is typical for GARCH models. These models are very popular in economics because of their intuitive appeal and relative simplicity. Long-memory GARCH models are also known meanwhile [Baillie, Bollerslev and Mikkelsen (1996), Ling and Li (1996); also se Ding and Granger (1996)]. Thus, GARCH models are serious competitors to heavy tailed processes. Methods will need to be developed to distinguish convincingly between GARCH-type models and time series processes with infinite moments. Also, GARCH models with stable innovation distributions may be useful.

## REFERENCES


BAILIE, R. T., BOLLERSLEV, T. and MIKKELSEN, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroscedasticity. *J. Econometrics*. To appear.

BERAN, J. (1993). Fitting long-memory models by generalized linear regression. *Biometrika* **80** 817–822.

BERAN, J. (1994). *Statistics for Long-memory Processes.* Chapman and Hall, New York.

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroscedastiscity. *J. Econometrics* **31** 307–327.

DING, Z. and GRANGER, C. W. J. (1996). Modeling volatility persistence of speculative returns: a new approach. *J. Econometrics* **63** 185–215.

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50** 987–1008.

GEWEKE, J. and PORTER-HUDAK, S. (1983). The estimation and application of long-memory time series models. *J. Time Ser. Anal.* **4** 221–238.

GIRAITIS, L., ROBINSON, P. M. and SAMAROV, A. (1996). Rate optimal semiparametric estimation of the memory parameter of the Gaussian time series with long-range dependence. Preprint.

GRAF, H. P. (1983). Long-range correlations and estimation of the self-similarity parameter. Ph.D. dissertation, ETH Zürich.

GRAF, H. P., HAMPEL, F. R. and TACIER, J. (1984). The problem of unsuspected serial correlations. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 127–145. Springer, New York.

LING, S.-Q. and LI, W. K. (1996). Fractional ARIMA-GARCH time series models. Preprint.
ROBINSON, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *Ann. Statist.* **23** 1630–1661.

UNIVERSITÄT KONSTANZ
FAKULTÄT F. WIRTSCHAFTSWISSENSCHAFTEN
UND STATISTIK
KONSTANZ 78434
P.O. BOX 5560
GERMANY
E-MAIL: jberan@iris.rz.uni-konstanz.de

## DISCUSSION

WALTER WILLINGER AND VERN PAXSON

*AT&T Labs–Research and Lawrence Berkeley National Laboratory*

As network researchers who have spent many long days in the past few years poring over gigabytes of network-related measurements to try to get some basic understanding of network traffic, we are very pleased to see this article and to be able to comment on it. Historically, the field of networking and communications research has suffered from a severe and constant shortage of traffic measurements [11]; however, during the past five to ten years, an abundance of enormous-sized data sets of high-quality network traffic measurements have become available and keep on being collected in ever-increasing quantities and from ever-faster networks. Unfortunately, the development of statistical techniques and methods for efficiently dealing with this "flood" and for effectively handling phenomena such as heavy tails and long-range dependence, hitherto unknown in the networking arena, has not kept up with the rate at which newer and larger data sets of traffic measurements are being captured, nor with the pace at which the networks themselves change. Professor Resnick's paper is therefore extremely timely and highly welcome by traffic analysts who are stymied by the many unfamiliar features that appear in networking-related data sets, and typically lack even the most basic tools to sensibly deal with them.

At the same time, Professor Resnick's paper also serves as useful reminder that collected data often determine the utility and relevance of one data analysis or modeling approach over another. To illustrate this, we first point out to the interested statisticians and data analysts some of the features that make data sets of traffic measurements from today's networks unique and challenging, especially when compared to available data sets from other areas of science and engineering (e.g., hydrology, medicine, biophysics, economics, finance). Given the special nature of the data at hand, we then argue—in contrast to Resnick—for abandoning the *black box modeling approach* from traditional time series analysis and focusing instead on *structural models* that

take into account the context in which the data arose in the first place, namely the highly intertwined hierarchies of networking functions that form the basis of modern computer communication. While we readily admit that black box models can be and are useful in other contexts, we strongly believe that they are essentially of no use for our main purposes of trying to understand the dynamic and complex nature of traffic in today's packet networks and, subsequently, of exploiting this understanding to design, manage and control these networks.

**1. Not your usual data sets.** Although Resnick points out that today's communication networks have recently become highly prolific providers of large amounts of traffic data, he does not dwell further into what kinds of data have actually been collected. We provide a more detailed account of the data sets recently measured on current communication networks, such as Common Channel Signaling Networks (CCSNs) used by the telephone system, Ethernet Local Area Networks (LANs) ubiquitous in shared computing environments, and IP links comprising the explosively growing global Internet. These are not the usual data sets that statisticians have been dealing with in the past. Our goal is to point out their unique features and properties, since these drive the entire subsequent discussion.

Thanks to sophisticated measurement devices, themselves often full-blown computers, the data sets collected from these networks are not only unique with respect to size, but also in terms of the quality and amount of information recorded for every observation. Depending on the network under consideration, an observation can be a *telephone call*, a *message* or a *packet*. Today's properly designed and well-tested measurement devices record exactly the bits and bytes that are transported over the network, resulting in complete, accurate and error-free data. The only uncertainties come from time stamp resolution and the possibility of "dropping" a measurement because the recording device is too slow. Both of these can be controlled by the diligent researcher. Furthermore, once such a device has been built and tested, there are no limitations on the length of the measurement periods, that is, on the amount of data that can be collected from a "live" network (in practice, though, available disc space often *does* impose a limit on the length of the period over which traffic can be recorded continuously).

In this context, some numbers might be telling: half a day of monitoring a single 56 kilobit/sec channel of an A-link to an end office in a CCS subnetwork resulted in about 500,000 calls (January 1993); one hour of measuring traffic on a 10 megabit/sec Ethernet LAN at Bellcore (August 1989) yielded about 1,500,000 packets; a 37-minute trace of traffic at the busy Internet exchange point FIX-WEST consisted of more than 35,000,000 packets (June 1995); and a week's worth of Internet traffic into and out of the University of California at Berkeley consisted of 439,000,000 packets and 89 gigabytes of data (January 1995).

In terms of the information recorded, for a CCS network a typical calling record consists of call arrival time, call holding time, caller and callee number,

and possibly further data associated with the caller profile. For an Ethernet or Internet setting, the recorded information on a per-packet basis consists of a time stamp and the full packet "header" information, including source and destination address of the packet, size, protocol, protocol-specific information such as sequencing position, and possibly the entire packet data contents, giving all of the application details.

These high-quality and high-volume measurements result in data sets that can easily extend into the giga- or even terabyte range. On the one hand, such voluminous data sets pose obvious challenges for researchers interested in exploratory data analysis or "data mining," where the latter phrase is used to denote "sensible digging into data to try to reveal what they are saying" rather than "torturing the data till they confess" (e.g., see the discussion in [2]). On the other hand, they offer data analysts and modelers unique opportunities, mainly because of the richness of available information at all levels of interest (e.g., at the level of aggregate network traffic, at the level of individual sources and destinations, at the transport protocol level, at the application level). Examples of some data sets from existing networks can be found in the *Internet Traffic Archive* at `http://www.acm.org/sigcomm/ITA/` .

**2. Not your usual modeling world.** Given the recent abundance of traffic data, traffic modeling is faced with a number of pressing new problems. Solving them will require a close collaboration between data analysts, applied mathematicians and networking experts. To avoid any misunderstanding: for us, the main objective of traffic analysis and modeling is to try to gain a good understanding of the actual dynamics of network traffic and to make use of this know-how when designing, managing and controlling existing or future networks.

For the present discussion, it is also crucial to keep in mind that modern communication networks are highly dynamic entities that undergo constant changes (e.g., network topology, user population, services and applications, network technologies, protocols). Just consider today's Internet and compare it with the network that existed four years ago (when it had practically no WWW traffic), or two years ago (just before the decommissioning of the NSFNET), or just a month or week ago (when it was still possible to connect to your favorite sites). Such change is nothing new, either. For example, the USENET Internet "news" system has exhibited striking, sustained exponential growth of 75%/year since 1984 [12]. Even at fixed points in time, connection characteristics vary significantly from site to site [13]. Furthermore, not only do networks quickly change in many ways, so too can patterns of use of well-established applications. For example, in October 1992 the median size of an FTP file transfer measured at a large research institute was 4,500 bytes. This statistic is presumably highly robust, as it was drawn from a sample of more than 60,000 transfers. Yet only five months later, the median fell by more than a factor of two, to 2,100 bytes, in a sample of more than 80,000 transfers [13].

Clearly, life for traffic data analysts and modelers is extremely challenging and difficult. What does it mean in this setting to pick the "best-fitting" model

for a given traffic trace, when the underlying data set keeps changing over time, across the same network, across different networks, and so on? The task is comparable to chasing a moving target—blindfolded! That is, the traffic data analyst is asked to detect some unknown characteristic features in enormous-sized empirical records, where the records are constantly augmented by new traffic measurements from new (and generally faster) networks that carry new (and typically more bandwidth-intensive) services and applications, and are being clogged up by an exponentially growing (and increasingly diverse) user population.

Successful traffic modeling for modern communications networks also has to face the popular belief that since actual network traffic is commonly considered to be highly complex in nature, only complicated and highly parameterized models are likely to result in accurate approximations of reality. However, in network engineering practice, which is the ultimate application of traffic modeling, such models are viewed as essentially useless, because there is no hope of forming solid estimates for numerous parameters in such a changing world. Instead, a traffic model is considered useful only if it (1) is simple and accurate, (2) has a physical explanation in the network context and hence provides new insights into the dynamics of network traffic, (3) can be inferred from operational measurements, and (4) has measurable and practical impact on system performance. Unfortunately, these practical criteria for traffic modeling have all but been ignored in the past, where traffic modeling was traditionally done without access to any data and where the resulting models were primarily judged by how well they could be analyzed mathematically and hardly ever by how relevant they are in engineering practice. In contrast, the ever-increasing size of the latest data sets of traffic measurements from today's networks, the dynamic nature of these networks and the complexity of the traffic they carry, all argue strongly in favor of modeling network traffic based on the *principle of parsimony*, also known as *Ockham's Razor* (e.g., see [6]).

The idea behind parsimonious modeling is to explain facts in as economical a way as possible. As a result, the quest for parsimonious traffic models forces the data analysts to concentrate on features that are common among the many large data sets and hence robust under changing networking conditions (the quest for "traffic invariants"); it offers the traffic modelers the opportunity to build traffic models around these invariants and to come up with plausible physical explanations for the empirically observed phenomena; and it provides network researchers with an improved understanding of the complex nature of network traffic, giving them tools to explore network performance without being hit by the "curse of dimensionality" (i.e., a large number of parameters that have little physical meaning). Naturally, this quest for parsimony will be successful the more that data analysts, traffic modelers and network engineers interact with one another.

2.1. *Searching for traffic invariants: heavy tails.* Despite its great diversity, measured network traffic has revealed a number of surprises (and candi-

date invariants) that make it specially interesting to data analysts and modelers. One of these surprises, which is to a large degree the motivation behind Professor Resnick's article, is the prevalence of the *infinite variance phenomenon* (also known as the *Noah Effect*), or more generally, of heavy tailed distributions in networking-related activities (e.g., CPU processor time, file size, video frame size, call holding times, interpacket times for interactive applications, burst sizes in data bulk transfer, WWW item sizes). On the one hand, this observation is good news for network researchers who have seen traditional (i.e., telephone network-based) traffic invariants either vanish or change into highly variable and unpredictable quantities in the context of other data networks (such as the Internet), where none of the following are any longer constant: network size, topology, transfer rates, congestion levels, dominant applications, traffic mixes across sites and over time, and connection characteristics. Traffic analysts and modelers are *desperate* for traffic invariants because they provide the basis for understanding the dynamic nature of network traffic and are crucial in narrowing down the search space for models both consistent with measurements and useful in engineering practice. To this end, heavy tails are a godsend: something not changing in a sea of change.

On the other hand, the empirically observed omnipresence of heavy tails has caused serious concerns among network researchers, not only because of their inexperience with heavy tailed phenomena, but more importantly because of an almost complete lack of readily available and practically useful tools and techniques for dealing with them. Here is where Professor Resnick's paper (especially Section 3) not only fills a real need, but does so admirably well and in a surprisingly(?) frank manner. The paper not only surveys the growing body of techniques for inferring and modeling heavy tailed phenomena, but equally important, it emphasizes the limitations and illustrates the pitfalls associated with a blind belief in these statistical techniques. Such a combined good news/bad news survey is especially (but not exclusively!) valuable for the field of engineering sciences, where one often encounters a strong belief (blame the statisticians?) in the notions of universally applicable statistical techniques and absolute numbers. Professor Resnick's discussions detailing the conditions under which the different Hill estimator techniques "work" in theory and exploring the "gray area" where they should work in theory but don't work in practice, are as valuable for practitioners as are the mathematical results (and his comment that the underlying assumptions are essentially unverifiable in practice) showing that the Hill estimator has nice theoretical properties. As practitioners, we cannot overemphasize Resnick's implicit warnings that (1) no statistical technique works in all cases, (2) you have to know what you are doing and (3) relying on a portfolio of different techniques is always preferable (often much more work, though) to sticking with a single method. We wish that more importance would be given in the statistical literature and in statistical teaching to exploring scenarios under which a given technique works well or doesn't, the pitfalls in using the tech-

nique, and possible alternative methods for dealing with the same problem (and their advantages and shortcomings).

Finally, our only criticism concerning Section 3 is that none of the presented methods really exploits or makes use of the availability of the enormous-sized data sets mentioned earlier. This obscures the different statistical regimes in which network research operates. For example, Figure 1 shows on the left a log-log complementary distribution plot for a data set of WWW transfer sizes (226,386 observations), and on the right the upper 14% tail consisting of those transfers exceeding 10 kilobytes (32,630 points). The line fitted to these argues for a Pareto tail with $\alpha \approx 1.35$. Surely, there must be ways of taking advantage of the fact that for measured network traffic data, the upper heavy tails themselves contain more data points than the entire data sets considered in this paper! (Note: we have 900 additional WWW data sets like this one available for analysis.)

2.2. *Black box models vs. structural models.* Besides heavy tails, another invariant feature that has been consistently observed in recent traffic data sets is *long-range dependence* (LRD), also known as the *Joseph Effect*. The mere mention of LRD usually causes strong (predominantly negative) reactions from statisticians as well as mathematical modelers and is typically accompanied by endless philosophical discussions about the (im)possibility of inferring "true" LRD from a finite data set (e.g., see [7, 5, 10]). In sharp contrast, in the present context of network traffic analysis, there exist very pragmatic reasons for moving beyond philosophical discussions towards the challenging problem of providing phenomenological explanations. For real-world analysis, the philosophical points are moot: the traditional theoretical framework based on Poisson assumptions of fleeting correlations lies in shambles [9, 14], and

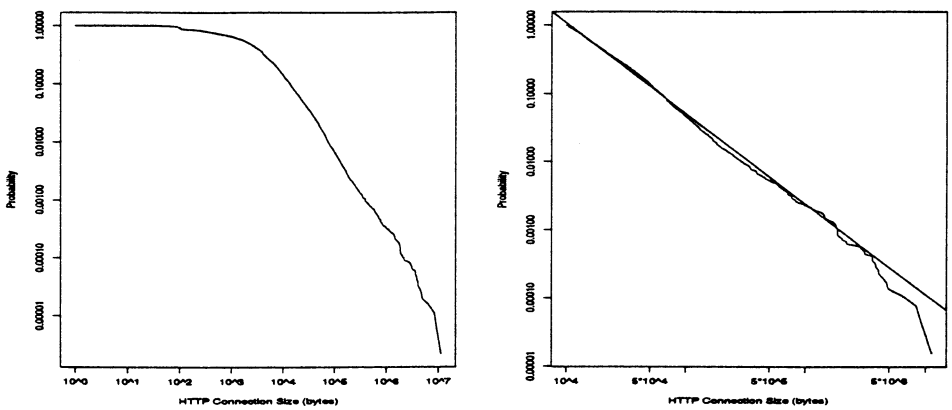


FIG. 1. *Log-log complementary distribution plot of* WWW *transfer sizes*: *full dataset* (*left*), *upper tail* (*right*).

networking practitioners find themselves giving serious consideration to LRD models as *good approximations* if nothing more.

Searching for plausible physical explanations for empirically observed phenomena such as LRD is intimately related to abandoning the black box models from traditional time series analysis in favor of *structural models*. The unique feature of structural traffic models is that they are capable of explicitly accounting for the hierarchical nature of today's network architectures and hence can capture the intertwined mechanisms and modes (at the different layers in the networking architecture) that determine the flow of packets across a network. For example, above the physical layer (i.e., the raw media such as copper or fiber over which data is sent) is the network interface or "link" layer with its mechanisms and protocols controlling how packets are sent over the media. Above it is the internetworking layer responsible for getting a packet through a series of different networks (e.g., the Internet Protocol, IP). At a yet higher level, the "transport" layer provides the functions needed to exchange data between two applications (e.g., the Transport Control Protocol, TCP), such as assuring that packets are reliably delivered and/or arrive at the application in the order they were sent. And above it, the application layer incorporates services designed to support a range of applications (e.g., Telnet, FTP, WWW).

Accurate accounting of this multilevel hierarchy in measured network traffic is possible because all the relevant information can be obtained unambiguously from looking "inside" the collected packets—checking the header of each recorded packet usually suffices. In contrast, (univariate) time series models typically treat these packets as black boxes. That is, by focusing on the mere existence of these packets (i.e., the corresponding time stamps, or the resulting time series of counts of packets) and not on their "meaning" as revealed by their headers, they ignore most of the gathered information. Even when replacing univariate by multivariate time series models, where the latter can account for covariates associated with the individual packets, it is not clear how the hierarchical and interconnected structure of network traffic can be captured. On the other hand, successful structural modeling makes use of potentially all of the recorded information at all layers of the network architecture. It results in traffic models that are consistent with that data at all levels of interest (e.g., internetworking layer, transport layer, application layer). And it offers practical answers to the following:

1. The data analysts' dilemma of making sense of the ever-increasing size of available network traffic measurements.
2. The network researchers' quest for traffic invariants.
3. The traffic modelers' emphasis on model parsimony.
4. The network engineers' insistence on models that are simple, physically meaningful, and relevant to engineering practice.

It does so by "carefully mining the available data" in the sense of Cox (see [2]), that is, determining to what extent important conclusions from the

data lie on the surface; by "broadening the basis" in the sense of Tukey [15], page 277 (see also [4]); that is, a traffic invariant is verified by checking against a wide range of similar data sets; by applying the principle of parsimony and abstracting out features of the data that do not significantly contribute to our understanding of network traffic; by fully exploiting subject matter considerations (achieved by close collaborations between network researchers and data analysts) and by iterating these procedures, if deemed necessary.

The appeal for structural models in the network setting stems from the observation that network hosts (e.g., computers, routers) behave in a highly predictable fashion in a given mode (e.g., request/response vs. bulk transfer vs. interactive vs. multimedia applications, file servers vs. clients). What can change drastically is the behavior in the different modes. Yet *all you have to do to recognize the different modes is look inside the packet header*—structural modeling at its simplest. The data set "Packet" considered in the paper is an example where all information concerning the different modes that generated this traffic trace has been ignored and where only the time series of arrival times of the successive packets is considered. As such, the data set is not terribly interesting for actual network engineering. On the other hand, with structural modeling that makes use of all available information on a per-packet basis, ISDN packet arrivals are known in advance to have a nontrivial dependence structure, due to the nature of the dominant applications (e.g., word-processing) and the transport protocol used. In particular, the jump at lag 6 in the autocorrelation function is almost certainly a result of the transport protocol used or of fragmentation boundaries (if we were given the packet headers we would immediately know which), and is of only minor interest for network engineering ("statistically significant but practically irrelevant"). For examples of successful structural modeling approaches in the context of network traffic, we refer to the recent papers by Willinger, Paxson and Taqqu [16] and Willinger, Taqqu, Sherman and Wilson [18, 17] Kurtz [8], and Crovella and Bestavros [3]. These papers also elaborate on the connection between the empirically observed phenomena of heavy tails and LRD at different levels of the network structure and show, as alluded to by Professor Resnick, how structural modeling is able to provide a unified framework within which both phenomena can be explained.

Clearly, our strong preference for structural traffic models over black box models clashes with Professor Resnick's decision in Section 5 to stick with the more traditional black box modeling approach from time series analysis. At the same time, his arguments in favor of pursuing black box models (tradition; requires no subject matter expertise; applicable across a broad range of disciplines) are not all that convincing, especially when recalling the title of the article, which explicitly mentions the application area of interest. In fact, as network researchers, we are more than willing to apply our networking know-how when it comes to constructing sensible structural models; moreover, we are much less concerned about the models' potential applicability in other areas of science, but worry immensely about the models' usefulness and practical relevance in the network context. This leaves Resnick with tradition

as the remaining argument for black box models, and we would love to see him argue for tradition in front of a group of Internet researchers.

**3. Not your usual topics for future work.** Judging from the recent past, it is a safe to predict that network researchers will continue to be faced with practically unlimited amounts of traffic measurements from the latest packet networks. Our experience has been that traditional traffic modeling (including conventional time series analysis) has little to offer for effectively and sensibly dealing with this situation, and it has not been for want of trying. Instead, we have found that the rather vaguely defined concepts of (1) careful data mining, (2) broadening the basis, (3) insistence on model parsimony, and (4) reliance on subject matter expertise do a much better job of addressing the task of practical network traffic analysis and modeling. In the process, one garners both feasible alternatives for dealing with the available data and some basic understanding of the nature of actual network traffic (and, as a result, useful traffic models).

All four concepts relate in one way or another to the basic problem of dealing effectively with large, diverse data sets, and in all four cases, the main research challenges center around the quest for techniques that scale (in the number of available observations). Beran and Terrin's [1] recent work is an example that shows how a technique (i.e., Whittle's method for estimating the degree of LRD) that is computationally infeasible in the presence of a large number of observations can be turned into a method that scales (and hence can be applied even to large data sets) by proving new central limit-type results and by exploiting modern high-performance computing and communication capabilities. Similarly, [18] presents an algorithm for quickly generating fractional Gaussian noise that also relies on a combination of new convergence results for stochastic processes and a highly parallel computing architecture. We strongly believe that such combinations of new probabilistic results and advanced parallel or distributed computing capabilities will be a driving force behind future progress in areas associated with the four mentioned concepts.

Another development in the area of network traffic analysis and modeling that results from the ease with which large amounts of network traffic measurements can be obtained these days is a steady shift away from *statistical inference*, with its traditional emphasis on a single data set and on testing models, and toward *scientific inference*. Scientific inference (as discussed for example in [2]) typically involves many data sets (e.g., traffic collected on the same network over different periods in time and at different points in the network) and attempts to identify features in the data that generalize to different conditions (e.g., different network loads, different mix of users or applications). As a result, in modern network research, reproducibility (also referred to as replication or repetition) of traffic measurements studies by other researchers, in other, different networks and at different times is becoming an essential part of model specification, model selection and model verification. In this sense, network research is starting to adopt a concept that has a long tradition in the physical sciences but has been all but ignored in the social sciences

and in the mainstream statistics literature. Put succinctly: how do we deal with huge data sets? How do we deal with huge numbers of huge data sets? How can we do reliable estimation in the presence of strong correlations and infinite variance? How can we more formally assess heuristic arguments of similarity in inference across different data sets?

Although it is worthwhile to always heed Hampel's [5] advice that "no model is ever correct—all are but better or worse approximations of reality" (the same idea is often expressed in the well-known saying "All models are wrong, but some are useful"), our decision as network researchers is obvious when given a choice between a black box model that fits a single data set "perfectly," and a model that "works" under a wide range of networking conditions and makes sense on physical grounds. However, we emphasize that our criticism of the black box modeling approach and the underlying "Box-Jenkins machinery" is subject-specific (but so is the paper under discussion) and should not be interpreted as an argument in favor of writing off time series analysis altogether. At the same time, it is very disconcerning to see that despite the drastic changes in terms of available network traffic data, much work in this area is still done "the same old way," where data was a scarce resource and where "squeezing the available data set dry" (see [2]) made sense. One cannot help but be highly critical of such work when there is now plenty of data to go around and when squeezing a single data set dry would become a researcher's lifetime job. Indeed, things change so fast on the networking scene that spending too much time studying a particular dataset brings with it a genuine risk of rendering the ultimate findings obsolete. Surely, there must be other ways for dealing with the currently available data! In this sense, the ideas put forward by Tukey in [15] that "data analysis, like calculations, can profit from repeated starts and fresh approaches" and that "there is not just one analysis for a substantial problem" are highly relevant to network traffic analysis and modeling, especially now that the data at hand gives new meaning to statistical inference and makes structural modeling a viable alternative to traditional time series analysis.

## REFERENCES

[1] BERAN, J. and TERRIN, N. (1994). Estimation of the long memory parameter, based on a multivariate central limit theorem. *J. Time Ser. Anal.* **15** 269–278.

[2] CHATFIELD, C. (1995). Model uncertainty, data mining and statistical inference. *J. Roy. Statist. Soc. Ser. A* **158** 419–466.

[3] CROVELLA, M. and BESTAVROS, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In *Proc. ACM SIGMETRICS '96, Philadelphia*, 160–169. ACM Press.

[4] DRAPER, D., HODGES, J. S., MALLOWS, C. L. and PREGIBON, D. (1993). Exchangeability and data analysis. *J. Roy. Statist. Soc. Ser. A* **156** 9–37.

[5] HAMPEL, F. R. (1987). Data analysis and self-similar processes. In *Proceedings of the 46th Session of the International Statistical Institute, Tokyo, Japan, September 1987*. International Statistical Institute.

[6] JEFFERYS, W. H. and BERGER, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist* **80** 64–72.

[7] KLEMES, V. (1974). The Hurst phenomenon: a puzzle? *Water Resources Research* **10** 675–688.

[8] KURTZ, T. G. (1996). Limit theorems for workload input models. In *Stochastic Networks; Theory and Applications* (F. P. Kelly, S. Zachary and I. Ziedins, eds.). Oxford Univ. Press.
[9] LELAND, W., TAQQU, M., WILLINGER, W. and WILSON, D. (1994). On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* **2** 1–15.
[10] LIEBOVITCH, L. S. (1989). Testing fractal and Markov models of ion channel kinetics. *Biophysics Journal* **55** 373–377.
[11] PAWLITA, P. (1989). Two decades of data traffic measurements: a survey of published results, experiences and applicability. In *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12* (M. Bonatti, ed.) 230–238. North-Holland, Amsterdam.
[12] PAXSON, V. (1994). Growth trends in wide-area TCP connections. *IEEE Network* **8** 8–17.
[13] PAXSON, V. (1994). Empirically-derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking* **2** 316–336.
[14] PAXSON, V. and FLOYD, S. (1995). Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking* **3** 226–244.
[15] TUKEY, J. W. (1986). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In *The Collected Works of John W. Tukey* **3** (L. W. Jones, ed.). Wadsworth, Monterey.
[16] WILLINGER, W., PAXSON, V. and TAQQU, M. S. (1997). Self-similarity and heavy tails: structural modeling of network traffic. In *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions* (R. Adler, R. Feldman and M. S. Taqqu, eds.). Birkhäuser, Boston. To appear.
[17] WILLINGER, W., TAQQU, M. S., SHERMAN, R. and WILSON, D. V. (1995). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *Computer Communications Review* **25** 100–113.
[18] WILLINGER, W., TAQQU, M. S., SHERMAN, R. and WILSON, D. V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level (extended version). *IEEE/ACM Transactions on Networking* **5** 71–86.

AT&T LABS–RESEARCH
FLORHAM PARK, NEW JERSEY 07932
E-MAIL: walter@research.att.com

LAWRENCE BERKELEY NATIONAL LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720
E-MAIL: vern@ee.lbl.gov

## REJOINDER

SIDNEY I. RESNICK

*Cornell University*

I am grateful to the discussants for the stimulating and informative comments on the paper.

Drs. Willinger and Paxson argue strongly for abandoning black box time series models in favor of structural models. Their criticisms deserve to be taken quite seriously for at least two reasons. First, as industry insiders, they are in an excellent position to judge what is useful, and second, one of the dirty little secrets of heavy tailed ARMA modeling, is that there are few if any success stories where real data with dependencies have been successfully modeled by a linear time series model. There are many potential reasons for the failure of linear models to successfully capture dependencies in heavy tailed data [see

Feigin and Resnick (1996) and Resnick (1997) for some discussion] but suffice it to say, there now exist both adequate theoretical results and software tools for ARMA models so that if ARMA modeling were going to work, it would have produced some successes by now.

Over the past years, it has been natural to experiment with traditional time series models in nontraditional contexts, one of which was heavy tailed phenomena. This yielded an abundance of research suggesting estimators for coefficients in heavy tailed ARMA models, and this research was worthwhile. You cannot know, for instance, that a heavy tailed autoregressive model will not adequately fit a data set without reasonably sophisticated theoretical and software tools giving estimation and goodness-of-fit techniques. Hopefully, the theory and software developed for the ARMA process will provide a useful basis for examining other heavy tailed time series models such as nonlinear models, hidden Markov chain models and the GARCH models rightly mentioned by Professor Beran. Ideally, these alternate models will have the flexibility to capture dependencies across a broad range of fields, not only in teletraffic engineering but also in finance and insurance where an abundance of large heavy tailed data sets is available to those with proper contacts. See, for example, Embrechts, Klüppelberg and Mikosch (1997), McNeil (1997) and Resnick (1997). These data sets rival in quality and length those emerging in teletraffic engineering and, for instance, the tick-by-tick data obtained for currency exchange rates can be arbitrarily large.

The focus of Drs. Willinger and Paxon on structural models is absolutely appropriate considering their obligations and goals. Some thought will have to be given to either statistical or qualitative measures so that we have some guidance in deciding when a parsimonious structural model is an adequate description of reality. How do we know we can believe what a structural model tells us? What if some predictions of a model match reality for the wrong reasons? *On–off* transmission models are very appealing structural models but how do we decide if it is satisfactory to neglect correlations in the data of on–off times? Also, how does one balance the desire for simplicity and parsimony with the fact that ever increasing supplies of data should make fitting parametric models with larger numbers of parameters feasible?

Structural modeling of teletraffic data is certainly worthwhile, and we look forward to the leadership industry experts provide in suggesting specific structural models for investigation. Academic researchers whose interests extend to heavy tailed modeling in other fields besides teletraffic engineering may wish to pursue other models as well. The mix of approaches is healthy and likely to be fruitful.

The comments of Drs. Willinger and Paxon highlight the differences between doing research as an industry insider versus as an outsider. Academic researchers need to ally themselves as closely as possible with the people close to the source of the problems. Getting bootlegged data third or fourth hand is not always a prescription for research success. Inevitable questions about the data, lack of documentation and lack of completeness are all possible pitfalls unless there is a close tie with the source.

Professor Adler's comments, as was to be expected, are sharp and pertinent. Why not use transformations? The easy answer is that it did not occur to me. The Jewish response (answer a question with a question) would be "Which transformations should we use?" The fuller response is that if you grow up in a Gaussian world, transformations may seem like a natural tool. If you grow up in an extreme value world, the tails make the subject interesting and special and obliterating tails by transforming is off-putting. This is more a statement of prejudice than science, but I do not think it is the right way to go. In any case, modeling dependence structure currently challenges the heavy tailed time series researcher, and transformations are unlikely to clarify the dependence structure.

Why not use stable? What do we do when we get estimates of $\alpha \in (3, 4)$, as frequently happens in economics? A lot of heavy tailed data looks much more Pareto than stable. However, this is definitely worth pursuing. One has to deal with finding reliable goodness-of-fit tests to insure that modeling with stable distributions is a good strategy, and one also has to investigate how robust a stable designed procedure would be to departures from the stable assumption. This could presumably be done and is worthwhile. Again, the issue is not so much modeling the marginal distribution but rather successfully capturing the tail and the dependence structure.

I am not ready to write off the Hill-like estimators. Using a Hill cocktail of several procedures, refinements and graphical methods seems to provide a reasonable estimate. Maybe further improvements will be forthcoming.

The comments by Professor Adler on rates of convergence are interesting. I doubt the rate of convergence is slow in all cases and probably, if extreme value theory is a guide [see de Haan and Resnick (1996)], the rate can be anything depending on the second-order regular variation of the underlying distribution. Stable does not behave particularly well for rates of convergence; the Pareto distribution would do better.

Bootstrapping heavy tailed phenomena is not likely to yield practical triumphs. If the sample size is $n$, you have to reduce the bootstrap sample size to $k = k(n) = o(n)$, and then problems similar to what arises with the Hill estimator occur. How do you choose $k$? I doubt this can be made practical. More comments on this are in Feigin and Resnick (1997) and can be consulted by those with stamina.

Professor Beran points out fascinating parallels between tail estimation and spectral density problems involving estimating the Hurst parameter. I have often wondered if there are deeper reasons for the parallels. Professor Beran provides a useful service when he points out that there is much relevant literature in robust time series methods. Concerning heteroscedastic models, it is known that the ARCH(1) model has a heavy Pareto tail. The proof is deep, relying on results of Kesten (1973). [See also Goldie (1991), Vervaat (1979) and Embrechts, Küppelberg and Mikosch (1997).] This result is probably true in much greater generality [some evidence is in Embrechts, Samorodnitsky, Dacorogna and Muller (1996)] and provides a class of models where input variables are light tailed but output process marginals are heavy tailed. This

is in contrast with the usual ARMA model assumptions where innovations are heavy tailed and hence process marginals are heavy tailed. The usual truncation of series methods that work for ARMA models fail for the random walk-based methods applied to ARCH and much subtler techniques are required. Such models will undoubtedly be relevant for modeling finance data which often have the charming feature that the data appear uncorrelated but the absolute values or the squares of the data appear almost to exhibit long range dependence.

I look forward to being in further touch with the discussants and other interested parties about these topics of interest.

## REFERENCES

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Heidelberg. To appear.

EMBRECHTS, P., SAMORODNITSKY, G., DACOROGNA, M. and MULLER, U. A. (1996). How heavy are the tails of a stationary HARCH(k) process? A study of the moments. Unpublished manuscript. Available at http://www.orie.cornell.edu/trlist/trlist.html as TR1172.ps.Z.

FEIGIN, P. and RESNICK, S. (1997). Linear programming estimators and bootstrapping for heavy tailed phenomena. *Adv. in Appl. Probab.* To appear.

FEIGIN, P. and RESNICK, S. (1996). Pitfalls of fitting autoregressive models for heavy-tailed time series. Unpublished manuscript. Available at http://www.orie.cornell.edu/trlist/trlist.html as TR1163.ps.Z.

GOLDIE, C. M. (1991). Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.* **1** 126–166.

DE HAAN, L. and RESNICK, S. (1996). Second-order regular variation and rates of convergence in extreme-value theory. *Ann. Probab.* **24** 97–124.

KESTEN, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta. Math.* **131** 207–248.

MCNEIL, A. (1997). Estimating the tails of loss severity distributions using extreme value theory. *Astin Bulletin*. To appear.

RESNICK, S. (1997). Discussion of the Danish data on large fire insurance losses. *Astin Bulletin*. To appear. Available at http://www.orie.cornell.edu/trlist/trlist.html as TR1178.ps.Z.

VERVAAT, W. (1979). On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Adv. in Appl. Probab.* **11** 750–783.

CORNELL UNIVERSITY
SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
ETC BUILDING
ITHACA, NEW YORK 14853
E-MAIL: sid@orie.cornell.edu