



On the controversy over tailweight of distributions

C.C. Heyde^{a,b,*}, S.G. Kou^c

^aDepartment of Statistics, Columbia University, New York, NY 10027, USA

^bMathematical Sciences Institute, Australian National University, ACT 0200, Australia

^cDepartment of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA

Received 23 September 2003; received in revised form 26 October 2003; accepted 28 October 2003

Abstract

Although understanding tail behavior of distributions is important in many areas, such as telecommunications network analysis and finance, there is considerable controversy about distinctions between exponential-type and power-type tails. This paper explains why the distinctions are surprisingly difficult for popular methods in the literature, and why particularly large samples are needed for clear discrimination.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Tail probability plots; Hill ratio plots; Mean excess function; Queuing theory; Insurance risk analysis; Finance

1. Introduction

The behavior of most probabilistic models is critically influenced by the tails of the distribution(s) which drive the models. This is especially important in such areas as asset pricing, value-at-risk, reliability, insurance risk, queueing theory and telecommunications network analysis. The literature, however, reveals considerable uncertainty, controversy, and error in connection with the distinctions between tail weights. In this paper we will explain why the distinctions are surprisingly difficult, why particularly large samples, perhaps in the tens of thousands or even hundreds of thousands, are necessary for clear discrimination and why theoretical errors have confused the discrimination.

To illustrate the qualitatively different behavior engendered by different tailweights we first examine the record behavior of a sample X_1, X_2, \dots, X_n of i.i.d. random variables with distribution function F . Clearly, $P(\max_{1 \leq k \leq n} X_k > x) = 1 - F^n(x)$. If $1 - F(x) = cx^{-\alpha}$, $\alpha > 0$, for $x > 0$, it is easily checked that $P(\max_{1 \leq k \leq n} X_k > (ny)^{1/\alpha}) \rightarrow 1 - e^{-c/y}$. Thus, for a power tail, $n^{-1/\alpha} \max_{1 \leq k \leq n} X_k$ converges in distribution, that is the records increase at a power rate. If, on the other hand, the tail is exponential, $1 - F(x) = ke^{-x/\beta}$, $\beta > 0$ for $x > 0$, then it is easily checked that $P(\max_{1 \leq k \leq n} X_k > (y/\beta) \log n) \rightarrow 1 - e^{-k/y}$, that is, $\beta(\log n)^{-1} \max_{1 \leq k \leq n} X_k$ converges in distribution; the records increase at a logarithmic rate. As another illustration, a simple queueing example in which tail behavior can influence performance unexpectedly concerns the waiting time process $\{W_n\}$ of the customers in a stationary M/G/1 queue. Here the usual asymptotic normality of $n^{-1/2}(W_1 + \dots + W_n - nEW)$ breaks down if the service time S has $ES^4 = \infty$, not as might be

* Corresponding author. Department of Statistics, Columbia University, New York, NY 10027, USA.

E-mail addresses: chris@stat.columbia.edu (C.C. Heyde), sk75@columbia.edu (S.G. Kou).

expected $ES^2 = \infty$ (see [14]). The distinction between power and exponential tails may appear to be easy to observe, but it is not.

The idea that there is a true statistical model for any real data set is a (convenient) theoretical idealization. Cox [4] has usefully pointed to the role and purpose of statistical models being to provide a “concise description of the aspects of the data judged relevant for interpretation”. Kac [20] has written that “Models are, for the most parts, caricatures of reality, but if they are good, then, like good caricatures, they portray, though perhaps in distorted manner, some of the features of the real world.”

We should not expect more than a parsimonious description. Indeed, any assumed distributional form is likely to be rejected in a test of adequacy when one has a (sufficiently) large data set, this usually being a reflection of departures, which may be quite small, in the middle part of the distribution. Perhaps even 5000 observations may be sufficient for rejection, and in some contexts, such as telecommunications, 50,000 observations may be routinely available. Introducing additional model parameters, up to three or four, may appear to help a little, but does not resolve the intrinsic difficulty. In fact, if a true distribution is assumed, it is known to be impossible to consistently estimate its density in the extreme tail (see [2]).

Therefore, it seems to us to be more appropriate to ask what type of distribution is suitable for the data than to seek some best-fitting distribution; and we will make this issue the focus of the paper. This apparently straightforward distinction is surprisingly difficult to make, and requires a substantially larger sample than what is intuitively expected to achieve good statistical precision. It turns out that the distinction rests on an extremely small percentage of the sample, perhaps less than 0.01%! See Sections 3 and 4.

First we need to be clear about the terminology, because different specialist areas have established somewhat different norms. For example, what may be described as a light tail in the queueing literature could be called a semi-heavy tail in the finance literature. In this paper, we will settle on the simple dichotomy of calling tails light or heavy according to whether or not a finite moment generating function exists. For a detailed classification of tails see [9, p. 50], and references therein.

Heavyness of tails is a subject of controversy in some areas, but especially in finance. For example, it is accepted that real returns data has tails which are heavier than those of the normal distribution, but one school of thought believes the tails to be light (typically exponential) and the other believes that they are heavy (typically powers). Representatives of the former school of thought are [1,8]. The advocacy of power tails dates back at least to [24]; for recent advocacy in book form see [25] and for associated modeling insights see [10]. For further substantive discussion see, e.g., [11,12,15–18,21–23]. Some authors conclude that both tail types are required (e.g. Chapter 13 in [26]).

Differences of opinion also appear in studies on World Wide Web traffic. Heavy tails are widely advocated (e.g. [5,6,28]), but also Weibull and exponential tails (e.g. [3]). See [13] for a nonparametric estimation of tail probabilities for a single server queue.

2. Standard methods of discrimination and estimation

2.1. The methods

We shall begin by mentioning six of the most widely used methods, together with their strengths and weaknesses. These methods, and variants of them, essentially span the spectrum of standard theory. The first five are discussed in considerable detail in the literature, for example in the books of [9,26], but insufficient attention has been given to their intrinsic capacity to discriminate. The sixth method, although widely used in practice, has a serious theoretical flaw, and should not be employed except under very special circumstances.

Throughout the study below, we will assume independent identically distributed samples, Y_1, Y_2, \dots . To make a further simplification, we shall assume the distribution is either one-sided or symmetric. Therefore, in the two-sided case, to study tail behavior, we can simply consider the random variables $X_i = |Y_i|$ instead of the original random variables Y_i , $i \geq 1$, and focus on the right tail; this effectively doubles the sample size, compared to considering both tails. If the distribution is asymmetric, then the problem becomes even harder, requiring larger sample sizes.

Method 1: The *tail-probability plot* has been widely used in the literature to study the tail behavior. The method relies on the fact that for exponential tails with mean θ

$$P(X \geq u) \approx \int_u^\infty \frac{1}{\theta} e^{-x/\theta} dx = e^{-u/\theta},$$

and for power tails with the tail index $\alpha > 1$, we have that

$$P(X \geq u) \approx C \int_u^\infty \frac{1}{x^\alpha} dx = C \frac{1}{\alpha - 1} u^{-\alpha+1}.$$

Therefore, taking the logarithm we have for exponential type distributions

$$\log P(X \geq u) \approx -u/\theta,$$

and for power type distributions

$$\log P(X \geq u) \approx \log C_\alpha - (\alpha - 1) \log u.$$

This suggests that if we have the log–log plot of the tail probability, then for power type tails it should display a straight line, but not for exponential type tails.

Method 1 has the appearance of a robust discriminator, but the clarity that comes from taking log–log plots is compromised by the fact that the range of u values for which data is available is generally small. If the data is standardized to make the standard deviation one, then the typical range of u values might be $1 \leq u \leq 6$, or at most $1 \leq u \leq 8$, observations of more than 8 standard deviations being very rare. But for such a range, the distinction between $C_1 u$ and $C_2 \log u$, with unknown constants C_1, C_2 , may be intrinsically difficult to discern.

Method 2: The popular *mean excess function* method is based on the function $e(u) = E(X - u | X > u)$. For exponential tails with $X = |Y|$,

$$\begin{aligned} E[X | X \geq u] &= \frac{E(XI(X \geq u))}{P(X \geq u)} \\ &\approx \frac{\int_u^\infty (x/\theta) e^{-x/\theta} dx}{\int_u^\infty (1/\theta) e^{-x/\theta} dx} \\ &= \frac{e^{-u/\theta}(u + \theta)}{e^{-u/\theta}} = u + \theta, \end{aligned}$$

so that

$$e(u) = \theta,$$

while for power tails with the power index $\alpha > 2$, we have that with $X = |Y|$,

$$E(XI(X \geq u)) \approx C \int_u^\infty x \frac{1}{x^\alpha} dx = C \frac{1}{\alpha - 2} u^{-\alpha+2},$$

$$\begin{aligned} E[X | X \geq u] &= \frac{E(XI(X \geq u))}{P(X \geq u)} \\ &\approx \frac{[1/(\alpha - 2)]u^{-\alpha+2}}{[1/(\alpha - 1)]u^{-\alpha+1}} = \frac{\alpha - 1}{\alpha - 2} u \end{aligned}$$

and

$$e(u) = \frac{u}{\alpha - 2}, \quad \alpha > 2.$$

This suggests that we can plot the empirical value of $e(u)$ against u : (a) If the plot is a linear curve, then it may be either a power type or exponential type distribution; otherwise it may reflect some other type of tail behavior. (b) If the slope of the linear curve is greater than zero and the intercept is zero, then it suggests a power type; otherwise, if the slope is equal to zero, it suggests an exponential type.

Method 2 avoids the problem of the compression of scale which comes about by taking logarithms in Method 1. However, the problem with the small range of u values available in practice remains. Furthermore, there are families of widely used distributions, such as the Weibull, which have an intermediate behavior of the form $e(u) \approx Ku^{1-\beta}$ for some $K > 0$ and whose presence may further cloud the choice of model.

Method 3: The *moment generating function* method. For exponential tails

$$\begin{aligned} E(e^{\gamma X} I(X \geq u)) &\approx \int_u^\infty e^{\gamma x} \frac{1}{\theta} e^{-x/\theta} dx \\ &= \begin{cases} \frac{1}{1 - \gamma\theta} e^{-u(1-\gamma\theta)/\theta}, & \theta\gamma < 1, \\ \infty, & \theta\gamma > 1 \end{cases} \end{aligned}$$

and for $\gamma < 1/\theta$ we have

$$\begin{aligned} E[e^{\gamma X} | X \geq u] &= \frac{E(e^{\gamma X} I(X \geq u))}{P(X \geq u)} \\ &\approx \frac{e^{-u(1-\gamma\theta)/\theta}}{e^{-u/\theta}} \cdot \frac{1}{1 - \gamma\theta} = \frac{e^{\gamma u}}{1 - \gamma\theta}, \end{aligned}$$

so that

$$\log E[e^X | X \geq u] \approx \gamma u - \log(1 - \gamma\theta).$$

For power tails with $\alpha > 2$ we have that

$$E(e^{\gamma X} I(X \geq u)) \approx C \int_u^\infty e^{\gamma x} \frac{1}{x^\alpha} dx = \infty,$$

for any γ . This suggests that one can plot, for small γ , $\log E[e^{\gamma X} | X \geq u]$ against u , and check the slope: (a) If the plot is approximately linear in u with slope γ , then the distribution may be an exponential type. (b) If the curve blows up, then it may not be an exponential type, and it is possibly a power type.

Here there is an advantage of the linearity in u , but the disadvantage is the uncertainty about what increase constitutes, in practice, a blowing-up for distributions for which a moment generating function does not exist.

Method 4: The next popular method is the *max-sum ratio plot*. The idea is based on the fact that for stationary $\{Y_i\}$,

$$R_n(p) = \frac{\max(X_1^p, \dots, X_n^p)}{X_1^p + \dots + X_n^p} \rightarrow 0 \quad \text{a.s.}$$

as $n \rightarrow \infty$, if and only if $E[X_1^p] < \infty$. Therefore, if one plots $R_n(p)$ against n for various p , and finds that $R_n(p)$ jumps up at some large n , then it is an indication that $E[X_1^p] = \infty$, which means that X has a power type tail.

Method 4 has the advantage of requiring minimal assumptions about the distribution of the data, only a moment condition which can be varied flexibly for exploratory investigation. However, in practice $R_n(p)$ may jump significantly at extreme values of the process, and convergence may not seem convincing. The larger p , the more dramatic is the effect. For example, even for $p = 8$, the difference between observations of 6 and 6.5 magnifies to $6^8 = 1.68 \times 10^6$ compared with $(6.5)^8 = 3.19 \times 10^6$.

Methods 1–4 continue to apply even if the processes generating the data are (strictly) stationary, rather than of independent and identically distributed variables. This is a consequence of the ergodic theorem, and no assumptions about asymptotic independence are needed. Unfortunately, they all have the disadvantage of being qualitative in nature, with no formal assessment of statistical precision.

Method 5: The fifth method is the *generalized Hill ratio plot*. The method uses the asymptotic result that if the tail probability is regularly varying

$$P(X > x) = \frac{L(x)}{x^\beta},$$

where $L(x)$ is a function which is slowly varying at infinity and $\beta > 0$, then

$$a_n(x_n) = \frac{\sum_{i=1}^n I(X_i > x_n)}{\sum_{i=1}^n \log(X_i/x_n) I(X_i > x_n)},$$

I being the indicator function, is a consistent estimator of β , if x_n is chosen so that $P(X > x_n) \rightarrow 0$ and $nP(X > x_n) \rightarrow \infty$. The standard Hill estimator corresponds to the particular case where the observations are ordered $X_{(n)} \leq X_{(n-1)} \leq \dots \leq X_{(1)}$ and $x_n = X_{(k_n+1)}$, where k_n is an integer (which tends to infinity as n tends to infinity).

In practice, one plots $a_n(x_n)$ against x_n , and hopes to find some stable region from which to determine a value of $a_n(x_n)$ as an estimator of α . There is a significant trade-off between small values of x_n , for which the estimation bias is large, and large values of x_n which produce small bias but large variance.

The Hill estimator and its variants are widely used in practice, often with apparent success, but the performance is erratic despite the availability of various consistency and asymptotic normality results. Hill “horror plots”, which have been published in [27,9, p. 194], of poor behavior, even under conditions for which the theory applies. The additional complications of dependent data, such as one meets in finance, may further cloud the performance picture.

Method 6: The last method is based on the use of *likelihood* and models the whole of the distribution, not just the tails. Necessarily this places some focus on the center of the distribution, for which most information is available. However, although it has been widely used (e.g. [17,18,25, Chapter 4]) it turns out that there is a serious problem with this method. Of course a likelihood approach could be arranged to focus directly on the tails, using conditional distributions of values over thresholds as in Method 5. The same problem that is indicated below with the Kullback–Leibler entropy continues to apply.

If the hypothesized density for the sample $\{Y_i, i = 1, 2, \dots, n\}$, which is assumed to be of independent variables, is $g(y)$ then the computed likelihood based on g is $L_g = \prod_{i=1}^n g(Y_i)$. The approach taken in the literature has been to say that the distribution with the largest likelihood from within a target class of competitors should be chosen. This apparently plausible approach masks a serious, possibly fatal, flaw.

Suppose that the (unknown) true density is $f(y)$. We have $\log L_g = \sum_{i=1}^n \log g(Y_i)$ and using the strong law of large numbers,

$$\log L_g \sim nE_f \log g(Y) \quad \text{a.s.}$$

the expectation being taken with respect to the true density. Now,

$$\begin{aligned} E_f \log g(Y) &= \int_{-\infty}^{\infty} \log g(y) f(y) dy \\ &= \int_{-\infty}^{\infty} \log f(y) f(y) dy - K(g, f), \end{aligned}$$

where

$$K(g, f) = - \int_{-\infty}^{\infty} \log(g(y)/f(y)) f(y) dy \geq 0$$

is the Kullback–Leibler entropy.

Therefore, the investigators who have used this method have, in essence, been seeking from among their target distributions, the one which minimizes $K(g, f)$. Unfortunately, the Kullback–Leibler entropy is not a metric, as is easily seen by noting that $K(g, f) \neq K(f, g)$, and there can be various unfortunate consequences of attempting to treat it as a distance.

To illustrate the difficulties, suppose that we have two target distributions, namely normal and Laplace, and that these are standardized to have mean zero and variance one. The Laplace (symmetric double exponential) distribution with variance one has density

$$\frac{1}{\sqrt{2}} e^{-\sqrt{2}|y|}, \quad -\infty < y < \infty. \quad (1)$$

Suppose also that the true underlying distribution of Y is t_v , $v > 2$, with mean zero and variance rescaled to one. This has the density

$$\begin{aligned} &\frac{1}{\sqrt{\pi(v-2)}} \cdot \frac{\Gamma((v+1)/2)}{\Gamma(v/2)(1+y^2/(v-2))^{(v+1)/2}}, \\ &-\infty < y < \infty. \end{aligned} \quad (2)$$

Then, if g_1, g_2 and f are, respectively, the densities of the normal target distribution, the Laplace target distribution and the standardized t_v true distribution, it is shown in the appendix that

$$\begin{aligned} &-E_f \log g_1(Y) < -E_f \log g_2(Y), \\ &\text{for } v = 8, 9, 10, \dots \end{aligned} \quad (3)$$

That is, while the normal distribution has tails which are much lighter than the Laplace and is further removed from those of the t_8 distribution, it could erroneously be preferred. This anomaly, appearing unexpectedly for the t_v distribution with degrees of freedom bigger than seven, highlights the risk of a wrong conclusion if this method is employed.

2.2. A numerical illustration

As an example to illustrate the difficulty in distinguishing between distributions we will perform a small simulation study. It should be emphasized that the purpose here is not to use the simulation to deliver a rejection to the standard methods. Rather, the goal is to gain some intuition for the further discussion of the sample size requirement in Sections 3 and 4. Furthermore, the calculation in the next two sections shows that the intuition revealed in the simulation may be quite general, and it is not the result of an unusual sample.

We shall consider a Laplace distributed random variable with variance 1 and a t random variable with d.f. $v=6$ rescaled to have variance 1, the densities being given by (1) and (2), respectively. Of course, the rescaling, which aims at giving the fairest comparison of the tails of the two distributions, does not alter the basic tail behavior of the distribution.

After generating 5000 samples of Y_i 's from each distribution, we then take absolute values to get $X_i = |Y_i|$. For simplicity, in the paper we shall refer to the two samples as Sample E (derived from 5000 observations from the Laplace distribution) and Sample T (derived from 5000 observations from the normalized t distribution with d.f. 6). The Laplace and t_6 distributions are chosen as typical of exponential and power tail cases respectively.

With 5000 sample size it is very easy to distinguish between Samples E and T via Kolmogorov–Smirnov goodness of fit test (for Sample E the test statistic gives 0.0414 with p -value being 0.00, and for Sample T the test statistic gives 0.0092 with p -value being 0.80), which seems to have a good power thanks to the large sample size here. However, it is important to note that we are able to do this mainly because of the small differences in the middle part (which is magnified due to the large sample size) of the two distributions, not

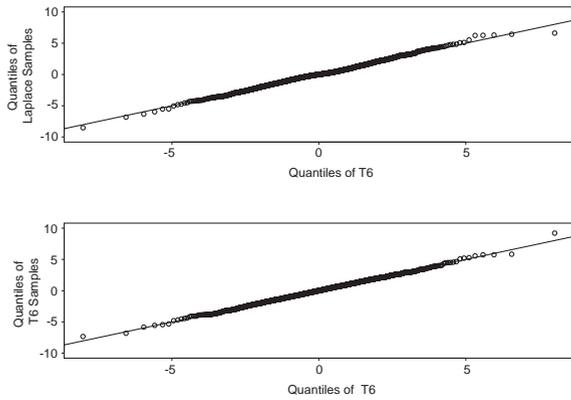


Fig. 1. QQ-plots for Samples E (upper) and T (lower).

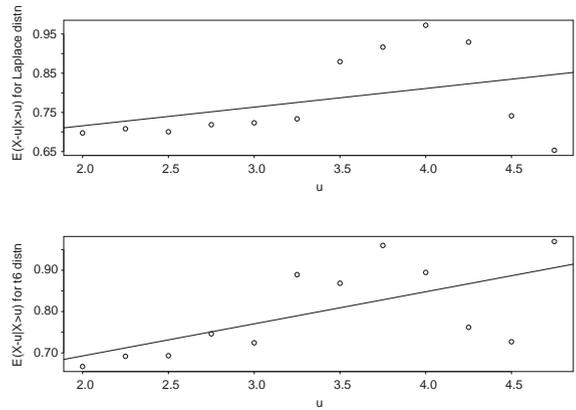


Fig. 3. Mean excess function plots for Samples E (upper) and T (lower).

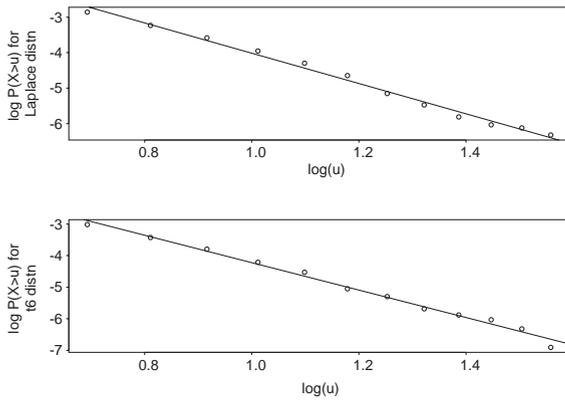


Fig. 2. Tail probability plots for Samples E (upper) and T (lower).

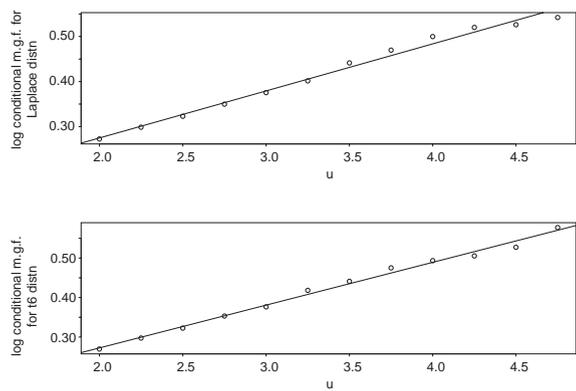


Fig. 4. Conditional m.g.f. plots for Samples E (upper) and T (lower).

because of the differences in the tail parts; the qq-plots in Fig. 1 illustrate this point.

To see how the log–log plot of Method 1 works, we apply it to the absolute values of Samples E and T. Taking absolute values should only increase the resolution of the method, as more observations are available in the right tails. However, Fig. 2 indicates that the method cannot distinguish the exponential tails from power tails, both samples exhibit a linear relationship (the R^2 being 0.9917 and 0.9914 for Samples E and T, respectively) in the log–log plot. Furthermore, the plot estimates $\hat{\alpha} = 5.26$ for Sample E (which is entirely wrong) and $\hat{\alpha} = 5.34$ for Sample T (while the true α is 7).

Method 2 with the mean excess function plot does not seem to be effective in our case. Fig. 3 illustrates

this point. In both cases, we do not see a strong linear relationship between $E[X - u | X \geq u]$ and u ; in fact R^2 are 0.1524 and 0.4078 for the exponential and t distribution cases, respectively. In the exponential case, the plot makes a mistake by estimating $\hat{\alpha} = 23.01$; while in the t distribution case, the estimate is $\hat{\alpha} = 14.86$, and the intercept is mistakenly regarded as nonzero.

In the case of the moment generating function plot of Method 3, we see a strong linear relationship with R^2 are 0.9882 and 0.9924 for the exponential (which should be linear) and t distribution cases (which should not be linear), respectively. Fig. 4 demonstrates the method for Samples E and T using $\gamma = 0.1$. The estimated slopes are, respectively, 0.1044 and

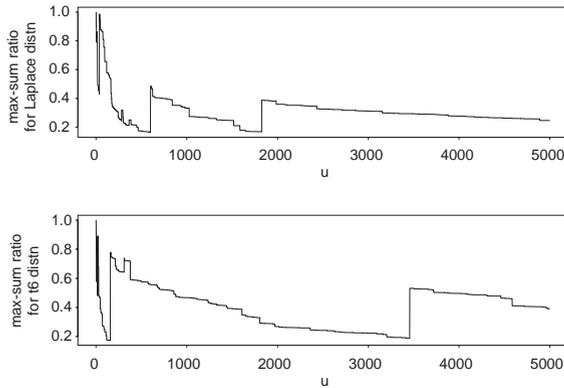


Fig. 5. Max-sum plots for Samples E (upper) and T (lower).

0.1085; both values are quite close to $\gamma = 0.1$, thus failing to distinguish the two tails.

In the case of the max-sum plot of Method 4, with $p = 6$, Fig. 5 shows that the exponential tails can also

have jumps for large n as well. This is mainly due to the fact that $R_n(p)$ is very sensitive to a single large value, especially for p not too small.

In the case of Method 5 the two simulations of both Samples E and T in Fig. 6 show that the plot can be highly variable, and clearly may fail to produce a reliable estimator of α . We have also tried different rate parameters for the Laplace distribution, and obtained similar simulation results. It is important to note that Hill estimation may not give any clue if it is inappropriately applied to a distribution which does not have power tails.

3. An explanation

After showing that the arguably most popular methods may fail to distinguish the difference between power type tails and exponential type tails for samples of size 5000, one may reasonably ask how large the

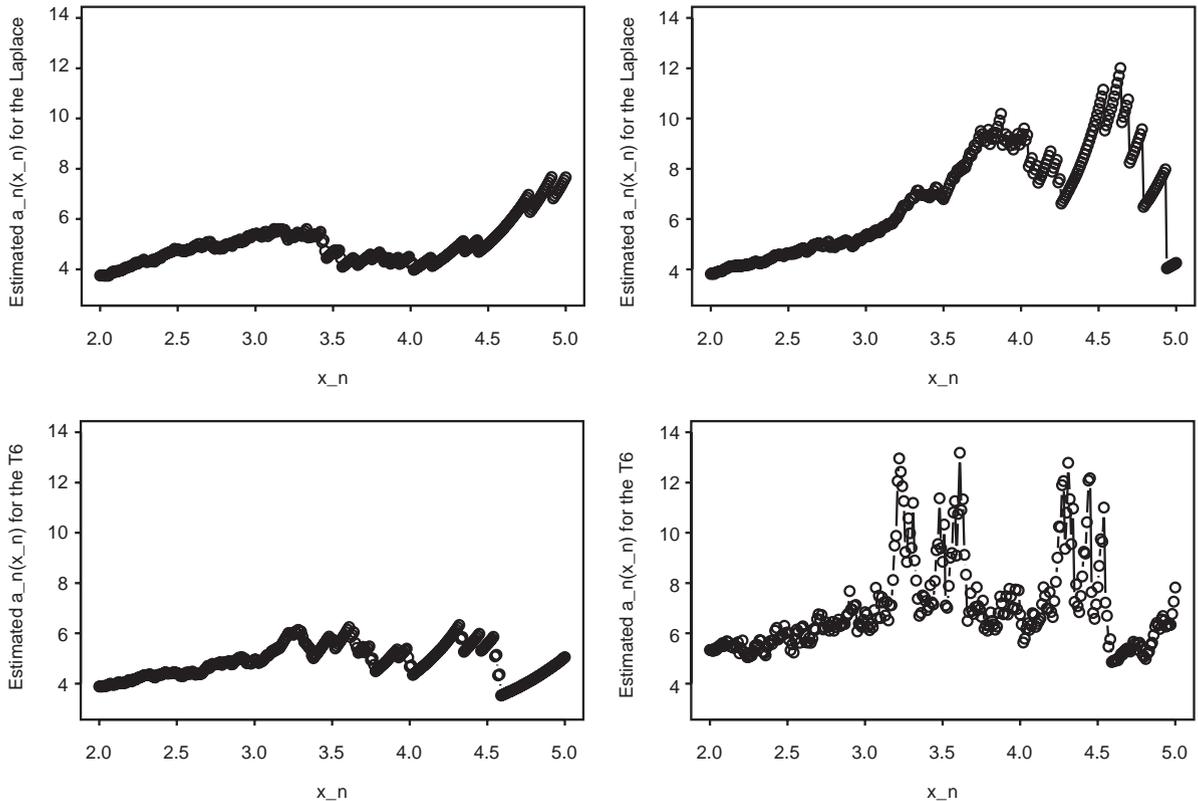


Fig. 6. Hill ratio plots for two simulated Samples E (upper) and T (lower).

Table 1
The (right) quantiles for the Laplace and normalized t densities as in (1) and (2)

Prob. (%)	Laplace	t7	t6	t5	t4	t3
1	2.77	2.53	2.57	2.61	2.65	2.62
0.1	4.39	4.04	4.25	4.57	5.07	5.90
0.01	6.02	5.97	6.55	7.50	9.22	12.82
0.001	7.65	8.54	9.82	12.04	16.50	27.67

Table 2
The (right) quantiles for the Laplace and t densities normalized by their interquartile ranges

Prob. (%)	Laplace	t7	t6	t5	t4	t3
1	2.82	2.11	2.19	2.31	2.53	2.97
0.1	4.48	3.37	3.63	4.02	4.81	6.66
0.01	6.14	4.93	5.56	6.63	8.78	14.51
0.001	7.80	7.08	8.36	10.68	15.74	31.33

sample size should be in order to make the distinction with appropriate confidence.

A reasonable impression may be obtained by simply looking at the quantile tables for both standardized Laplace and standardized t distributions. The quantiles for the Laplace and normalized t densities as in (1) and (2) are given in Table 1.

Table 1 shows that the Laplace distribution may have higher tail probabilities than those of t distributions with low degrees of freedom. For example, the 99.9% quantile of the Laplace distribution is actually larger than that of t distribution with d.f. 6 and 7! Thus, regardless of the sample size, the Laplace distribution may appear to be heavier tailed than a t -distribution with d.f. 6 or 7, up to the 99.9% quantile. In order to distinguish the distributions it is necessary to use quantiles with very low p values and correspondingly large samples, typically in the tens of thousands or even hundreds of thousands, as revealed in Table 1. Some specific assessments and general advice is given below in Section 4.

It may be thought that standardization using the variance is not an innocuous choice since the variance itself is sensitive to the extreme tail of the distribution. To this end, an illustration is given in Table 2 which adopts standardization using unit interquartile range. This standardization has the virtue of being able to

be used if some of the comparator distributions have infinite variance, such as is the case for non-normal stable laws. However, the results turn out to be very similar in Tables 1 and 2.

Overlapping of quantiles even for very low p values makes the job of distinguishing power and exponential type tails extremely difficult in practice. And the problem is further exacerbated by the fact that most large data sets are likely to be contaminated by departures from stationarity or autocorrelation structures.

4. Making assessments of the minimum sample size

Theoretical quantile results are useful to give a guideline of the *theoretical minimum sample size* required to distinguish two tails. For example, suppose we have two distributions, with the first one having asymptotically heavier tails. If the second distribution has a larger q th quantile, then obviously we cannot distinguish which one has the heavier tail at least up to the q th quantile. Therefore, the theoretical minimum sample size required is at least $1/(1 - q)$.

For example, to distinguish the Laplace distribution from the t distribution with d.f. 7, the theoretical minimum sample size should at least be $1/(0.01\%) = 10,000$, if the normalization is taken to be the standard deviation (as in Table 1 the 99.99% quantile for the Laplace distribution is 6.02 versus 5.97 for the t -distribution with d.f. 7), and at least be $1/(0.001\%) = 100,000$, if the normalization is taken to be interquartile ranges (as in Table 2 the 99.999% quantile for the Laplace distribution is 7.80 versus 7.08 for the t -distribution with d.f. 7).

Note that the computation of the theoretical minimum sample size is linked to the calculation of the quantiles, hence to the VaR literature (see, for example, [7,12]), which makes the calculation feasible.

In practice, the theoretical minimum sample size may not be large enough, as the quantiles can only be observed empirically (hence with a standard deviation) rather than observed precisely. One easy way to handle with this difficulty is by considering the asymptotic distributions of sample quantiles. For example, it is well known that for an i.i.d. sample of size n from an absolutely continuous distribution with density f , the $[np]$ (integer part of np) order statistic $X_{[np]}$ is

Table 3
Asymptotic standard deviation for (right) quantile values $X_{[np]}$ for $n = 5000$ and densities as in (1) and (2)

Prob. (%)	Laplace	t7	t6	t5	t4	t3
1	0.10	0.08	0.09	0.10	0.12	0.14
0.1	0.32	0.32	0.38	0.46	0.60	0.90
0.01	1.00	1.34	1.67	2.22	3.33	6.07
0.001	3.16	5.73	7.55	10.93	18.54	41.22

asymptotically normally distributed with mean ξ_p , the p th quantile, and variance $p(1-p)/(nf^2(\xi_p))$ (e.g. [29]).

Table 3 gives the standard deviations for the distributions used in Table 1 and for a sample size of 5000, typical of finance data sets. The standard deviations for sample size n can be calculated by dividing the table values by $(n/5000)^{1/2}$. Thus for sample size 50,000 the table entries are obtained by dividing each by $10^{1/2} \approx 3.16$.

It is immediately clear that there is no prospect of distinguishing the tabulated distributions on the basis of the even the 99.9% quantile with a sample of 5000. This is because, from Tables 1 and 3, the asymptotic 95% confidence interval of the 99.9% quantile are $(4.39 \pm 1.96 \times 0.32)$, $(4.04 \pm 1.96 \times 0.32)$, $(4.25 \pm 1.96 \times 0.38)$, $(4.57 \pm 1.96 \times 0.46)$, $(5.07 \pm 1.96 \times 0.60)$, $(5.90 \pm 1.96 \times 0.90)$, for Laplace and t7 to t3 distributions, respectively; and therefore the observed 99.9% quantile of the Laplace distribution may overlap with that of the other t -distributions. Conservative general advice (with the confidence level at 99% rather than 95%), to allow for rather broader families than are tabulated, is that one needs to use at least the 99.9% quantile and 50,000 observations for discrimination.

As a concrete example of the use of the tables consider the Hill plots in the upper section of Fig. 6. They are suggestive of a tail index of around 6 even though they are derived from a Laplace distribution. But Tables 1 and 3 make it evident that this is to be expected; it is not the result of an unusual sample.

As a final example, suppose that one has data suggestive of a large and possibly unstable kurtosis, which might be due to an infinite fourth moment. What sample size would one require to be reasonably sure that the data did not come from a distribution with a finite moment generating function? The t_ν distribution

has a finite or infinite fourth moment according as $\nu > 4$, $\nu \leq 4$ and the comparison between the Laplace and t_4 distributions is indicative of the distinction required. Consulting Tables 2 and 3 we see again that the use of the 99.9% or higher quantile would be required and the order of 50,000 observations would be necessary even at 95% confidence level.

Acknowledgements

We thank Professor Vadim Linetsky for many helpful suggestions. This research is supported in part by NSF grants DMS-0074637 and DMI-0216979.

Appendix

Sketched Proof of (3) Note that

$$-E_f \log g_1(Y) = \frac{1}{2} \log 2\pi + \frac{1}{2} E_f(Y^2) = \frac{1}{2} \log 2\pi + \frac{1}{2},$$

$$-E_f \log g_2(Y) = \frac{1}{2} \log 2 + \sqrt{2} E_f|Y|.$$

It follows from (2) (or e.g. from [19, p. 366]) that

$$E_f|Y| = \frac{2}{\nu-1} \sqrt{\frac{\nu-2}{\pi}} \frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)},$$

and then it is easily checked that $E_f|Y|$ is monotone increasing in ν along each of the subsequences $\nu = 3, 5, 7, \dots$ and $\nu = 4, 6, 8, \dots$. We then find that $-E_f \log g_1(Y) > -E_f \log g_2(Y)$ for $\nu = 3, 4, \dots, 7$, but $-E_f \log g_1(Y) < -E_f \log g_2(Y)$ for $\nu = 8, 9, 10, \dots$

References

- [1] O.E. Barndorff-Nielsen, N. Shephard, Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion), *J. Roy. Statist. Soc. Ser. B* 63 (2001) 167–241.
- [2] J. Beirlant, L. Devroye, On the impossibility of estimating densities in the extreme tail, *Statist. Probab. Lett.*, 43, 1999, 57–64.
- [3] J. Cao, W.S. Cleveland, D. Lin, D.S. Sun, Internet traffic tends towards Poisson and independent as the load increases, in: D.D. Denison, M.H. Hansen, C.C. Holmes, B. Yu (Eds.), *Nonlinear Estimation and Classification*, Springer Lecture Notes in Statistics, Vol. 171, Springer, New York, 2003.

- [4] D.R. Cox, Role of models in statistical analysis, *Statist. Sci.* 5 (1990) 169–174.
- [5] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, *IEEE/ACM Trans. Networking* 5 (1990) 835–846.
- [6] M.E. Crovella, M.S. Taqqu, A. Bestavros, Heavy-tailed probability distributions in the World Wide Web, in: R. Adler, R. Feldman, M.S. Taqqu (Eds.), *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Birkhauser, Boston, 1998, pp. 3–26.
- [7] D. Duffie, J. Pan, Analytical value at risk with jumps and credit risk, *Finance Stochastics* 5 (2001) 155–180.
- [8] E. Eberlein, U. Keller, Hyperbolic distributions in finance, *Bernoulli* 1 (1995) 281–299.
- [9] P. Embrechts, C. Klüppelberg, T. Mikosch, *Modelling Extremal Events*, Springer, Berlin, 1997.
- [10] X. Gabaix, P. Gopikrishnan, V. Plerou, H.E. Stanley, A theory of power law distributions in financial market fluctuations, *Nature* 423 (2003) 267–270.
- [11] P. Glasserman, P. Heidelberger, P. Shahabuddin, Variance reduction techniques for estimating value-at-risk, *Management Sci.* 46 (2000) 1349–1364.
- [12] P. Glasserman, P. Heidelberger, P. Shahabuddin, Portfolio value-at-risk with heavy-tailed risk factors, *Math. Finance* 12 (2002) 239–270.
- [13] P. Glynn, M. Torres, Nonparametric Estimation of tail probabilities for the single-server queue, in: P. Glasserman, K. Sigman, D.D. Yao (Eds.), *Stochastic Networks: Stability and Rare Events*, Lecture Notes in Statistics, Vol. 117, Springer, Berlin, 1996, pp. 109–138.
- [14] C.C. Heyde, Asymptotic efficiency results for the method of moments with application to estimation for queueing processes, in: J.W. Cohen, O.J. Boxma, R. Syski (Eds.), *Queueing Theory and its Applications. Liber Amicorum for*, North-Holland, Amsterdam, 1988, pp. 405–412.
- [15] C.C. Heyde, A risky asset model with strong dependence through fractal activity time, *J. Appl. Probab.* 36 (2000) 1234–1239.
- [16] C.C. Heyde, S. Liu, Empirical realities for a minimal description risky asset model: the need for fractal features, *J. Korean Math. Soc.* 38 (2001) 1047–1059.
- [17] S.R. Hurst, E. Platen, The marginal distribution of returns and volatility, in: Y. Dodge (Ed.), *L₁-Statistical Procedures and Related Topics*, IMS Lecture Notes Monograph Series, Vol. 31, IMS, Hayward, CA, 1997, pp. 301–314.
- [18] S.R. Hurst, E. Platen, S.T. Rachev, Subordinated Markov models, a comparison, *Fin. Eng. Japanese Markets* 4 (1997) 97–124.
- [19] N. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2, 2nd Edition, Wiley, New York, 1995.
- [20] M. Kac, Some mathematical models in science, *Science* 166 (1969) 695–699.
- [21] S.C. Kou, S.G. Kou, Modeling growth stocks via birth-death processes, *Adv. Appl. Probab.* 35 (2003) 641–664.
- [22] S.G. Kou, A jump diffusion model for option pricing, *Management Sci.* 48 (2002) 1086–1101.
- [23] A. Pagan, The econometrics of financial markets, *J. Empirical Finance* 15 (1996) 15–102.
- [24] P. Praetz, The distribution of share price changes, *J. Business* 45 (1972) 49–55.
- [25] S. Rachev, S. Mittnik, *Stable Paretian Models in Finance*, Wiley, New York, 2000.
- [26] R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values*, 2nd Edition, Birkhauser, Basel, 2001.
- [27] S.I. Resnick, Heavy tail modeling and teletraffic data, *Ann. Statist.* 25 (1997) 1805–1869.
- [28] K. Sigman, (Guest Editor). Queues with Heavy-tailed Distributions. *Queueing Systems Theory Appl.* 33 (1999).
- [29] A.M. Walker, A note on the asymptotic distribution of sample quantiles, *J. Roy. Statist. Soc., Ser. B* 30 (1968) 570–575.