In *Methodology and Computing in Applied Probability,* Vol 1 No. 1 (1999)

Estimating the Heavy Tail Index from Scaling Properties

Mark E. Crovella*
Computer Science Department
Boston University

Murad S. Taqqu†
Mathematics Department
Boston University

November 13, 1998

Abstract

This paper deals with the estimation of the tail index α for empirical heavy-tailed distributions, such as have been encountered in telecommunication systems. We present a method (called the "scaling estimator") based on the scaling properties of sums of heavy-tailed random variables. It has the advantages of being nonparametric, of being easy to apply, of yielding a single value, and of being relatively accurate on synthetic datasets. Since the method relies on the scaling of sums, it measures a property that is often one of the most important effects of heavy-tailed behavior. Most importantly, we present evidence that the scaling estimator appears to increase in accuracy as the size of the dataset grows. It is thus particularly suited for large datasets, as are increasingly encountered in measurements of telecommunications and computing systems.

1 Introduction

The presence of power-law behavior in the tail of a distribution has important implications for the behavior of a random variable: it may suggest the presence of infinite moments for example. For many purposes, the particular value of the exponent in such a power-law is also of prime importance. In this paper we present a new method for estimating such exponents in empirical data, and contrast it with previous methods. Our method is based on the *scaling behavior of sums* of random variables with power-law tails.

We say here that a random variable X follows a heavy-tailed distribution (with tail index α) if

$$P[X > x] \sim cx^{-\alpha}, \quad \text{as } x \to \infty, \;\; 0 < \alpha < 2,$$

where c is a positive constant, and where $\sim$ means that the ratio of the two sides tends to 1 as $x \to \infty$. This distribution has infinite variance; and if $\alpha \leq 1$ it has infinite mean.

The particular value of α is important in many practical situations, and a number of methods are commonly used to estimate α. The simplest and most important method is to examine the data directly. This can be done by plotting the complementary distribution (CD) function $\overline{F}(x) = 1 - F(x) = P[X > x]$ on log-log axes. Plotted in this way, heavy-tailed distributions have the property that

$$\frac{d \log \overline{F}(x)}{d \log x} \sim -\alpha,$$

for large x. Linear behavior on the plot for the upper tail is evidence of a heavy-tailed distribution. If such evidence exists, one can form an estimate for α by plotting the CD plot of the dataset and

*Partially supported by NSF grants CCR-9501822 and CCR-9706685 and by Hewlett-Packard Laboratories.
†Partially supported by NSF grants NCR-9404931, DMS-9404093, and ANI-9805623 at Boston University.

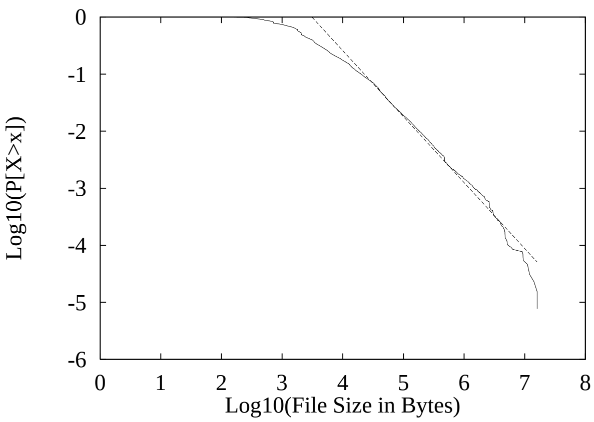


Figure 1: CD Plot for Web files transferred over network.

selecting a minimal value x_0 of x above which the plot appears to be linear. Estimating the slope for points greater than x_0 then gives an estimate of α.

A second approach to estimating the heavy tail index is by using the Hill estimator [4]. The Hill estimator gives an estimate of α as a function of the k largest elements in the data set; it is defined as

$$\mathcal{H}_{k,n} = \left(\frac{1}{k} \sum_{i=0}^{k-1} (\log X_{(n-i)} - \log X_{(n-k)}) \right)^{-1}$$

where $X_{(1)} \leq ... \leq X_{(n)}$ denote the dataset's order statistics, that is, the data items arranged according to size. In practice the Hill estimator is plotted against increasing values of k; if the estimator stabilizes to a consistent value this provides an estimate of α.

These two methods, while important, have the disadvantage that one must determine some point x_0 in the tail at which power-law behavior begins. This can be difficult, because often datasets do not show clear demarcations between power-law and non-power-law behavior in their empirical distributions. The proper choice of x_0 is important because it can have a significant impact on the estimate of α obtained using either the Hill or slope estimation methods.

As an illustration of this difficulty, consider the dataset shown in Figure 1. The Figure shows the CD plot of the sizes of files transferred through the World Wide Web (as described in [1, 2]). A straight line has been fitted to the tail of the distribution, using least squares fitting to only those points greater than 10^4 (*i.e.*, x_0 was chosen by eye to be 10^4).

Note that the fitted line deviates from the empirical distribution at both ends. Some deviation on the right is expected because there are very few sample observations in that range. However the deviation on the left presents more difficulty; the point at which the deviation begins depends on the choice of x_0. More importantly the slope of the line (an estimate of α) depends on the value of x_0 as well. Since the empirical CD plot shows gradual curvature in this region the proper choice of x_0 is not clear.

These problems associated with estimating α for empirical datasets motivate the work described in this paper. We present a method that helps identify the portion of a dataset's tail that exhibits power-law behavior. The method is based on the fact that the shape of the tail of a heavy-tailed distribution determines the scaling properties of the dataset when it is aggregated. By aggregating a dataset of n observations $X_i, i = 1, ..., n$ we refer to the process of summing non-overlapping blocks of observations of size m:

$$X_i^{(m)} = \sum_{j=(i-1)m+1}^{im} X_j.$$

By observing the distributional properties of $X^{(m)}$ we can make inferences about where in the tail power-law behavior begins. Based on these determinations we have the basis for forming an estimate of the tail index α.

While our method (called the *scaling* method) is useful in detecting and measuring heavy tailed behavior in a dataset, it can only suggest that such behavior is present; it cannot conclusively confirm heavy tailed behavior. It should be used in combination with inspection of the dataset. In order to support this mode of analysis we show how to use the estimation method graphically. We have developed a tool that outputs such graphical aids to assist in analyzing data; in particular, it can show the segment of the tail over which heavy-tailed behavior appears to be present.

We present results showing the performance of the scaling estimator on synthetic datasets of various lengths and drawn from a variety of distributions. Typical results indicate that on datasets of length 1,000, the estimator can produce a reasonably accurate estimate. Furthermore, when used on datasets of length 10,000 or 100,000 it is generally quite accurate. This property (increasing accuracy with increasing sample size) is especially important for datasets resulting from measurements of data communication and computing systems, since the nature of the data collection process in these systems allows for datasets of very large size (hundreds of thousands to millions of measurements) to be easily collected.

We also empirically compare the performance of the scaling estimator to that of Hill estimator (for fixed values of k). Our results show that the Hill estimator is more accurate than the scaling estimator on Pareto distributions; this is to be expected since for such distributions the Hill estimator is very accurate, and the choice of k generally has little effect on its value. However, for α-Stable distributions, we show that the scaling estimator is more effective over a broader range of α values than is the Hill estimator for any fixed k.

2 Measuring Scaling Properties

2.1 Background

We use the notation $X \stackrel{d}{=} Y$ to indicate that the random variables X and Y have the same distribution. Then $X \stackrel{d}{=} aY + b$ means that the distributions of X and Y differ only by location and scale parameters. For any random variable X we define Σ_n as a random variable that is the sum of n independent random variables each with the same distribution as X.

The usual central limit theorem describes the behavior of sums of random variables, but applies only to random variables with finite variance. For heavy-tailed distributions with tail index α, similar limit theorems may be formulated showing that sums of such variables converge to a stable distribution with the same α [3, 5, 6].

A distribution is stable (in the strict sense) if for each n there exist constants $c_n > 0$ such that

$$\Sigma_n \stackrel{d}{=} c_n X.$$

Note that the Normal distribution with zero mean is strictly stable with $c_n = \sqrt{n}$. This follows from the addition rule for variances, and the central limit theorem implies that the Normal distribution is the only stable distribution with finite variance. However, other stable distributions exist, and the constants c_n are always of the form $n^{1/\alpha}$. For any strictly α-Stable distribution $\mathcal{S}_\alpha$ the following holds:

$$s^{1/\alpha}X_1 + t^{1/\alpha}X_2 \stackrel{d}{=} (s+t)^{1/\alpha}X \tag{1}$$

whenever s and t are non-negative, and X, X_1, and X_2 have distribution $\mathcal{S}_\alpha$. When $\alpha = 2$, the strictly stable distribution $\mathcal{S}_\alpha$ is the normal distribution $N(0, \sigma^2)$.

Consider then a set of random variables drawn from a strictly stable distribution $\mathcal{S}_\alpha$. Then Equation 1 implies

$$\Sigma_n \stackrel{d}{=} n^{1/\alpha}X. \tag{2}$$

We refer to this property as the *scaling* property of sums of stable random variables.

The implication of Equation 2 is that the tail index is unchanged when independent stable random variables are summed. Furthermore, the relevant limit theorems state that, asymptotically, the tail index is unchanged when heavy-tailed random variables are summed. This effect can be observed in empirical data and is the basis for our method.

2.2 Observing the Scaling Property in Practice

It is relatively straightforward to observe the scaling property in the tail of the distribution. Figure 2 shows an example. In this example we started with 100,000 samples from a Pareto distribution with $\alpha = 1.1$. The Pareto distribution has complementary cumulative distribution function

$$\overline{F}(x) = P[X > x] = (k/x)^\alpha \quad x \geq k, \;\; \alpha, k > 0$$

and so is power-law over its entire range. After subtracting the empirical mean (which in this case was 8.107), the dataset was successively aggregated by factors of two (*i.e.,* adjacent points were summed). This process was repeated nine times, resulting in ten datasets. The complementary distribution of each dataset was then plotted on log-log axes.

The qualitative effect on which our method is based can be observed in the figure. Tails of successive datasets are approximately parallel, with slope approximately $-\alpha$.

2.3 Using the Scaling Property to Estimate α

The difficulty in applying a method based on the scaling property in practice lies in determining (1) the best aggregation factor to use and (2) where in the tail to measure α. Our approach is based on the following argument.

Suppose first that X_i is a set of i.i.d. random variables whose distribution is $\mathcal{S}_\alpha$, a strictly α-stable distribution with $0 < \alpha \leq 2$. Form $X_i^{(m)}$ as before, that is,

$$X_i^{(m)} = \sum_{j=(i-1)m+1}^{im} X_j$$

and let $X^{(m)}$ denote any of the $X_i^{(m)}$. We shall plot the complementary distribution $P[X^{(m)} > x]$ as a function of x and consider two curves for different values of m: $m_1 < m_2$. This situation is shown schematically in Figure 3. Then we can define two quantities, δ and τ, corresponding to the horizontal and vertical distances between the CD curves taken from a point $(\ln x_1, \ln P[X^{(m_1)} > x_1])$ on the m_1 curve.

We first evaluate the horizontal distance

$$\delta = \ln x_2 - \ln x_1, \tag{3}$$

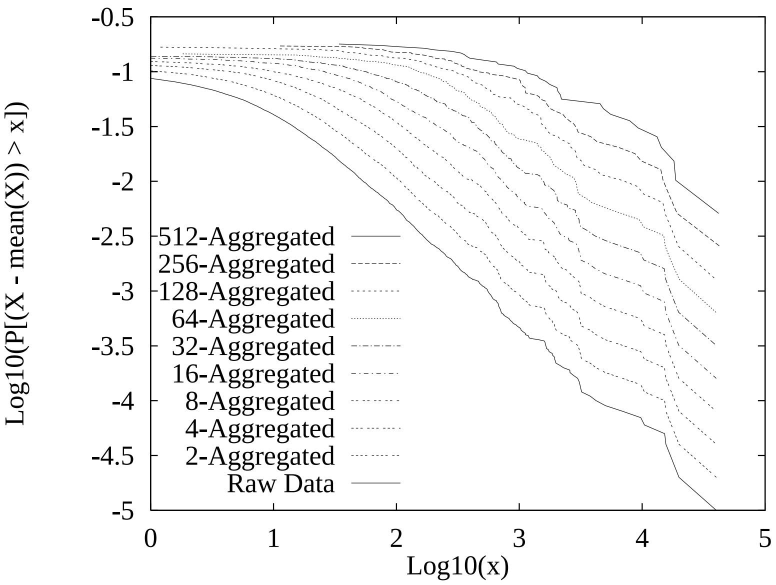


Figure 2: Scaling behavior in a synthetic dataset.

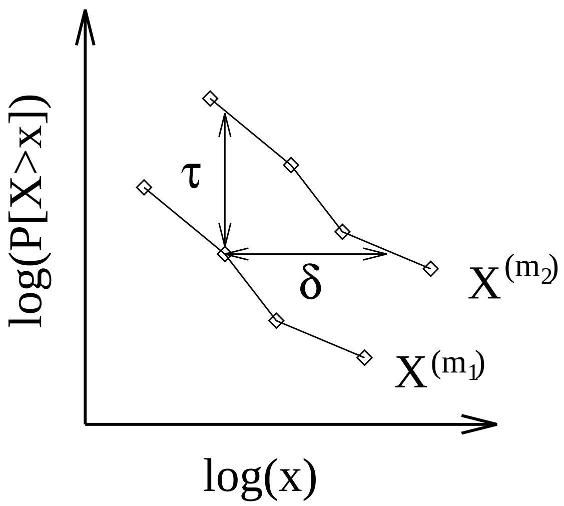


Figure 3: Measurements used in estimator.

where x_2 is such that $P[X^{(m_2)} > x_2] = P[X^{(m_1)} > x_1]$. Because of the scaling relation (2) one has

$$\frac{1}{m_2^{1/\alpha}} X^{(m_2)} \stackrel{d}{=} \frac{1}{m_1^{1/\alpha}} X^{(m_1)}$$

or

$$P[X^{(m_2)} > \left(\frac{m_2}{m_1}\right)^{1/\alpha} x_1] = P[X^{(m_1)} > x_1].$$

Therefore $x_2 = (m_2/m_1)^{1/\alpha} x_1$ and thus

$$\begin{aligned} \delta &= \ln x_2 - \ln x_1 \\ &= 1/\alpha \ \ln \frac{m_2}{m_1} + \ln x_1 - \ln x_1 \\ &= 1/\alpha \ \ln \frac{m_2}{m_1}. \end{aligned} \tag{4}$$

The horizontal distance δ can be estimated by using the empirical distribution functions of $X^{(m_1)}$ and $X^{(m_2)}$ and evaluated at different points x_1. In principle, one could use any value of x_1, measure the corresponding δ, and then use Equation (4) to obtain an estimate for α. Using an arbitrary x_1 makes sense only for perfectly scaling distributions (such as the stable or normal distributions) for which Equation (2) holds strictly (that is, throughout the body of the distribution: $P[\Sigma_n > x] = P[n^{1/\alpha} X > x]$ for all x). Since we want to use the method for distributions that have asymptotically power tails, and that may therefore only be in the domain of attraction of a stable distribution, we shall restrict the range of possible x_1 values. We will require x_1 to be in the tail of the distribution, and in addition, that the vertical distance

$$\tau = \ln P[X^{(m_2)} > x_1] - \ln P[X^{(m_1)} > x_1], \tag{5}$$

falls in a suitable range. In order to evaluate this vertical distance τ, we have to distinguish between the heavy-tailed case $0 < \alpha < 2$ and the Normal case $\alpha = 2$.

a) Consider first the heavy-tailed case $\alpha < 2$. If $P[X > x] \sim cx^{-\alpha}$ for large x, then

$$P[X^{(m)} > x] \sim mcx^{-\alpha} \quad \text{for large } x \tag{6}$$

(see [3, p. 278]).

A plot of the complementary distribution of $X_i^{(m)}$ on log-log axes will then show a line in the portion of the tail exhibiting the scaling property that is determined by:

$$\ln P[X^{(m)} > x] = \ln c + \ln m - \alpha \ln x.$$

Therefore, for a large family of heavy-tailed distributions if x is large enough, we have:

$$\begin{aligned} \tau &= \ln P[X^{(m_2)} > x_1] - \ln P[X^{(m_1)} > x_1] \\ &= (\ln c + \ln m_2 - \alpha \ln x_1) - (\ln c + \ln m_1 - \alpha \ln x_1) \\ &= \ln m_2/m_1. \end{aligned} \tag{7}$$

In principle, if x_1 is large enough to be in the scaling part of the tail of both $X^{(m_1)}$ and $X^{(m_2)}$, α could be estimated by the slope τ/δ, where τ and δ are measured using the empirical distribution functions of $X^{(m_2)}$ and $X^{(m_1)}$. The idea behind our method is to use only those x_1 for which τ is close to $\ln m_2/m_1$ (see Equation (7)) and then estimate α by

$$\hat{\alpha} = \ln(m_2/m_1)/\delta. \tag{8}$$

In this way one recovers the estimate one would use for the stable distribution. By focusing on the tails of the distribution, we can even consider small values of m_1 and m_2. This is because Equation (6) and the relation $P[X^{(m_2)} > x_2] = P[X^{(m_1)} > x_1]$ imply that δ approaches $1/\alpha \ln m_2/m_1$ as x_1 and x_2 tend to infinity.

Thus, each point along the tail of $X^{(m_1)}$ might potentially be used to form an $\hat{\alpha}$. This is the key to our method, because it allows us to identify a portion of the tail of $X^{(m_1)}$ for which the scaling property is holding.

If we set m_2/m_1 to a fixed aggregation step size f, Equation 8 becomes

$$\hat{\alpha} = (\ln f)/\delta \tag{9}$$

and Equation 7 reduces to

$$\tau = \ln f. \tag{10}$$

b) Finite-variance variables, when aggregated, will tend to a Normal distribution. The scaling relation (2) (self-similarity) holds when $\alpha = 2$ for the Normal distribution and hence we could forget about the τ criterion, estimate δ at arbitrary points x_1, and then use Equation (8) to obtain the corresponding estimate $\hat{\alpha}$. Recall, however, that our method uses only points for which $\tau \approx \ln m_1/m_2$. This is fine for the Normal distribution, but if X has an arbitrary distribution with finite variance, and if the number of points is not large enough to allow m_1 and m_2 to be large, and hence for the central limit theorem to hold, the corresponding estimate can be unreliable. As we will show below (see Section 2.6) our estimator allows one to discount spurious situations by assessing visually the region where scaling occurs.

2.4 An Algorithm

An algorithm employing this method is as follows (the algorithm is given more formally in Figure 4). Starting from an input data set X_i of length n and mean $\overline{X}$, subtract the mean from each value in the dataset, yielding new values for X_i. (Thus for the rest of this discussion X will refer to the original dataset with its mean subtracted.) Next choose an aggregation step size (f) and some number of aggregations to use (l). Then form $X^{(f^i)}$ for $i = 1, ..., l$. For example, taking $f = 2$ and $l = 10$, form $X^{(2)}, X^{(4)}, ..., X^{(1024)}$. Thus f is equal to the ratio m_2/m_1 as discussed in the previous paragraphs. In our experiments we use $f = 2$ since it is the smallest useful value, although larger values of f are permissible. The length of the most highly aggregated dataset will be approximately n/f^l; keeping this value large enough to form a reasonable CD plot helps suggest a value for l.

Having formed the datasets $X^{(f^i)}, i = 1, ..., l$, the next step is to construct the empirical log-log complementary distributions (CD) of each dataset. (For a given dataset, $P[X > x]$ is estimated by the proportion of data points exceeding x.) Then, repeat the following process for each CD plot:

1. For each point in the tail of the CD plot (which we define to be the upper 90th percentile of the dataset) measure the values δ and τ (the horizontal and vertical distances to the CD plot for the next higher aggregation level). These are defined in Equations (3) and (5) respectively.

2. Use δ to calculate a trial estimate of α using Equation 9, namely
$$\hat{\alpha} = (\ln f)/\delta.$$

3. Compare τ to $\ln f$ (see Equation 10); if they are close enough (we use a relative error of $\theta =$ 10%) then accept the trial estimate of α. Thus
$$\text{Accept } \hat{\alpha} \text{ if } |\tau - \ln f| < \theta \ln f \tag{11}$$

When this process has been completed for all l CD plots, report the *average value* of the trial α values that were accepted.

2.5 Discussion

The method takes into account:

(a) regions where there is power law shape, and

(b) regions where the distribution, properly rescaled, is invariant under aggregation (self-similarity).

For perfectly scaling distributions such as the Stable or Normal, one can forget about (a), and (b) will hold everywhere. For imperfectly scaling distributions (those that belong to the domain of attraction of a Stable distribution) the regions where (a) and (b) hold can be found, in theory, far in the tail and for large levels of aggregation m. In practice, when m is large, inaccuracies abound because one does not have many points. The method provides a compromise by using the relative error factor θ (see Equation 11). Because of the presence of θ, we can identify regions where (a) approximately holds when m is very small. Thus, for the Pareto distribution, the estimator yields reasonably good estimates even for small m. For the Stable distribution, while the power law shape (a) occurs only far out in the tail, effect (b) occurs over the whole distribution at all levels of aggregation.

Another issue affecting the robustness of the estimator is the operation of subtracting the empirical mean of the dataset at the outset of the estimation procedure. When $\alpha > 1$, a dataset with non-zero mean will not yield accurate estimates unless the empirical mean is subtracted at the outset. This is because the "central limit theorem" for random variables in the domain of attraction of a Normal or Stable distribution with $\alpha > 1$ requires the subtraction of the mean. Numerically, as the dataset is aggregated, its mean shifts, which can obscure the scaling relation used to estimate α (Equation 4). On the other hand, when $\alpha < 1$, the central limit theorem calls for no subtraction. Subtracting the empirical mean in this case is undesirable, because, if, for example the data is positive, it can result in making a large fraction (even the vast majority) of the data values negative. Negative values can not contribute to the estimate since the method operates on the upper tail of the distribution. Thus there is need for care in applying the estimator, and deciding about whether to subtract the empirical mean, when α is near 1 (*i.e.,* when α might be larger or smaller than 1). In this case it is important to examine the graphical output of the estimator to make sure that enough points are used to form a valid α estimate. (In tables and figures below, the empirical mean is subtracted even when $\alpha < 1$.)

2.6 Using the Method

In practice, power-law shape should exist over a significant fraction of the tail in order for the dataset to exhibit heavy-tailed behavior. Using the estimator described here it is possible to discount spurious cases by assessing visually which region of the dataset's distribution exhibits scaling, and by determining whether that scaling is occurring in the tail of the distribution.

In Figure 5 we show that same dataset as in Figure 2, but here we plot also the points used by the method in obtaining its α estimate. That is, those points are plotted for which $|\tau - \ln f| / \ln f < 0.10$. This highlights the region of the dataset's tail over which the scaling property is holding. As can be seen from the figure, this region encompasses the majority of the tail, spanning three orders of magnitude (which is to be expected since this dataset is indeed drawn from a heavy-tailed distribution).

Inputs

Input data points: $X_i \quad i \in 1, .., n; \quad \overline{X} = 0$
Aggregation Step Size: f (default: $f = 2$)
Number of Aggregation Steps: l (default: $l = 10$)
Maximum Allowable Relative Error in Forming α Estimate: θ (default: $\theta = 0.10$)

Algorithm

1. For $d \in 1, ..., l$ **[Form Aggregated Datasets]**
   (a) Let $m = f^d$
   (b) Form $X^{(m)}$: $X_j^{(m)} = \sum_{i=(j-1)m+1}^{jm} X_i$
2. For $d \in 0, ..., l$ **[Form Log Complementary Distributions]**
   (a) Let $m = f^d$
   (b) Let $U^{(m)}$ be the unique elements in $X^{(m)}$, in increasing order
   (c) Let u_m = the number of elements in $U^{(m)}$
   (d) For $i \in 1, ..., u_m$
      i. $P_i^{(m)} = \log(P[X^{(m)} > U_i^{(m)}])$
      ii. $L_i^{(m)} = \log(U_i^{(m)})$
3. Let $c = 0$
4. For $d \in 0, ..., (l-1)$ **[Collect trial αs]**
   (a) Let $m = f^d$
   (b) For $i = 1, ..., u_m$
      i. If $P_i^{(m)} > \log(1/10)$ goto next i [Only consider the 10% tails]
      ii. At point $(L_i^{(m)}, P_i^{(m)})$, measure τ and δ [see Figure 3]
      iii. Let $A_c = \log(f)/\delta$ [trial α]
      iv. Let $E_c = |\tau - \log(f)|/\log(f)$ [relative error of $\log(f)$]
      v. Increment c
5. Let $I = \{i | E_i < \theta\}$ **[Form $\hat{\alpha}$]**
   (a) If $|I| = 0$
         return **fail**
      else
         return $\hat{\alpha} = \frac{\sum_{i \in I} A_i}{|I|}$

Figure 4: Algorithm for α estimation.

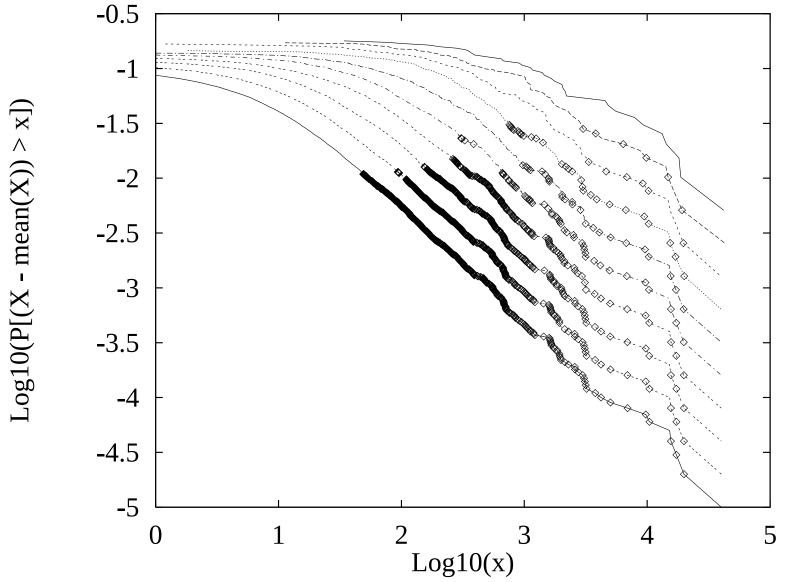


Figure 5: Power-Law shape region of tail for Pareto dataset.

Consider a Symmetric α-Stable distribution with $\alpha = 1.8$ (See Figure 6(a)). The points in the tail of the distribution for which $|\tau - \ln f|/\ln f < \theta$ with $\theta = 0.10$ are again highlighted. Observe that the vast majority of them do not lie in the far tail of the distribution. As it is well-known, a Stable distribution with an α close to 2 exhibits power-law shape only in the tail extremes. Nevertheless, the resulting estimate $\hat{\alpha} = 1.793$ is quite good. If we increase θ to 0.90 many more points are included in the estimation (See Figure 6(b)). Because the scaling relation (2) holds precisely for such a distribution, the estimate improves slightly and becomes $\hat{\alpha} = 1.801$. Too large a value of θ would worsen the estimate because points where the empirical distribution functions behave erratically would also be included. As a rule it is best to use the default 10% value.

The Normal distribution case (see Figure 7) displays a highlighting pattern similar to that of the Stable distribution with large α (Figure 6(a)), but with no highlighting in the extreme tails. When faced with a highlighting of the type of Figure 7, it is best in practice to conclude that there are no heavy tails. If α is less but close to two and one notices highlighting in the extreme tails (as in Figure 6) one may suspect a Stable distribution. To confirm this, it is best to follow up with a parametric test (one that assumes that the distribution is exactly Stable) and then compare the respective estimates of α.

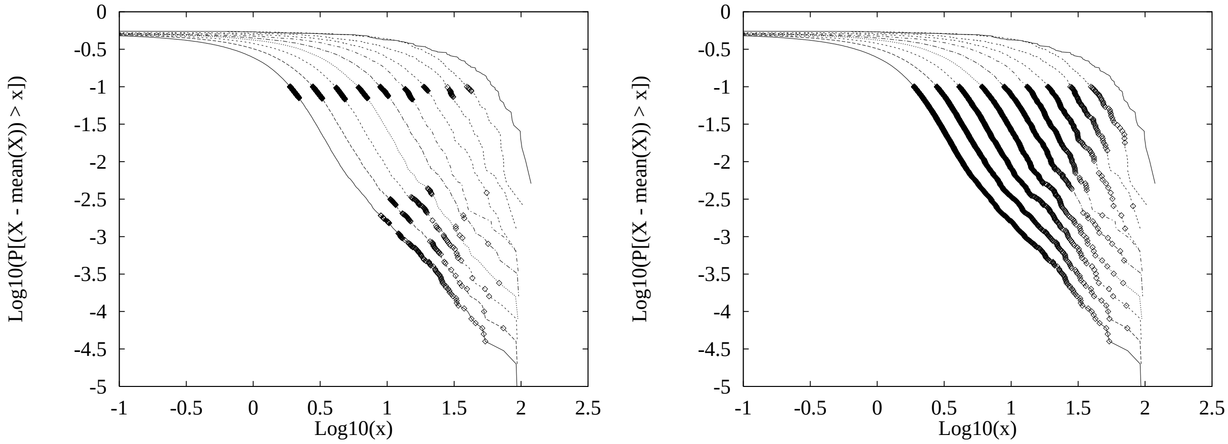
Finally, we show an example where the estimator highlights many points where $|\tau - \ln f|/\ln f < 0.10$, although the underlying distribution is not strictly heavy-tailed. Figure 8 shows the behavior of the estimator on a dataset of 100,000 samples drawn from a Lognormal distribution, in which $\ln X$ is normally distributed with $\mu = 0$ and $\sigma = 2$. In this case the tails of the dataset are so heavy that it appears to exhibit power-law shape over a large portion of the tail.

2.7 Using the Program

The scaling method has been implemented by the authors as a program written in C for use on Unix systems and called `aest`.[1] In this section we describe some details of the program's use.

An example invocation of the program might be:

[1] Source code for the program can be obtained from `http://www.cs.bu.edu/faculty/crovella/aest.html`.



(a) (b)

Figure 6: Scaling region of Symmetric-α-Stable; (a) $\theta = 0.10$; (b) $\theta = 0.90$.

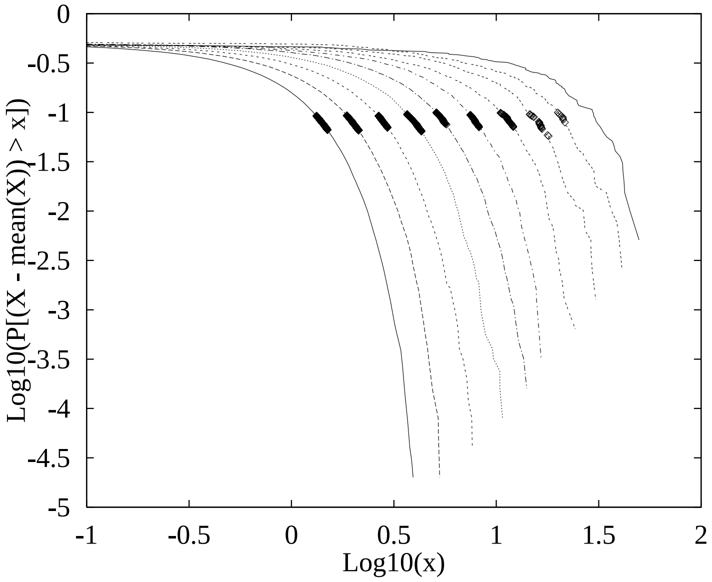


Figure 7: Scaling region for Normal dataset.

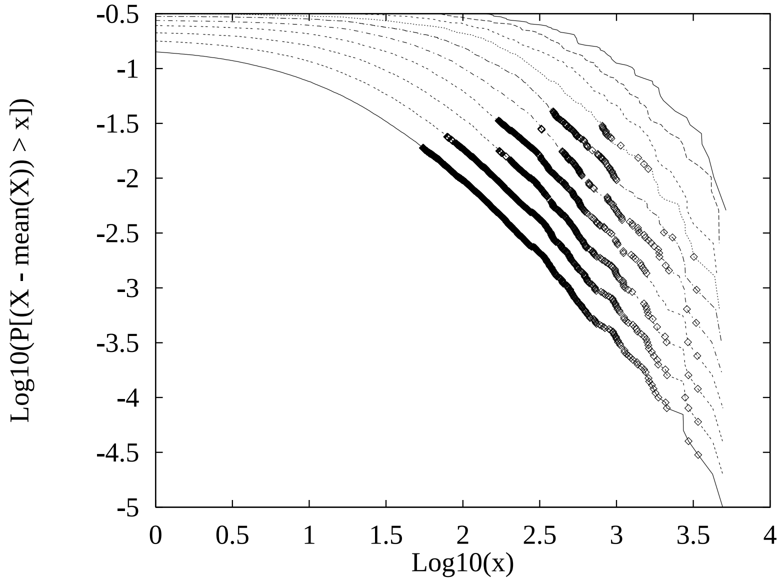


Figure 8: Power-Law region of tail for Lognormal dataset.

```
aest -f datafile -n 5000 -a 2 -l 10
```

This command means: run the estimator on the dataset contained in `datafile`, which contains 5000 points; use an aggregation factor $f = 2$, and aggregate over $l = 10$ levels. The result of running this command is output like:

```
Estimate: 1.068391
Subtracted mean: 8.107132
```

which reports the resulting α estimate and the empirical mean that was subtracted from all data points in the first step.

It is important to inspect the range of the distributional tail over which the scaling method appears to detect power-law behavior. For this reason the program outputs data suitable for direct plotting using the `gnuplot` plotting package. Invoking the program with the `-w` option will output files necessary to create all of the CD plots used by the program; adding the `-g` option will also show those points that met the relative error criterion and were used in forming the estimate. Since the data needed to create these plots can be large (larger in size than the original input dataset) the program accepts the `-s` option which will subsample the CD plots, yielding the same shape on visual inspection but eliminating redundantly plotted points. Thus, to form figures like those shown in this paper, a typical command would be:

```
aest -f datafile -n 5000 -a 2 -l 10 -w -g -s
```

This command will generate a set of files; each file consists of the `x,y` pairs for a single CD plot. There is also one file that contains `x,y` pairs for all of the points used in forming the α estimate (dark points in Figure 8). These files are referenced in the generated file `datafile.aest.gp` which includes commands for the plotting program `gnuplot`. For example, in order to view a figure like Figure 8 on the screen the user can simply run `gnuplot` and execute the command:

```
load "datafile.aest.gp"
```

To aid further exploration, the program can output the average of the trial $\hat{\alpha}$ values at each level of aggregation. The program can be also compiled in a form suitable for dynamic loading into the `Splus` statistical analysis package. This allows it to be used in conjunction with the other data analysis methods present in `Splus`.

3 Empirical Evaluation

This section presents the results of applying the algorithm shown in Figure 4 to a variety of synthetic datasets.

Table 1 shows a summary of the results of applying the algorithm to datasets drawn from a variety of heavy-tailed distributions: the Pareto distribution (defined in Section 2.2) with $k = 1$ and $\alpha = 0.7$, 0.9, 1.1, 1.5, and 1.8; and the symmetric α-Stable distribution with median zero, and $\alpha = 0.7$, 0.9, 1.1, 1.5, and 1.8 (obtained using the `rstab()` function in Splus).

The estimator was applied to datasets of length 1,000, 10,000, and 100,000. In each case the estimator was applied to 250 different datasets. The table shows under "% Estimates" the percent of times in each case that the estimator returned an estimate, and for those cases in which it did so, it shows the sample mean and sample standard deviation of the α estimates returned.

The table shows a distinction between stable and non-stable distributions. For stable distributions the estimator is fairly accurate over a wide range of α values and sample sizes. However, for the non-stable Pareto distributions, there are two trends: the accuracy of the estimator increases as datasets grow larger, and the accuracy of the estimator increases as α grows smaller. For the Pareto datasets, it appears that when α is small (close to or below 1), the estimator is usually fairly accurate except when datasets are quite small. As α approaches 2, the estimator shows some downward bias.

Table 2 shows the performance of the estimator when applied to datasets drawn from a variety of non-heavy-tailed distributions. Again, each row corresponds to the results of 250 trials, and the "% Estimates" column counts the percent of times the estimator returned a value.

The first two sections of Table 2 show the estimator's performance on Normal distributions with unit variance and the exponential distribution with CDF $P[X \leq x] = 1 - e^{-\lambda x}$ for $\lambda = 1$. This shows that finite-variance distributions, which tend to Normal when aggregated, can show scaling behavior with α close to 2.

The next two sections of Table 2 shows the estimator's performance on the Lognormal distribution:

$$X = e^{\sigma Z}$$

where $Z \sim N(\mu, \sigma)$. For these distributions μ (the mean of $\ln X$) was 0 and σ (the standard deviation of $\ln X$) was either 1 or 2. Note that when $\sigma = 2$ the estimator cannot distinguish the asymptotically Normal scaling taking place from heavy-tailed scaling.

The final section of Table 2 shows the estimator's performance on the Weibull distribution with CDF $P[X \leq x] = 1 - \exp(-(x/a))^b$. In these tests $a = 1$ and $b = e^{-1}$.

The performance of the estimator on Pareto datasets is summarized in Figure 9. In these figures, the distribution of differences between the estimated value of α and the true value of α is plotted using boxplots. In these boxplots, the central line shows the median value; the surrounding box shows the limits of the middle half of the data; and the whiskers show the full data range.

The figure shows that for large datasets drawn from Pareto distributions with small values of α, the estimator is usually accurate. For example, in the α= 0.7 case, the estimator always returns a value within 0.05 of the true value more than 50% of the time, as indicated by the range of the

Distribution	α	Samples	% Estimates	$\mu_{\hat{\alpha}}$	$\sigma_{\hat{\alpha}}$
Pareto	0.7	100,000	100	0.711	0.059
Pareto	0.7	10,000	98.8	0.765	0.157
Pareto	0.7	1,000	92	0.820	0.287
Pareto	0.9	100,000	100	0.911	0.052
Pareto	0.9	10,000	100	0.960	0.109
Pareto	0.9	1,000	94.8	1.014	0.251
Pareto	1.1	100,000	100	1.086	0.041
Pareto	1.1	10,000	100	1.121	0.112
Pareto	1.1	1,000	98.4	1.153	0.253
Pareto	1.5	100,000	100	1.380	0.037
Pareto	1.5	10,000	100	1.398	0.107
Pareto	1.5	1,000	100	1.344	0.212
Pareto	1.8	100,000	100	1.560	0.039
Pareto	1.8	10,000	100	1.561	0.102
Pareto	1.8	1,000	100	1.510	0.252
Symm-α-stable	0.7	100,000	99.2	0.761	0.136
Symm-α-stable	0.7	10,000	94.7	0.839	0.247
Symm-α-stable	0.7	1,000	82.9	0.860	0.352
Symm-α-stable	0.9	100,000	100	0.878	0.217
Symm-α-stable	0.9	10,000	98.4	0.959	0.150
Symm-α-stable	0.9	1,000	100	0.878	0.217
Symm-α-stable	1.1	100,000	100	0.954	0.361
Symm-α-stable	1.1	10,000	100	1.141	0.101
Symm-α-stable	1.1	1,000	96.8	1.225	0.292
Symm-α-stable	1.5	100,000	100	1.506	0.018
Symm-α-stable	1.5	10,000	100	1.524	0.055
Symm-α-stable	1.5	1,000	98.4	1.583	0.232
Symm-α-stable	1.8	100,000	100	1.804	0.026
Symm-α-stable	1.8	10,000	100	1.809	0.072
Symm-α-stable	1.8	1,000	99.2	1.863	0.233

Table 1: Summary of Performance of α-estimator on heavy-tailed distributions (250 trials each case).

Distribution	Samples	% Estimates	$\mu_{\hat{\alpha}}$	$\sigma_{\hat{\alpha}}$
Normal $\sigma = 1$	100,000	100	1.998	0.023
Normal $\sigma = 1$	10,000	100	1.995	0.076
Normal $\sigma = 1$	1,000	98.8	2.009	0.262
Exponential $\lambda = 1$	100,000	100	2.328	0.061
Exponential $\lambda = 1$	10,000	100	2.330	0.171
Exponential $\lambda = 1$	1,000	99.2	2.214	0.314
Lognormal $\sigma = 1$	100,000	100	2.078	0.048
Lognormal $\sigma = 1$	10,000	100	2.026	0.130
Lognormal $\sigma = 1$	1,000	99.2	1.876	0.273
Lognormal $\sigma = 2$	100,000	100	1.384	0.031
Lognormal $\sigma = 2$	10,000	100	1.381	0.092
Lognormal $\sigma = 2$	1,000	99.6	1.336	0.220
Weibull $b = e^{-1}$	100,000	100	2.097	0.054
Weibull $b = e^{-1}$	10,000	100	2.023	0.128
Weibull $b = e^{-1}$	1,000	100	1.785	0.267

Table 2: Summary of Performance of α-estimator on non-heavy-tailed distributions (250 trials each case).

box. As α increases for large datasets, the estimator's variance decreases slightly, but it begins to show significant bias for $\alpha \geq 1.5$.

Figure 10 shows the corresponding results obtained when applying the estimator to datasets drawn from Symmetric α-Stable distributions. This figure shows that the estimator has better performance on Stable distributions than on Pareto, which is to be expected. The figure shows that when the underlying dataset is drawn from a Stable distribution, the performance of the estimator is relatively unaffected by the particular values of α, except for a slight decrease in variance as α grows. In general, for datasets of size 100,000, the estimator nearly always returns a value within 0.1 of the true value, regardless of the particular value of α.

An important feature of the estimator evident from Figures 10, 9, and Table 1 is that both bias and variance decrease with increasing sample size. This feature makes it especially attractive for use on datasets taken from computing and telecommunications systems, where large sample sizes are common.

We also compare the effectiveness of scaling estimator to that of a commonly used alternative: the Hill estimator (as defined in Section 1). Using the Hill estimator requires the specification of k; the Hill estimator uses the k largest values in the dataset in forming its estimate. We consider three possible settings of k: either 0.01, 0.05, or 0.1 times the number of data points in the sample. These values correspond to applying the Hill estimator to the upper 1%, 5%, or 10% tail of the dataset.

In Figure 11 we show the performance of the Hill estimator side by side with that of the scaling estimator. The values for the scaling estimator are the same as shown in Figures 9 and 10 and are repeated for reference. In these plots the results for the scaling estimator are denoted by "S", and results for the Hill estimator using 1%, 5%, or 10% of the upper tail are denoted by "H1," "H5," and "H10" respectively.

The plots on the left side of the figure compare the two estimators for Pareto datasets. These plots show that the variance of the scaling estimator is larger than any version of the Hill estimator,

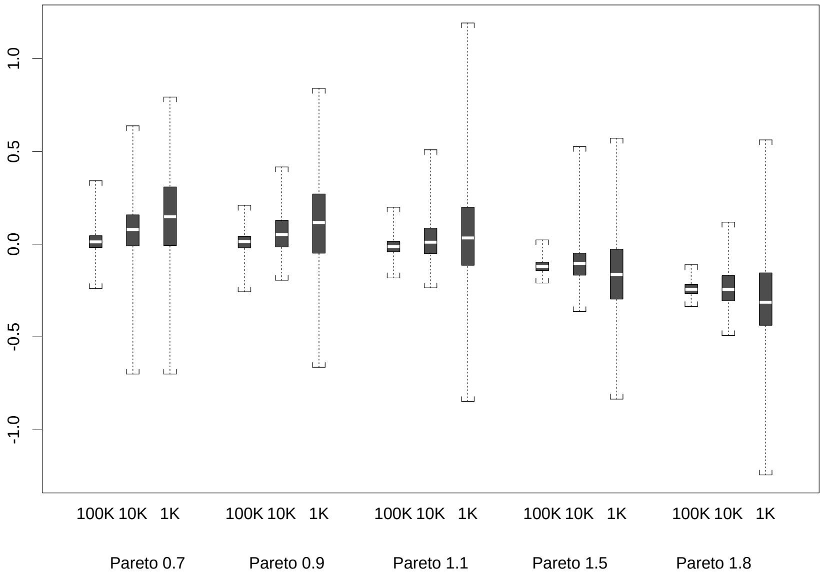


Figure 9: Distribution of $\hat{\alpha} - \alpha$ for Pareto datasets.

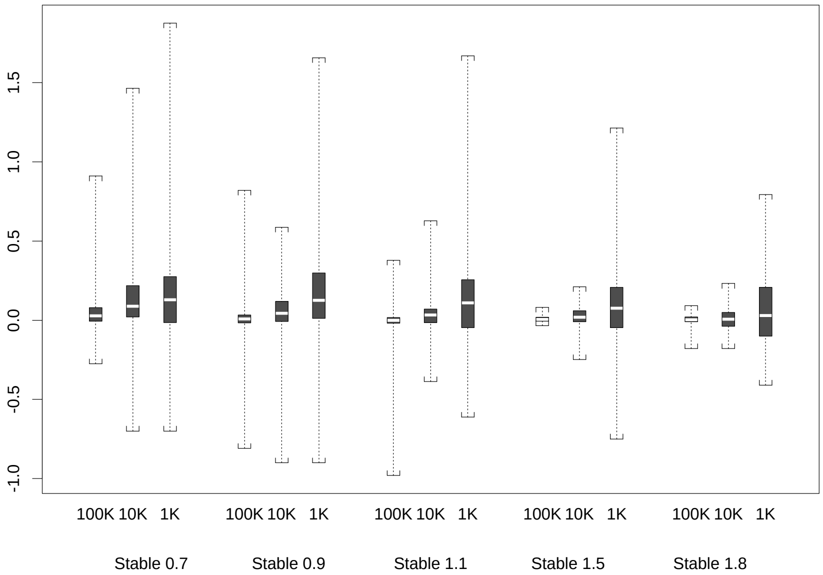
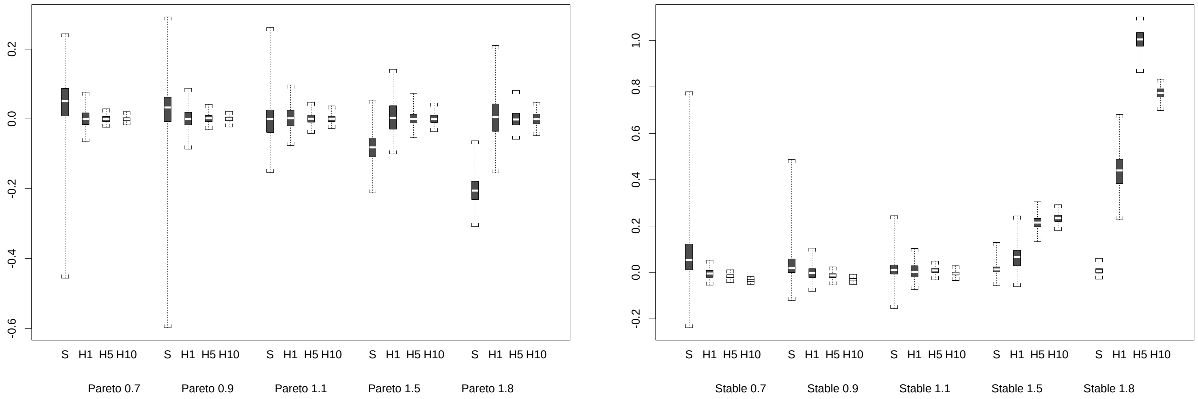
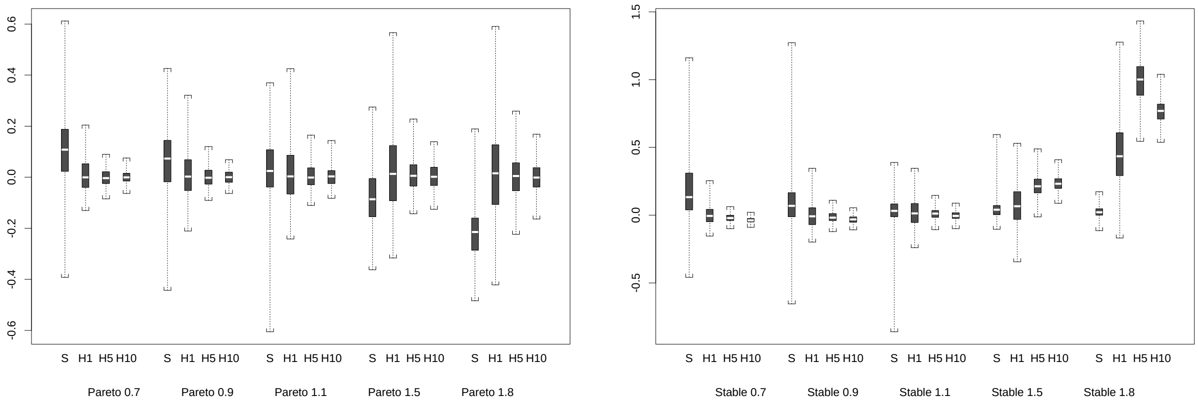


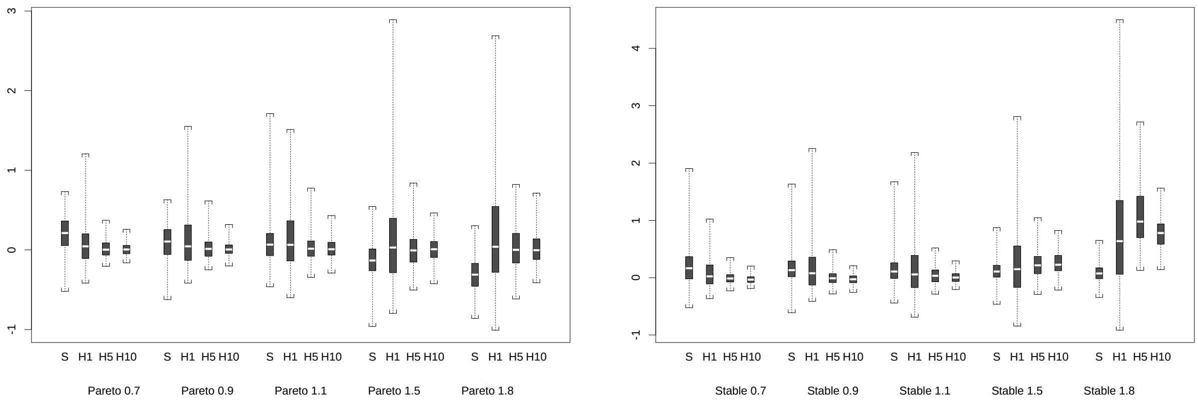
Figure 10: Distribution of $\hat{\alpha} - \alpha$ for α-Stable datasets.



100,000 Points



10,000 Points



1,000 Points

Figure 11: Comparison of Scaling and Hill Estimators; left: Pareto datasets; right: α-Stable datasets. The vertical axis corresponds to $\hat{\alpha} - \alpha$.

and that the scaling estimator exhibits some bias for large α that the Hill estimator does not show. These results are not surprising; since the Pareto distribution has the same power-law shape over its entire range, the Hill estimator yields a consistent value for nearly any choice of k. For such distributions, identifying the α value of the tail is not difficult in general.

The plots on the right side of the figure compare the two estimators for α-Stable distributions; these plots show different results. They show that the Hill estimator, for any of the fixed k values used, becomes quite inaccurate for distributions with large α values. In contrast, the scaling estimator remains accurate for large α. This effect occurs because the body of the α-Stable distribution does not have a power-law shape, and the Hill estimator's output becomes influenced by the shape of the distributional body.

4 Examples

To give an example of the utility of the scaling estimator in practice, we show its use on two different datasets.

First, we return to the example dataset introduced in Section 1. This dataset consists of the sizes of 130,140 files transferred over the World Wide Web. The CD, Hill, and scaling plots for this dataset are shown in Figure 12 (the CD plot shown is the same as Figure 1; it is repeated here for reference).

As noted in Section 1, estimating α from the CD plot is complicated by the need to select the proper x_0. Furthermore, the figure shows that the Hill estimator shows variability over the range 1.0 to 1.3; it does not consistently settle at any particular value. The scaling estimator gives an estimate $\hat{\alpha} = 1.08$, and shows that scaling behavior is clearly present in approximately the 1% upper tail.

A contrasting example is shown in Figure 13. This dataset is a record of about 4 weeks of data transfer activity of the Unitree Mass Storage System at NASA Goddard Space Flight Center, Greenbelt, MD. This system stores datasets used and generated by large-scale scientific computations. Each data point is the size of one retrieval or storage of an entire file. Normal traffic was about 5,000 transfers a day; this dataset contains 148,852 data points. The CD, Hill, and scaling plots for this dataset are shown in Figure 13.

Figure 13 shows that for this dataset, there is little evidence of heavy-tailed behavior. The CD plot shows that the upper tail appears linear, but the slope method yields $\hat{\alpha} = 2.79$ — which indicates a finite variance condition. The results from the Hill estimator are consistent with a finite variance conclusion; its estimates are very erratic, generally in the neighborhood of 2.0. The scaling estimator returns a value of $\hat{\alpha} = 1.4$ and shows that the only significant scaling is taking place in a region far from the extreme tail. We can, moreover, exclude the possibility that we are dealing with a Stable distribution with $\alpha = 1.4$. Such a distribution with 148,852 points would yield a very different type of scaling plot, as shown in Figure 14.

5 Conclusion

In this paper we have described a new method for addressing a problem that arises in characterizing empirical datasets: forming an estimate of the heavy tail index α. It has the advantage of being nonparametric, *i.e.,* it is not necessary to specify the form of the underlying distribution. Since the method relies on the scaling of sums, it measures a property that is often one of the most important effects of heavy-tailed behavior. Most importantly, the method increases in accuracy as the size

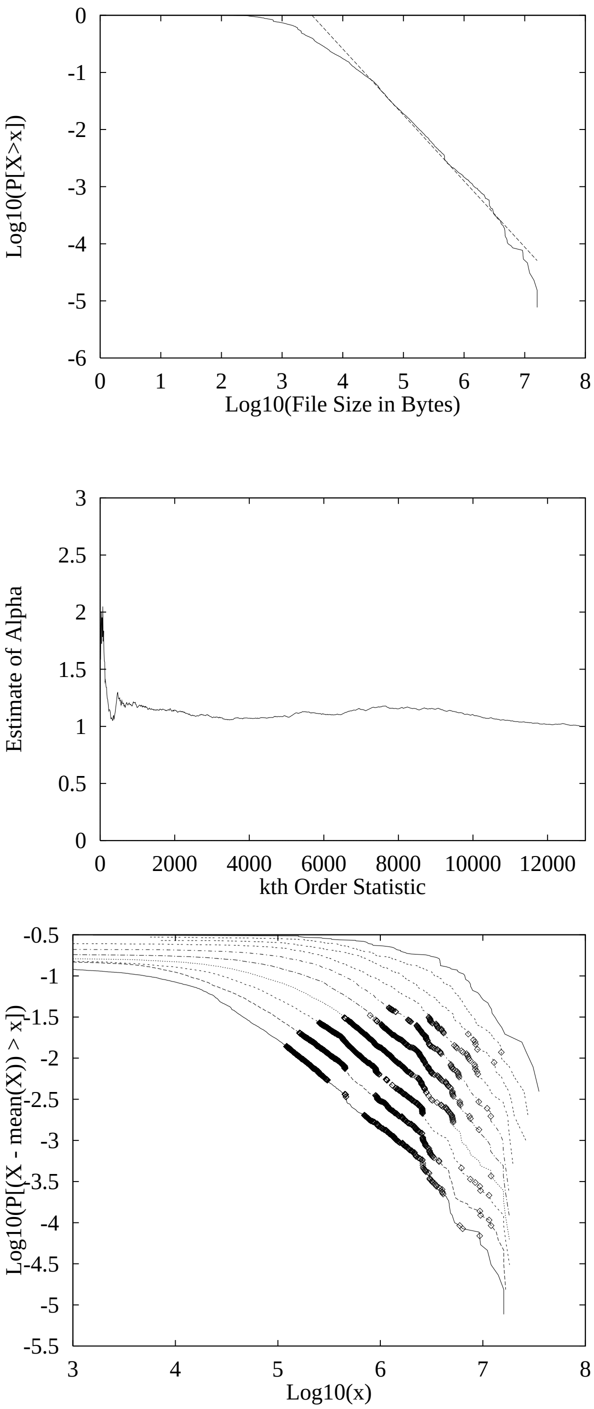


Figure 12: CD, Hill, and Scaling Plots for Web files transferred over network.

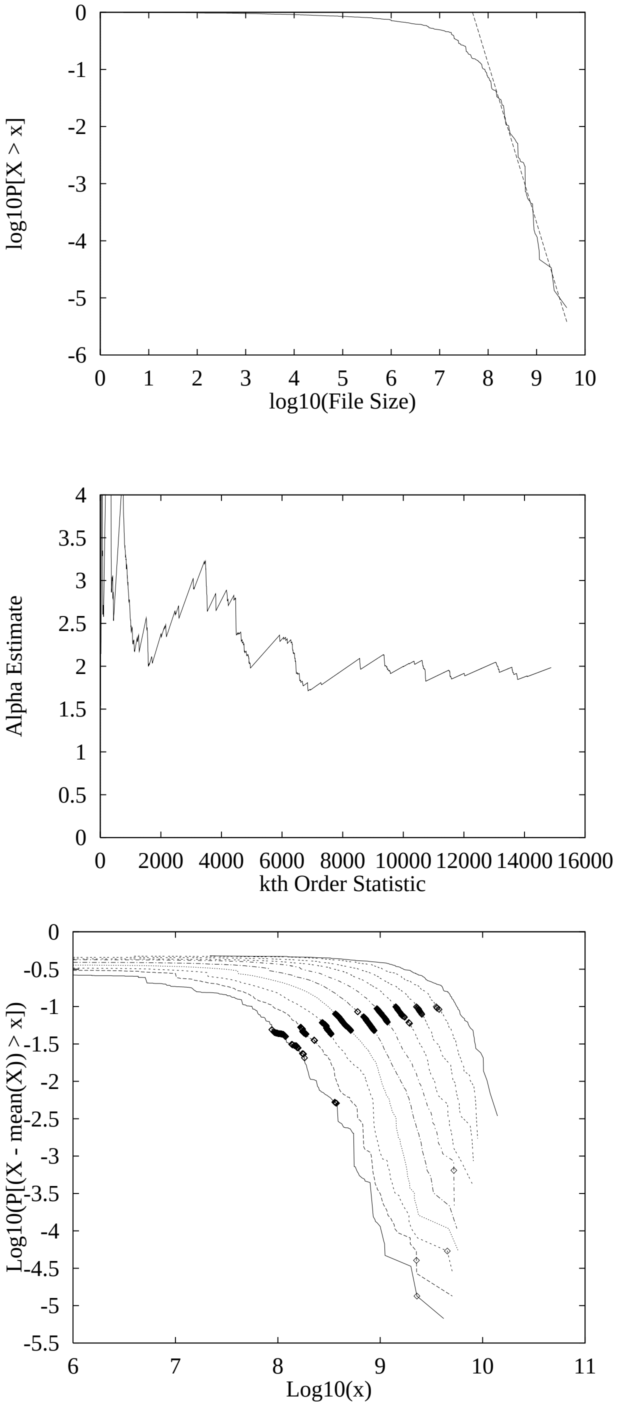


Figure 13: Slope, Hill, and Scaling plots for Unitree Transfers.

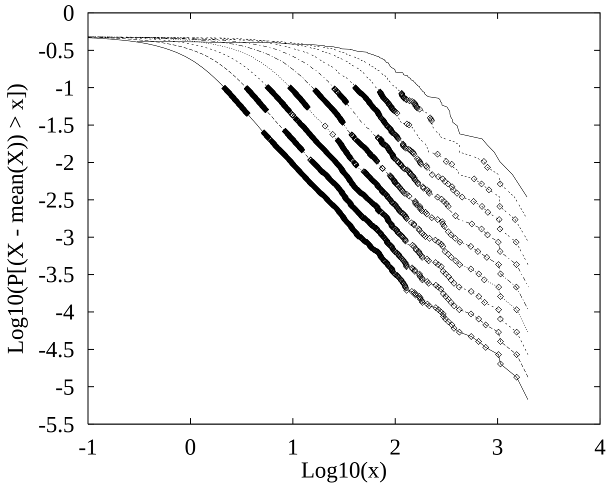


Figure 14: Scaling plots for Symmetric α Stable with α= 1.4.

of the dataset grows, meaning that it is particularly suited for large datasets, as are increasingly encountered in measurements of telecommunications and computing systems.

This paper presents a general method for estimating α based on scaling properties of a dataset. In addition we have described and evaluated a particular algorithm that uses this method in practice. However, we have not attempted to show that the algorithm we use is the best possible embodiment of the use of the scaling method for α estimation. A theoretical study of the variance and bias of our algorithm, and of the scaling method in general, is still required.

Clearly, parametric methods can be more accurate than the scaling estimator, which is to be expected since they use more information about the dataset. For example, if it is known that the underlying distribution is α-Stable, one could use α estimation methods based on dataset percentiles; likewise if it is known that the underlying distribution is Pareto, one could use the slope of the tail on a log-log plot. In contrast, the scaling method, being nonparametric, has the advantage of working for any heavy-tailed distribution.

Because this method assumes that observations are independent, its accuracy is affected by any correlations that may exist among the X_i's. For this reason we suggest that the estimator should be applied a number of times to different random permutations of the dataset, in an attempt to disrupt any correlation structure present.

The estimator is available as C source code from the authors.[2] It can produce graphical output as an aid to interpretation. It also can be compiled in a form suitable for dynamic loading into the `Splus` statistical analysis package.

References

[1] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

[2] Mark E. Crovella, Murad S. Taqqu, and Azer Bestavros. Heavy-tailed probability distributions in the World Wide Web. In *A Practical Guide To Heavy Tails*, chapter 1, pages 3–25. Chapman & Hall, New York, 1998.

[2] Source code can be obtained from `http://www.cs.bu.edu/faculty/crovella/aest.html`

[3] William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley and Sons, second edition, 1971.

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

[5] Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes.* Stochastic Modeling. Chapman and Hall, New York, 1994.

[6] V.M. Zolotarev. *One-dimensional Stable Distributions*, volume 65 of *Translations of mathematical monographs.* American Mathematical Society, 1986.