A Diagnostic Plot for Estimating the Tail Index of a Distribution

Bruno DE SOUSA and George MICHAILIDIS

The problem of estimating the tail index in heavy-tailed distributions is very important in many applications. We propose a new graphical method that deals with this problem by selecting an appropriate number of upper order statistics. We also investigate the method's theoretical properties are investigated. Several real datasets are analyzed using this new procedure and a simulation study is carried out to examine its performance in small, moderate and large samples. The results suggest that the new procedure overcomes many of the shortcomings present in some of the most common techniques—for example, the Hill and Zipf plots—used in the estimation of the tail index, and it performs very competitively when compared with other adaptive threshold procedures based on the asymptotic mean squared error of the Hill estimator.

Key Words: Heavy-tailed distributions; Sum plot; Tail index.

1. INTRODUCTION

There has been a lot of interest over the last few years for the problem of estimating the tail index (also known as the tail exponent) of a heavy-tailed distribution. This renewed interest stems from new applications where heavy-tailed distributions are present, such as in computer science and telecommunications (Adler, Feldman, and Taqqu 1998; Resnick 1997; Chen et al. 2002), finance and economics (Adler, Feldman, and Taqqu 1998; Jansen and de Vries 1991), and insurance (see Adler et al. 1998). Many approaches have been proposed over the years, including those of Zipf (1949), Hill (1975), Csörgő, Deheuvels, and Mason (1985) and more recently of Kratz and Resnick (1996), Beirlant, Vynckier, and Teugels (1996), Feuerverger and Hall (1999), and Crovella and Taqqu (1999), to name just a few.

The setting of the problem is as follows: the distribution F of the random variable of

Bruno de Sousa is Post Doctoral Fellow, Department of Statistics, The University of Toronto, 100 St. George Street, Room 6018, Toronto ON M5S 3G3 (E-mail: desousa@utstat.toronto.edu). George Michailidis is Assistant Professor, Department of Statistics, The University of Michigan, Ann Arbor, MI 48109-1092 (E-mail: gmichail@umich.edu).

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 1–22
DOI: 10.1198/106186004X12335

interest X satisfies

$$1 - F(x) \sim cx^{-\alpha}, \quad \text{as} \quad x \to \infty, \tag{1.1}$$

where c is a positive constant, $\alpha > 0$ corresponds to the tail index, and the symbol $\sim$ indicates that the limit of the two functions is 1 as $x \to \infty$. The random variable X is said to have a heavy-tailed distribution. The estimation of the parameter α based on a random sample of size n $\{X_i\}_{i=1}^n$ from F constitutes a basic statistical problem.

Although many of the methods for estimating the tail index over the years exhibit optimal asymptotic properties (i.e., consistency), their performance in finite samples is a different issue. The majority of them rely on plotting the statistic of interest against the number of the sample upper order statistics and then inferring an appropriate value of the tail index α from properties of the resulting graph. We discuss next some of the graphing techniques and their rationale.

Let $X^{(1)} > \cdots > X^{(n)}$ denote the order statistics of a random sample coming from a distribution F that satisfies (1.1), that is, for a large enough value of x the tail of the distribution behaves as the tail of a Pareto distributed random variable. The most popular and frequently used approaches in practice are based on the Zipf plot (Zipf 1949) [and its variation, the QQ-estimator (Kratz and Resnick 1996)], the Hill estimator (Hill 1975), and the CD plot (log-log complementary distribution plot), the latter being very popular among engineers.

In the Zipf plot, the quantity $\left\{\log \frac{n+1}{i}\right\}$ is plotted against $\{\log X^{(i)}\}$ for $i = 1, \ldots, n$, and the estimate of the parameter of interest α^{-1} is given by the least squares estimate of the slope for the part of the plot that exhibits a linear behavior (Kratz and Resnick 1996).

The Hill estimator is given by

$$H_{k,n} = \left\{ k^{-1} \sum_{i=1}^{k} \left(\log X^{(i)} - \log X^{(k+1)} \right) \right\}^{-1}, \quad \text{for} \quad 1 \le k < n, \tag{1.2}$$

and the Hill plot is based on graphing $H_{k,n}$ against k. The value of α is inferred by identifying a stable region in the graph.

Finally, in the CD plot, $\log(x)$ is graphed against $\log \bar{F}(x)$, where $\bar{F}(x) = P(X > x)$, and an estimate of α is obtained by estimating the slope of the linear part of the CD plot, since distributions satisfying (1.1) have the property that $\frac{d \log \bar{F}(x)}{d \log(x)} \sim -\alpha$, for large x. If $F(x)$ is estimated by its empirical counterpart $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)$, where $\mathbf{1}(.)$ denotes the indicator function, then plotting $\log \bar{F}_n\left(X^{(i)}\right) = -\log \frac{n}{i-1}$ against $\log X^{(i)}$, gives almost identical results to the Zipf plot. Hence, in this study only the Zipf plot results will be shown.

In Figures 1 and 2 the Hill plot and the Zipf plot of a random sample of size 5,000 drawn from a Pareto distribution and an Inverted Gamma distribution with $\alpha = 1.5$ are presented. The Pareto distribution represents the ideal case, since it satisfies (1.1) throughout its domain, while the Inverted Gamma only in its tail. It can be seen that for the ideal case of our setting, to a large extent both plots perform satisfactorily allowing the data analyst to identify correctly the underlying value of the tail index. However, this is not always the case

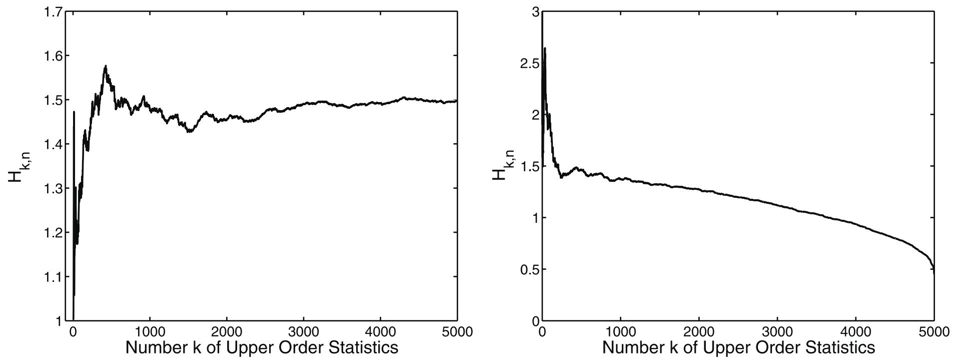

Figure 1. The Hill plot for the Pareto (left panel) and the Inverted Gamma (right panel) distributions.

as the graph of the so-called "horror" Hill plot (see Resnick 1997, p. 1,818) suggests. It should also be noted that the high variability in the right region (the one determined by the largest order statistics) of the Zipf plot is not a welcome feature, since it makes more difficult the proper selection of the number of upper order statistics involved in the estimation of the tail index. An important question that often arises in practice is whether one should ignore those observations, thus ignoring useful information about the behavior of the tail, or include them and get a biased estimate of α.

On the other hand, the situation is not clear at all for the Inverted Gamma distribution. The Hill plot does not quite stabilize throughout its range and therefore it is hard to come up with a value for α. For the Zipf plot the decision to exclude the largest 34 observations changes the estimate of α from 1.944 to 1.371. A nice feature of the Zipf plot (CD plot) is that the lack of linearity at the left part of the graph suggests a departure from a Pareto tail behavior. Based on the Zipf plot, Kratz and Resnick (1996) introduced the QQ-plot as an alternative to the Hill plot. Although the QQ-plot tends to be smoother than the Hill plot, problems can still occur when departures from the Pareto distribution occur. Such an example can be observed in the left hand-side of the Figure 3 where no value for α

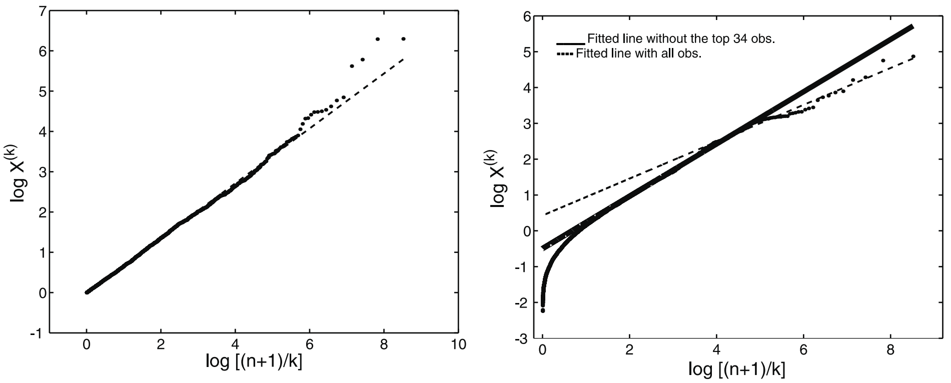

Figure 2. The Zipf plot for the Pareto (left panel) and the Inverted Gamma (right panel) distributions.

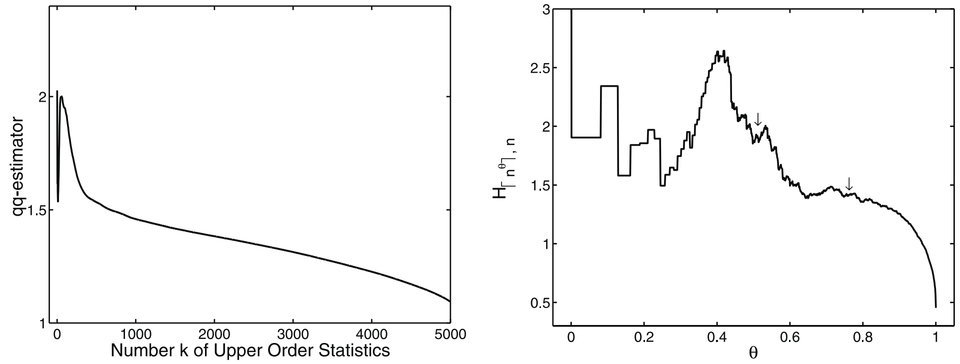

Figure 3. The QQ-plot (left panel) and the alternative Hill plot (right panel) for a sample of size 5,000 drawn from an Inverted Gamma distribution with tail index 1.5

is suggested from the QQ-plot, since no stable region can be inferred from it. The nice theoretical properties of the Hill estimator—consistency (Deheuvels, Haeusler, and Mason 1988; Mason 1982) and asymptotic normality (Hall 1982)—have led researchers (Drees, de Haan, and Resnick 2000; Resnick and Stărică 1997) to consider graphs closer to the Hill plot, such as the alternative Hill plot (Resnick and Stărică 1997), that overcomes some of the difficulties previously mentioned. Some of these modifications work well for the Pareto case, but some still present problems for data drawn from other heavy-tailed distributions, as Figure 3 shows (right panel). The graph suggests two possible stable regions corresponding to values of $\alpha = 1.9$ and 1.4, respectively.

The previous discussion indicates that the volatility of the Zipf plot especially in the region of the largest order statistics, and the lack of stability that the Hill plot and its alternatives exhibit throughout their range, make the correct identification of the tail index a rather challenging task in finite samples. We propose next an alternative plot that overcomes some of these difficulties and examine its theoretical properties as well as its performance in practical situations and through an extensive simulation study.

The objectives and structure of the article are as follows: at the *methodological* level, it develops a graphical procedure, called the *Sum plot*, that effectively determines the number of upper order statistics where the Pareto tail behavior occurs. At the technical level, it establishes several theoretical properties of the graph that prove useful when leveraged in data analytic situations. In Section 2 the new plot is introduced and its properties discussed. The results of the simulation study are given in Section 3, while its application to several real datasets is presented in Section 4. Some final remarks are followed in Section 5.

2. THE SUM PLOT

Let the random variables S_k, for $k = 1, \ldots, n$ be defined as

$$S_k = \sum_{i=1}^{k} iV_i = \sum_{i=1}^{k} i \log \frac{X^{(i)}}{X^{(i+1)}}, \tag{2.1}$$

where the order statistics $X^{(i)}$'s come from a random sample $(X_1, \ldots, X_n)$.

It is shown next that for the value of k such that for all $x \geq X^{(k+1)}$, $1 - F(x) = cx^{-\alpha}$, with $c > 0$ and $\alpha > 0$, the S_k's are Gamma distributed with parameters (k, α). Hence, $\mathbf{E}(S_k) = k\alpha^{-1}$ and by plotting S_k against k, it is expected that the resulting graph should be linear in the corresponding region of the upper order statistics where relationship (1.1) holds. The tail index can then be estimated by the inverse of the slope of the linear part of such a graph. We call this graph the *Sum plot*.

This defining property of the Sum plot is based on the following derivation. Let the random variables $Y_i = X_i^{-1}$, for $i = 1, \ldots, n$. Then $F_Y(y) = P[Y_i \leq y] = P[X_i \geq y^{-1}] = cy^{\,\alpha}$, for $y < (X^{(k+1)})^{-1} = Y^{(n-k)}$. From Rényi's representation (Rényi 1953), $Y^{(n-i+1)}$ can be expressed as

$$Y^{(n-i+1)} = F_Y^{-1}\left[\exp\left\{-\sum_{j=1}^{n-i+1} \frac{E_{n-j+1}}{n-j+1}\right\}\right] \quad \text{for} \quad i = 1, \ldots, n, \tag{2.2}$$

where the E_i's are independent exponential random variables with unit mean.

Because $-\log F_Y(Y^{(n-i+1)}) = \sum_{j=1}^{n-i+1} \frac{E_{n-i+1}}{n-i+1}$, and conditioning on the value of k such that $1 - F(x) = cx^{-\alpha}$ for $c > 0$, $\alpha > 0$ and $x \geq X^{(k+1)}$, the random variables E_i can be written as follows:

$$\begin{aligned} E_i &= i\left[\log F_Y(Y^{(n-i)}) - \log F_Y(Y^{(n-i+1)})\right] \\ &= i\left[\log \frac{c(Y^{(n-i)})^\alpha}{c(Y^{(n-i+1)})^\alpha}\right] = \alpha i \log \frac{X^{(i)}}{X^{(i+1)}}, \end{aligned}$$

for $i = 1, \ldots, k$. Therefore, $\alpha i V_i = E_i$, $i = 1, \ldots, k$. Hence, the random variables iV_i are independent exponential distributed with mean equals to α^{-1} and, consequently, the random variables S_k are Gamma distributed with parameters (k, α).

2.1 EXAMPLES

Some examples of the *Sum plot* are shown in Figures 4 and 5 for a sample of size 5,000 drawn from a Pareto and an Inverted Gamma distribution, and a sample of size 10,000 generated from a symmetric α-Stable distribution, all with tail index $\alpha = 1.5$.

The two most striking features of the Sum plot are, first, the stability of the generated pattern, and second, its linearity for the Pareto distribution (top panel). This strongly linear pattern suggests that *all* the observations in the sample should be used for estimating the tail index. Furthermore, the strong linear pattern in the left part of the Sum plot (the region that corresponds to the largest order statistics, i.e., the one that captures the behavior of the tail of the distribution) exhibited in the other two plots (middle and bottom panels), suggests that a procedure that identifies where the linear behavior stops effectively, identifies the value of k from which the assumption of the Pareto tail behavior is violated.

In Figure 5 the results of this strategy are shown. For comparison purposes, the Hill and Zipf plots for an appropriately chosen number of upper order statistics are also given.

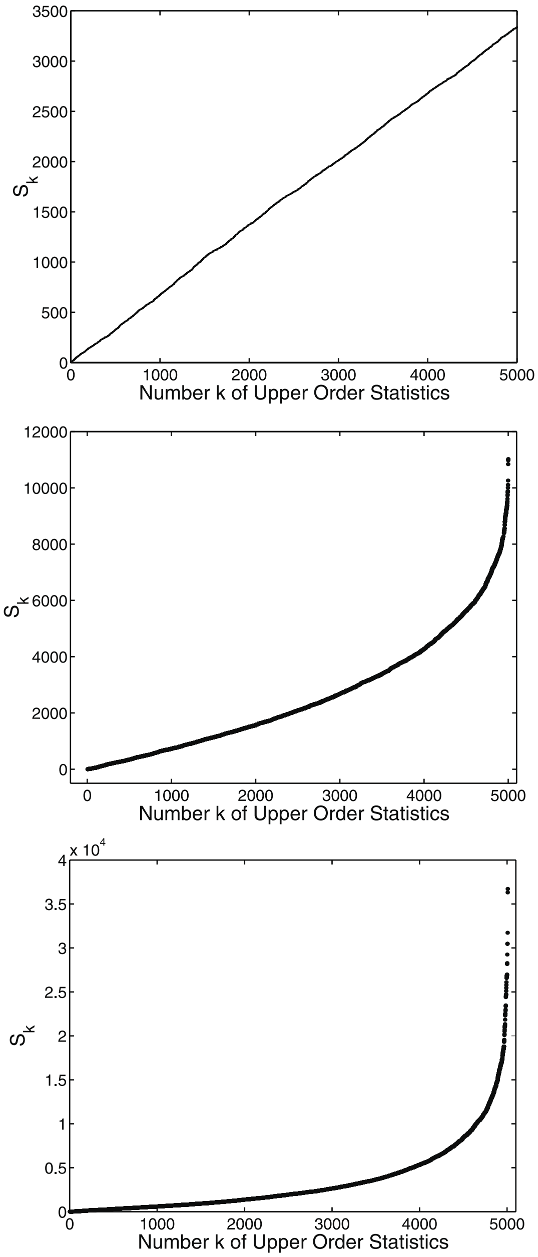

Figure 4. The Sum plot for a Pareto (top), an Inverted Gamma (middle), and an α-Stable (bottom) distribution.

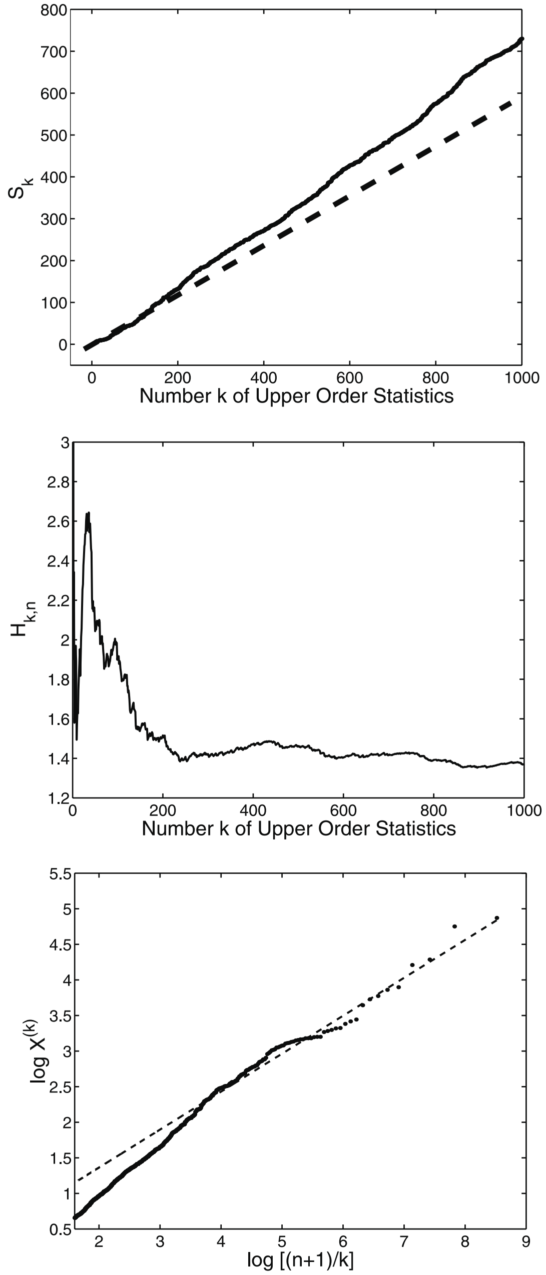

Figure 5. The Sum plot (top), the Hill plot (middle), and the Zipf plot (bottom) for an Inverted Gamma distribution focusing only on the top 1,000 observations.

Notice that the stability of the Sum plot facilitates spotting departures from linearity. An algorithm that formalizes how to detect such departures is given in Section 2.3.

For this example, the Sum plot suggests $k = 160$. Then, one can use this value of k in the Hill estimator (which is the conditional maximum likelihood one) and estimate a value for the tail index by $H_{160,5000} = 1.579$. The Hill plot does not quite stabilize in any particular region. The erratic behavior of the largest order statistics shown in the right part of the Zipf plot suggest a value for $k = 34$, with an estimate for $\hat{\alpha} = 1.944$ based on the QQ-estimator.

It can be seen that the Sum plot overcomes the difficulty of identifying a stable region in the Hill plot, and the unstable behavior of the Zipf plot in the area of the largest observations. We show next that the observed linearity is an inherent property of the Sum plot, which helps the data analyst in identifying departures from a Pareto tail behavior. We also discuss a connection between the estimate of α obtained from the Sum plot and the Hill estimator.

2.2 Properties of the Sum Plot

Proposition 1. *Consider a random variable X that satisfies $1 - F(x) = c^{\alpha}x^{-\alpha}$, for $x > c, \alpha > 0$ and $c > 0$. Then the random variables $\{S_k, k = 1, \ldots, n-1\}$ defined in (2.1) are stochastically increasing linear (SIL).*

The proof of this Proposition, which is based on stochastic convexity results, is given in the electronic version at www.ingenta.com.

The next result establishes a connection between the generalized least squares estimator of the slope of the Sum plot under (1.1) and the Hill estimator.

The fact the $\mathbf{E}(S_k) = k\alpha^{-1}$ suggests that the parameter of interest α^{-1} can be estimated by the regression coefficient for the slope of the simple linear regression model $S_i = \beta_0 + \beta_1 i + \epsilon_i,\ i = 1, \ldots, k$ for an appropriately chosen k. The covariance matrix of the error terms is given by $\Omega = \alpha^{-2}\left[\min(i,j)\right]_{i,j=1}^{k}$.

This result follows from the following facts: (1) the S_k's are Gamma distributed random variables with parameters (k, α); (2) $S_j = S_i + S^{'}_{j-i}$ for every $i < j$ with $S^{'}_{j-i} = \sum_{k=i+1}^{j} kV_k$ independent of $S_i = \sum_{k=1}^{i} kV_k$; and (3)

$$\begin{aligned}\text{cov}\ \left(1, S_i, S_j\right) &= \mathbf{E}[S_i(S_i + S^{'}_{j-i})] - \mathbf{E}[S_i]E[S_j] = \mathbf{E}(S_i^2) + \mathbf{E}(S_i)\mathbf{E}(S^{'}_{j-i}) - \frac{ij}{\alpha^2} \\ &= \frac{i(i+1)}{\alpha^2} + \frac{i(j-i)}{\alpha^2} - \frac{ij}{\alpha^2} = \frac{i}{\alpha^2}. \qquad (2.3)\end{aligned}$$

Therefore, the generalized least squares estimator is given by $\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}y$, where $y = (S_1, \ldots, S_k)'$, $X = [1\ i]_{i=1}^{k}$ and $\hat{\beta}_{\text{GLS}} = (\hat{\beta}_0,\ \hat{\beta}_1)'$.

Proposition 2. *Suppose that $1 - F(x) = cx^{-\alpha}$ for $x \geq X^{(k+1)}$ for some $1 \leq k < n$, and consider the regression model $y = X\beta + \varepsilon$, with y, X and β defined as above. Then, the parameter α^{-1} can be estimated by the GLS estimate of the slope, $\hat{\beta}_1$, and can be written* *as*

$$\hat{\alpha}^{-1} = \hat{\beta}_1 = \frac{k}{k-1} H_{k,n}^{-1} - \frac{1}{k-1} \log X^{(1)}, \tag{2.4}$$

where $H_{k,n}$ is the Hill estimator of α.

The proof of this proposition is given in the electronic version at www.ingenta.com.

Proposition 2 establishes the connection between an estimator based on the Sum plot and the conditional maximum likelihood Hill estimator. In case $\beta_0 = 0$ the GLS estimator corresponds to the inverse of the Hill estimator.

Corollary 1. *Assuming the conditions of the proposition above, and if β_0 is zero, then $\hat{\alpha}^{-1} = \hat{\beta}_{GLS} = H_{k,n}^{-1}$, where $H_{k,n}$ is the Hill estimator of α.*

The proof of the corollary is given in the electronic version at www.ingenta.com

The Hill estimator has served as the basis of other statistics for the problem at hand. Some examples include the $H^{(k)}$ and the $K^{(k)}$ statistics proposed in (Hill 1975) and further studied in (Hsieh 1999; de Sousa 2002); however, unlike the Sum plot they exhibit a more erratic behavior in finite samples. Other procedures for selecting the threshold in the estimation of the tail index were proposed, for example, by Matthys and Beirlant (2000), Danielsson, de Haan, Peng, and de Vries (2001), and Guillou and Hall (2001). These procedures are based on bootstrap methods and will be included in a comparison study on the estimation of the tail index in a future project.

The above properties suggest the following strategy for estimating α: (1) use the Sum plot to identify the correct value of k^* and (2) plug the value of k^* in $H_{k,n}$ to obtain the desired estimate. The second step avoids the computationally expensive procedure for large datasets of calculating the GLS estimator. For the first step the following procedure identifies the correct value of k, by detecting departures from linearity.

2.3 Algorithm for Identifying *k* From the Sum Plot

The idea behind the algorithm is to determine where there is a distinct break from linearity. Hence, a sequential testing procedure based on the regression model previously discussed is a natural candidate for this task. The following statistic discussed by McGee and Carleton (1970) tests the hypothesis that a new point y_0 is a point adjacent to the left or to the right of the set of points $y = (y_1, y_2, \ldots, y_k)$

$$F = s^{-2} \left[(y_0 - \hat{y}_0^*)^2 + \sum_{i=1}^{k} (\hat{y}_i - \hat{y}_i^*)^2 \right], \tag{2.5}$$

where the * represents the predictions based on $k+1$ observations, and $s^2 = \frac{y'y - \hat{\beta}X'y}{k-2}$. The null hypothesis is rejected if $F \geq F_{1,\,k-2,\,\alpha}$, where $F_{1,\,k-2,\,\alpha}$ is the $1-\alpha$ percentile of an $F(1, k-2)$ distribution. Experience has shown that the results are very similar for datasets with up to 10,000 observations for the 10%, 5%, and 1% levels of significance.

To find the value of k used in the estimation of the tail index consider the following stepwise procedure:

Algorithm 1.

1. Fit a least-squares regression line to the initial $k = \beta n$ top observations, $y = [y_i]_{i=1}^{k}$.
2. Using the test statistic (2.5), determine whether a new point $y_0 = y_j$ for $j > k$, belongs to the original set of points y ($k \times 1$ matrix). Continue adding points until the hypothesis that y_0 is a point adjacent to the set of points y is rejected.
3. Set $k_{\text{new}} = \max\left(0, \{j : F < F_{1,k-2,\alpha}\}\right)$. If $k_{\text{new}} > 0$, return to Step 1 with this new value for k. If $k_{\text{new}} = 0$, that is, no new points are added to the set of points y, go to Step 4.
4. Estimate the tail index using the k top observations of the dataset.

A good choice for the value of the proportion β of points used in the first step of the algorithm is .02 for datasets with up to 10,000 observations. Some experience with larger sets consisting of 100,000 points suggest that the value has to be adjusted to $\beta = .002$.

Other approaches of the use of the Sum plot in the selection of k were discussed by de Sousa (2002). The results of these procedures are to a large extent similar to the ones presented here.

3. A SIMULATION STUDY

This section presents the results of the simulation study where the observations are drawn from a Pareto, an Inverted Gamma, and a α-Stable distribution. Small ($n = 200$), moderate ($n = 1,000$), and large (5,000) sample sizes are considered, and the values of the tail index examined are $\alpha = 1.1,\ 1.5$, and 1.9. For all possible combinations, 100 samples were generated and the results were compared in terms of the standard deviation (STD), the bias (BIAS), and the mean squared error (MSE). Although not presented here, similar results were obtained for values of α equal to 1.3 and 1.7.

The techniques used to estimate the tail index are briefly described next. For the Zipf plot we have $X = \left[1 \;\; \log \frac{n+1}{i}\right]_{i=1}^{k}$ and $y = [\log X^{(i)}]_{i=1}^{k}$ in Algorithm 1. The tail index is estimated by the inverse of the least squares estimator for the slope of the linear part in the graph. This estimator was proposed by Kratz and Resnick (1996) and is known as the QQ-estimator.

Recall from the previous section that the covariance matrix of the error terms when $y_i = S_i$ is given by the matrix $\Omega = \alpha^{-2}[\min(i,j)]_{i,j=1}^{k}$. Using the Cholesky decomposition, we have seen that the matrix $[\min(i,j)]_{i,j=1}^{k}$ could be written as $\Theta\Theta'$ (see proof of Proposition 2, available in the electronic version at www.ingenta.com). Hence, cov $\left(1, \Theta^{-1}\epsilon\right) = \Theta^{-1}\text{cov}\,(1,\epsilon)\,(\Theta')^{-1} = \alpha^{-2}\Theta^{-1}\Omega(\Theta')^{-1} = \alpha^{-2}I$. Therefore for the Sum plot, Algorithm 1 will be applied to $\Theta^{-1}y$ with $y = [S_i]_{i=1}^{k}$, and $\Theta^{-1}X$ with $X = [1\ i]_{i=1}^{k}$, and the tail index will be estimated by the Hill estimator.

In analyzing the above approaches, it was considered that at least 2% of the upper order statistics should be used in the estimation of the tail index. This assumption seems quite reasonable, not only for the Inverted Gamma and the Stable distributions studied here, but also in general for the sample sizes considered in this study. With less than 2% of the data, any of the methods considered exhibit large variability and it becomes hard to justify a particular choice for $\hat{\alpha}$. The level of significance used in the stepwise procedure described in Section 2.3 was 1%, 5%, and 10%. Since the results obtained were very similar for these three cases, only the 5% significance level case is reported.

Two other adaptive procedures are also included in the simulation study. These procedures are based on the asymptotic mean squared error of the Hill estimator. Under certain conditions on model (1.1), Feuerverger and Hall (1999), and Beirlant, Dierckx, and Stărică (2002) determined that

$$i\left(\log X^{(i)} - \log X^{(i+1)}\right) \approx \left(\alpha^{-1} + b_{n,k}\left(\frac{i}{k+1}\right)^{-\rho}\right) e_i, \quad \text{for} \quad i = 1, \ldots, k, \tag{3.1}$$

where the random variables e_i's are independent exponential distributed with unit mean, α is the tail index, $b_{n,k} = b\left(\frac{n+1}{k+1}\right)$ for a positive function b such that $b(x) \to 0$ as $x \to \infty$, and $k = 1, \ldots, n-1$.

Therefore, the asymptotic mean squared error of the Hill estimator can be determined by $\text{AMSE}_{H^{-1}_{k,n}} = \left(\frac{b_{n,k}}{1-\rho}\right)^2 + \frac{\alpha^{-2}}{k}$, for $k = 1, \ldots, n-1$.

The optimal value of k is calculated by $k_{\text{op}} = \min_{1 \le k < n} \text{AMSE}_{H^{-1}_{k,n}}$. It can be shown that the $\text{AMSE}_{H^{-1}_{k,n}}$ is minimal for

$$k_{\text{op}} \sim b_{n,k}^{-\frac{2}{1-2\rho}} (k+1)^{-\frac{2\rho}{1-2\rho}} \left(\frac{(1-\rho)^2 \alpha^{-2}}{-2\rho}\right)^{\frac{1}{1-2\rho}}.$$

Algorithms 2 and 3 (Beirlant, Dierckx, and Stărică 2002; Matthys and Beirlant 2000) are based on these results. Furthermore it was found that for most distributions the algorithms seem to perform better for a fixed value of ρ, even if this parameter is misspecified. Following the suggestion by Matthys and Beirlant (2000) we consider $\rho = -1$.

Algorithm 2.

1. In the exponential regression model (3.1) consider $\rho = -1$ and determine the least-squares estimates for $\hat{\alpha}^{-1}$ and $\hat{b}_{n,k}$ for $k = 3, \ldots, n$.
2. Calculate an estimate of the $\text{AMSE}_{H^{-1}_{k,n}}$ by substituting ρ, α^{-1}, and $b_{n,k}$ with the values determined in Step 1.
3. Determine $\hat{k}^1_{\text{op}}$ by the value of k that minimizes the estimates of the $\text{AMSE}_{H^{-1}_{k,n}}$ in Step 2, and estimate α by $H_{\hat{k}^1_{\text{op}},n}$.

Algorithm 3.

1. In the exponential regression model (3.1) consider $\rho = -1$ and determine the least-squares estimates for $\hat{\alpha}^{-1}$ and $\hat{b}_{n,k}$ for $k = 3, \ldots, n$.
2. Calculate $\hat{k}_{\text{op},k}$ according to equation k_{op} defined earlier, substituting ρ, α^{-1}, and $b_{n,k}$ by the values determined in Step 1, for $k = 3, \ldots, n$.
3. Determined $\hat{k}^2_{\text{op}} = \text{median}\left\{\hat{k}_{\text{op},k},\ k = 3, \ldots, \frac{n}{2}\right\}$, and estimate α by $H_{\hat{k}^2_{\text{op}},n}$.

Table 1. Simulation Results for the Pareto(1,α) Distribution With $n = 200$ and 5,000

	200						*5,000*					
n	*A1*$_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_{100}	QQ_{100}	$A1_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_{100}	QQ_{100}
$\alpha =$ **1.1**												
STD	.1005	.0890	.0953	.2892	.0834	.1033	.0194	.0163	.0150	.0499	.0154	.0211
BIA	−.0384	.0126	.0131	.0451	.0050	−.0291	.0015	.0024	.0021	.0120	.0011	−.0018
MSE	.0116	.0081	.0093	.0857	.0070	.0115	.0004	.0003	.0002	.0026	.0002	.0004
$\alpha =$ **1.5**												
STD	.1527	.1224	.1628	.4927	.1100	.1413	.0319	.0231	.0202	.0726	.0209	.0294
BIAS	−.0395	.0124	.0286	.1197	.0067	−.0374	.0009	.0047	.0038	.0162	.0013	-.0033
MSE	.0249	.0151	.0273	.2570	.0122	.0214	.0010	.0006	.0004	.0055	.0004	.0009
$\alpha =$ **1.9**												
STD	.5705	.1331	.1652	.5347	.1304	.1736	.0389	.0298	.0273	.0846	.0275	.0381
BIAS	.0786	.0222	.0358	.0887	.0151	−.0373	−.0054	.0023	.0013	.0064	.0008	−.0062
MSE	.3317	.0182	.0286	.2937	.0172	.0315	.0015	.0009	.0007	.0072	.0008	.0015

In what follows, the performance of Algorithm 1 applied to the Zipf plot and the Sum plot is going to be compared to the two adaptive procedures defined in Algorithm 2 and Algorithm 3. These algorithms suggest a value of k for the estimation of the tail index. Except for the Zipf plot, the estimation of the tail index is determined by the Hill estimator, $H_{k,n}$.

3.1 Pareto Distribution

The observations drawn from a Pareto distribution with parameters (c, α) as defined in Proposition 1 were generated using the inversion method assuming $c = 1$. The approaches to be compared are: Algorithm 1 and the Zipf plot together with the QQ-estimator ($\mathrm{A1_{ZQQ}}$), Algorithm 1 and the Sum plot together with the Hill estimator ($\mathrm{A1_{SH}}$), Algorithm 2 (A2) and Algorithm 3 (A3). Since for this particular distribution the optimal value of k coincides with the sample size n, we also included the results of the Hill estimator (H_{100}) and the QQ-estimator ($\mathrm{QQ_{100}}$) based on all the observations. The results are presented in Table 1. The case of $n = 1{,}000$ can be seen in the electronic version at www.ingenta.com.

In terms of mean squared error, the results based on $\mathrm{A1_{SH}}$ are almost identical to those of H_{100}, although the Sum plot procedure exhibits a little higher bias. It is also worth noting that the bias and the mean squared error are of the same magnitude for all the different values of α. These results suggest that the proposed iterative procedure for fitting a line to the Sum plot identifies correctly the number of upper order statistics to be used.

The $\mathrm{A1_{ZQQ}}$ procedure exhibits the largest bias and mean squared error in most situations. Only Algorithm 3 performs, in some cases, worse than $\mathrm{A1_{ZQQ}}$. This behavior is not surprising due to the nature of the Zipf plot, where the estimation procedure is affected by the "jumpiness" present in the upper order statistics region of the graph. Even considering all observations, $\mathrm{QQ_{100}}$ exhibits larger mean square error than H_{100} and $\mathrm{A1_{SH}}$.

In general, Algorithm 3 tends to perform worse than Algorithm 2, except for moderate

Table 2. Simulation Results for the IG(α,1) Distribution With $n = 200$ and 5,000

	200						5,000					
n	$A1_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_5	H_{10}	$A1_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_5	H_{10}
α= **1.1**												
STD	.6003	.7303	.5932	.3333	.4506	.2622	.0460	.0701	.0325	.0599	.0687	.0482
BIAS	−.1357	−.1783	−.1371	.0557	.1094	.0205	−.1461	−.0145	−.1502	−.0199	−.0130	−.0330
MSE	.3788	.5651	.3707	.1142	.2150	.0692	.0235	.0051	.0236	.0040	.0049	.0034
α= **1.5**												
STD	.6604	.9980	1.0196	.4501	.6071	.3447	.0900	.1056	.0514	.0830	.0918	.0618
BIAS	−.2894	−.3503	−.1461	.0346	.0923	−.0414	−.2479	−.0611	−.2599	−.0727	−.0572	−.0994
MSE	.5198	1.1187	1.0609	.2038	.3770	.1205	.0695	.0149	.0702	.0122	.0117	.0137
α= **1.9**												
STD	1.3957	.8256	1.6259	.7945	.7003	.4036	.0680	.1290	.0710	.1161	.1111	.0736
BIAS	−.3735	−.4130	−.0939	.0675	.0586	−.1157	−.4167	−.1239	−.3881	−.1170	−.1195	−.1905
MSE	2.0875	.8522	2.6524	.6357	.4938	.1763	.1783	.0320	.1556	.0272	.0266	.0417

sample sizes ($n = 1,000$). In such cases, there does not exist a uniformly better procedure for all the different values of α. Overall, all the approaches considered perform reasonably well, as expected since the Pareto distribution represents the baseline case.

3.2 INVERTED GAMMA DISTRIBUTION

The Inverted Gamma random variables X, IGa(α, β), are generated from Gamma random variables Y, Ga(α, β), and then computing $X = Y^{-1}$. In this study, the value of β was fixed to 1. The methods compared are the same as the ones in the preceding section, except for the last two procedures, H_{100} and QQ$_{100}$. Instead the value of α is estimated from the Hill estimator by considering the largest 5% (H_5) and 10% (H_{10}) observations, respectively, a procedure commonly used in practice. The full results are given in the electronic version at www.ingenta.com, and the cases on $n =$ 200 and $n =$ 5,000 in Table 2.

Algorithm 3 and the fixed rule H_{10} give the best results for small and moderate sample sizes. Notice that for large sample sizes, $n = 5,000$, H_5 seems to be the preferable choice for $\alpha = 1.5$ and 1.9. For $n = 5,000$, the approaches A1$_{\text{SH}}$, A3, H_5, and H_{10} give very similar results. It is clear that Algorithm 1 applied to the Zipf plot together with the QQ-estimator is the procedure that gives in general higher mean squared errors. Also notice that for $n =$ 1,000 and 5,000, the results produced from Algorithm 1 together with the Sum plot and the Hill estimator are in general between the two adaptive procedures A2 and A3.

It is quite interesting that a simple fixed rule such as H_{10} or H_5 gives such good results. As can been seen in Figure 6, where the Pareto and the Inverted Gamma densities are graphed for $\alpha = 1.5$, the departure of the Pareto tail behavior occurs between the 90th and the 95th percentiles. Therefore, it does not come as a surprise that these particular choices for the

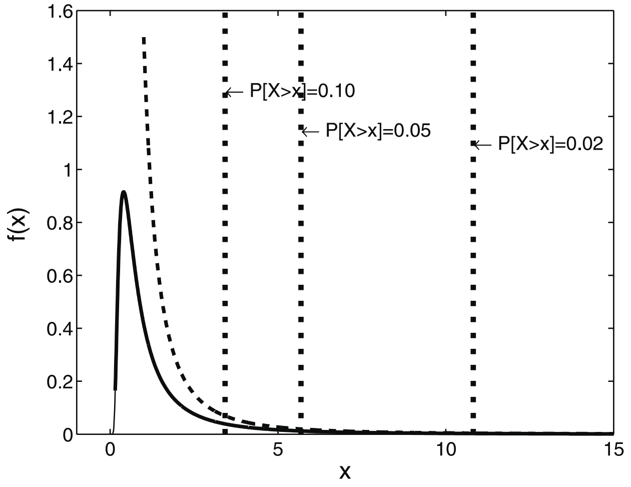

Figure 6. Probability density functions of a P(1,1.5) (−−) and an IGa(1.5,1) (___) distribution.

fixed rules will perform reasonably well for the Inverted Gamma distribution. However, as the results for the α-Stable distribution show, this good performance cannot be taken for granted.

3.3 α-Stable Distribution

The observations are generated from a symmetric α-Stable distribution using the method of Chambers, Mallows, and Struck (1976). The procedures analyzed for the α-Stable distribution were the same as the ones in the previous section and are given in Table 3, and in its full extent in the electronic version at www.ingenta.com. The results for large sample sizes show why relying on fixed rules is not a particularly good practice, since they have very high MSE. The adaptive procedure A3 outperforms its competitors for $\alpha = 1.1$, while for the remaining values of α both A2 and A3 perform well for moderate and large

Table 3. Simulation Results for the Symmetric α-Stable Distribution With $n = 400$ and 5,000

	200						*5,000*					
n	$A1_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_5	H_{10}	$A1_{ZQQ}$	$A1_{SH}$	*A2*	*A3*	H_5	H_{10}
STD	.4759	.9250	.7045	.3427	.4539	.2806	.0731	.0686	.0389	.0605	5.3822	1.7885
BIAS	−.1589	−.0476	.0391	.0921	.1448	.0697	−.0586	.0262	−.0596	.0145	1.2291	.5758
MSE	.2517	.8579	.4978	.1259	.2270	.0836	.0088	.0054	.0051	.0039	30.4782	3.5302
STD	.7148	.4578	.6183	.4685	.6850	.4425	.1049	.1006	.0774	.1061	6.7560	2.0452
BIAS	−.3005	−.2592	.1838	.2994	.3370	.2971	.0743	.2063	.1398	.1512	1.6475	.7076
MSE	.6012	.2768	.4161	.3091	.5827	.2841	.0165	.0527	.0255	.0341	48.3580	4.6836
STD	2.0651	1.3476	2.9124	2.2986	1.9511	1.0248	.3353	.2669	.0804	.1053	5.2538	2.6140
BIAS	.4760	.4444	2.0264	2.7208	2.4106	1.9065	.7587	1.6443	−.2635	−.2503	1.9633	1.0009
MSE	4.4914	2.0136	12.5884	12.6864	9.6176	4.6849	.6881	2.7750	.0759	.0737	31.4575	7.8348

sample sizes. The $\mathrm{A1}_{\mathrm{SH}}$ performs well for smaller values of α and for moderate and large sample sizes, while the $\mathrm{A1}_{\mathrm{ZQQ}}$ procedure performs satisfactorily in similar settings. It can also be seen that among the two adaptive procedures, A3 tends to outperform A2, especially in the presence of very heavy tails ($\alpha = 1.1$).

This simulation study illustrates that the proposed method performs well for different heavy tailed distributions. It also shows that the adaptive procedures are very competitive, while fixed rules can be misleading. The advantage of the proposed method is that is accompanied by a diagnostic graph that identifies successfully where the behavior of the tail that is consistent with (1.1) ends. On the other hand, a close examination of several of the generated samples used in the simulation study shows that when the Hill plot does not quite stabilize, the results of the adaptive algorithms are not particularly good.

4. DATA EXAMPLES

In this section the Sum plot is used on three real datasets. Because the Sum plot is used together with the Hill estimator, in some of the following graphs we will present both plots in a single graph, making it easier to determine the value of k and the estimated value for α given by $H_{k,n}$.

4.1 AUSTRALIA COMMUNITY SIZE DATA

A famous example of city size data is the Australian Community Size dataset, used in the original work of Zipf (Zipf 1949) and more recently re-examined by Feuerverger and Hall (1999). The dataset includes all the communities with more than 2,000 people in 1921, and the sample size is 256. Feuerverger and Hall (1999) suggested using the transformation $Y = N_0 - X$, where N_0 is a large positive value. In practice, N_0 can correspond to the size of the 5th, the 10th, or the 20th largest community. We settled on $N_0 = X^{(10)}$ and used the Y^{-1} values as our data. In Figure 7, the Hill, the Sum, and the Zipf plots for this dataset are shown. The Hill plot is not particularly informative, since it fails to stabilize; possible candidate values for α range from 1.15 to 1.30. On the other hand, the Sum plot exhibits a strong linear pattern for the largest order statistics. The Sum plot clearly indicates a break from linearity at the order $k = 50$, which corresponds to a value of $H_{50,246} = 1.291$, which is in agreement with values given by Feuerverger and Hall (1999). Algorithm 2 suggests a value of 1.5333 for α based on the 71 largest observations, and Algorithm 3 estimates the tail index by $H_{29,246} = 1.267$.

The Zipf plot shows large volatility at the top observations, which will highly influence the outcome of the procedure introduced in the simulation study. The value suggested for k is 5, with the QQ-estimator giving a value of .709 for α. Removing the top three observations, we obtain an estimate for the tail index of 1.185, based on the 20 largest observations of this reduced dataset. The two different fitted lines can be seen in the right panel of Figure 7.

All the values suggested are within the expected results for this data and in accordance

Corrupt file.

Figure 7. The Sum and Hill plots (left panel) and the Zipf plot (right panel).

to the results obtained by Zipf (1949) and Feuerverger and Hall (1999). The great advantage of the Sum plot is its stability towards the top observations in a dataset. The Sum plot uses all the information available in making the difficult decision of selecting the number of order statistics in the estimation of the tail index, without the subjectivity that takes place with the removal of the top three observations in the Zipf plot. It is important to notice that the results from the Sum plot together with the Hill estimator are within the ones obtained from the two adaptive procedures (A2 and A3). We are currently investigating whether the perturbed Pareto suggested in Feuerverger and Hall (1999) gives rise to a concave Sum plot, as shown in Figure 7. It is clear that our approach properly identifies the portion of the data exhibiting a heavy-tailed behavior, providing a good estimate for the tail exponent.

4.2 Network Connection Data

The next dataset describes the degree of connectivity between autonomous systems (AS—networks under a single administrative authority) on the Internet for the year 2000 and is provided by the National Laboratory for Applied Network Research. The information has been used to characterize the topology of the Internet (Faloutsos, Faloutsos, and Faloutsos 1999; Chen et al. 2002) and its impact on TCP/IP protocol dynamics (Feldmann, Huang, Gilbert, and Willinger 1999). The size of the data is 6,474 and each observation indicates to how many other ASs any system is connected to. The histogram plotted on a log-scale indicates that the vast majority of ASs are connected to only one of their peers, but there are a few ASs that are connected to almost 30% of their peers. This finding contradicts the long held hypothesis that the topology of the Internet can be captured by that of a random graph, to be replaced by the hypothesis of a power-law graph (Barabasi and Albert 1999). The Hill, the Sum, and the Zipf plots are given next.

In Figure 8 (middle panel), we take a closer look at the first 500 observations from the Sum and Hill plots (top panel). Once again the Hill plot fails to stabilize, while the Sum plot

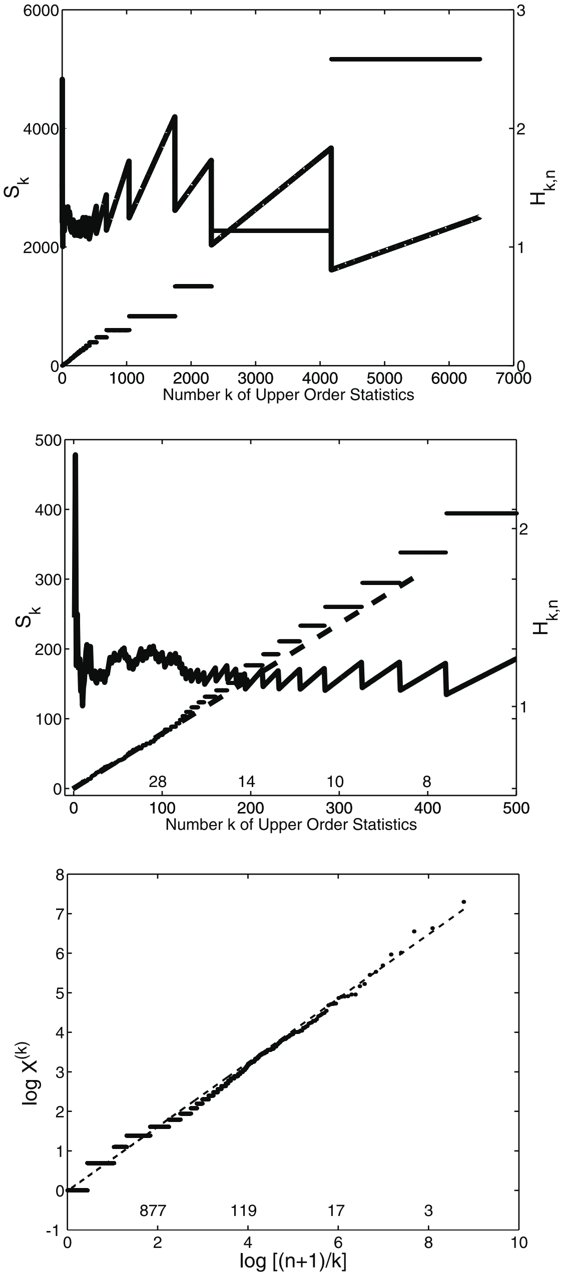

Figure 8. The Sum and Hill plots based on all (top), on the largest 500 (middle) observations, and the Zipf plot (bottom).

exhibits a strong initial linear pattern. The Sum plot suggests using around 125 observations, estimating the parameter α by $H_{125,6474} = 1.203$, which confirms the results reported by Faloutsos, Faloutsos, and Faloutsos (1999). It is interesting to note the sawtooth pattern in both graphs, which is a consequence of the repeated integer values of the observations. However, there is an unmistakable linear pattern for the largest order statistics, which helps us to identify the value of k. We are currently investigating ways to best make such plots for this type of data.

The Zipf plot in Figure 9 (bottom panel) is less affected by the nature of the data, because we are dealing with a log-log scale graph. The suggested value for the tail index is 1.233, based on all the 6,474 observations. Algorithms 2 and 3 suggest a value for the tail index of .8258 and 1.1651, respectively. The proposed estimate from Algorithm 2 seems to be a little too low since all the other procedures suggest values above 1. As can be seen in Matthys and Beirlant (2000) and confirmed by our simulation study, Algorithm 3 is usually preferable to Algorithm 2.

4.3 World Natural Gas Data

The data consist of the volumes of 369 natural gas world provinces. The data can be found in Table 1 at http://greenwood.cr.usgs.gov/energy/WorldEnergy/OF97-463. The study of the patterns in these types of data will help in understanding the development of future natural gas resources leading to better assessments of the reserve growth potential of the world's provinces. We show next the Sum, the Hill and the Zipf plots for this dataset.

By looking at Figure 9 we realize that the Hill plot does not quite stabilize in any particular region, making it extremely hard to decide upon a specific value for α simply from this graph. The Zipf plot shows as usual a little variability towards the larger observations, suggesting a value of k equals to 122 with a QQ-estimate of .7994. The Sum plot clearly detects the departure of linearity around $k = 19$ with a value of α given by $H_{19,369} = 1.2448$. Also, the values of the tail index given by A2 and A3 are 1.3931 and 1.1098, respectively. It can be seen that the shape of the Sum plot suggests the existence of a Pareto like tail in the data and provides to the data analyst a good estimate for the tail index.

5. CONCLUDING REMARKS

The main contribution of this article is the development of a diagnostic plot (the Sum plot) that allows the data analyst to identify successfully the order statistic that signals the beginning of a power law behavior in the tail of the data's distribution. The Sum plot overcomes many of the problems encountered in previous approaches, such as the difficulty in locating a stable region in the Hill plot and the high variability present in the region of the upper order statistics in the Zipf plot. The main advantage of the Sum plot is that the power law behavior in the tail coincides with a strong linear pattern; thus, transforming the problem of identifying the number of upper order statistics to be used in the estimation of the tail

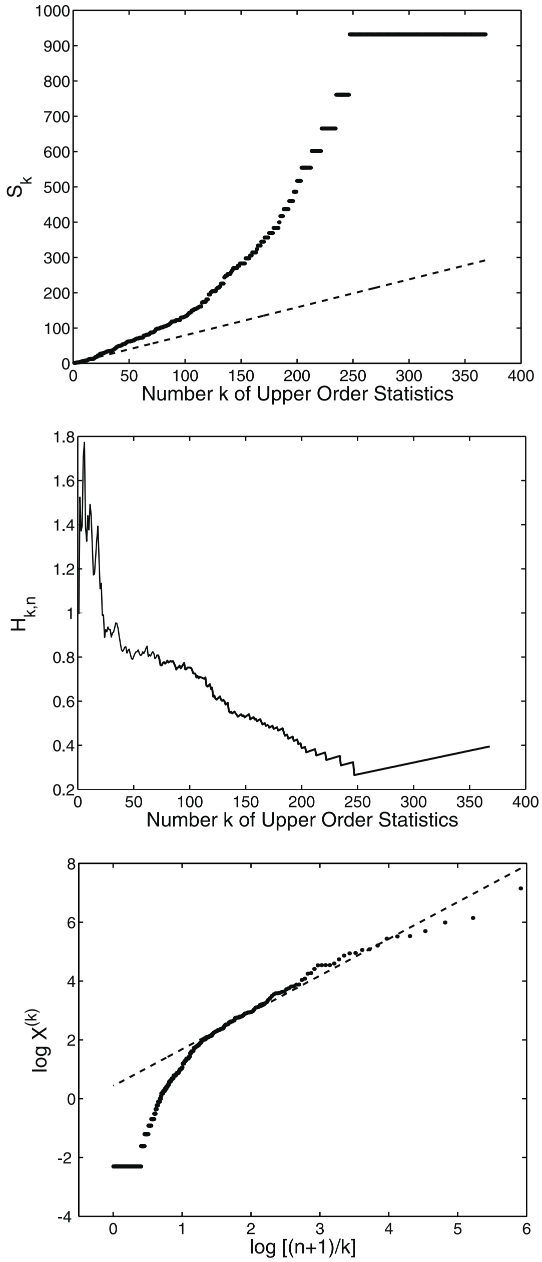

Figure 9. *The Sum plot (top), the Hill plot (middle), and the Zipf plot (bottom).*

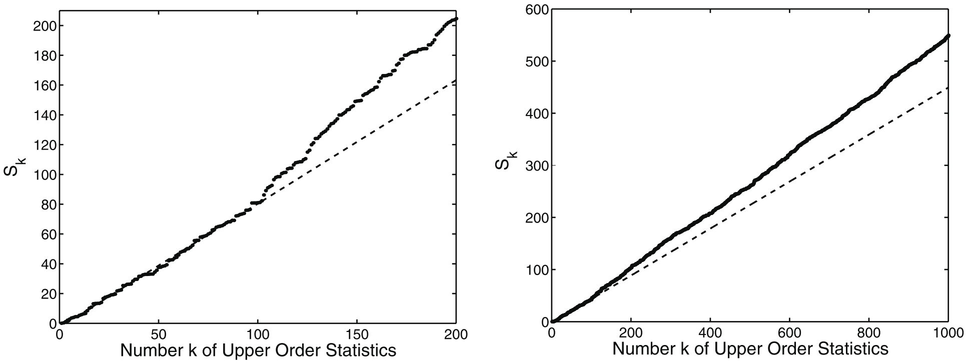

Figure 10. The Sum plot based on the largest 200 (left), and the largest 1,000 (right) observations from a sample of size 1,000 and 1,000,000, respectively, from a lognormal distribution with $\mu = 0$ and $\sigma = 2$.

index to one of detecting departures from linearity. Moreover, an extensive simulation study confirms that the Sum plot combined with the Hill estimator outperforms approaches based on the Zipf plot and its variants and on fixed rules favored in practice. It is also competitive against adaptive procedures that minimize the asymptotic mean squared error of the Hill estimator. Although the Sum plot has shown to be very useful in the applications presented in this study, where the volatile behavior in the Hill plot and in the top of the Zipf plot is taken away, its performance has not been tested for certain situations. For example, when mixtures occur in the tails of a distribution, such as in insurance, the Sum plot should be used with caution. The Sum plot should always be interpreted as one more tool to be used in the estimation of the tail index and compared to the methods available to date.

An extension currently under investigation is to study the patterns produced by some specific distributions that have been suggested in the literature as good models for capturing the behavior of certain aspects of Internet traffic [e.g., the double-Pareto lognormal distribution (Reed 2001)].

We conclude with a short discussion regarding the lognormal distribution, which has proved to be an extremely interesting case in the study of heavy-tailed distributions. In many studies (Crovella and Taqqu 1999; Adler, Feldman, and Taqqu 1998) the techniques used for estimating the tail index failed to correctly assess the non-heaviness of the tail for data drawn from a lognormal distribution with $\sigma = 2$. These approaches generally suggest values of $\alpha < 1.5$. We briefly discuss the difficulties that the lognormal distribution presents. In Figure 10 the Sum plots of a small number of upper order statistics from samples of size 1,000 (left panel) and 1,000,000 (right panel) of a lognormal distribution with mean zero and standard deviation 2 are shown.

It can be seen that for $n = 1{,}000$, the Sum plot fails to detect the presence of a light tail, since the suggested value of k is 100 corresponding to a $\hat{\alpha} = 1.2349$. However, the same holds true for all the other approaches discussed in the article. On the other hand, in the presence of a very large sample the Sum plot clearly detects the departure from linearity at around $k = 110$, with a corresponding value for the tail index of 2.1768. On the other hand, the A2 and A3 algorithms suggest values for $\hat{\alpha} = 1.54$ and 1.74, respectively, while the QQ-estimator gives a value for $\hat{\alpha} = 1.438$. These results further illustrate the usefulness of the Sum plot in data analysis studies.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two anonymous referees for their useful suggestions that improved the presentation of the article. The first author was partially supported by FCT/POCTI/FEDER, and the work of the second author was supported in part by NSF grant IIS 9988095.

[Received February 2002. Revised September 2003.]

REFERENCES

Adler, R. J., Feldman, R. E., and Taqqu, M. S. (1998), *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Boston, MA: Birkhauser.

Barabasi, A. L., and Albert, R. (1999), “Emergence of Scaling in Random Networks,” *Science*, 256, 509–512.

Beirlant, J., Dierckx, G., and Stărică, C. (2002), “On Exponential Representation of Log-Spacings of Extreme Order Statistics,” *Extreme*, 5, 157–180.

Beirlant, J., Vynckier, P., and Teugels, J. L. (1996), “Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics,” *Journal of the American Statistical Association*, 91, 1659–1667.

Chambers, J. M., Mallows, C. L., and Stuck, B. W. (1976), “A Method for Simulating Stable Random Variables,” *Journal of the American Statistical Association*, 71, 340–344.

Chen, Q., Chang, H., Govindan, R., Jamin, S., Shenker, S., and Willinger, W. (2002), “The Origin of Power-Laws in Internet Topology Revisited,” *Proceedings of IEEE Infocom 2002*.

Crovella, M. E., and Taqqu, M. S. (1999), “Estimating the Heavy Tail Index from Scaling Properties,” *Methodology and Computing in Applied Probability*, 1, 55–79.

Csörgő, S., Deheuvels, P., and Mason, D. M. (1985), “Kernel Estimates of the Tail Index of a Distribution,” *Annals of Statistics*, 13, 1050–1077.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001), “Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation,” *Journal of of Multivariate Analysis*, 76, 226–248.

Deheuvels, P., Haeusler, E., and Mason, D. M. (1988), “Almost Sure Convergence of the Hill Estimator,” *Mathematical Proceedings of the Cambridge Philosophical Society*, 104, 371–381.

de Sousa, B. (2002), “A Contribution to the Estimation of the Tail Index of Heavy-Tailed Distributions,” Ph.D. Dissertation, University of Michigan.

Drees, H., and Kaufmann, E. (1998), “Selecting the Optimal Sample Fraction in Univariate Extreme Value Estimation,” *Stochastic Processes and Their Applications*, 75, 149–172.

Drees, H., de Haan, L., and Resnick, S. (2000), “How to make a Hill plot,” *The Annals Statistics*, 28, 254–274.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999), “On Power-law Relationships of the Internet Topology,” in *Proceedings of ACM Sigcomm 99*, pp. 251–262.

Feldmann, A., Huang, P., Gilbert, A., and Willinger, W. (1999), “Dynamics of IP Traffic: A Study of the Role of Variability and the Impact of Control,” in *Proceedings of ACM Sigcomm 99*, pp. 301–313.

Feuerverger, A., and Hall, P. (1999), “Estimating a Tail Exponent by Modelling Departure from a Pareto Distribution,” *The Annals of Statistics*, 27, 760–781.

Guillou, A., and Hall, P. (2001), “A Diagnostic for Selecting the Threshold in Extreme Value Analysis,” *Journal of the Royal Statistical Society*, Ser. B, 63, 293–305.

Hall, P. (1982), “On Some Simple Estimates of an Exponent of Regular Variation,” *Journal of the Royal Statistical Society*, Series B, 44, 37–42.

Hill, B. M. (1975), “A Simple General Approach to Inference About the Tail of a Distribution,” *The Annals of Statistics*, 3, 1163–1174.

Hsieh, P.-H. (1999), “Robustness of Tail Index Estimation,” *Journal of Computational and Graphical Statistics*, 8, 318–332.

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.

Kratz, M., and Resnick, S. (1996), “The QQ-Estimator and Heavy Tails,” *Stochastic Models*, 12, 699–724.

Mason, D. (1982), “Laws of Large Numbers for Sums of Extreme Values,” *Annals of Probability*, 10, 754–764.

Matthys, G., and Beirlant, J. (2000), “Adaptive Threshold Selection in Tail Index Estimation,” in *Extremes and Integrated Risk Management*, ed. P. Embrechts, London: Risk Books.

McGee, V. E., and Carleton, W. T. (1970), “Piecewise Regression,” *Journal of the American Statistical Association*, 65, 1109–1124.

Reed, W. J. (2001), “The Double Pareto-Lognormal Distribution—A New Parametric Model for Size Distributions,” [on-line], http://www.math.uvic.ca/faculty/reed/index.html.

Rényi, A. (1953), “On the Theory of Order Statistics,” *Acta Mathematica Academiae Scientiarukm Hungaricae*, 4, 191–232.

Resnick, S. (1997), “Heavy Tail Modeling and Teletraffic Data,” *The Annals of Statistics*, 25, 1805–1869.

Resnick, S., and Stărică, C. (1997), “Smoothing the Hill Estimator, *Applied Probability*, 29, 271–293.

Sen, A., and Srivastava, M. (1990), *Regression Analysis: Theory, Methods and Application*, New York: Springer.

Shaked, M., and Shanthikumar, G. (1988), “Stochastic Convexity and its Applications,” *Advances in Applied Probability*, 20, 427–448.

U.S. Department of the Interior Geological Survey [on-line]. Available at http://greenwood.cr.usgs.gov/energy/WorldEnergy/OF97-463.

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Reading, MA: Addison-Wesley.

APPENDIX

Proof of Proposition 1: If the random variable X follows the Pareto distribution as defined in the proposition, then the distribution function of the random variable $Y = c^{-1}X$ is $P[Y > y] = P[c^{-1}X > y] = P[X > cy] = y^{-\alpha}$, for $y > 1$, that is, the random variable Y has a Pareto$(1, \alpha)$ distribution. Hence, without loss of generality, for the remainder of the proof we will assume that the constant $c = 1$.

The proof of this proposition is based on the following two results given by Shaked and Shanthikumar (1988).

Definition 1. *Let $\{X(\theta),\ \theta \in \Theta\}$ be a set of random variables with distribution function F_θ. If for any $\theta_i \in \Theta,\ i$ =1, 2, 3, 4, such that $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4$ and $\theta_1 + \theta_4 = \theta_2 + \theta_3$, there exist four random variables $\hat{X}_i,\ i$ =1, 2, 3, 4 which satisfy:*

1. *$\hat{X}_i =_{st} X(\theta_i)$, for i = 1, 2, 3, 4. The symbol $=_{st}$ denotes equality in distribution,*
2. *$[\hat{X}_1, \hat{X}_2, \hat{X}_3] \leq \hat{X}_4$almost surely, that is, all the random variables in the brackets are $\leq$ than the random variable on the right side of the inequality almost surely*
3. *$\hat{X}_1 + \hat{X}_4 = \hat{X}_2 + \hat{X}_3$almost surely, then $\{X(\theta),\ \theta \in \Theta\}$ is said to be stochastically increasing and linear in the sample path sense, SIL(sp).*

Proposition 3. *If $\{X(\theta),\ \theta \in \Theta\} \in SIL(sp)$, then $\{X(\theta),\ \theta \in \Theta\} \in SIL$. In other words, for A, the class of all increasing and linear real functions on R, and A_θ, the class of all increasing and linear real functions on Θ, we have $\phi \in A \Rightarrow E\phi(X(.)) \in A_\Theta$.*

See Shaked and Shanthikumar (1988) for a proof of this result.

Let $i_1 \leq i_2 \leq i_3 \leq i_4$ such that $i_1 + i_4 = i_2 + i_3$. Consider the four random variables $S_{i_1},\ S_{i_2},\ S_{i_3},\ S_{i_4}$ such that:

$$\begin{aligned}
S_{i_1} &= \sum_{i=1}^{i_1} iV_i, \\
S_{i_2} &= \sum_{i=1}^{i_2} iV_i = S_{i_1} + S'_{i_2-i_1}, \quad \text{where} \quad S'_{i_2-i_1} = \sum_{i=i_1+1}^{i_2} iV_i, \\
S_{i_3} &= S_{i_1} + S'_{i_2-i_1} + S'_{i_3-i_2}, \quad \text{where} \quad S'_{i_3-i_2} = \sum_{i=i_2+1}^{i_3} iV_i, \\
S_{i_4} &= S_{i_1} + S'_{i_2-i_1} + S'_{i_3-i_2} + S'_{i_4-i_3}, \quad \text{where} \quad S'_{i_4-i_3} = \sum_{i=i_3+1}^{i_4} iV_i.
\end{aligned}$$

Considering the Pareto distribution, the random variables $\{iV_i, i = 1, \ldots, n-1\}$ are independent exponentially distributed with mean equal to α^{-1} (see Section 2). Since $i_2 - i_1 = i_4 - i_3$, the random variables $S'_{i_2-i_1}$ and $S'_{i_4-i_3}$ are equally Ga$(i_2 - i_1, \alpha)$ distributed. Consider now the following random variables: $\hat{S}_{i_1} = S_{i_1}$, $\hat{S}_{i_2} = S_{i_2}$, $\hat{S}_{i_3} = S_{i_1} + S'_{i_3-i_2} + S'_{i_4-i_3}$, and $\hat{S}_{i_4} = S_{i_4}$. Then Conditions 1 and 2 from the previous definition are obviously satisfied. For Condition 3, notice that $\hat{S}_{i_2} + \hat{S}_{i_3} = 2S_{i_1} + S^{'}_{i_2-i_1} + S^{'}_{i_3-i_2} + S^{'}_{i_4-i_3}$ and $\hat{S}_{i_1} + \hat{S}_{i_4} = 2S_{i_1} + S^{'}_{i_2-i_1} + S^{'}_{i_3-i_2} + S^{'}_{i_4-i_3}$. Hence $\hat{S}_{i_1} + \hat{S}_{i_4} = \hat{S}_{i_2} + \hat{S}_{i_3}$, almost surely. By Definition 1 we conclude that the random variables $\{S_k, k = 1, \ldots, n-1\}$ are SIL(sp), and from Proposition 1 we can conclude the SIL of the random variables in question. □

Proof of Proposition 2: The matrix Ω can be decomposed as $\alpha^{-2}\Theta'\Theta$, where the matrices $\Theta = \left[\mathbf{1}_{\{j \leq i\}}\right]_{i,j=1}^{k}$ and $\Theta^{-1} = I - \left[\mathbf{1}_{\{j-i=-1\}}\right]_{i,j=1}^{k}$ with I the identity matrix and $\mathbf{1}_{\{.\}}$ the indicator function. Substituting Ω by its decomposition and proceeding with the definition of the generalized least squares estimator, $\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, the result follows immediately. The estimator of α^{-1} is the estimated slope of the fitting line, that is, β_1. Thus,

$$\begin{aligned}
\hat{\alpha}^{-1} &= \frac{1}{k-1}\sum_{i=2}^{k} iV_i = \frac{1}{k-1}\left\{\sum_{i=2}^{k}\log X^{(i)} - k\log X^{(k+1)}\right\} \\
&= \frac{1}{k-1}\left\{\sum_{i=1}^{k}\log X^{(i)} - k\log X^{(k+1)} - \log X^{(1)}\right\} \\
&= \frac{k}{k-1}H_{k,n}^{-1} - \frac{1}{k-1}\log X^{(1)}.
\end{aligned}$$

□

Proof of Corollary 1: The proof is identical to the previous one, with $X = [i]_{i=1}^{k}$ a $k \times 1$ matrix and $\beta = (\beta_1)$ an unknown regression parameter. Calculating the generalized least squares estimator we have

$$\hat{\alpha}^{-1} = \frac{1}{k}\sum_{i=1}^{k} iV_i = \frac{1}{k}\left\{\sum_{i=1}^{k}(\log X_i - \log X_{k+1})\right\} = H_{k,n}^{-1}.$$

□

TABLES

Table A.1. Simulation Results for the Pareto(1,α) distribution with $n = 200$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} =$ **1.1**						
STD	.1005	.0890	.0953	.2892	.0834	.1033
BIAS	−.0384	.0126	.0131	.0451	.0050	−.0291
MSE	.0116	.0081	.0093	.0857	.0070	.0115
$\boldsymbol{\alpha} =$ **1.5**						
STD	.1527	.1224	.1628	.4927	.1100	.1413
BIAS	−.0395	.0124	.0286	.1197	.0067	−.0374
MSE	.0249	.0151	.0273	.2570	.0122	.0214
$\boldsymbol{\alpha} =$ **1.9**						
STD	.5705	.1331	.1652	.5347	.1304	.1736
BIAS	.0786	.0222	.0358	.0887	.0151	−.0373
MSE	.3317	.0182	.0286	.2937	.0172	.0315

Table A.2. Simulation results for the Pareto(1,α) Distribution With $n = 1{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} =$ **1.1**						
STD	.0498	.0365	.0366	.1341	.0346	.0475
BIAS	−.0105	.0022	.0020	.0261	.0007	−.0093
MSE	.0026	.0013	.0187	.0013	.0012	.0023
$\boldsymbol{\alpha} =$ **1.5**						
STD	.0511	.0306	.0413	.1271	.0478	.0658
BIAS	−.0098	−.0015	.0131	.0072	.0004	−.0149
MSE	.0027	.0009	.0163	.0018	.0023	.0046
$\boldsymbol{\alpha} =$ **1.9**						
STD	.0859	.0682	.0673	.1952	.0615	.0875
BIAS	−.0137	.0071	.0067	.0170	.0036	−.0140
MSE	.0076	.0046	.0046	.0384	.0038	.0079

Table A.3. Simulation Results for the Pareto(1,α) Distribution with $n = 5{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha = 1.1}$						
STD	.0194	.0163	.0150	.0499	.0154	.0211
BIAS	.0015	.0024	.0021	.0120	.0011	−.0018
MSE	.0004	.0003	.0002	.0026	.0002	.0004
$\boldsymbol{\alpha = 1.5}$						
STD	.0319	.0231	.0202	.0726	.0209	.0294
BIAS	.0009	.0047	.0038	.0162	.0013	−.0033
MSE	.0010	.0006	.0004	.0055	.0004	.0009
$\boldsymbol{\alpha = 1.9}$						
STD	.0389	.0298	.0273	.0846	.0275	.0381
BIAS	−.0054	.0023	.0013	.0064	.0008	−.0062
MSE	.0015	.0009	.0007	.0072	.0008	.0015

Table A.4. Simulation Results for the IG(α,1) Distribution with $n = 200$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha = 1.1}$						
STD	.6003	.7303	.5932	.3333	.4506	.2622
BIAS	−.1357	−.1783	−.1371	.0557	.1094	.0205
MSE	.3788	.5651	.3707	.1142	.2150	.0692
$\boldsymbol{\alpha = 1.5}$						
STD	.6604	.9980	1.0196	.4501	.6071	.3447
BIAS	−.2894	−.3503	−.1461	.0346	.0923	−.0414
MSE	.5198	1.1187	1.0609	.2038	.3770	.1205
$\boldsymbol{\alpha = 1.9}$						
STD	1.3957	.8256	1.6259	.7945	.7003	.4036
BIAS	−.3735	−.4130	−.0939	.0675	.0586	−.1157
MSE	2.0875	.8522	2.6524	.6357	.4938	.1763

Table A.5. Simulation Results for the IG(α, 1) Distribution with $n = 1{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha = 1.1}$						
STD	.1714	.2401	.0851	.1419	.1581	.1070
BIAS	−.1721	−.0187	−.1775	.0077	.0004	−.0298
MSE	.0590	.0580	.0388	.0202	.0250	.0123
$\boldsymbol{\alpha = 1.5}$						
STD	.1725	.2188	.1108	.1620	.2134	.1377
BIAS	−.3046	−.1622	−.3211	−.0464	−.0271	−.0866
MSE	.1226	.0742	.1154	.0284	.0463	.0264
$\boldsymbol{\alpha = 1.9}$						
STD	.2520	.2978	.1576	.2268	.2430	.1648
BIAS	−.4436	−.2070	−.4446	−.1136	−.0974	−.1850
MSE	.2603	.1316	.2225	.0643	.0685	.0614

Table A.6. Simulation Results for the IG(α, 1) Distribution with $n = 5{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} = \mathbf{1.1}$						
STD	.0460	.0701	.0325	.0599	.0687	.0482
BIAS	−.1461	−.0145	−.1502	−.0199	−.0130	−.0330
MSE	.0235	.0051	.0236	.0040	.0049	.0034
$\boldsymbol{\alpha} = \mathbf{1.5}$						
STD	.0900	.1056	.0514	.0830	.0918	.0618
BIAS	−.2479	−.0611	−.2599	−.0727	−.0572	−.0994
MSE	.0695	.0149	.0702	.0122	.0117	.0137
$\boldsymbol{\alpha} = \mathbf{1.9}$						
STD	.0680	.1290	.0710	.1161	.1111	.0736
BIAS	−.4167	−.1239	−.3881	−.1170	−.1195	−.1905
MSE	.1783	.0320	.1556	.0272	.0266	.0417

Table A.7. Simulation Results for the Symmetric α-Stable Distribution with $n = 400$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} = \mathbf{1.1}$						
STD	.4759	.9250	.7045	.3427	.4539	.2806
BIAS	−0.1589	−0.0476	.0391	.0921	.1448	.0697
MSE	.2517	.8579	.4978	.1259	.2270	.0836
$\boldsymbol{\alpha} = \mathbf{1.5}$						
STD	.7148	.4578	.6183	.4685	.6850	.4425
BIAS	−.3005	−.2592	.1838	.2994	.3370	.2971
MSE	.6012	.2768	.4161	.3091	.5827	.2841
$\boldsymbol{\alpha} = \mathbf{1.9}$						
STD	2.0651	1.3476	2.9124	2.2986	1.9511	1.0248
BIAS	.4760	.4444	2.0264	2.7208	2.4106	1.9065
MSE	4.4914	2.0136	12.5884	12.6864	9.6176	4.6849

Table A.8. Simulation Results for the Symmetric α-Stable Distribution with $n = 2{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} = \mathbf{1.1}$						
STD	.1848	.1693	.9667	.1209	.1657	.1147
BIAS	−.0829	.0040	.0157	.0222	.0289	.0173
MSE	.0410	.0287	.9348	.0151	.0283	.0135
$\boldsymbol{\alpha} = \mathbf{1.5}$						
STD	.2123	.2243	.1513	.2288	.2575	.1834
BIAS	−.0170	.1660	.1078	.1990	.1748	.2404
MSE	.0454	.0779	.0345	.0920	.0969	.0914
$\boldsymbol{\alpha} = \mathbf{1.9}$						
STD	.9740	1.0268	.6015	1.0929	.6675	.3886
BIAS	.8901	1.3679	1.3003	1.8677	1.8631	1.6840
MSE	1.7410	2.9253	2.0527	4.6821	3.9166	2.9867

Table A.9. Simulation Results for the Symmetric α-Stable Distribution with $n = 10{,}000$

	$A1_{ZQQ}$	$A1_{SH}$	$A2$	$A3$	H_{100}	QQ_{100}
$\boldsymbol{\alpha} = \mathbf{1.1}$						
STD	.0731	.0686	.0389	.0605	5.3822	1.7885
BIAS	−.0586	.0262	−.0596	.0145	1.2291	.5758
MSE	.0088	.0054	.0051	.0039	30.4782	3.5302
$\boldsymbol{\alpha} = \mathbf{1.5}$						
STD	.1049	.1006	.0774	.1061	6.7560	2.0452
BIAS	.0743	.2063	.1398	.1512	1.6475	.7076
MSE	.0165	.0527	.0255	.0341	48.3580	4.6836
$\boldsymbol{\alpha} = \mathbf{1.9}$						
STD	.3353	.2669	.0804	.1053	5.2538	2.6140
BIAS	.7587	1.6443	−.2635	−.2503	1.9633	1.0009
MSE	.6881	2.7750	.0759	.0737	31.4575	7.8348