Pareto Tail Index Estimation Revisited

Mark Finkelstein,* Howard G. Tucker,† and Jerry Alan Veeh‡

Abstract

An estimator of the tail index of a Pareto distribution is given that is based on the use of the probability integral transform. This new estimator provides performance that is comparable to the best robust estimators, while retaining conceptual and computational simplicity. A tuning parameter in the new estimator can be adjusted to control the tradeoff between robustness and efficiency. The method used to compute the estimator also can be used to find a confidence interval for the tail index that is guaranteed to have the nominal confidence level for any given sample size. Guidelines for the use of the new estimator are provided.

1. Introduction

The Pareto distribution, whose distribution function is

$$F(x) = 1 - \left(\frac{\sigma}{x}\right)^{\alpha}, \quad x \geq \sigma$$

for fixed constants $\alpha > 0$ and $\sigma > 0$, is an often used parametric model for loss random variables. In this context, the parameter σ is treated as known, and the tail index parameter α is to be estimated from sample data. The recent paper of Brazauskas and Serfling (2000) reviewed many of the commonly used estimators of α, such as the maximum likelihood estimator, the method of moments estimator, trimmed mean estimators, regression-based estimators, least squares estimators, and estimators based on quantile matching. The deficiencies of these commonly used estimators were identified, and a new generalized median estimator for α was constructed that was shown to have superior resistance to the effects of contaminated data while maintaining a reasonable level of efficiency. In the subsequent discussion of the Brazauskas and Serfling paper, Bilodeau (2001a) introduced two additional estimators and expanded the comparison. Victoria-Feser and Ronchetti (1994) also developed an estimator of α with some optimality properties by following ideas of Hampel et al. (1986).

Here another estimator of α is presented that is shown to provide comparable performance to these other estimators, while being both conceptually and computationally simpler. Indeed, the new estimator with desired performance characteristics can be designed and computed using readily available spreadsheet software, at least for sample sizes up to 1000 or so. This new estimator is based on the probability integral transform. The methodology used in computing the estimator is also easily used to find confidence intervals for α that provably have the nominal level of confidence even for small sample sizes.

The comparisons given here also serve to make clear the fact that nonrobust estimation of the parameter α is extremely dangerous when even small amounts of contamination are present. The use of a robust estimator therefore should become standard practice.

2. Criteria for Efficiency and Robustness

A more detailed discussion of the concepts for efficiency and robustness discussed here can be found in Huber (1981) or Hampel et al. (1986).

Suppose $X_1, \ldots, X_n$ is a random sample on the Pareto distribution with a known scale parameter

* Mark Finkelstein is an Associate Professor is in the Department of Mathematics, 103 Multipurpose Science and Technology Building, University of California, Irvine, Irvine, CA 92697-3875, mfinkels@math.uci.edu.

† Howard G. Tucker is a Professor in the Department of Mathematics, 103 Multipurpose Science and Technology Building, University of California, Irvine, Irvine, CA 92697-3875.

‡ Jerry Alan Veeh is a Professor in the Department of Mathematics and Statistics, 232 Parker Hall, Auburn University, Auburn, AL 36849-5310.

σ and an unknown tail index $\alpha > 0$. The maximum likelihood estimator $\hat{\alpha}_{ML,n}$ based on this sample of size n is computed easily to be

$$\hat{\alpha}_{ML,n} = \frac{1}{n^{-1} \sum_{j=1}^{n} \ln(X_i) - \ln(\sigma)}.$$

The maximum likelihood estimator has two desirable properties. First, $\hat{\alpha}_{ML,n}$ is asymptotically consistent in the sense that $\hat{\alpha}_{ML,n} \to \alpha$ as $n \to \infty$ with probability 1. Second, the variance of $\hat{\alpha}_{ML,n}$ is asymptotic to α^2/n as $n \to \infty$, and this variance is asymptotically the smallest variance of all unbiased estimators of α, as is seen by computing the Cramér-Rao lower bound. Any other proposed estimator $\hat{\alpha}_n$ of α certainly should share the asymptotic consistency of the maximum likelihood estimator. Also the ratio $\mathrm{Var}(\hat{\alpha}_{ML,n})/\mathrm{Var}(\hat{\alpha}_n)$ should have a limit as close to one as possible. This limit is called the *asymptotic relative efficiency* of $\hat{\alpha}_n$ and measures the relative accuracy of the estimator $\hat{\alpha}_n$ compared to the maximum likelihood estimator based on the same sample size. Because of the optimal quality of the maximum likelihood estimator, the asymptotic relative efficiency of any other estimator cannot exceed 1.

The efficiency of the maximum likelihood estimator, and of other estimators as well, is bought at a price. That price is the sensitivity of the estimator to contamination of the sample. A simple way of measuring sensitivity to contamination is by means of the breakdown point of the estimator. In the Pareto setting, the most severe types of contamination can be idealized as occurring when one or more observations tend to the values σ or infinity.

Suppose that for a sample of size n the integer $1 \le k_n \le n$ is the smallest integer with the property that sending k_n of the observations to infinity forces the estimator $\hat{\alpha}_n \to 0$. The fraction k_n/n is called the *finite sample upper breakdown point* of the estimator $\hat{\alpha}_n$. The *upper breakdown point* is the limit $\lim_{n\to\infty} k_n/n$ of the finite sample upper breakdown points.

In a similar way, if $1 \le k_n \le n$ is the smallest integer with the property that sending k_n of the observations to σ forces $\hat{\alpha}_n \to \infty$, then k_n/n is the *finite sample lower breakdown point* of the estimator $\hat{\alpha}_n$. The *lower breakdown point* is the limit of the finite sample lower breakdown points as n tends to infinity.

For all of the estimators examined here, the sum of the upper breakdown point and lower breakdown point is unity. For this reason, and because in insurance applications upper breakdown is of greater interest, only upper breakdown points will be discussed. An estimator with a high upper breakdown point should be robust under contamination of the sample by unusually large values.

The earlier formula for the maximum likelihood estimator shows that the finite sample upper breakdown point is $1/n$, so that the upper breakdown point is 0. The objective is to develop an estimator with a higher upper breakdown point, while not paying a high price in terms of efficiency.

Another useful measure of robustness is *gross error sensitivity,* which intuitively measures the maximum impact on the estimator of an arbitrary change of a *single* observation, when n is large. Ideally, the effect of changes in a single observation should have minimal impact on the estimator. An estimator with small gross error sensitivity should be more robust than an estimator with larger gross error sensitivity. A technical description of gross error sensitivity can be found in Hampel et al.

3. A Probability Integral Transform Statistic

The motivation for the new estimator stems from the fact that since the distribution function F of the Pareto distribution is continuous and strictly increasing, the random variables $F(X_1), \ldots, F(X_n)$ form a random sample on the uniform distribution on the interval (0,1). With an eye toward the contamination issue, notice that any infinite observation X_j transforms to the value 1. Thus even infinite contamination has a bounded effect on the transformed data. This observation will be used to construct a family of estimators of α. Define

$$G_{n,t}(\beta) = n^{-1} \sum_{j=1}^{n} \left(\frac{\sigma}{X_j}\right)^{\beta t},$$

where $t > 0$ is a tuning parameter that later will be used to adjust the balance between efficiency and breakdown point. To maintain intuition, notice that when $\beta = \alpha$, $(\sigma/X_j)^{\alpha} = 1 - F(X_j)$ is a random variable with the uniform distribution.

Denote by $U_1, \ldots, U_n$ a random sample from the uniform distribution. Then $G_{n,t}(\beta)$ behaves probabilistically like $n^{-1} \sum_{j=1}^{n} U_j^{\beta t/\alpha}$, and when $\beta = \alpha$ this sum should behave like $n^{-1} \sum_{j=1}^{n} U_j^t$. The probabilistic behavior of this last sum does not depend on any unknown parameters. Values of β for which $G_{n,t}(\beta)$ behaves probabilistically like this last sum therefore must be values of β that are near α.

To turn this intuition into a practical tool, a standard of measuring similar probabilistic behavior must be used. The Strong Law of Large Numbers shows that $n^{-1} \sum_{j=1}^{n} U_j^t \to \mathrm{E}[U^t] = 1/(t + 1)$ as $n \to \infty$ with probability 1. So borrowing from the method of moments, the new estimator $\hat{\alpha}_n$ is defined to be the solution of the equation

$$G_{n,t}(\beta) = \frac{1}{t + 1}.$$

Lemma 1 below shows that this equation has exactly one solution for any fixed $t > 0$. The discussion following Lemma 1 shows that the bisection method can be used to easily compute the value of the estimator once the data are given.

As shown in Theorem 1 below, the new estimator $\hat{\alpha}_n \to \alpha$ as $n \to \infty$ with probability 1. Theorem 2 shows that the upper breakdown point of $\hat{\alpha}_n$ is $t/(t + 1)$. So the upper breakdown point approaches 0 as t approaches 0.

As shown following Theorem 3 below, the asymptotic relative efficiency of $\hat{\alpha}_n$ is $(2t + 1)/(t + 1)^2$. By taking t positive and near 0, the relative efficiency of $\hat{\alpha}_n$ can be made arbitrarily close to 1.

Theorem 4 shows that gross error sensitivity is $\alpha \max\{1 + 1/t, 1 + t\}$.

Taken together, the facts in the preceding paragraphs show how the tradeoff between efficiency, upper breakdown point, and gross error sensitivity is controlled by the tuning parameter t: for t near 0 the estimator is efficient but has a low upper breakdown point and high gross error sensitivity, while for large t the estimator loses efficiency but also becomes more resistant to upper contamination, while again increasing gross error sensitivity.

Because of the simple form of the formulas above, the value of t corresponding to the desired breakdown point, efficiency, or gross error sensitivity can be found easily. Once t is known, the value of the estimator for a given data set can be computed easily using the goal seek tool available in spreadsheet software, at least for sample sizes up to 1000 or so. The other estimators considered here are not so easily designed or computed.

4. COMPARISONS

The probability integral transform statistic (PITS) now will be compared with several other estimators. An informal description of the estimators is given here. A more detailed technical description is provided in the Appendix.

Brazuaskas and Serfling (2000) developed a generalized median estimator (GM) that is, for a sample of size n, the appropriately scaled median of the maximum likelihood estimator computed for all subsamples of size k of the given sample of size n. They presented compelling evidence that the generalized median estimator is superior to the maximum likelihood estimator, regression estimators, least squares estimators, method of moments estimators, and estimators based on quantile matching. Consequently, of these estimators only the generalized median estimator will be examined further here. The maximum likelihood estimator (MLE) will be used as a reference estimator. Notice that the generalized median estimator depends on the choice of the integer parameter k, which acts as a tuning parameter to control the tradeoff between efficiency and robustness.

Bilodeau (2001a, 2001b) developed two estimators that also have compelling properties for consideration. His M estimator (BM) depends on a tuning parameter ε that controls the balance between efficiency and robustness, as measured by upper breakdown point. Bilodeau's CM estimator (BCM) depends on ε and C, with the constant C allowing, in some cases, greater efficiency for a given upper breakdown point.

Victoria-Feser and Ronchetti (1994) explore the properties of the standardized optimal bias robust estimator (SOBRE). This estimator gives the highest efficiency for a given value of the standardized gross error sensitivity. Hampel et al. (1986) also described an unstandardized optimal bias robust estimator (UOBRE), which gives the highest efficiency for a given unstandardized gross error sensitivity. Each of these estimators depends on the choice of the bound C placed on the respective gross error sensitivity.

A sensible way of comparing the new estimator given here to these other estimators is to compare the upper breakdown point and gross error sensitivity when the asymptotic relative efficiency of all of the estimators is the same. Since the generalized median estimator is determined by an integer tuning parameter k, the other estimators will be adjusted to match the relative efficiency of the generalized median estimator for $2 \leq k \leq 4$. The reason for this seemingly narrow range of k values will become apparent in the subsequent discussion.

The comparison in Table 1 shows that BCM has the highest upper breakdown point, while SOBRE and UOBRE have the lowest gross error sensitivity. With the exception of SOBRE, the differences in gross error sensitivity are not large; with the exception of GM and BCM, the differences in upper breakdown point also are not large. Recall that for all of the estimators, the sum of the upper and lower breakdown points is 1. So with the exception of GM and BCM, the lower breakdown points are also quite similar.

Table 1

Comparison of Estimators Asymptotic Relative Efficiency (ARE), Upper Breakdown Point (UBP), and Gross Error Sensitivity (GES)

		ARE	UBP	GES
PITS	t:			
	0.883	0.78	0.469	2.13α
	0.531	0.88	0.347	2.88α
	0.394	0.92	0.283	3.54α
	0.324	0.94	0.245	4.08α
GM	k:			
	2	0.78	0.293	1.90α
	3	0.88	0.206	2.27α
	4	0.92	0.159	2.60α
	5	0.94	0.129	2.88α
BM	ε:			
	0.475	0.78	0.475	1.94α
	0.370	0.88	0.370	2.50α
	0.314	0.92	0.314	2.96α
	0.279	0.94	0.279	3.34α
BCM	ε, C:			
	0.50, 5.31	0.78	0.500	2.23α
	0.45, 6.35	0.88	0.450	2.59α
	0.40, 7.23	0.92	0.400	2.94α
	0.40, 7.95	0.94	0.400	3.23α
UOBRE	C:			
	1.84	0.78	0.451	1.84α
	2.27	0.88	0.359	2.27α
	2.58	0.92	0.314	2.58α
	2.82	0.94	0.287	2.82α
SOBRE	C:			
	1.63	0.78	0.449	0.583
	2.13	0.88	0.358	1.20
	2.48	0.92	0.314	1.68
	2.73	0.94	0.288	2.02

Source: Data for estimators are from Brazauskas and Serfling (2000), Bilodeau (2001a), and computations.

Given the closeness of these measures for all of the estimators, a natural question is whether the differences that do exist are of practical significance. All of these measures of efficiency and robustness are asymptotic in nature, so meaningful small sample comparisons are of interest.

Following a suggestion of Hampel et al., the behavior of the estimators will be examined on four small artificial data sets. All of the estimators considered here can be expressed easily in terms of the variables $(\sigma/X_j)^\alpha$, so the behavior of the ratio $\hat{\alpha}_n/\alpha$ can be examined readily. This frees the analysis from dependence on the true values of both α and σ.

The four data sets are each a sample of size 20 and were constructed as follows:

Sample I. Divide the Pareto distribution into 21 equiprobable intervals. The data points are the 20 quantiles so determined.

Sample II. The first data set was altered by increasing the largest datum by a factor of $10^{1/\alpha}$ and making the second largest datum equal to the former value of the largest datum.

Sample III. The largest two data of the original sample were each increased by a factor of $10^{1/\alpha}$, and the next two data were set equal to the former value of the largest two data.

Sample IV. Data set II was modified by decreasing the smallest datum by a factor of $10^{1/\alpha}$ and making the second smallest datum equal to the former value of the smallest datum.

Thus sample I is an ideal Pareto sample, and an ideal estimator should yield a value for $\hat{\alpha}_n/\alpha$ of 1. Sample II represents a small amount of upper contamination, while sample III represents a moderate amount of upper contamination. Sample IV represents a small amount of both upper and lower contamination.

Much of the literature on robust estimation suggests that robustness can be obtained with only a small sacrifice in efficiency. Table 2 shows that none of the estimators with 94% efficiency can cope with the corruption of data set III, since all provide only marginal improvement over the 20% error of the MLE in this case. Notice that this occurs even though the corruption is far lower than the upper breakdown point for all but

Table 2
Comparison of Estimators with 94% Efficiency on Artificial Samples of Size 20

	I	II	III	IV
PITS	1.041	0.963	0.854	0.970
GM	1.028	0.966	0.771	0.970
BM	1.041	0.965	0.852	0.972
BCM	1.042	0.964	0.848	0.972
UOBRE	1.040	0.984	0.842	0.990
SOBRE	1.040	0.984	0.843	0.990
MLE	1.078	0.928	0.802	0.932

Notes: See text for description of samples. Value of an ideal estimator is 1 in all cases.

Table 3
Comparison of Estimators with 92% Efficiency on Artificial Samples of Size 20

	I	II	III	IV
PITS	1.035	0.968	0.865	0.976
GM	1.023	0.986	0.807	0.990
BM	1.035	0.972	0.864	0.980
BCM	1.037	0.971	0.858	0.979
UOBRE	1.032	0.988	0.868	0.993
SOBRE	1.032	0.988	0.867	0.993
MLE	1.078	0.928	0.802	0.932

Notes: See text for description of samples. Value of an ideal estimator is 1 in all cases.

Table 4
Comparison of Estimators with 88% Efficiency on Artificial Samples of Size 20

	I	II	III	IV
PITS	1.025	0.975	0.886	0.986
GM	1.018	1.012	0.868	1.015
BM	1.025	0.984	0.890	0.993
BCM	1.027	0.984	0.878	0.994
UOBRE	1.025	1.013	0.903	1.020
SOBRE	1.025	1.013	0.903	1.020
MLE	1.078	0.928	0.802	0.932

Notes: See text for description of samples. Value of an ideal estimator is 1 in all cases.

Table 5
Comparison of Estimators with 78% Efficiency on Artificial Samples of Size 20

	I	II	III	IV
PITS	1.005	0.982	0.926	0.998
GM	1.012	1.012	1.012	1.012
BM	1.005	0.994	0.943	1.007
BCM	1.006	1.016	0.941	1.030
UOBRE	1.008	1.008	0.987	1.017
SOBRE	1.008	1.008	0.986	1.017
MLE	1.078	0.928	0.802	0.932

Notes: See text for description of samples. Value of an ideal estimator is 1 in all cases.

the GM estimator. The UOBRE and SOBRE estimators have a slight edge for the small amounts of corruption of samples II and IV.

Tables 3, 4, and 5 exhibit the same general pattern of behavior for asymptotic relative efficiencies of 92%, 88%, and 78%. Acceptable performance for sample II is attained only when estimators with 88% efficiency are used; acceptable performance for sample III requires a reduction to estimators with 78% efficiency.

In summary,

1. Using the maximum likelihood estimator is a dangerous practice. Even the small amount of contamination of sample II causes a significant deterioration of performance. The maximum likelihood estimator provides no protection against significant deviations from the model.
2. Insisting on high efficiency is a dangerous practice. None of these estimators provides adequate protection against contamination in the 94% efficiency case. Reasonable protection can be attained only with efficiency of 88% or less.
3. The upper breakdown point provides little information of value about the robustness of these estimators.
4. All of the estimators behave similarly for efficiencies below 88%.

In view of item 4, using PITS makes sense because of its conceptual simplicity and computational ease.

5. CONFIDENCE INTERVALS

Since all of the estimators considered here are asymptotically normal with mean α and known variance, confidence intervals for α with prescribed nominal confidence level could be obtained based on this asymptotic normality. However, the actual confidence level of such intervals could deviate widely from the nominal level. The results in Finkelstein, Tucker, and Veeh (2000), coupled with the earlier intuition for the present estimator, provide an easy way to find confidence intervals for α that *provably* have the nominal level of confidence no matter the sample size. The results of that paper apply since $G_{n,t}(\beta)$ is a monotone decreasing function of β.

To see how such intervals are found, consider again the intuition developed earlier. The random variable $G_{n,t}(\beta)$ behaves probabilistically like $n^{-1} \Sigma_{j=1}^{n} U_j^{\beta t/\alpha}$, and when $\beta = \alpha$ this sum should behave like $n^{-1} \Sigma_{j=1}^{n} U_j^t$. The probabilistic behavior of this last sum does not depend on any unknown parameters. Values of β for which $G_{n,t}(\beta)$ behaves probabilistically like this last sum therefore must be values of β that are near α. A two-sided 95% confidence interval for α then could be found in the following way. For the given value of t, find the $2\frac{1}{2}$ and $97\frac{1}{2}$ percentiles, $\pi_{0.025}$ and $\pi_{0.975}$, of the distribution of $n^{-1} \Sigma_{j=1}^{n} U_j^t$. Because this distribution does not depend on any unknown parameters, these percentiles could be determined accurately by simulation. If R satisfies $G_{n,t}(R) = \pi_{0.025}$ and L satisfies $G_{n,t}(L) = \pi_{0.975}$, then $[L, R]$ is a confidence interval for α that has confidence level 95%.

Keep in mind that the discussion in this section did not consider the effect of contamination, which is the focus of the next section.

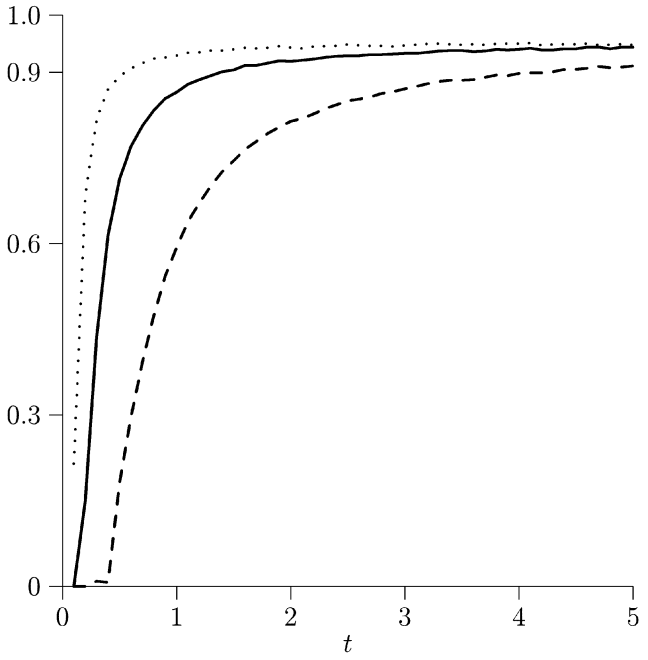
6. GUIDELINES FOR PRACTICAL APPLICATION

In view of the preceding theory, how might one begin with the data and select an appropriate value of the tuning parameter to carry out an analysis? The selection of the tuning parameter is a central problem for all of these estimators and is influenced by the amount of contamination present in the data. A quantile-quantile plot provides some useful information about the presence and amount of contamination and serves as a qualitative check on the appropriateness of the Pareto model.

The quantile-quantile plot is based on the following informal reasoning. The $100p$-th percentile of a Pareto distribution is the solution, x, of the equation $F(x) = p$, and hence the $100p$-th percentile is $\sigma(1 - p)^{-1/\alpha}$. Now, the true percentiles of a distribution may be estimated using the sample percentiles and, in particular, the order statistics of the sample. Denote by $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ the observations arranged in increasing order. One may consider $X_{(j)}$ to be a crude estimate of the $100j/(n + 1)$ percentile. From the crude approximation $X_{(j)} \approx \sigma(1 - j/(n + 1))^{-1/\alpha}$, the points $(-\ln(1 - j/(n + 1)), \ln(X_{(j)}/\sigma))$, $1 \leq j \leq n$, should lie approximately on a straight line with slope $1/\alpha$. If this is approximately true for the data at hand, the assumption that the data come from some Pareto distribution may be considered reasonable. Further, points lying substantially above the line for j near n would represent observations that are unusually large, that is, upper contamination. This plot also provides a rough estimate of the contaminating fraction f of the sample, at least in the case of a somewhat large sample.

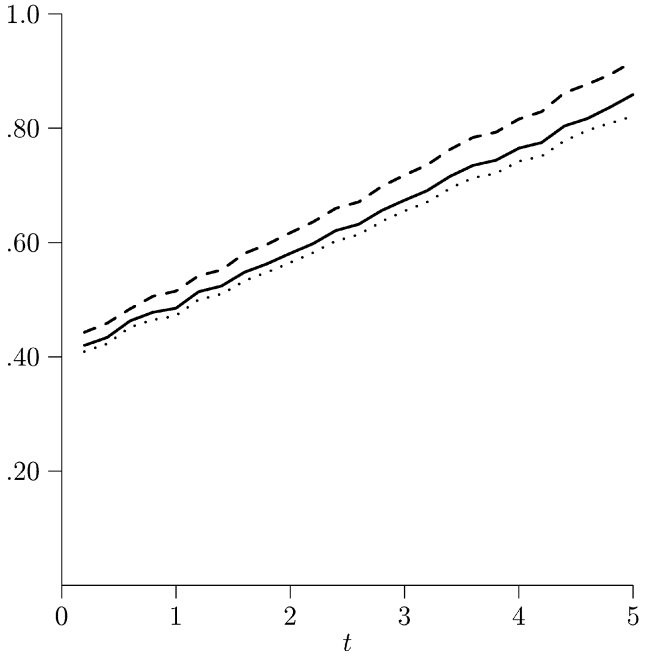
As mentioned above, the choice of the tuning parameter determines a tradeoff between efficient estimation of the unknown parameter α and robustness of that estimate. If the objective of the analysis is to obtain a (nominal) 95% confidence interval for α, this tradeoff manifests itself in the actual confidence level of the interval obtained and the length of that confidence interval. As shown below, the actual coverage probability of a nominal 95% confidence interval can be approximated easily as a function of f, t, and n. Figure 1 shows such a plot for $n = 100$ and three contamination levels. If the preliminary analysis sug-

Figure 1
Asymptotic Coverage Probability as a Function of the Tuning Parameter



Note: The nominal coverage probability is 95%. The levels of contamination are 2.5% (*dotted line*), 5% (*solid line*), and 10% (*dashed line*).

Figure 2
Expected Length of a Nominal 95% Confidence Interval



Note: Plot shows expected length divided by α computed when $n = 100$. The levels of contamination are 2.5% (*dotted line*), 5% (*solid line*), and 10% (*dashed line*).

gested about 5% contamination, a practitioner who wanted to retain at least a 90% confidence level would choose a tuning parameter t of at least 2. The downside of choosing larger values of t is seen in Figure 2, which plots the expected length of the confidence interval divided by the (unknown) α. Each unit increase in t is seen to increase the expected length of the confidence interval by about $\alpha/10$. In the present scenario, the choice of $t = 2$ thus represents a reasonable tradeoff between efficiency and robustness.

APPENDIX

THEOREMS AND PROOFS

Ideas from calculus are applied to show that the estimator $\hat{\alpha}_n$ does in fact exist.

Lemma 1

In the notation above, for any fixed $t > 0$ the equation $G_{n,t}(\beta) = 1/(t + 1)$ has exactly one solution.

PROOF

Since $G_{n,t}(0) = 1 > 1/(t + 1)$ and $\lim_{\beta\to\infty} G_{n,t}(\beta) = 0 < 1/(t + 1)$, the Intermediate Value Theorem shows that the equation has at least one solution. The derivative $G'_{n,t}(\beta) = n^{-1} \sum_{j=1}^{n} (\sigma/X_j)^{\beta t} t \ln(\sigma/X_j)$ is negative since each of the ratios $\sigma/X_j < 1$. Thus the equation has only one solution.

The proof of the lemma suggests a simple way of actually computing $\hat{\alpha}_n$ given a data set. Since $G_{n,t}(0) > 1/(t + 1)$, first search for a value of β for which $G_{n,t}(\beta) < 1/(t + 1)$ by simply evaluating $G_{n,t}$ at successive integers. Once such a value of β is found, the bisection method is applied to find the solution $\hat{\alpha}_n$ of the equation.

All of the estimators considered here except for GM belong to the class of M estimators. A brief review of the basic properties of M estimators is given here in the context of estimating the Pareto parameter α.

Motivation for M estimators comes from the maximum likelihood estimators. If $\psi_{\mathrm{MLE}}(x, \alpha) = d/d\alpha \ln (d/dx) F(x)$ is the log likelihood for the Pareto distribution, the maximum likelihood estimator of α is the value $\hat{\alpha}_{\mathrm{ML},n}$ satisfying

$$\sum_{j=1}^{n} \psi_{\mathrm{MLE}}(X_j, \hat{\alpha}_{\mathrm{ML},n}) = 0.$$

The other M estimators here arise as the solution of the same equation, but with a different choice of ψ. The PITS estimator is the M estimator corresponding to the choice $\psi_{\mathrm{PITS}}(x, \beta) = (\sigma/x)^{t\beta} - 1/(t + 1)$. A detailed discussion of the other estimators is given below.

Huber (1981) establishes several properties of M estimators. A central role in his theory is played by the function $\lambda(\beta) = \int_{\sigma}^{\infty} \psi(x, \beta) dF(x)$. For the PITS estimator direct computation gives $\lambda(\beta) = t(\alpha - \beta)/(\alpha + t\beta)(t + 1)$.

One of the desirable properties of any estimator is consistency.

Theorem 1

For any fixed $t > 0$, the estimator $\hat{\alpha}_n$ converges to α as $n \to \infty$ with probability 1.

PROOF

For the PITS estimator, $\lambda(\beta) > 0$ if $\beta < \alpha$ and $\lambda(\beta) < 0$ if $\beta > \alpha$. Since the PITS estimator is uniquely defined, consistency follows from Prop-

osition 2.1 and Corollary 2.2 of Chapter 3 of Huber's book. This completes the proof.

The breakdown points of the PITS estimator are not easily obtained from Huber's arguments. A direct argument establishes formulas for the upper and lower breakdown points. Denote by $\lceil x \rceil$ the smallest integer greater than or equal to x.

Theorem 2

The finite sample upper breakdown point is $\lceil nt/(t + 1) \rceil/n$, and the finite sample lower breakdown point is $\lceil n/(t + 1) \rceil/n$. The upper breakdown point is $t/(t + 1)$, and the lower breakdown point is $1/(t + 1)$.

Proof

The defining equation for $\hat{\alpha}_n$ gives $1/(t + 1) = n^{-1} \sum_{j=1}^{n} (\sigma/X_j)^{t\hat{\alpha}_n} = n^{-1} \sum_{j=1}^{K} (\sigma/X_j)^{t\hat{\alpha}_n} + n^{-1} \sum_{j=K+1}^{n} (\sigma/X_j)^{t\hat{\alpha}_n}$ for any integer $1 \le K \le n$. The effect on $\hat{\alpha}_n$ of taking $X_1, \ldots, X_K$ to infinity is to drive $\hat{\alpha}_n$ to the solution β of the equation $1/(t + 1) = n^{-1} \sum_{j=K+1}^{n} (\sigma/X_j)^{t\beta}$, and this solution will be positive if and only if $(n - K)/n > 1/(t + 1)$, that is, $K < nt/(t + 1)$. This proves the assertion about the finite sample upper breakdown point. The assertion about the upper breakdown point follows by letting $n \to \infty$ in the finite sample upper breakdown point formula. Similarly, the effect on $\hat{\alpha}_n$ of taking $X_1, \ldots, X_K$ to σ is to drive $\hat{\alpha}_n$ to the solution of the equation $1/(t + 1) = K/n + n^{-1} \sum_{j=K+1}^{n} (\sigma/X_j)^{t\beta}$, and this solution will be finite if and only if $K/n < 1/(t + 1)$, that is, $K < n/(t + 1)$. The finite sample lower breakdown point is therefore $\lceil n/(t + 1) \rceil/n$, and the lower breakdown point follows by letting $n \to \infty$.

To compute the asymptotic relative efficiency of $\hat{\alpha}_n$ the asymptotic distribution of $\hat{\alpha}_n$ is found.

Theorem 3

For any fixed $t > 0$, the estimator $\sqrt{n}\,(\hat{\alpha}_n - \alpha)$ is asymptotically normal with mean 0 and variance $\alpha^2(t + 1)^2/(2t + 1)$.

Proof

Direct computation gives $\lambda'(\alpha) = -t/\alpha(t + 1)^2 < 0$, and Huber's auxillary function $\sigma_0^2 = \int_\sigma^\infty \psi_{\text{PITS}}(x, \alpha)^2 dF(x) - \lambda(\alpha)^2 = t^2/(t + 1)^2(2t + 1)$. Application of Huber's Corollary 2.5 of Chapter 3 establishes the result.

Since the variance of the maximum likelihood estimator is α^2/n, the asymptotic relative efficiency of $\hat{\alpha}_n$ is $(2t + 1)/(t + 1)^2$. By taking t positive and near 0 the relative efficiency of $\hat{\alpha}_n$ can be made arbitrarily close to 1.

Theorem 4

The gross error sensitivity is $\alpha \max\{1 + 1/t, 1 + t\}$.

Proof

Applying the general formula for the influence curve of M estimators given in equation (2.13) of Chapter 3 of Huber gives the gross error sensitivity of the PITS estimator as

$$\sup_{x \ge \sigma} \left| \psi_{\text{PITS}}(x, \alpha) \Big/ \int_\sigma^\infty \frac{\partial}{\partial \alpha} \psi_{\text{PITS}}(y, \alpha) dF(y) \right| .$$

The denominator integral is computed easily to be $t/\alpha(t + 1)^2$. Since ψ is monotone in x, the maximum value occurs either when $x = \sigma$ or as $x \to \infty$. Making these substitutions gives the desired formula.

A brief technical description of the other M estimators considered here will be given now. A detailed discussion of the GM estimator can be found in Brazauskas and Serfling (2000).

The idea behind both of the optimal biased robust estimators UOBRE and SOBRE is that the estimator should be designed to have a preassigned bound on the gross error sensitivity. The unstandardized gross error sensitivity for an M estimator is given by

$$\sup_{x \ge \sigma} \left| \psi(x, \alpha) \Big/ \int_\sigma^\infty \frac{\partial}{\partial \alpha} \psi(y, \alpha)\, dF(y) \right|$$

in terms of the psi function defining the estimator (see Huber 1981, ch. 3). The UOBRE estimator is designed to make the gross error sensitivity $C\alpha$, for some user-selected constant C. This is accomplished by choosing

$$\psi_{\text{UOBRE}}(x, \beta) = \frac{C}{\beta} H\left(\frac{A}{C} \left((1 + \ln(\sigma/x)^\beta) - \alpha \right) \right),$$

where A and α are chosen so that the two conditions

$$\int_0^1 H\left(\frac{A}{C} (1 + \ln u - \alpha) \right) du = 0$$

and

$$C \int_0^1 H\left(\frac{A}{C}(1 + \ln u - a)\right)(1 + \ln u)du = 1$$

are satisfied. The two side conditions are solved numerically for A and α given a value of C. The function H is defined by the formula

$$H(y) = \begin{cases} -1 & y \le -1 \\ y & -1 \le y \le 1 \\ 1 & y \ge 1. \end{cases}$$

The SOBRE estimator is designed to satisfy a bound on the standardized gross error sensitivity, which in the Pareto case is given by

$$\sup_x |\psi(x, \alpha)| \Big/ \left(\int_\sigma^\infty \psi^2(x, \alpha)\, dF(x)\right)^{1/2}.$$

If the standardized gross error sensitivity is bounded by a user-selected constant C, the resulting M estimator has

$$\psi_{\text{SOBRE}}(x, \beta) = \frac{C}{A\beta} H\left(\frac{A}{C}(1 + \ln(\sigma/x)^\beta - a)\right),$$

where A and a are constants chosen so that

$$\int_0^1 H\left(\frac{A}{C}(1 + \ln(u) - a)\right) du = 0$$

and

$$C^2 \int_0^1 H^2\left(\frac{A}{C}(1 + \ln(u) - a)\right) du = 1.$$

The constants A and a are found numerically once a value of C is selected.

The asymptotic properties of these estimators can be found directly using the theory developed by Huber and proceding along the lines given in the earlier proofs for the PITS estimator. Further details of the construction of OBRE-type estimators may be found in Hampel (1986).

Bilodeau noticed that if X is Pareto, then $(\sigma/X)^\alpha$ is uniform, and thus $-\ln((\sigma/X)^\alpha)$ has the exponential distribution with mean 1. Let $\mathfrak{E}$ denote an exponential random variable with mean 1. Then $(1/\alpha)\mathfrak{E} = -\ln(\sigma/X)$, and the problem of estimating α is equivalent to the problem of estimating the scale factor $\theta = 1/\alpha$ for the exponential random variable. Since $\hat{\theta} = 1/\hat{\alpha}_n$, properties of the associated estimator $\hat{\alpha}_n = 1/\hat{\theta}$ can be found easily from those of $\hat{\theta}$ by using transformation rules for influence functions and asymptotic variances, together with the delta method.

Bilodeau constructed two estimators by making use of the auxillary function

$$\rho(x) = \begin{cases} 0 & x < 0 \\ 3x - 3x^2 + x^3 & 0 \le x \le 1 \\ 1 & x \ge 1. \end{cases}$$

His first estimator is an M estimator constructed as follows. The notation used here differs from that used in his paper.

Let $0 < \varepsilon < 1$ be given. Let λ_L be the unique value of λ for which $E[\rho(\mathfrak{E}/\lambda)] = \varepsilon$. The estimator BM then is defined using the psi function

$$\psi_{\text{BM}}(x, \theta) = \rho(x/\lambda_L\theta) - \varepsilon.$$

The upper breakdown point is ε, and the lower breakdown point is $1 - \varepsilon$. Keep in mind that the estimator found using this psi function and the observations $\ln(\sigma/X_i)$ is the estimator of the scale factor θ. The gross error sensitivity of $\hat{\alpha}_n = 1/\hat{\theta}$ is

$$\alpha \frac{\max\{\varepsilon, 1 - \varepsilon\}}{E[(\mathfrak{E}/\lambda_L)\rho'(\mathfrak{E}/\lambda_L)]},$$

and the asymptotic variance of $\sqrt{n}(\hat{\alpha}_n - \alpha)$ is

$$\alpha^2 \frac{E[\rho^2(\mathfrak{E}/\lambda_L)] - \varepsilon^2}{(E[(\mathfrak{E}/\lambda_L)\rho'(\mathfrak{E}/\lambda_L)])^2}.$$

Bilodeau built up a constrained M estimator using the BM estimator as the key building block. Let ε and λ_L be defined as above, and let $C > 0$ be arbitrary. Define λ_0 as a value of λ that minimizes the function $E[C\rho(\mathfrak{E}/\lambda)] + \ln(\lambda)$ on the interval $[\lambda_L, \infty)$.

If $\lambda_0 = \lambda_L$, the value of λ that gives equality in the constraint inequality is the value of λ at which the minimum is attained. Thus the CM estimator is the M estimator described earlier when $\lambda_0 = \lambda_L$.

If $\lambda_0 > \lambda_L$, the minimum of the objective function must be attained at a point where the derivative of the objective function with respect to λ is 0. So in this case the CM estimator corresponds to the psi function $\psi_{\text{BCM}}(x, \theta) = 1 - (Cx/\lambda_0\theta)\rho'(x/\lambda_0\theta)$. Again, using this psi function with the data $\ln(\sigma/X_i)$ gives an estimate of the scale factor θ. The gross error sensitivity of $\hat{\alpha}_n = 1/\hat{\theta}$ is

$$\frac{\alpha \max\{1, |1 - 4C/9|\}}{C\, E[(\mathfrak{E}/\lambda_0)^2\rho''(\mathfrak{E}/\lambda_0) + (\mathfrak{E}/\lambda_0)\rho'(\mathfrak{E}/\lambda_0)]},$$

and the asymptotic variance of $\sqrt{n}(\hat{\alpha}_n - \alpha)$ is

$$\frac{\alpha^2\,\mathrm{E}[(1-(C\mathfrak{E}/\lambda_0)\rho'(\mathfrak{E}/\lambda_0))^2]}{C^2\,(\mathrm{E}[(\mathfrak{E}/\lambda_0)^2\rho''(\mathfrak{E}/\lambda_0)+(\mathfrak{E}/\lambda_0)\rho'(\mathfrak{E}/\lambda_0)])^2}.$$

For some values of ε there may be a value of C for which the BCM estimator has a higher upper breakdown point or greater efficiency than the BM estimator with the same ε. Indeed, Table 1 shows that the BCM estimators have a higher upper breakdown point while maintaining the same relative efficiency, and about the same gross error sensitivity.

As shown in a more general setting in Finkelstein, Tucker, and Veeh (2000), if the $2\frac{1}{2}$ and $97\frac{1}{2}$ percentiles of the distribution of $n^{-1}\sum_{j=1}^{n} U_j^t$ are $\pi_{0.025}$ and $\pi_{0.975}$, then a 95% confidence interval for α is given by $\{\beta : \pi_{0.025} \le G_{n,t}(\beta) \le \pi_{0.975}\}$. If there is no contamination, this interval covers the true value of the unknown parameter α with probability 0.95. By making use of the Central Limit Theorem, an expression for the approximate coverage probability of this interval can be found even when there is contamination. This approximation will be valid only when the sample size n is large. When n is large, the average $n^{-1}\sum_{j=1}^{n} U_j^t$ has a distribution that is approximately normal. Thus the percentiles are approximately $\pi_{0.025} \approx \mathrm{E}[U^t] - 1.96\sqrt{\mathrm{Var}(U^t)/n}$ and $\pi_{0.975} \approx \mathrm{E}[U^t] + 1.96\sqrt{\mathrm{Var}(U^t)/n}$. Also, when nf of the observations are infinite, $G_{n,t}(\alpha) = n^{-1}\sum_{j=1}^{n(1-f)}(\sigma/X_j)^{t\alpha} = n^{-1}\sum_{j=1}^{n(1-f)} U_j^t$. When n is large, this random variable also has approximately a normal distribution with mean $(1-f)\mathrm{E}[U^t]$ and variance $(1-f)\mathrm{Var}(U^t)/n$. The coverage probability of the confidence interval is $P[\alpha \in \{\beta : \pi_{0.025} \le G_{n,t}(\beta) \le \pi_{0.975}\}] = P[\pi_{0.025} \le G_{n,t}(\alpha) \le \pi_{0.975}]$, and, making use of the two normal approximations, this last probability is approximately the probability that a standard normal random variable lies in the interval with endpoints $f\mathrm{E}[U^t]/\sqrt{(1-f)\mathrm{Var}(U^t)/n} \pm 1.96/\sqrt{1-f}$. Since the expectation and variance of U^t are easily computed, the approximate coverage probability also can be found easily for any given f and t as long as n is large.

The expected length of confidence intervals also can be found. The key observation is that the length of an interval is the same as the integral of the indicator function of the interval. Denote by $1_A(x)$ the indicator function of the set A. Let $A = \{\beta : \pi_{0.025} \le G_{n,t}(\beta) \le \pi_{0.975}\}$ be the confidence interval for α. The expected length of this interval is then $\mathrm{E}[\int_0^\infty 1_A(x)\,dx] = \int_0^\infty \mathrm{E}[1_A(x)]\,dx = \int_0^\infty P[\pi_{0.025} \le G_{n,t}(x) \le \pi_{0.975}]\,dx$. From here, the normal approximations used in the discussion of the coverage probability can be used to approximate this length as the integral of standard normal probabilities.

Doing this and making a simple change of variable gives the expected length as $\alpha\int_0^\infty P(x)dx$, where $P(x)$ is the probability that a standard normal random variable lies in the interval with endpoints $(\mathrm{E}[U^t] \pm 1.96\sqrt{\mathrm{Var}(U^t)/n} - (1-f)\mathrm{E}[U^{tx}])/\sqrt{(1-f)\mathrm{Var}(U^{tx})/n}$. This expression can be evaluated numerically for given f, t, and n.

REFERENCES

BILODEAU, MARTIN. 2001a. Discussions of Papers Already Published. *North American Actuarial Journal* 5(3): 123–28.

———. 2001b. Robust Estimation of the Tail Index of a Pareto Distribution. Technical Report, Université de Montréal.

BRAZAUSKAS, VYTARAS, AND ROBERT SERFLING. 2000. Robust and Efficient Estimation of the Tail Index of a Single-Parameter Pareto Distribution. *North American Actuarial Journal* 4(4): 12–27.

VICTORIA-FESER, MARIA-PIA, AND ELVEZZIO RONCHETTI. 1994. Robust Methods for Personal Income Distribution Models. *Canadian Journal of Statistics* 22(2): 247–58.

FINKELSTEIN, MARK, HOWARD G. TUCKER, AND JERRY ALAN VEEH. 2000. Conservative Confidence Intervals for a Single Parameter. *Communications in Statistics: Theory and Methods* 29(8): 1911–28.

HAMPEL, FRANK R., ELVEZZIO RONCHETTI, PETER J. ROUSSEEUW, AND WERNER A. STAHEL. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

HUBER, PETER J. 1981. *Robust Statistics*. New York: John Wiley.

Discussions on this paper can be submitted until July 1, 2006. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.