The Annals of Statistics
1999, Vol. 27, No. 2, 760–781

ESTIMATING A TAIL EXPONENT BY MODELLING DEPARTURE FROM A PARETO DISTRIBUTION

BY ANDREY FEUERVERGER AND PETER HALL

University of Toronto and University of Toronto and Australian National University

We suggest two semiparametric methods for accommodating departures from a Pareto model when estimating a tail exponent by fitting the model to extreme-value data. The methods are based on approximate likelihood and on least squares, respectively. The latter is somewhat simpler to use and more robust against departures from classical extreme-value approximations, but produces estimators with approximately 64% greater variance when conventional extreme-value approximations are appropriate. Relative to the conventional assumption that the sampling population has exactly a Pareto distribution beyond a threshold, our methods reduce bias by an order of magnitude without inflating the order of variance. They are motivated by data on extrema of community sizes and are illustrated by an application in that context.

1. Introduction. Estimating the tail exponent, or shape parameter, of a distribution is motivated by a particularly wide variety of practical problems, in areas ranging from linguistics to sociology and from hydrology to insurance. See, for example, Zipf (1941, 1949), Todorovic (1978), Smith (1984, 1989), NERC (1985), Hosking and Wallis (1987) and Rootzén and Tajvidi (1997). Many models and estimators have been proposed, including those of Hill (1975), Pickands (1975), de Haan and Resnick (1980), Teugels (1981), Csörgő, Deheuvels and Mason (1985) and Hosking, Wallis and Wood (1985). The goodness of fit of a Pareto model to the tail of a distribution can often be explored visually by simply plotting the logarithms of extreme order-statistics against the logarithms of their ranks. The plot should be approximately linear if the Pareto model applies over the range of those order statistics, and in this case the negative value of the slope of the line is an estimate of the Pareto exponent. More efficient estimators are available, however, and those in use today are generally of two types, based on likelihood and the method of moments, respectively.

Both these approaches are susceptible to errors in the assumed model for the distribution tail. As Rootzén and Tajvidi (1997) point out, none of the standard approaches is "robust against departures from the assumption that the tail of the distribution is approximated by a GP [generalized Pareto]

Received November 1997; revised February 1999.

Key words and phrases. Bias reduction, extreme-value theory, log-spacings, maximum likelihood, order statistics, peaks-over-threshold, regression, regular variation, spacings, Zipf's law.

distribution. If there are marked deviations from a GP tail, the results will be misleading." On the other hand, if one confines attention to data that are so far out in the tail that the Pareto assumption is valid, then the effective sample size can be small and the estimator of tail exponent may have relatively large variance.

Moreover, in some applications of the Pareto model, data that provide evidence of departure from the model are of greater interest than those to which the model fits well. A case in point is Zipf's (1941, 1949) classic study of the dynamics of community sizes. A linear plot of the logarithm of community size against the logarithm of rank, as in the case of U.S. cities during most of the twentieth century [see, e.g., Zipf (1941), Chapter 10; Hill (1975)] has been suggested as evidence of "stable intranational equilibrium"; while nonlinear plots, for example in the context of Austrian and Australian communities 60 to 80 years ago, have been interpreted as implying instability [Zipf (1949)]. As Zipf argues, it can be of greater sociological interest to analyze data from countries or eras that depart from the benchmark of "intranational equilibrium," than it is to analyze those which achieve the benchmark.

In this paper we propose a simple and effective way of reducing the bias that arises if one uses extreme-value data relatively deeply into the sample. We show that the main effects of bias may be accommodated by modelling the scale of log-spacings of order statistics. With this result in mind, we suggest two bias-reduction methods, both developed from a simple scale-change model for log-spacings, and based on likelihood and on least squares, respectively. In the first method, a three-parameter approximation to likelihood is suggested for the distribution of log-spacings. One of the parameters is the desired tail exponent. In our least-squares approach, log–log spacings play the role of response variables, the explanatory variables are ranks of order statistics and the regression "errors" have a type-3 extreme-value distribution, with exponentially light tails.

These techniques reduce bias by an order of magnitude, without affecting the order of variance. Therefore, they can lead to significant reductions in mean squared error. Moreover, they allow us to model extreme-value data that depart markedly from the standard asymptotic regime. This advantage will be illustrated in Section 3 by application to highly non-Pareto data on Australian community sizes. We are not aware of kernel or moment methods that are competitive with our approach; they would be awkward to construct in the context of a flexible and practicable class of distributions that represent departures from the Pareto model.

Statistical properties of estimators of tail exponents in Pareto models (sometimes referred to as models for Zipf's law) have been studied extensively. Early work in a parametric setting includes that of Hill (1970, 1974), Hill and Woodroofe (1975) and Weissman (1978). Smith (1985) points out the anomalous behaviour of estimators of certain exponents in the context of generalized Pareto distributions. Hall (1982, 1990) and Csörgő, Deheuvels and Mason (1985), among others, consider the effect of choice of threshold on performance of tail-exponent estimators. Rootzén and Tajvidi (1997) compare the performances of different approaches to tail-parameter estimation.

Davison (1984) and Smith (1984) provide statistical accounts of peaks-over-threshold, or POT, methods. Those techniques are of course not identical to methods based on extreme-order statistics, but the strong duality between the two approaches means, as Smith (1984) notes, that virtually all methods for one context have versions for the other. Our bias-correction methods are no exception. However, for the sake of brevity we shall discuss them only in the case of extreme order-statistics. Section 2 will introduce our methods, and Section 3 will describe their numerical properties. Theoretical performance will be outlined in Section 4, and technical arguments behind that work will be summarized in Section 5.

2. Methodology.

2.1. *Modelling the source of bias.* We shall introduce methodology in the case where the tail that is of interest is at the origin. Our methods extend immediately from there to the case of a tail at infinity. Suppose the distribution function F admits the approximation $F(x) \sim Cx^{\alpha}$ as $x \downarrow 0$, or more explicitly,

$$F(x) = Cx^{\alpha}\{1 + \delta(x)\}, \tag{2.1}$$

where C, α are positive constants and δ denotes a function that converges to 0 as $x \downarrow 0$. We wish to estimate α from a random sample $\mathscr{X} = \{X_1, \dots, X_n\}$ drawn from the distribution F. Often, one would proceed by assuming the particular Pareto model

$$F_0(x) = Cx^{\alpha}, \tag{2.2}$$

assumed for $0 < x < \varepsilon$, say, rather than (2.1), and alleviating bias problems caused by discrepancies between (2.1) and (2.2) by using only particularly small order-statistics from $\mathscr{X}$. However, this approach can have a detrimental effect on performance, since it ignores information about α that lies further into the sample. That information would be usable if we knew more about the function δ. In later work we shall refer to the model at (2.1) as a "perturbed" Pareto distribution, when it is necessary to distinguish it from the Pareto distribution at (2.2).

One approach to accessing the information is to model δ, for example in the fashion

$$\delta(x) = Dx^{\beta} + o(x^{\beta}) \tag{2.3}$$

as $x \to 0$, where $\beta > 0$ and $-\infty < D < \infty$ are unknown constants. In principle we could drop the "small oh" remainder term at (2.3), substitute the resulting formula into (2.1) and estimate the four parameters C, D, α, β by maximum likelihood. This means, in effect, working with the model

$$F_1(x) = C_1 x^{\alpha_1} + C_2 x^{\alpha_2}, \tag{2.4}$$

assumed for $0 < x < \varepsilon$, say, where C_1, α_1, α_2 are all positive and the model is made identifiable by insisting that $\alpha = \alpha_1 < \alpha_2$. The type of departure from a Pareto model suggested by (2.3) or (2.4) argues that the true distribution is, to a first approximation, a mixture of two Pareto distributions. This is sometimes implicitly assumed in accounts of departures from the Pareto. See for example Zipf's (1949), page 423 discussion of the contributions made by distinct urban and rural communities to the overall distribution of community sizes, and the remarks in the Appendix in this paper on data from countries that have been formed from mergers of autonomous states.

2.2. *Likelihood and least-squares approaches.* Our methods are based on the observation that, to a good approximation, the normalized log-spacings of small order-statistics in the sample $\mathscr{X}$ are very nearly rescaled exponential variables, where the scale change may be simply represented in terms of the model at (2.1). Specifically, let $X_{n1} \le \cdots \le X_{nn}$ denote the order statistics from $\mathscr{X}$ and define

$$U_i = i(\log X_{n,i+1} - \log X_{ni}).$$

Then, for a function δ_1 that can be expressed in terms of the δ of (2.1), it may be shown that

$$U_i \approx Z_i \theta \{1 + \delta_1(i/n)\} \approx Z_i \theta \exp\{\delta_1(i/n)\}, \tag{2.5}$$

where $\theta = 1/\alpha$ and the variables $Z_1, Z_2, \dots$ are independent and exponentially distributed with unit mean. It can be proved that if δ satisfies (2.1) and (2.3) then

$$\delta_1(y) = D_1 y^{\beta_1} + o(y^{\beta_1}) \tag{2.6}$$

as $y \downarrow 0$, where $\beta_1 = \beta/\alpha$ and $D_1 = -(\beta/\alpha)C^{-\beta/\alpha}D$. In view of (2.5), this suggests regarding the variables U_i as exponential with mean $\theta \exp\{D_1(i/n)^{\beta_1}\}$, and estimating θ, D_1, β_1 by maximum likelihood. In this case the negative log-likelihood is

$$r \log \theta + D_1 \sum_{i=1}^{r} (i/n)^{\beta_1} + \theta^{-1} \sum_{i=1}^{r} U_i \exp\{-D_1(i/n)^{\beta_1}\}. \tag{2.7}$$

(Here, r is the threshold or smoothing parameter, about which we shall say more in Section 2.4.) Differentiating with respect to θ, equating to zero and solving for θ, we obtain $\theta = T(D_1, \beta_1)$, say. Substituting back into (2.7) we conclude that $(\hat{D}_1, \hat{\beta}_1)$ should be chosen to minimize

$$L(D_1, \beta_1) = D_1 r^{-1} \sum_{i=1}^{r} (i/n)^{\beta_1} + \log\left[r^{-1} \sum_{i=1}^{r} U_i \exp\{-D_1(i/n)^{\beta_1}\}\right]. \tag{2.8}$$

An alternative approach to inference may be based on the fact that, by (2.5), the log–log spacings $V_i = \log U_i$ satisfy

$$V_i \approx \mu + \delta_1(i/n) + \varepsilon_i, \tag{2.9}$$

where $\mu = \log\theta + \mu_0$, μ_0 is the mean of the distribution of $\log Z_1$ (thus, $\mu_0 = -0.5772\ldots$, the negative of the value of Euler's constant), and $\varepsilon_i = \log Z_i - \mu_0$ (for $1 \le i \le n$) may be interpreted as an error in the approximate regression model (2.9). Given that δ_1 may be expressed by (2.6), we may (for fixed β_1) compute *explicit* estimators $\hat{\mu}(\beta_1)$, $\hat{D}_1(\beta_1)$ of the unknowns μ, D_1 that minimize

$$S(\mu, D_1, \beta_1) = \sum_{i=1}^{r} \left\{V_i - \mu - D_1(i/n)^{\beta_1}\right\}^2. \tag{2.10}$$

Despite their having larger asymptotic variance than maximum likelihood methods (see Section 4), we found in numerical studies that the least-squares approaches were sometimes less biased and so could have superior overall performance (Section 3).

The case where β is known, or is a known function of α (see Section 3.1), is of major interest. If we can reduce the number of unknown parameters, then the variances of our estimates of those that remain should also be reduced. Two canonical cases deserve special mention. In the first, one may think of X as being generated in the fashion $X = |Y|^{1/\alpha}$, where Y has a density that is nonzero and differentiable at the origin. Here, $\alpha = \beta$ and so $\beta_1 = 1$. The second example is generated in the form $F(x) = G(x)^\alpha$, where the distribution function G is supported on the positive half-line and (confined to that domain) has a density that is nonzero and differentiable at the origin. Here, $\beta = 1$ and so $\beta_1 = 1/\alpha$. In practice one may obtain empirical evidence for either of these cases by computing pilot estimates of D_1, α, β_1. See Section 3.1 for an example.

2.3. *An exploratory least-squares approach.* To obtain a preliminary approximation to β_1 it is sometimes helpful to fit the semiparametric model at (2.1) and (2.3) directly to log-spacings, as follows. Observe from (2.5) and (2.6) that

$$\log X_{n,i+1} \approx \log X_{ni} + i^{-1} Z_i \theta\left\{1 + D_1(i/n)^{\beta_1}\right\}.$$

Hence, for $i \ge j$ and with $x_{ij} = \log\{i/(j-1)\}$ and $y_{ij} = \log X_{n,i+1} - \log X_{nj}$,

$$y_{ij} \approx \theta x_{ij} + D_2 \exp(\beta_1 x_{ij}) + D_3, \tag{2.11}$$

where the "constants" D_2, D_3 depend on j and n. (The approximation is generally good, at least for i sufficiently greater than j, if D_3 is omitted. We do, however, need j moderately large in order to neglect the stochastic component.) We may fit the model at (2.11) by ordinary or weighted least-squares for values i satisfying $j < i \le r$, say.

2.4. *Estimators of* α. The analysis in Section 2.2 suggests three estimators of α, of which the first is based on maximum likelihood and the others on least-squares. The likelihood-based estimator is found by taking $(\hat{D}_1, \hat{\beta}_1)$ to minimize $L(D_1, \beta_1)$, defined at (2.8), and putting

$$(2.12) \qquad \hat{\alpha}_1 = T\left(\hat{D}_1, \hat{\beta}_1\right)^{-1} = \left[r^{-1} \sum_{i=1}^{r} U_i \exp\left\{-\hat{D}_1(i/n)^{\hat{\beta}_1}\right\}\right]^{-1}.$$

To derive the least-squares estimators, let $V_i = \log U_i = \log(\log X_{n,i+1} - \log X_{ni}) + \log i$, choose $(\hat{\mu}, \hat{\beta}_1, \hat{D}_1)$ to minimize $S(\mu, D_1, \beta_1)$ [defined at (2.10)] and let

$$(2.13) \qquad W_i = i(\log X_{n,i+1} - \log X_{ni})\exp\left\{-\hat{D}_1(i/n)^{\hat{\beta}_1}\right\},$$

$\overline{W} = r^{-1}\sum_{i \le r} W_i$ and $\mu_0 = \int_{x>0}(\log x)e^{-x}\,dx = -0.5772157$. Then two least-squares estimators of α are $\hat{\alpha}_2 = \exp(\mu_0 - \hat{\mu})$ and $\hat{\alpha}_3 = 1/\overline{W}$. When β_1 is known, or estimated separately, we replace $\hat{\beta}_1$ by that value at all its appearances, for example at (2.12) and (2.13). If we take $\beta_1 = 1/\alpha$ (see Sections 2.2 and 3.1), then we should again make the obvious changes to the algorithm, maximizing the likelihood or minimizing the sum of squares under the constraint.

Given an estimator $\hat{\alpha}$ of α, one may obtain an estimator of C quite simply, by substituting into the formula $\hat{C} = r(X_{nr})^{-\hat{\alpha}}/n$. As expected, the estimators of α do not require the value of n, while those of C do. In practice, when only extreme-order statistics are recorded, the value of n is usually unknown.

The value of r at (2.8) and (2.10) plays the role of a threshold, or smoothing parameter, determining the depth into the data that we are prepared to go when fitting the model defined by (2.1) and (2.3). For conventional estimators, such as those of Hill (1975) and also for our estimators $\hat{\alpha}_j$, increasing r results in an increase in bias, owing to departure of (2.1) from its "ideal" form (2.2), but this is accompanied by a decrease in variance. One virtue of our approach is that for it, bias does not increase so rapidly with increasing r, and so r may be chosen an order of magnitude larger, producing an improvement in mean-square performance by an order of magnitude. These properties will be demonstrated numerically in Section 3 and theoretically in Section 4; see, for example, Remark 4.4.

3. Numerical properties.

3.1. *Example*: *community sizes*. Zipf [(1949), page 139], showed graphically that data on large Australian community sizes in 1921 departed markedly from a Pareto distribution with regularly varying tail at infinity. He neither tabulated his data nor gave a source, but they were apparently taken from Wickens (1921). See the Appendix of the present paper for details. Figure 1 graphs the logarithm of community size against the logarithm of rank for these data for all 256 Australian communities that had 2000 or more inhabitants in 1921. (Zipf considered only communities of more than 3000 people.) The graph would of course be linear if the sampling distribution were Pareto.

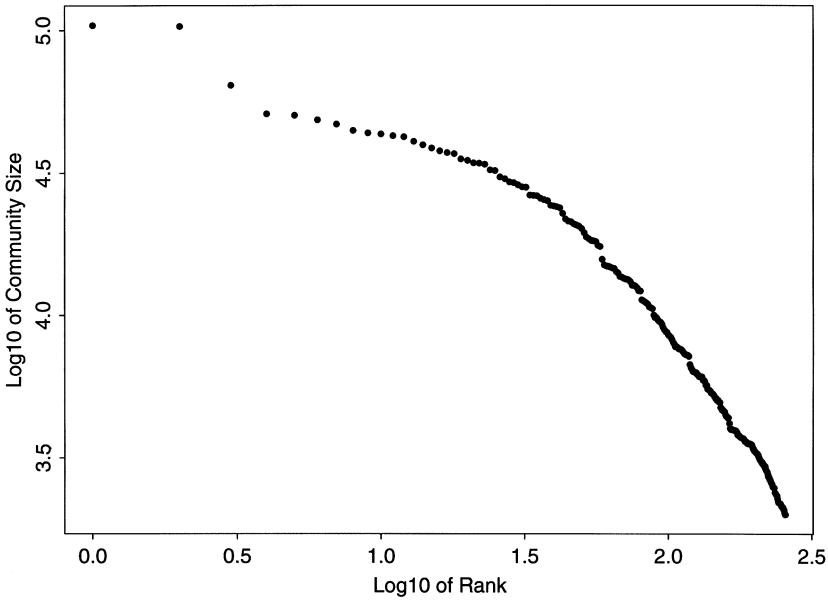

FIG. 1. *Logarithms of Australian community sizes, I. Plot of logarithm of community size against logarithm of rank for all* 256 *Australian communities having* 2000 *or more inhabitants in* 1921.

Nevertheless, it is plausible that the data were generated as $X = Y^{-1}$, where X denotes the population of a randomly chosen Australian community and Y has the perturbed Pareto distribution modelled by (2.1) and (2.3). We analyzed the data from that viewpoint and also as though they were from a perturbed Pareto distribution with an upper bound at a fixed number N_0, say. That is, we subtracted each population size from N_0 and regarded the new data $Z = N_0 - X$ as coming from a population whose distribution function had the form described by (2.1) and (2.3). This approach has apparently not been used before with community-size data and would be inimical to Zipf's ideas, but there are nevertheless good empirical reasons for adopting it. We shall report here only this type of analysis, since it produced residuals whose empirical distribution was closer to the exponential than was that of residuals obtained using the first approach.

Figure 2 shows a plot of values of the negative of the logarithm of Z against the logarithm of rank. It represents the analogue of Figure 1 in this context and would be linear if the single-component Pareto distribution at (2.2) were an appropriate model for departures of community size from an upper bound.

When applying the exploratory least-squares approach suggested in Section 2.3, we found that if we took $j \geq 5$ then a plot of y_{ij} against $\exp(\beta_1 x_{ij})$, for $j \leq i \leq 256$ and with β_1 chosen by least-squares, gave very nearly a

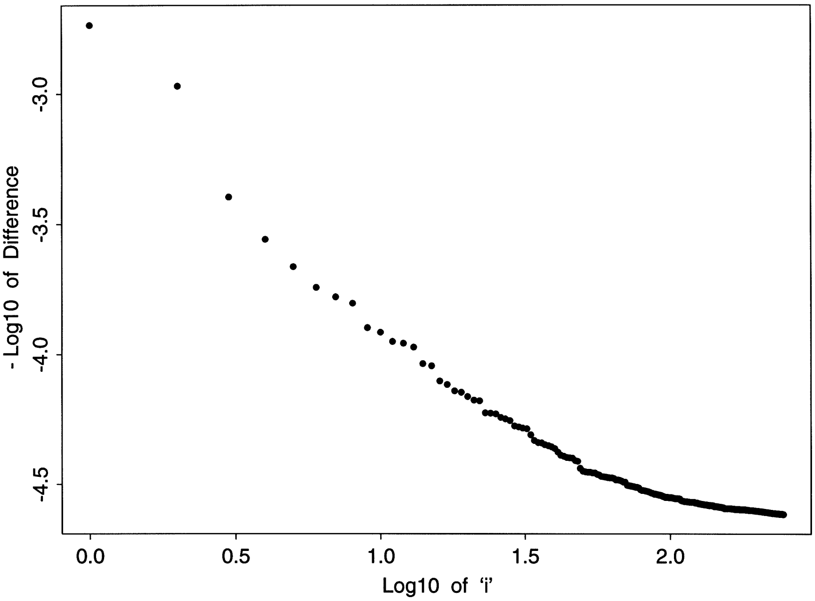

FIG. 2. *Logarithms of Australian community sizes, II. Plot of negative value of the logarithm of the difference between size of* 10*th largest community and size of the* (10 + *i*)*th largest community, for* $1 \le i \le 245$, *against the logarithm of* i.

straight line with negative slope. This suggests that for some $N_0 > 0$, the distribution function G of city size has very nearly the property,

$$
\begin{aligned}
1 - G(N_0 - x) &= A_1 \,|\log(1 - A_2 x)|^{\alpha} \\
&= A_1 A_2^{\alpha} x^{\alpha}\{1 + \alpha A_2 x + O(x^2)\},
\end{aligned}
\tag{3.1}
$$

for positive constants α, A_1, A_2 and small positive values of x. This is the perturbed Pareto model suggested by (2.1) and (2.3), with $\beta = 1$. To appreciate why (3.1) follows from linearity of the plot, note that linearity suggests that to a good approximation, $X_{n,i+1} = B_1 \exp(-B_2 i^{\beta_1})$ for constants B_1, $B_2 > 0$. Taking $X_{n,i+1} \approx G^{-1}\{1 - (i/n)\}$ and writing x for i/n, we obtain $1 - x = G\{B_1 \exp(-B_3 x^{\beta_1})\}$. This gives (3.1) exactly, with $\alpha = 1/\beta_1$.

Therefore, in our subsequent analysis, we took $\beta = 1$ at (2.3), or equivalently, $\beta_1 = 1/\alpha$ at (2.6). (Of course, this β_1 differs from that in the paragraph immediately above.) We used the maximum likelihood (ML) and least-squares (LS) methods suggested in Sections 2.2 and 2.4, with β_1 there constrained to equal $1/\alpha$. For brevity we took the LS estimator to be $\hat{\alpha}_2$; results for $\hat{\alpha}_3$ are similar. For the sake of simplicity we took N_0 to equal the population of the smallest community that was not used in the analysis. That is, if attention was confined to communities whose size ranked in the range $k + 1 \le i \le k + r + 1$, then we took N_0 to equal the size of the kth largest community. Removal of large community sizes was necessary to avoid errors

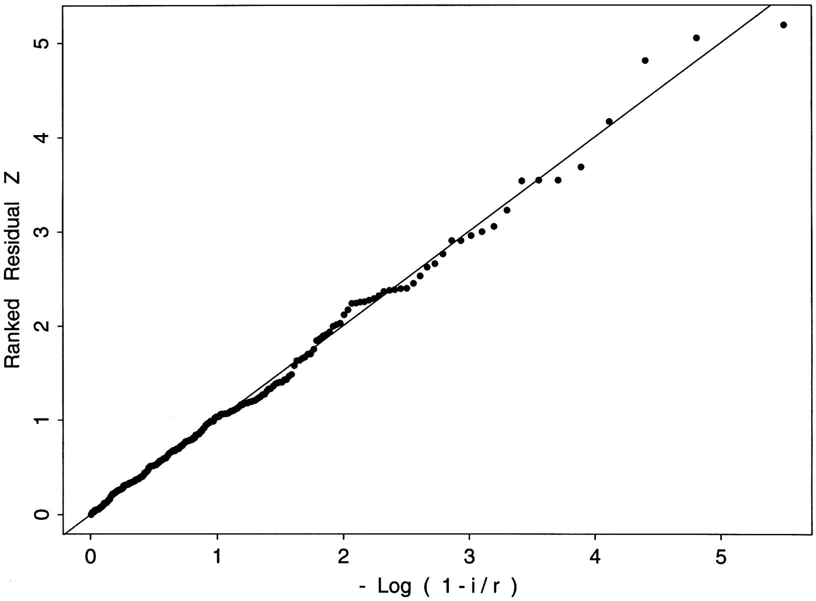

FIG. 3. *Residual q-q plot. Ranked values of the residuals $\hat{Z}_i$, defined by* (3.2), *plotted against corresponding quantiles of the exponential distribution with unit mean. The line "$y = x$" is superimposed on the plot. Raw data were those used to generate Figure* 2.

due to arbitrariness of community boundaries in large Australian cities. The definitions of Sydney and Sydney North, and Melbourne and Melbourne South as distinct communities are among the obvious examples of this problem.

The goodness-of-fit of the exponential distribution to residuals is apparent from q–q plots, of which Figure 3 is typical. There we graph $\log \hat{Z}_{(i+k+1)}$ against $-\log\{1 - (i/r)\}$ for $1 \le i \le r$, where

$$\hat{Z}_i = U_i \Big/ \Big[(1/\hat{\alpha})\exp\Big\{\hat{D}_1(i/n)^{1/\hat{\alpha}}\Big\}\Big] \tag{3.2}$$

is our estimate of the quantity Z_i appearing at (2.5), the parameter estimates on the right-hand side of (3.2) were obtained by maximum likelihood under the constraint that $\beta = 1$ and $\hat{Z}_{(i)}$ denotes the ith largest value of Z_i. For Figure 3 we took N_0 to be the size of the 10th largest community (so that $k = 10$) and used all remaining communities of size 2000 or more (so that $r = 256 - 11 = 245$). We do not of course know the value of n in (3.2), but that is of no concern since it is absorbed into the estimator of D_1.

The resulting estimates of α, for different values of r and the cut-off k, range from about 0.9 to 1.2 and are given in Table 1. The ML estimates are a little less variable than those based on LS and also tend to be a little larger. Not much can be read into this, however, since all estimates were computed from the same data. Both ML and LS estimates are strikingly resistant to substantial changes in r. The main source of variation appears to be random fluctuation, rather than systematic variation with r. This reflects the very low bias arising from the excellent fit of the model when β is constrained to equal 1; see Figure 3. By way of contrast, the Hill (1975) estimate of α increases virtually monotonically with increasing r, and more than doubles in size (from 1.57 to 3.53) within the range of Table 1. This substantial systematic error reflects the very poor fit to data offered by the single-component Pareto model at (2.2). The poor fit is exemplified by a highly nonlinear q–q plot (analogous to that of Figure 3 and not given here) for residuals under the single-component model.

In the same setting as Figure 3, the "full" (i.e., without β constrained) ML estimate of α equals 0.64, and the estimate of β is 0.54. To assess the significance of these results we conducted a simulation study using data generated from the distribution at (3.3), with $\alpha = \beta = 1$, $n = 20{,}000$ and $r = 250$ and values of B that produced "Zipf plots" having moderate curvatures, as in Figure 2. (The case of higher curvature will be treated in Section 3.2.) These values of n and r were chosen because they lead to Zipf plots broadly similar to that in Figure 2. For the sake of simplicity we shall continue to use the same n and r in Section 3.2.

TABLE 1
Estimates of α for different values of k and r[*]

		$r + k$			
	k	**256**	**200**	**150**	**100**
6	ML	1.215	1.051	1.053	1.157
	LS	1.152	0.910	0.873	0.980
	Hill	3.531	3.007	2.458	1.921
8	ML	1.149	0.986	0.964	0.993
	LS	1.097	0.874	0.811	0.865
	Hill	3.146	2.666	2.165	1.669
10	ML	1.166	1.001	0.981	1.017
	LS	1.212	0.883	0.839	0.891
	Hill	3.173	2.693	2.188	1.687
12	ML	1.211	1.043	1.043	1.149
	LS	1.152	0.918	0.851	1.004
	Hill	3.353	2.861	2.337	1.821
15	ML	1.165	0.997	0.980	1.026
	LS	1.125	0.892	0.846	0.935
	Hill	3.073	2.616	2.125	1.638
20	ML	1.167	0.995	0.974	1.007
	LS	1.134	0.901	0.853	0.952
	Hill	2.968	2.529	2.051	1.573

[*] In each cell in the table, the maximum likelihood estimate (denoted by ML, and computed under the constraint $\beta = 1$) is listed above the corresponding least-squares estimate (LS), which in turn is listed above the Hill (1975) estimate.

We found that (1) the "full" ML and LS estimators of α have much higher variance than the "constrained" estimators (i.e., with β constrained to equal 1); (2) the differences in bias are relatively minor; (3) the "constrained" ML estimator of α is substantially more accurate than the Hill (1975) estimator; and (4) the "full" ML estimator of β is biased downwards by a factor close to $\frac{1}{2}$, with variance being less of a problem. This bias appears to be due to third-order effects, which are not captured by second-order models such as the combination of (2.1) and (2.3). In view of this bias, we do not find the value $\beta = 0.54$, obtained using the "full" ML method, to be problematical. Overall, the "full" ML estimator does not perform well for these data, since departure from a single Pareto distribution is not sufficiently great; but the "constrained" ML method is effective.

We conclude that the true value of α is close to 1. This is the value claimed by Zipf (1949) for countries such as the United States that satisfy his law of "intranational equilibrium," although of course he was concerned with regular variation at infinity, not at an upper bound to city size.

3.2. *Summary of numerical properties*. When fitting mixture models, it is generally found that a multicomponent model produces improved performance only if there is clear evidence that more than one component is necessary. When a single component is adequate, fitting a mixture of two or more components typically leads to poor performance, because (1) the additional nuisance parameters use up information that would otherwise be available for estimating the main parameters of interest, and (2) there are problems of identifiability when the components are close. This gives rise to relatively poor performance of the "full" ML and LS methods, noted in Section 3.1, when the Zipf curve has only moderate curvature.

However, when fitting a single Pareto distribution to data such as those on which Figure 2 is based, it is relatively important to determine the threshold, r, by empirical means. It has been observed previously that, in a range of settings, empirical choice of r can increase root mean squared error by a factor of about 2 when β is unknown; see, for example, Hall and Welsh (1985). As shown in Section 3.1 (Table 1), when fitting a Pareto mixture the estimator of α is relatively robust against systematic effects resulting from choice of r. Therefore, even if "full" or "constrained" ML or LS methods, for fixed values of r, produce estimates that perform similarly to Hill's estimator when the latter is computed at an optimal threshold, the ML or LS methods can be superior to the Hill estimator in practice. Analysis of threshold-choice methods for Hill's estimator is beyond the scope of this paper.

Since β_1 is the parameter of a high-order term in the model, the likelihood surface is relatively flat near the maximizing value of β_1. Figure 4 depicts a typical graph of $-L(D_1, \beta_1)$, defined at (2.8), and shows this property clearly. Similar behavior may be observed for the function $S(\mu, D_1, \beta_1)$, defined at (2.10), and in this case one must also take care that the estimate of β_1 is not taken to be a pathological extremum at infinity. These problems are not as

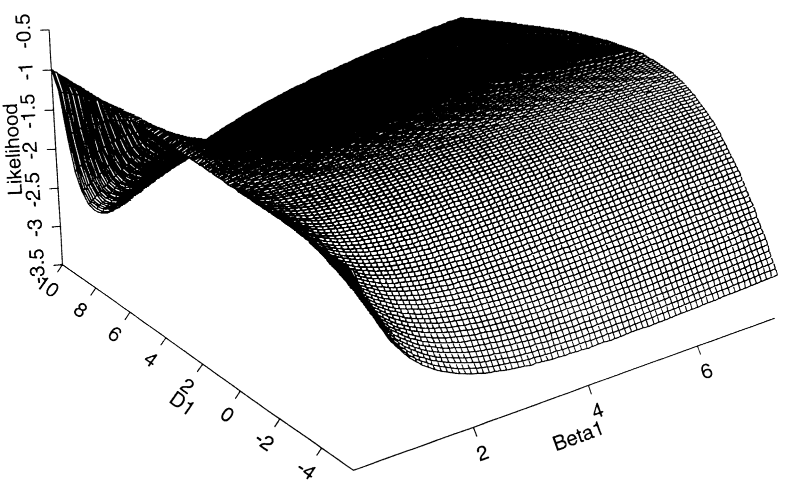

FIG. 4. *Likelihood surface. Typical plot of* $-L(D, \beta_1)$ *for data generated in the simulation study in Section* 3.2.

serious when data sets are analyzed individually, however. In our simulation study we used grid search to approximate the minimum, but due to the flatness of L and S it sometimes happened that the value we obtained was a long way from the true minimum. Additionally, in some samples where the Zipf plot was approximately linear, despite the average Zipf curve being nonlinear, the estimate of β_1 was a large distance from the true value of β_1. For these reasons we use median absolute deviation (MAD) instead of mean squared error to describe performance.

We simulated data from the distribution

$$X = U^{1/\alpha} \exp(BU^{\beta/\alpha}), \tag{3.3}$$

where $\alpha, \beta > 0$, $B = -D/\alpha$, $-\infty < D < \infty$ and U had a Uniform distribution on the interval $[0, 1]$. The parameters α, β, D have the meanings ascribed to them in Section 2, and the value of C there is now 1. In particular, the perturbed Pareto model defined at (2.1) and (2.3) is valid. In the context of (3.3), and for $n = 20{,}000$ and $r = 250$, Table 2 gives median absolute deviations of estimates computed by "full" or "constrained" ML and LS methods. The "constrained" estimators are generally slightly superior, and the ML and LS methods perform similarly. Throughout this section, the results reported represent averages over 100 simulated independent samples.

The MAD of estimators suggested by Hill (1975), for $r = 250$, exceeds that of "full" ML estimators by a factor of between 1.3 and 3.6 across the range of Table 2. If the Hill estimator is computed at the value of r that gave it optimal MAD performance in the simulation study, then its MAD is generally

TABLE 2
Values of mean absolute deviation[*]

	$\boldsymbol{\beta_1 = 1}$		$\boldsymbol{\beta_1 = 1/\alpha}$	
$\alpha = \frac{1}{2}$	$(B = 500)$		$(B = 15{,}000)$	
	$\mathrm{ML}^{(0)}$	0.12	$\mathrm{ML}^{(0)}$	0.061
	$\mathrm{LS1}^{(0)}$	0.13	$\mathrm{LS1}^{(0)}$	0.071
	$\mathrm{LS2}^{(0)}$	0.13	$\mathrm{LS2}^{(0)}$	0.071
	$\mathrm{ML}^{(1)}$	0.10	$\mathrm{ML}^{(2)}$	0.040
	$\mathrm{LS1}^{(1)}$	0.10	$\mathrm{LS}^{(2)}$	0.068
	$\mathrm{LS2}^{(1)}$	0.10		
$\alpha = 1$	$(B = 300)$		$(B = 300)$	
	$\mathrm{ML}^{(0)}$	0.30	$\mathrm{ML}^{(0)}$	0.30
	$\mathrm{LS1}^{(0)}$	0.36	$\mathrm{LS1}^{(0)}$	0.36
	$\mathrm{LS2}^{(0)}$	0.35	$\mathrm{LS2}^{(0)}$	0.35
	$\mathrm{ML}^{(1)}$	0.23	$\mathrm{ML}^{(2)}$	0.47
	$\mathrm{LS1}^{(1)}$	0.22	$\mathrm{LS}^{(2)}$	0.38
	$\mathrm{LS2}^{(1)}$	0.23		
$\alpha = 2$	$(B = 200)$		$(B = 30)$	
	$\mathrm{ML}^{(0)}$	0.87	$\mathrm{ML}^{(0)}$	1.0
	$\mathrm{LS1}^{(0)}$	0.58	$\mathrm{LS1}^{(0)}$	0.93
	$\mathrm{LS2}^{(0)}$	0.59	$\mathrm{LS2}^{(0)}$	0.92
	$\mathrm{ML}^{(1)}$	0.61	$\mathrm{ML}^{(2)}$	1.0
	$\mathrm{LS1}^{(1)}$	0.57	$\mathrm{LS}^{(2)}$	1.0
	$\mathrm{LS2}^{(1)}$	0.60		

[*] The estimates are $\hat{\alpha}_1$ (abbreviated here to ML), $\hat{\alpha}_2$ (LS1) and $\hat{\alpha}_3$ (LS2), and are as defined in Section 2.4. Their unconstrained forms, constrained forms subject to $\beta_1 = 1$, and constrained forms subject to $\beta_1 = 1/\alpha$, are indicated by the superscripts $^{(0)}$, $^{(1)}$ and $^{(2)}$, respectively. When the constraint is $\beta_1 = 1/\alpha$, the distinction between LS1 and LS2 is lost; there, the LS estimator of α is defined by minimizing $S(\mu_0 - \log\alpha, 1/\alpha, D_1)$ with respect to (α, D_1).

close to that of the "full" ML and LS estimates. However, this does not take into account the need in practice to choose r empirically so as to achieve good performance. We made no attempt to optimize our ML or LS estimates over r. When $\alpha = 1$ or 2, in particular, performance can be improved significantly by using smaller values of r. Guidance as to the appropriate r may be gained by examining q–q plots; see Figure 3.

In each case, B in (3.1) was chosen so that average values of Zipf plots showed curvature more marked than that in Figure 2 and of the opposite sign. Subject to this constraint, B was selected so that curvatures were visually similar in all the settings of the table. Figure 5 depicts average values of Zipf plots in the case of the first column of Table 2, each curve representing the mean of 100 independent synthetic samples.

If curvature is decreased then the performance of "full" ML and LS methods relative to their "constrained" versions deteriorates, since the relative contribution of stochastic error to MAD increases. As noted in Section 3.1, the ML and LS estimators are relatively robust against changes in r.

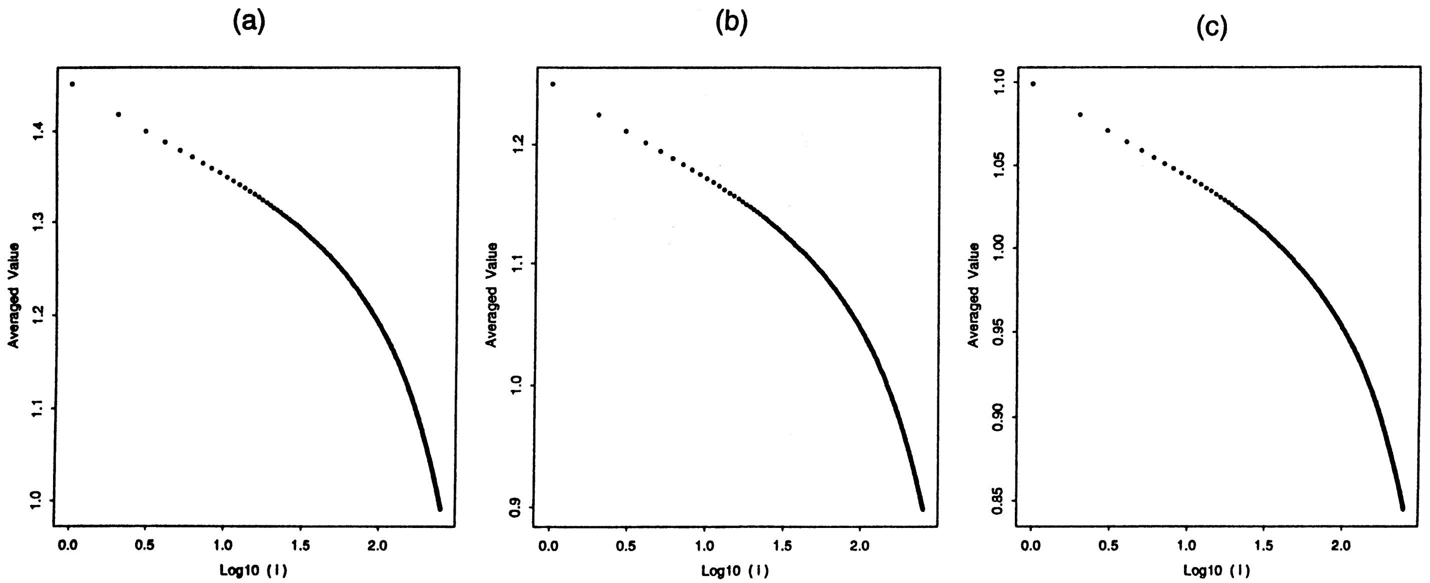

FIG. 5. *Average Zipf plots. Plot of expected value of the negative of the logarithm of the ith smallest simulated data value, against the logarithm of i. The parameter values for panels (a), (b) and (c) are those of the first three blocks, respectively, in the first column of Table 2.*

4. Theoretical properties. Assume that (2.1) holds, where $\alpha > 0$ and the function δ is twice-differentiable on $(0, \infty)$ and satisfies

$$\delta^{(i)}(x) = (\partial/\partial x)^i (Dx^{\beta}) + O(x^{\gamma - i}) \quad \text{for } i = 0, 1, 2 \tag{4.1}$$

as $x \downarrow 0$, with $0 < \beta < \gamma < \infty$ and $-\infty < D < \infty$. Put $\beta_1 = \beta/\alpha$ and $\gamma_1 = \gamma/\alpha$, and let $\sigma_1, \dots, \sigma_6$ denote positive constants. (In Remark 4.1 we shall give the values of σ_j^2 for $j = 1, 2, 4, 5$.) Define the estimators $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ as in Section 2.4.

THEOREM. *Assume condition* (4.1), *and that* $r = r(n) \to \infty$ *at a rate such that* $r^{-1} + (r/n) = O(n^{-\varepsilon})$ *for some* $\varepsilon > 0$. *If* β_1 *is estimated as part of the likelihood or least-squares procedure then, for* $j = 1, 2, 3$,

$$\hat{\alpha}_j = \alpha\Big[1 + r^{-1/2} N_{nj} + O_p\big\{(r/n)^{\min(2\beta_1, \gamma_1)} (\log n)^3\big\}\Big] \tag{4.2}$$

as $n \to \infty$, *where the random variable* N_{nj} *is asymptotically Normal* $N(0, \sigma_j^2)$. *If, in the estimation procedure, we substitute for* β_1 *a random variable that is within* $O_p(\eta)$ *of its true value, where* $0 \le \eta = O(n^{-\varepsilon})$ *for some* $\varepsilon > 0$, *then* (4.2) *remains true provided* (1) *we interpret* N_{nj} *as an asymptotically Normal* $N(0, \sigma_{j+3}^2)$ *random variable, and* (2) *we replace the remainder term by* $O_p\{(r/n)^{\min(2\beta_1, \gamma_1)} + \eta \log n\}$.

REMARK 4.1 (Values of σ_j^2). For integers $j \ge 1$ and $k \ge 0$, define

$$\kappa_{jk} = \int_0^1 x^{j\beta_1} (\log x)^k \, dx,$$

and put $\lambda_0 = (\kappa_{20} - \kappa_{10}^2)(\kappa_{22} - \kappa_{11}^2) - (\kappa_{21} - \kappa_{10}\kappa_{11})^2$, $\lambda_1 = \kappa_{20}\kappa_{22} - \kappa_{21}^2$, $\lambda_2 = \kappa_{11}\kappa_{21} - \kappa_{10}\kappa_{22}$ and $\lambda_3 = \kappa_{10}\kappa_{21} - \kappa_{11}\kappa_{20}$. Then,

$$\begin{aligned}\sigma_1^2 &= \lambda_0^{-2} \int_0^1 \big(\lambda_1 + \lambda_2 x^{\beta_1} + \lambda_3 x^{\beta_1} \log x\big)^2 \, dx \\ &= \lambda_0^{-2} \big\{\lambda_1^2 + \lambda_2^2 \kappa_{20} + \lambda_3^2 \kappa_{22} + 2(\lambda_1 \lambda_2 \kappa_{10} + \lambda_1 \lambda_3 \kappa_{11} + \lambda_2 \lambda_3 \kappa_{21})\big\},\end{aligned}$$

and $\sigma_2^2 = \sigma_0^2 \sigma_1^2$, where $\sigma_0^2 = \pi^2/6 = 1.644934$ is the variance of the logarithm of an exponential random variable. There is no such elementary relationship in the cases of σ_3^2 or σ_6^2, for which the formulas are particularly complex and, for brevity, are not given here. More simply, however, $\sigma_4^2 = (\kappa_{20} - \kappa_{10}^2)^{-1}$ and $\sigma_5^2 = \sigma_0^2 \sigma_4^2$.

REMARK 4.2 (Bias reduction). Condition (4.1) asks that to first order, δ decrease like x^{β} as $x \downarrow 0$, and to second order, decrease like x^{γ}. In these circumstances the biases of more conventional estimators of α are asymptotic to a constant multiple of $(r/n)^{\beta_1}$; see for example Hall (1982) and Csörgő, Deheuvels and Mason (1985). Our theorem shows that this level of bias has been eliminated completely from the estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$ and that the new bias is of order only $(r/n)^{\min(2\beta_1, \gamma_1)}$, multiplied by a logarithmic factor. This represents an improvement by an order of magnitude.

REMARK 4.3 (Variance). The variance of conventional estimators is of size r^{-1} [Hall (1982); Csörgő, Deheuvels and Mason (1985)]. It follows from the theorem that this level of variance is preserved by our bias-reduced estimators. Moreover, it may be proved that under the assumption that the model at (2.4) holds exactly for x in some interval $[0, \varepsilon]$, the estimator $\hat{\alpha}_1$ has asymptotic minimum variance among all estimators based on $X_{n1}, \dots, X_{nr}$.

REMARK 4.4 (Mean squared error reduction). The theoretically smallest order of mean squared error is achieved by selecting the threshold, r, so that squared asymptotic bias is of the same size as asymptotic variance. In view of the results noted in Remarks 4.2 and 4.3, this will produce a mean squared error that is an order of magnitude less for our estimators $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$ than in cases of conventional methods. In the conventional cases mentioned in Remarks 4.2 and 4.3, this balance would be achieved with a value r of the same size as $(n/r)^{2\beta_1}$, but for our estimators, the optimal r is an order of magnitude larger.

REMARK 4.5 (Estimators of C, D_1 and β_1). By extending arguments in Section 5 it may be shown that the estimator $\hat{C}_j = r(X_{nr})^{-\hat{\alpha}_j}/n$ has asymptotic variance of size $r^{-1}(\log n)^2$ and asymptotic bias of size $(r/n)^{\min(2\beta_1, \gamma_1)}$ multiplied by a power of $\log n$ (the latter depending on j). Our estimators of D_1 and β_1, derived by either maximum likelihood or least squares, have variances of size $r^{-1}(r/n)^{-2\beta_1}(\log n)^2$ and $r^{-1}(r/n)^{-2\beta_1}$, respectively. Therefore, these estimators will not be consistent unless $r(r/n)^{2\beta_1} \to \infty$. In view of Remark 4.4, this requires r to be an order of magnitude larger than would typically be used for conventional estimators of α. Note, however, that despite our likelihood and least-squares estimators being derived as functions of estimators of D_1 and β_1, consistent estimation of these quantities is not required for consistent estimation of α.

5. Derivation of theorem. For brevity we treat only the case of $\hat{\alpha}_1$, in the setting where α, β_1, D_1 are estimated together. Let $Z_1, Z_2, \dots$ be independent exponential random variables with unit mean and define

$$S_i = \sum_{j=1}^{n-i+1} Z_{n-j+1}/(n-j+1)$$

and $T_i = \exp(-S_i)$. By Rényi's representation for order statistics [e.g., David (1970), page 18], we may choose the Z_i's so that $X_{ni} = F^{-1}(T_i)$ for $1 \le i \le n$. Observe too that

$$S_i = \log(n/i) + O_p(i^{-1/2}) \tag{5.1}$$

uniformly in $1 \le i \le r$ and that the representation of F at (2.1) may equivalently be written as

$$\log F^{-1}(x) = \theta \log x + C_1 + \delta_2(x), \tag{5.2}$$

where $\theta = \alpha^{-1}$, $C_1 = -\theta \log C$, and by (4.1), defining $\xi = \min(2\beta_1, \gamma_1)$,

$$\delta_2 = -\theta C^{-\beta_1} D x^{\beta_1} + O(x^{\xi})$$

as $x \downarrow 0$. Furthermore, $S_{i+1} - S_i = -i^{-1} Z_i$. It follows from the latter result and (5.2) that

$$U_i \equiv i(\log X_{n,i+1} - \log X_{ni}) = \theta Z_i + i\{\delta_2(T_{i+1}) - \delta_2(T_i)\},$$

whence, defining $\mu_0 = E(\log Z_1)$, $\mu = \log\theta + \mu_0$, $\varepsilon_i = \log Z_i - \mu_0$ and

$$\Delta_i = \log\big[1 + (\theta Z_i)^{-1} i\{\delta_2(T_{i+1}) - \delta_2(T_i)\}\big],$$

we have

$$U_i = \theta Z_i \exp(1 + \Delta_i). \tag{5.3}$$

Put $\delta_3(z) = \delta_2(e^{-z})$. Then by (4.1), $\delta_3^{(i)}(z) = O\{\exp(-\beta_1 z)\}$ for $i = 1, 2$, as $z \to \infty$. In view of (5.1), $\exp(-S_i) = (i/n)\{1 + O_p(i^{-1/2})\}$ uniformly in $1 \le i \le r$, and so, defining $\rho_i = (i/n)^{\beta_1}$, we have

$$\begin{aligned} \delta_2(T_{i+1}) - \delta_2(T_i) &= \delta_3(S_{i+1}) - \delta_3(S_i) \\ &= -i^{-1} Z_i \delta_3'(S_i) + O_p\big\{(Z_i/i)^2 \rho_i\big\}. \end{aligned} \tag{5.4}$$

Moreover, by (5.1),

$$\delta_3'(S_i) = \delta_3'\{\log(n/i)\} + O_p\big(i^{-1/2}\rho_i\big). \tag{5.5}$$

Define $a_i = -\delta_3'\{\log(n/i)\} = (i/n)\delta_2'(i/n)$. Combining (5.4) and (5.5), we deduce that

$$\delta_2(T_{i+1}) - \delta_2(T_i) = i^{-1} Z_i a_i + O_p\big\{i^{-3/2} Z_i (Z_i + 1)\rho_i\big\}.$$

Therefore, $\theta\Delta_i = a_i + O_p\{i^{-1/2}(Z_i + 1)\rho_i + (i/n)^{\xi}\}$, uniformly in $1 \le i \le r$. Now, $a_i = -(\beta/\alpha^2) C^{-\beta/\alpha} D (i/n)^{\beta_1} + O\{(i/n)^{\xi}\}$. Hence,

$$\Delta_i = D_1 \beta_1 (i/n)^{\beta_1} + O_p\big\{i^{-1/2}(Z_i + 1)\rho_i + (i/n)^{\xi}\big\}, \tag{5.6}$$

uniformly in $1 \le i \le r$, where $D_1 = -(\beta/\alpha) C^{-\beta/\alpha} D$. Substituting (5.6) into (5.3) we see that

$$U_i = \theta Z_i \exp\big\{D_1 (i/n)^{\beta_1}\big\} + O_p\big\{i^{-1/2} Z_i (Z_i + 1)\rho_i + (i/n)^{\xi}\big\}, \tag{5.7}$$

uniformly in $1 \le i \le r$.

From this point it is convenient to write μ^0, D_1^0, β_1^0 rather than μ, D_1, β_1 for the true values of μ, D_1, β_1 and to write μ, D_1, β_1 for general candidates for μ^0, D_1^0, β_1^0. In this notation, put $\Delta_\mu = \mu - \mu^0$, $\Delta_D = (D_1 - D_1^0)/D_1^0$ and $\Delta_\beta = \beta_1 - \beta_1^0$, and note that $\rho_i = (i/n)^{\beta_1^0}$ and $\xi = \min(2\beta_1^0, \gamma_1)$. Now,

$$\begin{aligned} D(i/n)^{\beta_1} &= \big\{D_1^0 + \big(D - D_1^0\big)\big\}(i/n)^{\beta_1^0} \exp\big\{\big(\beta_1 - \beta_1^0\big)\log(i/n)\big\} \\ &= D_1^0 (i/n)^{\beta_1^0} + D_1^0 (i/n)^{\beta_1^0}\big\{\Delta_D + \Delta_\beta \log(i/n)\big\} \\ &\quad + O\Big[(i/n)^{\beta_1^0}\Big\{\Delta_D^2 + \big(\Delta_\beta \log n\big)^2\Big\}\Big], \end{aligned}$$

uniformly in $1 \le i \le r$. Therefore, by (5.7),

$$\begin{aligned} V_i &\equiv U_i \exp\big\{-D_1(i/n)^{\beta_1}\big\} \\ &= \theta Z_i\big\{1 - D_1^0\,\rho_i(\Delta_D + \Delta_\beta l_i)\big\} + O_p\{Q_i(D_1, \beta_1)\}, \end{aligned} \tag{5.8}$$

$$\begin{aligned} V_i(i/n)^{\beta_1} = {}& \theta Z_i\,\rho_i\big\{1 - D_1^0\,\rho_i(\Delta_D + \Delta_\beta l_i)\big\} + \theta Z_i\big\{(i/n)^{\beta_1} - \rho_i\big\} \\ &+ O_p\{\rho_i Q_i(D_1, \beta_1)\} \end{aligned} \tag{5.9}$$

uniformly in $1 \le i \le r$, where $l_i = \log(i/n)$ and

$$Q_i(D_1, \beta_1) = i^{-1/2} Z_i(Z_i + 1)\rho_i + (i/n)^{\xi} + \rho_i\Big\{\Delta_D^2 + \big(\Delta_\beta \log n\big)^2\Big\}.$$

For $j = 1, 2$, let B_j equal the average over $1 \le i \le r$ of the left-hand sides of (5.8) and (5.9), respectively, and let B_3 equal the average over $1 \le i \le r$ of the left-hand side of (5.9) multiplied by l_i. Let $\overline{Z}$, $\overline{\rho Z}$ and $\overline{\rho m Z}$ equal the averages of Z_i, $\rho_i Z_i$ and $\rho_i m_i Z_i$, respectively, over the same range. Put $l = l_r$ and $m_i = \log(i/r)$, and observe that $l_i = l + m_i$. For nonnegative integers k, define $\tau_{jk} = r^{-1}\sum_{i \le r} \rho_i^j m_i^k$, and let $q = r^{-1/2}\rho + (r/n)^{\xi} + \rho\{\Delta_D^2 + (\Delta_\beta \log n)^2\}$. In this notation we have, by (5.8) and (5.9),

$$\begin{aligned} \theta^{-1} B_1 &= \overline{Z} - D_1^0\{\tau_{10}\Delta_D + (\tau_{11} + l\tau_{10})\Delta_\beta\} + O_p(q), \\ \theta^{-1} B_2 &= \overline{\rho Z} - D_1^0\{\tau_{20}\Delta_D + (\tau_{21} + l\tau_{20})\Delta_\beta\} \\ &\qquad + r^{-1}\sum_{i=1}^{r} Z_i\big\{(i/n)^{\beta_1} - \rho_i\big\} + O_p(\rho q), \\ \theta^{-1} B_3 &= \overline{\rho m Z} + l\overline{\rho Z} - D_1^0\big\{(\tau_{21} + l\tau_{20})\Delta_D \\ &\qquad\qquad + \big(\tau_{22} + 2l\tau_{21} + l^2\tau_{20}\big)\Delta_\beta\big\} \\ &\qquad + r^{-1}\sum_{i=1}^{r} Z_i\big\{(i/n)^{\beta_1} - \rho_i\big\} l_i + O_p(\rho q \log n). \end{aligned} \tag{5.10}$$

Also,

$$\begin{aligned} b_1 &\equiv r^{-1}\sum_{i=1}^{r} (i/n)^{\beta} = \tau_{10} + r^{-1}\sum_{i=1}^{r}\big\{(i/n)^{\beta_1} - \rho_i\big\}, \\ b_2 &\equiv r^{-1}\sum_{i=1}^{r} (i/n)^{\beta}\log(i/n) = \tau_{11} + l\tau_{10} + r^{-1}\sum_{i=1}^{r}\big\{(i/n)^{\beta_1} - \rho_i\big\} l_i. \end{aligned}$$

Consequently,

$$\begin{aligned} \theta^{-1} b_1 B_1 &= \tau_{10}\big[\overline{Z} - D_1^0\{\tau_{10}\Delta_D + (\tau_{11} + l\tau_{10})\Delta_\beta\}\big] \\ &\qquad + \overline{Z} r^{-1}\sum_{i=1}^{r}\big\{(i/n)^{\beta_1} - \rho_i\big\} + O_p(\rho q), \\ \theta^{-1} b_2 B_1 &= (\tau_{11} + l\tau_{10})\big[\overline{Z} - D_1^0\{\tau_{10}\Delta_D + (\tau_{11} + l\tau_{10})\Delta_\beta\}\big] \\ &\qquad + \overline{Z} r^{-1}\sum_{i=1}^{r}\big\{(i/n)^{\beta_1} - \rho_i\big\} l_i + O_p(\rho q \log n), \end{aligned}$$

whence

$$\theta^{-1}(b_1B_1 - B_2) = \tau_{10}\overline{Z} - \overline{\rho Z} + D_1^0\Big[\big(\tau_{20} - \tau_{10}^2\big)\Delta_D + \big\{\tau_{21} + l\tau_{20} - \tau_{10}(\tau_{11} + l\tau_{10})\big\}\Delta_\beta\Big] + O_p\Big(\rho q + r^{-1/2}\rho|\Delta_\beta|\log n\Big), \tag{5.11}$$

$$\theta^{-1}(b_2B_1 - B_3) = (\tau_{11} + l\tau_{10})\overline{Z} - \big(\overline{\rho m Z} + l\overline{\rho Z}\big) + D_1^0\Big[\big\{\tau_{21} + l\tau_{20} - \tau_{10}(\tau_{11} + l\tau_{10})\big\}\Delta_D + \Big\{\tau_{22} + 2l\tau_{21} + l^2\tau_{20} - (\tau_{11} + l\tau_{10})^2\Big\}\Delta_\beta\Big] + O_p\Big\{\rho q\log n + r^{-1/2}\rho|\Delta_\beta|(\log n)^2\Big\}. \tag{5.12}$$

Here we have used the fact that, for $j = 0, 1$,

$$(\overline{Z} - 1)r^{-1}\sum_{i=1}^{r}\big\{(i/n)^{\beta_1} - \rho_i\big\}l_i^j = O_p\Big\{r^{-1/2}\rho|\Delta_\beta|(\log n)^{j+1}\Big\}.$$

Let $M = (m_{jk})$ denote the 2×2 symmetric matrix with $m_{11} = \tau_{20} - \tau_{10}^2$, $m_{12} = \tau_{21} - \tau_{10}\tau_{11} + l(\tau_{20} - \tau_{10}^2)$ and

$$m_{22} = \tau_{22} - \tau_{11}^2 + 2l(\tau_{21} - \tau_{10}\tau_{11}) + l^2\big(\tau_{20} - \tau_{11}^2\big).$$

Then, $\underline{\text{det } M} = (\tau_{20} - \tau_{10}^2)(\tau_{22} - \tau_{11}^2) - (\underline{\tau_{21}} - \tau_{10}\underline{\tau_{11})^2.}$ Define $(w_1, w_2, w_3) = E(\overline{Z}, \overline{\rho Z}, \overline{\rho m Z})$, $(W_1, W_2, W_3) = (\overline{Z} - w_1, \overline{\rho Z} - w_2, \overline{\rho m Z} - w_3)$, $A_1 = \tau_{10}W_1 - W_2$ and $A_2 = (\tau_{11} + l\tau_{10})W_1 - (W_3 + lW_2)$. Let $R_1, R_2, \ldots$ denote generic random variables each of which equals $1 + o_p(1)$. In this notation,

$$\begin{aligned} W &\equiv (\det M)(\tau_{10}, \tau_{11} + l\tau_{10})M^{-1}(A_1, A_2)^T \\ &= \Big[\tau_{10}\big\{\tau_{10}\big(\tau_{22} - \tau_{11}^2\big) - \tau_{11}(\tau_{21} - \tau_{10}\tau_{11})\big\} + \tau_{11}(\tau_{11}\tau_{20} - \tau_{10}\tau_{21})\Big]W_1 \\ &\quad - \big\{\tau_{10}\big(\tau_{22} - \tau_{11}^2\big) - \tau_{11}(\tau_{21} - \tau_{10}\tau_{11})\big\}W_2 - (\tau_{11}\tau_{20} - \tau_{10}\tau_{21})W_3. \end{aligned} \tag{5.13}$$

Hence,

$$\begin{aligned} &W_1 + (\det M)^{-1}W \\ &\quad = (\det M)^{-1}\big\{\big(\tau_{20}\tau_{22} - \tau_{21}^2\big)W_1 + (\tau_{11}\tau_{21} - \tau_{10}\tau_{22})W_2 + (\tau_{10}\tau_{21} - \tau_{11}\tau_{20})W_3\big\}. \end{aligned} \tag{5.14}$$

Since $\tau_{jk} \sim \rho^j\kappa_{jk}$ as $n \to \infty$ then the quantity at (5.14) is asymptotically Normal with zero mean and variance $r^{-1}\sigma_1^2$, where σ_1^2 is as defined in Remark 4.1.

Let $p = r^{-1/2}\rho + (r/n)^{\xi}$ denote the part of q that does not involve Δ_D or Δ_β. If, in the quantity on the far left-hand side of (5.13), we replace (A_1, A_2) by

$$(A'_1, A'_2) = \big(A_1 + O_p(\rho p), A_2 + O_p(\rho p \log n)\big),$$

then the net change to the far right-hand side of (5.13) is to add a term $O_p(t)$, where $t = \rho^4 p(\log n)$. Hence, the net change to (5.14) is to add a term of size $(r/n)^{\xi}(\log n)^3$. We claim that this gives the claimed limit theorem in the case of $\hat{\alpha}_1$. To appreciate why, let $\hat{\Delta}_D$, $\hat{\Delta}_\beta$ denote the versions of Δ_D, Δ_β in which D_1, β_1 are replaced by $\hat{D}_1$, $\hat{\beta}_1$, respectively. Note that the function $L(D_1, \beta_1)$, defined at (2.13), is minimized when $b_1 B_1 - B_2 = 0$ and $b_2 B_1 - B_3 = 0$. Therefore, $\hat{D}_1$, $\hat{\beta}_1$ are given asymptotically by the equations formed by setting the right-hand sides of (5.11) and (5.12) equal to zero. This shows that $(\hat{\Delta}_D, \hat{\Delta}_\beta)^T$ equals $-(D_1^0)^{-1}M^{-1}(A_1, A_2)^T$ [note the appearance of $M^{-1}(A_1, A_2)^T$ on the left in (5.13)], plus terms that are either negligible or of size $\rho^{-1}(r/n)^{\xi}(\log n)^2$. Moreover, our likelihood-based estimator $\hat{\theta} = \hat{\alpha}^{-1}$ of θ equals $B_1 = B_1(D_1, \beta)$, evaluated at $(\hat{D}_1, \hat{\beta}_1)$. Using the expansion (5.10) of B_1 we see that $\theta^{-1}\hat{\theta}$ equals

$$\bar{Z} - D_1^0(\tau_{10}, \tau_{11} + l\tau_{10})\big(\hat{\Delta}_D, \hat{\Delta}_\beta\big)^T,$$

plus terms that are either negligible or of size $(r/n)^{\xi}(\log n)^3$. [Note the appearance of $(\tau_{10}, \tau_{11} + l\tau_{10})$ on the left in (5.13).] Hence, by (5.13) (modified as suggested earlier), we deduce that $\theta^{-1}\hat{\theta}$ equals the quantity at (5.14), plus a term of size t. The desired central limit theorem for $\hat{\alpha}_1$ follows.

APPENDIX

Notes on the data. The data analyzed in Section 3.1 were extracted from tables prepared following the census of the Commonwealth of Australia on the nights of 3 and 4 April, 1921. See Wickens (1921). As definitions of communities we took "Municipalities" for New South Wales, Western Australia and Tasmania, "Cities, Towns and Boroughs" for Victoria, "Cities and Towns" for Queensland, and "Corporations" for South Australia. Alternative definitions, incorporating sparser communities in rural districts, would include data on "Shire" populations in New South Wales, Victoria and Queensland, "District Councils" for South Australia, and "Road Districts" for Western Australia. However, our choice appears to reproduce exactly the data presented graphically by Zipf [(1949), page 439].

The populations of the two internal territories in 1921, the Federal Capital Territory and the Northern Territory, were not tabulated by Wickens (1921) in a form which is readily comparable with that for the states, although detailed geographic distributions were given. We judged from the latter that the territories did not include any communities with populations exceeding 2000 and so did not include them in our data set.

Explanations alternative to those of Zipf (1941, 1949) are possible for nonlinear plots of log-community size against log-rank. They include the hypothesis that the data are drawn from mixtures of Pareto distributions. For example, the distribution of Australian community sizes in 1921 would have reflected the strictures of development in six largely autonomous British colonies, which had been federated into a “mixture” only 20 years previously.

Acknowledgment. We are particularly grateful to Ms. R. Boyce for obtaining the Australian population data used in Section 3.

REFERENCES

CSÖRGŐ, S., DEHEUVELS, P. and MASON, D. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.* **13** 1050–1077.

DAVID, H. A. (1970). *Order Statistics.* Wiley, New York.

DAVISON, A. C. (1984). Modelling excesses over high thresholds, with an application. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 461–482. Reidel, Dordrecht.

DE HAAN, L. and RESNICK, S. I. (1980). A simple asymptotic estimate for the index of a stable distribution. *J. Roy. Statist. Soc. Ser. B* **42** 83–88.

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

HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32** 177–203.

HALL, P. and WELSH, A. H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13** 331–341.

HILL, B. M. (1970). Zipf’s law and prior distributions for the composition of a population. *J. Amer. Statist. Assoc.* **65** 1220–1232.

HILL, B. M. (1974). The rank frequency form of Zipf’s law. *J. Amer. Statist. Assoc.* **69** 1017–1026.

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

HILL, B. M. and WOODROOFE, M. (1975). Stronger forms of Zipf’s law. *J. Amer. Statist. Assoc.* **70** 212–229.

HOSKING, J. R. M. and WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29** 339–349.

HOSKING, J. R. M., WALLIS, J. R. and WOOD, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27** 251–261.

NERC (1975). *Flood Studies Report* **1**. Natural Environment Research Council, London.

PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131.

ROOTZEN, H. and TAJVIDI, N. (1997). Extreme value statistics and wind storm losses: a case study. *Scand. Actuar. J.* 70–94.

SMITH, R. L. (1984). Threshold methods for sample extremes. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 621–638. Reidel, Dordrecht.

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90.

SMITH, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statist. Sci.* **4** 367–393.

TEUGELS, J. L. (1981). Limit theorems on order statistics. *Ann. Probab.* **9** 868–880.

TODOROVIC, P. (1978). Stochastic models of floods. *Water Resource Research* **14** 345–356.

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815.

WICKENS, C. H. (1921). *Census of the Commonwealth of Australia, 1921* **1**. H. J. Green, Government Printer, Melbourne.

ZIPF, G. K. (1941). *National Unity and Disunity: the Nation as a Bio-Social Organism*. Principia Press, Bloomington, IN.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO
CANADA M55 3G3

CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au