



On Statistical Inference Based on Record Values

ANDREY FEUERVERGER

Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3

PETER HALL

Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200 Australia

Abstract. We develop methodology for conducting inference based on record values and record times derived from a sequence of independent and identically distributed random variables. The advantage of using information about record times as well as record values is stressed. This point is a subtle one, since if the sampling distribution F is continuous then there is no information at all about F in the record times alone; the joint distribution of any number of them does not depend on F . However, the record times and record values jointly contain considerably more information about F than do the record values alone. Indeed, in the case of a distribution with regularly varying tails, the rate of convergence of the exponent of regular variation is two orders of magnitude faster if information about record times is included. Optimal estimators and convergence rates are derived under simple, specific models, and shown to be surprisingly robust against significant departures from those models. However, even under our special models the estimators have irregular properties, including an undefined information matrix. To some extent these difficulties may be alleviated by conditioning and by considering the relationship between maximum likelihood and maximum probability estimators.

Key words. central limit theorem, convergence rate, extreme value, information matrix, order statistic, regular variation, record time

AMS 1991 Subject Classifications. Primary—60G70, 62G05;
Secondary—62G09.

1. Introduction

Some statistical data are most easily accessed in terms of record values. That is to say, a particular data point is archived for posterity if it exceeds all others recorded in the past, but perhaps is not stored so accessibly otherwise. Indeed, data that are not records are sometimes not available at all. The times achieved in athletic events are a case in point, although of course one is then interested in the equivalent problem of minima rather than maxima. Such data have been analyzed as record values of sequences of independent and identically distributed random variables; see particularly Tryfos and Blackmore (1985), and also De Haan and Verkade (1987) and Smith (1988) who looked at independent observations with a trend. Related analyses of sporting events have been conducted by Chatterjee and Chatterjee (1982), Morton (1983) and Ballerini and Resnick (1985, 1987).

Record value data arise naturally in a variety of other contexts. For example, Smith (1988) noted that some hydrological and materials-testing data are of this form, and similar data also arise in real-time machine monitoring, where only record values are stored. There is a particularly extensive literature on the study of record values and related

data in terms of stochastic processes, to which two major contributors have been Shorrock (1972, 1973, 1974, 1975) and Resnick (1973a,b,c; 1975). A thorough account of major aspects of the stochastic theory of records may be found in Chapter 4 of Resnick (1987). Contributions of a more statistical nature are discussed for example by Smith and Miller (1986) and Smith (1988).

Our interest in this topic was aroused by a seminal paper of Berred (1992), who (apparently for the first time) considered estimation of tail parameters from record value data. There are at least two striking aspects to Berred's work. First of all, his estimates of the exponent of regular variation are based solely on record values, not at all on inter-record times, and indeed the best-performing of his estimates depends only on the most recent record value. Secondly, despite Berred's estimators being particularly close to those developed by other authors (e.g. Hill 1975, Pickands 1975) based on a full set of data, Berred does not need to employ a "smoothing parameter" in their construction. (A reader familiar with the analysis of statistical properties of Hill-type estimators, for example that given by Hall (1982) and Csörgő, Deheuvels and Mason (1985), will be aware that those estimators do not achieve good performance unless they are confined to a relatively small number of extreme values.)

It turns out that the best-performing of Berred's (1992) estimators is asymptotically equivalent to a maximum likelihood estimator under a particularly simple model for regular variation, assuming that only record value data are available; and that this estimator is remarkably robust against even substantial departures from the model. However, the estimator's convergence rate improves strikingly, by two orders of magnitude, if information about inter-record times is incorporated in an appropriate way. This is remarkable, since there is no information in record times themselves about the sampling distribution. If the sampling distribution F is continuous then the joint distribution of record times does not depend in any way on F . Nevertheless, there is crucial information about F in the joint distribution of inter-record times and record values, and it is this which makes possible the substantial improvement in convergence rates.

The estimators that attain this good performance are based on maximum likelihood estimators under specific models, but have such a high degree of robustness that they continue to enjoy excellent properties even when those models are significantly in error. However, even under the special models the maximum likelihood estimators are of particular interest because of their irregular features. The information matrix is not well-defined, and the usual Cramér–Rao theory is inapplicable, although the estimators are asymptotically equivalent to those derived by the maximum probability method and so have asymptotically minimum variance among estimators that are asymptotically Normal. Somewhat similar difficulties in a related setting have been noted by Smith (1985).

All these results are achieved without any need for "statistical smoothing." Nevertheless, unlike earlier work our results do demonstrate the effect that departures from the assumed model (which produced the estimators in the first place) have on bias, and we note that this impact can be reduced by smoothing appropriately. We choose the manner of departure from the central model so as to produce bias terms of an explicit polynomial order, which may be readily compared with the error-about-the-mean term.

Our primary emphasis is on inference in models whose tails decrease in a polynomial

fashion, although related work for exponentially decreasing tails will be mentioned. Results are similar in both cases, except that in the context of exponential tails there is a greater variety of situations to consider and so a thorough examination is more tedious. Note particularly that the distribution of the number of record values observed in a sequence of n independent and identically distributed, continuous random variables does not depend at all on the sampling distribution, and so the number of record values available for inference is not affected by assumptions about the sampling distribution.

Our main results are described in Section 2. Numerical work confirming their features and elucidating small-sample properties is presented in Section 3. All technical arguments are deferred to Section 4.

2. Main results

2.1. Summary

We begin by describing the type of data available. Let Y_1, Y_2, \dots and $L(1), L(2), \dots$ denote the consecutive record values and record times, respectively, observed in a sequence of independent random variables $\{X_1, X_2, \dots\}$ with common continuous distribution function F . By convention, $L(1) = 1$ and $Y_1 = X_1$, and then for $i \geq 2$, $L(i) = \inf\{j > L(i-1) : X_j > X_{L(i-1)}\}$, and $Y_i = X_{L(i)}$. Assume that the first N of the Y_i 's are observed.

Two cases may be identified—that where N denotes the number of records in a data sequence $\{X_1, \dots, X_n\}$ of given length n , so that N is a random variable; and that where N is predetermined. Statistical inference in the latter context has been considered before. See Berred (1992).

Sections 2.2 through 2.4 treat inference based on distributions with regularly varying tails, although we usually interpret the term “regularly varying” very generally, not in the restricted, classical sense discussed by, for example, Bingham, Goldie and Teugels (1987). Only in Section 2.3 do we demand a strict form of regular variation, and that is solely for the purpose of motivating estimators, not for data modeling. Our main interest is in estimating the shape parameter, α , of a distribution function F given by

$$F(x) = 1 - x^{-\alpha}K(x) \text{ where } \log K(x) = o(\log x) \quad (2.1)$$

as $x \rightarrow \infty$, and we do not insist that K be slowly varying. Similar methods and results in the case of distributions with exponentially decreasing tails are addressed in Section 2.5.

Section 2.2 discusses estimators based solely on the record values Y_1, Y_2, \dots , while Sections 2.3 and 2.4 study inference when both the Y_i 's and the record times $L(i)$ are available. Section 2.3 treats maximum likelihood estimation under the special model $F(x) = 1 - dx^{-\alpha}$, principally for the purpose of producing estimators whose special properties are later investigated under very general models in Section 2.4.

The estimators obtained assuming the special model are consistent under particularly general conditions. In fact, if both record values and record times are employed then our

estimator of α converges at rate $O_p(N^{-1})$, under even the very mild assumption (2.1). The convergence rate equals $O_p(N^{-3/2})$ if we ask of K in (2.1) that $K(x) = \text{const.} + O\{(\log x) - 1/2\}$ as $x \rightarrow \infty$. In general, failure of the special model influences principally the bias of the estimator, and does not have a first-order effect on variance, which continues to be close to its optimum.

The convergence rate $O_p(N^{-3/2})$ is a marked improvement on the optimal rate of only $O_p(N^{-1/2})$ available using record values alone. However, the constant multiplier of variance is greater by a factor of 12 in the case where the actual rate is faster, indicating that the estimator based on both record values and record times may not always have superior performance in small samples.

2.2. Estimators based on record values alone

We begin by suggesting *ad hoc* estimators based on weighted averages of logarithms of record values, and describe their performance under models that are substantially more general than asking that the upper tail of F be regularly varying. Then we observe that the best-performing of these estimators may be motivated as (approximately) a maximum likelihood estimator under a restricted model for regular variation, thereby explaining its relatively low variance. Nevertheless, under our more general models the latter estimator has bias similar to that of many competitors that are based on weighted means of logged record values. Furthermore, its variance, but not its bias, is robust against departures from the model.

One class of estimators of α may be defined as follows. Given a nonnegative, piecewise-continuous function w on $[0, 1]$ with left- and right-hand limits everywhere, and not identically zero, define $w_i = w(i/N)$ for $1 \leq i \leq N$, and let

$$\tilde{\alpha} = \left(\sum_{i=1}^N i w_i \right) / \left(\sum_{i=1}^N w_i \log Y_i \right),$$

the inverse of a weighted sum of logged record values. Consider also the version of $\tilde{\alpha}$ that arises when $w_N = 1$ and $w_i = 0$ otherwise; call the resulting estimator $\tilde{\alpha}_0$. This is one of the estimators proposed by Berred (1992), and of all the estimators that he considered it has, in the context of his assumptions, least asymptotic variance. One might suggest that judicious choice of w in the definition of α should produce an estimator that outperforms $\tilde{\alpha}_0$, but this argument does not take into account the particularly high positive correlation among successive record values.

We assume of the distribution F that

$$1 - F(x) = x^{-\alpha} K(x) \text{ where } \log K(x) = c(\log x)^\beta + o\{(\log x)^\beta\} \quad (2.2)$$

as $x \rightarrow \infty$, with $\alpha > 0$, $0 \leq \beta < 1$ and $-\infty < c < \infty$. This condition is more general than restrictions imposed by Berred, and encompasses cases where the contribution of bias

terms to mean squared error dominates that deriving from variance. To describe these properties, let

$$I_\xi = \int_0^1 u^\xi w(u) du, \quad J = \int_0^1 \left\{ \int_u^1 w(u) du \right\}^2 du,$$

$$b = -c\alpha^{1-\beta} I_\beta I_1^{-1} \text{ and } s^2 = \alpha^2 J I_1^{-2}.$$

Theorem 2.1: *Assume condition (2.2), and let ζ denote a random variable whose asymptotic distribution is Normal $N(0, 1)$. (a) If N is non-random then we may write $\tilde{\alpha} - \alpha = bN^{\beta-1} + sN^{-1/2}\zeta + o_p(N^{\beta-1} + N^{-1/2})$ as N increases. (b) If N equals the number of records in a sequence of n independent data values X_i , then $\tilde{\alpha} - \alpha = b(\log n)^{\beta-1} + s(\log n)^{-1/2}\zeta + o_p\{(\log n)^{\beta-1} + (\log n)^{-1/2}\}$ as n increases. (c) The asymptotic variance s^2 always exceeds α^2 , and indeed $\tilde{\alpha}$ never achieves the asymptotic performance of $\tilde{\alpha}_0$, which in context (a) above satisfies $\tilde{\alpha}_0 - \alpha = bN^{\beta-1} + \alpha N^{-1/2}\zeta + o_p(N^{\beta-1} + N^{-1/2})$, with an analogous formula (where $\log n$ replaces N) in case (b).*

The superior performance of $\tilde{\alpha}_0$ is more easily appreciated when it is noted that, under the restricted model $F(x) = 1 - dx^{-\alpha}$ for a constant d , $\tilde{\alpha}_0$ is asymptotically equivalent to the maximum likelihood estimator based solely on the record values (Y_1, \dots, Y_N) . Indeed, under this model, and regarding N as fixed, the random variables in the sequence

$$\mathcal{A} = \{\log(Y_1/d^{1/\alpha}), \log(Y_2/Y_1), \log(Y_3/Y_2), \dots, \log(Y_N/Y_{N-1})\}$$

are independent and exponentially distributed with mean α^{-1} , whose maximum likelihood estimator (for the data in \mathcal{A}) is the sequence mean,

$$N^{-1} \left\{ \log(Y_1/d^{1/\alpha}) + \sum_{i=1}^{N-1} \log(Y_{i+1}/Y_i) \right\} = \tilde{\alpha}_0^{-1} - (\alpha N)^{-1} \log d.$$

Under the restricted model $F(x) = 1 - dx^{-\alpha}$ the parameter d cannot be estimated consistently from record values alone.

2.3. Maximum likelihood estimation using both record values and inter-record times

We temporarily assume the restricted model $F(x) = 1 - dx^{-\alpha}$, say for $x \geq x_0$. Let $f(x) = \alpha dx^{-\alpha-1}$ denote the corresponding probability density. Given integers $1 \leq m \leq n$ and $1 = l(1) < \dots < l(m) \leq n$, and real numbers y_1, \dots, y_m satisfying $x_0 \leq y_1 < \dots < y_m < \infty$, write

$$\mathcal{L}_1\{\alpha, d|m; l(2), \dots, l(m); y_1, \dots, y_m\} dy_1 \dots dy_m$$

for the probability that $N = m$, $L(j) = l(j)$ for $2 \leq j \leq m$, and $Y_j \in (y_j, y_j + dy_j)$ for $1 \leq j \leq m$. If N equals the number of record values in a sequence of n independent data, and so is random, put $L(N+1) = n+1$; and if N is nonrandom, take $L(N+1) = L(N) + 1$. (This notation will greatly simplify our formulae, but to avoid ambiguity it should not be assumed when we make assertions about quantities such as $E\{L(i+1) - L(i) - 1\}$, for general i .) Then $\mathcal{L}_1 = R_1 \mathcal{L}_2$, where R_1 denotes a function of the data alone, not depending on α or d , and

$$\mathcal{L}_2\{N; L(2), \dots, L(N); Y_1, \dots, Y_N\} = \prod_{i=1}^N \{f(Y_i)F(Y_i)^{L(i+1)-L(i)-1}\}. \quad (2.3)$$

Indeed, \mathcal{L}_2 is proportional to the likelihood of $\mathbf{V} = (L(2), \dots, L(N), Y_1, \dots, Y_N)$, conditional on $(N, L(2), \dots, L(N))$, since the marginal distribution of the latter vector does not depend on the unknown F and so is subsumed into R_1 .

The fact that the parametric form assumed of F is only valid for $x \geq x_0$ is problematical. Even if that form were available for the entire distribution, the value of x_0 would be a function of unknown parameters and so inference would be awkward. This difficulty may be alleviated, but not eliminated, by conditioning on the smallest data value, Y_1 . The likelihood of \mathbf{V} conditional on Y_1 is given by $R_2 \mathcal{L}_3$, where R_2 does not depend on α or d , and

$$\mathcal{L}_3\{N; L(2), \dots, L(N); Y_1, \dots, Y_N\} = F(Y_1)^{L(2)-2} \prod_{i=2}^N \{f(Y_i)F(Y_i)^{L(i+1)-L(i)-1}\}.$$

Defining $\mathcal{L} = \log(\mathcal{L}_3 \prod_{i=2}^N Y_i)$ we have

$$\begin{aligned} \mathcal{L}(\alpha, d) &= \mathcal{L}\{\alpha, d | N; L(2), \dots, L(N); Y_1, \dots, Y_N\} \\ &= (N-1) \log(\alpha d) + \sum_{i=1}^N \{L(i+1) - L(i) - 1\} \\ &\quad \times \log(1 - dY_i^{-\alpha}) - \alpha \sum_{i=2}^N \log Y_i. \end{aligned}$$

Differentiating twice with respect to α and d we deduce that $\mathcal{L}(\alpha, d)$ is concave in both variables, and so is maximized at a unique point $(\hat{\alpha}, \hat{d})$ which solves the following two equations in (α, d) :

$$\sum_{i=1}^N \{L(i+1) - L(i) - 1\} (1 - dY_i^{-\alpha})^{-1} = L(N+1) - 2, \quad (2.4a)$$

$$\begin{aligned} & (N - 1)\alpha^{-1} + \sum_{i=1}^N \{L(i + 1) - L(i) - 1\} (1 - dY_i^{-\alpha})^{-1} \log Y_i \\ &= \sum_{i=1}^N \{L(i + 1) - L(i)\} \log Y_i - \log Y_1. \end{aligned} \tag{2.4b}$$

This account is an oversimplification, since it ignores the fact that $F(x) = 1 - dx^{-\alpha}$ must be nonnegative. Even if the latter parametric form were available for the entire distribution we would require $d^{1/\alpha} \leq x_0$, and hence we should impose the restriction $\hat{d}^{1/\hat{\alpha}} \leq Y_1$ on the estimators $(\hat{\alpha}, \hat{d})$. Of course, with probability tending to one as N increases this inequality will be satisfied, but nevertheless the solution of equations (2.4) is not necessarily the strictly-defined maximum likelihood estimator of (α, d) . This difficulty makes itself felt by rendering the (unconditional) information matrix indefinite, as we shall show in the next paragraph. Additionally it makes invalid the standard argument for employing the inverse of the information matrix as an asymptotic variance bound, although as we shall shortly see there are ways of circumventing this problem.

We claim that each element of the information matrix is infinite. To appreciate why, let us consider the simpler case where N is non-random. There, with $M_i = L(i + 1) - L(i) - 1$ we have

$$-E(\partial^2 \mathcal{L} / \partial \alpha^2) = \alpha^{-2}(N - 1) + \sum_{i=1}^{N-1} E\{M_i(1 - dY_i^{-\alpha})^{-2} dY_i^{-\alpha} (\log Y_i)^2\}.$$

Now, $E(M_i | Y_1, Y_2, \dots) = (dY_i^{-\alpha})^{-1} - 1$ for $i \geq 1$. (See Section 5 of Shorrock (1972), and note that in identities such as this we are not using the special notation for M_N .) Hence,

$$-E(\partial^2 \mathcal{L} / \partial \alpha^2) = \alpha^{-2}(N - 1) + \sum_{i=1}^{N-1} E\{(1 - dY_i^{-\alpha})^{-1} (\log Y_i)^2\}. \tag{2.5a}$$

Similarly,

$$-E(\partial^2 \mathcal{L} / \partial d^2) = d^{-2}(N - 1) + d^{-1} \sum_{i=1}^{N-1} E\{(1 - dY_i^{-\alpha})^{-1} Y_i^{-\alpha}\}, \tag{2.5b}$$

$$-E(\partial^2 \mathcal{L} / \partial \alpha \partial d) = -d^{-1} \sum_{i=1}^{N-1} E\{(1 - dY_i^{-\alpha})^{-1} \log Y_i\}. \tag{2.5c}$$

The first term in each of the series on the right-hand sides of equations (2.5) is infinite. This problem is not alleviated by constructing the information matrix in its alternative form, $E(\mathbf{W}\mathbf{W}^T)$, where $\mathbf{W} = (\partial \mathcal{L} / \partial \alpha, \partial \mathcal{L} / \partial d)^T$.

However, all subsequent terms (i.e. those corresponding to $i \geq 2$) are finite. Thus, if in conducting inference we were to suppress the first record value Y_1 (which is only a record

by convention, and not in the more accepted sense of exceeding the most recent record) then we would obtain a finite information matrix. Denote the corresponding values of $-E(\partial^2 \mathcal{L} / \partial \alpha^2)$, $-E(\partial^2 \mathcal{L} / \partial d^2)$ and $-E(\partial^2 \mathcal{L} / \partial \alpha \partial d)$ by a_1, a_2 and a_3 , respectively. Since, as $i \rightarrow \infty$, $E\{(1 - dY_i^{-\alpha})^{-1}(\log Y_i)^k\} \sim (\alpha^{-1} \log i)^k$ for $k = 1, 2$, and $E\{(1 - dY_i^{-\alpha})^{-1}Y_i^{-\alpha}\} \rightarrow 0$, then $a_1 \sim \frac{1}{3}\alpha^{-2}N^3$, $a_2 \sim d^{-2}N$ and $a_3 \sim -\frac{1}{2}\alpha^{-1}d^{-1}N^2$. Inverting the resulting matrix we find that the asymptotic variances of $\hat{\alpha}$ and \hat{d} are suggested to be given by

$$\text{asyp. var}(\hat{\alpha}) \sim 12\alpha^2N^{-3}, \quad \text{asyp. var}(\hat{d}) \sim 4d^2N^{-1}, \quad (2.6)$$

and the correlation to be $3^{1/2}/2$. The indicated asymptotic variance of $\hat{\alpha}$ is *two* orders of magnitude smaller than that of the estimators considered in Theorem 2.1, being of order N^{-3} rather than N^{-1} .

One way of overcoming the problem of an ill-defined covariance matrix is to calculate the conditional information, given Y_1 , thus removing the expectation operator from the first term in each series on the right-hand sides of equations (2.5). We also remark that for purposes of inference conditional on the record times, the expectations in (2.5) may be evaluated conditionally on the observed record time values; the required conditional distributions are readily derived from the joint distribution results given in Shorrock (1972).

Next we note that, even though the unconditional information matrix is not well-defined, the maximum likelihood estimators do enjoy optimal properties.

Theorem 2.2: *Assuming the model $1 - F(x) = dx^{-\alpha}$ for all sufficiently large x , the estimators of α and d defined by solving equations (2.4) are asymptotically equivalent to maximum probability estimators, and so enjoy minimum variance among asymptotically Normal estimators.*

The theory of maximum probability estimators was developed by Weiss and Wolfowitz (1967, 1973, 1974). A proof of Theorem 2.2 is relatively straightforward and so is not given here. That our estimators are indeed asymptotically Normal follows from Theorem 2.3 below.

We conclude this section by noting an approximation to the likelihood equations (2.4) which allows d to be eliminated. Put $Q_1 = L(N + 1) - 2$,

$$Q_2 = \sum_{1 \leq i \leq N} \{L(i + 1) - L(i)\} \log Y_i - \log Y_1 - (N - 1)\alpha^{-1}.$$

By Taylor expansion, equations (2.4) may be written as

$$\sum_{i=1}^N M_i(1 + dY_i^{-\alpha}) = Q_1 + O_p(1), \quad \sum_{i=1}^N M_i(1 + dY_i^{-\alpha}) \log Y_i = Q_2 + O_p(1).$$

(Here we have used the fact that $\sum_{1 \leq i \leq N} M_i Y_i^{-2\alpha} \log Y_i = O_p(1)$, which may be established under the model $F(x) = 1 - dx^{-\alpha}$.) If we ignore the $O_p(1)$ terms above then the approximate likelihood equations are linear in d , which may then be eliminated in the obvious way. This leads to the following equation in α alone:

$$(N - 1) \left(\sum_{i=1}^N M_i Y_i^{-\alpha} \log Y_i \right) - \left\{ \sum_{i=2}^N \log Y_i - (N - 1)\alpha^{-1} \right\} \left(\sum_{i=1}^N M_i Y_i^{-\alpha} \right) = 0. \quad (2.7)$$

The left-hand side is not a convex function of α , but with probability tending to 1 equation (2.7) has a solution $\bar{\alpha}$ in the neighborhood of the true value of α which, under the model $F(x) = 1 - dx^{-\alpha}$, admits the same first-order limiting behavior as $\hat{\alpha}$. Indeed, like $\hat{\alpha}$, $\bar{\alpha}$ is consistent for α under conditions that are substantially more general than those of our special model, as we shall show in the next section.

2.4. Asymptotic theory under general models

We begin by demonstrating asymptotic Normality of the estimators $(\hat{\alpha}, \hat{d})$, defined as the solution of equations (2.4), under models that are substantially more general than that which lead to (2.4). We ask that for constants $\alpha, \alpha_1, d > 0$ and $-\infty < d_1 < \infty$,

$$1 - F(x) = x^{-\alpha} K(x) \text{ where } K(x) = d[1 + d_1(\log x)^{-\alpha_1} + o\{(\log x)^{-\alpha_1}\}]. \quad (2.8)$$

Our next theorem shows that the asymptotic normality anticipated by Theorem 2.2 holds under this condition, which is considerably more general than that assumed there. In fact, if $\alpha_1 > \frac{1}{2}$ then the estimators $\hat{\alpha}$ and \hat{d} are asymptotically Normal with means α and d and variances given by (2.6).

Define $b_\alpha = -6\alpha^{1-\alpha_1} d_1 (1 - \alpha_1)^{-1}$ and $b_d = -4\alpha^{-\alpha_1} d d_1 (1 - \alpha_1)^{-1}$ if $0 < \alpha_1 \leq \frac{1}{2}$, and $b_\alpha = b_d = 0$ otherwise.

Theorem 2.3: *Assume condition (2.8), and let (ζ_1, ζ_2) denote random variables with a limiting bivariate Normal distribution having zero means, unit variances and correlation coefficient $3^{1/2}/2$. (a) If N is non-random then we may write*

$$\hat{\alpha} - \alpha = b_\alpha N^{-(\alpha_1+1)} + 12^{1/2} \alpha N^{-3/2} \zeta_1 + o_p(N^{-(\alpha_1+1)} + N^{-3/2}), \quad (2.9a)$$

$$\hat{d} - d = b_d N^{-\alpha_1} + 2d N^{-1/2} \zeta_2 + o_p(N^{-\alpha_1} + N^{-1/2}) \quad (2.9b)$$

as N increases. (b) If N equals the number of records in a sequence of n independent data values X_i , then (2.9) continues to hold provided that N is replaced throughout by $\log n$.

It may also be proved that under (2.8) there exists, with probability tending to 1, a

solution $\bar{\alpha}$ of equation (2.7) in any given neighborhood of α ; and that any sequence of such solutions satisfies (2.9a), with $\bar{\alpha}$ there replacing $\hat{\alpha}$.

For relatively small values of α_1 , where bias dominates error about the mean, it is possible to reduce the effect of bias and reduce the order of mean squared error by employing only relatively extreme order statistics in the construction of $\hat{\alpha}$. In particular, one could parallel arguments of Hall (1982) in the context of estimating α from a full data set. However, there are at least three reasons for not pursuing that matter in the present context. First, determination of the optimal cut-off demands estimation of second-order terms in (2.8), and in the context of inference using record values one typically has relatively little data with which to do this. Secondly, if it should happen (as is commonly assumed in the case of a full data set) that the function K in (2.8) satisfies $K(x) = d + O(x^{-\varepsilon})$ for some $d, \varepsilon > 0$, then (2.8) holds with $\alpha_1 > \frac{1}{2}$ and so the bias contribution is negligible. And thirdly, even in the extreme case $\alpha_1 = 0$ the estimator $\hat{\alpha}$ converges at rate $O_p(N^{-1})$ (see Theorem 2.4), so there is not a pressing need for such modification.

The first terms on the right-hand sides of equation (2.9), i.e. those of orders $N^{-(\alpha_1+1)}$ and $N^{-\alpha_1}$, represent the dominant contributions to the biases of $\hat{\alpha}$ and \hat{d} , respectively. The terms involving ζ_j , and of orders $N^{-3/2}$ and $N^{-1/2}$ respectively, are the results of errors about the mean. In each case these errors dominate those from bias if and only if $\alpha_1 > \frac{1}{2}$. It is clear that as α_1 gets closer to zero, the error in \hat{d} that results from bias increases without bound until, in the limit $\alpha_1 = 0$, \hat{d} is no longer consistent. This comes as no surprise, but it is perhaps unexpected that the same should not be true of the estimator of α . Our next result investigates this matter further, and shows that both $\hat{\alpha}$ and the estimator $\bar{\alpha}$ defined by (2.7) admit a relatively fast convergence rate to α under very general conditions.

Theorem 2.4: *Assume condition (2.1), and let $\mathcal{N} = (\alpha - \varepsilon, \alpha + \varepsilon)$ denote any neighborhood of the true parameter value α , where $\varepsilon > 0$. Then with probability tending to one there exists a solution $\bar{\alpha}$ of (2.7) which lies within \mathcal{N} ; and if $\bar{\alpha} = \bar{\alpha}_N$ denotes any sequence of such solutions, $\bar{\alpha} - \alpha = O_p(N^{-1})$. The same rate of convergence applies to $\hat{\alpha}$, the solution of equations (2.4).*

2.5. Distributions with exponentially small tails

In many respects the situation here is similar to that for distributions with regularly varying tails, and so we consider only the main features. In place of (2.1) or (2.2) let us assume that

$$1 - F(x) = c_1 \exp\{-x^\alpha K(x)\} \text{ where } K(x) = c + o(1)$$

as $x \rightarrow \infty$, and $\alpha, c, c_1 > 0$. Curiously, in this setting a surprisingly minor modification of the estimator $\tilde{\alpha}$ defined in Section 2.2 is consistent for α , although it is so heavily biased that it has a poor convergence rate. Indeed, defining here

$$\tilde{\alpha} = \left(\sum_{i=1}^N w_i \log i \right) / \left(\sum_{i=1}^N w_i \log Y_i \right),$$

one may show that for a very wide range of choices of the weights, w_i , $\tilde{\alpha} \rightarrow \alpha$ as N increases. As in Section 2.2 the optimal choice of weights is $w_N = 1$ and $w_i = 0$ for $1 \leq i \leq N - 1$, and in that case, $\tilde{\alpha} = \alpha + \alpha(\log c)(\log N)^{-1} + o_p\{(\log N)^{-1}\}$.

Of course, $\tilde{\alpha}$ is no longer related to a maximum likelihood estimator of α , and the logarithmic convergence rate that it exhibits can be substantially improved upon. The approach suggested by work in earlier sections is to develop a maximum likelihood estimator based on a specific model, and then determine the extent to which the estimator's properties are preserved under departures from the model. While this method promises to work well in many instances of distributions with exponentially decreasing tails, a detailed account of its main features is complicated by the relatively wide variety of models available. If the model depended only on shape and scale then it might be appropriate to employ only two parameters, so that the assumed density of the sampling distribution could be $c_1(\alpha, c) \exp(-cx^\alpha)$ where the function c_1 was given. However, in other circumstances c_1 might not be expressible *a priori* as a known function of α and c , and so a model involving at least three parameters would need to be fitted. Nevertheless, the convergence rates of estimators derived under such models are generally polynomial rather than logarithmic, although the polynomials can include logarithmic factors. As in the case of distributions with regularly varying tails, the rates improve if record times as well as record values are included in the estimators in an appropriate way.

3. Numerical properties

The procedures discussed in Section 2 were implemented using the S statistical software (Becker, Chambers and Wilks, 1988). Maximum likelihood estimation was implemented in two distinct ways: firstly using a Newton-Raphson procedure based on equations (2.4), with due attention to numerical stability and parameter boundaries, and secondly using a successive grid search procedure based on the corresponding log-likelihood function. We also implemented the estimator $\bar{\alpha}$ of equation (2.7), again using a Newton-Raphson method, as well as the estimator $N/\log Y_N$ due to Berred (1992). Random vectors (Y_1, Y_2, \dots, Y_N) were generated from exponentially distributed random variables using Proposition 4.1(ii) of Resnick (1987, p. 165); a description of this method is given in the proof of Theorem 2.1 below. The corresponding vectors (L_1, L_2, \dots, L_N) were then generated using the conditional geometric distribution for inter-record times (Shorrock, 1972, Section 5).

Table 1 summarizes our Monte Carlo trials for estimation of α by means of the Newton-Raphson based MLE, Newton-Raphson solution for the estimate $\bar{\alpha}$ based on (2.7), and Berred's estimator. For each combination of $\alpha = 1, 2, 5$ and $N = 5, 10, 20$, we generated a trial of 1000 samples from the distribution $F(x) = 1 - dx^{-\alpha}$ using the value $d = 1$. All iterations were started at the known true values. In the case of the MLE's, only those trials

Table 1. Summary of Monte Carlo trials for estimation of the parameter α by means of MLE (Newton-Raphson), equation (2.7), and Berred's estimator.

Estimator	α	N	Mean	Variance	MSE	No. trials
MLE	1.0	5	1.143	0.287	0.307	608
		10	1.0011	0.0147	0.0147	740
		20	0.9962	0.00159	0.00160	795
	2.0	5	2.29	0.82	0.91	561
		10	2.025	0.065	0.066	723
		20	1.993	0.0065	0.0066	815
	5.0	5	5.84	6.32	7.03	603
		10	5.054	0.47	0.477	738
		20	4.988	0.043	0.043	818
Eqn (2.7)	1.0	5	1.71	2.04	2.54	836
		10	1.150	0.098	0.120	956
		20	1.025	0.032	0.0038	998
	2.0	5	3.33	5.17	6.94	837
		10	2.34	0.54	0.66	956
		20	2.046	0.012	0.014	991
	5.0	5	8.40	34	45	832
		10	5.80	3.8	4.4	969
		20	5.12	0.077	0.090	996
Berred	1.0	5	1.27	0.78	0.85	1000
		10	1.111	0.16	0.17	1000
		20	1.060	0.062	0.065	1000
	2.0	5	2.52	2.09	2.36	1000
		10	2.26	0.59	0.66	1000
		20	2.11	0.26	0.27	1000
	5.0	5	6.36	11	14	1000
		10	5.63	3.7	4.2	1000
		20	5.23	1.5	1.6	1000

for which the Newton-Raphson algorithm converged successfully to a value on the $\hat{d}Y_1^{-\hat{\alpha}} \leq 1$ permissible region were used in computing the summary statistics; the number of such successful trials (out of 1000) is shown in the final column. For each Monte Carlo trial we computed the mean, variance, and MSE for each of the three estimators. As may be seen by comparisons with the asymptotic variance in (2.6), the trials behave more or less as expected once N increases beyond the value 5, with the asymptotics becoming very accurate for higher values of N . As may also be seen, the MLE outperforms the two other estimators for all values of N tabulated. The results also suggest that bias is a negligible problem—at least when the model specification is correct. Finally, we note that the estimator $\bar{\alpha}$ of (2.7) converged successfully more frequently than did the MLE.

These Newton-Raphson based trials were augmented by Monte Carlo experiments involving optimization of the log-likelihood function by means of a detailed successive grid search algorithm. Table 2 provides the results of these MLE trials for the values $d = 1$, $\alpha = 1, 2, 5$ and $N = 5, 8, 10, 12$. For the experiments reported here our grid search algorithm never failed to converge, but typically took more than ten times as long as Newton-Raphson optimization.

Table 2. Summary of Monte Carlo trials for estimation of the parameter α by means of MLE via successive grid-search.

Parameter values	N	Mean	Variance	MSE	No. trials
$\alpha = 1$	5	1.12	0.23	0.24	200
$d = 1$	8	1.06	0.051	0.055	200
	10	1.017	0.015	0.015	200
	12	1.009	0.007	0.007	200
$\alpha = 2$	5	2.50	0.85	1.10	100
$d = 1$	8	2.042	0.113	0.115	100
	10	2.075	0.062	0.068	100
	12	1.981	0.037	0.037	100
$\alpha = 5$	5	5.64	4.76	5.17	100
$d = 1$	8	5.36	0.83	0.96	100
	10	5.015	0.42	0.42	100
	12	5.075	0.18	0.19	100

To save space, we do not report here the Monte Carlo results for estimation of d by the two MLE procedures. In broad features, however, d proved to be much more variable and so more difficult to estimate than α , with sample sizes in excess of $N = 8$ being required to attain even tolerably accurate estimates. This is in accordance with the asymptotic variance result (2.6) and so was not unexpected.

Finally, in order to assess performance of the estimators under departures from the model, we implemented a Monte Carlo experiment by means of grid search for data from the distribution $F(x) = 1 - d(x^2 + x)^{-\alpha/2}$. This distribution was chosen in part for computational simplicity, but more importantly because it provides a high degree of departure from the MLE model. The results for these simulations, given in Table 3, show a marked degree of bias for α . (The estimates for d , of course, were entirely unreliable.) Interestingly, the results for Berred's estimator (not included here) show that it possesses less bias than the MLE in this case, for the sample sizes shown. This can be explained by the fact that it uses only the largest value of Y , whereas the distributional distortion introduced is greater at the lower end of values for the data. In practice, if the underlying

Table 3. Summary of Monte Carlo MLE via successive grid search for estimation of the parameter α in the model $F(x) = 1 - d(x^2 + x)^{-\alpha/2}$.

Parameter values	N	Mean	Variance	MSE	No. trials
$\alpha = 1$	5	0.34	0.019	0.45	200
$d = 1$	10	0.34	0.018	0.45	200
	15	0.43	0.017	0.34	200
$\alpha = 2$	5	0.61	0.038	1.97	200
$d = 1$	10	0.66	0.060	1.85	200
	15	0.75	0.063	1.61	200
$\alpha = 5$	5	1.18	0.11	14	200
$d = 1$	10	1.33	0.18	13	200
	15	1.85	0.34	10	200

model is highly uncertain, one may wish to use estimators based on the larger record values only, employing some form of ‘‘statistical smoothing’’; however, the ensuing lines of enquiry are not pursued here.

Further details of the numerical properties of the procedures may be found in a technical report available from the authors.

4. Proofs

4.1. Proof of Theorem 2.1

Put $\gamma = \alpha^{-1}$, $\tilde{\gamma} = \tilde{\alpha}^{-1}$ and $\tilde{\gamma}_0 = \tilde{\alpha}_0^{-1}$. It suffices to derive versions of the theorem for the estimators $\tilde{\gamma}$ and $\tilde{\gamma}_0$. To this end, define $G(x) = -\log\{1 - F(x)\}$ and $H(x) = G^{-1}(x) = \inf\{y : G(y) \geq x\}$. By Proposition 4.1 of Resnick (1987, p. 165), the random variables S_i defined by $Y_i = H(S_i)$, $i \geq 1$, are representable as $S_i = \sum_{1 \leq j \leq i} Z_j$, where the Z_j 's are independent and identically distributed with densities e^{-z} , $z > 0$. By (2.2), $H(x) = e^{\gamma x} K_1(x)$ where $\log K_1(x) = c\gamma^{\beta+1}x^\beta + o(x^\beta)$ as $x \rightarrow \infty$. Therefore,

$$\log H(x) = \gamma x + c\gamma^{\beta+1}x^\beta + o(x^\beta). \quad (4.1)$$

Furthermore, since the w_i 's are bounded then

$$\sum_{i=1}^N w_i \left(\sum_{j=1}^i Z_j \right)^\beta = N^{\beta+1} I_\beta + o_p(N^{\beta+1}).$$

From these results we may deduce that, with $V_i = Z_i - 1$ and

$$\Delta = \sum_{i=1}^N V_i \sum_{j=i}^N w_j,$$

we have

$$\sum_{i=1}^N w_i \log Y_i = \gamma \sum_{i=1}^N i w_i + \gamma \Delta + N^{\beta+1} c\gamma^{\beta+1} I_\beta + o_p(N^{\beta+1}).$$

This expansion and the fact that $\sum_{i=1}^N i w_i = N^2 I_1 + o_p(N^2)$ permit us to conclude that

$$\tilde{\gamma} - \gamma = -\gamma^2 b N^{\beta-1} + \gamma I_1^{-1} N^{-2} \Delta + o_p(N^{\beta-1} + N^{-2} |\Delta|). \quad (4.2)$$

If the data are recorded in such a manner that N is non-random, then Δ is asymptotically Normally distributed with zero mean and variance $N^3 J$, and so result (a) in the theorem follows directly from (4.2). Result (b) is a consequence of the fact that if the length n of the

data set is given the $N/\log n \rightarrow 1$ almost surely as n increases (see Propositions 4.1.3 and 4.1.4 of Resnick (1987, p. 190)); and Δ is asymptotically Normal $N\{0, (\log n)^3 J\}$. The latter result may be derived as a corollary of an invariance principle for the random function $\Delta(t)$, defined by replacing N in the definition of Δ by the integer part of $t \log n$, where $t \in (1 - \varepsilon, 1 + \varepsilon)$ for some fixed $\varepsilon > 0$.

The claimed asymptotic formulae for $\tilde{\alpha}_0$ are easily verified using a similar argument, and the result $s^2 \geq \alpha^2$ may be derived by standard variational methods.

4.2. Proof of Theorem 2.3

Let α and d denote the true values of those parameters, and initially, write $\hat{\alpha}$ and \hat{d} for general functions of the data

$$\mathcal{X} = \{L(2), \dots, L(N), Y_1, \dots, Y_N\}.$$

Given $\varepsilon > 0$, and defining $A = \hat{\alpha} - \alpha$ and $D = \hat{d} - d$, put

$$\mathcal{E} = \{|A| \leq N^{-1-\varepsilon}, |D| \leq N^{-\varepsilon}\}. \tag{4.3}$$

We shall prove that with probability tending to 1 the event \mathcal{E} applies to the maximum likelihood estimators of α and d . (At this point in our argument we shall take $(\hat{\alpha}, \hat{d})$ to be those maximum likelihood estimators.) Then we shall show that conditional on \mathcal{E} , the maximum likelihood estimators possess the limit laws contained in the theorem. This proves the theorem. During our argument, all remainder terms in our formulae that depend on i , $1 \leq i \leq N$, will be of the stated orders uniformly in i , given that the event \mathcal{E} obtains. These qualifiers will generally be omitted.

In view of (4.1), $\log Y_i = \log H(S_i) = O_p(N)$ uniformly in $1 \leq i \leq N$. (Of course, (2.8) implies (2.1), and hence (4.1), with $c = 0$ and any $\beta < 1$.) Therefore on \mathcal{E} ,

$$\begin{aligned} Y_i^{-\hat{\alpha}} &= Y_i^{-\alpha} [1 - A \log Y_i + O_p\{(NA)^2\}], \\ \hat{d} Y_i^{-\hat{\alpha}} &= d Y_i^{-\alpha} (1 + d^{-1} [D - dA \log Y_i + O_p\{(NA)^2 + N|AD|\}]). \end{aligned}$$

Therefore, with $M_i = L(i + 1) - L(i) - 1$ and for $k = 0$ or 1 ,

$$\begin{aligned} &\sum_{i=1}^N M_i (1 - \hat{d} Y_i^{-\hat{\alpha}})^{-1} (\log Y_i)^k - \sum_{i=1}^N M_i (1 - d Y_i^{-\alpha})^{-1} (\log Y_i)^k \\ &= \sum_{i=1}^N M_i (1 - d Y_i^{-\alpha})^{-2} Y_i^{-\alpha} (\log Y_i)^k [D - dA \log Y_i + O_p\{(NA)^2 + N|AD|\}]. \end{aligned}$$

More simply, $N\hat{\alpha}^{-1} - N\alpha^{-1} = O_p(N^{-\varepsilon})$. Hence, since $(\hat{\alpha}, \hat{d})$ solves equations (2.4), those equations are equivalent to $U_i = 0$ for $i = 1, 2$, where

$$U_1 = \sum_{i=1}^N M_i (1 - dY_i^{-\alpha})^{-2} Y_i^{-\alpha} [D - dA \log Y_i + O_p\{(NA)^2 + N|AD|\}] + V_1, \quad (4.4)$$

$$U_2 = \sum_{i=1}^N M_i (1 - dY_i^{-\alpha})^{-2} Y_i^{-\alpha} (\log Y_i) [D - dA \log Y_i + O_p\{(NA)^2 + N|AD|\}] + V_2 + O_p(N^{-\varepsilon}), \quad (4.5)$$

$$\begin{aligned} V_1 &= \sum_{i=1}^N M_i (1 - dY_i^{-\alpha})^{-1} - \{L(N+1) - 2\} \\ &= \sum_{i=1}^N M_i dY_i^{-\alpha} (1 - dY_i^{-\alpha})^{-1} - (N-1), \end{aligned}$$

$$\begin{aligned} V_2 &= (N-1)\alpha^{-1} + \sum_{i=1}^N M_i (1 - dY_i^{-\alpha})^{-1} \log Y_i - \sum_{i=1}^N (M_i + 1) \log Y_i + \log Y_1 \\ &= (N-1)\alpha^{-1} + \sum_{i=1}^N M_i dY_i^{-\alpha} (1 - dY_i^{-\alpha})^{-1} \log Y_i - \sum_{i=2}^N \log Y_i. \end{aligned}$$

Let $\mathcal{Y} = \{Y_1, Y_2, \dots\}$, and note that $\mu_i \equiv E(M_i/\mathcal{Y}) = \{1 - F(Y_i)\}^{-1} - 1$. We replace M_i in the definition of V_j by μ_i , for $1 \leq i \leq N-1$ (but not for $i = N$), and denote the resulting quantity by V'_j . It will emerge that V'_1 and V'_2 produce the dominant contributions to the bias of $\hat{\alpha}$ and \hat{d} . Now, under condition (2.8),

$$[\{1 - F(y)\}^{-1} - 1] dy^{-\alpha} (1 - dy^{-\alpha})^{-1} = 1 - d_1 (\log y)^{-\alpha_1} + o\{(\log y)^{-\alpha_1}\}$$

as $y \rightarrow \infty$. Therefore, since $i^{-1} \log Y_i \rightarrow \alpha^{-1}$ in probability as $i \rightarrow \infty$,

$$\begin{aligned} V'_1 &\equiv \sum_{i=1}^{N-1} [\{1 - F(Y_i)\}^{-1} - 1] dY_i^{-\alpha} (1 - dY_i^{-\alpha})^{-1} \\ &\quad + M_N dY_N^{-\alpha} (1 - dY_N^{-\alpha})^{-1} - (N-1) \\ &= \sum_{i=1}^{N-1} [1 - d_1 (\log Y_i)^{-\alpha_1} + o_p\{(\log Y_i)^{-\alpha_1}\}] + M_N (d^{-1} Y_N^\alpha - 1)^{-1} - (N-1) \\ &= - \sum_{i=1}^{N-1} \{d_1 (\alpha^{-1} i)^{-\alpha_1} + o_p(i^{-\alpha_1})\} + M_N (d^{-1} Y_N^\alpha - 1)^{-1}, \end{aligned} \quad (4.6)$$

$$\begin{aligned}
 V_2' &\equiv (N-1)\alpha^{-1} + \sum_{i=1}^{N-1} [\{1 - F(Y_i)\}^{-1} - 1] dY_i^{-\alpha} (1 - dY_i^{-\alpha})^{-1} \log Y_i \\
 &\quad + M_N dY_N^{-\alpha} (1 - dY_N^{-\alpha})^{-1} \log Y_N - \sum_{i=2}^N \log Y_i \\
 &= - \sum_{i=1}^{N-1} \{d_1(\alpha^{-1}i)^{-\alpha_1} + o_p(i^{-\alpha_1})\} \log Y_i + M_N (d^{-1}Y_N^\alpha - 1)^{-1} \log Y_N \\
 &\quad + (N-1)\alpha^{-1} + \log Y_1 - \log Y_N.
 \end{aligned} \tag{4.7}$$

We claim that

$$M_N (d^{-1}Y_N^\alpha - 1)^{-1} = O_p(1). \tag{4.8}$$

In the case of nonrandom N this result is trivial, since then, $M_N = L(N) + 1 - L(N) - 1 = 0$. When N is random, $M_N = n - L(N) \leq n$ and $(d^{-1}y^\alpha - 1)^{-1} \sim 1 - F(y)$ as $y \rightarrow \infty$, and so in that case (4.8) will follow if we prove that $n\{1 - F(Y_N)\} = O_p(1)$. But $1 - F(Y_N)$ has the distribution of the largest order statistic of a sample of size n from the Uniform distribution on the interval $(0, 1)$, and so $1 - F(Y_N) = O_p(n^{-1})$. This completes the proof of (4.8).

Combining (4.6)–(4.8) we deduce that

$$V_1' = aN^{\alpha_2} + o_p(N^{\alpha_2} + N^{1/2}), \quad V_2' = O_p(N), \tag{4.9}$$

where $(a, \alpha_2) = (-d_1\alpha^{-\alpha_1}(1 - \alpha_1)^{-1}, 1 - \alpha_1)$ if $0 < \alpha_1 \leq \frac{1}{2}$, and $a = 0$ otherwise.

The dominant contributions to the errors of $\hat{\alpha}$ and \hat{d} about their means are produced by $V_j'' = V_j - V_j'$ for $j = 1, 2$. To study these quantities in detail, observe that

$$\begin{aligned}
 V_j'' &= \sum_{i=1}^{N-1} (M_i - \mu_i) dY_i^{-\alpha} (1 - dY_i^{-\alpha})^{-1} (\log Y_i)^{j-1} \\
 &= \sum_{i=1}^{N-1} (M_i - \mu_i) \mu_i^{-1} A(Y_i) (\log Y_i)^{j-1},
 \end{aligned}$$

where $A(y) \equiv \{1 - F(y)\}^{-1} dy^{-\alpha} (1 - dy^{-\alpha}) \rightarrow 1$ as $y \rightarrow \infty$.

Note particularly that the mean of M_i , conditional on \mathscr{Y} , diverges to infinity as $i \rightarrow \infty$. If the random variable Q is geometrically distributed with mean μ then as $\mu \rightarrow \infty$, Q/μ is asymptotically distributed as Z , say, which enjoys the exponential distribution with unit mean. Furthermore, all the moments of Q/μ converge to the corresponding moments of Z . It follows that conditional on \mathscr{Y} , V_j'' is asymptotically distributed as

$$T_j = \sum_{i=1}^{N-1} (Z_i - 1)(\alpha^{-1}i)^{j-1},$$

where Z_1, Z_2, \dots are independent and identically distributed as Z ; or equivalently, that the conditional distribution of V_j'' is random. Normal with zero mean and variance $\sigma_j^2 = \alpha^{-2(j-1)}(2j-1)^{-1}N^{2j-1}$. This limiting distribution depends on \mathscr{Y} only through the value of N , and then only in the case where N is random. Moreover, in the case of random N , there exists a nonrandom sequence c_n such that $N/c_n \rightarrow 1$ in probability as $n \rightarrow \infty$. Hence, the unconditional asymptotic distribution of V_j'' is also Normal $N(0, \sigma_j^2)$. Combining this result with (4.9) we see that we may write

$$V_1 = aN^{\alpha_2} + T_1 + o_p(N^{\alpha_2} + N^{1/2}), \quad V_2 = T_2 + o_p(N^{3/2}). \quad (4.10)$$

Now we return to the definitions of U_1 and U_2 at (4.4) and (4.5), and simplify the first series in each. Employing the argument in the previous paragraph we see that for $j \geq 0$,

$$\sum_{i=1}^N (M_i - \mu_i)(1 - dY_i^{-\alpha})^{-2} Y_i^{-\alpha} (\log Y_i)^j = O_p(N^{(2j+1)/2}). \quad (4.11)$$

More simply, using the fact that (a) conditional on \mathscr{Y} the M_i 's are geometrically distributed with respective means μ_i ; and (b) in the case where N is a random variable, there exists a nonrandom sequence c_n such that $N/c_n \rightarrow 1$ in probability; we may deduce that

$$\sum_{i=1}^N M_i (1 - dY_i^{-\alpha})^{-2} Y_i^{-\alpha} |\log Y_i|^j = O_p(N^{j+1}). \quad (4.12)$$

The argument leading to (4.6) and (4.7), together with the facts that $\mu_i = \{1 - F(Y_i)\}^{-1} - 1$ and

$$[\{1 - F(y)\}^{-1} - 1]y^{-\alpha}(1 - dy^{-\alpha})^{-2} = d^{-1} + O\{(\log y)^{-\alpha_1}\},$$

produce the result

$$\begin{aligned} & \sum_{i=1}^N \mu_i (1 - dY_i^{-\alpha})^{-2} Y_i^{-\alpha} (\log Y_i)^j \\ &= d^{-1} \alpha^{-j} (j+1)^{-1} N^{j+1} + o_p(N^{j+1}). \end{aligned} \quad (4.13)$$

Combining results (4.11)–(4.13), and noting the relations between U_j and V_j at (4.4) and (4.5), we deduce that

$$U_1 - V_1 = d^{-1}ND - \frac{1}{2}\alpha^{-1}N^2A + o_p(N^2|A| + N|D|), \quad (4.14a)$$

$$U_2 - V_2 = \frac{1}{2}\alpha^{-1}d^{-1}N^2D - \frac{1}{3}\alpha^{-2}N^3A + o_p(N^3|A| + N^2|D|). \quad (4.14b)$$

The “ o ” remainder terms here are of the stated orders uniformly in those functions $\hat{\alpha}$ and \hat{d} of the data \mathcal{X} that satisfy the event \mathcal{E} , defined at (4.3).

Recall that, because of a convexity property, the likelihood equations (2.4) have no more than one solution; and that this is the solution of $(U_1, U_2) = (0, 0)$. In view of (4.10) and (4.14) there exists a solution of the likelihood equations that, with probability tending to 1, satisfies event \mathcal{E} , and so this must be the maximum likelihood estimator. It then follows from (4.10) and (4.11) that the latter estimator, for which we now reserve the notation $(\hat{\alpha}, \hat{d}) = (a + A, d + D)$, satisfies the matrix equation

$$\begin{pmatrix} -\frac{1}{2}\alpha^{-1} & d^{-1} \\ -\frac{1}{3}\alpha^{-2} & \frac{1}{2}\alpha^{-1}d^{-1} \end{pmatrix} \begin{pmatrix} N^2A \\ ND \end{pmatrix} = \begin{pmatrix} aN^{\alpha_2} + T_1 \\ N^{-1}T_2 \end{pmatrix} + \begin{pmatrix} o_p(N^{1/2}) \\ o_p(N^{1/2}) \end{pmatrix}.$$

Solving for A and D we deduce that

$$A = 6\alpha N^{\alpha_2-2} + N^{-2}6\alpha(T_1 - 2\alpha N^{-1}T_2) + o_p(N^{\alpha_2-2} + N^{-3/2}), \quad (4.15a)$$

$$D = 4adN^{\alpha_2-1} + N^{-1}2d(2T_1 - 3\alpha N^{-1}T_2) + o_p(N^{\alpha_2-1} + N^{-1/2}). \quad (4.15b)$$

The random vector $(N^{-1/2}(a_1T_1 + a_2\alpha N^{-1}T_2), N^{-1/2}(b_1T_1 + b_2\alpha N^{-1}T_2))$ has an asymptotic bivariate Normal distribution with zero means, respective variances $a_1^2 + a_1a_2 + \frac{1}{3}a_2^2$ and $b_1^2 + b_1b_2 + \frac{1}{3}b_2^2$, and covariance $a_1b_1 + \frac{1}{2}(a_1b_2 + a_2b_1) + \frac{1}{3}a_2b_2$. Theorem 2.3 follows from this result and (4.15).

4.3. Proof of Theorem 2.4

For the sake of brevity we treat only the case of $\bar{\alpha}$, and assume that N is non-random. Put $J = \log K$, and use notation introduced during the proofs of Theorems 2.1 and 2.3, except that $\bar{\alpha}$ replaces $\hat{\alpha}$. In particular, let $A = \bar{\alpha} - \alpha$ where α is the true parameter value and, initially, $\bar{\alpha}$ denotes *any* function of the data satisfying $|\bar{\alpha}| \leq C$ for any given constant $C > 0$. Since $1 - F(x) = x^{-\alpha} \exp\{J(x)\}$ then

$$x^{-\bar{\alpha}} = \{1 - F(x)\} \exp\{-A \log x - J(x)\}.$$

Furthermore, $\alpha i^{-1} \log Y_i \rightarrow 1$ in probability, and with $\mu_i = E(M_i | \mathcal{B}), \mu_i \{1 - F(Y_i)\} \rightarrow 1$. Hence, for $j = 0$ or 1 ,

$$\begin{aligned} s_j(\bar{\alpha}) &\equiv \sum_{i=1}^N M_i Y_i^{-\bar{\alpha}} (\log Y_i)^j = \sum_{i=1}^{N-1} M_i Y_i^{-\bar{\alpha}} (\log Y_i)^j \\ &= \{1 + o_p(1)\} \sum_{i=1}^{N-1} (M_i / \mu_i) \exp\{-A \log Y_i - J(Y_i)\} (\alpha^{-1} i)^j \\ &\quad + O_p\{\exp(-A \log Y_1)\}. \end{aligned} \quad (4.16)$$

The “ $o_p(1)$ ” and “ $O_p(1)$ ” terms in (4.16) are of that order uniformly in functions $\bar{\alpha}$ of \mathcal{X} that satisfy $|\bar{\alpha}| \leq C$.

By following the argument leading to (4.1), with only minor modification, we may deduce that under condition (2.1), $\alpha \log H(x) = x + o(x)$ as $x \rightarrow \infty$. Therefore,

$$\log Y_i = \log H(S_i) = \alpha^{-1} S_i + o_p(S_i) = \alpha^{-1} i \{1 + o_p(1)\}.$$

Since in addition, by (2.1), $J(x) = o(\log x)$, then in view of (4.16), and writing $\Delta = \Delta(\bar{\alpha}) = NA/\alpha = N\{(\bar{\alpha}/\alpha) - 1\}$, we have

$$s_j(\bar{\alpha}) = \{1 + o_p(1)\} \sum_{i=1}^{N-1} (M_i/\mu_i) \exp[-(i/N)\Delta\{1 + o_p(1)\}](\alpha^{-1}i)^j,$$

where the “ $o_p(1)$ ” terms are of that order uniformly in $1 \leq i \leq N - 1$ and functions $\bar{\alpha}$ satisfying $|\bar{\alpha}| \leq C$. More simply,

$$\sum_{i=2}^N \log Y_i - (N - 1)\alpha^{-1} = \{1 + o_p(1)\}_2^{\frac{1}{2}} \alpha^{-1} N^2.$$

Hence, if we replace α by $\bar{\alpha}$ on the left-hand side of (2.7) then that quantity may be written as $N^2 \alpha^{-1} t(\hat{\alpha})$, where

$$\begin{aligned} t(\bar{\alpha}) &= \{1 + o_p(1)\} \sum_{i=1}^{N-1} (M_i/\mu_i) \exp[-(i/N)\Delta\{1 + o_p(1)\}](i/N) \\ &\quad - \{1 + o_p(1)\}_2^{\frac{1}{2}} \sum_{i=1}^{N-1} (M_i/\mu_i) \exp[-(i/N)\Delta\{1 + o_p(1)\}] \\ &\quad + O_p\{\exp(-A \log Y_1)\}, \end{aligned} \quad (4.17)$$

where again the “ $o_p(1)$ ” terms are of that order uniformly in $1 \leq i \leq N - 1$ and functions $\bar{\alpha}$ satisfying $|\bar{\alpha}| \leq C$.

By inspection of (2.7) and (2.16) it may be proved that if $C > \alpha$ is sufficiently large then with probability tending to 1 there exists at least one solution $\bar{\alpha}$ of (2.7) satisfying $|\bar{\alpha}| \leq C$, and that any sequence of such solutions satisfies $\bar{\alpha} - \alpha = o_p(1)$. We shall show that any sequence $\bar{\alpha} = \bar{\alpha}_N$ of solutions satisfies $\Delta(\bar{\alpha}) = O_p(1)$. For this it suffices to show that for each sequence $\lambda = \lambda(N)$ of positive constants diverging to infinity, and any sequence $\bar{\alpha} = \bar{\alpha}_N$ of solutions of (2.7) that satisfy $\bar{\alpha} - \alpha = o_p(1)$, $P\{|\Delta(\bar{\alpha})| > \lambda\} \rightarrow 0$. Let \mathcal{E}_+ , \mathcal{E}_- denote the events $\{\Delta(\bar{\alpha}) > \lambda\}$, $\{\Delta(\bar{\alpha}) < -\lambda\}$ respectively. We may deduce from (4.17) that on \mathcal{E}_{\pm} ,

$$t(\bar{\alpha}) = \mp \{1 + o_p(1)\}_2^{\frac{1}{2}} \sum_{i=1}^{N-1} (M_i/\mu_i) \exp[-\Delta(i/N)\{1 + o_p(1)\}], \quad (4.18)$$

where the $+$, $-$ signs are to be taken respectively. Unless it is also true that $P(\mathcal{E}_{\pm}) \rightarrow 0$, (4.18) contradicts the assumption that $\bar{\alpha}$ solves $t(\bar{\alpha}) = 0$.

Acknowledgment

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada.

References

- [1] Ballerini, R. and Resnick, S., "Records from improving populations," *J. Appl. Probab.* 22, 487–502, (1985).
- [2] Ballerini, R. and Resnick, S., "Records in the presence of a linear trend," *Adv. Appl. Probab.* 19, 801–828, (1987).
- [3] Berred, M., "On record values and the exponent of a distribution with regularly varying upper tail," *J. Appl. Probab.* 29, 575–586, (1992).
- [4] Becker, R.A., Chambers, J.M., and Wilks, A.R., *The New S Language*, Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.
- [5] Bingham, N.H., Goldie, C., and Teugels, J., *Regular Variation. Encyclopedia of Mathematics and its Applications*, 27, Cambridge University Press, Cambridge, 1987.
- [6] Chatterjee, S. and Chatterjee, S., "New lamps for old: an exploratory analysis of running times in Olympic games," *Appl. Statist.* 31, 14–22, (1982).
- [7] Csörgő, S., Deheuvels, P., and Mason, D., "Kernel estimates of tail index of a distribution," *Ann. Statist.* 13, 1050–1077, (1985).
- [8] De Haan, L. and Verkade, E., "On extreme value theory in the presence of a trend," *J. Appl. Probab.* 24, 62–76, (1987).
- [9] Hall, P., "On some simple estimates of an exponent of regular variation," *J. Roy. Statist. Soc. Ser. B* 44, 37–42, (1982).
- [10] Hill, B.M., "A simple general approach to inference about the tail of a distribution," *Ann. Statist.* 3, 1163–1174, (1975).
- [11] Morton, R.H., "Letter to the Editor," *Appl. Statist.* 33, 317–318, (1983).
- [12] Pickands, J. III, "Statistical inference using extreme order statistics," *Ann. Statist.* 3, 119–131, (1975).
- [13] Resnick, S.I., "Limit laws for record values," *Stoch. Proc. Appl.* 1, 67–82, (1973a).
- [14] Resnick, S.I., "Record values and maxima," *Ann. Probab.* 1, 650–662, (1973b).
- [15] Resnick, S.I., "Extremal processes and record value times," *J. Appl. Probab.* 10, 864–868, (1973c).
- [16] Resnick, S.I., "Weak convergence to extremal processes," *Ann. Probab.* 3, 951–960, (1975).
- [17] Resnick, S.I., *Extreme Values, Regular Variation, and Point Processes*, Springer, New York, 1987.
- [18] Shorrock, R.W., "A limit theorem for inter-record times," *J. Appl. Probab.* 9, 219–223, (1972).
- [19] Shorrock, R.W., "Record values and inter-record times," *J. Appl. Probab.* 10, 543–545, (1973).
- [20] Shorrock, R.W., "On discrete time extremal processes," *Adv. Appl. Probab.* 6, 580–592, (1974).
- [21] Shorrock, R.W., "Extremal processes and random measures," *J. Appl. Probab.* 12, 316–323, (1975).
- [22] Smith, R.L., "Maximum likelihood estimation in a class of nonregular cases," *Biometrika* 72, 67–92, (1985).
- [23] Smith, R.L. and Miller, J.E., "A non-Gaussian state space model and application to the prediction of records," *J. Roy. Statist. Soc. Ser. B* 48, 79–88, (1986).
- [24] Smith, R.L., "Forecasting records by maximum likelihood," *J. Amer. Statist. Assoc.* 83, 331–338, (1988).
- [25] Tryfos, P. and Blackmore, R., "Forecasting records," *J. Amer. Statist. Assoc.* 80, 46–50, (1985).
- [26] Weiss, L. and Wolfowitz, J., "Maximum probability estimators," *Ann. Inst. Statist. Math.* 19, 193–206, (1967).

- [27] Weiss, L. and Wolfowitz, J., "Maximum likelihood estimation of a translation parameter of a truncated distribution," *Ann. Statist.* 1, 944–947, (1973).
- [28] Weiss, L. and Wolfowitz, J., *Maximum Probability Estimators and Related Topics*, Lecture Notes in Math. 424, Springer, Berlin, 1974.