ELSEVIER

Statistics & Probability Letters 43 (1999) 57–64

STATISTICS & PROBABILITY LETTERS

On the impossibility of estimating densities in the extreme tail

Jan Beirlant[a], Luc Devroye[b,*,1]

[a]Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium
[b]School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6

Received March 1998; received in revised form August 1998

Abstract

We give a short proof of the following result. Let $X_1,\dots,X_n$ be independent and identically distributed observations drawn from a density f on the real line. Let f_n be any estimate of the density g_n of $\max(X_1,\dots,X_n)$. We show that there exists a unimodal infinitely many times differentiable density f such that

$$\inf_n \boldsymbol{E}\{\int |f_n(x)\text{-}g_n(x)|\,\mathrm{d}x\} \geqslant \tfrac{1}{49}.$$

Thus, in the total variation sense, universally consistent density estimates do not exist. A similar result is derived concerning the supremum norm.

MSC: primary: 62G07; secondary: 62G05; 62F12; 60F25

Keywords: Density estimation; Nonparametric estimation; Lower bounds; Rate of convergence; Maxima; Extreme values

1. Introduction and main result

We are given data $X_1,\dots,X_n$, an independent and identically distributed (i.i.d.) sample drawn from an unknown density f on the real line. The maximum $Y_n=\max(X_1,\dots,X_n)$ has density $g_n=nfF^{n-1}$, where F is the cumulative distribution function for f. There have been many attempts in the literature at estimating quantities related to the tail of f that make sense in the context of the study of Y_n. Well-known references in this respect are e.g. Hill (1975), Pickands III (1975), Hall (1982), Csörgö et al. (1985), Dekkers and de Haan (1989) among many others. In this last reference a practical example of the estimation of small return period of a yearly maximum is given. Recently, Hall and Weissman (1997) have proposed a bootstrap procedure to estimate extreme tail probabilities. All of these references do assume that the distribution of Y_n is in the domain of attraction of an extreme value distribution. Also, de Haan and Resnick (1996) have studied rates of convergence of the distribution of Y_n to its limit distribution in the uniform metric and the total variation distance.

* Corresponding author. E-mail: luc@cs.mcgill.ca.

1 Work supported by NSERC Grant A3456 and FCAR Grant 90-ER-0291.

0167-7152/99/$ – see front matter
PII: S0167-7152(98)00246-6

In this paper we show that without such a domain of attraction condition on the underlying distribution f, the general problem of designing an estimate f_n of g_n, respectively F_n of F^n, that is consistent in total variation is unsolvable. This program will be carried out considering the total variation distance

$$\int |f_n(x) - g_n(x)| \, \mathrm{d}x,$$

respectively the uniform metric

$$\sup_x |F_n(x) - F^n(x)|.$$

Concerning the estimation of a small survival probability $1 - F(x)$ itself using a nonparametric density estimator $\hat{f}_n$ of f itself, the method of proof used in deriving these results also leads to a similar result for the distance

$$\int \left| \frac{\hat{f}_n}{f} - 1 \right| \mathrm{d}F^n(x).$$

This distance measure discusses the relative error of density estimators in the extreme tail area.

In extreme value theory, few results are available on rates of convergence of estimation procedures. In Hall and Welsh (1984) and Drees (1995) the rates of convergence for estimates of the extreme value index are discussed.

When related to the general literature on density estimation, we recall the following. An estimate $\hat{f}_n$ of f is a mapping from $\mathbb{R}^{n+1}$ to $\mathbb{R}$. Given the data, $f(x)$ is estimated by $\hat{f}_n(x; X_1, \ldots, X_n)$. The following "slow rate of convergence" result was shown in Devroye (1983).

Theorem 1. *Let $\{\hat{f}_n\}$ be a given sequence of estimates and let $a_n \downarrow 0$ be a sequence of real numbers. Then there exists a density f such that*

$$\boldsymbol{E} \int |\hat{f}_n - f| \, \mathrm{d}x > a_n$$

infinitely often. The density f may also be taken from the class of unimodal densities with infinitely many times continuous derivatives.

The theorem states that to study rates of convergence in density estimation, we need at least some combination of a tail condition and a smoothness condition. Nevertheless, the fact that the result referred to some unknown subsequence prompted Birgé (1986) to improve the above theorem as follows.

Theorem 2. *Let $a_n \to 0$ such that $\sup a_n \in (2/39, 2/13)$. For any sequence $\{\hat{f}_n\}$, there exists a density f on $[0,1]$ bounded by two such that*

$$\boldsymbol{E} \int |f_n - f| \, \mathrm{d}x \geqslant a_n$$

for all n.

With a_n monotonically decreasing and $a_1 \leqslant 1/32$, a short proof of the above theorem may be found in Devroye (1995). It is the method of proof of that paper that is used here again.

Main Theorem. *Let $X_1, \ldots, X_n$ be i.i.d. observations drawn from a density f on the real line. Let f_n be any estimate of the density g_n of $\max(X_1, \ldots, X_n)$. Then there exists a density f such that*

$$\inf_n \boldsymbol{E} \left\{ \int |f_n(x) - g_n(x)| \, \mathrm{d}x \right\} \geqslant \frac{1}{49}.$$

It should be noted that by transformations of the axis, we can pick our density f from all infinitely many continuously differentiable densities. Also, as should be obvious from the proof, no attempt was made to optimize the constant in the lower bound. Finally, at the expense of a smaller constant, we may pick f from the class of bounded unimodal densities. It is only by jointly placing monotonicity or regular variation conditions on the tail together with rates of decrease that one can obtain consistent density estimates.

A similar results yields an interpretation concerning the estimation of a density f in the extreme upper tail.

Second Theorem. *Let $X_1,\ldots,X_n$ be i.i.d. observations drawn from a density f on the real line. Let $\hat{f}_n$ be any estimate of the density f. Then there exists a density f such that*

$$\inf_n \boldsymbol{E}\left\{\int\left|\frac{\hat{f}_n(x)}{f(x)}-1\right|\,\mathrm{d}F^n(x)\right\}\geqslant\frac{1}{49}.$$

In the supremum norm, $\sup_x|F_n(x)-F^n(x)|$, where F_n is any estimate of the distribution function F^n of Y_n, and F is the distribution function of X_1, we also have poor performance for any estimate. As $\sup_x|F_n(x)-F^n(x)|\leqslant(1/2)\int|f_n-nfF^{n-1}|$, where f_n is the density estimate corresponding to F_n, the Main theorem is of little help. Also, F_n need not necessarily have a density. We only give a crude version of the lower bound (in the limit supremum sense) in this case. Thus, even for this weak norm, one cannot hope to ever universally estimate tail probabilities for maxima, regardless of sample size, without making assumptions on the tail behavior of F.

Third Theorem. *Let $X_1,\ldots,X_n$ be i.i.d. observations drawn from a distribution with distribution function F on the real line. Let $F_n=F_n(x,\mathscr{X}_n)$ be any distribution function estimate of F^n, the distribution function of $Y_n=\max(X_1,\ldots,X_n)$. Then there exists a distribution function F such that*

$$\limsup_{n\to\infty}\boldsymbol{E}\left\{\sup_x|F_n(x)-F^n(x)|\right\}\geqslant\frac{1}{2e^3}.$$

2. Proofs

Proof of the Main Theorem. First we construct a family of densities f. Let $b=0.b_1b_2b_3\ldots$ be a real number on $[0,1]$ with the shown expansion, where each b_i takes values in $\{0,\ldots,2^{l_i}-1\}$ and may be represented by l_i bits $(b_{i0},\ldots,b_{i,l_i-1})$. The choice of the positive integers l_i will be left open for now. Thus,

$$b=\sum_{i=1}^{\infty}\frac{b_i}{2^{l_1+\cdots+l_i}}.$$

Let B be a random variable uniformly distributed on $[0,1]$ with expansion $B=0.B_1B_2B_3\ldots$. It should be noted that this corresponds to taking all B_{ij}'s independent Bernoulli $(1/2)$ random variables. And each B_i is uniformly distributed on $\{0,\ldots,2^{l_i}-1\}$ and independent of B_j, $j\neq i$.

Let us define a random variable W with

$$\boldsymbol{P}\{W=j\}=p_j\ (j\geqslant 1),$$

where $p_j=2^{-j}$.

Define an i.i.d. sequence of uniform $[0,1]$ random variables $U_1,U_2,\ldots$. Define another i.i.d. sequence $W_1,W_2,\ldots$ drawn from the distribution of W. Define a third sequence of i.i.d. uniform random variables $Z_1,Z_2,\ldots$ on $\{0,\ldots,l_i-1\}$. These sequences are used to construct coupled data sequences. Each $b\in[0,1)$ describes a different distribution. With b replaced by B we have a random distribution. The real line is partitioned into blocks of length $2l_1$, $2l_2,\ldots$ respectively, where the i-th block is $A_i=[2\sum_{j<i}l_j,2\sum_{j\leqslant i}l_j)\stackrel{\text{def}}{=}[L_i,R_i)$.

The i-th block is picked with probability p_i. Within a block, Z_i picks the location, U_i moves X_i over within the given location, and $b_{W_iZ_i}$ is a polarity bit. We thus define

$$X_i = L_{W_i} + Z_i + U_i + b_{W_iZ_i}.$$

It is not difficult to see that if $W_i=j$, $Z_i=k$, and $b_{jk}=0$, then X_i is uniformly distributed on $[L_j+k, L_j+k+1)$. If $b_{jk}=1$, then it is uniformly distributed on $[L_j+k+1, L_j+k+2)$.

We write f_b to denote the density of X_1 for a given parameter value b. Let $g_{n,b}$ denote the density of Y_n for fixed b. Introduce the shorthand notation $\mathscr{U}_n=(U_1,U_2,\ldots,U_n)$, $\mathscr{W}_n=(W_1,W_2,\ldots,W_n)$, $\mathscr{Z}_n=(Z_1,Z_2,\ldots,Z_n)$, and $\mathscr{X}_n=(X_1,X_2,\ldots,X_n)$. We write $\mathscr{U}_\infty$ for the infinite sample. Observe that in this manner we have defined an infinite number of samples, one for each value of b. Define $k_n = 2 + \lceil \log_2 n \rceil$, and denote the L_1 error by

$$\begin{aligned} J_n(b) &= \int |f_n(x,\mathscr{X}_n) - g_{n,b}(x)|\,\mathrm{d}x \\ &= \sum_{i=1}^{\infty} \int_{A_i} |f_n(x,\mathscr{X}_n) - g_{n,b}(x)|\,\mathrm{d}x \\ &\geqslant \int_{A_{k_n}} |f_n(x,\mathscr{X}_n) - g_{n,b}(x)|\,\mathrm{d}x \\ &\stackrel{\text{def}}{=} K_n(b). \end{aligned}$$

Thus,

$$\sup_b \inf_n \boldsymbol{E} J_n(b) \geqslant \sup_b \boldsymbol{E}\left\{\inf_n J_n(b)\right\} \geqslant \boldsymbol{E}\left\{\inf_n K_n(B)\right\}.$$

Consider now the conditional expectation $\boldsymbol{E}\{\inf_n K_n(B)|\mathscr{U}_\infty,\mathscr{W}_\infty,\mathscr{Z}_\infty\}$. Then for any $c>0$,

$$\begin{aligned} \boldsymbol{E}\left\{\inf_n K_n(B)\middle| \mathscr{U}_\infty,\mathscr{W}_\infty,\mathscr{Z}_\infty\right\} &\geqslant c\boldsymbol{P}\left\{\bigcap_{n=1}^{\infty}[K_n(B)\geqslant c]\middle| \mathscr{U}_\infty,\mathscr{W}_\infty,\mathscr{Z}_\infty\right\} \\ &\geqslant c\left(1 - \sum_{n=1}^{\infty} \boldsymbol{P}\{K_n(B)<c]|\mathscr{U}_\infty,\mathscr{W}_\infty,\mathscr{Z}_\infty\}\right) \\ &= c\left(1 - \sum_{n=1}^{\infty} \boldsymbol{P}\{K_n(B)<c]|\mathscr{U}_n,\mathscr{W}_n,\mathscr{Z}_n\}\right). \end{aligned}$$

We bound the conditional probabilities inside the sum: let $\mathscr{D}_n$ denote $\mathscr{U}_n,\mathscr{W}_n,\mathscr{Z}_n,B'$, where B' denotes B, B_{k_n} excepted. With all this fixed, let $N=\sum_{i=1}^n I_{W_i=k_n}$. It is important to note that on A_{k_n}, the interval of interest to us, f_n can at most be one of 2^N possible functions, if we consider all possible values for B_{k_n} (note that B_{k_n} is the only random variable left after conditioning). Let $S=\{j\colon j=Z_i, W_i=k_n, i=1,\ldots,n\}$ be the index set of the collection of subintervals of A_{k_n} occupied by points X_i. Clearly, $|S|\leqslant N\leqslant n$. If we change b_{k_nj} for $j\notin S$, then f_n remains unchanged, yet $g_{n,b}$ changes a lot. For such j, denoting the two possible vectors b_{k_n} by b_+ and b_- respectively, (and keeping all other $l_{k_n}-1$ bits the same), we have with $A_{k_n}=[L_{k_n}+j, L_{k_n}+j+2)$

$$\begin{aligned} \max\left(\int_{A_{k_n,j}} |f_n - g_{n,b_+}|, \int_{A_{k_n,j}} |f_n - g_{n,b_-}|\right) &\geqslant \frac{1}{2}\left(\int_{A_{k_n,j}} |f_n - g_{n,b_+}| + \int_{A_{k_n,j}} |f_n - g_{n,b_-}|\right) \\ &\geqslant \frac{1}{2}\int_{A_{k_n,j}} |g_{n,b_-} - g_{n,b_+}| \end{aligned}$$

$$
\begin{aligned}
&= \int_{A_{k_n,j}} g_{n,b_-} \\
&\geqslant n \frac{p_{k_n}}{l_{k_n}} \left(p_1 + \cdots + p_{k_n-1}\right)^{n-1} \\
&\stackrel{\text{def}}{=} q_n.
\end{aligned}
$$

If $j \in S$, we can bound the same maximum from below by zero. We first verify the value of q_n. By our choice of p_i and k_n,

$$
\begin{aligned}
q_n &= \frac{n}{2^{k_n} l_{k_n}} \left(1 - 2/2^{k_n}\right)^{n-1} \\
&\geqslant \frac{n}{2^{k_n} l_{k_n}} \left(1 - 2(n-1)/2^{k_n}\right) \\
&\geqslant \frac{n}{8n l_{k_n}} \left(1 - \frac{2(n-1)}{4n}\right) \\
&\geqslant \frac{1}{16 l_{k_n}}.
\end{aligned}
$$

Consider thus $T \stackrel{\text{def}}{=} q_n \sum_{i=1}^{l_{k_n}-|S|} V_i$, where the V_i's are independent Bernoulli (1/2) random variables. Then given $\mathscr{D}_n$, $K_n(B)$ is stochastically greater than T. Thus, by Hoeffding's inequality (Hoeffding, 1963),

$$
\begin{aligned}
\boldsymbol{P}\{K_n(B) < c | \mathscr{D}_n\} &\leqslant \boldsymbol{P}\left\{ q_n \sum_{i=1}^{l_{k_n}-|S|} V_i < c \right\} \\
&\leqslant \boldsymbol{P}\left\{ \sum_{i=1}^{l_{k_n}-n} V_i < c/q \right\} \\
&\leqslant \boldsymbol{P}\left\{ \sum_{i=1}^{l_{k_n}-n} (V_i - \boldsymbol{E}V_i) < \frac{c}{q} - \frac{l_{k_n}-n}{2} \right\} \\
&\leqslant \boldsymbol{P}\left\{ \sum_{i=1}^{l_{k_n}-n} (V_i - \boldsymbol{E}V_i) < 16 c l_{k_n} - \frac{l_{k_n}-n}{2} \right\} \\
&\leqslant \boldsymbol{P}\left\{ \sum_{i=1}^{l_{k_n}-n} (V_i - \boldsymbol{E}V_i) < \frac{n}{2} - \frac{l_{k_n}}{6} \right\} \\
&\quad \text{(when } c = 1/48\text{)} \\
&\leqslant \boldsymbol{P}\left\{ \sum_{i=1}^{l_{k_n}-n} (V_i - \boldsymbol{E}V_i) < -\frac{l_{k_n}}{12} \right\} \\
&\quad \text{(when } l_{k_n} \geqslant 6n\text{)} \\
&\leqslant 2\mathrm{e}^{-2(l_{k_n}-n)^{-1}(l_{k_n}/12)^2} \\
&\quad \text{(by Hoeffding's inequality)} \\
&\leqslant 2\mathrm{e}^{-l_{k_n}/72}.
\end{aligned}
$$

As this bound does not depend upon the variables on which we conditioned, we obtain, by concatenation of bounds,

$$\begin{aligned}\sup_b \inf_n \mathbf{E} J_n(b) &\geqslant \mathbf{E}\left\{\inf_n K_n(B)\right\}\\ &\geqslant (1/48)\left(1-\sum_{n=1}^{\infty}\mathbf{P}\{K_n(B)<(1/48)|\mathscr{U}_n,\mathscr{W}_n,\mathscr{X}_n\}\right)\\ &\geqslant (1/48)\left(1-2\sum_{n=1}^{\infty}\mathrm{e}^{-l_{k_n}/72}\right)\\ &\geqslant (1/49)\end{aligned}$$

if $l_{k_n}\geqslant 72n\log 50+72\log 2$. So, it suffices to pick l_k in such a way that for all n

$$l_{2+\lceil \log_2 n\rceil}\geqslant 72n\log 50+72\log 2.$$

Clearly, this is possible by making l_k increase exponentially quickly with k. This concludes the proof of the main theorem. □

Proof of the Third Theorem. First we construct a family of densities F, following the lead of the proof of the Main Theorem. Let $b=0.b_1b_2b_3\ldots$ be a real number on $[0,1]$ with the shown expansion, where each b_i takes values in $\{0,1\}$. Thus,

$$b=\sum_{i=1}^{\infty}\frac{b_i}{2^i}$$

represents a real number on $[0,1]$. Let B be a random variable uniformly distributed on $[0,1]$ with expansion $B=0.B_1B_2B_3\ldots$. It should be noted that this corresponds to taking all B_i's independent Bernoulli $(1/2)$. Let us define a random variable W with

$$\mathbf{P}\{W=j\}=2^{-j}\ (j\geqslant 1).$$

Define an i.i.d. sequence $W_1,W_2,\ldots$ drawn from the distribution of W. These sequences are used to construct coupled data sequences. Each $b\in[0,1)$ describes a different distribution. With b replaced by B we have a random distribution. We define

$$X_i=2W_i+B_{W_i}.$$

Thus, X_i is either $2W_i$ or $2W_i+1$, depending upon the value of B_{W_i}. We write F_b to denote the distribution function of X_1 for parameter b. Let $F_{n,b}$ denote the distribution function of Y_n for fixed b. Introduce the shorthand notation $\mathscr{W}_n=(W_1,W_2,\ldots,W_n)$, and $\mathscr{X}_n=(X_1,X_2,\ldots,X_n)$. We write $\mathscr{W}_\infty$ for the infinite sample. Observe that in this manner we have defined an infinite number of samples, one for each value of b. Define $k_n=\lceil \log_2 n\rceil$, and denote the sup-norm error by

$$\begin{aligned}J_n(b)&=\sup_x |F_n(x,\mathscr{X}_n)-F_{n,b}(x)|\\ &\geqslant \max_{x\in\{2k_n+1/2,\,2k_n+3/2\}}|F_n(x,\mathscr{X}_n)-F_{n,b}(x)|\\ &\stackrel{\text{def}}{=} K_n(b).\end{aligned}$$

Thus,

$$\begin{aligned}
\sup_b \limsup_{n\to\infty} \boldsymbol{E} J_n(b) &\geqslant \boldsymbol{E}\left\{\limsup_{n\to\infty} \boldsymbol{E}\{J_n(B)|B\}\right\}\\
&\geqslant \limsup_{n\to\infty} \boldsymbol{E}\left\{\boldsymbol{E}\{J_n(B)|B\}\right\}\\
&\quad \text{(by Fatou's lemma and the boundedness of } J_n)\\
&= \limsup_{n\to\infty} \boldsymbol{E}\left\{J_n(B)\right\}\\
&\geqslant \limsup_{n\to\infty} \boldsymbol{E}\left\{K_n(B)\right\}.
\end{aligned}$$

Consider now the following conditional expectation:

$$\begin{aligned}
\boldsymbol{E}\left\{K_n(B)|\mathscr{W}_\infty\right\} &\geqslant \inf_b |F_{n,b}(2k_n+3/2) - F_{n,b}(2k_n-1/2)| I_{D_n}\\
&\quad \text{(where } D_n \text{ is the event that no } X_i \text{ is } 2k_n \text{ or } 2k_n+1)\\
&\geqslant \left[\left(1-\frac{1}{2^{k_n}}\right)^n - \left(1-\frac{2}{2^{k_n}}\right)^n\right] I_{D_n}\\
&\geqslant \frac{n+1}{2^{k_n}}\left(1-\frac{2}{2^{k_n}}\right)^{n-1} I_{D_n}\\
&\geqslant \frac{n+1}{2n}\left(1-\frac{2}{n}\right)^{n-1} I_D\\
&\sim \frac{1}{2e^2} I_{D_n}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\liminf_{n\to\infty} \boldsymbol{E} K_n(B) &\geqslant (1/(2e^2)) \liminf_{n\to\infty} \boldsymbol{P}\{D_n\}\\
&= (2e^2)^{-1} \liminf_{n\to\infty} (1-1/2^{k_n})^n\\
&\geqslant (2e^2)^{-1} \liminf_{n\to\infty} (1-1/n)^n\\
&\sim \frac{1}{2e^3}.
\end{aligned}$$

This concludes the proof of the Second Theorem. □

References

Birgé, L., 1986. On estimating a density using Hellinger distance and some other strange facts. Probab. Theory Related Fields 71, 271–291.

Csörgö, M., Deheuvels, P., Mason, D., 1985. Kernel estimates of the tail index of a distribution. Ann. Statist. 13, 1050–1077.

Dekkers, A., de Haan, L., 1989. On the estimation of the extreme-value index and large quantile estimation. Ann. Statist. 17, 1795–11832.

de Haan, L., Resnick, S., 1996. Second-order regular variation and rates of convergence in extreme-value theory. Ann. Probab. 24, 97–124.

Devroye, L., 1983. On arbitrarily slow rates of global convergence in density estimation. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 62, 475–483.

Devroye, L., 1995. Another proof of a slow convergence result of Birgé. Statist. Probab. Lett. 23, 63–67.

Drees, H., 1995. Optimal rates of convergence for estimates of the extreme value index. Technical Report, University of Cologne, November.

Hall, P., 1982. On simple estimates of an exponent of regular variation. J. Roy. Statist. Soc. B 42, 37–42.

Hall, P., Weissman, I., 1997. On the estimation of extreme tail probabilities. Ann. Statist. 25, 1311–1326.

Hall, P., Welsh, A., 1984. Best attainable rates of convergence for estimates of parameters of regular variation. Ann. Statist. 12, 1079–1084.

Hill, B.M., 1975. A simple approach to inference about the tail of a distribution. Ann. Statist. 3, 1163–1174.

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13–30.

Pickands III, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3, 119–131.