Statistica Sinica 3(1993), 501-515

A PENALIZED MAXIMUM LIKELIHOOD ESTIMATE OF $f(0+)$ WHEN f IS NON-INCREASING

Michael Woodroofe and Jiayang Sun

The University of Michigan

Abstract: The problem of estimating the value at 0+ of a non-increasing density f (on $(0, \infty)$) is considered. It is shown, by example, that the problem is interesting, and it is noted that the nonparametric maximum likelihood estimator is inconsistent. A penalized maximum likelihood estimator is derived as an alternative, and its properties studied through simulations and asymptotic analysis. In particular, the penalized maximum likelihood estimator is shown to be consistent.

Key words and phrases: Nonparametric maximum likelihood estimation, consistency, asymptotic distributions, simulation.

1. Introduction

Let f denote a left continuous density for which $f(x) = 0$ for all $-\infty < x \leq 0$ and is non-increasing in $0 < x < \infty$; and consider the problem of estimating f from a random sample, say $X_1, \ldots, X_n$. For this problem, the nonparametric maximum likelihood estimator, $\tilde{f}_n$ say, of f is well known and may be described as follows: letting $0 = x_0 < x_1 < \cdots < x_n < \infty$ denote the ordered values of 0 and $X_1, \ldots, X_n$, $\tilde{f}_n$ is a step function for which $\tilde{f}_n(x) = f_n(x_k)$ for all $x_{k-1} < x \leq x_k$, where

$$\tilde{f}_n(x_k) = \min_{0 \leq r \leq k-1} \max_{k \leq s \leq n} \frac{s - r}{n(x_s - x_r)} \tag{1}$$

for all $k = 1, \ldots, n$, and $\tilde{f}_n(x) = 0$ for other values of x. See, for example, Prakasa Rao (1983, p.354) and/or Robertson, Wright and Dykstra (1988, pp.326–328), hereafter RWD.

It is known that $\tilde{f}_n$ is a consistent estimator of f in that $\tilde{f}_n(x) \to f(x)$ w.p.1 for all $0 < x < \infty$ at which f is continuous. See, for example, Prakasa Rao (1983, p.352) or RWD, p.330. It does not follow, however, that $\tilde{f}_n(0+) = \tilde{f}_n(x_1)$ is a consistent estimator of $f(0+) = \lim_{x \searrow 0} f(x)$. In fact, it is easily seen that if $0 < f(0+) < \infty$, then

$$\frac{\tilde{f}_n(0+)}{f(0+)} \Rightarrow \sup_{1 \leq k < \infty} \frac{k}{\Gamma_k}, \tag{2}$$

where $\Gamma_1, \Gamma_2, \ldots$ are partial sums of i.i.d. standard exponential random variables and $\Rightarrow$ denotes convergence in distribution. See Remark 3 at the end of Section 5 for an outline of the derivation. Observe that $\tilde{f}_n(0+)$ is simply too big, since $P\{\sup_{1\le k<\infty} k/\Gamma_k > 1\} = 1$, by the Strong Law of Large Numbers. For example, the probability that $\tilde{f}_n(0+) > 2f(0+)$ approaches $P\{\sup_{1\le k<\infty} k/\Gamma_k > 2\} \ge P\{1/\Gamma_1 > 2\} = 1 - e^{-\frac{1}{2}} \cong .393$. The simulations presented in Section 4 provide more detail on the inconsistency.

One way to decrease the size of $\tilde{f}_n(0+)$ is to penalize the nonparametric likelihood function for large values of $f(0+)$, following the general approach of Good and Gaskins (1971). This approach is developed here. It is shown that appropriate penalized nonparametric maximum likelihood estimators of f do lead to consistent estimators of $f(0+)$ and are not much more difficult to compute than $\tilde{f}_n$.

The paper proceeds as follows. In Section 2, reasons are offerred to convince the reader that estimating $f(0+)$ is an interesting problem. The penalized maximum likelihood estimator is presented in Section 3, and studied through simulations in Section 4. Consistency is established in Sections 5 and 6, and an asymptotic distribution is derived in Section 7.

2. Why the Problem is Interesting

In the present context, $f(0+)$ is the (right hand) value of the density at the mode, assuming that f is non-increasing. Wegman (1970) considered the problem of estimating a mode, assuming only that f is unimodal, by fitting a non-decreasing density to the left and a non-increasing density to the right of x_k for each k, and then maximizing the likelihood over k. In the process, he discovered a spiking problem. The estimated value of the density at the mode was simply too big. It is expected that penalized maximum likelihood estimators, such as those considered here, may ameliorate this spiking problem.

Additional reasons for wanting to estimate $f(0+)$ are included in the following examples. For the first example, recall that a distribution function G is said to be arithmetic if it is supported by some positive multiple of the integers.

Example 1. Suppose that the times between breakdowns of a system are i.i.d. positive, random variables $Y_1, Y_2, \ldots$ having a common non arithmetic distribution function G with a finite positive mean $0 < \nu < \infty$. Suppose further that the system is inspected at time $t > 0$ and the time since the last breakdown (prior to t) is available; that is, suppose that $X = t - (Y_1 + \cdots + Y_N)$ is observed, where $N = N_t$ is the largest n for which $Y_1 + \cdots + Y_n \le t$. If t is large, then the distribution of X may be approximated by the distribution with density

$$f(x) = \frac{1}{\nu}[1 - G(x-)], \quad \forall \ 0 < x < \infty,$$

since the distribution of X converges to the latter distribution as $t \to \infty$. See, for example, Feller (1971, pp.355-356). Clearly, f is non-increasing. In this example, there is natural interest in the mean time ν between breakdowns; and ν is related to f by $\nu = 1/f(0+)$.

This example is adapted from Vardi (1989), who considers a more general model. It is reconsidered in Section 3.

Example 2. Lynden-Bell (1991) has described a model in which the probability with which a galaxy is observed depends on its observable angular diameter D in an unknown way. Suppose that consideration is restricted to galaxies for which $D \geq \underline{D} > 0$; and let $Y = D/\underline{D}$ denote the normalized angular diameter. If a uniform distribution of galaxies is assumed and the galaxy's distance R from earth is assumed to be independent of its true angular diameter RD, then Y has density $g(y) = 3/y^4$, $\forall\ 1 \leq y < \infty$, as in Woodroofe's (1991) discussion of Lynden-Bell's paper. Suppose now that a galaxy is observed with probability $w(y)$, where w is a non-decreasing function for which $\lim_{y\to\infty} w(y) = 1$. Then the conditional density of Y given that it is observed is

$$g^*(y) = \frac{3w(y)}{c(w)y^4}, \quad \forall \ 1 \leq y < \infty,$$

where

$$c(w) = \int_1^\infty \frac{3w(y)}{y^4} dy,$$

the unconditional probability that a galaxy for which $D \geq \underline{D}$ is included in the sample. There is natural interest in $c(w)$, since $1 - c(w)$ is the proportion of galaxies which were not observed.

To see how to estimate $c(w)$, let Y denote a random variable with density g^* and let $X = 1/Y^3$. Then X has a non-increasing density $f(x) = w(x^{-1/3})/c(w)$ for $0 < x \leq 1$; and $c(w) = 1/f(0+)$.

3. Penalized Maximum Likelihood Estimators

Derivation. Since n and $0 = x_0 < x_1 < \cdots < x_n$ are fixed throughout this section and the next, the subscript "n" is omitted from the notation. (It will reappear in Section 5.)

Let G denote the collection of all left continuous densities g for which $g(x) = 0$ for $-\infty < x \leq 0$ and $g(x)$ is non-increasing in $0 < x < \infty$. Then the penalized

nonparametric log likelihood functions considered here are of the form

$$\ell_\alpha(g) = \sum_{i=1}^{n} \log g(x_i) - n\alpha g(0+), \quad \forall\, g \in G, \tag{3}$$

where $\alpha > 0$ is a smoothing parameter. It is easily seen that the maximum occurs when g is a step function for which $g(x) = g_k$, say, for all $x_{k-1} < x \le x_k$ and all $k = 1, \ldots, n$ and $g(x) = 0$ for other values of x. So, only such functions need be considered. The condition that $\int_0^\infty g dx = 1$ may be written

$$\sum_{i=1}^{n} (x_i - x_{i-1}) g_i = 1. \tag{4}$$

So, the problem is to maximize $\ell_\alpha(g)$ with respect to $\infty > g_1 \ge \cdots \ge g_n > 0$, subject to the constraint (4).

For a fixed $0 < \alpha < \infty$ and all $0 \le \gamma \le 1$, let

$$g_k(\gamma) = \min_{1 \le r \le k} \max_{k \le s \le n} \frac{(s - r + 1)/n}{w_r + \cdots + w_s} \tag{5}$$

where $w_0 = 0$, $w_1 = w_1(\gamma) = \alpha + \gamma x_1$, and $w_k = w_k(\gamma) = \gamma(x_k - x_{k-1})$ for all $k = 2, \ldots, n$. Observe that

$$g_1(\gamma) = \max_{1 \le s \le n} \frac{s/n}{\alpha + \gamma x_s}$$

is non-increasing and convex in $0 \le \gamma \le 1$ and that $\gamma g_1(\gamma)$ is non-decreasing in $0 \le \gamma \le 1$.

Lemma 1. *If $x_n > \alpha$, then the equation $\gamma = 1 - \alpha g_1(\gamma)$ has a unique solution $0 < \hat{\gamma} < 1$; and*

$$\hat{\gamma} = \min_{1 \le s \le n} \left\{ \frac{1}{2}\left(1 - \frac{\alpha}{x_s}\right) + \sqrt{\left(\frac{\alpha}{2x_s}\right)^2 + \frac{\alpha}{2x_s}\left(1 - \frac{2s}{n}\right) + \frac{1}{4}} \right\}. \tag{6}$$

Proof. The equation may be written $\gamma = \min_{1 \le s \le n} g^s(\gamma)$, where $g^s(\gamma) = 1 - \alpha s/[n(\alpha + \gamma x_s)]$ for $0 < \gamma < 1$ and $s = 1, \ldots, n$. Clearly, each g^s is increasing and concave. Also $g^s(1) < 1$ for all $s \le n$, $g^s(0) = 1 - s/n > 0$ for $s < n$, and $(g^n)'(0) = x_n/\alpha > 1$. So, for all $s \le n$: the equation $\gamma = g^s(\gamma)$ has a unique solution, say $\gamma_s \in (0, 1)$; γ_s is given by the term in braces in (6); and $\gamma < g^s(\gamma)$ for all $0 < \gamma < \gamma_s$. Define $\hat{\gamma}$ by (6). Then $\hat{\gamma} \le g^s(\hat{\gamma})$ for all $s = 1, \ldots, n$ with equality for some s. That is, $\hat{\gamma} = \min_{1 \le s \le n} g^s(\hat{\gamma})$.

Theorem 1. *Suppose that $x_n > \alpha$ and define $\hat{\gamma}$ by (6). If $\hat{f}$ is a step function for which*

$$\hat{f}(x) = g_k(\hat{\gamma}), \quad \forall\, x_{k-1} < x \le x_k, \quad \forall\, k = 1, \ldots, n, \tag{7}$$

and $\hat{f}(x) = 0$ for other values of x, then $\hat{f} \in G$ and $\hat{f}$ maximizes $\ell_\alpha(g)$ with respect to $g \in G$.

Proof. Letting $\theta_k = \log g_k$ and introducing a Lagrange multiplier, called $n\gamma$, leads to the problem of maximizing

$$\begin{aligned} L_\gamma(\theta) &= \sum_{i=1}^{n} \theta_i - n\alpha e^{\theta_1} - n\gamma \sum_{i=1}^{n} (x_i - x_{i-1}) e^{\theta_i} \\ &= \sum_{i=1}^{n} \left[\theta_i - n w_i(\gamma) e^{\theta_i} \right] \end{aligned}$$

over $\boldsymbol{\theta} \in \Omega = \{\boldsymbol{\omega} \in \mathbf{R}^n : \infty > \omega_1 \ge \cdots \ge \omega_n > -\infty\}$ (with no constraint). Since L_γ is concave and differentiable on all of $\mathbf{R}^n$, a necessary and sufficient condition for L_γ to attain its maximum at a given $\hat{\boldsymbol{\theta}} \in \Omega$, is that $\nabla L_\gamma(\hat{\boldsymbol{\theta}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \le 0$ for all $\boldsymbol{\theta} \in \Omega$, where ∇ denotes gradient. See, for example, Rockafellar (1970, pp.270–271). Letting $\hat{g}_k = \exp(\hat{\theta}_k)$ and observing that $\partial L_\gamma(\boldsymbol{\theta})/\partial \theta_k = 1 - n w_k(\gamma) g_k$ for all $k = 1, \ldots, n$, this condition may be written as

$$\sum_{i=1}^{n} [1 - n w_i(\gamma) \hat{g}_i](\theta_i - \hat{\theta}_i) \le 0, \quad \forall\ \boldsymbol{\theta} \in \Omega,$$

or

$$\langle \boldsymbol{h} - \hat{\boldsymbol{g}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle \le 0, \quad \forall\ \boldsymbol{\theta} \in \Omega, \tag{8}$$

where

$$\boldsymbol{h} = \left[\frac{1}{n w_1(\gamma)}, \ldots, \frac{1}{n w_n(\gamma)} \right]'$$

and

$$\langle \boldsymbol{y}, \boldsymbol{z} \rangle = \sum_{i=1}^{n} w_i(\gamma) y_i z_i, \quad \forall\ \boldsymbol{y}, \boldsymbol{z} \in \mathbf{R}^n.$$

From Theorem 1.4.4 of RWD, it is known that $[g_1(\gamma), \ldots, g_n(\gamma)]'$ is the projection of $\boldsymbol{h}$ on Ω with respect to $\langle \cdot, \cdot \rangle$ and that, moreover, $[g_1(\gamma), \ldots, g_n(\gamma)]'$ satisfies (8). So, $\ell_\alpha(g) = L_\gamma(\theta)$ is maximized by letting $g_k = g_k(\gamma)$ for all $k = 1, \ldots, n$ for each fixed γ.

It remains to find a γ for which $\boldsymbol{g}(\gamma) = [g_1(\gamma), \ldots, g_n(\gamma)]'$ satisfies the constraint (4). Since $\boldsymbol{g}(\gamma)$ is the projection of $\boldsymbol{h}$ on Ω,

$$\alpha g_1(\gamma) + \sum_{i=1}^{n} \gamma (x_i - x_{i-1}) g_i(\gamma) = \sum_{i=1}^{n} w_i(\gamma) g_i(\gamma) = \sum_{i=1}^{n} w_i(\gamma) h_i(\gamma) = 1,$$

or

$$\sum_{i=1}^{n}(x_i - x_{i-1})g_i(\gamma) = \frac{1 - \alpha g_1(\gamma)}{\gamma},$$

using Theorem 1.3.6 of RWD. Thus, the constraint (4) is satisfied iff $\gamma = \hat{\gamma}$ and, the theorem follows easily.

Remark 1. Once $\hat{\gamma}$ is found, computing the penalized maximum likelihood estimator $\hat{f}$ is no more difficult than computing the unpenalized maximum likelihood estimator $\tilde{f}$. In fact, the penalized maximum likelihood estimator is equal to a maximum likelihood estimator with a deformed data set $\alpha + \hat{\gamma}x_k$, $k = 1, \ldots, n$.

Remark 2. Using (7) and Lemma 1 leads to $\hat{f}(0+) = (1 - \hat{\gamma})/\alpha$, where $\hat{\gamma}$ is as in (6).

Example 1. Revisited. The times (in weeks) since the last reboot of eight work stations at the University of Michigan were recorded on Aug. 30, 1991 and are listed in Table 1. These are of the form considered in Example 1. With the notation of that Example and $\alpha = \log n/2n$, the unconstrained and penalized maximum likelihood estimators of f are listed in Table 1. The penalized maximum likelihood estimate of ν is then $\hat{\nu} = 1/\hat{f}(0+) = 5.43$ weeks.

Table 1. Times since last reboot

x	Pen. MLE	MLE
.585	.184	.214
1.262	.184	.185
2.138	.155	.152
3.142	.155	.152
3.979	.155	.152
4.560	.155	.152
7.151	.049	.048
10.261	.041	.040

Notes: time in weeks since last reboot for eight workstations;
$\alpha = \log n/2n$; $\hat{\nu} = 5.43$ weeks.

Of course, the data set is too small for the estimate to be very reliable.

4. Simulations

In Table 2, $\hat{f}$ is compared to $\tilde{f}$ for a samples of size $n = 25, 50, 100$, and 200 from a half standard normal density (the density of the absolute value of a

standard normal random variable) and from a standard exponential density. Ten thousand samples of each size were simulated from both distributions. Average values and standard deviations of the L_1 distances $\int_0^\infty |\tilde{f} - f|dx$ and $\int_0^\infty |\hat{f} - f|dx$ are reported in columns 2–4, and average values and standard deviations of $\tilde{f}(0+)/f(0+)$ and $\hat{f}(0+)/f(0+)$ are reported in columns 5–7, both for selected values of α.

Table 2. Simulations

n	$\int_0^\infty \lvert\tilde{f} - f\rvert dx$	$\int_0^\infty \lvert\hat{f} - f\rvert dx$		$\tilde{f}(0+)/f(0+)$	$\hat{f}(0+)/f(0+)$	
		$\alpha = \frac{\log n}{n}$	n^{-pq}		$\alpha = \frac{\log n}{n}$	n^{-pq}
		The Standard Normal Case ($p = 3$)				
25	.280	.226	.224	9.985	1.066	1.042
	.097	.076	.074	119.3	.223	.209
50	.227	.191	.188	8.310	1.061	1.027
	.068	.057	.055	65.07	.188	.165
100	.182	.159	.156	7.904	1.059	1.014
	.048	.041	.040	54.96	.155	.131
200	.144	.130	.127	7.637	1.060	.988
	.033	.030	.028	55.12	.130	.102
		The Standard Exponential Case ($p = 2$)				
25	.315	.275	.275	8.56	.864	.886
	.088	.069	.070	88.20	.214	.226
50	.258	.231	.231	8.61	.887	.900
	.062	.052	.054	67.60	.184	.190
100	.209	.191	.191	14.26	.913	.912
	.044	.038	.038	352.74	.157	.156
200	.168	.156	.156	12.17	.942	.923
	.031	.028	.028	218.74	.137	.130

Notes: Columns 2, 3, and 4 list Monte Carlo estimates of the mean and standard deviation of the L^1 distances for $\tilde{f}$ and $\hat{f}$ for selected α. Columns 5, 6, and 7 list Monte Carlo estimates of the means and standard deviations of $\tilde{f}(0+)/f(0+)$ and $\hat{f}(0+)/f(0+)$ for selected α. The upper figure is the mean, and the lower is the standard deviation.

The large average and huge standard deviation of $\tilde{f}(0+)/f(0+)$ are to be expected since the limiting distribution (2) has an infinite mean. The corresponding

values for $\hat{f}(0+)/f(0+)$ are encouraging. Differences between averages and standard deviations of the L_1 distances of the estimators are much less dramatic; but $\hat{f}$ is consistently better. At the very least, the simulations suggest that penalizing the likelihood improves the estimation of $f(0+)$ by orders of magnitude, while not worsening global performance. This suggestion is established in Sections 5, 6, and 7 below.

5. Consistency at 0+

In the remainder of the paper, the quantities called $\hat{\gamma}, g_k(\gamma)$, and $\hat{f}$ in (5), (6), and (7) above are denoted by $\hat{\gamma}_n$, $g_{n,k}(\gamma)$, and $\hat{f}_n$. It is assumed throughout that

$$0 < f_0 = f(0+) < \infty, \quad 0 < \alpha = \alpha_n \searrow 0 \text{ and } n\alpha \nearrow \infty \tag{9}$$

as $n \to \infty$. Use is made of the fact that $[F(x_1), \ldots, F(x_n)]$ has the same distribution as $[\Gamma_1, \ldots, \Gamma_n]/\Gamma_{n+1}$ for each n, where F denotes the distribution function of f and $\Gamma_1, \Gamma_2, \ldots$ are partial sums of independent standard exponential random variables. Let

$$\Delta_n = \max_{1 \le k \le n} \frac{k/n}{F(x_k)}.$$

Then $\Delta_n \Rightarrow \Delta = \sup_{k \ge 1} k/\Gamma_k$ as $n \to \infty$. In particular, Δ_n is stochastically bounded.

The consistency of $\hat{f}_n(0+)$ is deduced as a corollary to

Theorem 2. *For any* $0 < \gamma_0 < 1$,

$$p\text{-}\lim_{n \to \infty} \sup_{\gamma_0 \le \gamma \le 1} |\gamma g_{n,1}(\gamma) - f_0| = 0.$$

Proof. Since $\gamma g_{n,1}(\gamma)$ is non-decreasing in $0 < \gamma \le 1$

$$\begin{aligned} & P\left\{ \sup_{\gamma_0 \le \gamma \le 1} |\gamma g_{n,1}(\gamma) - f_0| \ge \varepsilon \right\} \\ \le\ & P\{g_{n,1}(1) - f_0 \ge \varepsilon\} + P\{\gamma_0 g_{n,1}(\gamma_0) - f_0 \le -\varepsilon\} \end{aligned}$$

for all $\varepsilon > 0$. So, it suffices to show that $p\text{-}\lim_{n \to \infty}[g_{n,1}(1) - f_0]^+ = 0$ and $p\text{-}\lim_{n \to \infty}[\gamma g_{n,1}(\gamma) - f_0]^- = 0$ for all fixed $\gamma_0 \le \gamma \le 1$, where $[-]^+$ and $[-]^-$ denote the positive and negative parts of $[-]$.

For the second relation,

$$\gamma g_{n,1}(\gamma) = \max_{1 \le k \le n} \frac{\gamma k/n}{\alpha + \gamma x_k} \ge \frac{\gamma k/n}{\alpha + \gamma x_k}$$

for all $k = 1, \ldots, n$. Let $k = k_n$ be the least integer for which $k/n \geq \sqrt{\alpha}$. Then $F(x_k)/(k/n)$ has the same distribution as $n\Gamma_k/k\Gamma_{n+1}$, which approaches one in probability as $n \to \infty$. So,

$$\frac{f_0 x_k}{k/n} \xrightarrow{\text{P}} 1$$

as $n \to \infty$, since $k \geq n\sqrt{\alpha} \to \infty$ and $k/n \leq \sqrt{\alpha} + 1/n \to 0$ as $n \to \infty$. Next, since $\alpha = o(k/n)$, it follows that

$$p\text{-}\lim_{n\to\infty} \frac{\gamma k/n}{\alpha + \gamma x_k} = f_0$$

and, therefore, that $p\text{-}\lim_{n\to\infty}[\gamma g_{n,1}(\gamma) - f_0]^- = 0$.

For the first relation,

$$g_{n,1}(1) = \max_{1\leq k\leq n} \frac{F(x_k)}{\alpha + x_k} \times \frac{k/n}{F(x_k)} \leq \max_{1\leq k\leq n} \frac{f_0 x_k}{\alpha + x_k} \times \frac{k/n}{F(x_k)}.$$

Now,

$$\max_{1\leq k\leq m} \frac{f_0 x_k}{\alpha + x_k} \times \frac{k/n}{F(x_k)} \leq \Delta_n \times \max_{1\leq k\leq m} \frac{n f_0 x_k}{n\alpha + n x_k} \xrightarrow{\text{P}} 0$$

as $n \to \infty$ for any fixed integer m, independent of n, since $n\alpha \to \infty$ and $\max_{k\leq m} nx_k$ is stochastically bounded in $n = 1, 2, \ldots$; and $g_{n,1}(1) \geq \gamma_0 g_{n,1}(\gamma_0) \geq f_0 + o_p(1)$, since $\gamma g_{n,1}(\gamma)$ is non-decreasing in γ. So, with probability approaching one,

$$g_{n,1}(1) \leq \max_{m\leq k\leq n} \frac{f_0 x_k}{\alpha + x_k} \times \frac{k/n}{F(x_k)} \leq f_0 \max_{m\leq k\leq n} \frac{k/n}{F(x_k)} \Rightarrow f_0 \sup_{k\geq m} \frac{k}{\Gamma_k}$$

as $n \to \infty$ for all $m = 1, 2, \ldots$. That $p\text{-}\lim_{n\to\infty}[g_{n,1}(1) - f_0]^+ = 0$ now follows by letting $n \to \infty$ and then $m \to \infty$.

Corollary 1.

$$p\text{-}\lim_{n\to\infty} \frac{1 - \hat{\gamma}_n}{\alpha} = f_0.$$

Proof. If $x_n > \alpha$ and $\alpha g_{n,1}(\frac{1}{2}) < \frac{1}{2}$, then $1 - \alpha g_{n,1}(\frac{1}{2}) > \frac{1}{2}$ and, therefore, $\hat{\gamma}_n > \frac{1}{2}$. Since $P\{x_n \leq \alpha\} + P\{\alpha g_{n,1}(\frac{1}{2}) \geq \frac{1}{2}\} \to 0$, as $n \to \infty$, $1 - \hat{\gamma}_n = \alpha g_{n,1}(\hat{\gamma}_n) \leq \alpha g_{n,1}(\frac{1}{2}) \xrightarrow{\text{P}} 0$ and, therefore,

$$\frac{1 - \hat{\gamma}_n}{\alpha} = g_{n,1}(\hat{\gamma}_n) \xrightarrow{\text{P}} f_0.$$

Corollary 2.

$$p\text{-}\lim_{n\to\infty} \frac{\hat{f}_n(0+)}{f(0+)} = 1.$$

Proof. This is clear from Remark 2.

Remark 3. The derivation of (2) is similar to the proof of Theorem 2. For any fixed m,

$$\min_{1\le k\le m}\frac{F(x_k)}{x_k}\times\max_{1\le k\le m}\frac{k/n}{F(x_k)}\le\tilde{f}_n(0+)\le f_0\times\max_{1\le k\le n}\frac{k/n}{F(x_k)};$$

and (2) follows by letting $n\to\infty$ and then $m\to\infty$.

6. Global and Pointwise Consistency

It has been shown that $\hat{f}(0+)$ and $\tilde{f}(0+)$ differ substantially. That $\hat{f}_n(x)$ and $\tilde{f}_n(x)$ do not differ very much for $x>0$ is shown in this section.

Let F_n denote the empirical distribution function, and $\tilde{F}_n$ the least concave majorant of F_n. Then $\tilde{f}_n(x)=\tilde{F}_n'(x)$, the left hand derivative of $\tilde{F}_n$ at x, for all $0<x<\infty$. Let $\hbar$ denote the Hellinger metric, so that

$$\hbar^2(g_1,g_2)=\int_0^\infty\left(\sqrt{g_1}-\sqrt{g_2}\right)^2dx=2\left\{1-\int_0^\infty\sqrt{g_1g_2}dx\right\}$$

for densities $g_1,g_2\in G$.

Theorem 3.

$$\hbar^2(\tilde{f}_n,\hat{f}_n)\le\alpha[\tilde{f}_n(0+)-\hat{f}_n(0+)].$$

Proof. Since $\hat{f}_n$ maximizes the penalized likelihood function,

$$\begin{aligned}0&\le\ell_\alpha(\hat{f}_n)-\ell_\alpha(\tilde{f}_n)\\&=n\left\{\int_0^\infty\log\hat{f}dF_n-\int_0^\infty\log\tilde{f}dF_n\right\}+n\alpha[\tilde{f}_n(0+)-\hat{f}_n(0+)].\end{aligned}$$

From Theorem 1.2.1 of RWD, $\tilde{f}_n$ decreases only at values x_k for which $\tilde{F}_n(x_k)=F_n(x_k)$. It follows that $\int_0^\infty\log\tilde{f}_ndF_n=\int_0^\infty\log\tilde{f}_nd\tilde{F}_n$, and, therefore, that

$$\begin{aligned}0&\le\int_0^\infty\log\frac{\hat{f}_n}{\tilde{f}_n}d\tilde{F}_n+\int_0^\infty(\tilde{F}_n-F_n)d\log\hat{f}_n+\alpha[\tilde{f}_n(0+)-\hat{f}_n(0+)]\\&\le-\hbar^2(\tilde{f}_n,\hat{f}_n)+\alpha[\tilde{f}_n(0+)-\hat{f}_n(0+)],\end{aligned}$$

where the final inequality follows by noting that $\tilde{F}_n-F_n\ge0$, writing $\log x=2\log\sqrt{x}\le2(\sqrt{x}-1)$ for $0<x<\infty$, and using the second expression for squared Hellinger distance. The theorem follows immediately.

Corollary. *If $\alpha=\alpha_n\to0$ as $n\to\infty$, then $\hbar(f,\hat{f}_n)\to0$ in probability.*

The easy proof is left as an exercise.

Pointwise behavior is considered next.

Proposition 1. *If f is strictly decreasing near zero, then $\hat{f}_n(x) - \tilde{f}_n(x) = O_p(\alpha)$ as $n \to \infty$ for all $0 < x < \infty$.*

Proof (Outline). If f is strictly decreasing near zero, then it may be shown that $\min_{1 \le r \le k}$ may be replaced by $\min_{2 \le r \le k}$ in (5) for all $k \ge n\varepsilon$ with probability approaching one as $n \to \infty$ for any $\varepsilon > 0$. It then follows from the definition of $w_k(\gamma)$ that $\hat{f}_n(x) = \tilde{f}_n(x)/\hat{\gamma}_n$ with probability approaching one as $n \to \infty$ for fixed $x > 0$, so that

$$\hat{f}_n(x) - \tilde{f}_n(x) = \left(\frac{1}{\hat{\gamma}_n} - 1\right)\tilde{f}_n(x) = O_p(\alpha).$$

Corollary. *If $\alpha = o(n^{-1/3})$ as $n \to \infty$, then $n^{1/3}[\hat{f}_n(x) - f(x)]$ has the same limiting distribution, if any, as $n^{1/3}[\tilde{f}_n(x) - f(x)]$ for all $x > 0$.*

The latter distribution was found by Prakasa Rao (1969) and Groeneboom (1985), under modest additional conditions.

7. Asymptotic Distributions

In this section it is required that

$$F(x) = f_0 x - f_1 x^p + o(x^p) \quad \text{as} \quad x \searrow 0, \tag{10}$$

where $0 < f_0 = f(0+) < \infty$, $0 \le f_1 < \infty$, and $1 < p < \infty$. It is further required that

$$\alpha = cn^{-pq}, \tag{11}$$

where $q = 1/(2p-1)$ and $0 < c < \infty$. Let $r = (p-1)q = (p-1)/(2p-1)$. Then $0 < r < \frac{1}{2} < pq < 1$.

Theorem 4. *Let $\beta = f_1 f_0^{p-1}$. Then*

$$n^r\{\hat{f}_n(0+) - f_0\} \Rightarrow S_{c,\beta} = \sup_{0<t<\infty} \frac{W(t) - [c + \beta t^p]}{t}$$

as $n \to \infty$, where $W(t)$, $0 \le t < \infty$, denotes a standard Brownian motion and $\Rightarrow$ denotes convergence in distribution.

The proof of Theorem 4 is similar to that of Theorem 2.1 of Groeneboom (1985). The details are omitted here, but available in Woodroofe and Sun (1991).

Kernel estimates of the density of $\hat{f}(0+)/f(0+)$ are presented in Figures 1 and 2 for the same sample sizes and f's described in Section 4 and selected α.

Remark 4. If $f_1 = 0$, then the limiting distribution is exponential with failure rate $2c$. For then $P\{S_{c,0} > z\} = P\{W(t) > c + zt, \exists\, 0 < t < \infty\} = e^{-2cz}$ for all $0 < z < \infty$. See, e.g., Breiman (1968, pp.287-290).

Remark 5. For $\beta > 0$, $S_{c,\beta} \leq S_{c,0}$, so that $P\{S_{c,\beta} > z\} \leq e^{-2cz}$ for all $z > 0$. In fact, it is not difficult to see that for any $\beta > 0$,

$$P\{S_{c,\beta} > z\} \sim e^{-2cz}, \quad \text{as} \quad z \to \infty.$$

Remark 6. From the simulations of Table 2 and others not reported here, it appears that the global performance of $\hat{f}$ is insensitive to α, while that of $\hat{f}(0)$ is sensitive to α. Further, it appears that setting $\alpha = n^{-pq}$ provides a good choice for the half normal (and presumably other cases with $p = 3$). For the standard exponential ($p = 2$), setting $\alpha = n^{-pq}$ appears to oversmooth, and letting $\alpha = .7n^{-pq}$ provides a much better choice.

Acknowledgement

This research was supported by the National Science Foundation. Thanks to Geurt Jongbloed for Equation (6) and parts of Remark 1 and to a referee for helpful comments on an earlier draft of this manuscript.

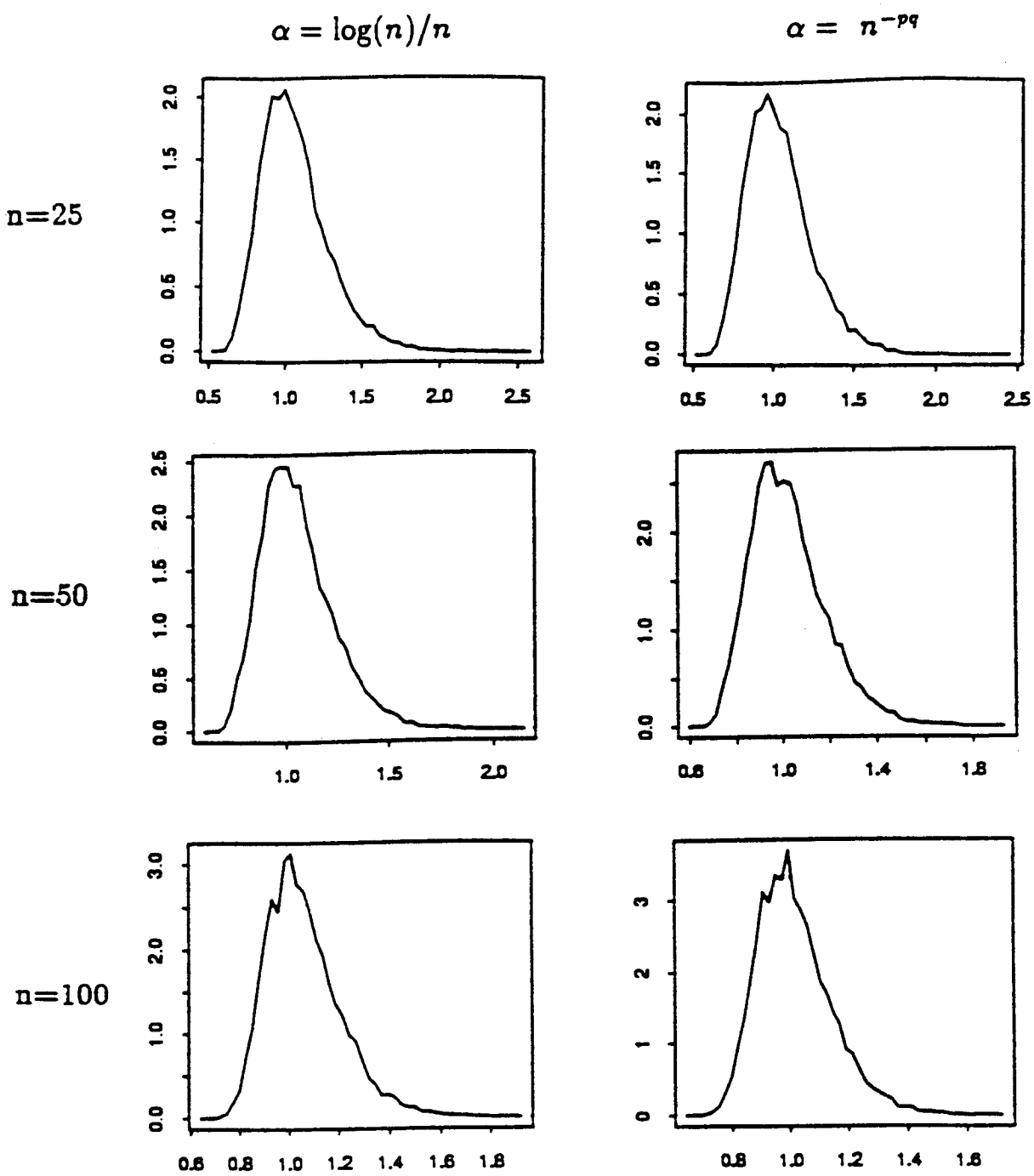

Figure 1. Kernel density estimates of $\hat{f}_n(0+)/f(0+)$ (window width = default in S)

Note: f is the half normal density. $\hat{f}_n$ is penalized MLE with penalizing parameter α based on data of size n from f. The simulation size is 10000 for all cases.

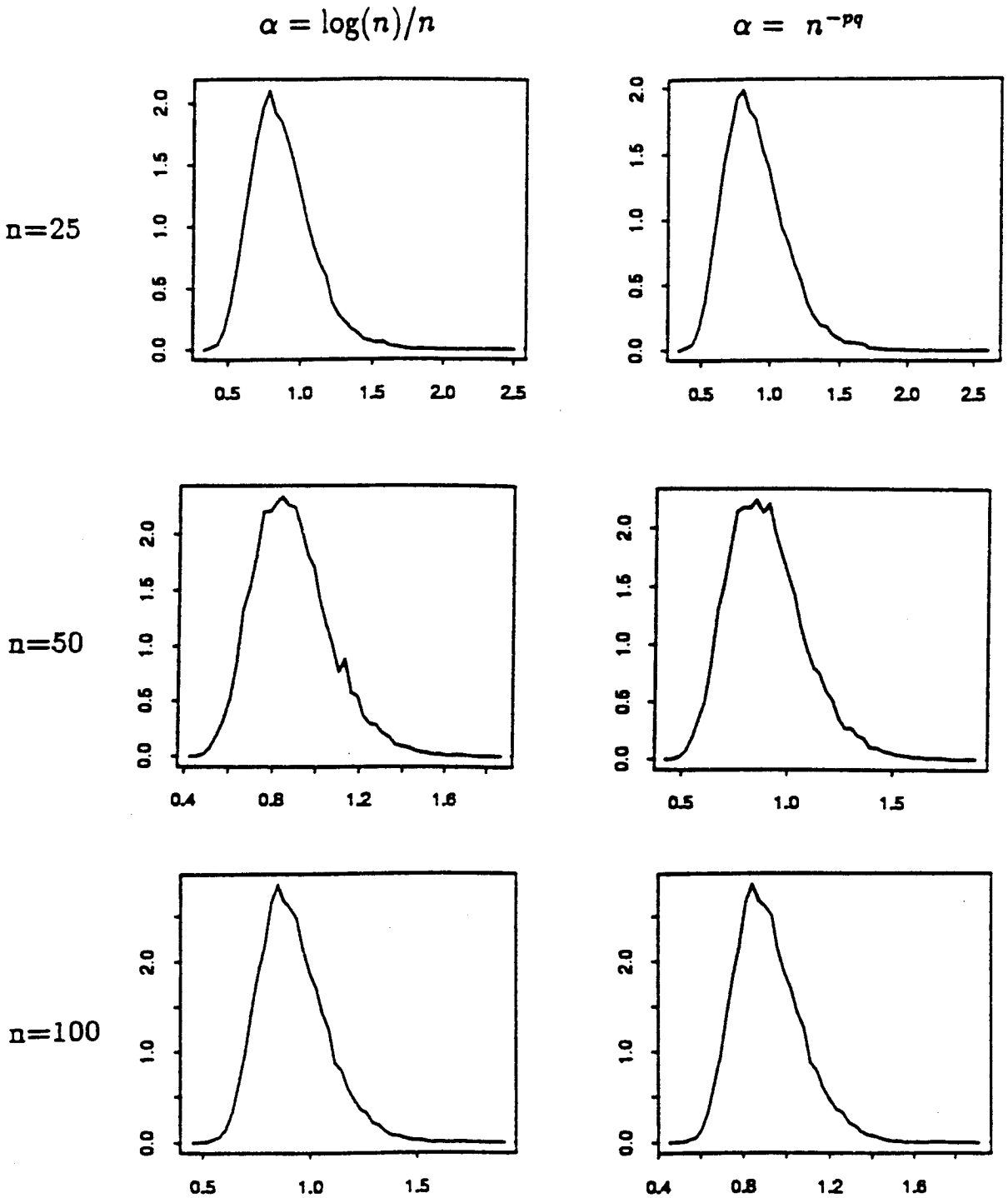

Figure 2. Kernel density estimates of $\hat{f}_n(0+)/f(0+)$ (window width = default in S)

Note: f is exponential density. $\hat{f}_n$ is penalized MLE with penalizing parameter α based on data of size n from f. The simulation size is 10000 for all cases. S is a statistical package supplied by AT&T.

References

Breiman, L. (1968). *Probability.* Addison Weseley.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, 2nd edition. John Wiley.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Groeneboom, P. (1985). Estimating a monotone density. *Proc. Conf. in Honor of Jerzy Neyman and Jack Kiefer* (Edited by L. LeCam and R. Olshen), **2**, 539–555.

Lynden-Bell, D. (1991). Eddington Malmquist bias, streaming motions, and the distribution of galaxies. To appear in *Statistical Challenges in Modern Astronomy* (Edited by J. Babu and E. Feigelson).

Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser.A* **31**, 23–36.

Prakasa Rao, B. L. S. (1983). *Nonparametric Function Estimation.* John Wiley.

Robertson, T., Wright, F. and Dykstra, R. (1988). *Order Restricted Inference.* John Wiley.

Rockafellar, R. (1970). *Convex Analysis.* Princeton.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76**, 751–761.

Wegman, E. J. (1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41**, 457–471.

Woodroofe, M. (1991). Discussion of "Eddington Malmquist bias, streaming motions, and the distribution of galaxies". *Statistical Challenges in Modern Astronomy* (Edited by J. Babu and E. Feigelson), 217–220.

Woodroofe, M. and Sun, J. (1991). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. Technical Report, Statistics Department, The University of Michigan.

Statistics Department, The University of Michigan, Ann Arbor, MI 48109-1027, U.S.A.

(Received October 1991; accepted January 1993)