THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS

PETER J. HUBER
SWISS FEDERAL INSTITUTE OF TECHNOLOGY

1. Introduction and summary

This paper proves consistency and asymptotic normality of maximum likelihood (ML) estimators under weaker conditions than usual.

In particular, (i) it is not assumed that the true distribution underlying the observations belongs to the parametric family defining the ML estimator, and (ii) the regularity conditions do not involve the second and higher derivatives of the likelihood function.

The need for theorems on asymptotic normality of ML estimators subject to (i) and (ii) becomes apparent in connection with robust estimation problems; for instance, if one tries to extend the author's results on robust estimation of a location parameter [4] to multivariate and other more general estimation problems.

Wald's classical consistency proof [6] satisfies (ii) and can easily be modified to show that the ML estimator is consistent also in case (i), that is, it converges to the θ_0 characterized by the property $E(\log f(x, \theta) - \log f(x, \theta_0)) < 0$ for $\theta \neq \theta_0$, where the expectation is taken with respect to the true underlying distribution.

Asymptotic normality is more troublesome. Daniels [1] proved asymptotic normality subject to (ii), but unfortunately he overlooked that a crucial step in his proof (the use of the central limit theorem in (4.4)) is incorrect without condition (2.2) of Linnik [5]; this condition seems to be too restrictive for many purposes.

In section 4 we shall prove asymptotic normality, assuming that the ML estimator is consistent. For the sake of completeness, sections 2 and 3 contain, therefore, two different sets of sufficient conditions for consistency. Otherwise, these sections are independent of each other. Section 5 presents two examples.

2. Consistency: case A

Throughout this section, which rephrases Wald's results on consistency of the ML estimator in a slightly more general setup, the parameter set Θ is a locally compact space with a countable base, $(\mathfrak{X}, \mathfrak{A}, P)$ is a probability space, and $\rho(x, \theta)$ is some real-valued function on $\mathfrak{X} \times \Theta$.

Assume that $x_1, x_2, \cdots$ are independent random variables with values in $\mathfrak{X}$ having the common probability distribution P. Let $T_n(x_1, \cdots, x_n)$ be any sequence of functions $T_n: \mathfrak{X}^n \to \Theta$, measurable or not, such that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i, T_n) - \inf_{\theta} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta) \to 0 \tag{1}$$

almost surely (or in probability—more precisely, outer probability). We want to give sufficient conditions ensuring that every such sequence converges almost surely (or in probability) toward some constant θ_0.

If $dP = f(x, \theta_0)\, d\mu$ and $\rho(x, \theta) = -\log f(x, \theta)$ for some measure μ on $(\mathfrak{X}, \mathfrak{A})$ and some family of probability densities $f(x, \theta)$, then the ML estimator of θ_0 evidently satisfies condition (1).

Convergence of T_n shall be proved under the following set of assumptions.

ASSUMPTIONS.

(A-1). *For each fixed $\theta \in \Theta$, $\rho(x, \theta)$ is $\mathfrak{A}$-measurable, and $\rho(x, \theta)$ is separable in the sense of Doob: there is a P-null set N and a countable subset $\Theta' \subset \Theta$ such that for every open set $U \subset \Theta$ and every closed interval A, the sets*

$$\{x|\rho(x, \theta) \in A, \forall \theta \in U\}, \qquad \{x|\rho(x, \theta) \in A, \forall \theta \in U \cap \Theta'\} \tag{2}$$

differ by at most a subset of N.

This assumption ensures measurability of the infima and limits occurring below. For a fixed P, ρ might always be replaced by a separable version (see Doob [2], p. 56 ff.).

(A-2). *The function ρ is a.s. lower semicontinuous in θ, that is,*

$$\inf_{\theta' \in U} \rho(x, \theta') \to \rho(x, \theta), \qquad \text{a.s.} \tag{3}$$

as the neighborhood U of θ shrinks to $\{\theta\}$.

(A-3). *There is a measurable function $a(x)$ such that*

$$\begin{aligned} &E\{\rho(x, \theta) - a(x)\}^- < \infty \quad \text{for all} \quad \theta \in \Theta, \\ &E\{\rho(x, \theta) - a(x)\}^+ < \infty \quad \text{for some} \quad \theta \in \Theta. \end{aligned} \tag{4}$$

Thus, $\gamma(\theta) = E\{\rho(x, \theta) - a(x)\}$ is well-defined for all θ.

(A-4). *There is a $\theta_0 \in \Theta$ such that $\gamma(\theta) > \gamma(\theta_0)$ for all $\theta \neq \theta_0$.*

If Θ is not compact, let ∞ denote the point at infinity in its one-point compactification.

(A-5). *There is a continuous function $b(\theta) > 0$ such that*

(i) $\inf_{\theta \in \Theta} \dfrac{\rho(x, \theta) - a(x)}{b(\theta)} \geq h(x)$

for some integrable h;

(ii) $\liminf_{\theta \to \infty} b(\theta) > \gamma(\theta_0)$;

(iii) $E\left\{\liminf_{\theta \to \infty} \dfrac{\rho(x, \theta) - a(x)}{b(\theta)}\right\} \geq 1.$

Θ is compact, then (ii) and (iii) are redundant.

EXAMPLE. Let $\Theta = \mathfrak{X}$ be the real axis, and let P be any probability distribution having a unique median θ_0. Then (A-1) to (A-5) are satisfied for $\rho(x, \theta) = |x - \theta|$, $a(x) = |x|$, $b(\theta) = |\theta| + 1$. (This will imply that the sample median is a consistant estimate of the median.)

Taken together, (A-2), (A-3), and (A-5) (i) imply by monotone convergence the following strengthened version of (A-2).

(A-2′). *As the neighborhood U of θ shrinks to $\{\theta\}$,*

$$E\{\inf_{\theta' \in U} \rho(x, \theta') - a(x)\} \to E\{\rho(x, \theta) - a(x)\}. \tag{5}$$

For the sake of simplicity we shall from now on absorb $a(x)$ into $\rho(x, \theta)$. Note that the set $\{\theta \in \Theta | E(|\rho(x, \theta) - a(x)|) < \infty\}$ is independent of the particular choice of $a(x)$; if there is an $a(x)$ satisfying (A-3), then one might choose $a(x) = \rho(x, \theta_0)$.

LEMMA 1. *If* (A-1), (A-3), *and* (A-5) *hold, then there is a compact set $C \subset \Theta$ such that every sequence T_n satisfying* (1) *almost surely ultimately stays in C.*

PROOF. By (A-5) (ii), there is a compact C and a $0 < \epsilon < 1$ such that

$$\inf_{\theta \notin C} b(\theta) \geq \frac{\gamma(\theta_0) + \epsilon}{1 - \epsilon}; \tag{6}$$

by (A-5) (i), (iii) and monotone convergence, C may be chosen so large that

$$E\left\{\inf_{\theta \notin C} \frac{\rho(x, \theta)}{b(\theta)}\right\} \geq 1 - \tfrac{1}{2}\epsilon. \tag{7}$$

By the strong law of large numbers, we have a.s. for sufficiently large n

$$\inf_{\theta \notin C} \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\rho(x_i, \theta)}{b(\theta)}\right\} \geq \frac{1}{n}\sum_{i=1}^{n} \inf_{\theta \notin C} \frac{\rho(x_i, \theta)}{b(\theta)} \geq 1 - \epsilon; \tag{8}$$

hence,

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta) \geq (1 - \epsilon)b(\theta) \geq \gamma(\theta_0) + \epsilon \tag{9}$$

for $\forall \theta \notin C$, which implies the lemma, since for sufficiently large n

$$\inf_{\theta} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta) \leq \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta_0) \leq \gamma(\theta_0) + \tfrac{1}{2}\epsilon. \tag{10}$$

By using convergence in probability in (1) and the weak law of large numbers, one shows similarly that $T_n \in C$ with probability tending to 1. (Note that a.s. convergence does not imply convergence in probability, if T_n is not measurable!)

THEOREM 1. *If* (A-1), (A-2′), (A-3), *and* (A-4) *hold, then every sequence T_n satisfying* (1), *and the conclusion of lemma* 1, *converges to θ_0 almost surely. An analogous statement is true for convergence in probability.*

PROOF. We may restrict attention to the compact set C. Let U be an open neighborhood of θ. By (A-2′), γ is lower semicontinuous; hence its infimum on the compact set $C \setminus U$ is attained and is—because of (A-4)—strictly greater than $\gamma(\theta_0)$, say $\geq \gamma(\theta_0) + 4\epsilon$ for some $\epsilon > 0$. Because of (A-2′), each $\theta \in C \setminus U$ admits a neighborhood U_θ such that

$$E\{\inf_{\theta' \in U_\theta} \rho(x, \theta')\} \geq \gamma(\theta_0) + 3\epsilon. \tag{11}$$

Select a finite number of points θ_s such that the $U_s = U_{\theta_s}$, $1 \leq s \leq N$, cover $C \setminus U$. By the strong law of large numbers, we have a.s. for sufficiently large n and all $1 \leq s \leq N$,

$$\inf_{\theta' \in U_s} \frac{1}{n} \sum \rho(x_i, \theta') \geq \frac{1}{n} \sum \inf_{\theta' \in U_s} \rho(x_i, \theta') \geq \gamma(\theta_0) + 2\epsilon \tag{12}$$

and

$$\frac{1}{n} \sum \rho(x_i, \theta_0) \leq \gamma(\theta_0) + \epsilon. \tag{13}$$

It follows that

$$\inf_{\theta \in C \setminus U} \frac{1}{n} \sum \rho(x_i, \theta) \geq \inf_{\theta \in U} \frac{1}{n} \sum \rho(x_i, \theta) + \epsilon, \tag{14}$$

which implies the theorem. Convergence in probability is proved analogously.

REMARKS. (1) If assumption (A-4) is omitted, the above arguments show that T_n a.s. ultimately stays in any neighborhood of the (necessarily compact) set $\{\theta \in \Theta | \gamma(\theta) = \inf_{\theta'} \gamma(\theta')\}$.

(2) Quite often (A-5) is not satisfied—for instance, if one estimates location and scale simultaneously—but the conclusion of lemma 1 can be verified quite easily by ad hoc methods. (This happens also in Wald's classical proof.) I do not know of any fail-safe replacement for (A-5).

3. Consistency: case B

Let Θ be locally compact with a countable base, let $(\mathfrak{X}, \mathfrak{A}, P)$ be a probability space, and let $\psi(x, \theta)$ be some function on $\mathfrak{X} \times \Theta$ with values in m-dimensional Euclidean space R^m.

Assume that $x_1, x_2, \cdots$ are independent random variables with values in $\mathfrak{X}$, having the common probability distribution P. We want to give sufficient conditions that any sequence $T_n: \mathfrak{X}^n \to \Theta$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \psi(x_i, T_n) \to 0 \tag{15}$$

almost surely (or in probability), converges almost surely (or in probability) toward some constant θ_0.

If Θ is an open subset of R^m, and if $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$ for some differentiable parametric family of probability densities on $\mathfrak{X}$, then the ML estimate of θ will satisfy (15). However, our ψ need not be a total differential.

Convergence of T_n shall be proved under the following set of assumptions.

ASSUMPTIONS.

(B-1). *For each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is $\mathfrak{A}$-measurable, and $\psi(x, \theta)$ is separable* (*see* (A-1)).

(B-2). *The function ψ is a.s. continuous in θ:*

$$\lim_{\theta' \to \theta} |\psi(x, \theta') - \psi(x, \theta)| = 0, \quad \text{a.s.} \tag{16}$$

(B-3). *The expected value* $\lambda(\theta) = E\psi(x, \theta)$ *exists for all* $\theta \in \Theta$, *and has a unique zero at* $\theta = \theta_0$.

(B-4). *There exists a continuous function which is bounded away from zero,* $b(\theta) \geq b_0 > 0$, *such that*

(i) $\sup_\theta \dfrac{|\psi(x, \theta)|}{b(\theta)}$ is integrable,

(ii) $\liminf_{\theta \to \infty} \dfrac{|\lambda(\theta)|}{b(\theta)} \geq 1$,

(iii) $E\left\{\limsup_{\theta \to \infty} \dfrac{|\psi(x, \theta) - \lambda(\theta)|}{b(\theta)}\right\} < 1.$

In view of (B-4) (i), (B-2) can be strengthened to

(B-2'). *As the neighborhood* U *of* θ *shrinks to* $\{\theta\}$

$$E(\sup_{\theta' \in U} |\psi(x, \theta') - \psi(x, \theta)|) \to 0. \tag{17}$$

It follows immediately from (B-2') that λ is continuous. Moreover, if there is a function b satisfying (B-4), one may obviously choose

$$b(\theta) = \max(|\lambda(\theta)|, b_0). \tag{18}$$

LEMMA 2. *If* (B-1) *and* (B-4) *hold, then there is a compact set* $C \subset \Theta$ *such that any sequence* T_n *satisfying* (15) *a.s. ultimately stays in* C.

PROOF. With the aid of (B-4) (i), (iii), and the dominated convergence theorem, choose C so large that the expectation of

$$v(x) = \sup_{\theta \notin C} \frac{|\psi(x, \theta) - \lambda(\theta)|}{b(\theta)} \tag{19}$$

is smaller than $1 - 3\epsilon$ for some $\epsilon > 0$, and that also (by (B-4) (ii))

$$\inf_{B \notin C} \frac{|\lambda(\theta)|}{b(\theta)} \geq 1 - \epsilon. \tag{20}$$

By the strong law of large numbers, we have a.s. for sufficiently large n,

$$\sup_{\theta \notin C} \frac{|n^{-1} \sum [\psi(x_i, \theta) - \lambda(\theta)]|}{b(\theta)} \leq \frac{1}{n} \sum v(x_i) \leq 1 - 2\epsilon; \tag{21}$$

thus,

$$\left|\frac{1}{n} \sum [\psi(x_i, \theta) - \lambda(\theta)]\right| \leq (1 - 2\epsilon) b(\theta) \leq \frac{1 - 2\epsilon}{1 - \epsilon} |\lambda(\theta)| \leq (1 - \epsilon)|\lambda(\theta)| \tag{22}$$

for $\forall \theta \notin C$, or

$$\left|\frac{1}{n} \sum \psi(x_i, \theta)\right| \geq \epsilon|\lambda(\theta)| \geq \epsilon(1 - \epsilon) b_0 \tag{23}$$

for $\forall \theta \notin C$, which implies the lemma.

THEOREM 2. *If* (B-1), (B-2'), *and* (B-3) *hold, then every sequence* T_n *satisfying* (15) *and the conclusion of lemma* 2 *converges to* θ_0 *almost surely. An analogous statement is true for convergence in probability.*

PROOF. We may restrict attention to the compact set C. For any open neighborhood U of θ_0, the infimum of the continuous function $|\lambda(\theta)|$ on the compact set $C \setminus U$ is strictly positive, say $\geq 5\epsilon > 0$. For every $\theta \in C \setminus U$, let U_θ be a neighborhood of θ such that by (B-2'),

$$E\{\sup_{\theta' \in U_\theta} |\psi(x, \theta') - \psi(x, \theta)|\} \leq \epsilon; \tag{24}$$

hence, $|\lambda(\theta') - \lambda(\theta)| \leq \epsilon$ for $\theta' \in U_\theta$. Select a finite subcover $U_s = U_{\theta_s}$, $1 \leq s \leq N$. Then we have a.s. for sufficiently large n

$$\begin{aligned} \sup_{\theta' \in C \setminus U} \left|\frac{1}{n} \sum [\psi(x, \theta') - \lambda(\theta')]\right| & \\ \leq \sup_{1 \leq s \leq N} \frac{1}{n} \sum \sup_{\theta' \in U_s} |\psi(x, \theta') - \psi(x, \theta_s)| & \\ + \sup_{1 \leq s \leq N} \left|\frac{1}{n} \sum [\psi(x, \theta_s) - \lambda(\theta_s)]\right| + \epsilon & \\ \leq 4\epsilon. & \end{aligned} \tag{25}$$

Since $|\lambda(\theta)| \geq 5\epsilon$ for $\theta \in C \setminus U$, this implies

$$\left|\frac{1}{n} \sum \psi(x_i, \theta)\right| \geq \epsilon \tag{26}$$

for $\theta \in C \setminus U$ and sufficiently large n, which proves the theorem. Convergence in probability is proved analogously.

4. Asymptotic normality

In the following, Θ is an open subset of m-dimensional Euclidean space R^m, $(\mathfrak{X}, \mathfrak{A}, P)$ is a probability space, and $\psi: \mathfrak{X} \times \Theta \to R^m$ is some function.

Assume that $x_1, x_2, \cdots$ are independent identically distributed random variables with values in $\mathfrak{X}$ and common distribution P. We want to give sufficient conditions ensuring that every sequence $T_n = T_n(x_1, \cdots, x_n)$ satisfying

$$(1/\sqrt{n}) \sum_{i=1}^{n} \psi(x_i, T_n) \to 0 \quad \text{in probability} \tag{27}$$

is asymptotically normal.

In particular, this result will imply asymptotic normality of ML estimators: let $f(x, \theta)$, $\theta \in \Theta$ be a family of probability densities with respect to some measure μ on $(\mathfrak{X}, \mathfrak{A})$, $dP = f(x, \theta_0)\, d\mu$ for some θ_0, $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$, then the sequence of ML estimators T_n of θ_0 satisfies (27).

Assuming that consistency of T_n has already been proved by some other means, we shall establish asymptotic normality under the following conditions.

ASSUMPTIONS.

(N-1). *For each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is $\mathfrak{A}$-measurable and $\psi(x, \theta)$ is separable* (*see* (A-1)).

Put

$$\lambda(\theta) = E\psi(x, \theta),$$
$$u(x, \theta, d) = \sup_{|\tau-\theta|\leq d} |\psi(x, \tau) - \psi(x, \theta)|. \tag{28}$$

Expectations are always taken with respect to the true underlying distribution P.

(N-2). *There is a* $\theta_0 \in \theta$ *such that* $\lambda(\theta_0) = 0$.

(N-3). *There are strictly positive numbers* a, b, c, d_0 *such that*

(i) $|\lambda(\theta)| \geq a \cdot |\theta - \theta_0|$ for $|\theta - \theta_0| \leq d_0$,

(ii) $Eu(x, \theta, d) \leq b \cdot d$ for $|\theta - \theta_0| + d \leq d_0$, $d \geq 0$,

(iii) $E[u(x, \theta, d)^2] \leq c \cdot d$ for $|\theta - \theta_0| + d \leq d_0$, $d \geq 0$.

Here, $|\theta|$ denotes any norm equivalent to Euclidean norm. Condition (iii) is somewhat stronger than needed; the proof can still be pushed through with $E[u(x, \theta, d)^2] \leq o(|\log d|^{-1})$.

(N-4). *The expectation* $E[|\psi(x, \theta_0)|^2]$ *is finite.*

Put

$$Z_n(\tau, \theta) = \frac{\left|\sum_{i=1}^{n} [\psi(x_i, \tau) - \psi(x_i, \theta) - \lambda(\tau) + \lambda(\theta)]\right|}{\sqrt{n} + n|\lambda(\tau)|}. \tag{29}$$

The following lemma is crucial.

LEMMA 3. *Assumptions* (N-1), (N-2), (N-3) *imply*

$$\sup_{|\tau-\theta_0|\leq d_0} Z_n(\tau, \theta_0) \to 0 \tag{30}$$

in probability, as $n \to \infty$.

PROOF. For the sake of simplicity, and without loss of generality, take $|\theta|$ to be the sup-norm, $|\theta| = \max(|\theta_1|, \cdots, |\theta_m|)$ for $\theta \in R^m$. Choose the coordinate system such that $\theta_0 = 0$ and $d_0 = 1$.

The idea of the proof is to subdivide the cube $|\tau| \leq 1$ into a slowly increasing number of smaller cubes and to bound $Z_n(\tau, 0)$ in probability on each of those smaller cubes.

Put $q = 1/M$, where $M \geq 2$ is an integer to be chosen later, and consider the concentric cubes

$$C_k = \{\theta \mid |\theta| \leq (1 - q)^k\}, \qquad k = 0, 1, \cdots, k_0. \tag{31}$$

Subdivide the difference $C_{k-1} \setminus C_k$ into smaller cubes (see figure 1) with edges of length

$$2d = (1 - q)^{k-1}q, \tag{32}$$

such that the coordinates of their centers ξ are odd multiples of d, and

$$|\xi| = (1 - q)^{k-1}\left(1 - \frac{q}{2}\right). \tag{33}$$

For each value of k there are less than $(2M)^m$ such small cubes, so there are $N < k_0 \cdot (2M)^m$ cubes contained in $C_0 \setminus C_{k_0}$; number them $C_{(1)}, \cdots, C_{(N)}$.

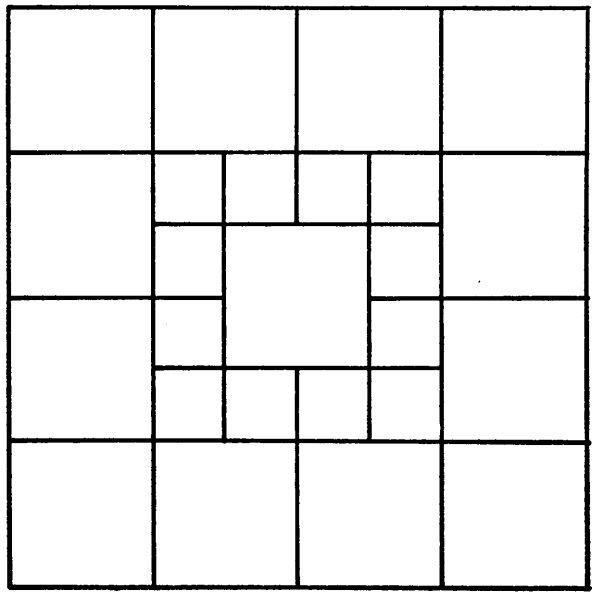

Figure 1

Now let $\epsilon > 0$ be given. We shall show that for a proper choice of M and of $k_0 = k_0(n)$, the right-hand side of

$$P(\sup_{\tau \in C_0} Z_n(\tau, 0) \geq 2\epsilon) \leq P(\sup_{\tau \in C_{k_0}} Z_n(\tau, 0) \geq 2\epsilon) + \sum_{j=1}^{N} P(\sup_{\tau \in C_{(j)}} Z_n(\tau, 0) \geq 2\epsilon) \tag{34}$$

tends to 0 with increasing n, which establishes lemma 1.

Actually, we shall choose

$$M \geq (3b)/(\epsilon a), \tag{35}$$

and $k_0 = k_0(n)$ is defined by

$$(1 - q)^{k_0} \leq n^{-\gamma} < (1 - q)^{k_0 - 1}, \tag{36}$$

where $\frac{1}{2} < \gamma < 1$ is an arbitrary fixed number. Thus

$$k_0(n) - 1 < \frac{\gamma \cdot \log n}{|\log (1 - q)|} \leq k_0(n), \tag{37}$$

hence

$$N = O(\log n). \tag{38}$$

Now take any of the cubes $C_{(j)}$, with center ξ and edges of length $2d$ according to (32) and (33). For $\tau \in C_{(j)}$ we have then by (N-3),

$$|\lambda(\tau)| \geq a|\tau| \geq a \cdot (1 - q)^k, \tag{39}$$

$$|\lambda(\tau) - \lambda(\xi)| \leq Eu(x, \xi, d) \leq bd \leq b(1 - q)^k q. \tag{40}$$

We have

$$Z_n(\tau, 0) \leq Z_n(\tau, \xi) + \frac{\left|\sum_{i=1} [\psi(x_i, \xi) - \psi(x_i, 0) - \lambda(\xi)]\right|}{\sqrt{n} + n|\lambda(\tau)|}, \tag{41}$$

hence

$$\sup_{\tau \in C_{(j)}} Z_n(\tau, 0) \leq U_n + V_n \tag{42}$$

with

$$U_n = \frac{\sum_{i=1}^{n} [u(x_i, \xi, d) + Eu(x, \xi, d)]}{na(1-q)^k}, \tag{43}$$

$$V_n = \frac{\left|\sum_{i=1}^{n} [\psi(x_i, \xi) - \psi(x_i, 0) - \lambda(\xi)]\right|}{na(1-q)^k}. \tag{44}$$

Thus,

$$P(U_n \geq \epsilon) = P\left\{\sum_{i=1}^{n} [u(x_i, \xi, d) - Eu(x, \xi, d)] \geq \epsilon na(1-q)^k - 2nEu(x, \xi, d)\right\}. \tag{45}$$

In view of (40) and (35),

$$\epsilon a(1-q)^k - 2Eu(x, \xi, d) \geq \epsilon a(1-q)^k - 2bq(1-q)^k \geq bq(1-q)^k, \tag{46}$$

hence (N-3) (iii) and Chebyshev's inequality yield

$$P(U_n \geq \epsilon) \leq \frac{c}{b^2q(1-q)} \cdot \frac{1}{n(1-q)^{k-1}}. \tag{47}$$

In a similar way,

$$P(V_n \geq \epsilon) \leq \frac{c}{9b^2q^2(1-q)^2} \cdot \frac{1}{n(1-q)^{k-1}}. \tag{48}$$

Hence, we obtain from (36), (42), (47), (48) that

$$P(\sup_{\tau \in C_{(j)}} Z_n(\tau, 0) \geq 2\epsilon) \leq K \cdot n^{\gamma-1} \tag{49}$$

with

$$K = \frac{c}{b^2q(1-q)} + \frac{c}{9b^2q^2(1-q)^2}. \tag{50}$$

Furthermore,

$$\sup_{\tau \in C_{k_0}} Z_n(\tau, 0) \leq \frac{\sum_{1}^{n} [u(x_i, 0, d) + Eu(x, 0, d)]}{\sqrt{n}} \tag{51}$$

with $d = (1-q)^{k_0} \leq n^{-\gamma}$. Hence,

$$P(\sup_{\tau \in C_{k_0}} Z_n(\tau, 0) \geq 2\epsilon) \leq P(\sum_{i=1}^{n} [u(x_i, 0, d) - Eu(x_i, 0, d)] \geq 2\sqrt{n}\epsilon - 2nEu(x, 0, d)). \tag{52}$$

Since $Eu(x, 0, d) \leq bd \leq bn^{-\gamma}$, there is an n_0 such that for $n \geq n_0$,

$$2\sqrt{n}\epsilon - 2nEu(x, 0, d) \geq \sqrt{n}\epsilon; \tag{53}$$

thus, by Chebyshev's inequality,

$$P(\sup_{\tau \in C_{k_0}} Z_n(\tau, 0) \geq 2\epsilon) \leq c \cdot \epsilon^{-2} \cdot n^{-\gamma}. \tag{54}$$

Now, putting (34), (38), (49) and (54) together, we obtain

$$P(\sup_{\tau \in C_0} Z_n(\tau, 0) \geq 2\epsilon) \leq O(n^{-\gamma}) + O(n^{\gamma-1} \log n), \tag{55}$$

which proves lemma 3.

THEOREM 3. *Assume that* (N-1) *to* (N-4) *hold and that* T_n *satisfies* (27). *If* $P(|T_n - \theta_0| \leq d_0) \to 1$, *then*

$$(1/\sqrt{n}) \sum_{i=1}^{n} \psi(x_i, \theta_0) + \sqrt{n}\lambda(T_n) \to 0 \tag{56}$$

in probability.

PROOF. Assume again $\theta_0 = 0$, $d_0 = 1$. We have

$$\sum_1^n \psi(x_i, T_n) = \sum_1^n [\psi(x_i, T_n) - \psi(x_i, 0) - \lambda(T_n)] + \sum_1^n [\psi(x_i, 0) + \lambda(T_n)]. \tag{57}$$

Thus, with probability tending to 1

$$\frac{\left|\sum_1^n [\psi(x_i, 0) + \lambda(T_n)]\right|}{\sqrt{n} + n|\lambda(T_n)|} \leq \sup_{|\tau| \leq 1} Z_n(\tau, 0) + (1/\sqrt{n})\left|\sum_1^n \psi(x_i, T_n)\right|. \tag{58}$$

The terms on the right-hand side tend to 0 in probability (lemma 1 and assumption (27)), so the left-hand side does also.

Now let $\epsilon > 0$ be given. Put $K^2 = 2E(|\psi(x, 0)|^2)/\epsilon$; then, by Chebyshev's inequality and for sufficiently large n, say $n \geq n_0$, the inequalities

$$\left|(1/\sqrt{n}) \sum_1^n \psi(x_i, 0)\right| \leq K \tag{59}$$

and

$$\left|\sum_1^n [\psi(x_i, 0) + \lambda(T_n)]\right| \leq \epsilon \cdot (\sqrt{n} + n|\lambda(T_n)|) \tag{60}$$

are violated with probabilities $\leq \frac{1}{2}\epsilon$; so both hold simultaneously with probability exceeding $1 - \epsilon$. But (60) implies

$$\sqrt{n}|\lambda(T_n)|(1 - \epsilon) \leq \epsilon + (1/\sqrt{n})\left|\sum_1^n \psi(x_i, 0)\right|, \tag{61}$$

hence, by (59), $\sqrt{n}|\lambda(T_n)| \leq (K + \epsilon)/(1 - \epsilon)$. Thus,

$$\left|(1/\sqrt{n}) \sum_1^n \psi(x_i, 0) + \sqrt{n}\lambda(T_n)\right| \leq (K + 1)\epsilon/(1 - \epsilon) \tag{62}$$

holds with probability exceeding $1 - \epsilon$ for $n \geq n_0$. Since the right-hand side of (62) can be made arbitrarily small by choosing ϵ small enough, the theorem follows.

COROLLARY. *Under the conditions of theorem* 3, *assume that* λ *has a nonsingular derivative* Λ *at* θ_0 *(that is,* $|\lambda(\theta) - \lambda(\theta_0) - \Lambda \cdot (\theta - \theta_0)| = o(|\theta - \theta_0|)$*). Then* $\sqrt{n}(T_n - \theta_0)$ *is asymptotically normal with mean* 0 *and covariance matrix* $\Lambda^{-1}C(\Lambda')^{-1}$, *(where* C *stands for the covariance matrix of* $\psi(x, \theta_0)$, *and* Λ' *is the transpose of* Λ*).*

PROOF. The proof is immediate.

REMARK. We worked under the tacit assumption that the T_n are measurable. Actually, this is irrelevant, except that for nonmeasurable T_n, some careful circumlocutions involving inner and outer probabilities are necessary.

Efficiency. Consider now the ordinary ML estimator, that is, assume that $dP = f(x, \theta_0)\, d\mu$ and that $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$ (derivative in measure). Assume that $\psi(x, \theta)$ is jointly measurable, and that (N-1), (N-3), and (N-4) hold locally uniformly in θ_0, and that the ML estimator is consistent.

We want to check whether the ML estimator is efficient. That is, we want to show that $\lambda(\theta_0) = 0$ and $\Lambda = -C = -I(\theta_0)$, where $I(\theta_0)$ is the information matrix. This implies, by the corollary to theorem 3, that the asymptotic variance of the ML estimator is $I(\theta_0)^{-1}$.

Obviously, with $\theta_t = \theta_0 + t \cdot (\theta - \theta_0)$, $0 \leq t \leq 1$,

$$
\begin{aligned}
0 &\geq \int [\log f(x, \theta) - \log f(x, \theta_0)] f(x, \theta_0)\, d\mu \\
&= \iint_0^1 \psi(x, \theta_t)\, dt\, f(x, \theta_0)\, d\mu \cdot (\theta - \theta_0) \\
&= \int_0^1 \lambda(\theta_t)\, dt \cdot (\theta - \theta_0).
\end{aligned} \tag{63}
$$

(The interchange of the order of the integrals is legitimate, since $\psi(x, \theta_t)$ is bounded in absolute value by the integrable $|\psi(x, \theta_0)| + u(x, \theta_0, |\theta - \theta_0|)$.)

Since λ is continuous, $\int_0^1 \lambda(\theta_t)\, dt \to \lambda(\theta_0)$ for $\theta \to \theta_0$, hence $\lambda(\theta_0) \cdot \eta \leq 0$ for any vector η, thus $\lambda(\theta_0) = 0$.

Now consider

$$
\begin{aligned}
\lambda(\theta) - \lambda(\theta_0) &= \int \psi(x, \theta) f(x, \theta_0)\, d\mu \\
&= -\int \psi(x, \theta)[f(x, \theta) - f(x, \theta_0)]\, d\mu \\
&= -\int \psi(x, \theta) \int_0^1 \psi(x, \theta_t) f(x, \theta_t)\, dt\, d\mu \cdot (\theta - \theta_0).
\end{aligned} \tag{64}
$$

But

$$
\begin{aligned}
\int \psi(x, \theta) \int_0^1 \psi(x, \theta_t) f(x, \theta_t)\, dt\, d\mu &= \iint_0^1 \psi(x, \theta_t)\psi(x, \theta_t) f(x, \theta_t)\, dt\, d\mu + r(\theta) \\
&= \int_0^1 I(\theta_t)\, dt + r(\theta),
\end{aligned} \tag{65}
$$

with

$$|r(\theta)| \leq \int_0^1 \int u(x, \theta_t, |\theta - \theta_0|)|\psi(x, \theta_t)|f(x, \theta_t)\, d\mu\, dt \leq O(|\theta - \theta_0|^{1/2}) \tag{66}$$

by Schwarz's inequality.

If $I(\theta)$ is continuous at θ_0, the assertion $\Lambda = -I(\theta_0)$ follows. If $I(\theta)$ should be discontinuous, a slight refinement of the above argument shows that $I(\theta_0)^{-1}$ is an upper bound for the asymptotic variance. I do not know whether (N-3) actually implies continuity of $I(\theta)$.

5. Examples

EXAMPLE 1. Let $\mathfrak{X} = \Theta = R^m (m \geq 2)$, and let $\rho(x, \theta) = |x - \theta|^p$, $1 \leq p \leq 2$, where $|\ |$ denotes Euclidean norm. Define T_n by the property that it minimizes $\sum_{i=1}^n \rho(x_i, T_n)$; or, if we put

$$\psi(x, \theta) = -\frac{1}{p}\frac{\partial}{\partial\theta}\rho(x, \theta) = |x - \theta|^{p-2}(x - \theta), \tag{67}$$

by the property $\sum_{i=1}^n \psi(x_i, T_n) = 0$. This estimator was considered by Gentleman [3].

A straightforward calculation shows that both u and u^2 satisfy Lipschitz conditions

$$u(x, \theta, d) \leq c_1 \cdot d \cdot |x - \theta|^{p-2} \tag{68}$$

$$u^2(x, \theta, d) \leq c_2 \cdot d \cdot |x - \theta|^{p-2} \tag{69}$$

for $0 \leq d \leq d_0 < \infty$. Thus, conditions (N-3) (ii) and (iii) are satisfied, provided that

$$E|x - \theta|^{p-2} \leq K < \infty \tag{70}$$

in some neighborhood of θ_0, which certainly holds if the true distribution has a bounded density with respect to Lebesgue measure. Furthermore, under the same condition (70),

$$\frac{\partial}{\partial\theta}\lambda(\theta) = E\frac{\partial\psi(x, \theta)}{\partial\theta}. \tag{71}$$

Thus

$$\operatorname{tr}\frac{\partial\lambda}{\partial\theta} = E \operatorname{tr}\frac{\partial\psi}{\partial\theta} = -(m + p - 2)E|x - \theta|^{p-2} < 0, \tag{72}$$

hence also (N-3) (i) is satisfied. Condition (N-1) is immediate, (N-2) and (N-4) hold if $E|x|^{2p-2} < \infty$, and consistency of T_n follows either directly from convexity of ρ, or from verifying (B-1) to (B-4) (with $b(\theta) = \max(1, |\theta|^{p-1})$).

EXAMPLE 2. (Cf. Huber [4], p. 79.)

Let $\mathfrak{X} = \theta = R$, and let $\rho(x, \theta) = \frac{1}{2}(x - \theta)^2$ for $|x - \theta| \leq k$, $\rho(x, \theta) = \frac{1}{2}k^2$ for $|x - \theta| > k$. Condition (A-4) of section 2, namely unicity of θ_0, imposes a restriction on the true underlying distribution; the other conditions are trivially sat-

isfied (with $a(x) \equiv 0, b(\theta) \equiv \frac{1}{2}k^2, h(x) \equiv 0$). Then, the T_n minimizing $\sum \rho(x_i, T_n)$ is a consistent estimate of θ_0.

Under slightly more stringent regularity conditions, it is also asymptotically normal. Assume for simplicity $\theta_0 = 0$, and assume that the true underlying distribution function F has a density F' in some neighborhoods of the points $\pm k$, and that F' is continuous at these points. Conditions (N-1), (N-2), (N-3) (ii), (iii), and (N-4) are obviously satisfied with $\psi(x, \theta) = (\partial/\partial\theta)\rho(x, \theta)$; if

$$\int_{-1}^{+k} F(dx) - kF'(k) - kF'(-k) > 0, \tag{73}$$

also (N-3) (i) is satisfied. One checks easily that the sequence T_n defined above satisfies (27), hence the corollary to theorem 3 applies.

Note that the consistency proof of section 3 would not work for this example.

REFERENCES

[1] H. E. DANIELS, "The asymptotic efficiency of a maximum likelihood estimator," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 151–163.

[2] J. L. DOOB, *Stochastic Processes*, New York, Wiley, 1953.

[3] W. M. GENTLEMAN, Ph.D. thesis, Princeton University, 1965, unpublished.

[4] P. J. HUBER, "Robust estimation of a location parameter," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 73–101.

[5] YU. V. LINNIK, "On the probability of large deviations for the sums of independent variables," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 2, pp. 289–306.

[6] A. WALD, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, Vol. 20 (1949), pp. 595–601.