THE ASYMPTOTIC EFFICIENCY OF A MAXIMUM LIKELIHOOD ESTIMATOR

H. E. DANIELS
UNIVERSITY OF BIRMINGHAM

1. Introduction and summary

The consistency of a maximum likelihood estimator has been established under very general conditions by Wald [6] and Wolfowitz [7]. Much more stringent conditions are needed for it to be asymptotically efficient, that is, consistent and asymptotically normal with variance equal to the Cramér-Rao lower bound. Typical conditions are given by Cramér [2], Gurland [3], Kulldorf [4], all of which restrict the behavior of at least the second derivative of the likelihood function. Authors such as, for example, Le Cam [5] and Bahadur [1] discuss large sample estimation in a more general context but still require regularity conditions on the second derivative of the likelihood for the maximum likelihood estimator to be asymptotically efficient.

However, cases are known which are not covered by these regularity conditions. The density function $f(x, \theta) = (1/2) \exp - |x - \theta|$ provides an example. The sample median is a maximum likelihood estimator of θ. It is known to be asymptotically normal with variance n^{-1}, which is the Cramér-Rao lower bound. But $\partial \log f/\partial\theta$ is discontinuous and $\partial^2 \log f/\partial\theta^2$ is zero for almost all x.

In the present paper weaker conditions for asymptotic efficiency are given which do not involve the second derivative of the likelihood. Two sets of sufficient conditions are stated. From the first, asymptotic efficiency can be proved directly without appeal to the Wald-Wolfowitz result but there is a convexity requirement which is frequently not satisfied. The second set of conditions dispenses with this requirement at the cost of some specialization elsewhere, but consistency has to be established by the Wald-Wolfowitz method. Finally a more general situation is considered where a modified maximum likelihood procedure is shown still to yield an asymptotically efficient estimator. The relation of this modified estimator to a class of smoothed estimators is indicated.

2. First set of sufficient conditions

We consider for simplicity a univariate distribution which has a probability density $f(x, \theta)$, where θ is a parameter which can take any value in an open interval Θ. With obvious changes the discussion will apply to discrete distributions also. Let $x_1, x_2, \cdots, x_n$ be a random sample S from such a distribution. Write $l(x, \theta) = \log f(x, \theta)$ and let $L(S, \theta) = \sum_{r=1}^{n} l(x_r, \theta)$ denote the log-likelihood of the sample S. The statistic $\hat{\theta}$ is said to be a maximum likelihood estimator of θ if $L(S, \theta) \leqq L(S, \hat{\theta})$ for all θ in Θ. It is not necessarily unique.

The asymptotic efficiency of $\hat{\theta}$ is now proved under the following conditions on $l(x, \theta)$ which are suggested by the example $f(x, \theta) = (1/2) \exp - |x - \theta|$. The symbol θ_0 refers to the true parameter value being estimated. The notation $E\{g(x)|\theta\} = \int g(x)f(x, \theta)\, dx$ is used.

CONDITIONS I.

(1) *$l(x, \theta)$ is continuous in θ throughout Θ. At every θ_0 there is a neighborhood such that for all θ, θ' in it,*

$$|l(x, \theta) - l(x, \theta')| < A(x, \theta_0)|\theta - \theta'| \tag{2.1}$$

where $E\{A^2|\theta_0\} < \infty$.

It is not difficult to show that this implies

$$\left|\frac{f(x, \theta)}{f(x, \theta')} - 1\right| < B(x, \theta_0)|\theta - \theta'|, \qquad E\{B^2|\theta_0\} < \infty. \tag{2.2}$$

(2) *At every θ, $\partial l(x, \theta)/\partial\theta$ exists and is continuous for almost all x. It is not almost everywhere zero.*

This is satisfied, as in the example quoted, when $\partial l/\partial\theta$ has a finite set of discontinuities at $\theta = \theta_j(x)$ where each $d\theta_j/dx$ exists and is not zero. On the other hand, it is not satisfied if the discontinuity points of $\partial l/\partial\theta$ are independent of x, as in the following example

$$f(x, \theta) = \begin{cases} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x - \theta)^2\right\}, & \theta \geqq 0, \\ (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x + \theta)^2\right\}, & \theta < 0. \end{cases} \tag{2.3}$$

Conventionally, $\partial l/\partial\theta$ is assumed to be continuous on the right in θ at every θ and x.

The third condition is of a more restrictive character.

(3) *$\partial l(x, \theta)/\partial\theta$ is a nowhere increasing and somewhere decreasing function of θ.*

We first observe that as a consequence of conditions I(1) and I(3), $l(x, \theta)$ is a convex function of θ, and so therefore is $L(S, \theta)$. Because of this, every $\hat{\theta}$ has the property that $\partial L/\partial\theta \geqq 0$ when $\theta \leqq \hat{\theta}$, and $\partial L/\partial\theta \leqq 0$ when $\theta \geqq \hat{\theta}$. There must be a random interval $(\hat{\theta}_L, \hat{\theta}_R)$ such that $\partial L/\partial\theta > 0$ when $\theta < \hat{\theta}_L$, while $\partial L/\partial\theta = 0$ when $\hat{\theta}_L \leqq \theta \leqq \hat{\theta}_R$, and $\partial L/\partial\theta < 0$ when $\theta > \hat{\theta}_R$. Every point of the interval is a $\hat{\theta}$ and every other point is not. Then

$$\begin{aligned} P\{\hat{\theta}_L > \theta|\theta_0\} &= P\left\{\frac{\partial L}{\partial\theta} > 0|\theta_0\right\} \\ P\{\hat{\theta}_R < \theta|\theta_0\} &= P\left\{\frac{\partial L}{\partial\theta} < 0|\theta_0\right\}. \end{aligned} \tag{2.4}$$

We next show that I(1) and I(2) are sufficient to ensure the following results holding for every θ, θ_0 in Θ

(a) $E\left\{\frac{\partial l(x,\theta_0)}{\partial\theta_0}\middle|\theta_0\right\} = 0.$

(b) $0 < E\left\{\left(\frac{\partial l(x,\theta_0)}{\partial\theta_0}\right)^2\middle|\theta_0\right\} = I(\theta_0) < \infty.$

(c) $E\left\{\frac{\partial l(x,\theta)}{\partial\theta}\middle|\theta_0\right\} = -(\theta-\theta_0)I(\theta_0) + o(\theta-\theta_0).$

(d) $E\left\{\left(\frac{\partial l(x,\theta)}{\partial\theta}\right)^2\middle|\theta_0\right\} = I(\theta_0) + o(1).$

The Lebesgue dominated convergence theorem is used; see, for example, Cramér [2], theorems (7.3.1), (7.3.2).

(a) Since $E\{B^2|\theta_0\} < \infty$ implies $E\{B|\theta_0\} < \infty$ it follows from I(1) that

$$0 = \lim_{\theta\to\theta_0}\int\frac{f(x,\theta)-f(x,\theta_0)}{\theta-\theta_0}\,dx = \int\frac{\partial f(x,\theta_0)}{\partial\theta_0}\,dx. \tag{2.5}$$

(b) $I(\theta_0)$ exists and is finite since $[f(x,\theta)-f(x,\theta_0)]^2/(\theta-\theta_0)^2f(x,\theta_0)$ is dominated by $B^2f(x,\theta_0)$. $I(\theta_0) > 0$ since $\partial l/\partial\theta_0$ is not almost everywhere zero.

(c) Under the standard regularity conditions this result is most naturally proved by differentiating $\int[\partial l(x,\theta)/\partial\theta]f(x,\theta_0)\,dx$ with respect to θ under the integral sign. But $\partial l/\partial\theta$ may now have discontinuities and the operation is not allowable. However, we may still differentiate the integral with respect to θ_0, obtaining

$$\frac{\partial}{\partial\theta_0}E\left\{\frac{\partial l}{\partial\theta}\middle|\theta_0\right\} = E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_0}\middle|\theta_0\right\}. \tag{2.6}$$

For in a neighborhood of θ_0 containing θ, θ_1, θ_2, the integrand of

$$\frac{[f(x,\theta_1)-f(x,\theta)]}{(\theta_1-\theta)f(x,\theta)}\frac{[f(x,\theta_2)-f(x,\theta_0)]}{(\theta_2-\theta_0)}\,dx \tag{2.7}$$

is dominated in modulus by $B^2f(x,\theta_0)$ and the limit as $\theta_1\to\theta$, $\theta_2\to\theta_0$ may be taken. Moreover, since $\partial l(x,\theta)/\partial\theta$ is continuous in θ for almost all x we also have with similar justification

$$\lim_{\theta\to\theta_0}E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_0}\middle|\theta_0\right\} = I(\theta_0). \tag{2.8}$$

Hence

$$E\left\{\frac{\partial l}{\partial\theta}\middle|\theta_0\right\} = E\left\{\frac{\partial l}{\partial\theta}\middle|\theta\right\} - (\theta-\theta_0)I(\theta_0) + o(\theta-\theta_0). \tag{2.9}$$

The first term on the right vanishes by (a) and (c) follows. In a similar way (d) is proved.

These results are now applied to (2.4). By (c) and (d),

$$E\left\{\frac{\partial L}{\partial\theta}\middle|\theta_0\right\} = -n(\theta-\theta_0)\{I(\theta_0)+o(1)\},$$

$$\text{(2.10)} \qquad \operatorname{Var}\left\{\frac{\partial L}{\partial \theta}\bigg|\theta_0\right\} = n\{I(\theta_0) + o(1)\}$$

and by the central limit theorem, for any fixed $w = [n\, I(\theta_0)]^{1/2}(\theta - \theta_0)$,

$$\text{(2.11)} \qquad P\left\{\frac{\partial L}{\partial \theta} > 0\bigg|\theta_0\right\} \sim \Phi(w) = \int_{-\infty}^{w} e^{-1/2t^2} \frac{dt}{\sqrt{2\pi}}$$

as n becomes large. Hence

$$\text{(2.12)} \qquad P\{\hat\theta_L > \theta|\theta_0\} \sim 1 - P\{\hat\theta_R < \theta|\theta_0\} \sim \Phi\{[n\, I(\theta_0)]^{1/2}(\theta - \theta_0)\}.$$

Since $P\{\hat\theta_L > \theta|\theta_0\} \leqq P\{\hat\theta > \theta|\theta_0\} \leqq P\{\hat\theta_R \geqq \theta|\theta_0\}$ it follows that $P\{\hat\theta > \theta|\theta_0\}$ tends to the same limit for every $\hat\theta$, and we have proved

THEOREM 1. *Under conditions* I *every maximum likelihood estimator* $\hat\theta$ *is asymptotically efficient.*

3. Second set of sufficient conditions

Conditions I are satisfied by $f(x, \theta) = (1/2) \exp - |x - \theta|$ and by similar densities such as

$$\text{(3.1)} \qquad f(x, \theta) = \begin{cases} \frac{1}{3} e^{x-\theta}, & x < \theta, \\ \frac{1}{3}, & \theta \leqq x \leqq \theta + 1, \\ \frac{1}{3} e^{\theta+1-x}, & \theta + 1 < x. \end{cases}$$

(In this case $\hat\theta$ is never unique.) But the convexity of $l(x, \theta)$ imposed by I(3) is a severe restriction and does not hold even in such regular cases as mixtures of normal densities. Moreover, convexity is not necessarily preserved under a transformation. For example, if θ in $f(x, \theta) = (1/2) \exp - |x - \theta|$ is converted to a scale parameter ϕ by the transformation $y = e^x$, $\phi = e^\theta$, the density becomes

$$\text{(3.2)} \qquad g(y|\phi) = \begin{cases} \frac{1}{2\phi}, & 0 < y < \phi, \\ \frac{\phi}{2y^2}, & 0 < \phi \leqq y, \end{cases}$$

and the log-likelihood is not convex in ϕ.

The purpose of I(3) was to enable (2.4) to hold for all θ, which ensured both consistency and asymptotic normality. But consistency was established under very general conditions by Wald [6] and if his conditions are satisfied we need only a local and possibly weaker equivalent of I(3) to hold near θ_0. Consider the following example.

$$(3.3) \qquad f(x,\theta) = \begin{cases} \dfrac{1}{(1+\theta)}, & 0 \leqq x < \theta, \\ \dfrac{1}{(1+\theta)} \exp(\theta - x), & \theta \leqq x < \infty, \end{cases}$$

where $0 < \theta < \infty$. In this case

$$(3.4) \qquad \frac{\partial l}{\partial \theta} = \frac{\theta}{(1+\theta)}, \quad 0 < \theta \leqq x, \qquad \frac{\partial l}{\partial \theta} = -\frac{1}{(1+\theta)}, \quad x < \theta < \infty,$$

and $\partial l/\partial\theta$ is an *increasing* function of θ except at the discontinuity point $\theta = x$. Nevertheless I(1) and I(2) still hold and imply (a), (b), (c), (d). Also, Wald's conditions are satisfied and $\hat{\theta}$ is consistent.

The graph of $(1/n)\ \partial L/\partial\theta$ against θ has the character of a random walk of n downward steps of $1/n$ at $\theta = x_r$ superposed on a continuous upward trend $\theta/(1+\theta)$. As n increases, the jumps become less severe and more numerous, and $(1/n)\ \partial L/\partial\theta$ tends near θ_0 to be contained within a narrow band of slope $-I(\theta_0)$. It can be shown that as n increases the width of the band decreases rapidly enough for a result similar to but weaker than (2.4) to be stated and the asymptotic efficiency of θ deduced. We use this idea to prove that θ is asymptotically efficient under the following conditions.

CONDITIONS II.

(1) *$l(x,\theta)$ is continuous in θ throughout Θ. At every θ_0 there is a neighborhood such that for all θ, θ' in it*

$$(3.5) \qquad |l(x,\theta) - l(x,\theta')| < A(x,\theta_0)|\theta - \theta'|$$

where $E\{A^3|\theta_0\} < \infty$.

This is stronger than I(1) and is introduced to impose some extra smoothness on $E\{\partial l/\partial\theta|\theta_0\}$, though a weaker condition would probably suffice.

(2) *At every θ, $\partial l(x,\theta)/\partial\theta$ exists for almost all x and is not almost everywhere zero. It is continuous in θ except at a finite number of discontinuity points at which it has finite jumps of either sign.*

(3) *The probability that the interval (θ,θ') contains a discontinuity point of $\partial l/\partial\theta$ is $O(\theta' - \theta)$ for any true value θ_0.*

Conventionally, $\partial l/\partial\theta$ is again assumed continuous on the right. Thus $\partial l/\partial\theta = c(x,\theta) + h(x,\theta)$ where $c(x,\theta)$ is continuous at every θ and $h(x,\theta)$ is a step function. We require a further condition on $c(x,\theta)$.

(4) $|c(x,\theta') - c(x,\theta)| < G(x,\theta_0)|\theta' - \theta|, \qquad E\{G^2|\theta_0\} < \infty.$

Since II(2) and II(3) together imply I(2), II(1) to II(3) are sufficient for (a), (b), (c), (d) to hold. Later, (c) will be replaced by a stronger result (c′).

We first appeal to the known consistency of $\hat{\theta}$. Wald proved under mild conditions that every $\hat{\theta}$ maximizing $L(S,\theta)$ almost certainly lies in a preassigned interval $(\theta_0 - \delta, \theta_0 + \delta)$ as $n \to \infty$. His conditions are more than covered by I or II, provided the end endpoints of Θ are taken care of by an additional condition to ensure compactness, since Wald assumes Θ to be a closed interval. Wald's proof depends on the inequality

$$E\{l(x, \theta) - l(x, \theta_0)|\theta_0\} < 0, \qquad \theta \neq \theta_0. \tag{3.6}$$

The strong law of large numbers then makes $L(S, \theta) < L(S, \theta_0)$ almost certainly true as $n \to \infty$ for every $\theta \neq \theta_0$. With the reasonable extra assumption that (3.6) still holds in the limit as θ tends to the possibly infinite endpoints of Θ it can be deduced that

$$\sup_{|\theta - \theta_0| > \delta} L(S, \theta) < L(S, \theta_0) \tag{3.7}$$

with probability 1 as $n \to \infty$. Thus every $\hat{\theta}$ must ultimately lie in $(\theta_0 - \delta, \theta_0 + \delta)$ and so $\hat{\theta}$ is strongly consistent.

Weak consistency, which is more relevant to asymptotic efficiency, means that $P\{\theta_0 - \delta < \hat{\theta} < \theta_0 + \delta|\theta_0\} \to 1$ for every $\hat{\theta}$ as $n \to \infty$. This was established by Wolfowitz [7] in a similar way, using instead the weak law of large numbers. Conditions I or II allow the result to be strengthened somewhat. It is permissible by I(1) to integrate (c) with respect to θ and obtain

$$E\{l(x, \theta) - l(x, \theta_0)|\theta_0\} = -\frac{1}{2}(\theta - \theta_0)^2\{I(\theta_0) + o(1)\}. \tag{3.8}$$

Also from (b),

$$E\{[l(x, \theta) - l(x, \theta_0)]^2|\theta_0\} = (\theta - \theta_0)^2\{I(\theta_0) + o(1)\}. \tag{3.9}$$

The central limit theorem then gives

$$P\{L(S, \theta) - L(S, \theta_0) < 0|\theta_0\} \sim \Phi\left\{\frac{1}{2}|\theta - \theta_0|[n\, I(\theta_0)]^{1/2}\right\}, \tag{3.10}$$

so that for each θ such that $\alpha n^{-1/2+\epsilon} < |\theta - \theta_0| < \delta$, with $\alpha > 0$, $\epsilon > 0$, we have $P\{L(S, \theta) - L(S, \theta_0) < 0|\theta_0\} \to 1$ as $n \to \infty$. It can then be deduced as before that for any preassigned $\alpha > 0$, $\epsilon > 0$, and for every $\hat{\theta}$,

$$P\{\theta_0 - \alpha n^{-1/2+\epsilon} < \hat{\theta} < \theta_0 + \alpha_n{}^{-1/2+\epsilon}|\theta_0\} \to 1 \tag{3.11}$$

as $n \to \infty$. In the subsequent discussion we may therefore confine our attention to values of θ in the interval $(\theta_0 - \alpha n^{-1/2+\epsilon}, \theta_0 + \alpha n^{-1/2+\epsilon})$ which we denote by $\mathscr{I}_n$.

4. Approximate local monotonicity of $\partial L/\partial\theta$ near θ_0

The next objective is to establish that as θ varies over $\mathscr{I}_n$, then $\partial L/\partial\theta$ tends to lie within a band of slope $-n\, I(\theta_0)$ which is narrow compared with the standard deviation of $\partial L/\partial\theta$ as n becomes large. The strongest result of this kind could evidently be deduced from random walk theory under suitable regularity conditions, since the probability of a jump in $\partial l/\partial\theta$ is approximately uniformly distributed over $\mathscr{I}_n$. However, we adopt a more elementary approach leading to a weaker result which is adequate for our purpose. The method has the advantage that it carries over to a more general situation discussed later.

Let $\{\theta_m\}$ be a discrete set of values of θ dividing $\mathcal{I}_n$ into N equal subintervals $\mathcal{I}_{n,m}$ of length $\theta_{m+1} - \theta_m = 2\alpha n^{-1/2+\epsilon}/N$, where N is taken to be a function of n such that the subdivision of $\mathcal{I}_n$ becomes increasingly fine as n becomes large. It will be shown that if we choose $N \sim n^{1/8+\epsilon}$ and $\epsilon < 1/16$ all differences $\partial L/\partial\theta_m - \partial L/\partial\theta_p$ tend simultaneously to lie within $O(n^{3/8})$ of their expectations. Fluctuations of this order are vanishingly small compared with $[n\, I(\theta_0)]^{1/2}$, which is the approximate standard deviation of each $\partial L/\partial\theta_m$. By disposing of "end effects" the same result is shown to hold for all $\partial L/\partial\theta - \partial L/\partial\theta'$ where θ, θ' range continuously over $\mathcal{I}_n$.

Consider the behavior of $\partial l/\partial\theta$ over the subinterval $\mathcal{I}_{n,m}$. If n is sufficiently large, $\mathcal{I}_{n,m}$ is small enough to contain at most one discontinuity point of $\partial l/\partial\theta$ since there is only a finite number of them. If $\mathcal{I}_{n,m}$ contains a discontinuity, $|\partial l/\partial\theta_{m+1} - \partial l/\partial\theta_m| < K < \infty$ by II(2), and this will occur with probability less than $M(\theta_{m+1} - \theta_m)$, where $M < \infty$, by II(3). If there is no discontinuity

$$(4.1) \qquad \left|\frac{\partial l}{\partial\theta_{m+1}} - \frac{\partial l}{\partial\theta_m}\right| < (\theta_{m+1} - \theta_m)G(x, \theta_0)$$

by II(4). So we have

$$(4.2) \qquad E\left\{\left(\frac{\partial l}{\partial\theta_{m+1}} - \frac{\partial l}{\partial\theta_m}\right)^2 \middle| \theta_0\right\} < (\theta_{m+1} - \theta_m)K^2M + (\theta_{m+1} - \theta_m)^2E\{G^2|\theta_0\}$$
$$= O(\theta_{m+1} - \theta_m)$$

and hence

$$(4.3) \qquad \operatorname{Var}\left\{\frac{\partial L}{\partial\theta_{m+1}} - \frac{\partial L}{\partial\theta_m}\middle|\theta_0\right\} = O[n(\theta_{m+1} - \theta_m)] < R\frac{n^{1/2+\epsilon}}{N}$$

for some $R < \infty$ which may be taken the same for all $\mathcal{I}_{n,m}$.

Let $N \sim n^{1/8+\epsilon}$, in which case $n(\theta_{m+1} - \theta_m) = O(n^{3/8})$. By the central limit theorem, for $\lambda > 0$ and large n_0,

$$(4.4) \qquad P\left\{\left|\frac{\partial L}{\partial\theta_{m+1}} - \frac{\partial L}{\partial\theta_m} - E\left[\frac{\partial L}{\partial\theta_{m+1}} - \frac{\partial L}{\partial\theta_m}\middle|\theta_0\right]\right| < \frac{\lambda n^{3/8}}{N}\right\}$$
$$> 2\Phi(\lambda R^{-1/2}n_0^{1/16-\epsilon}) - 1$$

for all $n > n_0$. We require that the inequalities on the left shall hold simultaneously for all m, that is, that

$$(4.5) \qquad \max_m \left|\frac{\partial L}{\partial\theta_{m+1}} - \frac{\partial L}{\partial\theta_m} - E\left\{\frac{\partial L}{\partial\theta_{m+1}} - \frac{\partial L}{\partial\theta_m}\middle|\theta_0\right\}\right| < \frac{\lambda n^{3/8}}{N},$$

since this implies the result we want, namely

$$(4.6) \qquad \max_{m,\,p} \left|\frac{\partial L}{\partial\theta_m} - \frac{\partial L}{\partial\theta_p} - E\left\{\frac{\partial L}{\partial\theta_m} - \frac{\partial L}{\partial\theta_p}\middle|\theta_0\right\}\right| < \lambda n^{3/8}.$$

By Boole's inequality the probability of (4.5), and hence that of (4.6), exceeds $1 - 2N(1 - \Phi)$ with Φ as in (4.4). Assuming $\epsilon < 1/16$ this probability tends to 1 as $n \to \infty$.

We now consider the "end effects" which come in when θ is allowed to vary continuously over $\mathcal{I}_n$. With the same K as before, define a random variable $z_m(x)$ to take the value K if $\mathcal{I}_{n,m}$ contains a discontinuity of $\partial l/\partial\theta$, and the value $(\theta_{m+1} - \theta_m)G(x, \theta_0)$ if it does not. Then

$$\max_{\theta\epsilon\mathcal{I}_{n,m}} \left|\frac{\partial l}{\partial\theta} - \frac{\partial l}{\partial\theta_m}\right| < z_m(x), \tag{4.7}$$

and hence if $Z_m(S) = \sum_{r=1}^{n} z_m(x_r)$,

$$\max_{\theta\epsilon\mathcal{I}_{n,m}} \left|\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta_m}\right| < Z_m(S). \tag{4.8}$$

From II(3) and II(4), both $E\{Z_m|\theta_0\}$ and Var $\{Z_m|\theta_0\}$ are

$$O\{n(\theta_{m+1} - \theta_m)\} = O(n^{3/8}).$$

It follows easily from the central limit theorem and Boole's inequality that if $\mu > 0$, then $|\partial L/\partial\theta - \partial L/\partial\theta_m| < \mu n^{3/8}$ holds, in probability, uniformly for all θ in $\mathcal{I}_{n,m}$ and all θ_m as $n \rightarrow \infty$. So the extra terms introduced by allowing θ to vary continuously can be absorbed by adjusting λ, and we can assert that for some $\lambda > 0$,

$$P\left\{\max_{\theta,\theta'\epsilon\mathcal{I}_n}\left|\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta'} - E\left\{\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta'}\middle|\theta_0\right\}\right| < \lambda n^{3/8}\theta_0\right\} \rightarrow 1. \tag{4.9}$$

Finally we examine the behavior of $E\{\partial L/\partial\theta - \partial L/\partial\theta'|\theta_0\}$ over $\mathcal{I}_n$ using a stronger form of (c). From (2.6) we have

$$E\left\{\frac{\partial l}{\partial\theta}\middle|\theta_0\right\} = -(\theta - \theta_0)E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\middle|\theta_1\right\} \tag{4.10}$$

for some θ_1 between θ_0 and θ, since $E\{(\partial l/\partial\theta)(\partial l/\partial\theta_0)|\theta_0\}$ is continuous in θ_0 by the previous argument. By II(1) the procedure may be repeated to give

$$E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\middle|\theta_1\right\} = E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\middle|\theta_0\right\} + (\theta_1 - \theta_0)E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\frac{\partial l}{\partial\theta_0}\middle|\theta_0\right\} + o(\theta_1 - \theta_0) \tag{4.11}$$

so that we now have

$$E\left\{\frac{\partial l}{\partial\theta}\middle|\theta_0\right\} = -(\theta - \theta_0)E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\middle|\theta_0\right\} + O\{(\theta - \theta_0)^2\}. \tag{4.12}$$

As before, II(3) and II(4) imply that $E\{[(\partial l/\partial\theta) - (\partial l/\partial\theta_0)]^2|\theta_0\} = O\{|\theta - \theta_0|\}$ and it follows on applying the Schwarz inequality that

$$E\left\{\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta_1}\middle|\theta_0\right\} = I(\theta_0) + O\{|\theta - \theta_0|^{1/2}\}. \tag{4.13}$$

Hence we can replace (c) by

(c′) $E\left\{\frac{\partial l}{\partial\theta}\middle|\theta_0\right\} = -(\theta - \theta_0)I(\theta_0) + O\{|\theta - \theta_0|^{3/2}\}.$

For all θ in $\mathcal{I}_n$ we have $|\theta - \theta_0| = O(n^{-1/2+\epsilon})$, so that

$$E\left\{\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta'}\middle|\theta_0\right\} = -n(\theta - \theta')I(\theta_0) + O(n^{1/4+3\epsilon/2}). \tag{4.14}$$

Since $\epsilon < 1/16$ this is more than enough to enable us to say that for some $\lambda > 0$,

$$P\left\{\max_{\theta,\theta'}\left|\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta'} + n(\theta - \theta')I(\theta_0)\right| < \lambda n^{3/8}\middle|\theta_0\right\} \to 1 \tag{4.15}$$

as $n \to \infty$, which is the required result.

5. Completion of proof of theorem 2

We are now in a position to apply an argument similar to that used for theorem 1. Let $\hat\theta_L$, $\hat\theta_R$ be the least and greatest maximum likelihood estimators. Notice that it is now not necessarily true that *every* point in the interval $(\hat\theta_L, \hat\theta_R)$ is a $\hat\theta$. The inequality

$$\left|\frac{\partial L}{\partial\theta} - \frac{\partial L}{\partial\theta'} + n(\theta - \theta')I(\theta_0)\right| < \lambda n^{3/8} \tag{5.1}$$

may be written

$$\frac{\partial L}{\partial\theta} - \lambda n^{3/8} < \frac{\partial L}{\partial\theta'} + n(\theta' - \theta)I(\theta_0) < \frac{\partial L}{\partial\theta} + \lambda n^{3/8}. \tag{5.2}$$

Suppose (5.2) to be satisfied for every θ, θ' in $\mathcal{I}_n$. Then $\partial L/\partial\theta < -\lambda n^{3/8}$ implies $\partial L/\partial\theta' < 0$ for every $\theta' > \theta$, and since $L(S, \theta')$ is continuous in θ' this in turn implies $\hat\theta_R < \theta$. Similarly $\partial L/\partial\theta > \lambda n^{3/8}$ implies $\partial L/\partial\theta' > 0$ for every $\theta' < \theta$ and this implies $\hat\theta_L > \theta$. Also $\hat\theta_R < \theta$ implies $\hat\theta_L \leqq \theta$. Hence, conditional on (5.2) which holds in probability for all θ, θ' in $\mathcal{I}_n$,

$$\begin{aligned} P\left\{\frac{\partial L}{\partial\theta} < -\lambda n^{3/8}\middle|\theta_0\right\} &\leqq P\{\hat\theta_R < \theta|\theta_0\} \\ &\leqq P\{\hat\theta_L \leqq \theta|\theta_0\} \leqq P\left\{\frac{\partial L}{\partial\theta} \leqq \lambda n^{3/8}\middle|\theta_0\right\}. \end{aligned} \tag{5.3}$$

Both $P\{\partial L/\partial\theta < -\lambda n^{3/8}|\theta_0\}$ and $P\{\partial L/\partial\theta \leqq \lambda n^{3/8}|\theta_0\}$ tend to

$$\Phi\{(\theta - \theta_0)[n\,I(\theta_0)]^{1/2}\} \tag{5.4}$$

as $n \to \infty$ for fixed $n^{1/2}(\theta - \theta_0)$. So therefore does $P\{\hat\theta \leqq \theta|\theta_0\}$ for every $\hat\theta$ and we have proved

THEOREM 2. *Under conditions* II *every maximum likelihood estimator is asymptotically efficient.*

6. A more general situation

Though the class of densities for which $\hat{\theta}$ is asymptotically efficient has been considerably widened there remain cases where the Cramér-Rao lower bound is neither zero nor infinite, but which are not covered by conditions I or II. Consider the example

$$(6.1)\qquad f(x, \theta) = \frac{1}{2\Gamma\left(1 + \frac{1}{\kappa}\right)} \exp\{-|x - \theta|^{\kappa}\}, \qquad \frac{1}{2} < \kappa < 1,$$

where $-\infty < x < \infty$, $-\infty < \theta < \infty$. In this case $\partial l/\partial\theta = \kappa(x, \theta)^{-(1-\kappa)}$ for $\theta < x$; $\partial l/\partial\theta = -(\theta - x)^{-(1-\kappa)}$ for $x < \theta$. Not only is $\partial l/\partial\theta$ an increasing function of θ where it is continuous, but it has an infinite discontinuity at $\theta = x$. If the ordered observations are $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, then $\partial L/\partial\theta$ increases from $-\infty$ to ∞ as θ goes from each $x_{(r)}$ to $x_{(r+1)}$. It has n infinite discontinuities separated by intervals whose average width is of order n^{-1}.

Nevertheless $I(\theta_0) = \kappa\Gamma(2 - 1/\kappa)/\Gamma(1 + 1/\kappa)$ is finite if $\kappa > 1/2$. Notice that the asymptotic efficiency of the median is $\sin \pi(1/\kappa - 1)/\pi(1/\kappa - 1)$, which decreases from 1 to 0 as κ decreases from 1 to 1/2. Also

$$(6.2)\qquad E\left\{\frac{\partial l}{\partial\theta}\bigg|\theta_0\right\} = -(\theta - \theta_0)I(\theta_0) + O|\theta - \theta_0|^{2\kappa},$$

$$(6.3)\qquad E\left\{\left(\frac{\partial l}{\partial\theta} - \frac{\partial l}{\partial\theta'}\right)^2\bigg|\theta_0\right\} = O\{|\theta - \theta'|^{2\kappa-1}\}.$$

So in spite of its irregular behavior, $\partial l/\partial\theta$ has quite reasonable average properties, and is actually more than continuous in mean square. We now show that in cases of this type an asymptotically efficient estimator of θ_0 can be found by maximizing $L(S, \theta)$ over a discrete set of values of θ separated by intervals which decrease faster than $n^{-1/2}$, but not too fast, as $n \to \infty$.

In the example, $l(x, \theta)$ is continuous in θ and it may also be verified that Wald's conditions are satisfied. We shall assume this to be true throughout the rest of the discussion. To avoid unnecessary complications the following conditions are stated in terms of differences rather than derivatives. They are satisfied by (6.1) with $\rho = 2\kappa - 1$. We need only consider $0 < \rho < 1$.

$$(6.4)\qquad E\{[l(x, \theta_0 + \omega) - l(x, \theta_0)]^2|\theta_0\} = \omega^2 I(\theta_0) + o(\omega^2),$$

$$(6.5)\qquad E\{l(x, \theta_0 + \omega) - l(x, \theta_0)|\theta_0\} = -\frac{1}{2}\omega^2 I(\theta_0) + O(\omega^{2+\rho}),$$

$$(6.6)\qquad E\{[l(x, \theta + 2\omega) - 2l(x, \theta + \omega) + l(x, \theta)]^2|\theta_0\} = O(\omega^{2+\rho}).$$

The argument used to prove theorem 2 is pursued as far as possible. Consider a mesh of equally spaced values $\{\theta_m\}$ ranging over the entire interval Θ with $\theta_{m+1} - \theta_m = \gamma n^{-1/2-\rho/8}$, where $\gamma > 0$. The true value θ_0 is not necessarily on the mesh. Since $E\{l(x, \theta_m) - l(x, \theta_0)|\theta_0\} < 0$ for all $\theta_m \neq \theta_0$, it can be shown as before that

$$\max_{|\theta_m - \theta_0| > \alpha n^{-1/2+\epsilon}} L(S, \theta_m) < L(S, \theta_0) \tag{6.7}$$

with probability tending to 1. Let θ_{m_0} be the value of θ_m nearest to θ_0. Then (6.7) is also true in probability if $L(S, \theta_0)$ is replaced by $L(S, \theta_{m_0})$, since $l(x, \theta)$ is continuous and $E\{l(x, \theta) - l(x, \theta_0)|\theta_0\}$ cannot approach arbitrarily close to zero except near θ_0. Hence if $\tilde{\theta}$ is a value of θ_m such that $L(S, \theta_m) \leqq L(S, \tilde{\theta})$ for all θ_m,

$$P\{\theta_0 - \alpha n^{-1/2+\epsilon} < \tilde{\theta} < \theta_0 + \alpha n^{-1/2+\epsilon}|\theta_0\} \to 1 \tag{6.8}$$

and again it is only necessary to consider values of θ_m in $\mathcal{G}_n$.

Let $N \sim n^{\rho/8+\epsilon}$ and $\gamma = 2\alpha$, so that $\theta_{m+1} - \theta_m = 2\alpha n^{-1/2+\epsilon}/N$ as before. The discussion now proceeds as for theorem 2 but with differences replacing derivatives. Write

$$\begin{aligned} d_m(x) &= \frac{l(x, \theta_{m+1}) - l(x, \theta_m)}{\theta_{m+1} - \theta_m}, \\ D_m(S) &= \sum_{r=1}^{n} d_m(x_r). \end{aligned} \tag{6.9}$$

From (6.6) we have $E\{(d_{m+1} - d_m)^2|\theta_0\} = O\{(\theta_{m+1} - \theta_m)^\rho\}$ and hence

$$\operatorname{Var} \{D_{m+1} - D_m|\theta_0\} = O\{n(\theta_{m+1} - \theta_m)^\rho\} = O(n^{1-\rho/2-\rho^2/8}). \tag{6.10}$$

So however large n_0 we have for some $C > 0$,

$$\begin{aligned} P\left\{|D_{m+1} - D_m - E\{D_{m+1} - D_m|\theta_0\}| < \frac{\lambda n^{1/2-\rho/8}}{N}\bigg|\theta_0\right\} \\ > 2\Phi(C n_0^{\rho^2/16-\epsilon}) - 1 \end{aligned} \tag{6.11}$$

for all $n > n_0$. If $\epsilon < \rho^2/16$ it follows as before that for all θ_m, θ_p in $\mathcal{G}_n$,

$$\max_{m,p} |D_m - D_p - E\{D_m - D_p|\theta_0\}| < \lambda n^{1/2-\rho/8} \tag{6.12}$$

in probability, as $n \to \infty$. Also from (6.5),

$$\begin{aligned} E\{l(x, \theta_{m+1}) - l(x, \theta_m)|\theta_0\} \\ = -(\theta_{m+1} - \theta_m)(\theta_m - \theta_0)I(\theta_0) + O\{(\theta_m - \theta_0)^{2+\rho}\} \end{aligned} \tag{6.13}$$

and it may be deduced that

$$E\{D_m - D_p|\theta_0\} = -n(\theta_m - \theta_p)I(\theta_0) + O\{n^{1/2-3\rho/8+(2+\rho)\epsilon}\}. \tag{6.14}$$

The remainder is less than $O(n^{1/2-\rho/8})$ and we can therefore state that for some $\lambda > 0$,

$$\max |D_m - D_p + n(\theta_m - \theta_p)I(\theta_0)| < \lambda n^{1/2-\rho/8} \tag{6.15}$$

with probability tending to 1.

The argument used for $\hat{\theta}$ can now be applied in the same way to the discrete set $\{\theta_m\}$ to prove that if $\tilde{\theta}_L$ and $\tilde{\theta}_R$ are the least and greatest values of $\tilde{\theta}$,

$$\begin{aligned} P\{D_m < -\lambda n^{1/2-\rho/8}|\theta_0\} &\leqq P\{\tilde{\theta}_R < \theta|\theta_0\} \\ &\leqq P\{\tilde{\theta}_L < \theta_0\} \leqq P\{D_m \leqq \lambda n^{1/2-\rho/8}|\theta_0\} \end{aligned} \tag{6.16}$$

and hence that

$$P\{\tilde{\theta} \leqq \theta | \theta_0\} \sim \Phi\{(\theta - \theta_0)[n\, I(\theta_0)]^{1/2}\} \tag{6.17}$$

for every $\tilde{\theta}$. Thus every $\tilde{\theta}$ is asymptotically efficient.

7. Smoothed maximum likelihood estimators

It is clearly not possible to complete the argument of theorem 2 in the more general situation, and the asymptotic efficiency of $\hat{\theta}$ itself remains an open question. The choice of $N \sim n^{\rho/8+\epsilon}$ was to some extent arbitrary, but N must certainly increase more slowly than $n^{1/2}$ for the proof to go through. Roughly speaking, the reason why $\tilde{\theta}$ is easier to handle than $\hat{\theta}$ is that by maximizing over the mesh instead of over Θ we reduce the chance of selecting one of the erratic cusps of the likelihood function. In fact $\tilde{\theta}$ is related to a class of smoothed maximum likelihood estimators defined in the following way. Let

$$\overline{L}_n(S, \theta) = \int L(S, \theta - u) g_n(u)\, du \tag{7.1}$$

where $g_n(u)$ is a normalized weight function such that $\int u^2 g_n(u)\, du \to 0$ as $n \to \infty$. We call $\check{\theta}$ a smoothed maximum likelihood estimator if $\overline{L}_n(S, \theta) \leqq \overline{L}_n(S, \check{\theta})$ for all θ. For the uniform weight function $g_n(u) = 1/2a_n$, $|u| \leqq a_n$, $g_n(u) = 0$, $|u| > a_n$, we have

$$\overline{L}_n(S, \theta) = \frac{1}{2a_n} \int_{\theta - a_n}^{\theta + a_n} L(S, \phi)\, d\phi \tag{7.2}$$

and

$$\frac{\partial \overline{L}_n}{\partial \theta} = \frac{1}{2a_n} \{L(S, \theta + a_n) - L(S, \theta + a_n)\}. \tag{7.3}$$

The estimator $\tilde{\theta}$ is evidently closely related to $\check{\theta}$ for a uniform weight function with $a_n = (1/2)\gamma n^{-1/2-\rho/8}$. The averaging of L over the interval will cause $\overline{L}_n$ to vary smoothly as θ goes from θ_m to θ_{m+1} so that the maximizations leading to $\tilde{\theta}$ and $\check{\theta}$ should give nearly the same result. Estimators of this type with a general weight function seem worthy of further study.

I have profited greatly from discussions with D. V. Lindley on the subject of the paper.

REFERENCES

[1] R. R. Bahadur, "On the asymptotic efficiency of tests and estimates," to appear in *Sankhyā*.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton, Princeton University Press, 1946.

[3] J. Gurland, "On regularity conditions for maximum likelihood estimators," *Skand. Aktuarietidskr.*, Vol. 37 (1954), pp. 71–76.

[4] G. KULLDORF, "On the conditions for consistency and asymptotic efficiency of maximum likelihood estimators," *Skand. Aktuarietidskr.*, Vol. 40 (1957), pp. 130–144.

[5] L. LE CAM, "On some asymptotic theory of estimation and testing hypotheses," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1955, Vol. 1, pp. 129–156.

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

[7] J. WOLFOWITZ, "On Wald's proof of the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, Vol. 20 (1949), pp. 601–602.