The Annals of Statistics
1978, Vol. 6, No. 4, 727–739

JACKKNIFING MAXIMUM LIKELIHOOD ESTIMATES[1]

BY JAMES A. REEDS

University of California, Berkeley

This paper proves the apparently outstanding conjecture that the maximum likelihood estimate (m.l.e.) "behaves properly" when jackknifed. In particular, under the usual Cramér conditions (1) the jackknifed version of the consistent root of the m.l. equation has the same asymptotic distribution as the consistent root itself, and (2) the jackknife estimate of the variance of the asymptotic distribution of the consistent root is itself consistent. Further, if the hypotheses of Wald's consistency theorem for the m.l.e. are satisfied, then the above claims hold for the m.l.e. (as well as for the consistent root).

1. Introduction. Consider a statistical estimation procedure which, given the sample data $X_1, X_2, \cdots, X_n$, yields as the estimate of some parameter that value of θ which maximizes $\bar{h}_n(\theta) \equiv 1/n \sum_1^n h(X_i, \theta)$, where h is a given fixed function. This framework certainly includes maximum likelihood estimation for a given parametric family of probability distributions, even if the true distribution of the X_i does not lie in the given family. Let us denote the maximizing value of θ—neglecting for the moment questions of existence and uniqueness—by $\theta_n^{\max} = \theta_n^{\max}(X_1, \cdots, X_n)$ and let us denote a root of the critical point equation $(\partial/\partial\theta)\,\bar{h}_n(\theta) = 0$ by $\theta_n^{\text{root}} = \theta_n^{\text{root}}(X_1, \cdots, X_n)$. (We will specify *which* root later on.) Given such an estimate we may "jackknife" it, that is, use θ_n to define a new estimate JK θ_n (where θ_n is either of $\theta_n^{\max}$ or θ_n^{root}) defined by

$$\begin{aligned}\text{JK }\theta_n(X_1, \cdots, X_n) &= \theta_n - \frac{n-1}{n} \sum_{j=1}^n (\theta_{n-1}(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n) - \theta_n(X_1, \cdots, X_n)) \\ &= \theta_n - \frac{n-1}{n} \sum_{j=1}^n (\theta_{n,-j} - \theta_n) \\ &= \theta_n - \frac{n-1}{n} \sum_{j=1}^n R_{nj}\,.\end{aligned}$$

We may also calculate a "jackknife estimate of variance"

$$\text{JKV }\theta_n(X_1, \cdots, X_n) = (n-1) \sum_{j=1}^n (R_{nj} - \bar{R}_n)(R_{nj} - \bar{R}_n)'\,,$$

where $\bar{R}_n = (1/n) \sum_{j=1}^n R_{nj}$ and $'$ denotes transpose.

Under suitable regularity conditions on the function h and on the distribution

Received March 1977; revised October 1977.

[1] This paper was prepared with support of U.S. Office of Naval Research Contract Number N00014-75-C-0444/NR 042-036.

AMS 1970 *subject classifications.* Primary 62F10; Secondary 62E20, 62F25.

Key words and phrases. Jackknife, maximum likelihood estimate, *M*-estimate, asymptotic normality, Banach space law of large numbers, reversion of series, Cramér conditions.

of the X_i we can expect that $\lim_{n\to\infty} P(\theta_n^{\text{root}} = \theta_n^{\max}) = 1$ and that $n^{\frac{1}{2}}(\theta_n - \theta^*) \to_{\mathscr{L}} N(0, \Sigma)$ for some value of θ^* and some matrix Σ, where θ_n is either of θ_n^{root} or $\theta_n^{\max}$. By analogy with known jackknife results we should also expect such convergences as

$$n^{\frac{1}{2}}(\text{JK}\,\theta_n - \theta_n) \to_{\text{a.s.}} 0 \tag{J.1}$$

$$\text{JKV}\,\theta_n \to_P \Sigma \tag{J.2}$$

$$\text{JKV}\,\theta_n \to_{\text{a.s.}} \Sigma \tag{J.3}$$

$$n^{\frac{1}{2}}(\text{JK}\,\theta_n - \theta^*) \to_{\mathscr{L}} N(0, \Sigma) \tag{J.4}$$

$$n(\text{JK}\,\theta_n - \theta^*)'(\text{JKV}\,\theta_n)^{-1}(\text{JK}\,\theta_n - \theta^*) \to_{\mathscr{L}} \chi^2\,. \tag{J.5}$$

This paper proves (J.1) and (J.2) under conditions on the function h that (in the maximum likelihood case) are slightly weaker than Cramér's conditions for the asymptotic normality of the consistent root of the likelihood equation (Cramér, (1946), page 500). We are unable to prove (J.3) without a strong moment condition. (J.4) follows from (J.1) and the known asymptotic properties of θ_n; (J.5) follows from (J.2) and (J.4). Our proof of (J.1) and (J.2) seems to be new (Miller, (1974), page 5).

Such convergences are of practical use: a confidence interval based on (J.2) or a test based on (J.5), for instance, has a degree of robustness under model-violations not shared by intervals and tests based on, say, the estimated Fisher information. This—and other features ot the jackknife—are surveyed in Miller (1974).

The present results should not be confused with those of Brillinger (1964), which concern a different kind of jackknife. Brillinger defines his jackknife by dividing the observations into a fixed number of groups, and successively omitting the groups from the formula for the estimate. As the sample size increases without bound, the number of groups remains fixed, but the number of observations in each group tends to infinity. This is contrasted with our jackknife, in which the number of groups tends to infinity and the number of observations per group remains fixed (namely, one per group).

A later paper of Brillinger does concern the present version of the jackknife (Brillinger, (1977)). Its results are very close to ours; Brillinger uses an asymptotic expansion of Chibisov (1973) to prove convergences (J.1) and (J.2). But the appeal to Chibisov's expansion requires much stronger moment and smoothness conditions than the usual Cramér conditions: in the maximum likelihood case, Brillinger requires, *inter alia*, the existence of eighth moments of the fourth derivative of the log likelihood function at the true parameter point. We are able to prove (J.1) and (J.2) under our much weaker hypotheses because we use a reversion-of-series result (Lemma 2) specially fitted to the jackknife application instead of a general purpose result like Chibisov's.

Like all known jackknife results, ours relies heavily on Taylor expansions and on a "delta method" argument—but with a difference. Where previous jackknife authors use Taylor approximations of functions of several real variables and apply these to sample means of real random variables or of finite dimensional random vectors, we apply a Taylor approximation of a function of a Banach space vector to a sample average of a random function. In other words, we carry out our delta method argument in an infinite dimensional setting. The idea is that θ_n depends differentiably upon the entire sample function $\bar{h}_n(\theta)$, which is viewed as a random element in a certain Banach space.

2. Notations, assumptions, results. Let $X_1, X_2, \cdots$ be i.i.d. random elements in some measurable space $(\mathscr{X}, \mathscr{A})$. Let $\Theta \subseteqq \mathbb{R}^q$ be a set. Suppose the function $h: \mathscr{X} \times \Theta \rightarrow \mathbb{R}$ is measurable in x for each θ. Suppose $\eta(\theta) = Eh(X_1, \theta) < \infty$ for each $\theta \in \Theta$ (the case $\eta(\theta) = -\infty$ not excluded) and suppose there exists a $\theta^* \in \Theta$ with $\eta(\theta^*) \neq -\infty$ such that if $\theta \in \Theta$ is unequal to θ^*, $\eta(\theta) < \eta(\theta^*)$. Let $\bar{h}_n: \Theta \rightarrow \mathbb{R}$ be defined by $\bar{h}_n(\theta) = (1/n)(h(X_1, \theta) + \cdots + h(X_n, \theta))$.

(We interpret Θ as a parameter space and θ^* as the "true value of θ" which is unknown to the statistician. Whatever its value, however, it is known that $\eta(\theta)$ is maximized at θ^*; since $\bar{h}_n(\theta)$ is in some sense an estimate of $\eta(\theta)$, we propose to estimate θ^* by maximizing $\bar{h}_n(\theta)$. In maximum likelihood estimation with a parametric family of probability distributions P_θ with $dP_\theta = f(x, \theta)\, d\mu(x)$ indexed by $\theta \in \Theta$, where the data $X_1, X_2, \cdots$ has distribution P_{θ^*}, the function $h(x, \theta)$ can be taken to be $\ln f(x, \theta)/f(x, \theta^*)$. Then, maximizing the likelihood is equivalent to maximizing $\bar{h}_n$; the requirement $\eta(\theta) \leqq \eta(\theta^*)$ is true by Jensen's inequality; and the requirement $\eta(\theta) \neq \eta(\theta^*)$ for $\theta \neq \theta^*$ is implied by the requirement that the family of measures P_θ is not over-parameterized—i.e., that the parameter θ is "identifiable.")

In addition to these notations and innocuous assumptions, we impose a set of stronger restrictions on the distribution of the X_i and on the behavior of the function h in the vicinity of θ^*. We refer to these extra assumptions as "L" because of their local nature.

ASSUMPTIONS L. There exists a compact neighborhood K—which can, without loss of generality, be taken to be cubical—$K \subseteqq \Theta$, such that:

(L.1) with probability 1, $h(X_1, \theta)$ is twice continuously differentiable (as a function in θ) on K;

(L.2) $E|(\partial/\partial\theta_i)\, h(X_1, \theta^*)|^2 < \infty$, $i = 1, \cdots, q$;

(L.3) $E|(\partial^2/\partial\theta_i\, \partial\theta_j)\, h(X_1, \theta^*)| < \infty$, $i, j = 1, \cdots, q$;

(L.4) $A = (a_{ij})_{i,j=1}^q$ is nonsingular, where $a_{ij} = E(\partial^2/\partial\theta_i\, \partial\theta_j)\, h(X_1, \theta^*)$; and

(L.5) there exists a function $M(x)$ and a constant $\lambda > 0$ such that $EM(X_1) < \infty$ and, with probability 1, for all s and $t \in K$,

$$\left|\frac{\partial^2}{\partial\theta_i\, \partial\theta_j} h(X_1, s) - \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} h(X_1, t)\right| \leqq M(X_1)|s - t|^\lambda .$$

When "L" is assumed we will, with further comment, adopt the notations $g(x, \theta) = (g_i(x, \theta))_{i=1}^q$, where $g_i(x, \theta) = (\partial/\partial\theta_i)h(x, \theta)$, and $\bar{g}_n(\theta) = (1/n) \sum_{j=1}^n g(X_j, \theta)$, and $S = (s_{ij})_{i,j=1}^q$ where $s_{ij} = Eg_i(X_1, \theta^*)g_j(X_1, \theta^*)$.

Note that condition (L.5) allows us to interchange differentiation and expectation, as in $Eg_i(X_1, \theta) = (\partial/\partial\theta_i)\,\eta(\theta)$, $a_{ij} = (\partial^2/\partial\theta_i\,\partial\theta_j)\,\eta(\theta^*)$, and $Eg_i(X_1, \theta^*) = 0$, etc.

PROPOSITION 1. *Assume "L". Then there exists a sequence of measurable functions* $\theta_n^{\text{root}}: \mathscr{X} \times \cdots \times \mathscr{X} \to K$, *and a compact neighborhood* $N \subseteqq K$ *of* θ^* *such that*:

(i) *with probability tending to* 1 *as* $n \to \infty$, $\theta_n^{\text{root}}(X_1, \cdots, X_n)$ *is the point where* $\bar{h}_n(\theta)$ *attains its unique maximum in* K, *as well as the unique root of* $\bar{g}_n(\theta) = 0$ *in* N;
(ii) $n^{\frac{1}{2}}(\theta_n^{\text{root}}(X_1, \cdots, X_n) - \theta^*) \to_{\mathscr{L}} N_q(0, \Sigma)$ *as* $n \to \infty$, *where* $\Sigma = A^{-1}SA^{-1}$.

This result is of course classical, and is essentially contained in Cramér ((1946), pages 500 ff.). Our main result is:

THEOREM. *Assume "L"; let* θ_n^{root} *be as in Proposition* 1. *Then* θ_n^{root} *obeys the jackknife property* (J.2). *If the constant* λ *in* (L.5) *can be taken to be greater than* $\frac{1}{2}$, *then* θ_n^{root} *also obeys* (J.1).

Our theorem addresses the jackknife behavior of the "consistent root" of $\bar{g}_n(\theta) = 0$; it would be desirable to have a corresponding result for the behavior of the location of the global maximum of $\bar{h}_n(\theta)$. The following simple results go part way to satisfying this want.

COROLLARY 1. *Assume "L"; let* θ_n^{root} *be as in Proposition* 1. *Assume that* $\theta_n: \mathscr{X} \times \cdots \times \mathscr{X} \to \Theta$ *is a sequence of statistics such that the following occurs with probability* 1:

For all n sufficiently large,

$$\theta_n(X_1, \cdots, X_n) = \theta_n^{\text{root}}(X_1, \cdots, X_n)$$

and

$$\theta_{n-1}(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n) = \theta_{n-1}^{\text{root}}(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n)$$

for all $j = 1, \cdots, n$. *Then the conclusions of the theorem apply to the sequence* θ_n *as well as to* θ_n^{root}.

COROLLARY 2. *Assume "L". Define* $\theta_n^{\max}: \mathscr{X} \times \cdots \times \mathscr{X} \to \Theta$ *by* $\bar{h}_n(\theta_n^{\max}) = \sup_\Theta \bar{h}_n(\theta)$. *If, for each neighborhood* $\mathscr{N} \subseteqq \Theta$ *of* θ^*, *with probability* 1 *for all sufficiently large n,*

$$\theta_n^{\max}(X_1, \cdots, X_n) \in \mathscr{N}$$

and

$$\theta_{n-1}^{\max}(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n) \in \mathscr{N}, \qquad j = 1, \cdots, n,$$

then the conclusions of the theorem apply to the sequence $\theta_n^{\max}$.

COROLLARY 3. *Assume "L"; let* $\theta_n^{\max}$ *be defined by* $\bar{h}_n(\theta_n^{\max}) = \sup_\Theta \bar{h}_n(\theta)$. *If the convergence* $\theta_n^{\max}(X_1, \cdots, X_n) \to_{\text{a.s.}} \theta^*$ *follows from any of the usual versions of Wald's consistency theorem for the maximum likelihood estimate—or, more generally, for minimum contrast estimates* (*Wald*, (1949); *Bahadur*, (1967), *page* 320; *Pfanzagl*, (1969))—*then the conclusions of our theorem apply to* $\theta_n^{\max}$.

Corollary 1 should be obvious; Corollary 2 follows from Corollary 1 and from Proposition 1. Corollary 3 may be proven by noticing that the usual proof is based (ultimately) on an appeal to the strong law of large numbers, which may be replaced by an appeal to a "jackknife strong law":

PROPOSITION 2. *Let $W_1, W_2, \cdots$ be* i.i.d. *random vectors in a separable Banach space, with $E\|W_1\| < \infty$. For each neighborhood $\mathscr{W}$ of EW_1, the following occurs with probability one: for all n sufficiently large, $\bar{W}_n = (1/n)\sum_{i=1}^n W_i \in \mathscr{W}$ and for each $j = 1, \cdots, n$, $\bar{W}_{nj} = (1/(n-1))\sum_{i=1;i\neq j}^n W_i \in \mathscr{W}$.*

SKETCH OF PROOF. $E\|W_1\| < \infty$ implies both $T_n = \max\{\|W_1\|, \cdots, \|W_n\|\}/n \to_{\text{a.s.}} 0$ and $\bar{W}_n \to_{\text{a.s.}} EW_1$. The estimate $\|\bar{W}_{nj} - \bar{W}_n\| \leqq (2n/(n-1))T_n$ completes the proof.

There are consistency theorems for the maximum likelihood estimate (m.l.e.) (and other minimum contrast estimates) which do not use the strong law of large numbers; Perlman (1972) presents several which rely instead on the martingale convergence theorem. We have not attempted to see if those theorems can be modified as we modify Wald's theorem in Corollary 3.

3. Spaces. Let W_1 and W_2 be Banach spaces. If $\mathscr{O} \subseteqq W_1$, we say $\psi: \mathscr{O} \to W_2$ is uniformly λ-Hölder continuous if there exists a constant M such that $x, y \in \mathscr{O}$ imply $|\psi(x) - \psi(y)| \leqq M|x-y|^\lambda$. If $\mathscr{O} \subseteqq W_1$ is open, we say $\psi: \mathscr{O} \to W_2$ is continuously differentiable if there is a continuous function $D\psi: \mathscr{O} \to L(W_1, W_2)$ (where $L(W_1, W_2)$ is the Banach space of continuous linear maps from W_1 to W_2, normed with the operator norm) such that $(d/dt)|_0 \psi(x+ty) = D\psi(x)y$ for all $x \in \mathscr{O}$, $y \in W_1$. It can be shown that such ψ are Fréchet differentiable at each $x \in \mathscr{O}$, with Fréchet derivative $D\psi(x)$. Let $V = \mathbb{R}^q$; in the special case $W_1 = W_2 = V$, $L(W_1, W_2) = V \otimes V^*$ is the space of q by q real matrices, and then $D\psi(x)$ is the matrix with ij entry $\partial\psi_i/\partial x_j$ where ψ_i is the ith coordinate function of ψ.

If $K \subseteqq V$ is compact, we say $\psi: K \to V$ is continuously differentiable if there is an open set $\mathscr{O} \supseteqq K$ and an extension $\psi_{\mathscr{O}}$ of ψ to $\mathscr{O}$ such that $\psi_{\mathscr{O}}: \mathscr{O} \to V$ is continuously differentiable.

If $\psi: K \to V$ is continuously differentiable, let

$$\|\psi\| = \sup\{|\psi(t)| + |D\psi(t)| : t \in K\} + \sup\{|D\psi(s) - D\psi(t)| \cdot |s-t|^{-\lambda} : s, t \in K, s \neq t\}.$$

Let

$$B = \{\psi : K \to V \text{ continuously differentiable with } \|\psi\| < \infty\};$$

let

$$B_0 = \{\psi \in B \text{ with } \psi(\theta^*) = 0 \text{ and } D\psi(\theta^*) = 0\}.$$

$B = (B, \|\ \|)$ is a Banach space; $(B_0, \|\ \|)$ is a sub-Banach space of B.

Let $U = V \oplus (V \otimes V^*) \oplus B_0$; we claim that this Banach(able) space is isomorphic to B. We check this by exhibiting two continuous linear maps $\alpha: B \to U$

and $\beta: U \rightarrow B$ which are inverse to each other. For $u = (y, z, \phi) \in U$, let $\beta(u)$ be the function of θ defined by

$$\beta(u)(\theta) = y + z(\theta - \theta^*) + \phi(\theta)$$

(here—and hereafter—$z(\theta - \theta^*)$ is matrix multiplication and $\phi(\theta)$ is argument-of-a-function). For $\psi \in B$ let $\alpha(\psi) = (\alpha_1(\psi), \alpha_2(\psi), \alpha_3(\psi)) \in U$ where

$$\alpha_1(\psi) = \psi(\theta^*)$$

$$\alpha_2(\psi) = D\psi(\theta^*)$$

and

$$\alpha_3(\psi)(\theta) = \psi(\theta) - \alpha_1(\psi) - \alpha_2(\psi)(\theta - \theta^*) .$$

The assumptions "L" imply that the random functions $g(X_i, \theta)$ are i.i.d. Borel measurable random elements $g(X_i)$ of the Banach space B, and that $E\|g(X_i)\| < \infty$. Hence the expectation of $g(X_i)$ exists; let it be denoted by γ. The continuity of α implies that $U_i = \alpha(g(X_i) - \gamma)$ is a random element of U, that $Y_i = \alpha_1(g(X_i) - \gamma)$ is a random element of V, that $Z_i = \alpha_2(g(X_1) - \gamma)$ is a random element of $V \otimes V^*$, and that $\phi_i = \alpha_3(gX_i) - \gamma)$ is a random element of B_0. (Of course, the claims about Y_i and Z_i were already known.) We define $\bar{U}_n = (U_1 + \cdots + U_n)/n$, $\bar{Y}_n = (Y_1 + \cdots + Y_n)/n$, $\bar{Z}_n = (Z_1 + \cdots + Z_n)/n$, and $\bar{\phi}_n = (\phi_1 + \cdots + \phi_n)/n$, so $\bar{U}_n = (\bar{Y}_n, \bar{Z}_n, \bar{\phi}_n)$. $\bar{Y}_n$, $\bar{Z}_n$ and $\bar{\phi}_n$ are sample averages of the mean centered i.i.d. random vectors Y_i, Z_i and ϕ_i, respectively.

The assumptions "L" ensure that $E|Y_i|^2 < \infty$, $E|Z_i| < \infty$ and $E\|\phi_i\| < \infty$. Cumbersome though it seems, it will be most convenient for us to rewrite the equation

$$\bar{g}_n(\theta) = 0$$

as

$$\gamma(\theta) + \bar{Y}_n + \bar{Z}_n(\theta - \theta^*) + \bar{\phi}_n(\theta) = 0 .$$

4. The function $f(y, z, \phi)$.

For small enough vectors $u = (y, z, \phi) \in U$, the equation

$$\gamma(\theta) + y + z(\theta - \theta^*) + \phi(\theta) = 0 \tag{4.1}$$

has a solution in θ, which we denote by $\theta = f(y, z, \phi)$. This solution is described in a pair of lemmas.

LEMMA 1. *There is a neighborhood $\mathscr{U}$ of* 0 *in U and a neighborhood $\mathscr{N}$ of θ^* in K and a continuously differentiable function $f: \mathscr{U} \rightarrow \mathscr{N}$ whose derivative is uniformly λ-Hölder continuous, such that*

(i) $f(0) = \theta^*$.

(ii) *For all $u = (y, z, \phi) \in \mathscr{U}$, $\theta = f(u)$ solves* (4.1).

(iii) *If $(u, \theta) \in \mathscr{U} \times \mathscr{N}$ satisfies* (4.1), *then $\theta = f(u)$.*

PROOF. This follows from a slight extension of the usual Banach space implicit function theorem stated, for example, in Dieudonné (1960) or Lang (1962). While the usual theorem states that a C^p equation has a C^p solution (where C^p

denotes p times continuously differentiable) it is also true that a $C^{p,\lambda}$ equation has a $C^{p,\lambda}$ solution, where $C^{p,\lambda}$ denotes C^p with the pth derivative uniformly λ-Hölder continuous. The proof of this sharpening is a trivial modification of the usual proof and hence is omitted.

Consider the map

$$m: (u, \theta) = (y, z, \phi, \theta) \mapsto \gamma(\theta) + y + z(\theta - \theta^*) + \phi(\theta) .$$

m is continuously differentiable; its first derivative is uniformly λ-Hölder continuous on bounded sets, $m(0, \theta^*) = 0$, and the partial derivative of m with respect to θ, evaluated at $(0, \theta^*)$, is nonsingular. Hence the implicit function theorem applies, and there are neighborhoods $\mathscr{U}$ and $\mathscr{N}$ and a function f as stated. □

Let the partial derivatives of $f(y, z, \phi)$ with respect to y and z be denoted by f_y and f_z, respectively. It is easy to check that $f_y(0) = -A^{-1}$ and $f_z(0) = 0$.

Lemma 2. *The neighborhood $\mathscr{U}$ in Lemma* 1 *can be chosen so that uniformly in (y, z, ϕ), $(y + \delta y, z, \phi)$, $(y, z + \delta z, \phi)$ and $(y, z, \phi + \delta\phi) \in \mathscr{U}$:*

(i) $f(y, z, \phi + \delta\phi) = f(y, z, \phi) + \mathscr{O}(\|\delta\phi\| |y|^{1+\lambda})$;
(ii) $f_z(y, z, \phi) = \mathscr{O}(|y|)$;
(iii) $f(y + \delta y, z, \phi) = f(y, z, \phi) + f_y(y, z, \phi)\, \delta y + \mathscr{O}(|\delta y|^{1+\lambda})$;
(iv) $f(y, z + \delta z, \phi) = f(y, z, \phi) + f_z(y, z, \phi)\, \delta z + \mathscr{O}(|\delta z|)^{1+\lambda}|y|)$; *and*
(v) $f_z(y + \delta y, z, \phi) = f_z(y, z, \phi) + \mathscr{O}(|\delta y|^\lambda |y| + |\delta y|)$.

Proof. For any $\phi \in B_0$ and any θ, the definition of $\|\phi\|$ implies that

$$|\phi(\theta)| \leqq \|\phi\| |\theta - \theta^*|^{1+\lambda} . \tag{4.2}$$

The nonsingularity of $D\gamma(\theta^*) = A$ implies the existence of a neighborhood $\mathscr{V}$ of θ^* and a constant $c > 0$ such that $s, t \in \mathscr{V}$ implies

$$|s - t| \leqq c|\gamma(s) - \gamma(t)| .$$

By continuity of f at 0 choose $\mathscr{U}$ so small that $f(\mathscr{U}) \subseteqq \mathscr{V}$. Then

$$|f(u_1) - f(u_2)| \leqq c|\gamma(f(u_1)) - \gamma(f(u_2))| ,$$

uniformly in $u_i \in \mathscr{U}$.

Further, choose $\mathscr{U}$ so small that $u = (y, z, \phi) \in \mathscr{U}$ implies $c|z| \leqq \frac{1}{3}$, $c\|\phi\| \leqq \frac{1}{3}$, and $|f(u) - \theta^*| \leqq 1$. Then

$$\begin{aligned}|f(u) - \theta^*| &\leqq c|\gamma(f(u)) - \gamma(\theta^*)| \\ &= c|\gamma(f(u))| \\ &= c|y + z(f(u) - \theta^*) + \phi(f(u))| \\ &\leqq c|y| + c|z|\,|f(u) - \theta^*| + c|\phi(f(u))| \\ &\leqq c|y| + c|z|\,|f(u) - \theta^*| + c\|\phi\|\,|f(u) - \theta^*|^{1+\lambda}\end{aligned}$$

by (4.2).

But $c\|\phi\| \leqq \frac{1}{3}$ and $|f(u) - \theta^*| \leqq 1$, so $c\|\phi\||f(u) - \theta^*|^{1+\lambda} \leqq \frac{1}{3}|f(u) - \theta^*|$ and hence

$$\begin{aligned} |f(u) - \theta^*| &\leqq c|y| + c|z||f(u) - \theta^*| + \tfrac{1}{3}|f(u) - \theta^*| \\ &\leqq c|y| + \tfrac{2}{3}|f(u) - \theta^*| , \end{aligned}$$

because $c|z| \leqq \frac{1}{3}$. This in turn implies

$$|f(u) - \theta^*| \leqq 3c|y| = \mathcal{O}(|y|) \,. \tag{4.3}$$

Let $u_1 = (y, z, \phi)$ and $u_2 = (y, z, \phi + \delta\phi)$; let $\theta_i = f(u_i)$. Then

$$\gamma(\theta_1) + y + z(\theta_1 - \theta^*) + \phi(\theta_1) = 0$$

and

$$\gamma(\theta_2) + y + z(\theta_2 - \theta^*) + \phi(\theta_2) + \delta\phi(\theta_2) = 0$$

so usingfformulas (4.2) and (4.3) in turn, we see

$$\begin{aligned} |\theta_1 - \theta_2| &\leqq c|z(\theta_1 - \theta_2) + \phi(\theta_1) - \phi(\theta_2)| + c|\delta\phi(\theta_2)| \\ &\leqq c|z||\theta_1 - \theta_2| + c\|\phi\||\theta_1 - \theta_2| + c\|\delta\phi\||\theta_2 - \theta^*|^{1+\lambda} \\ &\leqq \tfrac{2}{3}|\theta_1 - \theta_2| + c_2\|\delta\phi\||y|^{1+\lambda} \,. \end{aligned}$$

Hence, $|\theta_1 - \theta_2| \leqq 3c_2\|\delta\phi\||y|^{1+\lambda}$, where c_2 is some positive constant. This proves (i).

The remaining claims of the lemma follow from the formula for $f_z(u)$ obtained by implicit differentiation:

$$D\gamma(\theta)f_z + (\theta - \theta^*) + zf_z + D\phi(\theta)f_z = 0 \,,$$

or

$$f_z(u) = G(u)(\theta - \theta^*) \,, \tag{4.4}$$

where $\theta = f(u) = f(y, z, \phi)$ and $G(u) = -(D\gamma(\theta) + z + D\phi(\theta))^{-1}$.

In the sequel we will use the following simple consequence of (4.4):

$$\begin{aligned} f_z(u_1) - f_z(u_2) &= G(u_1)(\theta_1 - \theta^*) - G(u_2)(\theta_2 - \theta^*) \\ &= (G(u_1) - G(u_2))(\theta_1 - \theta^*) + G(u_2)(\theta_1 - \theta_2) \,, \end{aligned} \tag{4.5}$$

where $\theta_i = f(u_i)$.

If $\mathcal{U}$ is chosen small enough, $G(u)$ is uniformly bounded and uniformly λ-Hölder continuous on $\mathcal{U}$.

Then (4.4) together with (4.3) implies $|f_z(u)| = \mathcal{O}(|y|)$, which proves (ii).

This in turn implies

$$f(y, z_1, \phi) - f(y, z_2, \phi) = \mathcal{O}(|y||z_1 - z_2|) \,, \tag{4.6}$$

uniformly in (y, z_i, ϕ) in $\mathcal{U}$.

To prove (iii) and (iv) we use the Lagrange form of the Taylor remainder formula:

$$f(u + h) = f(u) + Df(u)h + \textstyle\int_0^1 (Df(u + th) - Df(u))h \, dt \,. \tag{4.7}$$

To prove (iii) we choose $h = (\delta y, 0, 0)$. Then the integrand in (4.7) is

$$(Df(u + th) - Df(u))h = (f_y(y + th, z, \phi) - f_y(y, z, \phi))\,\delta y \,,$$

which is bounded in norm by $|f_y(y + t\,\delta y, z, \phi) - f_y(y, z, \phi)|\,|\delta y|$. This is in turn bounded by a constant multiple of $|\delta y|^{1+\lambda}$, because f_y is uniformly λ-Hölder continuous (by Lemma 1). Thus, on integration,

$$f(y + \delta y, z, \phi) = f(y, z, \phi) + f_y(y, z, \phi)\,\delta y + \mathcal{O}(|\delta y|^{1+\lambda}),$$

which is (iii).

To prove (iv) we use (4.7) with $h = (0, \delta z, 0)$. Then the integrand $(Df(u + th) - Df(u))h$ is

$$(f_z(y, z + t\,\delta z, \phi) - f_z(y, z, \phi))\,\delta z = (f_z(u_1) - f_z(u_2))\,\delta z,$$

where $u_1 = (y, z + t\,\delta z, \phi)$ and $u_2 = (y, z, \phi)$. Let $\theta_i = f(u_i)$; then (4.5) and the uniform λ-Hölder continuity of G imply that

$$\begin{aligned}|f_z(u_1) - f_z(u_2)| &= \mathcal{O}(|u_1 - u_2|^\lambda|\theta_1 - \theta^*|) + \mathcal{O}(|G(u_2)|\,|\theta_1 - \theta_2|)\\ &= \mathcal{O}(|\delta z|^\lambda|y|) + \mathcal{O}(\theta_1 - \theta_2)\\ &= \mathcal{O}(|\delta z|^\lambda|y|) + O(|\delta z|\,|y|) \quad \text{(by (4.6))}\\ &= \mathcal{O}(|\delta z|^\lambda|y|).\end{aligned}$$

Thus the integrand in (4.7) is $\mathcal{O}(|\delta z|^{1+\lambda}|y|)$, and on integration we get (iv).

Finally, claim (v) follows directly from (4.5), with $u_1 = (y + \delta y, z, \phi)$ and $u_2 = (y, z, \phi)$. The first term in (4.5) is $\mathcal{O}(|u_1 - u_2|^\lambda|y|) = \mathcal{O}(|\delta z|^\lambda|y|)$ and the second is—by (4.3)—$\mathcal{O}(|\delta y|)$. □

5. Proofs.

Sketch of proof of Proposition 1. Let $\mathscr{U}$ be as in Lemma 2. Let θ_0 be some fixed, arbitrarily chosen point in Θ. If $\bar{U}_n \in \mathscr{U}$, let $\theta_n^{\text{root}}(X_1, \cdots, X_n) = f(\bar{U}_n)$; if $\bar{U}_n \notin \mathscr{U}$, let $\theta_n^{\text{root}}(X_1, \cdots, X_n) = \theta_0$. By the Banach space law of large numbers

$$P(\bar{U}_n \in \mathscr{U}) \to 1 \qquad \text{as} \quad n \to \infty.$$

Further application of the Banach space law of large numbers verifies the rest of (i).

Let

$$\begin{aligned}R_n &= n^{\frac{1}{2}}(f(\bar{U}_n) - f(0) - Df(0)\bar{U}_n)\\ &= n^{\frac{1}{2}}(\theta_n^{\text{root}} - \theta^* + A^{-1}\bar{Y}_n)\end{aligned}$$

if $\bar{U}_n \in \mathscr{U}$. If we can show that $R_n \to_P 0$ as $n \to \infty$, (ii) is established. But $f(\bar{U}_n) = f(\bar{Y}_n, \bar{Z}_n, \bar{\phi}_n)$ differs from $f(\bar{Y}_n, \bar{Z}_n, 0)$ by $\mathcal{O}(\|\bar{\phi}_n\|\,|\bar{Y}_n|^{1+\lambda})$; and $f(\bar{Y}_n, \bar{Z}_n, 0)$ differs from $f(\bar{Y}_n, 0, 0)$ by $\mathcal{O}(|\bar{Y}_n|\,|\bar{Z}_n|)$, by Lemma 2, provided $\bar{U}_n \in \mathscr{U}$.

Since $|n^{\frac{1}{2}}\bar{Y}_n|$ is tight, and $|\bar{Z}_n| \to_P 0$, it suffices to show that

$$R_n' = n^{\frac{1}{2}}(f(\bar{Y}_n, 0, 0) - f(0) - Df(0)(\bar{Y}_n, 0, 0)) \to_P 0.$$

By Lemma 2(iii), $R_n' = \mathcal{O}(n^{\frac{1}{2}}|\bar{Y}_n|^{1+\lambda})$, which converges to zero in probability. □

Proof of the theorem. Let $\bar{U}_{nj} = (n\bar{U}_n - U_j)/(n - 1)$, and similarly for $\bar{Y}_{nj}$, $\bar{Z}_{nj}$ and $\bar{\phi}_{nj}$. With probability 1, for all n sufficiently large, all $\bar{U}_n \in \mathscr{U}$ and

$\bar U_{nj} \in \mathscr{U}$, $j = 1, \cdots, n$. Hence, in proving (J.1) and (J.2) we may as well assume that $\bar U_n \in \mathscr{U}$ and all the $\bar U_{nj} \in \mathscr{U}$. Then the jackknifed version of θ_n^{root} is

$$\begin{aligned}\text{JK}\,\theta_n^{\text{root}} &= f(\bar U_n) - \frac{n-1}{n} \sum_{j=1}^n (f(\bar U_{nj}) - f(\bar U_n)) \\ &= f(\bar U_n) - \frac{n-1}{n} \sum_{j=1}^n R_{nj}\,.\end{aligned}$$

Making liberal use of Lemma 2, we expand R_{nj}:

$$\begin{aligned}R_{nj} &= f(\bar U_{nj}) - f(\bar U_n) \\ &= f(\bar Y_{nj}, \bar Z_{nj}, \bar\phi_{nj}) - f(\bar Y_n, \bar Z_n, \bar\phi_n) \\ &= f(\bar Y_{nj}, \bar Z_{nj}, \bar\phi_{nj}) - f(\bar Y_{nj}, \bar Z_{nj}, \bar\phi_n) + f(\bar Y_{nj}, \bar Z_{nj}, \bar\phi_n) - f(\bar Y_{nj}, \bar Z_n, \bar\phi_n) \\ &\qquad + f(\bar Y_{nj}, \bar Z_n, \bar\phi_n) - f(\bar Y_n, \bar Z_n, \bar\phi_n) \\ &= A_{nj} + f_z(\bar Y_{nj}, \bar Z_n, \bar\phi_n)(\bar Z_{nj} - \bar Z_n) + B_{nj} + f_y(\bar Y_n, \bar Z_n, \bar\phi_n)(\bar Y_{nj} - \bar Y_n) + C_{nj} \\ &= f_y(\bar U_n)(\bar Y_{nj} - \bar Y_n) + f_z(\bar U_n)(\bar Z_{nj} - \bar Z_n) + A_{nj} + B_{nj} + C_{nj} + D_{nj}\end{aligned}$$

where

$$\begin{aligned}A_{nj} &= \mathscr{O}(\|\bar\phi_{nj} - \bar\phi_n\| |\bar Y_{nj}|^{1+\lambda}) \\ B_{nj} &= \mathscr{O}(|\bar Z_{nj} - \bar Z_n|^{1+\lambda} |\bar Y_{nj}|) \\ C_{nj} &= \mathscr{O}(|\bar Y_{nj} - \bar Y_n|^{1+\lambda})\end{aligned}$$

and

$$D_{nj} = \mathscr{O}(|\bar Y_{nj} - \bar Y_n|^{\lambda} |\bar Z_{nj} - \bar Z_n| |\bar Y_n|) + \mathscr{O}(|\bar Y_{nj} - \bar Y_n| |\bar Z_{nj} - \bar Z_n|)\,.$$

The following lemma facilitates the task of checking when quantities like $n^{\frac{1}{2}} \sum_{j=1}^n |A_{nj}|$ or $n \sum_{j=1}^n |B_{nj}|^2$, etc. converge to 0 almost surely.

LEMMA 3. *Let the sequence of pairs of vectors* $(V_1, W_1), (V_2, W_2), \cdots$ *be independently and identically distributed, with* $E|V_1|^2 < \infty$ *and* $E|W_1| < \infty$. *Let* $\bar V_n = (V_1 + \cdots + V_n)/n$, *let* $\bar W_n = (W_1 + \cdots + W_n)/n$. *Let*

$$T_n = n^\alpha \sum_{j=1}^n \left|\frac{V_j - \bar V_n}{n}\right|^\beta \left|\frac{W_j - \bar W_n}{n}\right|^\gamma,$$

where β *and* γ *are nonnegative. If* $\beta \geqq 2\alpha$ *and* $\beta + \gamma > \alpha + 1$, *then* $T_n \to_{\text{a.s.}} 0$.

PROOF. It suffices to examine the behavior of $T_n' = n^\alpha \sum_{j=1}^n |V_j'/n|^\beta |W_j'/n|\gamma$, where $\{V_j'\}$ and $\{W_j'\}$ are two sequences of i.i.d. real random variables with $E|V_1'|^2 < \infty$ and $E|W_1'| < \infty$. For T_n is bounded by a weighted sum of four such quantities T_n', based on the choices $(V_i', W_i') = (1, 1), (1, |W_i|), (|V_i|, 1)$ and $(|V_i|, |W_i|)$, respectively, with weights which are monomial functions of the almost surely convergent random variables $|\bar V_n|$ and $|\bar W_n|$.

Consider the random variable $Z_i = |V_i'|^\beta |W_i'|^\gamma$; let $t = (\beta + \gamma - \alpha)^{-1}$. If $E|Z_1|^t < \infty$ and $t < 1$ (i.e., $\beta + \gamma > \alpha + 1$), Marcinkiewicz's law of large numbers (Stout, (1974), page 126) implies $T_n' = \sum Z_i/n^{1/t} \to_{\text{a.s.}} 0$. Hence it suffices to check that $\beta \geqq 2\alpha$ implies $E|Z_1|^t < \infty$. This follows from Hölder's inequality:

$$\begin{aligned} E|Z_1|^t &= E(|V_1'|^{\beta/(\beta+\gamma-\alpha)}|W_1'|^{\gamma/(\beta+\gamma-\alpha)} \\ &\leqq (E|V_1'|^{p\beta/(\beta+\gamma-\alpha)})^{1/p}(E|W_1'|^{q\gamma/(\beta+\gamma-\alpha)})^{1/q} \end{aligned}$$

for any choice of $p, q \geqq 1$ with $1/p + 1/q \leqq 1$; in particular, with $p = 2(\beta + \gamma - \alpha)/\beta$ and $q = (\beta + \gamma - \alpha)/\gamma$. □

The direct application of Lemma 3 results in the following (if $\lambda > 0$):

$$\begin{array}{ll} n^{\frac{1}{2}} \sum_{j=1}^n |A_{nj}| \rightarrow_{\text{a.s.}} 0\,, & n \sum_{j=1}^n A_{nj}^2 \rightarrow_{\text{a.s.}} 0\,, \\ n^{\frac{1}{2}} \sum_{j=1}^n |B_{nj}| \rightarrow_{\text{a.s.}} 0\,, & n \sum_{j=1}^n B_{nj}^2 \rightarrow_{\text{a.s.}} 0\,, \\ & n \sum_{j=1}^n C_{nj}^2 \rightarrow_{\text{a.s.}} 0\,, \\ n^{\frac{1}{2}} \sum_{j=1}^n |D_{nj}| \rightarrow_{\text{a.s.}} 0\,, & n \sum_{j=1}^n D_{nj}^2 \rightarrow_{\text{a.s.}} 0\,. \end{array}$$

Further, if $\lambda > \frac{1}{2}$, $n^{\frac{1}{2}} \sum |C_{nj}| \rightarrow_{\text{a.s.}} 0$.

As an example we show $n^{\frac{1}{2}} \sum |A_{nj}| \rightarrow_{\text{a.s.}} 0$. By Lemma 2 it suffices to prove that $n^{\frac{1}{2}} \sum \|\bar{\phi}_{nj} - \bar{\phi}_n\| |\bar{Y}_{nj}|^{1+\lambda} \rightarrow_{\text{a.s.}} 0$; since $\bar{Y}_{nj} = (n\bar{Y}_n - Y_j)/(n-1)$ and $\bar{\phi}_{nj} = (n\bar{\phi}_n - \phi_j)/(n-1)$ it suffices to prove that Q_1 and Q_2 converge to zero almost surely, where

$$Q_1 = n^{\frac{1}{2}} \sum \left|\frac{Y_j - \bar{Y}_n}{n}\right|^{1+\lambda} \left|\frac{\phi_j + \phi_n}{n}\right|$$

and

$$Q_2 = n^{\frac{1}{2}} \sum \left|\frac{\phi_j - \bar{\phi}_n}{n}\right| |\bar{Y}_n|^{1+\lambda}\,.$$

Q_1 clearly converges to zero a.s. by Lemma 3 with $\alpha = \frac{1}{2}$, $\beta = 1 + \lambda$, and $\gamma = 1$. To see that Q_2 converges to zero, note that $n^{\frac{1}{2}-\varepsilon}|\bar{Y}_n|$ converges to zero a.s., for any choice of $\varepsilon > 0$, in particular, for $\varepsilon = \lambda/5$. Then

$$\begin{aligned} Q_2 &= |n^{\frac{1}{2}-\varepsilon}\bar{Y}_n|^{1+\lambda} n^{(1+\lambda)(\varepsilon-\frac{1}{2})+\frac{1}{2}} \sum \left|\frac{\phi_j - \bar{\phi}_n}{n}\right| \\ &= |n^{\frac{1}{2}-\varepsilon}\bar{Y}_n|^{1+\lambda} n^{\varepsilon(1+\lambda)-\lambda/2} \sum \left|\frac{\phi_j - \bar{\phi}_n}{n}\right| \end{aligned}$$

is the product of a quantity converging to zero a.s. and a sum of the type described by Lemma 3, with $\alpha = \varepsilon(1 + \lambda) - \lambda/2$, $\beta = 0$, and $\gamma = 1$. The conditions $\beta \geqq 2\alpha$ and $\beta + \gamma > 1 + \alpha$ reduce to $\varepsilon(1 + \lambda) < \lambda/2$, which is certainly true if $\varepsilon = \lambda/5$ (since $0 < \lambda \leqq 1$). Hence the sum, and thus also Q_2, converges to zero almost surely. The other convergences claimed above follow by exactly the same type of reasoning.

Thus, if $\lambda > \frac{1}{2}$, $n^{\frac{1}{2}} \sum R_{nj} = n^{\frac{1}{2}} \sum (A_{nj} + B_{nj} + C_{nj} + D_{nj}) \rightarrow_{\text{a.s.}} 0$, proving (J.1).

The proof of (J.2) is slightly more intricate. The statement of (J.2) is that $(n-1) \sum_{j=1}^n (R_{nj} - \bar{R}_n)(R_{nj} - \bar{R}_n)' \rightarrow_P \Sigma$ where $\bar{R}_n = 1/n \sum_{j=1}^n R_{nj}$. It is clear that since $E(A^{-1}Y_1 Y_1'(A^{-1})') = \Sigma$ and since $f_y(\bar{U}_n) \to_{\text{a.s.}} f_y(0) = -A^{-1}$,

$$\frac{1}{n} \sum_{j=1}^{n} f_y(\bar{U}_n)(Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)' f_y(\bar{U}_n)' \to_{\text{a.s.}} \Sigma .$$

Hence, it suffices to show that

$$\frac{1}{n} \sum_{j=1}^{n} |nR_{nj} - f_y(\bar{U}_n)(Y_j - \bar{Y}_n)|^2 \to_P 0 ;$$

since $\bar{Y}_{nj} - \bar{Y}_n = (\bar{Y}_n - Y_j)/n - 1$, it suffices to show that

$$n \sum_{j=1}^{n} (|f_z(\bar{U}_n)(\bar{Z}_{nj} - \bar{Z}_n)|^2 + A_{nj}^2 + B_{nj}^2 + C_{nj}^2 + D_{nj}^2) \to_P 0 .$$

All that remains to be checked is that

$$T_n = n \sum_{j=1}^{n} f_z(\bar{U}_n)(\bar{Z}_{nj} - \bar{Z}_n)|^2 \to_P 0 .$$

But by Lemma 2(ii), we know that

$$\begin{aligned} T_n &= \mathcal{O}(n \sum_{j=1}^{n} |\bar{Y}_n|^2 |\bar{Z}_{nj} - \bar{Z}_n|^2) \\ &= \mathcal{O}(|n^{\frac{1}{2}} \bar{Y}_n|^2 \sum_{j=1}^{n} |\bar{Z}_{nj} - \bar{Z}_n|^2) . \end{aligned}$$

Finally, by Lemma 3 and by the central limit theorem, we know that $T_n \to_P 0$. $\square$

Note that if the moment condition in hypothesis (L.3) were strengthened so that $E|(\partial^2/\partial\theta_i\, \partial\theta_j)\, h(X_1, \theta^*)|^{1+\varepsilon} < \infty$, for some $\varepsilon > 0$, then the quantity T_n in the proof of (J.2) would actually converge to 0 almost surely, and instead of (J.2), we could draw the stronger conclusion that

$$\text{JKV}\ \theta_n \to_{\text{a.s.}} \Sigma .$$

REFERENCES

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303-324.

Brillinger, D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence-limits (the jackknife) when applied to maximum likelihood estimates. *Rev. Inst. Internat. Statist.* **32** 202-206.

Brillinger, D. R. (1977). Approximate estimation of the standard errors of complex statistics based on sample surveys. *New Zealand Statistician* **11** 35-41.

Chibisov, D. M. (1973). An asymptotic expansion for a class of estimators containing maximum likelihood estimators. *Theor. Probability Appl.* **18** 295-303.

Cramér, H. (1949). *Mathematical Methods of Statistics.* Princeton Univ. Press.

Dieudonné, J. (1960). *Foundations of Modern Analysis.* Academic Press, New York.

Lang, S. (1962). *Introduction to Differentiable Manifolds.* Wiley, New York.

Miller, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1-15.

Perlman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proc. Sixth Berkeley Symp. Math. Statist. Probability* **1** 263-281. Univ. of California Press.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika* **14** 247-272.

Stout, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595-601.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720