Ann. Inst. Statist. Math.
Vol. 44, No. 4, 687–701 (1992)

ONE-STEP JACKKNIFE FOR M-ESTIMATORS COMPUTED USING NEWTON'S METHOD*

JUN SHAO

Department of Mathematics, University of Ottawa,
585 King Edward, Ottawa, Ontario, Canada K1N 6N5

(Received October 4, 1991; revised February 6, 1992)

Abstract. To estimate the dispersion of an M-estimator computed using Newton's iterative method, the jackknife method usually requires to repeat the iterative process n times, where n is the sample size. To simplify the computation, one-step jackknife estimators, which require no iteration, are proposed in this paper. Asymptotic properties of the one-step jackknife estimators are obtained under some regularity conditions in the i.i.d. case and in a linear or nonlinear model. All the one-step jackknife estimators are shown to be asymptotically equivalent and they are also asymptotically equivalent to the original jackknife estimator. Hence one may use a dispersion estimator whose computation is the simplest. Finite sample properties of several one-step jackknife estimators are examined in a simulation study.

Key words and phrases: Asymptotic equivalence, asymptotic variance, computation of jackknife estimator, consistency, iteration, M-estimator, one-step estimator.

1. Introduction

The jackknife is one of the most popular method in estimating the dispersion of a point estimator. One important class of point estimators, including the maximum likelihood estimators, is the class of M-estimators $T(F_n)$, where F_n is the empirical distribution based on n i.i.d. samples from an unknown population F and T is a p-dimensional functional defined on a class of distribution functions. A more precise description is given in Section 2. Let $F_{n,i}$ be the empirical distribution based on the data with the i-th observation removed. The jackknife estimator of the dispersion of $T(F_n)$ is

$$V_n = \frac{n-1}{n}\sum_i [T(F_{n,i}) - \bar{T}][T(F_{n,i}) - \bar{T}]^\tau, \tag{1.1}$$

* The research was supported by Natural Sciences and Engineering Research Council of Canada.

where $\bar{T} = \sum_i T(F_{n,i})/n$ and the superscript τ denotes the transpose.

Brillinger (1964) and Reeds (1978) showed that under some conditions, V_n converges to the asymptotic covariance matrix of $T(F_n)$. For practical uses, the computation of V_n requires the evaluation of $T(F_{n,i})$, $i = 1, \ldots, n$. This may be cumbersome, since for most M-estimators, T is defined implicitly and each evaluation of T involves an iterative process.

Some efforts have been made in modifying the jackknife estimator to simplify its computation. For the least squares estimator in a nonlinear regression model, Fox *et al.* (1980) proposed a linear jackknife estimator which uses the linear term in the Taylor expansion to approximate the jackknife estimator. In the situation where the point estimator is a fixed point of an iterative process $T_{j+1} = g_n(T_j)$, where g_n is an explicitly known, data-dependent function from $\mathcal{R}^p$ to $\mathcal{R}^p$, Jorgensen (1987) considered the use of the result from a single step of iteration and suggested the following estimator

$$U_n(\epsilon) = \frac{1-\epsilon}{n^2\epsilon^2} \sum_i Q_n[T_1(G_{n,i}(\epsilon)) - \tilde{T}][T_1(G_{n,i}(\epsilon)) - \tilde{T}]^\tau Q_n^\tau, \tag{1.2}$$

where $\epsilon > 0$ is given, $G_{n,i}(\epsilon) = (F_n - \epsilon H_i)/(1-\epsilon)$, H_i is the distribution function corresponding to the point mass one at the i-th observation, $T_1(\cdot)$ denotes the result of the first step of the iteration in computing $T(\cdot)$, $\tilde{T} = \sum_i T_1(G_{n,i}(\epsilon))/n$, $Q_n = (I - J_n)^{-1}$, J_n is the Jacobian of g_n evaluated at $T(F_n)$ and I is the $p \times p$ identity matrix. Here, it is assumed that $T(F_n)$ is used as the initial point in computing $T_1(G_{n,i}(\epsilon))$, $i = 1, \ldots, n$. Jorgensen (1987) showed that for fixed n, $U_n(\epsilon)$ converges to the infinitesimal jackknife estimator as $\epsilon \to 0$. In practice ϵ is usually taken to be n^{-1} and therefore $G_{n,i}(\epsilon) = F_{n,i}$. The limiting behavior (as $n \to \infty$) of the estimator $U_n(\epsilon)$ with $\epsilon = n^{-1}$ is not studied in Jorgensen (1987).

We confine our attention to the situation where Newton's method is used for the iterative process in evaluating the M-functional T. In this case $T_1(F_n)$ is called the one-step M-estimator in the literature and its asymptotic property is well known. That is, both $T(F_n)$ and $T_1(F_n)$ are asymptotically normal with the same mean and covariance matrix. Let $V_{n,S}^{[1]}$ be the jackknife estimator of the dispersion of $T_1(F_n)$, i.e.,

$$V_{n,S}^{[1]} = \frac{n-1}{n} \sum_i [T_1(F_{n,i}) - \bar{T}_1][T_1(F_{n,i}) - \bar{T}_1]^\tau, \tag{1.3}$$

where S is the initial point used in computing $T_1(F_{n,i})$, $i = 1, \ldots, n$, and $\bar{T}_1 = \sum_i T_1(F_{n,i})/n$. Since $T(F_n)$ and $T_1(F_n)$ have the same asymptotic covariance matrix, $V_{n,S}^{[1]}$ can also be used to estimate the dispersion of $T(F_n)$. The computation of $V_{n,S}^{[1]}$ needs no iteration and is much simpler than the computation of V_n. The estimator $V_{n,S}^{[1]}$ will be called the *one-step jackknife estimator*.

Note that $V_{n,S}^{[1]}$ depends on the initial point S used in computing $T_1(F_{n,i})$, $i = 1, \ldots, n$. Define

$$V_n^{[1]} = V_{n,S}^{[1]} \quad \text{with} \quad S = T(F_n). \tag{1.4}$$

It is shown in Section 2 that $V_n^{[1]}$ is the same as $U_n(\epsilon)$ in (1.2) with $\epsilon = n^{-1}$. Practical reasons for using an initial point other than $T(F_n)$ in computing $T_1(F_{n,i})$, $i = 1, \ldots, n$, are

(1) $T(F_n)$ may not be available due to its computational complexity.

(2) If $T_1(F_n)$ instead of $T(F_n)$ is used as the point estimator, it is more natural to use $T_1(F_n)$ or the initial point used in computing $T_1(F_n)$ as the initial point in computing $T_1(F_{n,i})$, $i = 1, \ldots, n$, for the purpose of estimating the dispersion of $T_1(F_n)$.

(3) For some users, the point estimators are computed by using a standard program. Then it is easy for them to use the same initial point in computing $T(F_n)$, $T_1(F_n)$ and $T_1(F_{n,i})$, $i = 1, \ldots, n$.

However, it will be shown that if the initial point S is properly chosen, $V_{n,S}^{[1]}$ and $V_n^{[1]}$ are asymptotically equivalent, i.e., the difference between $V_{n,S}^{[1]}$ and $V_n^{[1]}$ is of the order $O_p(n^{-3/2})$. Note that both $V_{n,S}^{[1]}$ and $V_n^{[1]}$ are of the order $O_p(n^{-1})$. This gives us some flexibility: We may use a one-step jackknife estimator whose computation is the simplest and/or the most inexpansive.

Theorem 2.1 in Section 2 shows that as $n \to \infty$, $V_n^{[1]}$ is a consistent estimator of the asymptotic covariance matrix of $T(F_n)$. The asymptotic equivalence among V_n, $V_n^{[1]}$ and $V_{n,S}^{[1]}$ is established in Theorems 2.2 and 2.3 and therefore all these estimators converge to the same quantity. Although in real applications the sample size n cannot increase infinitely, the asymptotic theory provides a theoretical support (and often a guide) for the proper use of the methodology. The jackknife method is supported by not only the asymptotic theory but also a sound intuitive background (Tukey (1958), Efron (1982)) and an acceptable behavior in many finite sample simulation studies. Results from a simulation study of the finite sample performance of V_n, $V_n^{[1]}$ and $V_{n,S}^{[1]}$ are presented in Section 3.

Section 4 considers the case of regression M-estimation. We establish the consistency of the one-step jackknife estimator and show that for the least squares estimator in a nonlinear model, the one-step jackknife estimator is asymptotically equivalent to the linear jackknife estimator in Fox *et al.* (1980).

The last section contains technical proofs.

2. Asymptotic results

We start with a precise definition of the M-estimator. Let F be an unknown k-dimensional distribution function and θ be a p-dimensional parameter related to F. Suppose that $\theta = T(F)$ and T is a functional from $\mathcal{F}$ to $\mathcal{R}^p$ defined as the unique solution of

$$\int \rho(x, T(G))dG(x) = \min_{t \in \Omega} \int \rho(x, t)dG(x), \qquad G \in \mathcal{F},$$

where $\mathcal{F}$ is a convex class of k-dimensional distributions containing F and all degenerate distributions, $\Omega \subset \mathcal{R}^p$ is the set of all possible values of θ, and ρ is a real-valued function defined on $\mathcal{R}^k \times \mathcal{R}^p$. Let $X_1, \ldots, X_n$ be n i.i.d. samples from F and F_n be the empirical distribution based on $X_1, \ldots, X_n$. The M-estimator

of $\theta = T(F)$ is then $T(F_n)$. Examples of M-estimators can be found in Serfling (1980) and Huber (1981). The well-known maximum likelihood estimator is a special type of M-estimator when F has a density $f(x,\theta)$ with a known f and $\rho(x,t) = -\log[f(x,t)]$.

Suppose that for almost all x,

$$\psi(x,t) = \partial\rho(x,t)/\partial t \quad \text{and} \quad \Psi(x,t) = \partial\psi(x,t)/\partial t$$

exist for $t \in \mathcal{N}_\theta$, a compact neighborhood of θ. Let

$$\lambda(t,G) = \int \psi(x,t)dG(x) \quad \text{and} \quad \Lambda(t,G) = \partial\lambda(t,G)/\partial t.$$

Assume that $\lambda(t,F)$ and $\Lambda(t,F)$ are well defined for $t \in \mathcal{N}_\theta$ and the inverse of $\Lambda(\theta,F)$ exists. Under some conditions (Serfling (1980), Huber (1981)),

$$D_n^{-1/2}[T(F_n) - \theta] \to N(0,I) \quad \text{in law,} \tag{2.1}$$

where $D_n^{-1/2}$ is the inverse of the square root of

$$D_n = n^{-1}[\Lambda(\theta,F)]^{-1}\,\mathrm{Cov}[\psi(X_1,\theta)][\Lambda(\theta,F)]^{-\tau} \tag{2.2}$$

and $A^{-\tau}$ denotes the inverse of A^τ. The matrix D_n is called the asymptotic covariance matrix of $T(F_n)$. Suppose that $T(F_n)$ is computed by Newton's method. That is, having the result T_j from the j-th iteration, we compute

$$T_{j+1} = T_j - [\Lambda(T_j,F_n)]^{-1}\lambda(T_j,F_n), \quad j = 1,2,\ldots$$

and $T(F_n) = \lim_j T_j$. If we start with an initial estimator T_0 which is $n^{1/2}$-consistent, i.e., $T_0 - \theta = O_p(n^{-1/2})$, then (2.1) holds with $T(F_n)$ replaced by $T_1(F_n)$, the result from the first step iteration.

To compute the one-step jackknife estimator given by (1.3), we compute

$$T_1(F_{n,i}) = S - [\Lambda(S,F_{n,i})]^{-1}\lambda(S,F_{n,i}), \quad i = 1,\ldots,n,$$

with an initial estimator S. The initial estimator S can be any $n^{1/2}$-consistent estimator such as $T(F_n)$, $T_1(F_n)$ or T_0. The equivalence among $V_{n,S}^{[1]}$ using different initial estimators is studied in Theorem 2.2.

In the special case where $\rho(x,t) = (x-t)^\tau(x-t)$, $T(F_n)$ is the sample mean and V_n and $V_{n,S}^{[1]}$ are identical.

We now show that the matrix Q_n in (1.2) is equal to I and therefore $V_n^{[1]}$ in (1.4) is the same as Jorgensen's estimator $U_n(\epsilon)$ in (1.2) with $\epsilon = n^{-1}$. For Newton's method, the iterative process can be written as $T_{j+1} = g_n(T_j)$ with

$$g_n(t) = t - [\Lambda(t,F_n)]^{-1}\lambda(t,F_n).$$

The Jacobian matrix of $g_n(t)$ is

$$J_n(t) = I - [\Lambda(t,F_n)]^{-1}\partial\lambda(t,F_n)/\partial t + W_n$$

where W_n is a $p \times p$ matrix whose j-th row is $[\partial\zeta_j(t, F_n)/\partial t]^\tau \lambda(t, F_n)$, and $\zeta_j(t, F_n)$ is the j-th row of $[\Lambda(t, F_n)]^{-1}$. Since $\lambda(T(F_n), F_n) = 0$ and $\Lambda(t, F_n) = \partial\lambda(t, F_n)/\partial t$, $J_n(T(F_n)) = 0$ and therefore $Q_n = I$.

The following results show the limiting behavior of the one-step jackknife estimators and are useful for large sample statistical inferences. Proofs of these results are given in Section 5.

THEOREM 2.1. *Assume that for almost all x, $\psi(x,t)$ and $\Psi(x,t)$ are continuous on $\mathcal{N}_\theta$, for fixed t, $\psi(x,t)$ and $\Psi(x,t)$ are measurable, and there are real-valued functions $h_k(x)$, $k = 1, 2$, such that*

$$\sup_{t \in \mathcal{N}_\theta} \|\psi(x,t)\|^2 \le h_1(x), \tag{2.3}$$

$$\sup_{t \in \mathcal{N}_\theta} \|\Psi(x,t)\| \le h_2(x) \tag{2.4}$$

and

$$\int h_k(x) dF(x) < \infty, \qquad k = 1, 2, \tag{2.5}$$

where $\|a\| = (a^\tau a)^{1/2}$ for a vector a and $\|A\| = [\text{trace}(A^\tau A)]^{1/2}$ for a matrix A. Then as $n \to \infty$,

$$n(V_n^{[1]} - D_n) \underset{p}{\longrightarrow} 0, \tag{2.6}$$

where $V_n^{[1]}$ and D_n are given in (1.4) *and* (2.2), *respectively.*

Result (2.6) indicates that $V_n^{[1]}$ is a weakly consistent estimator of D_n. If in addition, $T(F_n) \to \theta$ almost surely (this holds in many situations), then it follows from the proof of Theorem 2.1 that $V_n^{[1]}$ is strongly consistent. However, (2.6) is sufficient for statistical inference.

THEOREM 2.2. *Assume the conditions in Theorem* 2.1 *and that S is $n^{1/2}$-consistent. Then*

$$V_{n,S}^{[1]} - V_n^{[1]} = o_p(n^{-1}). \tag{2.7}$$

If in addition, there is a function $h_3(x)$ such that

$$\int h_3(x) dF(x) < \infty$$

and

$$\sup_{t \in \mathcal{N}_\theta} \|\Psi(x,t)\|^2 \le h_3(x),$$

then

$$V_{n,S}^{[1]} - V_n^{[1]} = O_p(n^{-3/2}). \tag{2.8}$$

The consistency of the one-step jackknife estimators requires somewhat weaker technical conditions than the consistency of the original jackknife estimator V_n in (1.1). Reeds (1978) showed the consistency of V_n by assuming

$$\begin{aligned} &\|\Psi(x,t)-\Psi(x,s)\| \le M(x)\|t-s\|^{\epsilon}, \\ &\int M(x)dF(x) < \infty \end{aligned} \tag{2.9}$$

for an $\epsilon > 0$, which is restrictive and not required for the consistency of the one-step jackknife estimators. In fact, Theorem 2.3 shows that under (2.9), the one-step jackknife estimators and V_n are asymptotically equivalent.

THEOREM 2.3. *Assume the conditions in Theorem* 2.1 *and that* (2.9) *holds. Then*

$$V_n - V_n^{[1]} = O_p(n^{-(1+\epsilon/2)}), \tag{2.10}$$

where $\epsilon > 0$ *is given in* (2.9).

3. Simulation results

A simulation study of the finite sample performance of the one-step jackknife estimators was conducted. Let $X_1,\ldots,X_n$ be i.i.d. random variables having common distribution $F(x-\theta)$, where F is symmetric about zero but unknown and θ is the unknown parameter to be estimated. In the simulation,

$$F(x-\theta) = (1-\epsilon)\Phi(x-\theta) + \epsilon\Phi\left(\frac{x-\theta}{\tau}\right)$$

with $\theta = 1$, $\epsilon = 0.1$ and $\tau = 4$, where Φ is the standard normal distribution function ($F(x-\theta)$ is in the contaminated normal family). We considered the following two M-estimators of θ:

(1) Huber's estimator with

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & \text{if } |x-t| \le K \\ K|x-t| - \frac{1}{2}K^2 & \text{if } |x-t| > K, \end{cases}$$

where $K = 1.5$ as suggested in Lehmann ((1983), p. 376).

(2) The least p-th power estimator with $\rho(x,t) = |x-t|^p$. We considered only $p = 1.5$ in the simulation.

To compute the M-estimators $T(F_n)$, we used Newton's method and $\bar{X} = \sum_i X_i/n$ as the initial estimator. The true asymptotic variances of these two M-estimators are $1.18448/n$ and $1.09135/n$, respectively. Hence the M-estimators are much more efficient than the customary estimators sample mean and sample median, which have asymptotic variances $2.5/n$ and $1.83824/n$, respectively. The

simulation results also show that the M-estimators are more efficient even for the case where the sample size $n = 12$.

We examined the following four jackknife estimators:

(1) The jackknife estimator V_n.

(2) The one-step jackknife estimator $V_{n1}^{[1]} = V_n^{[1]}$ in (1.4).

(3) The one-step jackknife estimator $V_{n2}^{[1]}$ using $T_1(F_n)$ as the initial point S.

(4) The one-step jackknife estimator $V_{n3}^{[1]}$ using $\bar{X}$ as the initial point S.

Table 1 shows the simulation means and variances of these jackknife estimators under 5,000 replication for $n = 12$, 20 and 36. For comparison, the simulation means and variances and the asymptotic variances D_n of the two M-estimators $T(F_n)$ and the corresponding one-step M-estimators $T_1(F_n)$ are also given in Table 1.

Table 1. Simulation means and variances of the M-estimators and the jackknife estimators.

Part A: Huber's estimator

	$T(F_n)$	$T_1(F_n)$	V_n	$V_{n1}^{[1]}$	$V_{n2}^{[1]}$	$V_{n3}^{[1]}$
			$n = 12, D_n = 0.09871$			
Mean	1.02745	1.03045	0.09977	0.09977	0.09944	0.10133
Var	0.08663	0.08703	0.00419	0.00419	0.00408	0.00440
			$n = 20, D_n = 0.05922$			
Mean	1.02768	1.03029	0.05874	0.05874	0.05868	0.05949
Var	0.05135	0.05156	0.00073	0.00073	0.00073	0.00080
			$n = 36, D_n = 0.03290$			
Mean	1.03093	1.03348	0.03273	0.03273	0.03268	0.03313
Var	0.02873	0.02885	0.00012	0.00012	0.00012	0.00013

Part B: The least 1.5th power estimator

	$T(F_n)$	$T_1(F_n)$	V_n	$V_{n1}^{[1]}$	$V_{n2}^{[1]}$	$V_{n3}^{[1]}$
			$n = 12, D_n = 0.09095$			
Mean	1.00137	1.00092	0.09957	0.09957	0.10217	0.11384
Var	0.09200	0.09361	0.00545	0.00545	0.00522	0.00565
			$n = 20, D_n = 0.05457$			
Mean	1.00002	1.00026	0.05682	0.05682	0.05826	0.06315
Var	0.05461	0.05533	0.00113	0.00113	0.00106	0.00115
			$n = 36, D_n = 0.03032$			
Mean	1.00118	1.00127	0.03131	0.03131	0.03185	0.03365
Var	0.03037	0.03062	0.00020	0.00020	0.00019	0.00020

The simulation results are summarized as follows.

(1) The mean and the variance of the one-step jackknife estimator $V_{n1}^{[1]}$ is almost identical to the mean and the variance of the jackknife estimator V_n, re-

spectively. Note that $V_{n1}^{[1]}$ uses the M-estimator $T(F_n)$ as the initial estimator S.

(2) The means and the variances of the other two one-step jackknife estimators, $V_{n2}^{[1]}$ and $V_{n3}^{[1]}$, are not the same but close to the mean and the variance of V_n, respectively. Thus, in addition to the asymptotic results in Theorems 2.2 and 2.3, the one-step jackknife estimators are also close to the jackknife estimator for fixed and not very large sample sizes.

(3) The jackknife and one-step jackknife estimators are good estimators of the asymptotic variance D_n. For small n, the finite sample variance of $T(F_n)$ or $T_1(F_n)$ may not be close to D_n, especially for the case of Huber's estimator (Table 1, part A). In such cases, the jackknife and one-step jackknife estimators are not very good as estimates of the finite sample variances of $T(F_n)$ and $T_1(F_n)$.

(4) Overall, the one-step estimator $V_{n2}^{[1]}$, which uses $T_1(F_n)$ as the initial point S, is preferred in terms of its performance and computational simplicity.

4. Results for regression M-estimators

We extend some results in Section 2 to the regression problem. Let y_i, $i = 1, \ldots, n$, be the observations satisfying

$$y_i = f(z_i, \beta) + \epsilon_i, \tag{4.1}$$

where β is a p-dimensional vector of unknown parameters, z_i's are values of a q-dimensional regressor, f is a known real-valued function on $\mathcal{R}^q \times \mathcal{R}^p$, and ϵ_i's are random errors. If f is linear in β, then (4.1) is called a linear model; otherwise (4.1) is called a nonlinear model. The M-estimator of β is obtained by solving

$$\sum_i \varphi(y_i - f(z_i, \hat{\beta}_n)) = \min_{\beta \in \mathcal{B}} \sum_i \varphi(y_i - f(z_i, \beta)),$$

where φ is a real-valued function on $\mathcal{R}$ and $\mathcal{B}$ is the set of all possible values of β. When $\varphi(t) = t^2/2$, $\hat{\beta}_n$ is the ordinary least squares estimator.

We now study the estimation of the dispersion of the M-estimator $\hat{\beta}_n$. If z_i's are random and $X_i = (y_i, z_i)$, $i = 1, \ldots, n$, are i.i.d., the problem reduces to a special case of that in Section 2 with $\rho(x, t) = \varphi(y - f(z, t))$. Therefore, we focus on the case where z_i's are deterministic.

Assume that φ is twice continuously differentiable and that ϵ_i's are independent with

$$E[\varphi'(\epsilon_i)] = 0 \quad \text{for all } i. \tag{4.2}$$

Note that we do not assume that ϵ_i's are identically distributed. If φ is an even function and the distribution of ϵ_i is symmetric about zero, then $E[\varphi'(\epsilon_i)] = 0$. In the case of least squares estimation, $\varphi(t) = t^2/2$ and (4.2) reduces to the usual assumption that $E(\epsilon_i) = 0$ for all i.

Under some conditions (see Theorem 4.1), $\hat{\beta}_n$ is asymptotically normal with mean β and asymptotic covariance matrix

$$D_n = [M_n(\beta)]^{-1} \sum_i E[\varphi'(\epsilon_i)]^2 g(z_i, \beta)[g(z_i, \beta)]^\tau [M_n(\beta)]^{-1},$$

where $g(z, \gamma) = \partial f(z, \gamma)/\partial \gamma$ and

$$M_n(\beta) = \sum_i E[\varphi''(\epsilon_i)] g(z_i, \beta)[g(z_i, \beta)]^\tau. \tag{4.3}$$

Let $G(z, \gamma) = \partial g(z, \gamma)/\partial \gamma$, $r_i = y_i - f(z_i, \hat{\beta}_n)$,

$$H_{n,i} = \sum_{j \neq i} \{\varphi'(r_j) G(z_j, \hat{\beta}_n) - \varphi''(r_j) g(z_j, \hat{\beta}_n)[g(z_j, \hat{\beta}_n)]^\tau\}$$

and

$$\begin{aligned} t_{n,i} &= \hat{\beta}_n - H_{n,i}^{-1} \sum_{j \neq i} \varphi'(r_j) g(z_j, \hat{\beta}_n) \\ &= \hat{\beta}_n + H_{n,i}^{-1} \varphi'(r_i) g(z_i, \hat{\beta}_n). \end{aligned}$$

Then the one-step jackknife estimator of the dispersion of $\hat{\beta}_n$ is

$$V_n^{[1]} = \frac{n-1}{n} \sum_i \left(t_{n,i} - \frac{1}{n} \sum_i t_{n,i} \right) \left(t_{n,i} - \frac{1}{n} \sum_i t_{n,i} \right)^\tau.$$

Here we use $\hat{\beta}_n$ as the initial estimator in computing the one-step M-estimator after deleting the i-th pair (y_i, z_i). Hence $V_n^{[1]}$ is the same as Jorgensen's estimator $U_n(\epsilon)$ with $\epsilon = n^{-1}$.

Some special cases are studied in the following.

(1) $\varphi(t) = t^2/2$ and f is linear in β. In this case, model (4.1) is a linear model and $\hat{\beta}_n$ is the ordinary least squares estimator. The one-step jackknife estimator reduces to the usual jackknife estimator given in Miller (1974). Its consistency was shown by Miller (1974) for i.i.d. ϵ_i and by Shao and Wu (1987) for independent but non-indentically distributed ϵ_i.

(2) $\varphi(t) = t^2/2$ and f is nonlinear in β. In this case, model (4.1) is a nonlinear model and $\hat{\beta}_n$ is the ordinary least squares estimator. Fox *et al.* (1980) proposed the linear jackknife estimator

$$\tilde{V}_n = \frac{n-1}{n} \sum_i \left(\tilde{t}_{n,i} - \frac{1}{n} \sum_i \tilde{t}_{n,i} \right) \left(\tilde{t}_{n,i} - \frac{1}{n} \sum_i \tilde{t}_{n,i} \right)^\tau,$$

where

$$\tilde{t}_{n,i} = \hat{\beta}_n + \tilde{H}_n^{-1} r_i g(z_i, \hat{\beta}_n)$$

and

$$\tilde{H}_n = -\sum_i g(z_i, \hat{\beta}_n)[g(z_i, \hat{\beta}_n)]^\tau.$$

Note that $\varphi'(r_i) = r_i$ and $\varphi''(r_i) = 1$ in this case. It is shown in Section 5 that under the conditions stated in Theorem 4.1,

$$V_n^{[1]} - \tilde{V}_n = o_p(n^{-1}). \tag{4.4}$$

(3) f is linear in β but φ is arbitrary. In this case $\hat{\beta}_n$ is the M-estimator in a linear model (Huber (1981)). The estimator $V_n^{[1]}$ is asymptotically equivalent to the one given by (6.6) and (6.15) in Huber ((1981), Chapter 7) with $w_i = [\varphi'(r_i)]^2$ to handle the unequality of $E[\varphi'(\epsilon_i)]^2$.

The proof of the following theorem is given in Section 5.

THEOREM 4.1. *Assume the conditions on φ and ϵ_i previously stated and that $f(z,\gamma)$, $g(z,\gamma)$ and $G(z,\gamma)$ are continuous functions on $\{\|z\| \le z_\infty\} \times \mathcal{R}^p$, where $z_\infty = \sup_i \|z_i\|$ is finite. Assume further that there is a function h such that for all i,*

$$\sup_{\gamma \in \mathcal{N}_\beta} [\varphi'(y_i - f(z_i,\gamma))]^2 \le h(y_i),$$

$$\sup_{\gamma \in \mathcal{N}_\beta} |\varphi''(y_i - f(z_i,\gamma))| \le h(y_i)$$

and

$$\sup_i E[h(y_i)]^{1+\delta} < \infty$$

with a constant $\delta > 0$, where $\mathcal{N}_\beta = \{\gamma : \|\gamma - \beta\| \le c\}$ for a constant $c > 0$. Let $M_n(\beta)$ be given by (4.3). Assume that $\hat{\beta}_n \xrightarrow[p]{} \beta$ and $\liminf_n [n^{-1} M_n(\beta)] > 0$. Then

$$n(V_n^{[1]} - D_n) \xrightarrow[p]{} 0.$$

5. Proofs

We first establish the following lemma.

LEMMA 5.1. *Assume the conditions on Ψ stated in Theorem 2.1. Then for any sequence of random vectors ξ_n satisfying $\xi_n \xrightarrow[p]{} \theta$,*

$$\max_{i \le n} \|\Lambda(\xi_n, F_{n,i}) - \Lambda(\theta, F)\| \xrightarrow[p]{} 0.$$

PROOF. From the conditions on Ψ and the uniform strong law of large numbers,

$$\|\Lambda(\xi_n, F_n) - \Lambda(\theta, F)\| \xrightarrow[p]{} 0.$$

The result follows from

$$\max_{i\le n}\|\Lambda(\xi_n,F_{n,i})-\Lambda(\xi_n,F_n)\|\underset{p}{\longrightarrow}0. \tag{5.1}$$

Note that

$$\Lambda(\xi_n,F_{n,i})-\Lambda(\xi_n,F_n)=\frac{1}{n-1}[\Lambda(\xi_n,F_n)-\Psi(X_i,\xi_n)]. \tag{5.2}$$

Hence for $\xi_n\in\mathcal{N}_\theta$,

$$\begin{aligned}\max_{i\le n}\|\Lambda(\xi_n,F_{n,i})-\Lambda(\xi_n,F_n)\|&\le\frac{1}{n-1}\left[\|\Lambda(\xi_n,F_n)\|+\max_{i\le n}\|\Psi(X_i,\xi_n)\|\right]\\&\le\frac{1}{n-1}\|\Lambda(\xi_n,F_n)\|+\frac{1}{n-1}\max_{i\le n}[h_2(X_i)]\underset{p}{\longrightarrow}0.\end{aligned}$$

Hence (5.1) follows from $P(\xi_n\in\mathcal{N}_\theta)\to 1$. □

PROOF OF THEOREM 2.1. Since $T(F_n)\underset{p}{\longrightarrow}\theta$, we may focus on the event $\{T(F_n)\in\mathcal{N}_\theta\}$. Let

$$\begin{aligned}&\Lambda=\Lambda(\theta,F),\quad \Lambda_n=\Lambda(S,F_n),\quad \lambda_n=\lambda(S,F_n),\\&\Lambda_{n,i}=\Lambda(S,F_{n,i}),\quad \lambda_{n,i}=\lambda(S,F_{n,i}),\end{aligned} \tag{5.3}$$

$$A_n=(n-1)\sum_i(\Lambda_{n,i}^{-1}\lambda_{n,i}-\Lambda_n^{-1}\lambda_n)(\Lambda_{n,i}^{-1}\lambda_{n,i}-\Lambda_n^{-1}\lambda_n)^\tau \tag{5.4}$$

and

$$B_n=n(n-1)\left(\frac{1}{n}\sum_i\Lambda_{n,i}^{-1}\lambda_{n,i}-\Lambda_n^{-1}\lambda_n\right)\left(\frac{1}{n}\sum_i\Lambda_{n,i}^{-1}\lambda_{n,i}-\Lambda_n^{-1}\lambda_n\right)^\tau. \tag{5.5}$$

Then

$$nV_n^{[1]}=A_n-B_n$$

with $S=T(F_n)$. For $S=T(F_n)$, $\lambda_n=0$. Hence

$$A_n=(n-1)\sum_i\Lambda_{n,i}^{-1}(\lambda_{n,i}-\lambda_n)(\lambda_{n,i}-\lambda_n)^\tau\Lambda_{n,i}^{-\tau}.$$

Let

$$C_n=(n-1)\sum_i\Lambda^{-1}(\lambda_{n,i}-\lambda_n)(\lambda_{n,i}-\lambda_n)^\tau\Lambda^{-\tau}.$$

Since

$$\begin{aligned}\lambda_{n,i}-\lambda_n&=\frac{1}{n-1}\sum_{j\ne i}\psi(X_j,S)-\frac{1}{n}\sum_j\psi(X_j,S)\\&=\frac{1}{n-1}[\lambda_n-\psi(X_i,S)],\end{aligned} \tag{5.6}$$

we have

$$C_n = \frac{1}{n-1}\Lambda^{-1}\sum_i [\psi(X_i, S) - \lambda_n][\psi(X_i, S) - \lambda_n]^\tau \Lambda^{-\tau}.$$

By conditions (2.4), (2.6), $S \xrightarrow[p]{} \theta$ and the uniform strong law of large numbers, $C_n - nD_n \xrightarrow[p]{} 0$. By Lemma 5.1, $A_n - C_n \xrightarrow[p]{} 0$ and therefore $A_n - nD_n \xrightarrow[p]{} 0$. It remains to show that $B_n \xrightarrow[p]{} 0$, which follows from

$$b_n = \sum_i \Lambda_{n,i}^{-1}\lambda_{n,i} - n\Lambda_n^{-1}\lambda_n \xrightarrow[p]{} 0.$$

From (5.2), (5.6) and $\Lambda_{n,i}^{-1} - \Lambda_n^{-1} = \Lambda_{n,i}^{-1}(\Lambda_n - \Lambda_{n,i})\Lambda_n^{-1}$,

$$\begin{aligned}\|b_n\| &= \left\| \frac{1}{n-1}\sum_i \Lambda_{n,i}^{-1}[\Psi(X_i, S) - \Lambda_n]\Lambda_n^{-1}\lambda_{n,i} \right\| \\ &\le \|\Lambda_n^{-1}\| \max_{i\le n}\|\lambda_{n,i}\| \max_{i\le n}\|\Lambda_{n,i}^{-1}\| \frac{1}{n-1}\sum_i \|\Psi(X_i, S) - \Lambda_n\|.\end{aligned}$$

Hence $b_n \xrightarrow[p]{} 0$ follows from (5.6), Lemma 5.1 and condition (2.4)–(2.6). □

PROOF OF THEOREM 2.2. We use the notation given in (5.3)–(5.5). Let

$$Z_n = (n-1)\sum_i (\Lambda_{n,i}^{-1} - \Lambda_n^{-1})\lambda_n\lambda_n^\tau(\Lambda_{n,i}^{-\tau} - \Lambda_n^{-\tau}).$$

Then A_n in (5.4) is equal to

$$Z_n + (n-1)\sum_i \Lambda_{n,i}^{-1}(\lambda_{n,i} - \lambda_n)(\lambda_{n,i} - \lambda_n)^\tau \Lambda_{n,i}^{-\tau} + \text{cross product terms}.$$

For (2.7), it suffices to show that $Z_n \xrightarrow[p]{} 0$. Since $\Lambda_{n,i}^{-1} - \Lambda_n^{-1} = (1/(n-1))\Lambda_{n,i}^{-1}[\Psi(X_i, S) - \Lambda_n]\Lambda_n^{-1}$,

$$\|Z_n\| \le \|\lambda_n\|^2\|\Lambda_n^{-1}\|^2 \max_{i\le n}\|\Lambda_{n,i}^{-1}\|^2 \frac{1}{n-1}\sum_i \|\Psi(X_i, S) - \Lambda_n\|^2. \tag{5.7}$$

From condition (2.4), for $S \in \mathcal{N}_\theta$,

$$\sum_i \|\Psi(X_i, S) - \Lambda_n\|^2 \le 2\sum_i [h_2(X_i)]^2 + 2n\|\Lambda_n\|^2.$$

From condition (2.5), $\sum_i [h_2(X_i)]^2/n^2 \xrightarrow[p]{} 0$. Hence

$$\frac{1}{n(n-1)}\sum \|\Psi(X_i, S) - \Lambda_n\|^2 \xrightarrow[p]{} 0. \tag{5.8}$$

By the mean-value theorem and the fact that $\lambda(T(F_n), F_n) = 0$,

$$\lambda_n = \Lambda(\xi_n, F_n)[S - T(F_n)],$$

where ξ_n is between S and $T(F_n)$. From Lemma 5.1, $\Lambda(\xi_n, F_n) \xrightarrow[p]{} \Lambda$. Then $\lambda_n = O_p(n^{-1/2})$, since both S and $T(F_n)$ are $n^{1/2}$-consistent. Thus

$$\|\lambda_n\|^2 = O_p(n^{-1}). \tag{5.9}$$

Hence $Z_n \xrightarrow[p]{} 0$ by (5.7)–(5.9).

For (2.8), it suffices to show that

$$Z_n = O_p(n^{-1}) \quad \text{and} \quad B_n = O_p(n^{-1}).$$

From (5.7), (5.9) and Lemma 5.1, $Z_n = O_p(n^{-1})$ is implied by

$$\sum_i \|\Psi(X_i, S) - \Lambda_n\|^2 = O_p(n^{-1}),$$

which holds under the assumed conditions on Ψ. The proof of $B_n = O_p(n^{-1})$ is similar. □

PROOF OF THEOREM 2.3. By the mean value theorem and $\lambda(T(F_n), F_n) = 0$,

$$\lambda_{n,i} = \Lambda(\xi_{n,i}, F_{n,i})[T(F_n) - T(F_{n,i})],$$

where $\xi_{n,i}$ is between $T(F_n)$ and $T(F_{n,i})$. Then

$$V_n = \frac{n-1}{n} \sum_i [\Lambda(\xi_{n,i}, F_{n,i})]^{-1} \lambda_{n,i} \lambda_{n,i}^\tau [\Lambda(\xi_{n,i}, F_{n,i})]^{-\tau} - \frac{n-1}{n^2} \left\{ \sum_i [\Lambda(\xi_{n,i}, F_{n,i})]^{-1} \lambda_{n,i} \right\} \left\{ \sum_i [\Lambda(\xi_{n,i}, F_{n,i})]^{-1} \lambda_{n,i} \right\}^\tau.$$

Comparing the above expression with the A_n and B_n in the proof of Theorem 2.1, we conclude that (2.10) is implied by

$$\max_{i \le n} \|\Lambda_{n,ii}^{-1} - \Lambda_{n,i}^{-1}\| = O_p(n^{-\epsilon/2}), \tag{5.10}$$

where $\Lambda_{n,ii} = \Lambda(\xi_{n,i}, F_{n,i})$ and $\Lambda_{n,i} = \Lambda(T(F_n), F_{n,i})$. From condition (2.9),

$$\begin{aligned} \|\Lambda_{n,ii}^{-1} - \Lambda_{n,i}^{-1}\| &\le \|\Lambda_{n,ii}^{-1}\| \|\Lambda_{n,i}^{-1}\| \|\Lambda_{n,ii} - \Lambda_{n,i}\| \\ &\le \|\Lambda_{n,ii}^{-1}\| \|\Lambda_{n,i}^{-1}\| \|T(F_n) - \xi_{n,i}\|^\epsilon \left[\frac{1}{n-1} \sum_{j \ne i} M(X_j) \right] \\ &\le \|\Lambda_{n,ii}^{-1}\| \|\Lambda_{n,i}^{-1}\| \|T(F_n) - T(F_{n,i})\|^\epsilon \left[\frac{1}{n-1} \sum_j M(X_j) \right]. \end{aligned}$$

From Reeds (1978), $n(V_n - D_n) \underset{p}{\longrightarrow} 0$. Hence

$$\max_{i \le n} \|T(F_n) - T(F_{n,i})\|^2 \le \sum_i \|T(F_n) - T(F_{n,i})\|^2 = O_p(n^{-1}).$$

This and Lemma 5.1 imply (5.10). □

PROOF OF THEOREM 4.1. By using the following lemma, the result can be shown using the same argument used in the proof of Theorem 2.1. Therefore, the detailed proof is omitted. □

LEMMA 5.2. *Let y_i, $i = 1, 2, \ldots$ be a sequence of independent random variables, z_i, $i = 1, 2, \ldots$ be a sequence of deterministic q-vectors satisfying $\sup_i \|z_i\| = z_\infty < \infty$ and $u(y, z, \gamma)$ be a continuous function on $\mathcal{R} \times \{\|z\| \le z_\infty\} \times \mathcal{N}_\beta$. Suppose that there is a function $h(y)$ such that*

$$\begin{aligned} &\sup_{\|z\| \le z_\infty, \gamma \in \mathcal{N}_\beta} |u(y, z, \gamma)| \le h(y), \\ &\sup_i E[h(y_i)]^{1+\delta} < \infty \end{aligned} \tag{5.11}$$

for a constant $\delta > 0$. Let γ_n, $n = 1, 2, \ldots$ be a sequence of random p-vectors satisfying $\gamma_n \underset{p}{\longrightarrow} \beta$. Then

$$\frac{1}{n} \sum_i \{u(y_i, z_i, \gamma_n) - E[u(y_i, z_i, \beta)]\} \underset{p}{\longrightarrow} 0.$$

PROOF. Under the given conditions, we need only to show that

$$\frac{1}{n} \sum_i |u(y_i, z_i, \gamma_n) - u(y_i, z_i, \beta)| \underset{p}{\longrightarrow} 0. \tag{5.12}$$

Let $\chi(A)$ be the indicator function of the set A. Then the left side of (5.12) is bounded by the sum of

$$\frac{1}{n} \sum_i |u(y_i, z_i, \gamma_n) - u(y_i, z_i, \beta)| \chi(|y_i| > m) \tag{5.13}$$

and

$$\frac{1}{n} \sum_i |u(y_i, z_i, \gamma_n) - u(y_i, z_i, \beta)| \chi(|y_i| \le m) \tag{5.14}$$

for any $m > 0$. Under condition (5.11), the quantity in (5.13) can be made arbitrarily small by choosing m large. For any fixed m, the quantity in (5.14) converges to zero in probability by the uniform continuity of the function u on $[-m, m] \times \{\|z\| \le z_\infty\} \times \mathcal{N}_\beta$ and the condition that $\gamma_n \underset{p}{\longrightarrow} \beta$. This proves (5.12). □

PROOF OF (4.4). Comparing $t_{n,i}$ with $\tilde{t}_{n,i}$, we need only to show

$$\max_{i\le n} \|\tilde{H}_n^{-1} H_{n,i} - I\| \xrightarrow[p]{} 0. \tag{5.15}$$

Note that for $\varphi(t) = t^2/2$, $\varphi'(t) = t$ and $\varphi''(t) = 1$. From the definition of $H_{n,i}$ and $\tilde{H}_n$, the left side of (5.15) is equal to

$$\max_{i\le n}\left\|\tilde{H}_n^{-1}\left\{g(z_i,\hat{\beta}_n)[g(z_i,\hat{\beta}_n)]^\tau + \sum_{j\neq i} r_j G(z_j,\hat{\beta}_n)\right\}\right\|$$
$$\le \|\tilde{H}_n^{-1}\|\left\{\max_{i\le n}\|g(z_i,\hat{\beta}_n)\|^2 + \max_{i\le n}|r_i|\|G(z_i,\hat{\beta}_n)\| + \left\|\sum_i r_i G(z_i,\hat{\beta}_n)\right\|\right\}.$$

From the continuity of the functions f, g and G,

$$\frac{1}{n}\max_{i\le n}\|g(z_i,\hat{\beta}_n)\|^2 \xrightarrow[p]{} 0$$

and

$$\frac{1}{n}\max_{i\le n}|r_i|\|G(z_i,\hat{\beta}_n)\| \xrightarrow[p]{} 0.$$

From Lemma 5.2, $\|\tilde{H}_n\| = O_p(n)$ and

$$\frac{1}{n}\sum_i r_i G(z_i,\hat{\beta}_n) \xrightarrow[p]{} 0.$$

Hence (5.15) holds. □

REFERENCES

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

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, Philadelphia.

Fox, T., Hinkley, D. and Larntz, K. (1980). Jackknifing in non-linear regression, *Technometrics*, **22**, 29–33.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Jorgensen, M. A. (1987). Jackknifing fixed points of iterations, *Biometrika*, **74**, 207–211.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Miller, R. M. (1974). An unbalanced jackknife, *Ann. Statist.*, **2**, 880–891.

Reeds, J. A. (1978). Jackknifing maximum likelihood estimates, *Ann. Statist.*, **6**, 727–739.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Shao, J. and Wu, C. F. J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models, *Ann. Statist.*, **15**, 1563–1579.

Tukey, J. (1958). Bias and confidence in not quite large samples, *Ann. Math. Statist.*, **29**, 614.