ADAPTIVE ESTIMATES FOR AUTOREGRESSIVE PROCESSES*

RUDOLF BERAN

(Received Dec. 4, 1972; revised Jan. 6, 1975)

Abstract

Let $\{X_t: t=0, \pm 1, \pm 2, \cdots\}$ be a stationary rth order autoregressive process whose generating disturbances are independent identically distributed random variables with marginal distribution function F. Adaptive estimates for the parameters of $\{X_t\}$ are constructed from the observed portion of a sample path. The asymptotic efficiency of these estimates relative to the least squares estimates is greater than or equal to one for all regular F. The nature of the adaptive estimates encourages stable behavior for moderate sample sizes. A similar approach can be taken to estimation problems in the general linear model.

1. Introduction

A discrete stochastic process $\{X_t: t=0, \pm 1, \pm 2, \cdots\}$ may be called a stationary rth order autoregressive process if it has the following properties: the process is strictly stationary and satisfies a difference equation of the form

$$X_t-\mu=\sum_{j=1}^{r} \alpha_j(X_{t-j}-\mu)+E_t , \tag{1.1}$$

where $\{E_t\}$ is a sequence of independent identically distributed random variables with mean 0 and finite variance σ^2, the parameters μ, $\{\alpha_j\}$ are real-valued, and the roots of the polynomial equation

$$x^r=\sum_{j=1}^{r} \alpha_j x^{r-j} \tag{1.2}$$

all have modulus less than one. Under these assumptions, the difference equation (1.1) has a unique solution, expressible in the form

* This research was partially supported by National Science Foundation Grant GP-31091X. American Mathematical Society 1970 subject classification. Primary 62N10; Secondary 62G35. Key words and phrases: autoregressive process, adaptive estimates, robust estimates.

$$X_t = \mu + \sum_{j=0}^{\infty} \beta_j E_{t-j} , \tag{1.3}$$

the sum on the right converging in mean square (Anderson [1] or Mann and Wald [7]).

Stationary autoregressive processes serve as convenient parametric models for certain time series; their prediction theory is particularly simple. Estimation of the parameters from the observed portion $(X_1, X_2, \cdots, X_N)$ of a sample path is commonly carried out by writing (1.1) in the form

$$X_t = \alpha_0 + \sum_{j=1}^{r} \alpha_j X_{t-j} + E_t \tag{1.4}$$

and by applying the method of least squares. If X is the $(N-r)\times(r+1)$ matrix whose ith row is $(1, X_{r+i-1}, X_{r+i-2}, \cdots, X_i)$ and Y is the $(N-r)\times 1$ vector $(X_{r+1}, X_{r+2}, \cdots, X_N)^T$, the least squares estimate (LSE) of $\rho=(\alpha_0, \alpha_1, \cdots, \alpha_r)^T$ based upon $(X_1, X_2, \cdots, X_N)$, $N\geqq 2r+1$, is

$$\hat{\rho}_N = (X^T X)^{-1} X^T Y . \tag{1.5}$$

The asymptotic distribution of $N^{1/2}(\hat{\rho}_N - \rho)$ as $N\to\infty$ is normal $(0, \sigma^2\Gamma^{-1})$, where $\Gamma=(N-r)^{-1}\,\mathrm{E}\,(X^T X)$ (Anderson [1] or Mann and Wald [7]).

Let F denote the marginal distribution function of E_t and let $(\cdot, \cdot)$ and $\|\cdot\|$ denote the inner product and norm in $L_2(F)$. Under regularity conditions (cf. Section 2), there exists an estimate $\hat{\rho}_N(\phi_F)$, depending on

$$\phi_F(x) = -f'(x)/f(x) , \tag{1.6}$$

f being the density of F, such that the asymptotic distribution of $N^{1/2}(\hat{\rho}_N(\phi_F)-\rho)$ is normal $(0, \|\phi_F\|^{-2}\Gamma^{-1})$. The asymptotic efficiency of the LSE $\hat{\rho}_N$ relative to $\hat{\rho}_N(\phi_F)$ is less than or equal to one, with equality if and only if F is normal. Thus, the problem arises of constructing a practical estimate of ρ whose asymptotic performance will dominate that of the LSE for all F of interest. This paper develops one possible answer, under mild regularity assumptions on F.

2. Linearized estimates

Let ρ_0 denote the parameter vector of the autoregressive model under which the data is realized. For every $\rho=(\alpha_0, \alpha_1, \cdots, \alpha_r)^T \in R^{r+1}$, define the residual process

$$R_t(\rho) = X_t - \alpha_0 - \sum_{j=1}^{r} \alpha_j X_{t-j} , \qquad r+1\leqq t\leqq N . \tag{2.1}$$

Corresponding to any scalar-valued function ϕ defined on R^1, let $\Psi(\rho)$

denote the $(N-r)\times 1$ vector $(\psi\circ R_{r+1}(\rho), \psi\circ R_{r+2}(\rho), \cdots, \psi\circ R_N(\rho))^T$, where $\psi\circ R_t(\rho)$ denotes $\psi(R_t(\rho))$. Under the usual assumptions, a conditional maximum likelihood estimate for ρ_0 must be a solution to the equation

$$X^T\Phi_F(\rho)=0\ ; \tag{2.2}$$

Φ_F is associated with ϕ_F in the obvious way. Consequently an M-estimate for ρ_0 may be defined (Huber [4]) as any estimate which satisfies an equation of the form

$$X^T\Psi(\rho)=0 \tag{2.3}$$

for some score function ψ.

Application of Newton's method to (2.3), with the LSE $\hat\rho_N$ as starting point, suggests the linearized estimate

$$\hat\rho_N(\psi)=\hat\rho_N+[\hat D(\psi, F)]^{-1}(X^TX)^{-1}X^T\Psi(\hat\rho_N)\ , \tag{2.4}$$

where $\hat D(\psi, F)$ is a consistent estimate of a functional $D(\psi, F)$ defined below; one possibility is $D(\psi, F)=(\psi', 1)$. This technique for adjusting estimates is well-known in the literature if $\psi=\phi_F$ (for example, see LeCam [6]). We will study the asymptotic behavior of $\hat\rho_N(\psi)$ as a preliminary to the construction of adaptive estimates.

A basic result which underlies the asymptotic theory developed in this paper is

LEMMA 2.1. *If $\{X_t\}$ is a stationary rth order autoregressive process, then*

$$N^{-1/2}\max_{1\le t\le N}|X_t|\xrightarrow{p}0$$

$$N^{-1}X^TX\xrightarrow{p}\Gamma \qquad \textit{nonsingular}$$

as $N\to\infty$.

The first property follows from $\mathrm{E}\,X_t^2<\infty$, which is implied by our definition of X_t; the second is proved in Anderson [1], for example.

The following assumptions on ψ and F will be used:

A1. $\psi=\psi_+-\psi_-$, where $\psi_\pm\in L_2(F)$ and is monotone nondecreasing; $(\psi, 1)=0$.

A2. $\lim_{h\to 0}\|\psi_\pm(x+h)-\psi_\pm(x-h)\|^2=0$, and for some $\varepsilon>0$,
$$\sup_{|a|\le\varepsilon,\,|h|\le\varepsilon}|h|^{-1}([\psi_\pm(x+a+h)-\psi_\pm(x+a)], 1)<\infty.$$

A3. $\lim_{h\to 0}(2h)^{-1}([\psi_\pm(x+h)-\psi_\pm(x-h)], 1)=D(\psi_\pm, F)<\infty.$

In applications of our results, ψ will belong to a known, relatively small family of functions. With this family specified in advance, assumptions A1, A2, A3 amount to regularity assumptions on F. For the sake of brevity in exposition, any F satisfying A1, A2, A3 for a given family of functions ψ will be called *regular*.

For every $z \in R^{r+1}$, let $|z|$ denote $\max_{1 \leq i \leq r+1} |z_i|$. Let $D(\psi, F) = D(\psi_+, F) - D(\psi_-, F)$ and let $\hat{D}(\psi, F)$ denote a consistent estimate of $D(\psi, F)$.

THEOREM 2.1. *If* A1, A2, A3 *are satisfied, then for every* $C>0$

$$N^{-1/2} \sup_{|\rho-\rho_0| \leq CN^{-1/2}} |X^T\Psi(\rho) - X^T\Psi(\rho_0) + (X^TX)(\rho-\rho_0)\hat{D}(\psi, F)| \xrightarrow{P} 0 \tag{2.6}$$

in ρ_0*-probability as* $N \to \infty$.

The proof of this theorem is deferred to Section 5. A similar result for the general linear model has been established by Bickel [2].

THEOREM 2.2. *If* A1, A2, A3 *are satisfied, the asymptotic distribution of* $N^{1/2}(\hat{\rho}_N(\psi) - \rho_0)$ *as* $N \to \infty$ *is normal* $(0, V(\psi, F)\Gamma^{-1})$, *where*

$$V(\psi, F) = \|\psi\|^2 / D^2(\psi, F) . \tag{2.7}$$

PROOF. Since $N^{1/2}(\hat{\rho}_N - \rho_0)$ is bounded in probability asymptotically, it follows from (2.4) and Theorem 2.1 that

$$N^{1/2}(\hat{\rho}_N(\psi) - \rho_0) = N^{1/2}[\hat{D}(\psi, F)]^{-1}(X^TX)^{-1}X^T\Psi(\rho_0) + o_p(1) . \tag{2.8}$$

Let A_t be the σ-algebra generated by $(X_1, X_2, \cdots, X_t)$, $t \geqq 1$. For every $c \in R^{r+1}$, let

$$S_N(c) = c^T X^T \Psi(\rho_0) . \tag{2.9}$$

Because of A1 and (1.3), $\{S_N(c), A_N;\ N \geqq 2r+1\}$ is a martingale. Application of a suitable central limit theorem (Brown [3], p. 60) shows that the asymptotic distribution of $N^{-1/2}S_N(c)$ is normal $(0, \|\psi\|^2 c^T \Gamma c)$. The theorem follows.

Suppose that F has an absolutely continuous density f and ϕ_F satisfies A1, A2, A3 with $D(\phi_F, F) = (\phi_F', 1) = \|\phi_F\|^2$. Then the estimate $\hat{\rho}_N(\phi_F)$, defined according to (2.4), achieves the minimal asymptotic covariance matrix $\|\phi_F\|^{-2}\Gamma^{-1}$ mentioned earlier.

3. Adaptive estimates

Since F is usually not known, the estimate $\hat{\rho}_N(\phi_F)$ cannot be found in practice. A natural idea is to estimate ϕ_F from the data and use this estimate in place of ϕ_F in (2.4). While theoretically possible, this approach encounters the difficulty that consistent estimates of ϕ_F may converge very slowly as the sample size increases.

In estimating ϕ_F, as in estimating densities, there is a trade-off between asymptotic variance and asymptotic bias. This suggests a more modest approach that separates these two considerations: Replace ϕ_F with an approximation $\phi_{F,H}$ which is easier to estimate from the data but keeps $V(\phi_{F,H}, F)$ close to $V(\phi_F, F)$ for a range of interesting F. Estimate $\phi_{F,H}$ consistently by $\hat{\phi}_{F,H}$ and estimate ρ_0 by $\hat{\rho}_N(\hat{\phi}_{F,H})$.

This program motivates the developments in this section. However, the results will be formulated in a manner that does not involve ϕ_F.

A real-valued function g defined on R^1 will be said to satisfy condition C if $g=g_+-g_-$, where $g_\pm$ is monotone nondecreasing, and $\lim_{h\to 0}([g_\pm(x+h)-g_\pm(x-h)], 1)=0$. Let $\{\psi_i: 1\leqq i\leqq k\}$ be a family of score functions which fulfill some or all of the following assumptions, as required.

B1. Each ψ_i satisfies A1 and A2.

B2. Each $\psi_{i\pm}$ is absolutely continuous and $\psi'_{i\pm}$ satisfies condition C.

B3. Each $\psi_i\psi_j$ satisfies condition C.

B4. If $\left\|\sum_{j=1}^k c_j\psi_j\right\|=0$ for some constants $\{c_j\}$, then $c_j=0$, $1\leqq j\leqq k$.

Note that B2 implies that A3 holds for each ψ_i, with $D(\psi_{i\pm}, F)=(\psi'_{i\pm}, 1)$. Let H be the subspace of $L_2(F)$ spanned by all linear combinations of the $\{\psi_i\}$. Let W denote the $k\times k$ matrix whose (i, j)th element is (ψ_i, ψ_j) and let $v=((\psi'_1, 1), (\psi'_2, 1), \cdots, (\psi'_k, 1))^T$. Assumption B4 ensures that W is nonsingular. Define the vector $a=(a_1, a_2, \cdots, a_k)^T$ by $a=W^{-1}v$ and let

$$\phi_{F,H}=\sum_{i=1}^k a_i\psi_i \ . \tag{3.1}$$

LEMMA 3.1. *If each* $\psi_i\in L_2(F)$ *and* B2, B4 *are satisfied, then*

$$V(\phi_{F,H}, F)=\min_{\psi\in H} V(\psi, F) \ . \tag{3.2}$$

PROOF. If $\psi\in H$, there exist constants $c=(c_1, c_2, \cdots, c_k)^T$ such that $\psi=\sum_{i=1}^k c_i\psi_i$. Since $v=Wa$,

$$D(\psi, F)=(\psi', 1)=c^Tv=c^TWa=(\psi, \phi_{F,H}) \ . \tag{3.3}$$

Hence

$$V(\psi, F)=\frac{\|\psi\|^2}{(\psi, \phi_{F,H})^2}\geqq\frac{1}{\|\phi_{F,H}\|^2}=V(\phi_{F,H}, F) \ , \tag{3.4}$$

with equality if and only if ψ is proportional to $\phi_{F,H}$.

An interesting interpretation can be given to this lemma when F has an absolutely continuous density and $\phi_F\in L_2(F)$. In this case, $(\psi', 1)=(\psi, \phi_F)$ for every $\psi\in L_2(F)$ and $\phi_{F,H}$ is simply the projection of ϕ_F into

H. Note also that any multiple of $\phi_{F,H}$ will retain the minimizing property (3.2).

LEMMA 3.2. *If* B1, B2, B3 *are satisfied, then for every* $C>0$, $1 \leqq i, j \leqq K$,

$$(3.5)\qquad \begin{aligned} &(N-r)^{-1} \sup_{|\rho-\rho_0| \leqq CN^{-1/2}} \left| \sum_{t=r+1}^{N} \psi_i' \circ R_t(\rho) - \sum_{t=r+1}^{N} \psi_i'(E_t) \right| \xrightarrow{p} 0 \\ &(N-r)^{-1} \sup_{|\rho-\rho_0| \leqq CN^{-1/2}} \left| \sum_{t=r+1}^{N} [\psi_i \circ R_t(\rho)][\psi_j \circ R_t(\rho)] - \sum_{t=r+1}^{N} \psi_i(E_t)\psi_j(E_t) \right| \xrightarrow{p} 0 \end{aligned}$$

in ρ_0*-probability as* $N \to \infty$.

PROOF. Let $\rho_0 - \rho = (\Delta_0, \Delta_1, \cdots, \Delta_r)^T$. For each $\delta>0$, define $U_t(\delta)$ by

$$(3.6)\qquad U_t(\delta) = \begin{cases} X_t & \text{if } |X_t| \leqq \delta N^{-1/2} \\ 0 & \text{otherwise}. \end{cases}$$

Without loss of generality, assume ψ' is monotone nondecreasing. Under B2,

$$(3.7)\qquad \begin{aligned} &(N-r)^{-1} \mathrm{E}\left[\sup_{|\rho-\rho_0| \leqq CN^{-1/2}} \left| \sum_{t=r+1}^{N} \psi_i'\Big(E_t + \Delta_0 + \sum_{j=1}^{r} \Delta_j U_{t-j}(\delta)\Big) - \sum_{t=r+1}^{N} \psi_i'(E_t) \right| \right] \\ &\qquad \leqq \mathrm{E}\,[\psi_i'(E_t + CN^{-1/2} + rC\delta) - \psi_i'(E_t - CN^{-1/2} - rC\delta)] \to 0 \end{aligned}$$

as $N \to \infty$ and $\delta \to 0$. Moreover, because of Lemma 2.1,

$$(3.8)\qquad \begin{aligned} &\mathrm{P}\,[(U_1(\delta), U_2(\delta), \cdots, U_N(\delta)) \neq (X_1, X_2, \cdots, X_N)] \\ &\qquad \leqq \mathrm{P}\,[\max_{1 \leqq t \leqq N} |X_t| > \delta N^{1/2}] \to 0 \end{aligned}$$

for every $\delta>0$ as $N \to \infty$. The first line in (3.5) is implied by (3.7) and (3.8); the second is proved analogously using B3.

Let $\hat{v}$ be the $k \times 1$ vector whose ith component is $(N-r)^{-1} \sum_{t=r+1}^{N} \psi_i' \circ R_t(\hat{\rho}_N)$ and let $\hat{W}$ be the $k \times k$ matrix whose (i,j)th element is $(N-r)^{-1} \cdot \sum_{t=r+1}^{N} [\psi_i \circ R_t(\hat{\rho}_N)][\psi_j \circ R_t(\hat{\rho}_N)]$. From the lemma above, it follows that $\hat{v} \xrightarrow{p} v$ and $\hat{W} \xrightarrow{p} W$ as $N \to \infty$. Define $\hat{W}^{-1}$ as the inverse of $\hat{W}$ when possible and arbitrarily otherwise. Since W is nonsingular under B4, $\hat{W}^{-1} \xrightarrow{p} W^{-1}$ and $\hat{a} = \hat{W}^{-1}\hat{v} \xrightarrow{p} a$ as $N \to \infty$. The implied estimate of $\phi_{F,H}$ is

$$(3.9)\qquad \hat{\phi}_{F,H} = \sum_{i=1}^{k} \hat{a}_i \psi_i .$$

By setting $\psi = \hat{\phi}_{F,H}$ in the linearized estimate $\hat{\rho}_N(\psi)$ and noting that $\hat{a}^T \hat{W} \hat{a}$ is a consistent estimate of $D(\phi_{F,H}, F) = \|\phi_{F,H}\|^2 = a^T W a$, we arrive at the adaptive estimate

(3.10) $$\hat{\rho}_N(H)=\hat{\rho}_N+(\hat{a}^T\hat{W}\hat{a})^{-1}(X^TX)^{-1}X^T\hat{\Phi}_{F,H}(\hat{\rho}) ,$$

where $\hat{\Phi}_{F,H}(\hat{\rho}_N)$ is the $(N-r)\times 1$ vector of scored residuals $(\hat{\phi}_{F,H}\circ R_{r+1}(\hat{\rho}_N), \cdots, \hat{\phi}_{F,H}\circ R_N(\hat{\rho}_N))^T$.

THEOREM 3.1. *If* B1, B2, B3, B4 *are satisfied, then as* $N\to\infty$, $N^{1/2}(\hat{\rho}_N(H)-\hat{\rho}_N(\phi_{F,H}))\xrightarrow{p}0$ *in* ρ_0*-probability and the asymptotic distribution of* $N^{1/2}(\hat{\rho}_N(H)-\rho_0)$ *is normal* $(0, \|\phi_{F,H}\|^{-2}\Gamma^{-1})$.

PROOF. As in (2.2), let $\Phi_{F,H}$, Ψ_i denote $(N-r)\times 1$ vectors of scored residuals, the score functions being $\phi_{F,H}$, ψ_i respectively. Since

(3.11) $$\Phi_{F,H}(\rho)=\sum_{i=1}^{k} a_i\Psi_i(\rho) \qquad \hat{\Phi}_{F,H}(\rho)=\sum_{i=1}^{k}\hat{a}_i\Psi_i(\rho) ,$$

we may write

(3.12) $$N^{-1/2}X^T\hat{\Phi}_{F,H}(\hat{\rho}_N)-N^{-1/2}X^T\Phi_{F,H}(\hat{\rho}_N)=\sum_{i=1}^{k}(\hat{a}_i-a_i)(N^{-1/2}X^T\Psi_i(\hat{\rho}_N)) .$$

Theorem 2.1 implies that $N^{-1/2}X^T\Psi_i(\hat{\rho}_N)$ is bounded in probability asymptotically. Since also $\hat{a}\xrightarrow{p}a$, the difference (3.12) converges in probability to zero as $N\to\infty$. The theorem follows with the help of Theorem 2.2.

A desirable property possessed by the LSE $\hat{\rho}_N$ is invariance under rescaling of the observations in the sense that the mapping $X_t\to cX_t$, $c>0$, induces the mappings $\hat{\alpha}_0\to c\hat{\alpha}_0$ and $\hat{\alpha}_i\to\hat{\alpha}_i$ for $1\leqq i\leqq r$. For suitable H, the adaptive estimate $\hat{\rho}_N(H)$ is also scale invariant.

DEFINITION. A subspace H of $L_2(F)$ is said to be closed under scaling if $\psi(\cdot)\in H$ implies that $\psi(c\cdot)\in H$ for every scalar $c>0$.

THEOREM 3.2. *If H is closed under scaling and $\hat{W}$ is nonsingular, the adaptive estimate $\hat{\rho}_N(H)$ is scale invariant.*

PROOF. Let F_N denote the empirical distribution function of the residuals $\{R_t(\hat{\rho}_N): r+1\leqq t\leqq N\}$ and let

(3.13) $$\hat{V}(\psi, F_N)=\frac{\int\psi^2(t)dF_N(t)}{\left[\int\psi'(t)dF_N(t)\right]^2} .$$

An argument like that for Lemma 3.1 shows that $\hat{\phi}_{F,H}$ is characterized, up to a constant of proportionality, by the property

(3.14) $$\hat{V}(\hat{\phi}_{F,H}, F_N)=\min_{\psi\in H}\hat{V}(\psi, F_N) .$$

The scaling $X_t\to cX_t$, $1\leqq t\leqq N$, $c>0$, induces the following mappings: $R_t(\hat{\rho}_N)\to cR_t(\hat{\rho}_N)$ and $F_N(\cdot)\to F_N(\cdot/c)$. Consideration of (3.13), (3.14) and the closure of H yields

$$\hat{V}(\hat{\phi}_{F,H}(\cdot/c), F_N(\cdot/c)) = c^2\hat{V}(\hat{\phi}_{F,H}, F_N) = c^2 \min_{\phi\in H} \hat{V}(\phi, F_N) = \min_{\phi\in H} \hat{V}(\phi, F_N(\cdot/c)) . \tag{3.15}$$

Consequently, the scaling $X_t \to cX_t$ must map $\hat{\phi}_{F,H}(\cdot)$ into a multiple of $\hat{\phi}_{F,H}(\cdot/c)$. The theorem follows from this fact and the invariance of $\hat{\rho}_N$.

Under the assumptions of Lemma 3.2, $\hat{V}(\phi, F_N)$ is a consistent estimate of $V(\phi, F)$. Therefore, $\hat{\phi}_{F,H}$ is an element of H that minimizes, in the obvious sense, $N(X^TX)^{-1}\hat{V}(\phi, F_N)$, the estimated asymptotic covariance matrix of $\hat{\rho}_N(\phi)$. This fact makes $\hat{\rho}_N(H)$ an extended analogue of the adaptive L-estimates for location studied by Jaeckel [5].

4. Applications

Scale invariant adaptive estimates of ρ can be constructed as follows. Assume F is symmetric about the origin and take as a basis for H the set of functions $\phi_i(x) = |x|^{r_i} \operatorname{sign}(x)$, $r_i > 0$, $1 \leq i \leq k$. In this case, H is closed under scaling and the other assumptions required by Theorem 3.1 can be checked readily. Indeed, let $\phi(x) = |x|^r \operatorname{sign}(x)$, $r > 0$. If $\int |x|^{2r} dF(x) < \infty$, assumption B1 holds for ϕ, and if also $r \geq 1$, so does B2. On the other hand, if $0 < r < 1$, fulfillment of B2 is assured whenever F has a bounded density which is uniformly continuous in a neighborhood of the origin. Assumption B3 holds under a moment condition similar to that for B1, while B4 is satisfied if F is absolutely continuous and the exponents $\{r_i\}$ are distinct.

Whenever the particular score function $\phi(x) = x$ belongs to H, the asymptotic efficiency of $\hat{\rho}_N(H)$ relative to the LSE $\hat{\rho}_N$ is greater than or equal to one for all regular F; this is a consequence of Theorem 3.1 and Lemma 3.1.

The adaptive estimate $\hat{\rho}_N(H)$ can be applied to hypothesis testing. Let C be a $q \times (r+1)$ matrix constant of rank q, let $\tau = C\rho$ and let $\hat{\tau}_N = C\hat{\rho}_N(H)$. To test the hypothesis H: $\sum_{i=1}^{q} \tau_i^2 = 0$ versus K: $\sum_{i=1}^{q} \tau_i^2 > 0$, calculate

$$T_N = [\hat{\tau}_N^T (C\hat{\Gamma}_N^{-1}C')^{-1} \hat{\tau}_N] [\hat{a}^T \hat{W} \hat{a}] , \tag{4.1}$$

where $\hat{\Gamma}_N = N^{-1}X^TX$, and reject H for values of T_N that are large relative to the asymptotic χ_q^2 distribution implied by Theorem 3.1. Under a sequence of alternatives K_N : $\tau = N^{-1/2}\Delta$, Δ a non-null $q \times 1$ vector, the asymptotic efficiency of this test relative to the corresponding test based on $\hat{\rho}_N$ is the same as in the estimation problem.

The central ideas underlying the definition and properties of $\hat{\rho}_N(H)$ carry over unchanged to the general linear model. For counterparts of Theorems 2.1 and 2.2 in that case, see Bickel [2]. The analogues of Theorems 3.1 and 3.2 will be evident to the reader.

Practical aspects. The following practical suggestions are made on partly heuristic grounds and need further investigation.

1. For samples from a moderately contaminated normal distribution, try $\psi_1(x)=|x|^{1/2}\,\mathrm{sign}\,(x)$, $\psi_2(x)=x$ as a basis for H. If F is normal, $\hat{\rho}_N(H)$ is still fully efficient asymptotically. If F is actually double exponential, $\hat{\phi}_{F,H}(x)$ will converge in probability to $\phi_{F,H}(x)\doteq 1.90\psi_1(x)-.76\psi_2(x)$, and therefore $\hat{\rho}_N(H)$ will discount outlying residuals in large samples. Note that $\phi_{F,H}$ becomes negative-valued only far out in the tails of the double exponential distribution. With F double exponential, the asymptotic efficiency of $\hat{\rho}_N(H)$ relative to the best estimate is $(1/2)[1+\pi/(32-9\pi)]\doteq .92$. The efficiency of $\hat{\rho}_N$ in this case is only .50 and the efficiency of $\hat{\rho}_N(\psi_1)$ is .79.

2. If more serious departures from normality are anticipated, bring into H selected functions of the form $\psi(x)=|x|^r\,\mathrm{sign}\,(x)$, with $0\leqq r<1$ for heavier tailed F and $r>1$ for lighter tailed F. In some cases it will be necessary to replace $\hat{\rho}_N$ with a more robust estimate, such as the M-estimate corresponding to the score function $\psi(x)=\mathrm{sign}\,(x)$.

3. If the sample size N is large, some experimentation with the choice of H may be worthwhile. Plot $\hat{\phi}_{F,H}(x)$ and note changes that occur as functions are added to or removed from the basis of H. The aim is to discover, at least qualitatively, the shape of ϕ_F. Keep K small relative to N.

Numerical example. To check the numerical practicality of the adaptive estimator, a pseudo-random sample of size 50 was generated from the autoregressive process $X_t=.5+.5X_{t-1}+E_t$, where E_t has a double-exponential distribution with scale parameter .75. The first column of Table 1 records the sample values. For this data, the LSE of $\rho=(\alpha_0,\alpha_1)^T=(.5,.5)^T$ is $\hat{\rho}_N=(.457692, .534041)^T$ and the least squares residuals $R_t(\hat{\rho}_N)=X_t-\hat{\alpha}_0-\hat{\alpha}_1X_{t-1}$ are given by the second column of Table 1.

The basis for the subspace H consists of two functions: $\psi_1(x)=|x|^{1/2}\,\mathrm{sgn}\,(x)$ and $\psi_2(x)=x$. The estimates of W and a are

$$W=\begin{pmatrix}.714230 & .784176\\ .784176 & .959636\end{pmatrix}, \qquad \hat{a}=\begin{pmatrix}1.51483\\ -.19579\end{pmatrix}. \tag{4.2}$$

The actual values of W and a under the double-exponential model that generated the sample are

Table 1

Autoregressive series	Residuals	Scored residuals
1.13639	.524676	.99453
1.58924	.866306	1.24032
2.17272	− .375636	− .854877
1.24238	− .403473	− .883213
.7177	1.28047	1.46344
2.12145	.71641	1.1419
2.30704	.416844	.896409
2.10659	−2.3101	−1.85008
− .727401	.239953	.695056
.309182	− .130794	− .522236
.492013	2.8828	2.00756
3.60325	− .437629	− .916427
1.94435	1.92483	1.72477
3.42088	2.42519	1.88421
4.70978	− .70798	−1.13598
2.26493	.65645E−2	.121448
1.67382	−1.1386	−1.39347
.212983	1.49821	1.56083
2.06964	.815332E−1	.41658
1.6445	−1.77403	−1.67029
− .438107	− .731375	−1.15229
− .507651	.261263	.723134
.447848	− .955391	−1.29359
− .25853	− .473251	− .949439
− .153625	− .396044	− .875769
− .020395	− .805794E−1	− .41423
.36622	−1.6481	−1.62202
− .994836	.820102E−1	.41775
.841844E−2	.109264	.479334
.571451	.464126	.94113
1.227	− .811606E−1	− .415664
1.0318	− .722722	−1.1463
.285992	−1.18925	−1.41911
− .578829	1.18348	1.41623
1.33206	− .904759E−1	− .437934
1.07859	.337221	.813645
1.37092	−1.50854	−1.56519
− .318721	− .340982	− .817801
− .535011E−1	− .773312	−1.1807
− .344192	− .217533	− .663931
.563454E−1	.163325	.580217
.651107	− .313339	− .786601
.49207	.524206	.99413
1.24468	.116964	.495169
1.23937	.51387	.985287
1.63344	− .30782	− .78018
1.02219	− .39051	− .870168
.613075	.472668	.948911
1.25777	.406439	.886163
1.53583		

$$W=\begin{pmatrix} .750000 & .863432 \\ .863432 & 1.125000 \end{pmatrix}, \qquad a=\begin{pmatrix} 2.92981 \\ -1.35972 \end{pmatrix}. \tag{4.3}$$

Although $\hat{a}$ differs markedly from a, the components of $\hat{a}$ have the correct signs. Hence, scoring the residuals according to a has the desired effect of discounting the larger residual values. The scored residual vector $\hat{\Phi}_{F,H}(\hat{\rho}_N)$ is listed in the third column of Table 1.

The adaptive estimate $\hat{\rho}_N(H)$ is $(.451717, .487241)^T$. Comparison with the LSE and the actual parameter values shows that the adaptive estimate of α is slightly worse than the LSE of α_0, but the situation is reversed for α_1 and the gain in accuracy outweighs the loss.

5. Proof of Theorem 2.1

The random vector $X^t\Psi(\rho)$ may be written out as

$$\left(\sum_{t=r+1}^{N} \psi\circ R_t(\rho),\ \sum_{t=r+1}^{N} X_{t-1}\psi\circ R_t(\rho), \cdots, \sum_{t=r+1}^{N} X_{t-r}\psi\circ R_t(\rho)\right)^T .$$

Corresponding to these components, define

$$\begin{aligned} T_{N0}(\rho)&=N^{-1/2}\sum_{t=r+1}^{N}[\psi\circ R_t(\rho)-\mathrm{E}(\psi\circ R_t(\rho)\,|\,A_{t-1})] \\ T_{Ni}(\rho)&=N^{-1/2}\sum_{t=r+1}^{N}X_{t-i}[\psi\circ R_t(\rho)-\mathrm{E}(\psi\circ R_t(\rho)\,|\,A_{t-1})]\ , \qquad 1\leqq i\leqq r\ . \end{aligned} \tag{5.1}$$

The method of expansion adopted in this section uses ideas from Bickel [2].

LEMMA 5.1. *If* A1, A2 *are satisfied, then for every* $C>0$

$$\sup_{|\rho-\rho_0|\leqq CN^{-1/2}} |T_{Ni}(\rho)-T_{Ni}(\rho_0)|\xrightarrow{p}0\ , \qquad 0\leqq i\leqq r \tag{5.2}$$

in ρ_0-probability as $N\to\infty$.

PROOF. Without loss of generality, assume that $i\geqq 1$ and ψ is monotone nondecreasing. We begin by showing that if $|\rho-\rho_0|\leqq CN^{-1/2}$,

$$T_{Ni}(\rho)-T_{Ni}(\rho_0)\xrightarrow{p}0 \tag{5.3}$$

as $N\to\infty$. Indeed, let $\rho_0-\rho=(\Delta_0, \Delta_1, \cdots, \Delta_r)^T$ and let

$$\begin{aligned} R_t^\delta(\rho)&=E_t+\Delta_0+\sum_{j=1}^{r}\Delta_j U_{t-j}(\delta) \\ T_{Ni}^\delta(\rho)&=N^{-1/2}\sum_{t=r+1}^{N}X_{t-i}[\psi\circ R_t^\delta(\rho)-\mathrm{E}(\psi\circ R_t^\delta(\rho)\,|\,A_{t-1})]\ , \end{aligned} \tag{5.4}$$

where the $\{U_j(\delta)\}$ are defined as in (3.6). Because of (3.8),

$$(5.5)\qquad P[T_{Ni}(\rho)\neq T^{\delta}_{Ni}(\rho)]\rightarrow 0$$

as $N\rightarrow\infty$.

$$(5.6)\qquad \begin{aligned} &E[T^{\delta}_{Ni}(\rho)-T_{Ni}(\rho_0)]^2 \\ &=N^{-1}\sum_{t=r+1}^{N} E\{X^2_{t-i}[\psi\circ R^{\delta}_t(\rho)-\psi\circ R_t(\rho_0) \\ &\quad -E(\psi\circ R^{\delta}_t(\rho)-\psi\circ R_t(\rho_0)\,|\,A_{t-1})]^2\} \\ &\leqq N^{-1}E\sum_{t=r+1}^{N} X^2_{t-i}\int\Big[\psi\Big(x+\Delta_0+\sum_{j=1}^{r}\Delta_j U_{t-j}(\delta)\Big)-\psi(x)\Big]^2 dF(x) \\ &\leqq N^{-1}E\sum_{t=r+1}^{N} X^2_{t-i}\int[\psi(x+h)-\psi(x-h)]^2 dF(x)\ , \end{aligned}$$

where $h=CN^{-1/2}+rC\delta$. Now (5.3) follows because of A2 and (5.5).

Decompose the $r+1$ dimensional cube $B=\{\rho\colon |\rho-\rho_0|\leqq CN^{-1/2}\}$ into sub-cubes whose vertices are at the points $\rho_0+(j_1\varepsilon N^{-1/2}, j_2\varepsilon N^{-1/2},\cdots, j_{r+1}\varepsilon N^{-1/2})$, where $\varepsilon>0$ is chosen to give an even division of B into sub-cubes and $j_i=0,\pm1,\cdots,\pm M(\varepsilon)$. For each $\rho\in B$, let $V(\rho)$ denote the vertex nearest ρ_0 of the sub-cube containing ρ (or one of them, in case of ties). Then for each $\varepsilon>0$, from (5.3)

$$(5.7)\qquad \sup_{|\rho-\rho_0|\leqq CN^{-1/2}}|T_{Ni}(\rho_0)-T_{Ni}\circ V(\rho)|\xrightarrow{p}0$$

as $N\rightarrow\infty$.

Suppose that $\rho\in B^*$, a particular sub-cube of B, and let $\rho^*=V(\rho)$. Then

$$(5.8)\qquad \begin{aligned} &\sup_{\rho\in B^*}|T_{Ni}(\rho)-T_{Ni}\circ V(\rho)| \\ &\leqq\sup_{\rho\in B^*}\Big|N^{-1/2}\sum_{t=r+1}^{N} X_{t-i}(\psi\circ R_t(\rho)-\psi\circ R_t(\rho^*))\Big| \\ &\quad+\sup_{\rho\in B^*}\Big|N^{-1/2}\sum_{t=r+1}^{N} X_{t-i}\,E(\psi\circ R_t(\rho)-\psi\circ R_t(\rho^*)\,|\,A_{t-1})\Big| \\ &=D_1+D_2\ . \end{aligned}$$

By an argument like that for (5.3),

$$(5.9)\qquad D_1\leqq N^{-1/2}\sum_{t=r+1}^{N}|X_{t-i}|\,E[\psi(R_t(\rho^*)+\varepsilon S_t)-\psi(R_t(\rho^*)-\varepsilon S_t)\,|\,A_{t-1}]+o_p(1)\ ,$$

where $S_t=N^{-1/2}\Big(1+\sum_{j=1}^{r}|X_{t-j}|\Big)$. Hence

$$(5.10)\qquad \begin{aligned} &\sup_{\rho\in B^*}|T_{Ni}(\rho)-T_{Ni}\circ V(\rho)| \\ &\leqq 2N^{-1/2}\sum_{t=r+1}^{N}|X_{t-i}|\,E[\psi(R_t(\rho^*)+\varepsilon S_t)-\psi(R_t(\rho^*)-\varepsilon S_t)\,|\,A_{t-1}] \\ &\quad+o_p(1) \end{aligned}$$

$$\leqq 2N^{-1/2}\sum_{t=r+1}^{N}|X_{t-i}|\int[\psi(x+U_t+\varepsilon S_t)-\psi(x+U_t-\varepsilon S_t)]dF(x)$$
$$+o_p(1)\ ,$$

where $|U_t|\leqq CS_t$. In view of this, A2, and Lemma 2.1,

$$\sup_{\rho\in B^*}|T_{Ni}(\rho)-T_{Ni}\circ V(\rho)|=O_p(\varepsilon) \tag{5.11}$$

as $N\rightarrow\infty$. The lemma follows from (5.11) and (5.7).

PROOF OF THEOREM 2.1. Note that

$$\begin{aligned} N^{-1/2}\sum_{t=r+1}^{N}X_{t-i}\,\mathrm{E}\,(\psi\circ R_t(\rho)-\psi\circ R_t(\rho_0)\,|\,A_{t-1}) \\ =N^{-1/2}\sum_{t=r+1}^{N}X_{t-i}V_t\cdot V_t^{-1}\int[\psi(x+V_t)-\psi(x)]dF(x)\ , \end{aligned} \tag{5.12}$$

where $V_t=\Delta_0+\sum_{j=1}^{N}\Delta_jX_{t-j}$. As $N\rightarrow\infty$, the difference between (5.12) and $N^{-1/2}\left(\sum_{t=r+1}^{N}X_{t-1}V_t\right)D(\psi,F)$ converges in probability to zero because of A3 and Lemma 2.1. The theorem follows from Lemma 5.1.

UNIVERSITY OF CALIFORNIA, BERKELEY

REFERENCES

[1] Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, John Wiley & Sons.
[2] Bickel, P. J. (1975). One-step Huber estimates in the linear model, *J. Amer. Statist. Ass.*, **70**, 428-434.
[3] Brown, B. M. (1971). Martingale central limit theorems, *Ann. Math. Statist.*, **42**, 59-66.
[4] Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Statist.*, **35**, 73-101.
[5] Jaeckel, L. A. (1971). Some flexible estimates of location, *Ann. Math. Statist.*, **42**, 1540-1552.
[6] LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses, *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, **1**, 129-156.
[7] Mann, H. B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations, *Econometrica*, **11**, 217-226.