The Annals of Statistics
1982, Vol. 10, No. 2, 442–453

RECURSIVE COMPUTATION OF *M*-ESTIMATES FOR THE PARAMETERS OF A FINITE AUTOREGRESSIVE PROCESS

BY KATHERINE CAMPBELL

Los Alamos National Laboratory

Stochastic approximation methods are used to generate a sequence of "M-estimates" for the unknown parameters of an autoregressive process of known, finite order which may have heavy-tailed innovations. Weak dependence properties, which can be demonstrated for many autoregressive processes, are used in the proof that the sequence converges almost surely to the parameters. A brief Monte Carlo study verifies that bounded influence functions provide protection for recursive procedures against heavy-tailed innovations.

1. Introduction. The parameters $\boldsymbol{\theta}$ of a finite autoregressive process,

$$X_{k+1} = \theta_1 X_k + \cdots + \theta_q X_{k-q+1} + U_{k+1} = \langle \boldsymbol{\theta}, \pi \mathbf{X}_k \rangle + U_{k+1}, \tag{1.1}$$

where $\pi \mathbf{X}_k = (X_k, \cdots, X_{k-q+1})$ denotes the q most recent observations of the process, are usually fitted by least squares estimation, or else by solving the Yule-Walker equations using the least squares estimates of the autocorrelations. These two procedures are asymptotically equivalent, and are consistent even when the innovations U_k are non-Gaussian stable random variables with infinite variance. However, like least squares procedures generally, the traditional methods lose efficiency in the presence of a heavy-tailed innovation process.

This observation has led to the investigation of robust modifications of least squares, paralleling robust regression methods explored by Andrews (1974), Denby and Larsen (1977), and others. Instead of choosing $\hat{\boldsymbol{\theta}}$ to minimize

$$\textstyle\sum_{k=q}^{n-1} (x_{k+1} - \langle \hat{\boldsymbol{\theta}}, \pi \mathbf{x}_k \rangle)^2,$$

an M-estimate for $\boldsymbol{\theta}$ minimizes

$$\textstyle\sum_{k=0}^{n-1} \rho(x_{k+1} - \langle \hat{\boldsymbol{\theta}}, \pi \mathbf{x}_k \rangle),$$

where $\rho(t)$ is a function increasing less rapidly for large $|t|$ than does $\rho_{LS}(t) = t^2$. It is usually assumed that ρ has a derivative ψ, and the minimization problem thus becomes the problem of solving the "normal equations"

$$\textstyle\sum_{k=q}^{n-1} \psi(x_k - \langle \hat{\boldsymbol{\theta}}, \pi \mathbf{x}_k \rangle) \pi \mathbf{x}_k = \mathbf{0}.$$

Denby and Martin (1979) further allow the possibility of "influencing" the process $\{X_k\}$, so that the equation for their "generalized M-estimate" becomes

$$\textstyle\sum_{k=q}^{n-1} \psi(x_k - \langle \hat{\boldsymbol{\theta}}, \pi \mathbf{x}_k \rangle) \gamma(\pi \mathbf{x}_k) = \mathbf{0}. \tag{1.2}$$

The effect of this type of modification on the asymptotic variance of the estimator is very similar to the effect of robust M-estimation of location for i.i.d. observations whose common distribution F is that of the innovations U_k. In Beran's (1976) interesting modification of the M-estimator for the parameters of an autoregression, the data are used to select an influence function ψ from a subspace of $\mathscr{L}_2(F)$.

Received August 1980; revised November 1981.
AMS 1970 *subject classification.* Primary 62L12, 62M10; secondary 62G35, 62L20.
Key words and phrases. Robustness, stochastic approximation, weak dependence.

These procedures require one to have in hand all of the data to be used in estimating the parameters before starting. In order to update the estimate when new data are acquired, the entire calculation must be repeated. For many applications of time series this is not a satisfactory state of affairs. In some industrial situations a current estimate of the process is needed continuously as input to a control mechanism. In other situations where there is a great deal of data coming in, it simply becomes impractical to preserve all of the observations.

Recursive versions of least squares procedures take advantage of the fact that all of the information about past observations which is needed can be summarized in an estimated autocovariance matrix which is easily updated with each new observation. From the initial formulation of recursive least squares—ascribed to Plackett (1950)—through more recent work, these procedures rely implicitly on the linearity which has always made least squares computationally attractive. Even procedures described by Gardner (1964) and Davis and Koopmans (1973), which superficially resemble stochastic approximation, make use of the convergence of the estimated autocovariance function. An algorithm proposed by Saridis and Stein (1968) is a true stochastic approximation algorithm, but the proof of its convergence still requires linearity.

Unfortunately, no convenient summary of the past is available for the non-linear, robust analogs of least squares. For this reason we must use a stochastic approximation recursion of the form

$$\mathbf{T}_{n+1} = \mathbf{T}_n + a_n \mathbf{d}(\mathbf{T}_n; \pi\mathbf{X}_n, U_{n+1}), \tag{1.3}$$

where $a_n \to 0$ so that later estimates give more weight to the cumulative experience represented by $\mathbf{T}_n$ than to the newest information contained in the second term. $\mathbf{d}$ is chosen so that the function

$$\mathbf{D}(\mathbf{t}) = E\,\mathbf{d}(\mathbf{t}; \pi\mathbf{X}_n, U_{n+1}) \tag{1.4}$$

has a unique root at $\mathbf{t} = \boldsymbol{\theta}$. ($\mathbf{D}(\mathbf{t})$ does not depend on n for a strictly stationary process.)

Traditional proofs of convergence of stochastic approximation algorithms rest on the assumption that

$$E\{\mathbf{d}(\mathbf{T}_n; \pi\mathbf{X}_n, U_{n+1}) \mid X_1, \cdots, X_n\} = \mathbf{D}(\mathbf{T}_n). \tag{1.5}$$

However, this assumption fails when $\mathbf{d}$ has the form given in (1.3), because the conditional expected value depends on the past explicitly through $X_{n-q+1}, \cdots, X_n$ as well as through $\mathbf{T}_n$. That (1.5) is not necessary was noted by Dvoretzky (1956), who replaced it with the condition that

$$\sum a_n E\{\langle \boldsymbol{\theta} - \mathbf{T}_n, \mathbf{d}(\mathbf{T}_n; \pi\mathbf{X}_n, U_{n+1}) - \mathbf{D}(\mathbf{T}_n)\rangle \mid X_1, \cdots, X_n\}$$

be uniformly bounded and convergent. Close examination of Blum's (1954) proof of almost sure convergence shows that in fact

$$\sum a_n E\{\langle \boldsymbol{\theta} - \mathbf{T}_n, \mathbf{d}(\mathbf{T}_n; \pi\mathbf{X}_n, U_{n+1}) - \mathbf{D}(\mathbf{T}_n)\rangle\} < \infty \tag{1.6}$$

is sufficient. Therefore the problem is to find conditions weaker than (1.5) under which (1.6) holds.

The required conditions make use of a type of "mixing" or weak dependence of the process $\{X_k\}$ described Section 3. In Section 4 we prove a lemma required to carry out the proof of almost sure convergence of $\mathbf{T}_n$ to $\boldsymbol{\theta}$, which is completed in Section 5. In Section 6 some additional results are stated without proof, and the results of a simulation experiment are described briefly. In the following section the algorithm, the necessary conditions and the main result are set forth.

2. Statement of results. By analogy with (1.2), the function $\mathbf{d}$ in (1.3) is taken to be

$$\mathbf{d}(\mathbf{T}_n; \pi\mathbf{X}_n, U_{n+1}) = \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle),$$

where $\pi\mathbf{X}_n = (X_n, \cdots, X_{n-q+1})$ is the vector of the q most recent observation and $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the ℓ_2-inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{q} x_i y_i.$$

Explicitly, (1.3) becomes

$$\mathbf{T}_{n+1} = \mathbf{T}_n + a_n \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n \rangle). \tag{2.1}$$

Estimation begins at time zero, when there are $n_0 \geq 0$ observations already available. If $n_0 < q$, it is convenient to start things off by assuming that $X_{-q+1} = \cdots = X_{-n_0} = EX_n = 0$. The initial estimate $\mathbf{T}_0$ is a function of these n_0 observations.

Let $\mathbf{X}_n = (X_{-n_0+1}, \cdots, X_n)$ be the vector of $n_0 + n$ observations up to time $n \geq 0$, and let $\pi\mathbf{X}$ (with some abuse of notation) denote a random vector with the distribution of q consecutive variables of the stationary process.

Let $\Psi(a) = E_F\psi(U + a)$ and $\Psi_2(a) = E_F\psi^2(U + a)$, where F is again the common distribution of the innovations U_k in (1.1). Since U_{n+1} is independent of $\mathbf{X}_n$, while $\mathbf{T}_n$ and $\pi\mathbf{X}_n$ are functions of $\mathbf{X}_n$,

$$\begin{aligned} E\{\gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n \rangle) \,|\, \mathbf{X}_n\} &= \gamma(\pi\mathbf{X}_n)E\{\psi(U_{n+1} + \langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n \rangle) \,|\, \mathbf{X}_n\} \\ &= \gamma(\pi\mathbf{X}_n)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n \rangle). \end{aligned} \tag{2.2}$$

Similarly

$$E\{|\gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n \rangle)|^2 \,|\, \mathbf{X}_n\} = |\gamma(\pi\mathbf{X}_n)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n \rangle). \tag{2.3}$$

where $|\mathbf{x}|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$.

If $\mathbf{X}$ is a random vector on $(\Omega, \mathcal{B}, P)$ with values in $\mathcal{R}^q$, we will say that $\mathbf{X} \in \mathcal{K}_p(\Omega, \mathcal{B}, P)$ if the random variable $|\mathbf{X}|$ belongs to $\mathcal{L}_p(\Omega, \mathcal{B}, P)$. In this case, define

$$\|\mathbf{X}\|_p = \left(\int |\mathbf{X}|^p \, dP \right)^{1/p}.$$

$\|\cdot\|_p$ defines a norm on $\mathcal{K}_p$ provided that, as usual, we identify random vectors which are almost surely equal $[P]$. If $\mathbf{X} \in \mathcal{K}_p$ and $\mathbf{Y} \in \mathcal{K}_r$, where $1/p + 1/r = 1$, then

$$E|\langle \mathbf{X}, \mathbf{Y} \rangle| \leq E(|\mathbf{X}||\mathbf{Y}|) \leq \|\mathbf{X}\|_p \|\mathbf{Y}\|_r. \tag{2.4}$$

The set of conditions given below could undoubtedly be weakened to accommodate a variety of algorithms. In particular, some of the moment conditions might be weakened by working, as Blum does, with a real-valued function defined on $\mathcal{R}^q$ with continuous first and second partial derivatives other than the function $f(\mathbf{x}) = |\mathbf{x} - \boldsymbol{\theta}|^2$.

CONDITIONS.

A1: The function ψ satisfies $|\psi(x)| \leq M_\psi < \infty$ for all x.

A2: $\gamma(\mathbf{x})$ is of the form $\mathbf{x}g(\mathbf{x})$ for some non-negative function g of $\mathbf{x}$.

A3: $\|\gamma(\pi\mathbf{X})\|_r = M_\gamma < \infty$ for some $r > 2$.

A4: For $s = 2r/(r-2)$, with r as in A3, $(\int |\langle \pi\mathbf{X}, \gamma(\pi\mathbf{X}) \rangle|^s \, dP)^{1/s} = M_p < \infty$.

A5: $\|\mathbf{T}_0 - \boldsymbol{\theta}\|_2 < \infty$.

A6: There exists $\alpha > 0$ such that for all a and Δa, $|\Psi(a + \Delta a) - \Psi(a)| \leq \alpha|\Delta a|$.

A7: $a\Psi(a) \geq 0$ for all a.

A8: For any $\varepsilon > 0$, there exists $L_\varepsilon > 0$ such that $\inf_{|\theta - \mathbf{t}| \geq \varepsilon} \langle \boldsymbol{\theta} - \mathbf{t}, \mathbf{D}(\mathbf{t}) \rangle \geq L_\varepsilon$.

A9: $a_n > 0$, $\sum a_n = \infty$ and $\sum a_n^2 < \infty$.

It is further assumed that there exists a function N mapping the non-negative integers into themselves which satisfies the following conditions.

A10: $0 \leq N(n) \leq n$ for all $n \geq 0$.
A11: $\sum_n a_n \beta(N(n) - q + 1)^{1/2-1/r} < \infty$, where β is the "strong regularity" coefficient which will be defined in Section 3.
A12: $\sum_n a_n \sum_{k=n-N(n)}^{n-1} a_k < \infty$.

Conditions such as A7, A8 and A9 are part of every stochastic approximation algorithm. Condition A11 implies that $N(n) \to \infty$ as $n \to \infty$ and that the process is strongly regular. $N(n)$ must increase sufficiently fast to guarantee A11, yet slowly enough that A12 also holds. A traditional choice for the a_n is the harmonic series $a_n = 1/(n+1)$. If $N(n) = [\sqrt{n}]$, i.e. the largest integer which is less than or equal to $\sqrt{n}$, then A12 holds. If $\beta(n) = O(n^{\lambda} c^{-n})$, then for $\varepsilon > 0$,

$$a_n \beta(N(n) - q + 1)^{\varepsilon} = O(n^{\lambda'} d^{-\sqrt{n}})$$

for some λ', sufficiently large but finite, and the series of these terms also converges.

If $r \leq 4$, then $2r/(r-2) \geq 4$, and conversely, so A3 and A4 require the existence of fairly high-order moments.

The principal result of this paper is the following theorem.

THEOREM. *Under conditions* A1-A12, *the sequence of estimators defined by* (2.1) *converges almost surely to* $\boldsymbol{\theta}$.

3. Strong regularity. In this section only, $\{X_n\}$ is not necessarily an autoregressive process as in (1.1), but strict stationarity will be assumed. Let $\cdots, X_n, X_{n+1}, \cdots$ be a strictly stationary process of random variables defined on a probability space $(\Omega, \mathscr{B}, P)$, and let $\mathscr{B}_s^t$ be the sub-σ-algebra of $\mathscr{B}$ determined by $\{X_n, s \leq n \leq t\}$.

Various "mixing" conditions for stochastic processes, that is, conditions on the weakening of the dependence between two variables of the process, X_p and X_q, as $|p - q| \to \infty$, have been proposed in the literature. The condition called "strong regularity" below appears to have been suggested by Kolmogorov, and is investigated in detail by Volkonski and Rozanov (1959, 1961). It is stronger than "strong mixing"—introduced by Rosenblatt (1956) and also called "complete regularity"—but considerably weaker than "uniform mixing" (also called "ϕ-mixing"), which is not satisfied by many stochastic processes of interest (see Ibragimov, 1962, or Gastwirth and Rubin, 1975). Somewhat like uniform mixing, however, the coefficient $\beta(n)$ defined below is a measure of the difference between the probability of a set in $\mathscr{B}_{m+n}^{\infty}$ and its probability conditioned on $\mathscr{B}_{-\infty}^{m}$. Intuitively, weak dependence should mean that this difference goes to zero in some sense as $n \to \infty$.

Strong regularity can be defined for non-stationary processes, but here only strictly stationary processes are considered.

DEFINITION (Volkonski and Rozanov, 1959). A strictly stationary process $\{X_n\}$ is *strongly regular* if $\lim_{n\to\infty} \beta(n) = 0$, where

$$\beta(n) = E[\mathrm{Var}_{A \in \mathscr{B}_{m+n}} \{P(A \mid \mathscr{B}_{-\infty}^{m}) - P(A)\}]. \tag{3.1}$$

The quantity in parentheses is the total variation of the signed measure $P(\cdot \mid \mathscr{B}_{-\infty}^{m}) - P(\cdot)$ restricted to the sub-σ-algebra $\mathscr{B}_{m+n}^{\infty}$. This is a $\mathscr{B}_{-\infty}^{m}$-measurable random variable, and $\beta(n)$ is its expected value. $\beta(n)$ is independent of m by stationarity.

Conditions for a process to be strongly regular are given by Ibragimov and Solev (1969) and by Gastwirth and Rubin (1975). In particular, Gastwirth and Rubin show that first-order autoregressive processes with normal or Laplace innovations are strongly regular with $\beta(n) = O(c^{-n})$, while those with Cauchy innovations satisfy $\beta(n) = O(nc^{-n})$. These

results can be extended to autoregressions of any finite order, and by the remarks following the conditions given in Section 2, A11 will be satisfied for such processes if $N(n) = [\sqrt{n}]$.

Let $\pi_{m,n}$ denote the direct product of P restricted to $\mathscr{B}^m_{-\infty}$ with P restricted to $\mathscr{B}^\infty_{m+n}$.

PROPOSITION 3.1. $\beta(n) = \mathrm{Var}_{C\in\mathscr{B}^m_{-\infty}\times\mathscr{B}^\infty_{m+n}}\{P(C) - \pi_{m,n}(C)\}$.

PROOF. See Volkonski and Rosanov (1961, Section 4).

Now let $Y \in \mathscr{R}^k$ and $W \in \mathscr{R}^\ell$ be two vectors of random variables defined on $(\Omega, \mathscr{B}, P)$, where Y is $\mathscr{B}^m_{-\infty}$-measurable and W is $\mathscr{B}^\infty_{m+n}$-measurable. Let $\hat{P}$ denote the joint distribution of (Y, W) on $\mathscr{R}^k \times \mathscr{R}^\ell = \mathscr{R}^{k+\ell}$, and let $\hat{P}_Y$ and $\hat{P}_W$ be the marginal distributions of Y and W on $\mathscr{R}^k$ and $\mathscr{R}^\ell$, respectively. Then $S = \hat{P} - \hat{P}_Y \times \hat{P}_W$ is a finite signed measure on the Borel sets $\mathscr{A}^{k+\ell}$ in $\mathscr{R}^{k+\ell}$. Let $E^+ \cup E^- = \mathscr{R}^{k+\ell}$ be a Hahn decomposition for S. If $G \in \mathscr{A}^{k+\ell}$, then

$$G_y = \{w: (y, w) \in G\} \in \mathscr{A}^\ell.$$

Write

$$S(G) = \int Q(G, y)\, d\hat{P}_Y(y), \tag{3.2}$$

where

$$Q(G, y) = \hat{P}(G \mid Y = y) - \hat{P}_W(G_y). \tag{3.3}$$

$\hat{P}(G \mid Y = y)$ is a regular conditional probability of G given $Y = y$, so Q is a signed measure on $(\mathscr{R}^{k+\ell}, \mathscr{A}^{k+\ell})$ for fixed y, and a measurable function of y for fixed G. Define $Q^+(G, y) = Q(G \cap E^+, y)$ and $Q^-(G, y) = Q(G \cap E^-, y)$, and set

$$|Q|(G, y) = Q^+(G, y) - Q^-(G, y).$$

The total variation of the signed measure S is given by

$$|S|(G) = \int |Q|(G, y)\, d\hat{P}_Y(y). \tag{3.4}$$

PROPOSITION 3.2. $|S|(\mathscr{R}^{k+\ell}) \le \beta(n)$.

PROOF. $|S|(\mathscr{R}^{k+\ell})$ is the supremum over all Borel sets G of $\mathscr{R}^{k+\ell}$ of

$$|S(G)| = |\hat{P}(G) - \hat{P}_Y \times \hat{P}_W(G)|.$$

By definition,

$$\hat{P}(G) = P\{\omega \in \Omega: (Y(\omega), W(\omega)) \in G\}.$$

It is also easily verified that

$$(\hat{P}_Y \times \hat{P}_W)(G) = \pi_{m,n}\{\omega \in \Omega: (Y(\omega), W(\omega)) \in G\}.$$

As sets of the form $\{\omega: (Y(\omega), W(\omega)) \in G\}$ are a subcollection of the sets in $\mathscr{B}^m_{-\infty} \times \mathscr{B}^\infty_{m+n}$,

$$|S|(\mathscr{R}^{k+\ell}) \le \sup_{C\in\mathscr{B}^m_{-\infty}\times\mathscr{B}^\infty_{m+n}} |P(C) - \pi_{m,n}(C)| \le \mathrm{Var}_{C\in\mathscr{B}^m_{-\infty}\times\mathscr{B}^\infty_{m+n}}\{P(C) - \pi_{m,n}(C)\} = \beta(n).$$

In the proof of the lemma of Section 4, we will need a bound on $\| \mathbf{H} \|_2$, where $\mathbf{H}$ and $\mathbf{h}$ are q-vector valued functions related by

$$\mathbf{H}(y) = E_{\hat{P}}\{\mathbf{h}(Y, W) \mid Y = y\} - E_{\hat{P}_W}\{\mathbf{h}(y, W)\} = \int h(y, w)\, dQ. \tag{3.5}$$

PROPOSITION 3.3 *Let* $\mathbf{H}(y)$ *be defined as in* (3.5), *where* $\mathbf{h}$ *belongs to* $\mathscr{K}_r(\mathscr{R}^{k+\ell}, \mathscr{A}^{k+\ell}, |S|)$ *for some* $r \geq 2$. *Then* $\mathbf{H}$ *belongs to* $\mathscr{K}_2(\mathscr{R}^k, \mathscr{A}^k, \hat{P}_Y)$ *and*

$$\|\mathbf{H}\|_2 \leq \|\mathbf{h}\|_r \{\beta(n)\}^{1/2-1/r}.$$

PROOF. Using Jensen's inequality,

$$|\mathbf{H}(y)|^2 \leq \int |\mathbf{h}(y, w)|^2 \, d|Q| \quad \text{a.s.} \quad [\hat{P}_Y]. \tag{3.6}$$

Therefore

$$\|\mathbf{H}\|_2^2 \leq \int\int |\mathbf{h}(y, w)|^2 \, d|Q| \, d\hat{P}_Y = \int |\mathbf{h}(y, w)|^2 \, d|S| = \|\mathbf{h}\|_2^2. \tag{3.7}$$

The right-hand side of (3.7) is finite, as $\|\mathbf{h}\|_2^2 \leq \|\mathbf{h}\|_r^2 < \infty$. So $\mathbf{H} \in \mathscr{K}_2(\mathscr{R}_k, \mathscr{A}_k, \hat{P}_Y)$, and (3.7) completes the proof of the theorem for $r = 2$. If $r = \infty$, then from (3.6),

$$\|\mathbf{H}\|_2^2 \leq \|\mathbf{h}\|_\infty^2 \int d|Q| \, d\hat{P}_Y \leq \|\mathbf{h}\|_\infty^2 \beta(n),$$

using (3.4) and Proposition 3.2. If $2 < r < \infty$, set $1/t = 1 - 2/r > 0$, so that $1/r + 1/2t = 1$, and apply Hölder's inequality to (3.6) to get

$$|\mathbf{H}(y)|^2 \leq \left(\int |\mathbf{h}(y, z)|^r \, d|Q|\right)^{2/r} \left(\int d|Q|\right)^{1/t}. \tag{3.8}$$

Apply Hölder's inequality again to the integral of the product on the right-hand side of (3.8) with respect to $\hat{P}_Y$ to get

$$\|\mathbf{H}\|_2^2 \leq \left(\int\int |\mathbf{h}(y, w)|^r \, d|Q| \, d\hat{P}_Y\right)^{2/r} \left(\int\int d|Q| \, d\hat{P}_Y\right)^{1/t} \leq \|\mathbf{h}\|_r^2 \{\beta(n)\}^{1/t}. \tag{3.9}$$

As $1/2t = 1/2 - 1/r$, taking square roots in (3.9) completes the proof.

4. Lemma. The lemma of this section verifies that Conditions A1–A12 (together with one additional condition which will be independently established at the beginning of the proof of the theorem in Section 5) are sufficient to guarantee (1.6) when $\mathbf{T}_n$ is computed using the recursive formulas (2.1). The procedure is to rewrite the sum in (1.6) as

$$\textstyle\sum_{n=1}^{\infty} a_n E \langle \boldsymbol{\theta} - \mathbf{T}_{n-N(n)}, \mathbf{d}(\mathbf{T}_{n-N(n)}; \pi\mathbf{X}_n, U_{n+1}) - \mathbf{D}(\mathbf{T}_{n-N(n)}) \rangle + \sum_{n=1}^{\infty} a_n \sum_{k=n-N(n)}^{n-1} a_k E Z_k.$$

If EZ_k is bounded uniformly in k, then the second sum converges by A12. In the first sum, we will form the expectation of the nth term conditioned on the past up to the $n - N(n)$th observation to get

$$\textstyle\sum_{n=1}^{\infty} a_n E\langle \boldsymbol{\theta} - \mathbf{T}_{n-N(n)}, E\{\mathbf{d}(\mathbf{T}_{n-N(n)}; \pi\mathbf{X}_n, U_{n+1}) \,|\, \mathbf{X}_{n-N(n)}\} - \mathbf{D}(\mathbf{T}_{n-N(n)}) \rangle. \tag{4.1}$$

Because $\mathbf{X}_n$ is only weakly dependent on the distant past, up to the $n - N(n)$th observation,

$$E\{\mathbf{d}(\mathbf{T}_{n-N(n)}; \pi\mathbf{X}_n, U_{n+1}) \,|\, \mathbf{X}_{n-N(n)}\} \to \mathbf{D}(\mathbf{T}_{n-N(n)})$$

sufficiently rapidly that the sum (4.1) converges, even though $\sum a_n$ diverges.

This type of solution was proposed in some rather obscure work of Sakrison (1962, 1964, 1967), but in order to implement it he was obliged to assume that the function ψ increases linearly (or faster) with its argument, whereas we wish to consider bounded functions. Sakrison's mixing condition will here be replaced by strong regularity, specifically by A11.

For this section only, it will be convenient to introduce some further abbreviations. Let

$$\mathbf{h}_{n,j}(\mathbf{X}_{n-j}, \pi\mathbf{X}_n) = \gamma(\pi\mathbf{X}_n)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_{n-j}, \pi\mathbf{X}_n \rangle) = E\{\gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_{n-j}, \pi\mathbf{X}_n \rangle) \,|\, \mathbf{X}_n\}$$

for $j \geq 0$. If $Q_{n,j}$ is the signed conditional measure developed in Section 3, where $Y = \mathbf{X}_{n-j}$

and $W = \pi \mathbf{X}_n$, then let

$$\mathbf{H}_{n,j}(y) = \int \mathbf{h}_{n,j}(y,w)\, dQ_{n,j} = E\{\mathbf{h}_{n,j}(\mathbf{X}_{n-j}, \pi\mathbf{X}_n) \mid \mathbf{X}_{n-j} = y\} - E\mathbf{h}_{n,j}(y, \pi\mathbf{X}_n)$$

$$= E\{\gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_{n-j}, \pi\mathbf{X}_n\rangle) - \mathbf{D}(\mathbf{T}_{n-j}) \mid \mathbf{X}_{n-j} = y\}. \tag{4.2}$$

The random vectors $Y = \mathbf{X}_{n-j}$ and $W = \pi\mathbf{X}_n$ are separated by $j - q + 1$ indices in the process $\{X_n\}$. If $|\mathbf{h}_{n,j}(z)|$ possesses an rth moment $\|\mathbf{h}_{n,j}\|_r$ with respect to $|S_{n,j}|$ for some $r > 2$, where $|S_{n,j}|$ is the total variation of $\hat{P} - \hat{P}_Y \times \hat{P}_W$ for the above choices of Y and W, then

$$\|\mathbf{H}_{n,j}\|_2 \le \|\mathbf{h}_{n,j}\|_r \{\beta(j - q + 1)\}^{1/2-1/r} \tag{4.3}$$

by Proposition 3.3. Under Conditions A1–A12

$$\|\mathbf{h}_{n,m}\|_r \le M_\gamma M_\psi. \tag{4.4}$$

Iterating (2.1) gives

$$\mathbf{T}_n = \mathbf{T}_{n-m} + \textstyle\sum_{k=n-m}^{n-1} a_k \gamma(\pi\mathbf{X}_k)\psi(X_{k+1} - \langle \mathbf{T}_k, \pi\mathbf{X}_k\rangle). \tag{4.5}$$

LEMMA 4.1. *If Conditions* A1–A12 *hold and if* $\|\mathbf{T}_n - \boldsymbol{\theta}\|_2$ *is uniformly bounded in* n, *then the series* (1.6) *converges absolutely, where* $\mathbf{T}_n$ *is computed by the algorithm* (2.1).

PROOF. We begin by expanding the nth term in the series (1.6):

$$\begin{aligned} &E\{a_n\langle \mathbf{T}_n - \boldsymbol{\theta}, \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle) - \mathbf{D}(\mathbf{T}_n)\rangle\} \\ &= E\{a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle) - \mathbf{D}(\mathbf{T}_n)\rangle\} \\ &\quad + E\{a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k\psi(X_{k+1} - \langle \mathbf{T}_k, \pi\mathbf{X}_k\rangle)\langle \gamma(\pi\mathbf{X}_k), \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle) \\ &\quad - \mathbf{D}(\mathbf{T}_n)\rangle\} \\ &= W_1 + W_2, \end{aligned} \tag{4.6}$$

where (4.5) has been used to iterate back to $n - N(n)$.

The second term W_2 can be evaluated by taking the conditional expected value given $\mathbf{X}_n$. As everything in that sum is a function of $\mathbf{X}_n$, this merely allows us to use (4.2) to write

$$W_2 = E\{a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k\psi(X_{k+1} - \langle \mathbf{T}_k, \pi\mathbf{X}_k\rangle)\langle \gamma(\pi\mathbf{X}_k), \mathbf{H}_{n,0}(\mathbf{X}_n)\rangle\}.$$

Then

$$\begin{aligned} |W_2| &\le M_\psi a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k E|\langle \gamma(\pi\mathbf{X}_k), \mathbf{H}_{n,0}(\mathbf{X}_n)\rangle|, && \text{by A1,} \\ &\le M_\psi a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k \|\gamma(\pi\mathbf{X}_k)\|_2 \|\mathbf{h}_{n,0}\|_2, && \text{by (2.4) and (4.3),} \\ &\le M_\psi^2 M_\gamma^2 a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k, && \text{by (4.4) and A3.} \end{aligned} \tag{4.7}$$

Rewrite W_1 as

$$\begin{aligned} W_1 &= E[a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, E\{\gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle) \mid \mathbf{X}_n\} - \mathbf{D}(\mathbf{T}_n)\rangle] \\ &= E(a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{D}_1 + \mathbf{D}_2 + \mathbf{D}_3\rangle), \end{aligned}$$

where

$$\begin{aligned} \mathbf{D}_1 &= \gamma(\pi\mathbf{X}_n)\{\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle) - \Psi(\langle \boldsymbol{\theta} - \mathbf{T}_{n-N(n)}, \pi\mathbf{X}_n\rangle)\}, \\ \mathbf{D}_2 &= \gamma(\pi\mathbf{X}_n)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_{n-N(n)}, \pi\mathbf{X}_n\rangle) - \mathbf{D}(\mathbf{T}_{n-N(n)}), \\ \mathbf{D}_3 &= \gamma(\mathbf{T}_{n-N(n)}) - \mathbf{D}(\mathbf{T}_n). \end{aligned} \tag{4.8}$$

Since

$$\begin{aligned}|\mathbf{D}_1| &\le \alpha|\gamma(\pi\mathbf{X}_n)||\langle \mathbf{T}_n - \mathbf{T}_{n-N(n)}, \pi\mathbf{X}_n\rangle|, \quad \text{by A6,}\\ &\le \alpha|\gamma(\pi\mathbf{X}_n)||\pi\mathbf{X}_n||\mathbf{T}_n - \mathbf{T}_{n-N(n)}|\\ &\le \alpha M_\psi \textstyle\sum_{k=n-N(n)}^{n-1} a_k |\langle \pi\mathbf{X}_n, \gamma(\pi\mathbf{X}_n)\rangle||\gamma(\pi\mathbf{X}_k)|\end{aligned}$$

by (4.5), A1 and A2, we have

$$\|\mathbf{D}_1\|_2 \le \alpha M_\psi \textstyle\sum_{k=n-N(n)}^{n-1} a_k \left(\int |\gamma(\pi\mathbf{X}_k)|^2 |\langle \pi\mathbf{X}_n, \gamma(\pi\mathbf{X}_n)\rangle|^2 \, dP_n\right)^{1/2},$$

where P_n is the distribution of $\mathbf{X}_n$. As $2/r + 2/s = 1$, Hölder's inequality applied to the integral above leads to

$$\|\mathbf{D}_1\|_2 \le \alpha M_\psi M_\gamma M_p \textstyle\sum_{k=n-N(n)}^{n-1} a_k,$$

by A3 and A4. Thus

$$\begin{aligned}(4.9)\qquad |E\{a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{D}_1\rangle\}| &\le a_n\|\mathbf{T}_{n-N(n)} - \boldsymbol{\theta}\|_2 \|\mathbf{D}_1\|_2\\ &\le \alpha M_\psi M_\gamma M_p a_n \textstyle\sum_{k=n-N(n)}^{n-1} a_k \|\mathbf{T}_{n-N} - \boldsymbol{\theta}\|_2.\end{aligned}$$

Similarly

$$\begin{aligned}|\mathbf{D}_3| &\le E[|\gamma(\pi\mathbf{X})\{\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_{n-N(n)}, \pi\mathbf{X}\rangle) - \Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}\rangle)\}|]\\ &\le E\{\alpha|\gamma(\pi\mathbf{X})||\langle \mathbf{T}_{n-N(n)} - \mathbf{T}_n, \pi\mathbf{X}\rangle|\} \le \alpha M_\psi M_p \textstyle\sum_{k=n-N(n)}^{n-1} a_k |\gamma(\pi\mathbf{X}_k)|,\end{aligned}$$

so that the right-hand side of (4.9) bounds $|E(a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{D}_3\rangle)|$ as well. Finally

$$\begin{aligned}E(a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{D}_2\rangle) &= E\{a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, E(\mathbf{D}_2|\mathbf{X}_{n-N(n)})\rangle\}\\ &= E\{a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{H}_{n,N(n)}(\mathbf{X}_{n-N(n)})\rangle\}\end{aligned}$$

by (4.2), and

$$\begin{aligned}(4.10)\qquad |E(a_n\langle \mathbf{T}_{n-N(n)} - \boldsymbol{\theta}, \mathbf{D}_2\rangle)| &\le a_n\|\mathbf{T}_{n-N(n)} - \boldsymbol{\theta}\|_2 \|\mathbf{h}_{n,N(n)}\|_r \beta(N(n) - q + 1)^{1/2-1/r}\\ &\le a_n\|\mathbf{T}_{n-N(n)} - \boldsymbol{\theta}\|_2 M_\psi M_\gamma \beta(N(n) - q + 1)^{1/2-1/r}\end{aligned}$$

by (2.4), (4.3) and (4.4).

Combining (4.7), (4.9) and (4.10) and using the hypothesis that $\|\mathbf{T}_n - \boldsymbol{\theta}\|_2$ is uniformly bounded, we get a bound on the absolute value of the nth term in the series (1.6) of the form

$$a_n\{C_1\beta(N(n) - q + 1)^{1/2-1/r} + C_2 \textstyle\sum_{k=n-N(n)}^{n-1} a_k\}.$$

Thus, by A11 and A12, the series is absolutely summable.

5. Almost sure convergence. The proof of almost sure convergence of $\mathbf{T}_n$ to $\boldsymbol{\theta}$ follows that of Blum (1954), whose work appears to have been the first on stochastic approximation procedures to make use of the martingale-like properties of sequences of stochastic approximation estimators. These properties were formalized by Robbins and Siegmund (1971).

LEMMA 5.1. *Let $(\Omega, \mathcal{B}, P)$ be a probability space, and let $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots$ be an increasing sequence of sub-σ-algebras of $\mathcal{B}$. For each $n = 1, 2, \cdots$ let Z_n, α_n, ζ_n and ξ_n be non-negative, $\mathcal{B}_n$-measurable random variables such that*

$$E(Z_{n+1}|\mathcal{B}_n) = Z_n(1 + \alpha_n) + \xi_n - \zeta_n.$$

Then $\lim_{n\to\infty} Z_n$ *exists and is finite, and* $\sum \zeta_n < \infty$ a.s. *on* $C = \{\sum \alpha_n < \infty, \sum \xi_n < \infty\}$.

The sequence Z_n is "almost" a super-martingale on the set C. Note also that the increasing series $\sum \alpha_n$ and $\sum \xi_n$ converge a.s. $[P]$ if $\sum E\alpha_n < \infty$ and $\sum E\xi_n < \infty$.

THEOREM. *Under Conditions* A1–A12, *the sequence of estimators defined by* (2.1) *converges almost surely to* $\boldsymbol{\theta}$.

PROOF. From (2.1) we get

$$|\mathbf{T}_{n+1} - \boldsymbol{\theta}|^2 = |\mathbf{T}_n - \boldsymbol{\theta}|^2 + 2a_n\langle \mathbf{T}_n - \boldsymbol{\theta}, \gamma(\pi\mathbf{X}_n)\psi(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle)\rangle + a_n^2|\gamma(\pi\mathbf{X}_n)|^2\psi^2(X_{n+1} - \langle \mathbf{T}_n, \pi\mathbf{X}_n\rangle). \tag{5.1}$$

Take the conditional expected value of each side, given $\mathbf{X}_n$, and use (2.2) and (2.3) to get

$$\begin{aligned} E(|\mathbf{T}_{n+1} - \boldsymbol{\theta}|^2 \,|\, \mathbf{X}_n) &= |\mathbf{T}_n - \boldsymbol{\theta}|^2 + 2a_n\langle \mathbf{T}_n - \boldsymbol{\theta}, \gamma(\pi\mathbf{X}_n)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle)\rangle \\ &\quad + a_n^2|\gamma(\pi\mathbf{X}_n)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle) \\ &= |\mathbf{T}_n - \boldsymbol{\theta}|^2 - \zeta_n + \xi_n \end{aligned} \tag{5.2}$$

in the notation of Lemma 4.1. Here

$$\zeta_n = 2a_n\langle \theta - \mathbf{T}_n, \gamma(\pi\mathbf{X}_n)\rangle\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle)$$

is non-negative by A2 and A7, while

$$\xi_n = a_n^2|\gamma(\pi\mathbf{X}_n)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle)$$

is bounded in expected value by $a_n^2M_\psi^2M_\gamma^2$, by A1 and A3, so $\sum E\xi_n < \infty$ by A9. By Lemma 4.1 and the remark which follows it,

$$\lim_{n\to\infty}|\mathbf{T}_n - \boldsymbol{\theta}|^2 \text{ exists and is finite a.s.} \tag{5.3}$$

Taking the expected value of each side in (5.2) gives

$$\|\mathbf{T}_{n+1} - \boldsymbol{\theta}\|_2^2 = \|\mathbf{T}_n - \boldsymbol{\theta}\|_2^2 - 2a_nE\langle \boldsymbol{\theta} - \mathbf{T}_n, \gamma(\pi\mathbf{X}_n)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle)\rangle + a_n^2E|\gamma(\pi\mathbf{X}_n)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_n, \pi\mathbf{X}_n\rangle). \tag{5.4}$$

This is bounded above by $\|\mathbf{T}_n - \boldsymbol{\theta}\|_2^2 + a_n^2M_\psi^2M_\gamma^2$, since the second term is non-positive. Iterating back to $n = 0$ gives

$$\|\mathbf{T}_{n+1} - \boldsymbol{\theta}\|_2^2 \le \|\mathbf{T}_0 - \boldsymbol{\theta}\|_2^2 + \textstyle\sum_{k=0}^{\infty} a_k^2M_\psi^2M_\gamma^2 < \infty \tag{5.5}$$

by A5 and A9, so $\|\mathbf{T}_n - \boldsymbol{\theta}\|_2^2$ is bounded uniformly in n and Lemma 3.3 holds.

Next subtract and add the term $(2a_nE\langle \boldsymbol{\theta} - \mathbf{T}_n, \mathbf{D}(\mathbf{T}_n)\rangle)$ to the right-hand side of (5.4) and iterate back to $n = 0$ to get

$$\begin{aligned} 0 \le \|\mathbf{T}_{n+1} - \boldsymbol{\theta}\|_2^2 &= \|\mathbf{T}_0 - \boldsymbol{\theta}\|_2^2 - 2\textstyle\sum_{k=0}^{n} a_kE\langle \boldsymbol{\theta} - \mathbf{T}_k, \mathbf{D}(\mathbf{T}_k)\rangle \\ &\quad - 2\textstyle\sum_{k=0}^{n} a_kE\langle \boldsymbol{\theta} - \mathbf{T}_k, \gamma(\pi\mathbf{X}_k)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_k, \pi\mathbf{X}_k\rangle) - \mathbf{D}(\mathbf{T}_k)\rangle \\ &\quad + \textstyle\sum_{k=0}^{n} a_k^2E|\gamma(\pi\mathbf{X}_k)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_k, \pi\mathbf{X}_k\rangle). \end{aligned}$$

Rearrange this to get

$$\begin{aligned} \textstyle\sum_{k=0}^{n} a_kE\langle \boldsymbol{\theta} - \mathbf{T}_k, \mathbf{D}(\mathbf{T}_k)\rangle &\le \tfrac{1}{2}\|\mathbf{T}_0 - \boldsymbol{\theta}\|_2^2 \\ &\quad - \textstyle\sum_{k=0}^{n} a_kE\langle \boldsymbol{\theta} - \mathbf{T}_k, \gamma(\pi\mathbf{X}_k)\Psi(\langle \boldsymbol{\theta} - \mathbf{T}_k, \pi\mathbf{X}_k\rangle) - \mathbf{D}(\mathbf{T}_k)\rangle \\ &\quad + \tfrac{1}{2}\textstyle\sum_{k=0}^{n} a_k^2E|\gamma(\pi\mathbf{X}_k)|^2\Psi_2(\langle \boldsymbol{\theta} - \mathbf{T}_k, \pi\mathbf{X}_k\rangle). \end{aligned} \tag{5.6}$$

The second sum on the right-hand side of (5.6) converges by Lemma 4.1 (taking the expected value of the kth term conditioned on $\mathbf{X}_k$ inside the overall expected value does

not affect this result), while the third term is bounded above for all n as in the first part of the proof. Therefore the series on the left-hand-side, whose terms are non-negative by A8, must converge. Because $\sum a_n$ diverges, this implies that

$$\liminf E\,\langle \boldsymbol{\theta} - \mathbf{T}_n,\ \mathbf{D}(\mathbf{T}_n)\rangle = 0.$$

Let $\{n_k\}$ be a subsequence such that

$$\lim E\langle \boldsymbol{\theta} - \mathbf{T}_{n_k},\ \mathbf{D}(\mathbf{T}_{n_k})\rangle = 0.$$

As $\langle \boldsymbol{\theta} - \mathbf{T}_{n_k},\ \mathbf{D}(\mathbf{T}_{n_k})\rangle \geq 0$, this implies that

$$\langle \boldsymbol{\theta} - \mathbf{T}_{n_k},\ \mathbf{D}(\mathbf{T}_{n_k})\rangle \to_p 0,$$

and so there is a further subsequence $\{n'_k\}$ such that

$$\langle \boldsymbol{\theta} - \mathbf{T}_{n'_k},\ \mathbf{D}(\mathbf{T}_{n'_k})\rangle \to 0 \text{ a.s.}$$

By A8, this is possible only if $|\boldsymbol{\theta} - \mathbf{T}_{n'_k}| \to 0$ a.s. But $|\boldsymbol{\theta} - \mathbf{T}_{n'_k}|$ must converge a.s. to the finite limit whose existence was established in (5.3) and, consequently, this limit is almost surely zero. Thus the full sequence $|\boldsymbol{\theta} - \mathbf{T}_n|$ converges a.s. to zero and $\mathbf{T}_n \to \boldsymbol{\theta}$ a.s. This completes the proof.

6. Additional results and simulation. One way to weaken the strong moment conditions (in particular, A1 and A4) is to take advantage of a priori bounds on $|\boldsymbol{\theta}|$. For example, stationarity in (1.1) implies that $|\boldsymbol{\theta}| < 1$. When the algorithm (2.1) produces $\mathbf{T}_{n+1}$ such that $|\mathbf{T}_{n+1}|$ exceeds the prior bound, we can truncate $\mathbf{T}_{n+1}$ before continuing. With this modification, almost sure convergence can be shown under weaker conditions.

Campbell (1979) also considers the convergence of $\mathbf{T}_n$ to $\boldsymbol{\theta}$ in mean square. This requires one additional condition,

A13: For any $M > 0$, there exists a number $K_M > 0$ such that $\langle \mathbf{t} - \boldsymbol{\theta},\ \mathbf{D}(\mathbf{t})\rangle \geq K_M|\boldsymbol{\theta} - \mathbf{t}|^2$ for all $\mathbf{t}$ satisfying $0 \leq |\boldsymbol{\theta} - \mathbf{t}| \leq M$.

The simulation study reported in Campbell (1979) considered several algorithms:

(1) OLS: Ordinary least squares, a linear, non-recursive computation.

(2) RYW: Recursive solution of the Yule-Walker equations, in which the estimate of the covariance matrix is updated recursively after each observation (see Method II of Gardner, 1964.) The estimate was truncated, as discussed above.

(3) SA1: An algorithm meeting the conditions of the preceding sections,

$$\mathbf{T}_{n+1} = \mathbf{T}_n + a_n\gamma\left(\frac{\pi\mathbf{X}_n}{SX_0}\right)\psi\left(\frac{X_{n+1} - \langle \mathbf{T}_n,\ \pi\mathbf{X}_n\rangle}{SR_0}\right),$$

where

$$a_n = O\left(\frac{1}{n}\right), \qquad \gamma(\mathbf{x}) = \mathbf{x}g(|\mathbf{x}|), \qquad g(x) = \frac{x}{1 + (x/2.5)^2},$$

$$\psi(x) = \begin{cases} x & |x| < 2.5, \\ 2.5\ \operatorname{sgn}(x) & |x| \geq 2.5, \end{cases}$$

and SX_0 and SR_0 are scale estimates.

(4) SA2: Similar to SA1 except that ψ and g are interchanged. This choice of γ fails to meet Condition A4 when the innovations are Cauchy, but nevertheless the algorithm performs well there, again suggesting that the given moment conditions are too strong.

(5) LSA: A linearized version of the stochastic approximation algorithm (γ and ψ are identity functions), truncated.

Initial estimates of $\boldsymbol{\theta}$ and of scale are based on twenty observations.

These algorithms were applied to time series generated from the model

$$X_t = .5\,X_{t-1} + U_t,$$

TABLE 1.
Comparison of Algorithms

	N	algorithm	bias	variance	relative efficiency
Normal	125	OLS	−.0032	.0058	1.00
		RYW	−.0231	.0845	.69
		LSA	−.0171	.0053	1.10
		SA1	−.0078	.0063	.93
		SA2	−.0119	.0064	.91
	500	OLS	−.0035	.0016	1.00
		RYW	−.0084	.0025	.63
		LSA	−.0089	.0016	1.01
		SA1	−.0062	.0020	.79
		SA2	−.0079	.0018	.88
Contaminated Normal	125	OLS	−.0130	.0055	1.00
		RYW	−.0162	.0055	1.00
		LSA	−.0319	.0061	.91
		SA1	−.0127	.0044	1.26
		SA2	−.0125	.0033	1.64
	500	OLS	−.0062	.0013	1.00
		RYW	−.0070	.0014	.92
		LSA	−.0093	.0023	.57
		SA1	−.0056	.0013	1.04
		SA2	−.0033	.0009	1.47
Cauchy	125	OLS	−.0008	.0031	1.00
		RYW	−.2027	.4898	.006
		LSA	−.0307	.0351	.09
		SA1	−.0078	.0028	1.11
		SA2	−.0014	.0011	2.70
	500	OLS	−.0015	.0006	1.00
		RYW	−.2734	.5387	.001
		LSA	−.0182	.0145	.04
		SA1	−.0007	.0006	.98
		SA2	−.0004	.0002	3.13

TABLE 2.
Median Computation Times Compared to Least Squares

	N = 125	N = 500
OLS	1.00	1.00
RYW	3.40	3.47
LSA	3.48	3.54
SA	6.36	6.48

and it was assumed that the true order ($p = 1$) was known. The entries in Table 1 are averages over 300 blocks of data. The estimators $\hat{\mathbf{T}}_N$ are computed for series of lengths $N = 125$ and 500 (with an additional twenty observations at the beginning on which initial estimates are based.) The innovations U_t have the following symmetric distributions: Normal $N(0, 1)$; Contaminated Normal, $.9\ N(0, 1) + .1\ N(0, 25)$; Cauchy, density $1/\{\pi(1 + x^2)\}$.

The least squares estimate is known to converge even when the variance is infinite, and in fact is seen to perform creditably in all cases, as well as at low cost (Table 2). When there is no obstacle to storing all of the data and repeating the calculation each time an

estimate is required, there may well be no reason to look further. The recursive Yule-Walker algorithm appears to be a reasonable recursive alternative for the contaminated normal case, but converges much more slowly in the normal case and not at all when the innovations are Cauchy. On the other hand, the truncated linear stochastic approximation algorithm (LSA) converges slowly for contaminated normals and is very inefficient in the Cauchy case.

The second robust stochastic approximation algorithm handles all of the heavy-tailed cases, including the Cauchy innovations, very efficiently, although with rather large bias in some cases, and is adequate when the innovations are normal. SA1 is generally an improvement over the non-robust recursive procedures in the non-normal cases, but less spectacularly efficient than SA2, and frequently more biased as well. Its behavior might be improved by a different choice of the scaling constant (here 2.5). Unfortunately these two procedures are almost twice as expensive as the other two recursive procedures (Table 2), but in cases where protection against heavy-tailed innovations is desirable this may not be too high a price to pay.

REFERENCES

ANDREWS, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16** 523–531.

BERAN, R. (1976). Adaptive estimates for autoregressive processes. *Ann. Inst. Statist. Math.* **28** 77–89.

BLUM, J. R. (1954). Approximation methods which converge with probability one. *Ann. Math. Statist.* **25** 382–386.

CAMPBELL, K. (1979). Stochastic approximation procedures for mixing stochastic processes. Unpublished Ph.D. dissertation, The University of New Mexico.

DAVIS, H. T. and KOOPMANS, L. H. (1973). Adaptive prediction of stationary time series. *Sankhyā Ser. A* **35** 5–22.

DENBY, L. and LARSEN, W. A. (1977). Robust regression estimators compared via Monte Carlo. *Comm. Statist. A* **6** 335–362.

DENBY, L. and MARTIN, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *J. Amer. Statist. Assoc.* **74** 140–146.

DVORETZKY, A. (1956). On stochastic approximation. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** 39–56.

GARDNER, L. A. (1964). Adaptive predictors. *Trans. Third Prague Conf. on Information Theory, Statistical Decision Functions and Random Processes* 123–192.

GASTWIRTH, J. L. and RUBIN, H. (1975). The asymptotic distribution theory of the empiric CDF for mixing stochastic processes. *Ann. Statist.* **3** 809–824.

IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Theory Probability Appl.* **7** 349–382.

IBRAGIMOV, I. A. and SOLEV, V. N. (1969). A condition for regularity of a Gaussian stationary process. *Soviet Math. Doklady* **10** 371–375.

PLACKETT, R. L. (1950). Some theorems in least squares. *Biometrika* **37** 149–157.

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*. J. S. Rustagi, Ed. 233–257. Academic, New York.

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U.S.A.* **42** 43–47.

SAKRISON, D. J. (1962). Application of stochastic approximation methods to system optimization. Technical Report 391, M.I.T. Research Lab. Electronics.

SAKRISON, D. J. (1964). A continuous Kiefer-Wolfowitz procedure for random processes. *Ann. Math. Statist.* **35** 590–599.

SAKRISON, D. J. (1967). The use of stochastic approximation to solve the system identification problem. *IEEE Trans. Automat. Control* **AC-12** 563–567.

SARIDIS, G. N. and STEIN, G. (1968). Stochastic approximation algorithms for linear discrete-time system identification. *IEEE Trans. Automat. Control* **AC-13** 515–523 and 594–595.

VOLKONSKII, V. A. and ROSANOV, YU. A. (1959). Some limit theorems for random functions. I. *Theory Probability Appl.* **4** 178–197.

VOLKONSKII, V. A. and ROSANOV, YU. A. (1961). Some limit theorems for random functions. II. *Theory Probability Appl.* **6** 186–198.

STATISTICS GROUP (M.S. 600)
LOS ALAMOS NATIONAL LABORATORY
LOS ALAMOS, NM 87545