Recursive M-estimators of location and scale for dependent sequences

Recursive M-estimators of location and scale may be obtained via stochastic approximation algorithms. We consider the case when the observations can be described by a strictly stationary process satisfying certain strong mixing conditions and results on strong convergence are given. The asymptotic distributions of the estimators for sequences of independent observations are also discussed.
Recursive M-estimators
of location and scale
for dependent sequences

by

Jan-Eric Englund, Ulla Holst
and David Ruppert *

1986-9

November 1986

* The University of North Carolina
at Chapel Hill
Recursive M-estimators of location and scale may be obtained via stochastic approximation algorithms. We consider the case when the observations can be described by a strictly stationary process satisfying certain strong mixing conditions and results on strong convergence are given. The asymptotic distributions of the estimators for sequences of independent observations are also discussed.
1. INTRODUCTION.

P. Huber introduced the simultaneous M-estimates of location and scale, \( \eta \) and \( \sigma \), based on observations \( y_1, \ldots, y_n \), as a solution \((T_n, S_n)\) of

\[
\begin{align*}
\sum_{i=1}^{n} \psi(S_n^{-1}(y_i - T_n)) &= 0, \\
\sum_{i=1}^{n} \chi(S_n^{-1}(y_i - T_n)) &= 0,
\end{align*}
\]

where \( \psi \) and \( \chi \) are suitably chosen functions. In most cases \( \psi \) is an odd and \( \chi \) an even function. In particular he studied M-estimators generated by functions \( \psi \) and \( \chi \) of the form (Huber's Proposal 2)

\[
\begin{align*}
\psi(x) &= \text{sign}(x) \min(\|x\|, k), \\
\chi(x) &= \min(k^2, x^2) - \beta_k,
\end{align*}
\]

with \( \beta_k \) chosen to make \( E(\chi(z)) = 0 \) if the distribution of \( z \) is \( N(0,1) \). We refer to the books by Huber (1981) or Hampel et al (1986) for a review of the properties of the M-estimators. There it is proved that, if the observations are i.i.d. with a symmetric distribution, \( \psi \) is an odd and \( \chi \) an even function, it follows that

\[
\begin{align*}
\mathbf{N}(T_n - \eta) &\in \mathbb{R}N(0, \sigma^2E(\psi^2(z_1))/E(\psi'(z_1))^2), \\
\mathbf{N}(S_n - \sigma) &\in \mathbb{R}N(0, \sigma^2E(\chi^2(z_1))/E(z_1\chi'(z_1))^2),
\end{align*}
\]

where \( z_1 = \sigma^{-1}(y_1 - \eta) \). In this case \( T_n \) and \( S_n \) are asymptotically independent.

In real-time situations, where the estimate is updated when new observations are obtained, it is often preferable to use a recursive estimator. Martin and Masreliez (1975) pointed out the possibility of constructing recursive M-estimators using a stochastic approximation approach. The classical results for stochastic approximation algorithms
can be applied rather straightforwardly to investigate the asymptotic properties of recursive M-estimators when the observations are independent.

The behaviour of recursive M-estimators in dependent situations are less known. The pure location parameter case with m-dependent and strongly regular observations is studied in Holst (1980) and Holst (1984) respectively. For practical use some recursive estimator of scale must be constructed and coupled to the estimator of the location parameter. Recursive scale-estimators which are variants of the median absolute deviation are studied in Holst (1985).

A broader approach to the estimating problem is to construct recursive algorithms based on (1.1). In this paper we prove strong convergence of estimators of the form

\[
\begin{align*}
\eta_{n+1} &= \eta_n + (n+1)^{-1} \tilde{h}_n (\varphi \sigma_n^{-1} (y_{n+1} - \eta_n)), \\
\sigma_{n+1} &= \sigma_n + (n+1)^{-1} \tilde{h}_n \chi \sigma_n^{-1} (y_{n+1} - \eta_n), \\
\eta_0, \sigma_0, \varphi(1), \varphi(2) & \text{ arbitrary and finite,}
\end{align*}
\]

and mainly we discuss the following choice of \( H_n^{(1)} \) and \( H_n^{(2)} \):

\[
\begin{align*}
H_n^{(1)} &= (n^{-1} \sum_{i=1}^{n} \tilde{\psi} (\varphi_{i-1}^{-1} (y_i - \eta_{i-1})))^{-1} \\
H_n^{(2)} &= (n^{-1} \sum_{i=1}^{n} \tilde{\psi} (\varphi_{i-1}^{-1} (y_i - \eta_{i-1}) \chi (\varphi_{i-1}^{-1} (y_i - \eta_{i-1}))))^{-1}
\end{align*}
\]

With the notation \( \tilde{\psi} \) we mean \( \psi \) truncated above and below.

We consider the case when the observations \( \{y_i\}_{i=1}^\infty \) can be described by a strictly stationary process satisfying certain strong mixing conditions. For the analysis we assume that \( \psi \) and \( \chi \) satisfy some regularity conditions. These are introduced in Section 2.

Strong convergence of \( \eta_n \) and \( \sigma_n \) and also of the adaptive sequences \( H_n^{(1)} \) and \( H_n^{(2)} \) is proved in Section 3.
In Englund, Holst and Ruppert (1987) we prove a strong representation theorem for the estimators. It is possible to derive asymptotic distributions using this theorem together with suitable forms of the Central Limit Theorem. When the observations are a sequence of i.i.d. variables it follows that \((\gamma_n, \sigma_n)\) has the same asymptotic distribution as the nonrecursive estimator \((T_n, S_n)\). Comments on the asymptotic distribution is given in Section 4. Further we discuss whether our choice of \(H_n^{(1)}\) and \(H_n^{(2)}\) is optimal or if it is possible to find a better one. We consider a gain matrix which might be preferred, but this matrix contains unknown parameters which like \(a\) and \(b\) must be estimated, and this leads to an expansion of the dimension of the parameter.

In Section 5 we illustrate the behaviour of the estimates for Huber's Proposal 2 when the observations are i.i.d. with a contaminated normal distribution.

2. NOTATIONS AND ASSUMPTIONS.

To incorporate the adaptive sequences \(H_n^{(1)}\) and \(H_n^{(2)}\) we rewrite the algorithm in the following way

\[
\begin{align*}
\theta_{n+1} &= \theta_n + (n+1)^{-1}H_n h(\theta_n, y_{n+1}), \\
\theta_0, H_0 & \text{ arbitrary and finite},
\end{align*}
\]

where

\[
\theta_n = (\eta_n, \sigma_n, a_n, b_n)^T.
\]

Further

\[
h(\theta_n, y_{n+1}) = \begin{bmatrix}
\sigma_n \psi(u_{n+1}) \\
\sigma_n x(u_{n+1}) \\
\psi'(u_{n+1}) - a_n \\
u_{n+1} x'(u_{n+1}) - b_n
\end{bmatrix}
\]
with

$$u_n = \sigma_n^{-1}(y_n - \eta_n)$$

and

(2.2) \[ H_n = \text{diag}(\tilde{\sigma}_n^{-1}, \tilde{\sigma}_n^{-1}, 1, 1) \]

With the notation \( \tilde{\sigma}_n \) we mean \( \sigma_n \) truncated above by a large positive number \( \nu_2 \) and below by a small positive number \( \nu_1 \) so that

(2.3) \[ \tilde{\sigma}_n = \begin{cases} \nu_1 & \text{if } \tilde{\sigma}_n < \nu_1, \\
\sigma_n & \text{if } \nu_1 \leq \tilde{\sigma}_n \leq \nu_2, \\
\nu_2 & \text{if } \tilde{\sigma}_n > \nu_2. \end{cases} \]

Throughout the paper it is understood that \( \nu_1 \leq \sigma \leq \nu_2 \). The above notation will also be used for \( \tilde{a}_n \) and \( \tilde{b}_n \). Note that

$$a_n = n^{-1} \sum_{j=1}^{n} u_j$$

and

$$b_n = n^{-1} \sum_{j=1}^{n} u_j x'(u_j)$$

so that with \( H_n \) defined as in (2.2) we get algorithm (1.4) with \( H_n^{(1)} \) and \( H_n^{(2)} \) given by (1.5).

Define

$$\bar{h}(x) = E(h(x,y_1))$$

and let \( \theta \) be the solution of \( \bar{h}(\theta) = 0 \), where \( \theta = (\eta, \sigma, a, b)^T \), that is with \( z_1 = \sigma^{-1}(y_1 - \eta) \)

(2.4) \[
\begin{align*}
E(\psi(z_1)) &= 0, \\
E(x(z_1)) &= 0, \\
E(\psi'(z_1)) &= a, \\
E(z_1 x'(z_1)) &= b.
\end{align*}
\]
Let $F_{\mathbb{1}}^m = F(y_1, \ldots, y_m)$ be the $\sigma$-algebra generated by the random variables $y_1, \ldots, y_m$. The sequence of strong mixing coefficients $\alpha_i$ is defined

$$\alpha_i = \sup_{m} \sup_{F_{\mathbb{1}}^m} |P(FG) - P(F)P(G)|.$$

Further, we need the following notations $n(k) = \lfloor k^\delta \rfloor$ for some $\delta > 1$ and

$$n(k + 1) - 1 \geq \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} = O(k^{-1}).$$

The constant $C$ is positive and may change from line to line. For shortness we usually write $z$ instead of $z_1$ below.

Finally we list the following assumptions for later use.

A1. The sequence of observations $\{y_i\}_1^\infty$ is strictly stationary and strong mixing with $\sum_{i=1}^{\infty} \alpha_i^{1-\epsilon} < \infty$ for some $0 < \epsilon < 1$. The marginal distribution is symmetric, continuous and positive in a neighbourhood of $\eta$.

A2. The function $h(x,y)$ is bounded and Lipschitz-continuous both as a function of $x$ and $y$, i.e.

$$\|h(x_1,y) - h(x_2,y)\| \leq K_1 \|x_1 - x_2\|$$

$$\|h(x,y_1) - h(x,y_2)\| \leq K_2 \|y_1 - y_2\|$$

for some positive constants $K_1$ and $K_2$.

A3. The function $\psi(\cdot)$ is bounded, increasing (strictly increasing in a neighbourhood of zero) and odd. The function $\chi(\cdot)$ is bounded, increasing on $(0,\infty)$ (strictly increasing in a neighbourhood of zero) and even.

A4. The function $\bar{h}(\cdot)$ satisfies $\bar{h}(0) = 0$.

A5. The following functions exist and are bounded:

$\psi^{(k)}(x)$ for $1 \leq k \leq 3$, $\chi^{(k)}(x)$ for $1 \leq k \leq 2$, $\psi^{(3)}(x)$,

$\chi^{(k)}(x)$ for $1 \leq k \leq 2$ and $\chi^{(k)}(x)$ for $1 \leq k \leq 3$. 
Note that A2 holds if the functions $x^3(x)$, $x^4(x)$, $x^5(x)$ and $x^6(x)$ exist and are bounded and that A5 is a strong assumption which is used in Section 4 only.

3. ALMOST SURE CONVERGENCE.

In this section we study almost sure convergence of the algorithm (2.1). It is proved in Theorem 3.1 that $\theta_n \to \theta$ a.s., where $\theta$ solves $\bar{H}(\theta) = 0$. The proof consists of two parts. Following Ruppert (1983) we show that

$$\theta_{n(k+1)} = \theta_{n(k)} + \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)}) + o(k^{-1}).$$

This is accomplished by writing

$$\theta_{n(k+1)} = \theta_{n(k)} + \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)}) +$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

$$= \theta_{n(k)} + \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)}) +$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} (i+1)^{-1} \bar{H}_i \bar{R}(\theta_{n(k)})$$

say, and then we prove that $R_{k,n(k+1)-1}^2$, $S_{k,n(k+1)-1}$ and $T_{k,n(k+1)-1}$ all are $o(k^{-1})$. The most involved expression, $R_{k,n(k+1)-1}$, is handled in Lemma 3.3, which is a lemma by Ruppert (1983, Lemma 3.2). The second part of the proof is to show that (3.1) is sufficient to establish
convergence. This is verified using Lemma 3.4, which is proved by a technique similar to the one used by Blum (1954).

In Lemma 3.1 we prove that \( \{g(y_i)\}_{i=1}^{\infty} \) is a mixingale with parameters \( \psi_m \) of size \( \sim_1 \) and \( c_n \) a constant if \( g(\cdot) \) is a bounded function with \( E(g(y_1)) = 0 \). For a definition of mixingales and notations, see McLeish (1975). Also the result in Lemma 3.2 is a mixingale inequality by McLeish (1975, Theorem 1.6).

**Lemma 3.1** Let \( g(\cdot) \) be a bounded Borel-measurable function with \( E(g(y_1)) = 0 \). If \( A_1 \) holds then \( \{g(y_i)\}_{i=1}^{\infty} \) is a mixingale with parameters \( \psi_m \) of size \( \sim_1 \) and \( c_n \) a constant.

**Proof** Let \( F_n = F(y_1, \ldots, y_n) \). Lemma 2.1 by McLeish (1975) with \( p = 2 \) and \( r = \infty \) gives

\[
\|E(g(Y_n)|F_n^{\infty-\infty}) - E(g(Y_n))\|_2 = \|E(g(Y_n)|F_n^{\infty-\infty})\|_2 \\
\leq 2(2^2+1)^{1/2} \sqrt{n(\psi_m^{\infty-\infty})} \|g(Y_n)\|_\infty \\
\leq C_{n^{-1}} 
\]

that is \( \psi = \omega_{m}^{-1/1} \) and \( c_n = C \). The fact that \( \psi_{m}^{\infty-\infty} \sim_1 \) for some \( 0 < \epsilon < 1 \) implies that \( \psi_m \) is of size \( \sim_1 \) according to McLeish (1975, p. 831). This proves the lemma.

**Lemma 3.2** Let \( \{g(y_i)\}_{i=1}^{\infty} \) be defined as in Lemma 3.1. Then there exists a constant \( c \) such that

\[
E(\max_{n} \sum_{i=1}^{n} d_i g(y_i)^2) \leq c \sum_{i=1}^{\infty} d_i^2
\]

for all \( n \) and constants \( d_1, \ldots, d_n \).

**Proof** It is obvious from Lemma 3.1 that \( \{d_i g(y_i)\}_{i=1}^{\infty} \) is a mixingale with parameters \( \psi_m \) of size \( \sim_{1} \) and \( c_n = d_n C \). Theorem 1.6 by McLeish (1975) proves the lemma.
**Lemma 3.3** If A1–A2 hold then

\[
\sup_{x \in \mathbb{R}} \max_{\mu(k) \leq n(k) \leq n(k+1)} \sum_{i=n(k)}^{n(k+1)} i^{-1}(h(x,y_{i+1}) - h(x)) = 0
\]

when \( k \to \infty \).

**Proof** According to Ruppert (1983, Lemma 3.2), we have to verify his assumptions A3–A6. A3 is obvious, A4 is exactly Lemma 3.2 above if \( r = 0 \) and A5 is satisfied since \( h(x,y) \) is bounded. Finally A6 follows from the Lipschitz continuity of \( h(x,y) \) as a function of \( y \). This proves Lemma 3.3.

**Lemma 3.4** Let \( t(\cdot) \) be a bounded function from \( \mathbb{R}^n \) to \( \mathbb{R} \), \( x^{(j)} \) an element of \( \mathbb{R}^n \), and \( \{x_k\} \) a sequence of r.v. satisfying the following assumptions:

1. \( x_k^{(j)} - x_k^{(j)} = D_k t(x_k) + o(k^{-1}) \) a.s. for some positive sequence \( \{D_k\} \) satisfying \( K_1 k^{-1} \leq D_k \leq K_2 k^{-1} \) where \( K_1 \) and \( K_2 \) are positive constants.

2. For all \( \gamma > 0 \) there are \( \delta_1, \delta_2 > 0 \) and \( N_\gamma \) satisfying \( \sup t(x_k) = -\delta_1 \), where the supremum is for \( \{x_k : x_k^{(j)} > x^{(j)} + \gamma\} \), and \( \inf t(x_k) = \delta_2 \), where the infimum is for \( \{x_k : x_k^{(j)} < x^{(j)} - \gamma\} \), for all \( k \geq N_\gamma \).

Then \( x_k^{(j)} \to x^{(j)} \) a.s.

**Proof** Assume that \( x_k^{(j)} \to \infty \). The assumptions make it possible to find a constant \( N_1 \) such that \( D_k t(x_k) + o(k^{-1}) < 0 \) for \( k > N_1 \), and thus \( x_k^{(j)} < x_k^{(j)} \) and hence we get a contradiction. (The case \( x_k^{(j)} \to -\infty \) is treated in the same way.) Now assume that \( x_k^{(j)} \) doesn't converge, that is \( \liminf x_k^{(j)} < \limsup x_k^{(j)} \), and also assume that \( \limsup x_k^{(j)} > x^{(j)} \).

(The case \( \limsup x_k^{(j)} \leq x^{(j)} \) is handled by a similar argument.)

Define \( \gamma \) from the relation \( \limsup x_k^{(j)} = x^{(j)} + 3\gamma \). Take \( N_2 \) so large that \( -D_k \delta_1 + o(k^{-1}) < 0 \) for \( k > N_2 \). Then we can find \( N_2 \leq n, n > n+1 \).
such that $x(j) < x_n(j) < x(j) + \gamma$, $x(j) + \gamma x_k(j) \leq x(j) + 2\gamma$

for $k = n+1, \ldots, m-1$ and $x_m(j) > x(j) + 2\gamma$. This is possible since $x_{k+1} - x_k = 0$. Now

$$x_m(j) - x_n(j) = \sum_{k=n}^{m-1} (D_k(x_k) + o(k^{-1})) < D_{n+1}(x_n) + o(n^{-1})$$

and this quantity can be made arbitrarily small, which is a contradiction.

Theorem 3.1 will now be stated.

**Theorem 3.1** Let $\theta_n$ be generated by algorithm (2.1). If A1-A4 hold, then $\theta_n + \theta_n$ a.s. as $n \to \infty$.

**Proof** The first part is to prove that

$$\theta_{n(k)+1} = \theta_{n(k)} + \sum_{i=n(k)}^{n(k+1)-1} H_i h(\theta_i(y_{i+1}), y_{i+1})$$

For $n(k) \leq i < n(k+1)$ we have

$$\theta_{i+1} = \theta_i + (i+1)^{-1} H_i h(\theta_i, y_{i+1})$$

$$= \theta_n(k) + \sum_{i=n(k)}^{n(k+1)-1} i H_i h(\theta_i, y_{i+1})$$

$$= \theta_n(k) + \sum_{i=n(k)}^{n(k+1)-1} i H_i \bar{h}(\theta_n(k)) +$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} i H_i h(\theta_i(y_{i+1}) - \bar{h}(\theta_n(k))) +$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} i H_i(h(\theta_i, y_{i+1}) - h(\theta_n(k), y_{i+1})))$$

$$+ \sum_{i=n(k)}^{n(k+1)-1} i H_i h(\theta_i, y_{i+1}) + R_k, k \in S_k, k \in T_k, k \in \mathbb{N}.$$
that
\[ \sup_{n(k)<i<n(k+1)} \| H_i - H_n(k) \| = o(1). \]

This follows easily for our choice of \( H_i \). The term \( T_{k,i} \) is treated by writing

\[ \| T_{k,i} \| = \| \sum_{i=n(k)}^{k} (i+1)^{-1} H_i (h(\theta_i, y_{i+1}) - h(\theta_n(k), y_{i+1})) \| \]

\[ \leq C \sum_{i=n(k)}^{k} (i+1)^{-1} \| h(\theta_i, y_{i+1}) - h(\theta_n(k), y_{i+1}) \| \]

\[ \leq C \sum_{i=n(k)}^{k} (i+1)^{-1} \| \theta_i - \theta_n(k) \| \]

\[ \leq C \sum_{k}^{n(k)} \max_{n(k) \leq i \leq n(k+1)} \| \theta_i - \theta_n(k) \| \]

since \( h(x,y) \) is Lipschitz continuous in \( x \). Now we have proved that

\[ \| \theta_{i+1} - \theta_n(k) \| \leq C_1 \rho_k + o(\rho_k) + C_2 \rho_k \max_{n(k) \leq i \leq n(k+1)} \| \theta_i - \theta_n(k) \|. \]

The inequality

\[ \max_{n(k) \leq i \leq n(k+1)} \| \theta_i - \theta_n(k) \| \leq (C_1 \rho_k + o(\rho_k))/(1-C_2 \rho_k) \]

for large \( k \) gives \( \| T_{k,i} \| = o(k^{-1}) \). Summarizing we have verified (3.3).

The second part of the proof is to show that this gives the result stated in the theorem. We apply Lemma 3.4 to the components of the vector \( \tilde{h}(\theta_n(k)) \). It is obvious that B1 holds and it remains to verify B2 for all components. We start at \( \tilde{h}^{(1)}(\theta_n(k)) \) and take the components in order. The convergence of \( \eta_n(k) \) follows because \( \tilde{h}^{(1)}(\theta_n(k)) \) satisfies B2 since \( \psi^*(\cdot) \) is increasing and odd. For \( \sigma_n(k) \) we have

\[ \tilde{h}^{(2)}(\theta_n(k)) = \tilde{h}^{(2)}(\theta_n(k)) - \tilde{h}^{(2)}(\eta_n(k)) + \tilde{h}^{(2)}(\kappa_n(k)) \]

where

\[ \kappa_n(k) = (\eta_n(k), \sigma_n(k), a_n(k), b_n(k))^T. \]
Assumptions A1 and A4 imply that $\hat{h}^{(2)}(\kappa_{n(k)}) = -\delta$ if $\sigma_{n(k)} > \sigma + \gamma$ because $\chi(\cdot)$ is increasing on $(0, \infty)$ and even and the first part of the assumption is satisfied from the fact that

$$\|h^{(2)}(\theta_{n(k)}) - \hat{h}^{(2)}(\kappa_{n(k)})\| \leq C\|n_{n(k)} - n\| \leq \delta/2$$

if $k$ is large enough. This is due to the proved part above and the Lipschitz-continuity of $\hat{h}(x)$ as a function of $x$. The second part of the assumption follows in the same way. The convergence of $a_{n(k)}$ follows because

$$\hat{h}^{(3)}(\theta_{n(k)}) = \hat{h}^{(3)}(\theta_{n(k)}) - \hat{h}^{(3)}(\lambda_{n(k)}) + \hat{h}^{(3)}(\lambda_{n(k)})$$

$$= \hat{h}^{(3)}(\theta_{n(k)}) - \hat{h}^{(3)}(\lambda_{n(k)}) + a - a_{n(k)}$$

where

$$\lambda_{n(k)} = (n, \sigma, a_{n(k)}, b_{n(k)})^T.$$

The Lipschitz-continuity makes it possible to choose $N$ such that

$$\|\hat{h}^{(3)}(\theta_{n(k)}) - \hat{h}^{(3)}(\lambda_{n(k)})\| \leq C(\|n_{n(k)} - n\| + \|\sigma - \sigma_{n(k)}\|) < \delta/2$$

for all $k > N$, and this proves that $a_{n(k)} + a$. A similar argument shows that $b_{n(k)} + b$.

The relation $\|\theta_{n(k+1)} - \theta_{n(k)}\| + \theta_{n(k)}$ and the previous result $(3.4)$ proves the remaining part of the theorem.

\[ \square \]

4. COMMENTS ON THE ASYMPTOTIC DISTRIBUTION AND ON THE CHOICE OF THE ADAPTIVE MATRIX.

In Section 3 we proved strong consistency of the algorithm (2.1). In order to discuss our choice of $H_n$, we also need results for the asymptotic distribution of the algorithm.

The asymptotic distribution can be derived from a strong represen-
tation theorem which is proved in Englund, Holst and Ruppert (1987).
The same theorem is stated here without proof to facilitate the
discussion below. (By the notation ̂{\text{x}} in this section we mean a
continuous and differentiable version of (2.3).)

**Theorem 4.1** If A1, A3–A5 hold and  ̂{\theta}_n is given by the algorithm
(2.1), then there exists  \epsilon > 0 such that

\[
\frac{b}{n} (\hat{\theta} - \theta) = \frac{b}{n} \left( \sum_{k=1}^{n} a^{-1} \sigma_\psi(z_k) \right) + O(n^{-\epsilon}),
\]

where

\[
z_k = \sigma^{-1}(y_k - \eta), \quad d_1 = b^{-1}E(z_1\psi''(z_1)) \quad \text{and} \quad d_2 = 1 + b^{-1}E(z_1^2\psi''(z_1)).
\]

For a sequence of independent observations the theorem gives

\[
\frac{b}{n} (\hat{\theta} - \theta) \in As N(0, V),
\]

where

\[
V = \begin{pmatrix}
V_{11} & 0 & 0 & 0 \\
0 & V_{22} & V_{23} & V_{24} \\
0 & V_{32} & V_{33} & V_{34} \\
0 & 0 & V_{43} & V_{44}
\end{pmatrix}
\]

with variance elements

\[
V_{11} = a^{-2} \sigma^2 E\psi^2(z),
V_{22} = b^{-2} \sigma^2 E\chi^2(z),
V_{33} = V(\psi'(z)) + 2d_1^2 V(\chi(z)) - 2d_1 C(\chi(z), \psi'(z)),
\]

\[
V_{23} = V_{32} = V(\psi''(z)) + 2d_1 V(\chi(z)) - 2d_1 C(\chi(z), \psi'(z)).
\]
\[ V_{44} = \mathbf{V}(z\mathbf{x}'(z)) + 2d_2^2\mathbf{V}(z) - 2d_2C(z\mathbf{x}(z),z\mathbf{x}'(z)), \]

and covariance elements

\[ V_{23} = -d_1b^{-1}\sigma\mathbf{V}(z) + b^{-1}\sigma C(z,\mathbf{y}'(z)), \]
\[ V_{24} = -d_2b^{-1}\sigma\mathbf{V}(z) + b^{-1}\sigma C(z,\mathbf{x}'(z)), \]
\[ V_{34} = 2d_1d_2\mathbf{V}(z) - d_1C(z,\mathbf{x}'(z)) - d_2C(z,\mathbf{y}'(z)) + C(z\mathbf{x}'(z),\mathbf{y}'(z)). \]

In the remaining part of this section we discuss whether \( \mathbf{H}_n \) in Section 3 is optimal or if we can find a better one.

Given a recursive algorithm it is well known that the "optimal" adaptive matrix \( \mathbf{H}^{\text{opt}} \) is the negative inverse of the derivatives of \( \mathbf{h}(\theta) \).

Here we get

\[
\mathbf{H}^{\text{opt}} = \begin{bmatrix}
\mathbf{E}(\mathbf{y}'(z)) & \mathbf{E}(z\mathbf{y}'(z)-\mathbf{y}(z)) & 0 & 0 \\
\mathbf{E}(\mathbf{x}'(z)) & \mathbf{E}(z\mathbf{x}'(z)-\mathbf{x}(z)) & 0 & 0 \\
\mathbf{E}(\sigma^{-1}\mathbf{y}'(z)) & \mathbf{E}(\sigma^{-1}z\mathbf{y}'(z)) & 1 & 0 \\
\mathbf{E}(\sigma^{-1}(\mathbf{x}'(z)+z\mathbf{x}'(z))) & \mathbf{E}(\sigma^{-1}(z\mathbf{x}'(z)+z^2\mathbf{x}'(z))) & 0 & 1
\end{bmatrix}
\]

If \( \psi \) is odd and \( \chi \) even this reduces to

\[
\mathbf{H}^{\text{opt}} = \begin{bmatrix}
a^{-1} & 0 & 0 & 0 \\
0 & b^{-1} & 0 & 0 \\
0 & -(bc)^{-1}\mathbf{E}(z\mathbf{y}'(z)) & 1 & 0 \\
0 & -\sigma^{-1}(1+b^{-1}\mathbf{E}(z^2\mathbf{x}'(z))) & 0 & 1
\end{bmatrix}
\]

where as above \( a = \mathbf{E}(\mathbf{y}'(z)) \) and \( b = \mathbf{E}(z\mathbf{x}'(z)) \). The values of \( a, b, \mathbf{E}(z\mathbf{y}'(z)) \) and \( \mathbf{E}(z^2\mathbf{x}'(z)) \) are in general unknown and if we try to estimate \( \mathbf{E}(z\mathbf{y}'(z)) \) and \( \mathbf{E}(z^2\mathbf{x}'(z)) \) we get more elements in the parameter vector.

It is however worth noting that this more complicated algorithm may reduce the asymptotic variances of \( a_n \) and \( b_n \). If we assume that we
know the values of \( d_1 = b^{-1}E(z\psi''(z)) \) and \( d_2 = 1+b^{-1}E(z^2\psi''(z)) \) and insert them and the truncated estimates of \( a \) and \( b \) in the matrix \( \mathbf{g}^{\text{opt}} \) we get the algorithm

\[
\begin{aligned}
\mathbf{n}_{n+1}' &= \mathbf{n}_n' + (n+1)^{-1}\mathbf{g}_n'\psi(\mathbf{u}_{n+1}')/\mathbf{\sigma}_n', \\
\mathbf{\sigma}_{n+1}' &= \mathbf{\sigma}_n' + (n+1)^{-1}\mathbf{g}_n'\mathbf{X}(\mathbf{u}_{n+1}')/\mathbf{\sigma}_n', \\
\mathbf{a}_{n+1}' &= \mathbf{a}_n' + (n+1)^{-1}(\psi'(\mathbf{u}_{n+1}') - d_1 \mathbf{X}(\mathbf{u}_{n+1}') - \mathbf{a}_n'), \\
\mathbf{b}_{n+1}' &= \mathbf{b}_n' + (n+1)^{-1}(\psi'(\mathbf{u}_{n+1}') - d_2 \mathbf{X}(\mathbf{u}_{n+1}') - \mathbf{b}_n'),
\end{aligned}
\]

(4.2)

where \( \mathbf{u}_{n+1}' = (\mathbf{v}_{n+1}'-\mathbf{n}_n')/\mathbf{\sigma}_n' \). It is easy to prove that this algorithm satisfies \( \mathbf{a}_n' = 0 \) a.s. and from the technique used in Englund, Holst and Ruppert (1987) it also follows for independent observations that

\[
\mathbf{n}_n'(\mathbf{\theta}'-\mathbf{\theta}) \in \mathcal{N}(0,\mathbf{V}'),
\]

where

\[
\mathbf{V}' = \begin{bmatrix}
\mathbf{V}'_{11} & 0 & 0 & 0 \\
0 & \mathbf{V}'_{22} & \mathbf{V}'_{23} & \mathbf{V}'_{24} \\
0 & \mathbf{V}'_{33} & \mathbf{V}'_{34} \\
0 & \mathbf{V}'_{44}
\end{bmatrix}
\]

The only difference between \( \mathbf{V} \) and \( \mathbf{V}' \) is the elements

\[
\begin{aligned}
\mathbf{V}'_{33} &= \mathbf{V}(\psi'(z)) + d_1^2\mathbf{V}(\mathbf{x}(z)) - 2d_1\mathbf{C}(\mathbf{x}(z),\psi'(z)), \\
\mathbf{V}'_{44} &= \mathbf{V}(\mathbf{x}'(z)) + d_2^2\mathbf{V}(\mathbf{x}(z)) - 2d_2\mathbf{C}(\mathbf{x}(z),\mathbf{x}'(z)).
\end{aligned}
\]

\[
\begin{aligned}
\mathbf{V}'_{34} &= d_1d_2\mathbf{V}(\mathbf{x}(z)) - d_1\mathbf{C}(\mathbf{x}(z),\mathbf{x}'(z)) - d_2\mathbf{C}(\mathbf{x}(z),\psi'(z)) + \\
&\quad + C(\mathbf{x}(z),\psi'(z)).
\end{aligned}
\]

Note that the variances \( \mathbf{V}_{33}' \) and \( \mathbf{V}_{44}' \) for the algorithm in Section 2 both are larger than \( \mathbf{V}_{33}' \) and \( \mathbf{V}_{44}' \).

As an example we take Huber’s Proposal 2 with \( k = 1.5 \). Although the functions defined in (1.2) do not satisfy A2 and A5 it is conjectured in Englund, Holst and Ruppert (1987) that the theorem is
valid if $d_1$ and $d_2$ are interpreted as $d_1 = -2b^{-1}k f(k)$ and

$$d_2 = 2 - 4b^{-1} k^3 f(k),$$

where $f$ is the density of $z$. For this choice and independent $N(0,1)$ distributed random variables we get

$$\mathbf{v} = \begin{pmatrix}
1.0371 & 0 & 0 & 0 \\
0.6894 & 0.0621 & 0.2797 \\
0.1641 & 0.2262 \\
1.2326
\end{pmatrix}$$

and

$$\mathbf{v}' = \begin{pmatrix}
1.0371 & 0 & 0 & 0 \\
0.6894 & 0.0621 & 0.2797 \\
0.0600 & 0.2699 \\
1.2143
\end{pmatrix}$$

Observe that $b'_n = 2k^2 a'_n - 2(k^2 - b'_k)$ for Huber's Proposal 2, which implies that $\rho(a'_n, b'_n) = 1$ and hence the number of components of $\theta'_n$ reduces to three.

It is our intention to study algorithm (4.2) with estimates of $d_1$ and $d_2$ in the near future. For Huber's Proposal 2 we only have to estimate $d_1$ and this makes use of $R^{opt}$ more feasible.

5. A NUMERICAL EXAMPLE.

In this section we give a numerical example of the adaptive estimator defined in Section 2 when $(y_t)_{t=1}^{1000}$ is a sequence of independent r.v. with a contaminated normal distribution $0.9N(0,1) + 0.1N(0,25)$. We will use Huber's Proposal 2, defined in (1.2). The constant $k$ is chosen to 1.5 which makes $B_{1.5} = 0.7784$. The variables $a'_n, b'_n$ and $b'_n$ are all truncated below by 0.1 and above by 10. To avoid that bad early estimates of $\hat{a}_n, \hat{b}_n$ and $\hat{b}_n$ influence the results too much we take $R_n = 1$ and
if \( n \leq 50 \). The initial value is \( \theta_0 = (0,1,0,0)^T \) and the solution of (2.4) is \((\eta, \sigma, a, b)^T = (0, 1.1346, 0.8468, 0.8024)^T\).

The figures below are produced to give an impression of the behaviour of the recursive estimates. The performance of \( \eta_n, \sigma_n, a_n \) and \( b_n \) for \( n = 1, \ldots, 1000 \) is shown in Figures 5.1 - 5.4 respectively. Also the recursive least squares estimator of \( \eta \), the sample mean, is given in Figure 5.1 for comparison. The arrows in the figures indicate the convergence points.

**Fig. 5.1.**
1: \( \eta_n \)
2: sample mean
Fig. 5.2. $\sigma_n$

Fig. 5.3. $a_n$

Fig. 5.4. $b_n$
Finally we mention that the asymptotic variance is 1.3977 for the recursive estimator, while the least squares estimator has the asymptotic variance 3.4000.

6. REFERENCES.


Hampel et al. (1986). Robust statistics, the approach based on influence functions. Wiley-Interscience.


END
DATE
FILMED
DEC.
1987