THE INFORMATION MATRIX AND ROBUST M-ESTIMATES FOR ARMA MODELS

BY

C.H. LEE AND R. DOUGLAS MARTIN

TECHNICAL REPORT NO. 24
JUNE 1982

THIS RESEARCH IS PARTIALLY SUPPORTED
BY
NATIONAL SCIENCE FOUNDATION GRANTS
SOC 78-09474 and SES 80-15570.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON
98195
THE INFORMATION MATRIX AND ROBUST M-ESTIMATES

FOR ARMA MODELS

by

C.H. Lee¹ and R. Douglas Martin²

ABSTRACT

This paper discusses the stochastic process structure of certain differential transformations (DT's) associated with perfectly observed ARMA processes and uses DT's to obtain the asymptotic information matrix for possibly non-Gaussian situations. The DT's can also be applied to implement approximate M-estimate algorithms for the ARMA model parameters. M-estimates yield asymptotic efficiency robustness for perfectly observed ARMA processes. Both asymptotic efficiency results and some finite-sample Monte Carlo results are presented.

¹C.H. Lee is currently with the Verbex Corporation, Bedford, MA.
²The research of both authors was supported by the National Science Foundation under Grants SOC 78-09474 and SES 80-15570.
1. INTRODUCTION

This paper is concerned with robust estimation of mixed autoregressive and moving average (ARMA) model parameters using the maximum-likelihood type estimate, or M-estimate, approach introduced by Huber (1964) for the problem of estimating a location parameter. Rellés (1968) and Huber (1973) extended the M-estimate approach to the classical regression set-up, where the observation errors in the linear regression model are assumed independent and identically distributed (i.i.d.). Rather complete and general asymptotic results for M-estimate in the linear regression model have recently been obtained by Yohai and Maronna (1979).

Work on robust methods for time series has occurred only rather recently, and an overview may be found in Martin (1981). A motivation for the work presented here is Martin's treatment (1979, 1982) of M-estimates for the special case of purely autoregressive procedures.

We shall focus our attention on one particular type of time series outlier model, namely that of a perfectly observed ARMA model with outliers being generated by the innovations. This is the so-called innovations outlier (IO) model (cf. Fox, 1972; Martin, 1979, 1981)

$$y(t) + \sum_{i=1}^{p} \phi_i y(t-i) = \gamma + \varepsilon(t) + \sum_{j=1}^{q} \theta_j \varepsilon(t-j) \quad (1.1)$$

where we relax the assumption that the density $f$ of the i.i.d. innovations sequence $\varepsilon(t)$ is Gaussian, and sometimes allow it to be a heavy-tailed outlier-generating distribution.

Although other kinds of time series-outlier models exist which appear to be of considerable use, e.g., additive outlier (AO) models
(Martin 1979, 1981), the IO model is of sufficient interest in its own right to warrant a thorough treatment. For one thing, the asymptotic behaviors of least squares (LS) and M-estimates for such models have rather striking features not encountered in ordinary regression models. Secondly, it appears that the IO model may be quite appropriate for certain kinds of time series such as economic time series and digital speech signals.

The asymptotic behavior of M-estimates for pure autoregressions has been discussed recently by Martin (1979, 1982), who points out Whittle's (1953) early results on the asymptotic min-max robustness and distribution free property of the ordinary LS-estimates. However, the LS-estimate is not efficient under non-gaussian heavy-tailed IO distributions. More precise estimates can be obtained by using M-estimates.

An M-estimate of ARMA model parameters is a solution of the following non-linear optimization problem

\[
\begin{align*}
\min_{\alpha} & \sum_{t=1}^{N} \rho(\hat{e}(t,\alpha)) \\
\end{align*}
\]

where \( \rho \) is a typical robustifying loss function, \( \alpha = (\gamma, \phi, \theta) \) is the parameter vector in Model (1.1), and \( \{\hat{e}(t,\alpha)\} \) is an estimate of the residual process. In this paper we shall, in particular, implement an approximate M-estimate (AM-estimate) algorithm and discuss the efficiency robustness of the M-estimate for ARMA model parameters. This paper is organized in the following manner.

In Section 2, we study certain differential transformations (DT) for ARMA processes. This transformation appears briefly in Box and Jenkins (1976) in the context of the finite-sample-size nonlinear
optimization problem associated with least-squares estimation.

The differential transformation proves to be useful for establishing the asymptotic form of the information matrix and the Cramer-Rao lower bound for the estimation of ARMA model parameters. This is treated in Section 3.

In Section 4 we show how the differential transformation can be applied to implement an approximate M-estimate (AM-estimate) algorithm. Basically, an AM-estimate is a one-step Newton procedure of the form

\[ \hat{\alpha}_M = \tilde{\alpha} + \tilde{\mathbf{A}}^{-1}(\psi,F)(\tilde{\mathbf{D}}^T\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{D}}^T \psi(\tilde{\varepsilon}) \]  

(1.3)

where \( \tilde{\alpha} \) is a preliminary estimate (e.g., a LS estimate) for the true parameter \( \alpha_0 = (\gamma_0, \phi_0, \theta_0) \), \( \tilde{\mathbf{D}} \) is the estimated differential transformation matrix based on \( \tilde{\alpha} \), and \( \tilde{\varepsilon} \) is the vector of estimated residuals based on \( \tilde{\alpha} \). \( \psi = \rho' \) is an appropriate robustifying psi-function, and \( \tilde{\mathbf{A}}(\psi,F) \) is an estimate of \( \mathbf{A}(\psi,F) = E_F[\psi'(\varepsilon)] \). With the same DT device, we can also work out the asymptotics for the M-estimate. However, this involves considerable technical detail which is beyond the scope of this presentation. Proofs of the asymptotic results may be found in Lee and Martin, 1982a.

In Section 5, we present asymptotic efficiency results for M-estimate. In particular, the asymptotic covariance \( V_M \) of \( \hat{\beta}_M = (\hat{\gamma}_M, \hat{\theta}_M, \hat{\phi}_M) \) is

\[ V_M = V_{LS} \cdot V_{loc} \sigma_e^2 \]  

(1.4)

where \( V_{loc} = E\psi^2/E\psi' \) is the asymptotic variance of Huber's (1964) location M-estimate, and \( V_{LS} \) is the asymptotic covariance for the LS-estimate. This result reveals that M-estimates for general ARMA models have the same behavior noted by Martin (1979, 1982) for the pure autoregression
case: $V_{loc}$ is quite stable while the innovations variance $\sigma^2_{\varepsilon}$ can be quite large for heavy-tailed densities. The relative efficiency of the M-estimate with respect to the LS estimate is just the ratio $\sigma^2_{\varepsilon}/V_{loc}$, which can be quite large for heavy-tailed densities. Thus the M-estimate gives increased precision over the LS-estimate when there are innovations outliers due to heavy-tailed innovations distributions.

Finally, in Section 6, we present some small sample Monte Carlo results for the estimation of ARMA (1,1) model parameters using both LS and M-estimates. The results clearly display the efficiency robustness of the M-estimates.

2. DIFFERENTIAL TRANSFORMATIONS FOR ARMA PROCESSES

A common approach to time series modeling supposes the data is generated by a linear aggregation of random shocks. For practical purposes, it is desirable to employ models which use as few parameters as possible. Parameter parsimony may often be achieved by representation of the linear process in terms of a small number of autoregression (AR) and moving average (MA) terms, in the form of a mixed autoregressive-moving average (ARMA) process of orders $p$ and $q$ (Box and Jenkins, 1970)

$$ y(t) + \phi_1 y(t-1) + \ldots + \phi_p y(t-p) + \theta_0, \epsilon(t) + \epsilon_1 \epsilon(t-1) + \ldots + \epsilon_q, \epsilon(t-q), t \in \mathbb{Z}^+ $$

(2.1)

where $\gamma_0$ is an intercept term, and the sequence $(\epsilon(t))$ consists of zero-mean, independent and identically distributed (i.i.d.) random variables.

$^+\mathbb{Z}$ denotes the set of all integers.
variables with finite variance $\sigma^2$. Define AR and MA operators by

$$\phi(z^{-1}) = 1 + \phi_1 z^{-1} + \ldots + \phi_p z^{-p} \quad (2.2)$$
$$\theta(z^{-1}) = 1 + \theta_1 z^{-1} + \ldots + \theta_q z^{-q} \quad (2.3)$$

with $z^{-1}$ the backshift operator, and use a subscript 'o' to denote true parameter values. The ARMA $(p,q)$ (2.1) process can be expressed in the operator form

$$\phi_o(z^{-1}) y(t) = \gamma_o + \theta_o(z^{-1}) \epsilon(t). \quad (2.4)$$

It is assumed that this process is stationary and invertible, and further that $\phi_0(x) = 0$ and $\theta_0(x) = 0$ have no common roots. For notational convenience, let $S_\phi = \phi(1) = 1 + \sum_{i=1}^p \phi_i$ and $S_\theta = \theta(1) = 1 + \sum_{j=1}^q \theta_j$ denote the values of the AR and MA operators evaluated at the point $z = 1$.

Stationarity implies that the roots of characteristic equation $\phi_0(x) = 0$ lie outside the unit circle in the complex $z$-plane, and thus the process $\{y(t)\}$ has the infinite order-moving average representation

$$y(t) = \gamma_o / S_\phi,0 + \phi_o^{-1}(z^{-1}) \theta_o(z^{-1}) \epsilon(t). \quad (2.5)$$

By invertibility, we mean that all the roots of characteristic equation $\theta_0(x) = 0$ have magnitude greater than unity, so the process $\{y(t)\}$ also has the infinite-order autoregressive representation

$$\theta_o^{-1}(z^{-1}) \phi_o(z^{-1}) \left[y(t) - \gamma_o / S_\phi,0 \right] = \epsilon(t). \quad (2.6)$$
The equalities appearing in the MA representation (2.5) and the AR representation (2.6) are in the *mean-square sense* (Anderson, 1971).

Let \( a^T = (\gamma, \varphi_1, \ldots, \varphi_p, \theta_p, \ldots \theta_q) \) denote the \((p+q+1)\)-parameter vector, and \( S_1 \) be the subset of the \((p+q+1)\)-dimension parameter space such that all the roots of \( \phi(x) = 0 \) and \( \theta(x) = 0 \) have magnitude greater than unity. Throughout, we assume that \( \{y(t)\} \) is a realization of a stationary and invertible ARMA\((p,q)\) process with *true but unknown parameter* \( \alpha_0 \in S_1 \) which satisfies difference equation (2.4). Let \( Y^t = \{y_s; s \leq t\} \) be an arbitrary observation vector at time \( t \) which includes the present observation \( y_t \) and its remote past, then (2.6) is an expression of the true random shocks \( \{\varepsilon(t)\} \) as a function of the true parameter \( \alpha_0 \in S_1 \) and known observation vector \( Y^t \), i.e.,

\[
\varepsilon(t) = \sum_{j=0}^{\infty} \delta_{0,j} (y_{t-j} + \sum_{i=1}^{p} \varphi_{i} y_{t-j-i} - \gamma_{0})
\]

(2.7)

where coefficients \( \{\delta_{0,j}\} \) are functions of \( \alpha^T_0 = (\theta_0,\ldots,\theta_0,\theta_q) \) satisfying the following system of equations (Anderson, 1971)

\[
\begin{align*}
1 &= \theta_{0,0} \delta_{0,0} = \delta_{0,0} \\
0 &= \theta_{0,0} \delta_{0,1} + \theta_{0,1} \delta_{0,0} \\
&\vdots \\
0 &= \theta_{0,0} \delta_{0,q-1} + \cdots + \theta_{0,q-1} \delta_{0,0} \\
0 &= \theta_{0,0} \delta_{0,t} + \cdots + \theta_{0,q} \delta_{0,t-q} \text{ for } t > q
\end{align*}
\]

(2.8)
or simply

\[ \theta^{-1}(z^{-1}) = \sum_{j=0}^{\infty} \delta_j z^{-j} \quad (2.9) \]

Now, for an arbitrary vector \( \alpha = (\gamma, \phi, \theta) \in SI \), we can obtain coefficients \( \{\delta_j\} \) according to (2.8) with \( \theta \) replacing \( \theta_0 \) and \( \delta_j \) replacing \( \delta_0, j \). So a residual process \( \{r(t, \mathbf{y}_t, \alpha)\} \) can be constructed as a function of an arbitrary observation vector \( \mathbf{y}_t \) and an arbitrary vector \( \alpha \in SI \):

\[
r(t, \mathbf{y}_t, \alpha) = \sum_{j=0}^{\infty} \delta_j \left\{ \gamma - \sum_{i=1}^{p} \phi_i \mathbf{y}_{t-i} - \gamma \right\} \quad (2.10)
\]

The equality above is again in the mean square sense.

Since \( \alpha \in SI \), the \( \{\delta_j\} \) sequence in (2.8) also satisfies

\[
\theta^{-1}(z^{-1}) = \sum_{j=0}^{\infty} \delta_j z^{-j} \quad (2.11)
\]

and thus, the residual process \( \{r(t, \mathbf{y}_t, \alpha)\} \) satisfies the difference equation

\[
\theta(z^{-1})r(t, \mathbf{y}_t, \alpha) + \gamma = \phi(z^{-1})y(t) \quad (2.12)
\]

Clearly, for all \( t \in \mathbb{Z} \), and for any \( \alpha \in SI \), the residual \( r(t, \mathbf{y}_t, \alpha) \) is a continuously differentiable function of \( \alpha \), so we can evaluate the derivatives of \( r(t, \mathbf{y}_t, \alpha) \) with respect to \( \alpha \) in the mean square sense. These derivatives which we will call differential transformations, possess interesting and useful properties that will be developed in the following.

Corresponding to each residual \( r(t, \mathbf{y}_t, \alpha) \), we can define, for all \( \alpha \in SI \), the following first-order differential transformations:
\[ \Delta_{\gamma}(t, Y^t, \alpha) = \partial r(t, Y^t, \alpha) / \partial \gamma \] (2.13)
\[ \Delta_{\phi,i}(t, Y^t, \alpha) = \partial r(t, Y^t, \alpha) / \partial \phi_i , \quad i=1...p \] (2.14)
\[ \Delta_{\theta,j}(t, Y^t, \alpha) = \partial r(t, Y^t, \alpha) / \partial \theta_j , \quad j=1...q \] (2.15)

The above differentiations are computed assuming the observation sequence \( \{y(t)\} \) is fixed. For notational convenience, we subsequently drop the argument \( Y^t \) as long as no ambiguity occurs. Thus corresponding to each residual sequence \( \{r(t, \alpha)\} \), we have defined \((p+q+1)\) sequences \( \{\Delta_{\gamma}(t, \alpha)\}, \{\Delta_{\phi,i}(t, \alpha), \quad i=1...p\} \) and \( \{\Delta_{\theta,j}(t, \alpha), \quad j=1...q\} \) whose properties we will investigate.

A preview of results to come is obtained by considering the zero-mean pure autoregressive situation, with \( q = 0 \). Then \( r(t, \alpha) \) is a linear function of the parameters

\[ r(t, \alpha) = y(t) + \phi_1 y(t-1) + \ldots + \phi_p y(t-p) \] (2.16)

In this case, it is obvious that \( \Delta_{\phi,i}(t, \alpha) = y(t-i) \), and therefore the differential transformations are just shift versions of the original observed sequence \( \{y(t)\} \). Furthermore, \( \{\Delta_{\phi,i}(t, \alpha)\} \) forms a stationary AR(p) process for each \( i, 1 \leq i \leq p \). It turns out that for the general mixed ARMA(p,q) process, the sequences \( \{\Delta_{\phi,i}(t, \alpha)\} \) and \( \{\Delta_{\theta,j}(t, \alpha)\} \) possess the same "shift" and "stationary" properties, as will be demonstrated in the following lemmas.

**Lemma 2.1** For all \( t \in \mathbb{Z} \) and any \( \alpha \in \mathcal{S}_\theta \), we have

\[ \Delta_{\gamma}(t, \alpha) = -1/(1 + \sum_{j=1}^{q} \theta_j) = -1/S_\theta \] (2.19)
a constant independent of the data.

Proof. For any $\alpha \in S$, the residual sequence $\{y(t, \alpha)\}$ satisfies the difference equation (2.12). Now differentiate both sides of (2.14) with respect to $\gamma$, the result is immediate.

Lemma 2.2 (Shift Property of Differential Transformations)

For all $\alpha \in S$ and for all $t \in \mathbb{Z}$, we have

(a) $\Delta_{\phi, 1}(t, \alpha) = \Delta_{\phi, 2}(t+1, \alpha) = \ldots = \Delta_{\phi, p}(t+p-1, \alpha)$  \hspace{1cm} (2.18)

(b) $\Delta_{\theta, 1}(t, \alpha) = \Delta_{\theta, 2}(t+1, \alpha) = \ldots = \Delta_{\theta, q}(t+q-1, \alpha)$  \hspace{1cm} (2.19)

Proof. (a) Differentiate both sides of (2.12) with respect to $\phi_i$; then since the observation sequence $\{y(t)\}$ is fixed, we have

$$y(t-i) = \theta(z^{-1}) \frac{\partial r(t, \alpha)}{\partial \phi_i} .$$  \hspace{1cm} (2.20)

Therefore

$$\Delta_{\phi, i}(t, \alpha) = \frac{\partial r(t, \alpha)}{\partial \phi_i} = \theta^{-1}(z^{-1})y(t-i) .$$  \hspace{1cm} (2.21)

Only the difference $(t-i)$ involved in the expression for $\Delta_{\phi, i}(t, \alpha)$, and (2.18) follows immediately.

(b) Differentiate both sides of (2.12) with respect to $\theta_j$, then

$$\theta(z^{-1}) \frac{\partial r(t, \alpha)}{\partial \theta_j} + r(t-j, \alpha) = 0$$

or

$$\theta(z^{-1}) \frac{\partial r(t, \alpha)}{\partial \theta_j} = -r(t-j, \alpha) .$$  \hspace{1cm} (2.22)
Thus we have,

$$\Delta_{\theta,j}(t,\alpha) = \frac{\partial r(t,\alpha)}{\partial \theta_j} = -\theta^{-1}(z^{-1})r(t-j,\alpha). \quad (2.23)$$

Again, since only $(t-j)$ is involved, we have proved (2.19).

We now show that for arbitrary $\alpha \in \mathcal{SI}$, the differential transformations satisfy autoregressive type difference equations.

**Lemma 2.3** For all $\alpha \in \mathcal{SI}$, $t \in \mathcal{Z}$ and $1 \leq i \leq p$, $1 \leq j \leq q$, we have

(a) \hspace{1em} \Delta_{\phi,i}(t,\alpha) + \phi_1 \Delta_{\phi,i}(t-1,\alpha) + \ldots + \phi_p \Delta_{\phi,i}(t-p,\alpha) \\
= \gamma/S_\theta + r(t-i,\alpha) \hspace{2em} (2.24)

and

(b) \hspace{1em} \Delta_{\theta,j}(t,\alpha) + \theta_1 \Delta_{\theta,j}(t-1,\alpha) + \ldots + \theta_q \Delta_{\theta,j}(t-q,\alpha) \\
= -r(t-j,\alpha) \hspace{2em} (2.25)

**Proof.** Part (b) is just the result of (2.22). To show part (a) we use operator $\phi(z^{-1})$ on both sides of (2.21),

$$\phi(z^{-1})\Delta_{\phi,i}(t,\alpha) = \theta^{-1}(z^{-1})\phi(z^{-1})y(t-i) \hspace{2em}$$

$$= \theta^{-1}(z^{-1})[\gamma + \theta(z^{-1})r(t-i,\alpha)]$$

or

$$\phi(z^{-1})\Delta_{\phi,i}(t,\alpha) = \gamma/S_\theta + r(t-i,\alpha) \hspace{2em} (2.26)$$

Therefore, (2.24) is verified.
Let $\alpha_0 \in \Sigma$ be the true parameter which generates the observed \{y(t)\} according to model (2.1)

$$\phi_0(z^{-1})y(t) = \gamma_0 + \theta_0(z^{-1})\epsilon(t).$$

(2.27)

In Lemma 2.3, we show that the differential transformations satisfy some AR type difference equations with residuals $r(t, \alpha)$ as input. The next lemma shows that when we consider the random shock $\epsilon(t)$ as input, the residuals and differential transformations satisfy some ARMA type difference equations.

Lemma 2.4 For all $t \in \mathbb{Z}$ and $\alpha \in \Sigma$, the residuals and differential transformations satisfy the following ARMA-type difference equations. Define $\gamma_\epsilon = \phi(z^{-1})\gamma_\epsilon - \phi_0(z^{-1})\gamma$, we have

(a) $\phi_0(z^{-1})\theta(z^{-1})r(t, \alpha) = \gamma_\epsilon + \phi(z^{-1})\theta_0(z^{-1})\epsilon(t).$  

(2.28)

(b) For $1 \leq i \leq p$,

$$\phi_0(z^{-1})\theta(z^{-1})\Delta_{\phi, i}(t, \alpha) = \gamma_\epsilon + \phi(z^{-1})\theta_0(z^{-1})\epsilon(t-i).$$

(2.29)

(c) For $1 \leq j \leq q$,

$$\phi_0(z^{-1})\theta^2(z^{-1})\Delta_{\theta, j}(t, \alpha) = -\gamma_\epsilon - \phi(z^{-1})\theta_0(z^{-1})\epsilon(t-j).$$

(2.30)

Furthermore, $r(t, \alpha), \Delta_{\phi, i}(t, \alpha)$ and $\Delta_{\theta, j}(t, \alpha)$ are all stationary processes.

Proof. The proof is quite straightforward, see Lee (1981). □

It is obvious that when evaluated at the true parameter value $\alpha_0$, the residual $r(t, \alpha_0)$ is equal to the random input $\epsilon(t)$ which generates \{y(t)\}. Thus we have the following result.
Theorem 2.1 If the true parameter \( \alpha_0 \in \mathcal{S}_I \), then the differential transformations evaluated at the true parameter \( \alpha_0 \) are stationary processes of pure autoregressive type such that

(a) For each \( i, 1 \leq i \leq p \), \( \{ \Delta_{\phi,i}(t,\alpha_0) \} \) is a stationary AR(p) process with \( T_0 = (\phi_0,1,\ldots,\phi_0,p) \) as parameter, i.e.,

\[
\phi_0(z^{-1})\Delta_{\phi,i}(t,\alpha_0) = \gamma_0/S_{\theta,0} + \epsilon(t-i),
\]

(2.31)

and

(b) For each \( j, 1 \leq j \leq q \), \( \{ \Delta_{\theta,j}(t,\alpha_0) \} \) is a stationary AR(q) process with \( \theta_0 = (\theta_0,1,\ldots,\theta_0,q) \) as parameter, i.e.,

\[
\theta_0(z^{-1})\Delta_{\theta,j}(t,\alpha_0) = -\epsilon(t-j).
\]

(2.32)

Proof. Replace \( \alpha \) by \( \alpha_0 \) in Eqs. (2.24) and (2.25), the theorem follows immediately.

Corollary 2.1 The differential transformations \( \Delta_{\phi,i}(t,\alpha_0) \) and \( \Delta_{\theta,j}(t,\alpha_0) \) are independent of the present shock \( \epsilon(t) \).

Proof. By looking into (2.31) and (2.32), \( \Delta_{\phi,i}(t,\alpha_0) \) and \( \Delta_{\theta,j}(t,\alpha_0) \) depend only on the past shocks \( \{ \epsilon(s); s < t \} \), therefore the Corollary is immediate.

3. INFORMATION MATRIX AND CRAMER-RAO LOWER BOUND

Suppose that \( \{ y(t) \} \) is a realization of ARMA(p,q) process generated by model (2.1). Let \( h \) be the conditional density of the present observation \( y(t) \) given all its remote past \( Y^{t-1} = \{ y(s); s < t \} \), then we can
define the information matrix for the information contained in an observation at time $t$ as

$$I(t, \alpha_0) = E\left[\left(\frac{\partial}{\partial \alpha_0}\right) \log f(y_t | \gamma^{t-1}, \alpha_0) \cdot \left(\frac{\partial}{\partial \alpha_0}\right)^T \log f(y_t | \gamma^{t-1}, \alpha_0)\right], \quad (3.1)$$

where the expectation is evaluated at the true distribution and $\alpha_0$ is the true but unknown parameter.

Now assume the random shocks $\{e(t)\}$ in model (2.1) are zero-mean with common symmetric density $f$ having a finite variance $\sigma_e^2$. Then, as in (2.7), the unobservable random shock $r(t, \gamma^t, \alpha_0) = e(t)$ can be expressed as a function of known observation vector $\gamma^t$ and unknown parameter $\alpha_0$. For notational convenience we drop argument $\gamma^t$ and use $\alpha$ to denote the true parameter $\alpha_0$. Thus, the information matrix becomes

$$I(t, \alpha) = E\left[\left(\frac{\partial}{\partial \alpha}\right) \log f(r(t, \alpha)) \cdot \left(\frac{\partial}{\partial \alpha}\right)^T \log f(r(t, \alpha))\right]. \quad (3.2)$$

Denoting the efficient score function for location (Cox and Hinkley, 1974) by

$$\psi(e) = - f'(e)/f(e)$$

and the corresponding Fisher information by

$$i(f) = E[\psi^2(e)] \quad . \quad (3.3)$$

We have

$$I(t, \alpha) = E\left[\left(\frac{\partial}{\partial \alpha}\right) r(t, \alpha) \cdot \left(\frac{\partial}{\partial \alpha}\right)^T r(t, \alpha) \cdot \psi^2(r(t, \alpha))\right]$$

$$= E\left[\left(\frac{\partial}{\partial \alpha}\right) r(t, \alpha) \cdot \left(\frac{\partial}{\partial \alpha}\right)^T r(t, \alpha)\right] \cdot E[\psi^2(e)]$$

$$= E\left[\left(\frac{\partial}{\partial \alpha}\right) r(t, \alpha) \cdot \left(\frac{\partial}{\partial \alpha}\right)^T r(t, \alpha)\right] \cdot i(f) \quad .$$
The second equality results from the independence of \( \partial r(t,\alpha)/\partial \alpha \) and \( r(t,\alpha) = \varepsilon(t) \) established in Corollary 2.1.

Denote the gradient vector for the residual process as

\[
D(t,\alpha) = \frac{\partial}{\partial \alpha} r(t,\alpha) \tag{3.4}
\]

and the association covariance matrix as

\[
\Omega(t,\alpha) = \mathbb{E}[D(t,\alpha) \cdot D^T(t,\alpha)] \tag{3.5}
\]

By Lemma 2.4, \( \{D(t,\alpha)\} \) is a stationary vector process; therefore \( \Omega(t,\alpha) \) and hence \( I(t,\alpha) \) do not depend on time index \( t \).

To evaluate \( I(t,\alpha) = I(\alpha) \), we first compute \( \Omega(t,\alpha) = \Omega(\alpha) \). From (2.12), we know that

\[
\theta_0(z^{-1}) \frac{\partial}{\partial \gamma} r(t,\alpha) = -1
\]

or simply

\[
\frac{\partial}{\partial \alpha} r(t,\alpha) = -1/S_\theta \tag{3.6}
\]

which is a constant independent of the data.

Let \( \mu \) be the mean of process \( \{y(t)\} \), the model (2.1) can also be expressed in the following form:

\[
[y(t) - \mu] + \phi_1 [y(t-1) - \mu] + \ldots + \phi_p [y(t-p) - \mu] = \varepsilon(t) + \theta_1 \varepsilon(t-1) + \ldots + \theta_q \varepsilon(t-q) \tag{3.7}
\]

where
\[
\mu = \gamma/(1 + \sum_{i=1}^{p} \phi_i) = \gamma/S_\phi . 
\] (3.8)

Then, (2.31) becomes

\[
\psi(z^{-1})[\frac{\partial}{\partial \phi_i} r(t,\alpha) - \mu/S_{\Theta}] = r(t-i,\alpha) , 
\] (3.9)

Since \( r(t,\alpha) \) has zero-mean, therefore,

\[
E[\frac{\partial}{\partial \phi_i} r(t,\alpha)] = \mu/S_{\Theta} , 
\] (3.10)

and similarly from (2.32)

\[
E[\frac{\partial}{\partial \theta_j} r(t,\alpha)] = 0 . 
\] (3.11)

This gives

\[
E[\{\frac{\partial}{\partial \gamma} r(t,\alpha)\}^2] = 1/S_{\Theta}^2 
\] (3.12)

\[
E[\{\frac{\partial}{\partial \gamma} r(t,\alpha)\} \cdot \{\frac{\partial}{\partial \phi} r(t,\alpha)\}] = -\mu/S_{\Theta}^2 
\] (3.13)

and

\[
E[\{\frac{\partial}{\partial \gamma} r(t,\alpha)\} \cdot \{\frac{\partial}{\partial \theta_j} r(t,\alpha)\}] = 0 . 
\] (3.14)

Let \( \Sigma \) be the covariance matrix of the differential transformation vector for a zero-mean ARMA(p,q) process, with parameter \((\phi,\theta)\). Then \( \Sigma \) may be expressed in a partitioned form

\[
\Sigma = \begin{pmatrix}
\Sigma_{\phi \phi} & \Sigma_{\phi \theta} \\
\Sigma_{\theta \phi} & \Sigma_{\theta \theta}
\end{pmatrix} . 
\] (3.15)
Note from (3.10) and (3.11) that the elements of $\Sigma$ can be expressed as follows:

$$
\Sigma_{\phi_i,\phi_k} = E\left[\frac{\partial}{\partial \phi_i} r(t,\alpha) \cdot \frac{\partial}{\partial \phi_k} r(t,\alpha) - \mu/\theta_0\right]
$$

$$
\Sigma_{\phi_i,\theta_\lambda} = E\left[\frac{\partial}{\partial \phi_i} r(t,\alpha) \cdot \frac{\partial}{\partial \theta_\lambda} r(t,\alpha)\right] = \Sigma_{\phi_i,\phi_i}
$$

$$
\Sigma_{\theta_j,\theta_\lambda} = E\left[\frac{\partial}{\partial \theta_j} r(t,\alpha) \cdot \frac{\partial}{\partial \theta_\lambda} r(t,\alpha)\right].
$$

Therefore we have,

$$
E\left[\frac{\partial}{\partial \phi_i} r(t,\alpha)\right] \cdot \frac{\partial}{\partial \theta_\lambda} r(t,\alpha) = \Sigma_{\phi_i,\phi_k} + (\mu/\theta_0)^2
$$

and

$$
E\left[\frac{\partial}{\partial \phi_i} r(t,\alpha) \cdot \frac{\partial}{\partial \theta_\lambda} r(t,\alpha)\right] = \Sigma_{\phi_i,\theta_\lambda}.
$$

(3.16)

Now, the information matrix becomes

$$
I(\alpha) = \begin{pmatrix}
1/\theta_0^2 & -\mu/\theta_0 \cdot 1_\rho^T & 0_\rho^T \\
-\mu/\theta_0 \cdot 1_\rho & \mu^2/\theta_0^2 \cdot 1_\rho + \Sigma_{\phi\phi} & \Sigma_{\phi\theta} \\
0_\rho & \Sigma_{\theta\phi} & \Sigma_{\theta\theta}
\end{pmatrix} \cdot i(f)
$$

(3.17)

where $1_\rho$ is a $p$-vector with all components being 1, $0_\rho$ is a $q$-vector with all components being 0, and $I_m$ is an $m \times m$ identity matrix.

Consider the parameterization $\beta^T = (\nu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$, where the intercept $\gamma$ appearing in $\alpha$ is replaced by the location parameter. It is easy to check that $\alpha = g(\beta)$ is a one-to-one transformation under
stationary (i.e., $S_\phi \neq 0$). The Jacobian matrix of transforming $\alpha$, with element $J_{ij} = \partial g_{ij}(\beta)/\partial \beta_j$ is

$$J = \begin{pmatrix} S_\phi & 0 \\ \mu 1 & -p^T \\ 0 & -p+q \\ 0 & I_{p+q} \end{pmatrix} .$$

(3.18)

The information matrix $I(\beta)$ corresponding to parameter $\beta$ is given by (Cox and Hinkley, 1974)

$$I(\beta) = J^T I(\alpha) J$$

(3.19)

and so we have

$$I(\beta) = \begin{pmatrix} S_\phi^2 / S_\theta^2 & 0 \\ 0 & -p+q^T \\ 0 & I_{p+q} \end{pmatrix} i(f) .$$

(3.20)

It may be noted that the information matrix for estimates of the location parameter $\mu$ and ARMA parameters $(\phi, \theta)$ is block diagonal, which is not the case for estimates of intercept $\gamma$ and $(\phi, \theta)$. Thus for large samples, $\mu$ can be estimated independently of $(\phi, \theta)$ in non-Gaussian as well as in Gaussian situations, just as in the case of non-Gaussian autoregressions (Martin, 1978).

Correspondingly, the Cramer-Rao lower bound for $\beta = (\mu, \phi, \theta)$ is

$$V_{CR}(\beta) = I^{-1}(\beta)$$
or

\[
V_{CR}(\beta) = \begin{pmatrix}
(S_\phi/S_\phi)^2 \cdot \sigma_e^2 & 0_{p+q}^T \\
0_{p+q} & C(\phi, \theta)
\end{pmatrix} \begin{pmatrix} 1 \\ i(f) \cdot \sigma_e^2 \end{pmatrix}
\]

where \(C(\phi, \theta) = \Sigma^{-1} \sigma_e^2\) is the asymptotic covariance matrix of the LS estimate for the estimation of ARMA(\(p,q\) coefficients \(\phi\) and \(\theta\), which is independent of the innovations distributions.

The matrix \(\Sigma\) may be computed from the knowledge of the roots of the characteristic equations for the AR and MA operators (cf., Chapter 7 in Box and Jenkins, 1970). Suppose \(\{G_i : i=1, \ldots, p\}\) and \(\{H_j : j=1, \ldots, q\}\) are these corresponding roots then the \((i,j)\) element of the submatrix are

\[
\Sigma_{\phi\phi}(i,j) = (1-G_i G_j)^{-1} \cdot \sigma_e^2, \quad \Sigma_{\phi\theta}(i,j) = -(1-G_i H_j)^{-1} \cdot \sigma_e^2\]

and \(\Sigma_{\theta\theta}(i,j) = (1-H_i H_j)^{-1} \cdot \sigma_e^2\). The following is a simple example.

Example Consider the ARMA(1,1) case, the roots for AR and MA operators are \(-\phi\) and \(-\theta\) respectively, therefore the elements of matrix \(\Sigma\) are:

\[
\Sigma_{\phi\phi} = \sigma_e^2/(1-\phi^2), \quad \Sigma_{\phi\theta} = -\sigma_e^2/(1-\phi \theta), \quad \text{and} \quad \Sigma_{\theta\theta} = \sigma_e^2/(1-\theta^2).
\]

The information matrix for \(\beta = (\mu, \phi, \theta)\) is

\[
I(\beta) = \begin{pmatrix}
\frac{(1+\phi)^2}{(1+\theta)^2} & 0 & 0 \\
0 & \sigma_e^2/(1-\phi^2) & -\sigma_e^2/(1-\phi \theta) \cdot i(f) \\
0 & -\sigma_e^2/(1-\phi \theta) & \sigma_e^2/(1-\theta)^2
\end{pmatrix} (3.22)
\]
and the Cramer-lower bound is

\[
V_{CR}(\theta) = \begin{pmatrix}
\frac{(1+\theta)^2 \sigma^2}{1+\phi} & 0 & 0 \\
0 & C(\phi, \theta) & \frac{1}{i(f) \cdot \sigma^2} \\
0 & \frac{1}{i(f) \cdot \sigma^2} & \frac{1}{i(f) \cdot \sigma^2}
\end{pmatrix}
\]  

(3.23)

where

\[
C(\phi, \theta) = \Sigma^{-1} \sigma^2 = \begin{pmatrix}
(1-\phi^2)(1-\theta) & (1-\phi^2)(1-\theta^2) \\
(1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi)
\end{pmatrix}
\]  

(3.24)

is the asymptotic covariance matrix for the LS estimate of ARMA(1,1) parameter (\phi, \theta), which is independent of the innovations variance \sigma^2.

4. AM-ALGORITHM FOR ARMA MODEL PARAMETER ESTIMATION

A general stationary and invertible ARMA(p,q) process \{y(t)\} is generated from model (2.1). Now given the observed data vector \( \mathbf{y}^T = [y(1), \ldots, y(N)] \), we want to estimate the (p+q+1)-parameter vector \( \mathbf{a} = (\mathbf{y}, \phi, \theta) \). An M-estimate for \( \mathbf{a} \) is obtained by solving the following optimization problem

\[
\min_{\mathbf{a}} \sum_{t=1}^{N} \rho(r(t, \mathbf{y}, \mathbf{a}))
\]

(4.1)

where \( r(t, \mathbf{y}, \mathbf{a}) \) is the residual expressed as a function of the parameter \( \mathbf{a} \) and the data vector \( \mathbf{y} \), and \( \rho \) is a properly chosen robustifying loss function. For ARMA processes, the residual \( r(t, \mathbf{y}, \mathbf{a}) \) is a nonlinear function of \( \mathbf{a} \). With \( D(t, \mathbf{y}, \mathbf{a}) = \partial r(t, \mathbf{y}, \mathbf{a})/\partial \mathbf{a} \), the corresponding differential transformation vector, we have the estimating equation
\[ \sum_{t=1}^{N} D(t, Y, \hat{a}) \psi(r(t, Y, \hat{a})) = 0 \] (4.2)

where \( \psi = \rho' \), which we wish to solve (4.2) for \( \hat{a} \).

**Choice of scale estimate**

The solution \( \hat{a} \) to (4.2) is not scale invariant. To obtain a scale invariant estimate \( \hat{a} \), we modify the psi-function in (4.2) by

\[ \psi_s(x) = s \psi(x/s) \] (4.3)

where \( s \) is a positive innovations scale parameter, and (4.2) thus becomes,

\[ \sum_{t=1}^{N} D(t, Y, \hat{a}) \psi_s(r(t, Y, \hat{a})) = 0 . \] (4.4)

Since the scale parameter is usually unknown, it has to be estimated. In order to solve the nonlinear equation (4.4), the solution is invariably iterative, it is natural to compute a robust scale estimate \( s \) as part of the overall iterative procedure. Therefore at each iteration, we use the current estimate \( \tilde{a} \) of \( a \) to form residuals \( \{r(t, \tilde{a})\} \), and a robust scale estimate \( \hat{s} \) is then computed from these residuals. We shall actually use \( \hat{s} = \text{MAD}(r(t, \tilde{a}))/0.6745 \), where for any vector \( z = [z(1), \ldots, z(N)] \), \( \text{MAD}(z(t)) = \text{Median}|z_i - \text{Median}(z)| \). (This has been an often-used procedure for obtaining robust scale estimates, cf., Mosteller and Tukey, 1977.) Now, the next iteration \( \hat{a} \) is obtained by solving

\[ \sum_{t=1}^{N} D(t, Y, \hat{a}) \psi_s(r(t, Y, \hat{a})) = 0 \] (4.5)

with \( \hat{s} \) fixed. Another possibility for dealing with the scale problem is given by Huber (Proposal 2, 1964; 1973). From now on, we drop the argument \( Y \) in \( r(t, Y, \alpha) \) and \( D(t, Y, \alpha) \) for notational simplicity.
Choice of root-searching algorithms

There are a number of well known approaches for solving nonlinear regression problems of the type (4.5). For a brief review of the possibilities, see Chambers (1973, 1977). The usual Newton-Raphson (NR) procedure requires second order differential transformations (i.e., a Hessian matrix) in order to carry out the Taylor series expansion on both $D(t,a)$ and $\psi(r(t,a))$. We propose here a simpler algorithm which uses only the first order differential transformation $D(t,a)$. Given the current estimate $\hat{\alpha}$, we expand $\psi(t(t,a))$ about the point $\hat{\alpha}$, i.e.,

$$\psi(r(t,a)) = \psi(r(t,\hat{\alpha})) + \psi'(r(t,\hat{\alpha})) \cdot D(t,\hat{\alpha})(\hat{\alpha} - \hat{\alpha})$$  \hspace{1cm} (4.6)$$

and then approximate $D(t,\hat{\alpha})$ by $D(t,\hat{a})$. Thus (4.5) can now be approximated by

$$[- \sum_{t=1}^{N} \psi'(r(t,\hat{\alpha})) \cdot D(t,\hat{\alpha})D'(t,\hat{\alpha})] (\hat{\alpha} - \hat{\alpha})$$

$$= \sum_{t=1}^{N} D(t,\hat{\alpha})\psi(r(t,\hat{\alpha})) \hspace{1cm} (4.7)$$

Since we know from Corollary 2.1 that $r(t,\alpha_o)$ and $D(t,\alpha_o)$ are independent, we would expect that for not-too-small a sample size, $r(t,\hat{\alpha})$ and $D(t,\hat{\alpha})$ are approximately independent. Therefore, it should be reasonable to simplify (4.7) a bit more as follows:

$$-\hat{A}(\psi,F) \cdot \left[ \sum_{t=1}^{N} D(t,\hat{\alpha})D'(t,\hat{\alpha}) \right] (\hat{\alpha} - \hat{\alpha})$$

$$= \sum_{t=1}^{N} D(t,\hat{\alpha})\psi'(r(t,\hat{\alpha})) \hspace{1cm} (4.8)$$

where $\hat{\psi}(\psi,F) = N^{-1} \sum_{t=1}^{N} \psi'(r(t,\hat{\alpha}))$ is the natural estimate of $A(\psi,F) =$
\( E \psi'(\epsilon) \). With \( D^T = [D(1,a) \ldots D(N,a)] \) and \( \psi^T = [\psi_S(r(1,a)) \ldots \psi_S(r(N,a))] \), we have the "normal" equation

\[-\tilde{A}(\psi_S,F) \cdot [D^T D] \cdot (\tilde{a} - \tilde{\alpha}) = D^T \psi \quad (4.9)\]

The solution

\[ \tilde{\alpha}_{AM} = \tilde{\alpha} - \tilde{A}^{-1}(\psi_S,F) \cdot (D^T D)^{-1} D^T \psi \quad (4.10) \]

to (4.9) is called a one-step approximate M-estimate, or simply AM-estimate.

The usual NR algorithm is known to have quadratic rate of convergence, while the gradient method has only a linear rate of convergence (Luenberger, 1973). A little consideration reveals that the AM-algorithm should lie between the gradient and NR methods in terms of the derivative information used to obtain the solution. Thus one would expect that the convergence rate of the AM-estimate would lie somewhere between the linear and quadratic rates. Although the exact rate is difficult to compute, our Monte Carlo study shows that AM-algorithm has nearly a quadratic rate of convergence. This is hardly surprising in view of the following fact: it can be shown that the second order DT terms we discard in the AM-algorithm is asymptotically negligible, therefore the AM- and the NR-algorithms virtually yield the same results for large sample sizes. We might also add that the NR-algorithm sometimes exhibits a well-known overshooting or oscillation problem when the preliminary is far away from the true value (cf., Chambers, 1973), whereas the AM-algorithm does not suffer from these disadvantages. For more detail about the algorithm issues, the reader is referred to Lee (1981).
Recursive computations of $r(t, \alpha)$ and $D(t, \alpha)$

In order to evaluate the AM-estimate in (4.9), we need to compute $r(t, \alpha)$ and $D(t, \alpha)$ at the current parameter value $\alpha$. The special structure of ARMA model allows for a particularly simple recursive evaluation of the gradient (unlike those in general nonlinear regression problems). Referring to Lemma 2.2 we note that both $r(t, \alpha)$ and $D(t, \alpha)$ satisfy AR-type difference equations. Therefore $r(t, \alpha)$ and $D(t, \alpha)$ can be computed recursively provided appropriate initial conditions are used to start the recursions. The shift property established in Lemma 2.3 also suggests a saving of computational effort: instead of computing the entire $(p+q+1)$-dimensional DT vector sequence, only two sequences are required, one for the AR part called $\{v(t)\}$ and the other for the MA part called $\{w(t)\}$. The other elements of the DT vector are just shift versions of the sequences $\{v(t)\}$ and $\{w(t)\}$; see equations (4.13) - (4.15) to follow shortly.

Given some reasonable initial conditions

$$
\begin{align*}
\mathbf{r}^T &= [r(-d_0+1), \ldots, r(0)] \\
\mathbf{v}^T &= [v(-p+1), \ldots, v(0)] \\
\mathbf{w}^T &= [w(-q+1), \ldots, w(0)]
\end{align*}
$$

(4.11)

where $d_0 = \text{MAX}(p, q)$, then $\{r(t, \alpha) = r(t)\}$, $\{v(t)\}$, and $\{w(t)\}$ can be evaluated recursively as follows:

for $1 \leq t \leq N$,

$$
r(t) = y(t) + \sum_{i=1}^{p} \phi_i y(t-i) - \sum_{j=1}^{q} \theta_j r(t-j) - \gamma ;
$$

(4.12)
for \( 1 \leq t \leq N + p - 1 \),

\[
v(t) = r(t-p) - \sum_{i=1}^{p} \phi_i v(t-i) - \gamma/S_\theta \quad \text{; (4.13)}
\]

and for \( 1 \leq t \leq N + q - 1 \),

\[
w(t) = -r(t-q) - \sum_{j=1}^{q} \theta_j w(t-j) \quad . \quad \text{ (4.14)}
\]

And the DT vector can now be formed

\[
D^T(t) = [\Delta_y, \Delta_{\psi}, 1(t), \ldots, \Delta_{\phi}, p(t), \Delta_{\theta}, 1(t), \ldots, \Delta_{\theta}, q(t)] \quad \text{ (4.15)}
\]

with \( \Delta_y = -1/S_\theta \), \( \Delta_{\psi}, i(t) = v(t+p-i) \) and \( \Delta_{\theta}, j(t) = w(t+q-j) \).

Preliminary estimates and initial conditions

In order to realize the iterative procedure used to compute the AM-estimate we need a good preliminary estimate \( \hat{\alpha}^{(0)} \) for the ARMA parameter \( \alpha = (\gamma, \phi, \theta) \), and some reasonable initial conditions for the recursions (4.15) - (4.17).

For the initial estimate of ARMA coefficients \( \phi \) and \( \theta \) we first compute an initial robust location estimate \( \hat{\mu}^{(0)} \). In our algorithm, we choose \( \hat{\mu}^{(0)} \) to be the median of the data sequence. Now the initial \( \hat{\phi}^{(0)} \) and \( \hat{\theta}^{(0)} \) are obtained by applying Durbin's (1959) algorithm to the centered data \( \{\hat{y}(t) = y(t) - \hat{\mu}^{(0)}\} \). In view of the relationship \( \gamma = \mu \cdot (1 + \sum_{i=1}^{p} \phi_i) \) given by (3.8), it is natural to choose \( \hat{\gamma}^{(0)} = \hat{\mu}^{(0)}[1 + \sum_{i=1}^{p} \hat{\phi}_i^{(0)}] \) as an initial estimate of the intercept.

Now that the preliminary estimate \( \hat{\alpha}^{(0)} = [\hat{\gamma}^{(0)}, \hat{\phi}^{(0)}, \hat{\theta}^{(0)}] \) is available, we can estimate the initial conditions \( \tilde{r}, \tilde{v}, \) and \( \tilde{w} \) in (4.11) by doing backforecasting on the centered sequence \( \{\hat{y}(t)\} \) just as described.
in Chapter 7 in Box and Jenkins (1976). For cases with large sample
sizes and small parameter values, we would expect that the effect of the
initial conditions on the estimates to be relatively small. Our Prel-
liminary ARMA(1,1) Monte Carlo runs, with sample size $N = 100$,
intercept $\gamma = 0$ and ARMA(1,1) coefficients $0.5 \leq |\phi|, |\theta| \leq 0.8$, support
the above assumption. Therefore in our Monte Carlo studies we simply
choose the initial conditions to be their prior mean values, i.e., for
t $\leq 0$, set $r(t) = 0$, $v(t) = \mu/S_{\theta}$, and $\omega(t) = 0$.

**Estimation of location parameter $\mu$**

In some situations, we are more interested in the location parameter
$\mu$. For instance, consider estimating the location of the data sequence
$\{y(t)\}$

$$y(t) = \mu + v(t)$$  \hspace{1cm} (4.16)

where instead of having i.i.d. errors, we suspect that $v(t)$ is generated
by some ARMA process, i.e.,

$$v(t) + \sum_{i=1}^{p} \phi_i v(t-i) = \epsilon(t) + \sum_{j=1}^{q} \theta_j \epsilon(t-j) .$$  \hspace{1cm} (4.17)

Then Equation (4.16) can now be expressed as

$$[y(t)-\mu] + \sum_{i=1}^{p} \phi_i [y(t-i)-\mu] = \epsilon(t) + \sum_{j=1}^{q} \theta_j \epsilon(t-j)$$  \hspace{1cm} (4.18)

which is exactly an ARMA setup with the location parameter $\mu = \gamma/(1+ \sum_{i=1}^{p} \phi_i)$
as in Eq. (3.8).

In order to estimate $\mu$, there are several ways to proceed, for more
in-depth discussion about robust location estimates, see Andrews et. al.
(1972). Here we propose a natural way of estimating location parameters
by the relationship

\[ \hat{\mu}_{AM} = \hat{\gamma}_{AM} \frac{S'_{\phi,AM}}{p} \]  

(4.19)

where \( S'_{\phi,AM} = 1 + \sum_{i=1}^{p} \hat{\phi}_{i,AM} \). Martin (1978) called \( \hat{\mu}_{AM} \), the **natural M-estimate of location** in contrast to Huber's **ordinary M-estimate** \( \hat{\mu}_{OM} \) which solves the location estimating equation

\[ \sum_{t=1}^{N} \psi_{S}(y(t) - \hat{\mu}_{OM}) = 0. \]  

(4.20)

It can be shown (Lee, 1981) that the ordinary location M-estimate \( \hat{\mu}_{OM} \) has surprisingly low efficiency toward ARMA errors generated by heavy-tailed innovations; whereas the **natural location M-estimate** \( \hat{\mu}_{AM} \) has high efficiency robustness toward both Gaussian and non-Gaussian ARMA process errors.

**AM-algorithm summaries**

**Step 1.** Obtain a robust location estimate \( \tilde{\mu} \) from the original data sequence \( \{y(t)\} \) and compute the centered sequence \( \{\tilde{y}(t) = y(t) - \tilde{\mu}\} \);

**Step 2.** Apply Durbin's algorithm to \( \{\tilde{y}(t)\} \) to get a good preliminary estimate \( \tilde{\phi} \) and \( \tilde{\theta} \); compute \( \tilde{\gamma} = (1 + \sum_{i=1}^{p} \tilde{\phi}_{i})\tilde{\mu} \), so we have \( \tilde{\alpha} = (\tilde{\gamma}, \tilde{\phi}, \tilde{\theta}) \);

**Step 3.** Use \( \tilde{\alpha} \) and \( \{\tilde{y}(t)\} \) to do backforecasting and obtain reasonable starting residuals \( r(-d+1), \ldots, r(0) \), with \( d_0 = \text{MAX}(p, q) \); and good starting DT vector elements \( v(-p+1), \ldots, v(0) \) and \( w(-q+1), \ldots, w(0) \), or simply set \( r(t) = 0 \), \( v(t) = \tilde{\mu}/(1 + \sum_{j=1}^{q} \tilde{\theta}_{j}) \) and \( w(t) = 0 \) for \( t < 0 \);

**Step 4.** Evaluate the estimated residuals \( \{r(t)\} \) using the following recursion, for \( 1 \leq t \leq N \)
\[ r(t) = \tilde{y}(t) + \sum_{i=1}^{p} \phi_i \tilde{y}(t-i) - \sum_{j=1}^{q} \tilde{\theta}_j r(t-j); \]

**Step 5.** Evaluate the estimated DT vector elements \( \{v(t)\} \) and \( \{w(t)\} \) using the following recursions,

for \( 1 \leq t \leq N + p \)

\[ v(t) = r(t-p) - \sum_{i=1}^{p} \phi_i v(t-i) - \tilde{\gamma}/(1 + \sum_{j=1}^{q} \tilde{\theta}_j) \]

and for \( 1 \leq t \leq N + q \)

\[ w(t) = r(t-q) - \sum_{j=1}^{q} \tilde{\theta}_j w(t-j). \]

Then form the DT vector, for \( 1 \leq t \leq N \),

\[ D^T(t) = [-1/(1 + \sum_{j=1}^{q} \tilde{\theta}_j), v(t+p-1), \ldots, v(t), w(t+q-1), \ldots, w(t)]; \]

**Step 6.** Obtain a robust scale estimate, e.g.,

\[ \hat{s} = \text{MAD}(r(t))/0.6745 \]

from the estimate residual sequences \( \{r(t)\}; \)

**Step 7.** Choose an appropriate psi-function with proper tuning constant and evaluate the sequences \( \{\psi_S(r(t))\} \) and \( \{\psi_S^t(r(t))\} \), then compute

\[ A(\psi_S, F) = \frac{1}{N} \sum_{t=1}^{N} \psi_S^t(r(t)); \]

**Step 8.** Solve for the one-step improved estimate \( \hat{\alpha} = (\hat{\gamma}, \hat{\phi}, \hat{\theta}) \) in the following matrix equation

\[ -A(\psi_S, F) [ \sum_{t=1}^{N} D(t)D^T(t) ] (\hat{\alpha} - \alpha) = \sum_{t=1}^{N} D(t) \psi_S(r(t)) \]

and evaluate \( \hat{\mu} = \hat{\gamma}/(1 + \sum_{i=1}^{q} \phi_i); \)
Step 9. Use some stopping rule to check for convergence of $\hat{\alpha}$, if so, STOP!

Step 10. Replace $\hat{\alpha}$ by $\tilde{\alpha}$ and $\{\tilde{y}(t)\}$ by $\{\tilde{y}(t) + \tilde{\mu} - \hat{\mu}\}$, then go to Step 3 for next iteration.

The above AM-algorithm becomes an ALS-algorithm if we replace the psi-function in Step 7 by the identity function. Meanwhile Step 5 is not required because there is no longer a scaling problem.

5. ASYMPTOTIC EFFICIENCY RESULTS

Consider estimating the ARMA model parameter $\beta = (\mu, \phi, \theta)$ with $\hat{\beta} = (\hat{\mu}, \hat{\phi}, \hat{\theta})$ where $\mu$ is estimated as in (4.19). It can then be shown that under certain regularity conditions and a finite variance assumption, $\hat{\beta}$ is consistent and asymptotically normal. The asymptotic covariance matrix for AM-estimate is of the following block diagonal form (see Lee and Martin, 1982a)

$$V_M = \begin{bmatrix} V_{\mu} & 0^T \\ 0 & C(\phi, \theta) \end{bmatrix} \cdot V_{loc}(\psi, F)/\sigma^2$$

(5.1)

where $V_{loc}(\psi, F) = E\psi^2(\varepsilon)/E\psi'(\varepsilon)$ is the asymptotic variance of the ordinary location $M$-estimate, and $v_\mu$ is the asymptotic variance of the LS location estimate:

$$v_\mu = [(1 + \sum_{j=1}^q \theta_j)/(1 + \sum_{i=1}^p \phi_i)]^2 \cdot \sigma^2 = (S_\theta/S_\phi)^2 \cdot \sigma_c^2 \cdot \sigma_d^2.$$  

(5.2)

The $(p+q) \times (p+q)$ matrix $C(\phi, \theta)$ is the asymptotic covariance matrix for the LS estimate of $\phi$ and $\theta$ when $\mu = 0$; it depends only upon $\phi$ and $\theta$, which exhibits an asymptotic distribution-free property of the least-
squares. On the other hand, the presence of the factor $V_{10C}(ψ,F)/σ^2$ in $V_M$ shows that heavy-tailed innovations distribution can give rise to increased precision in the ARMA model case, just as was reported for the pure autoregression case by Martin (1979, 1981).

For $ψ$-function equals to the identity function, we obtain an approximate least-squares (ALS) estimate with asymptotic covariance matrix

$$V_{LS} = \begin{bmatrix} v & 0 \\ 0 & C(ϕ,θ) \end{bmatrix}^{-1} \cdot \frac{1}{i(f)σ^2} \quad \cdot \quad (5.3)$$

If we know the innovations density $f$ then we can choose $ψ$-function to be the efficient score function, i.e., choose $ψ(x) = -f'(x)/f(x)$. The AM-estimate now becomes asymptotically the maximum likelihood estimate (MLE) with the asymptotic covariance matrix being the Cramer-Rao matrix derived in (3.21), i.e.,

$$V_{MLE} = \begin{bmatrix} v & 0 \\ 0 & C(ϕ,θ) \end{bmatrix}^{-1} \cdot \frac{1}{i(f)σ^2} \quad \cdot \quad (5.4)$$

The differences between $V_M$, $V_{MLE}$, and $V_{LS}$ in (5.1) - (5.4) appear only in the factors $V_{10C}(ψ,F)/σ^2$ and $1/i(ψ,σ^2)$. To compare the performance of the LS- and M-estimates, we need a definition of efficiency (both relatively and absolutely) for vector parameter estimates. Among the possibilities are linear combination efficiency, trace efficiency, and determinational efficiency. In comparing these measures for pure autoregression, Zeh (1979) found that they gave similar results. In this paper, we use the determinantal efficiency measure.
The determinantal efficiency is defined here as the $k^{th}$ root of the ratio of the determinants of the covariance matrices of the two estimates. Thus, for example, the relative asymptotic efficiency of the $\hat{\beta}_{AM}$ with respect to the $\hat{\beta}_{ALS}$ is

$$\text{REFF}(M,LS) = (|V_{LS}|/|V_{M}|)^{1/k} \tag{5.5}$$

where $k = p + q + 1$ is the number of the parameters estimated.

Similarly, the absolute asymptotic efficiency of the $\hat{\beta}_{AM}$ and the $\hat{\beta}_{ALS}$ are defined as

$$\text{AEFF}(M) = (|V_{MLE}|/|V_{M}|)^{1/k} \tag{5.6}$$

and

$$\text{AEFF}(LS) = (|V_{MLE}|/|V_{LS}|)^{1/k} \tag{5.7}$$

For the ARMA case, (5.5) - (5.7) reduce to

$$\text{REFF}(M,LS) = \sigma_e^2 / V_{loc}(\psi,F) \tag{5.8}$$

$$\text{AEFF}(M) = 1/[i(f) \cdot V_{loc}(\psi,F)] \tag{5.9}$$

and

$$\text{AEFF}(LS) = 1/[i(f) \cdot \sigma_e^2] \tag{5.10}$$

The above asymptotic efficiency results are exactly the same as those for pure autoregression cases (Martin, 1979). These values depend on the type of psi-functions used, and also on the innovations distributions. For this paper, we consider two popular psi-functions, namely Huber's monotone psi-function.
\[
\psi_H(x) = \begin{cases} 
  x , & \text{for } |x| \leq cs \\
  cs \cdot \text{SGN}(x) , & \text{for } |x| > cs 
\end{cases} \quad (5.11)
\]

and Tukey's redescending bi-square psi-function

\[
\psi_{BS}(x) = \begin{cases} 
  x(1 - \frac{x^2}{c^2s^2})^2 , & \text{for } |x| \leq cs \\
  0 , & \text{for } |x| > cs 
\end{cases} \quad (5.12)
\]

The scale parameter \( s \) is chosen to be \( \text{MAD}/.6745 \) for the innovations density and the tuning constants are chosen to be \( c_H = 1.5 \) and \( c_{BS} = 6.0 \) so that the AM-estimate will have high efficiency when the innovations density is Gaussian.

To see how the AM-estimate outperforms the ALS-estimate under non-gaussian heavy-tailed innovations, we consider the popular contaminated normal innovations density (Tukey, 1960)

\[
CN(\delta, \sigma^2) = (1-\delta)N(0,1) + \delta N(0, \sigma^2) \quad (5.13)
\]

where \( \delta \) is the amount of contamination which is typically small and \( \sigma^2 \) is the contaminated variance which is usually larger than the nominal variance 1.

With \( \delta = 0.1, 0.25 \) and \( \sigma^2 = 9, 36, 100 \), we compute the efficiencies (5.8) - (5.10) for the \( CN(\delta, \sigma^2) \) innovations distributions. The results are displayed in Table 1 for both Huber's M-estimate with \( c_H = 1.5 \) and Tukey's M-estimate with \( c_{BS} = 6.0 \). It is obvious from the Table that the LS estimate is seriously lacking in efficiency robustness toward heavy-tailed innovations while the M-estimate is highly robust.
Table 1: Asymptotic Absolute and Relative Efficiencies for the LS- and M-estimates of Location under Various Error Distributions.

<table>
<thead>
<tr>
<th>Innovations</th>
<th>$\psi_H(c=1.5)$</th>
<th>$\psi_{BS}(c=6.0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{AEFF(LS)}$</td>
<td>$\text{AEFF(M)}$</td>
</tr>
<tr>
<td>$N(0,1)$</td>
<td>1.000</td>
<td>.964</td>
</tr>
<tr>
<td>CN(.1,9)</td>
<td>.698</td>
<td>.968</td>
</tr>
<tr>
<td>CN(.1,36)</td>
<td>.276</td>
<td>.875</td>
</tr>
<tr>
<td>CN(.1,100)</td>
<td>.111</td>
<td>.819</td>
</tr>
<tr>
<td>CN(.25,9)</td>
<td>.564</td>
<td>.904</td>
</tr>
<tr>
<td>CN(.25,36)</td>
<td>.173</td>
<td>.680</td>
</tr>
<tr>
<td>CN(.25,100)</td>
<td>.062</td>
<td>.567</td>
</tr>
</tbody>
</table>
6. MONTE CARLO RESULTS FOR ARMA(1,1) PROCESSES

In this section we present some small-sample Monte Carlo results for the estimation of ARMA(1,1) model

\[ y(t) - \mu + \phi [y(t-1) - \mu] = \epsilon(t) + \theta \epsilon(t-1) \]  

(6.1)

where \( \beta = (\mu, \phi, \theta) \) is the parameter to be estimated.

In order to obtain information concerning the validity of the asymptotic performance of the ALS and the AM-estimates for finite sample sizes, a Monte Carlo (100 replications) experiment was performed for a fixed sample size of \( N = 100 \) and the true parameter values were chosen to be \( \mu_0 = 0, \phi_0 = -0.8, \) and \( \theta_0 = 0.5 \) (these parameter values were used in Hannan, 1969). Our main goal here is to study the efficiency robustness of the AM-estimate under both Gaussian and non-Gaussian innovations processes. \( N(0,1) \{ \epsilon(t) \} \) were chosen to generate Gaussian \( \{ y(t) \} \), while \( CN(\delta, \sigma^2) \{ \epsilon(t) \} \) were used to generate non-Gaussian \( \{ y(t) \} \) according to model (6.1). \( \delta = 0.1, 0.25 \) and \( \sigma = 3, 6, 10 \) were chosen to form various pairs of \( (\delta, \sigma^2) \) to represent lightly and moderately heavy-tailed innovations \( CN(\delta, \sigma^2) \). Details concerning the random number generation of innovations \( \{ \epsilon(t) \} \) are given in Lee (1981).

The estimators for \( (\phi, \theta) \) included in the Monte Carlo studies are

1. ALS: Approximate least-squares estimate;
2. AM-H: Approximate M-estimate with \( \psi(\cdot) \) of the Huber type; and
3. AM-B: Approximate M-estimate with \( \psi(\cdot) \) of the bisquare type. They are computed using the AM-algorithm proposed in Section 4.

As for the location parameter \( \mu \), we find that the so-called natural M-estimate \( \hat{\mu}_{AM} \) obtained from the AM-algorithm always outperforms the
popular Huber's ordinary M-estimate under non-Gaussian situations. And for heavy-tailed IO cases, \( \hat{\mu}_{OM} \) gives both large bias and huge variance, whereas \( \hat{\mu}_{AM} \) take the ARMA error structure into account, therefore it produces robust estimate for the location estimate \( \mu \). The above location results may be found in a forthcoming paper by Lee and Martin (1982b).

For the remainder of this section, we will treat \( \mu \) as a nuisance parameter and focus our attention on the estimation of ARMA(1,1) parameters \( \phi \) and \( \theta \).

In order to compare the ALS and AM-estimates for vector parameter \( (\phi, \theta) \), we still adopt the determinational efficiency measure. Let \( M_{\alpha} \) be the Monte Carlo mean-squares-errors matrix for some estimator \( \hat{\alpha} = (\hat{\phi}, \hat{\theta}) \), then we here compute the absolute and relative determinantal Monte Carlo efficiencies as follows:

\[
MCAEFF(\hat{\alpha}) = \sqrt{V_{MLE}[1/|M_{\alpha}|^{1/2}]}, \quad \text{(6.2)}
\]

and

\[
MCREEFF(\hat{\alpha}_1, \hat{\alpha}_2) = \sqrt{\frac{|M_{\alpha_2}|}{|M_{\alpha_1}|}} \cdot \sqrt{V_{MLE}[1/|M_{\alpha_1}|^{1/2}]}, \quad \text{(6.3)}
\]

The Monte Carlo efficiency results are displayed in Table 2, where several IO distributions are considered. At this point, the reader is referred to Table 1 for the asymptotic results. The overall picture shows that the M-estimate is clearly superior to the LS estimate. Just as in the AR(1) IO case (Denby and Martin, 1979), the least-squares estimate gives much larger Monte Carlo bias. We also found that the Monte Carlo variance is smaller for the M-estimate, this is to be ex-
pected. Therefore, if the relative efficiencies were based on the Monte Carlo covariance matrix $V_{\alpha}$ only instead of $M_{\alpha}$, the M-estimate is still more efficient than the least-squares.

In Figure 1, we show two sets of efficiency results, they are all plotted against the contaminated normal standard deviation $\sigma$. The first set is based on the Monte Carlo results for $\text{CN}(\delta, \sigma^2)$, with amount of contamination $\delta = 0.1$, and the second set is for $\delta = 0.25$. Three plots were included in each set: (1) Monte Carlo efficiencies for the $\phi$-estimate, (2) Monte Carlo efficiencies for the $\theta$-estimate, and (3) Monte Carlo determinantal efficiencies. The $\sigma = 1$ situation is the case of Gaussian innovations, and the least-square performs only slightly better than the M-estimate in this Gaussian case. As $\sigma$ increases, the gap between the LS-curves and the M-estimate curves becomes larger, which indicates that the least-squares lacks efficiency robustness toward heavy-tailed innovations distributions. Except for the $\text{CN}(.25, 9)$ case, $AM-B$ always outperforms $AM-H$, especially for the cases with large $\sigma$ values, which is just as would be expected from the asymptotic results.
Table 2  Monte Carlo Absolute and Relative Efficiencies for the ALS and AM-estimates of \((\phi, \theta)\) for Model 10 with True \((\phi, \theta) = (-0.8, 0.5)\) and Sample Size = 100.

<table>
<thead>
<tr>
<th>Innovations</th>
<th>ALS</th>
<th>AM-H(c=1.5)</th>
<th>AM-B(c=6.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AEFF</td>
<td>AEFF</td>
<td>REFF(M,LS)</td>
</tr>
<tr>
<td>N(0,1)</td>
<td>.648</td>
<td>.635</td>
<td>.908</td>
</tr>
<tr>
<td>CN(.1,9)</td>
<td>.548</td>
<td>.713</td>
<td>1.301</td>
</tr>
<tr>
<td>CN(.1.36)</td>
<td>.226</td>
<td>.629</td>
<td>2.783</td>
</tr>
<tr>
<td>CN(.1.100)</td>
<td>.090</td>
<td>.537</td>
<td>5.967</td>
</tr>
<tr>
<td>CN(.25,9)</td>
<td>.390</td>
<td>.604</td>
<td>1.549</td>
</tr>
<tr>
<td>CN(.25,36)</td>
<td>.133</td>
<td>.434</td>
<td>3.263</td>
</tr>
<tr>
<td>CN(.25,100)</td>
<td>.054</td>
<td>.307</td>
<td>5.685</td>
</tr>
</tbody>
</table>
Monte Carlo efficiency results for ARMA(1,1) estimates

(a) $\phi$ - parameter

(b) $\theta$ - parameter

(c) Determinantal efficiencies

Figure 1
BIBLIOGRAPHY


