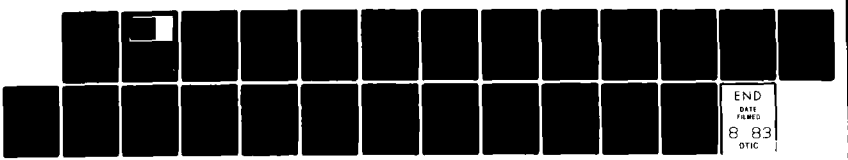
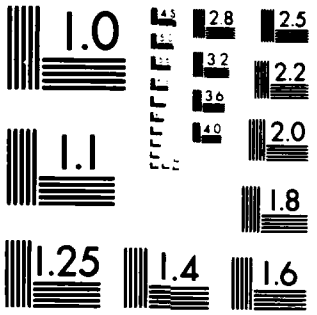
AD-A130 500 THE ESTIMATION OF TIME SERIES MODELS PART 1 YET ANOTHER 1/1
ALGORITHM FOR THE..(U) WISCONSIN UNIV-MADISON
MATHEMATICS RESEARCH CENTER G TUNNICLIFFE-WILSON
UNCLASSIFIED JUN 83 MRC-TSR-2528 DAAG29-80-C-0041 F/G 12/1 NL



END
DATE FILMED
8 83
DTIC



MRC Technical Summary Report #2528

THE ESTIMATION OF TIME SERIES MODELS PART I. YET ANOTHER ALGORITHM FOR THE EXACT LIKELIHOOD OF ARMA MODELS

G. Tunnicliffe-Wilson

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

June 1983

(Received April 27, 1983)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

THE ESTIMATION OF TIME SERIES MODELS. PART I. YET ANOTHER ALGORITHM FOR THE EXACT LIKELIHOOD OF ARMA MODELS

G. Tunnicliffe-Wilson

Technical Summary Report #2528
June 1983

ABSTRACT

This paper presents a method for calculating the likelihood function of autoregressive-moving average (ARMA) models for time series data. Model estimation requires maximization of the likelihood, and to assist in this, a method for calculating derivatives of the function is also presented. The computational efficiency is competitive with that of other algorithms for this purpose. Extensions which allow for seasonal models, missing data, and the estimation of a data transformation are also described.

AMS (MOS) Subject Classifications: 62M10, 65U05

Key Words: Autoregressive-moving average model; Time series estimation; Exact likelihood

Work Unit Number 4 (Statistics and Probability)

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

This paper is concerned with the statistical analysis of time series, i.e. records of observations of variables in fields as diverse as economics, engineering, meteorology ... . The class of ARMA models has been proved to be most successful in representing the dynamic structure of such series and are useful in both prediction and control, and in the investigation of relationships between series. The models are in the form of discrete time difference equations relating present values of the series to past values, with a prediction error term. It is important to have precise tools for the estimation of the coefficients in these equations, and for discriminating between different models. Computation of the exact likelihood of a model is important in this respect, and so is estimation of the coefficients by maximizing the likelihood. Several algorithms for computing the likelihood are now available, and the one presented in this paper is highly competitive on the grounds of numerical accuracy and efficiency. It has been implemented and proved useful for modelling a wide variety of data.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

THE ESTIMATION OF TIME SERIES MODELS. PART I. YET ANOTHER ALGORITHM FOR THE EXACT LIKELIHOOD OF ARMA MODELS

G. Tunnicliffe-Wilson

1. INTRODUCTION

The basic autoregressive-moving average model of orders p and q, or ARMA (p,q) model, for a time series x_t, is defined by

$$w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (1.1)$$

where $w_t = x_t - c$, the constant c representing the expectation of x_t. The unobserved series a_t is assumed to be a sequence of independent random variables from a Normal $(0,\sigma^2)$ distribution. In terms of the backward shift operator notation of Box and Jenkins (1976) the model is conveniently written $\phi(B)w_t = \theta(B)a_t$ where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$. The coefficients of the model are constrained by a stationarity condition that $\phi(B) \neq 0$ for $B \leq 1$. This ensures that the process reaches statistical equilibrium or stationarity. A further condition is that $\theta(B) \neq 0$ for $|B| < 1$, which allows identification of a_t with the linear innovations (or one step ahead prediction errors) of the process, i.e. $a_t = x_t - E(x_t|x_{t-1},x_{t-2}\ldots)$. This includes the borderline case when $\theta(B)$ may have zeros on $|B| = 1$.

Given a finite sample $x_1,\ldots,x_n$ (and assuming the process has reached stationarity), the likelihood of the model parameters $\beta = (c,\sigma^2,\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)$ may be formally presented using the covariance matrix Γ of the sample, which has elements $\Gamma_{ij} = \mathrm{cov}(x_i,x_j) = \gamma_{|i-j|}$. We may set $\Gamma_{ij} = \sigma^2 v_{ij}$ where V depends only on $\phi = (\phi_1 \cdots \phi_p)$ and $\theta = (\theta_1 \cdots \theta_q)$. Then the likelihood is

$$L(\beta) \propto \sigma^{-n} M^{-1/2} \exp(-\tfrac{1}{2} Q/\sigma^2)$$

where $Q = w'V^{-1}w$ is a quadratic form in $w = (w_1 \cdots w_n)'$ and $M = \det V$.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

When maximized w.r.t. σ^2, $L \propto D^{-n/2}$ where $D = M^{1/n}Q$. The M.L.E.s of c,ϕ,θ may then be found by minimizing D, which we shall call the deviance. The factor $M(\phi,\theta,n)$ does not depend on the data, and tends to a limit $M(\phi,\theta)$ as $n \to \infty$. Consequently $M^{1/n} \to 1$ for any fixed ϕ and θ, and is often neglected in large samples, but is important when n is small, or ϕ,θ are close to their constraint boundaries.

The form Q may be evaluated by many devices, most of which involve a sum of squares function roughly of the form $Q = \Sigma\, a_t^2$, where the terms a_t are regenerated from the data via the model (1.1). Lack of knowledge of x_t for $t < 0$ (the end effect) however, prevents exact evaluation of $a_1 \cdots a_n$, so they are estimated by various means ranging from the ingenious backforecasting scheme of Box and Jenkins (1976) to the general methods of Newbold (1974) and Ljung and Box (1979). More recently Q has been calculated as a weighted sum of squares of innovations based on the finite data set, $b_t = x_t - E(x_t|x_{t-1} \cdots x_1)$, so that

$$Q = \sum_1^n \left(\frac{b_t}{h_t}\right)^2 , \quad M = \prod_1^n h_t .$$

Although $a_t - b_t \to 0$ and $h_t \to 1$ as t increases, the residuals may differ appreciably for small values of t, e.g. $b_1 = x_1$ and $h_1^2 = \mathrm{var}(x_t)/\sigma^2$ always. The projection techniques for constructing b_t may be based on direct use of Cholesky factorization as in Ansley (1979) or on the Kalman filter as in Gardner et al. (1980). They are computationally efficient, and by exploiting the special structure of the ARMA model within the Kalman filter, Mélard (1983) has produced an algorithm which is extremely economical in its use of arithmetical operations and core storage. The algorithm presented here follows more closely the classical approach of Ljung and Box but exploits simple special methods which have been long known for pure AR(p) and MA(q) models. This simplicity allows convenient analytic computation of the derivatives of the deviance

for use in minimization routines. Not only is this more accurate than the use of numerical derivatives, it may also be computationally much cheaper when the orders p,q are greater than one.

2. DEVELOPMENT OF THE ALGORITHM

It is convenient to reproduce first the algorithms for the cases of the pure AR(p) and MA(q) models since these are the basis for the general ARMA(p,q) case.

For the AR(p) model define series

$$a_t = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p}, \qquad t = 1 \cdots n$$

$$b_t = w_t - \phi_1 w_{t+1} - \cdots - \phi_p w_{t+p}, \qquad t = (1-p) \cdots 0$$

where the values of w_t for $t < 0$ are all taken as 0. Then

$$Q = \sum_{t=1}^{n} a_t^2 - \sum_{t=1-p}^{0} b_t^2 .$$

The factor M remains constant for $n > p$ in this case, so is found from the case $n = p$ by extracting from Q the elements of $V^{-1} = F$ as

$$F_{ij} = \sum_{k=0}^{p-i} f_k f_{k+\ell} - \sum_{k=j}^{p-\ell} f_k f_{k+\ell} \quad \text{for} \quad i,j = 1 \cdots p \quad \text{and} \quad i > j ,$$

where $f_0 = 1$, $f_k = -\phi_k$ for $k = 1 \cdots p$ and $\ell = i - j$.

M may then be obtained as 1/det F, though we shortly provide a more efficient procedure for its calculation.

The above expression for Q is to be found in Ljung and Box (1979) equn. (4.4) and is derived by exploiting the time reversal symmetry of the model as in Box and Jenkins (1976), Appendix A7.5.

The calculation of det F follows a scheme based on the reverse of the Durbin-Levinson algorithm as presented by Tunnicliffe-Wilson (1979), p. 303. It is based on the reduction of $f_0 \cdots f_p$ to a new set, say $f'_0 \cdots f'_{p'}$, where $p' = p - 1$, by $f'_0 = 1$; $f'_j = (f_j - f_k f_{k-j})/\tau_p$ for $j = 1 \cdots p'$ where $\tau_p = (1 - f_p^2)$. This is repeated with f',p' replacing f,p until $p' = 1$. Then

$$\det F = \prod_{k=1}^{p} \tau_k^k . \qquad (2.1)$$

This also provides a check on the stationarity condition which is satisfied iff $\tau_k > 0$ - see Duffin (1969). For the MA(q) model, the first step is to introduce variables $(a_{1-q},\ldots,a_0) = a_L'$ corresponding to the pre-observation period innovations. If these are supplied, the remaining set $(a_1 \cdots a_n) = a_R'$ may be regenerated by

$$a_t = w_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}, \quad t = 1 \cdots n \tag{2.2}$$

and a sum of squares calculated:

$$S(a_L) = \sum_{1-q}^{n} a_t^2 .$$

Then $Q = \min_{a_L} S$. The minimizing set $\hat{a}_L$ may be found by linear least squares and M obtained as the determinant of the matrix in the least squares equations. Details are given in Box and Jenkins (1976) Appendix A7.4. Following ideas in Ljung and Box (1979) it is convenient to use the backforecasts $(w_{1-q},\ldots,w_0) = w_L'$ as an equivalent set of variables in place of a_L. The least squares equations are then of the form

$$X'X\hat{w}_L = X'a$$

where $a' = (a_L', a_R')$ with $a_L = 0$, and $X_{ij} = \pi_{i-j}$, the sequence π_k being the model π-weights. The Jacobian between a_L and w_L is 1, so again $M = \det(X'X)$.

The combined algorithm for the mixed model, ARMA(p,q), again requires the introduction of the set w_L, which is to be estimated by $\hat{w}_L$. In this process it is natural initially to set $w_L = 0$, but because estimation of the model parameters β is iterative, it is convenient to initialize w_L to the value $\hat{w}_L$ from the previous iteration and then to estimate a correction $\delta\hat{w}_L$. Thus we take w_L to be set to some initial value, not necessarily 0. The procedure is then to regenerate

$$\left.\begin{aligned} a_t &= w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}, \quad t = (1-q) \cdots n \\ b_t &= w_t - \phi_1 w_{t+1} \cdots - \phi_p w_{t+p} + \theta_1 b_{t-1} + \cdots + \theta_q b_{t-q}, \quad t = (1-p-q) \cdots (-q) \end{aligned}\right\} \tag{2.3}$$

taking $w_t = a_t = b_t = 0$ for $t < 1 - q$.

Setting $(a_{1-q} \cdots a_n)' = a$ etc. we shall represent (2.3) by

$$a = \frac{\phi(B)}{\theta(B)} w = \pi(B)w; \quad b = \frac{\phi(B^{-1})}{\theta(B)} w = \chi(B)w \tag{2.4}$$

the vectors being of lengths $n + q, p$ respectively in these two equations. The 'sum of squares' function is then

$$S(w_L) = \sum_{t=1-q}^{n} a_t^2 - \sum_{t=1-p-q}^{-q} b_t^2 , \tag{2.5}$$

and again, $Q = \min_{w_L} S = S(\hat{w}_L)$. This minimization requires calculation of the coefficients $(\pi_0,\ldots,\pi_{n+q-1}) = \pi'$ and $(\chi_{-p},\ldots,\chi_{-1}) = \chi'$, by applying the same equations (2.3) to an impulse sequence $I = (1,0,0\ldots)$ in place of w. Thus

$$\pi = \pi(B)I; \quad \chi = \chi(B)I . \tag{2.6}$$

The estimate $\hat{w}_L$ is $w_L + \delta\hat{w}_L$, where $\delta\hat{w}_L$ is the solution of $A\delta\hat{w}_L = -G$, the terms in this equation being given by

$$A_{ij} = \sum_{t=1}^{n+q} \pi_{t-i}\pi_{t-j} - \sum_{t=1-p}^{0} \chi_{t-i}\chi_{t-j}; \quad i,j = 1,\ldots,q \tag{2.7}$$

$$G_i = \sum_{t=1}^{n+q} \pi_{t-i}a_{t-q} - \sum_{t=1-p}^{0} \chi_{t-i}b_{t-q}; \quad i = 1,\ldots,q . \tag{2.8}$$

The minimum value Q may be obtained as $S(\hat{w}_L)$, or as

$$Q = S(w_L) - (\delta\hat{w}_L)'G .$$

The factor M in the likelihood is then given by $M = \det A/\det F$ where F is defined from the Autoregressive parameters as for the pure AR(p) model.

An example - the ARMA(1,1) model.

Taking $w_0 = 0$ calculate $a_t = w_t - \phi w_{t-1} + \theta a_{t-1}$; $t = 1 \cdots n$ and $b_{-1} = 0$. Also $\pi_0 = 1$, $\pi_1 = (\phi - \theta) \cdots$, $\pi_{n+1} = (\phi - \theta)\theta^n$, and $\chi_{-1} = -\phi$. Then $S(w_0) = \sum_1^n a_t^2$,

$$A = 1 + (\phi - \theta)^2(1 + \theta^2 + \cdots + \theta^{2n}) - \phi^2, \quad g = (\phi - \theta)\sum_1^n \theta^{t-1}a_t .$$

$$\hat{w}_0 = -g/A, \qquad Q = S(w_0) - \hat{w}_0 g .$$

$F = 1 - \phi^2$ and on simplifying ,

$$M = A/F = 1 + \begin{cases} (\phi - \theta)^2(1 - \theta^{2n})/\{(1 - \phi^2)(1 - \theta^2)\} & \text{if } \theta \neq 1 \\ n(1 - \phi)/(1 + \phi) & \text{if } \theta = 1 . \end{cases}$$

Verification of the algorithm. Express the model (1.1) using an intermediate series v_t, as $w_t = \theta(B)v_t$ where $\phi(B)v_t = a_t$. Take as before, $v_L = (v_{-q},\ldots,v_0)$, $v_R = (v_1,\ldots,v_n)$. Then we have the MVN pdf

$$f(v) = f(v_L,v_R) \propto \sigma^{-n} \det F^{1/2} \exp -1/2\ S/\sigma^2 \qquad (2.9)$$

where $S = \Sigma\, a_t^2 - \Sigma\, b_t^2$ with a_t, b_t derived from v as for the AR(p) model.

Defining now $(w_L,w_R) = \theta(B)(v_L,v_R)$, we have w_R as precisely the set of observed series values, whereas w_L must be considered as merely a linear transformation of v_L and not as a set of true series values, because in its definition v_t is set to 0 for $t < 1 - q$. The Jacobian of all these transformations is 1, so (2.9) is also the MVN pdf $f(w_L,w_R)$. Also, S as a function of w corresponds exactly to the constructions (2.3) and (2.5). By the properties of the MVN distribution, any set of variables (in this case w_L) may be integrated out by minimizing the exponential term S, which in effect expresses

$$S(w_L) = Q + (w_L - \hat{w}_L)A(w_L - \hat{w}_L)$$

where $Q = S(\hat{w}_L)$. The usual algebra then gives the pdf $f(w_R)$ by replacing S by Q in (2.9) and introducing the factor $1/\det A^{1/2}$ to give the stated form of M.

3. COMPUTATIONAL REQUIREMENTS

We present the dominant terms in the number of multiplications (mults) required for calculating the deviance. It is worthwhile considering here the possibility of 'sparse' seasonal models e.g. the multiplicative operator $(1 - \theta B)(1 - \Theta B^{12})$ used in the airline model. We shall use q for the total number of coefficients in such a moving average operator, e.g. two in this case, and s for the maximum order, e.g. 13 here. We use p in a similar manner to q. Thus to calculate one term a_t needs $p + q$ mults, but we need s 'backforecasts' w_L. Most of the calculations for the series b_t are small compared with those for a_t so are left out. Similarly we shall use n where in some cases we should strictly use $n + q$ e.g. for the length of a_t.

Step (2.3) to regenerate a_t requires $n(p + q)$ mults.

Step (2.6) to obtain π_i requires nq since for $i > p$ the AR terms are not required.

Step (2.7) to set up A requires $\frac{1}{2}ns^2$ mults as it stands and exploiting symmetry. However, following Ljung and Box there are two possible economies. Computing the first column of A, i.e. $A_{i,1}$, $i = 1 \cdots s$ requires only ns mults and other entries may be obtained by $A_{i+1,j+1} = A_{i,j} - \pi_{n+s-i}\pi_{n+s-j}$, requiring few mults. Furthermore the first column in A may itself be found as the first s value in the sequence determined as $\pi(B^{-1})\pi$. This has an advantage for seasonal models of requiring only nq mults, if organized carefully. Similarly, by constructing $\pi(B^{-1})a$ the elements of G can be obtained in nq rather than ns mults. To apply $\pi(B^{-1})$ efficiently in this case would be to compute e.g.

$$e_t = a_t + \theta_1 e_{t+1} + \cdots + \theta_q e_{t+q}, \qquad t = n \cdots 1 - q$$

taking $e_t = 0$ for $t > n$, then

$$G_i = e_{i-q} - \phi_1 e_{i+1-q} - \cdots - \phi_p e_{i+p-q}, \qquad i = 1 \cdots s .$$

Solution of the equations and evaluation of $\det A$, may be performed using a Cholesky decomposition with $(1/6)s^3$ mults. Computation of $\det F$ requires $\frac{1}{2}p^2$ mults which we neglect, giving a minimum number of $n(p + 4q) + (1/6)s^3$ mults.

The main penalty here is the term in s^3, and associated with it is a storage requirement of $\frac{1}{2}s^2$ real numbers for A. For many applications this is no burden, including seasonal models with typical values of $s = 13$. However, in some large organizations models of hourly data with a seasonal period of one week are routinely fitted (using at present the classical backforecasting scheme of Box and Jenkins) and here a value of $s = 168 + 24 + 1 = 193$ is common. In this context the method of Mélard would be expected to have a distinct advantage. The special structure of A does give some hope of treatment by special means such as have been presented by Dickinson (1978) which may reduce the number of multiplications to $O(s^2)$ and storage to $O(s)$.

For a pure moving average multiplicative seasonal model, Hillmer and Tiao (1979) also provide a scheme which reduces the number of operations in this stage to $O(q^3)$.

In circumstances where the sequence π_t can be considered to decay to 0 before $t = n$, it is possible to use the classical backforecasting scheme to generate $\hat{w}_f$, and the method of McLeod (1977) to determine the factor M. The formula (2.1) may be exploited in McLeod's procedure, so that the number of computations is $O(p + q)^2$.

Given that much of the execution of a statistical package is occupied with organizational aspects, computational efficiency in the main algorithm is not always vital. The advent of parallel processing may favor some algorithms which can exploit this feature, and which were previously uncompetitive. Nevertheless, problems are bound to arise which stretch the resources available, and effort put into algorithmic efficiency is then appreciated.

4. COMPUTATION OF DERIVATIVES

The derivative of D wrt the model parameters β may be expressed as

$$M^{1/n}\{(Q/n)(\partial/\partial\beta)(\log\det A - \log\det F) + \partial Q/\partial\beta\} = M^{1/n}G(\beta) \qquad (4.1)$$

where, now introducing β as a parameter,

$$Q(\beta) = \min_{w_L} S(w_L,\beta) = S(\hat{w}_L(\beta),\beta) \ .$$

From this, $\frac{\partial Q}{\partial\beta} = \frac{\partial S}{\partial\beta}(\hat{w}_L,\beta)$, since $\frac{\partial S}{\partial w} = 0$ at $\hat{w}_L$. Now writing e.g. $a\beta_t$ for $\frac{\partial a_t}{\partial\beta}$ we have simply

$$\frac{\partial S}{\partial\beta} = 2(\Sigma\ a_t a\beta_t - \Sigma\ b_t b\beta_t) \ , \qquad (4.2)$$

where the summations run as in (2.5). We now turn to the computation of the sequences $a\beta, b\beta$ and use operator notation such as in (2.4) to represent equations such as (2.3) which are actually used for the computation. Recall that when such equations are used all variables associated with times before the first point in the sequence are taken as 0.

For the model constant term c, $ac = -\pi(B)u$ where u is a sequence of ones from $t = 1$, with zeros for $t \leq 0$. In fact $ac_t = ac_{t-1} - \pi_{t-1}$ for $t > 1$, which saves on computation. (Note that for the sequence π, our indexing convention is chosen to agree with the notation of Box and Jenkins, so that the first term in the sequence π is π_0, even though it is associated with the first time point $t = 1 - q$.)

For all the autoregressive parameters we need generate only two series

$$a\phi = -1/\theta(B)w, \qquad b\phi = -B^{-p}/\theta(B)w \qquad (4.3)$$

from which may be obtained all the required derivatives

$$\partial a_t/\partial\phi_k = a\phi_{t-k}, \qquad \partial b_t/\partial\phi_k = b\phi_{t-(p-k)} \ . \qquad (4.4)$$

The shift B^{-p} in the operator generating $b\phi$ is a formality to ensure merely that we can continue to use the span of p values $b\phi_{1-p-q} \cdots b\phi_{-q}$ which would otherwise contain zeros.

For all the moving average parameters we generate series

$$a\theta = 1/\theta(B)a, \qquad b\theta = 1/\theta(B)b \qquad (4.5)$$

from which

$$\partial a_t/\partial\theta_k = a\theta_{t-k},\ \partial b_t/\partial\theta_k = b\theta_{t-k}\ . \tag{4.6}$$

Thus all the derivatives of S can be constructed with a further $n(p + 2q)$ mults including accumulation of the products, if (4.3) is used as first step in constructing a_t.

Turning now to $(\partial/\partial\beta)\log\det A = \mathrm{tr}(A^{-1}\partial A/\partial\beta)$, the matrix A^{-1} is obtained as a by product of solving the equations for $\delta\hat{w}_L$, with relatively few further operations. The elements of $\partial A/\partial\beta$ are of the form

$$\partial A_{ij}/\partial\beta_k = 2(\Sigma\ \pi_{t-i}\pi\beta_{t-j-k} - \Sigma\ \chi_{t-i}\chi\beta_{t-j-k}) \tag{4.7}$$

where for the autoregressive parameters ϕ_k,

$$\pi\phi = -1/\theta(B)I \quad \text{and} \quad \chi\phi = -B^{-P}/\theta(B)I \tag{4.8}$$

and k is replaced by $p - k$ in the second sum of (4.7), as in (4 4). For the moving average parameters,

$$\pi\theta = 1/\theta(B)\pi \quad \text{and} \quad \chi\theta = 1/\theta(B)\chi\ .$$

Thus only two new series of length n need to be constructed, and the sums in (4.7) may again be formed using the operator $\pi(B^{-1})$ as in the construction of A, with as few as $2nq$ further mults. Finally the trace terms may be accumulated with $\frac{1}{2}\ s^2(p + q)$ mults.

The derivative of $\log\det F$ is $\mathrm{tr}(F^{-1}\partial F/\partial\beta)$ and is required only for the AR parameters. In fact $F^{-1} = V$ which has elements $V_{ij} = v\rho_{|i-j|}$ for $i,j = 1 \cdots p$, where v is the variance of a series following the AR(p) model with $\sigma^2 = 1$, and ρ_k is the acf of the series. These may all be rapidly constructed as described by Tunnicliffe-Wilson (1979), following the construction of $\det F$ in (2.1). Then the required derivative wrt ϕ_k is, after some simplification, given by

$$h_k = -2v \sum_{i=0}^{p'} (p' - i - k)f_i\rho_{|k-i|} \tag{4.9}$$

where $p' = p - 1$. This agrees with the formula obtained from (A7.5.16) and (A7.5.17) in Box and Jenkins (1976) if the former is used to eliminate $\rho_{|k-p|}$.

In total therefore, some $n(p + 4q) + \frac{1}{2}s^2(p + q)$ mults provide all the first derivatives of the likelihood, barely more than is necessary to compute the likelihood itself.

Hessian approximation.

It is possible to obtain an approximation to the second derivatives of D, which is in many cases adequate for use in iterative minimization schemes similar to those of Marquardt (1963). The first step in this approximation is to retain only the term $M^{1/n}\partial^2 Q/\partial\beta_i\partial\beta_j$ on differentiating $D = M^{1/n}Q$. This is reasonable at points away from the boundary of the parameter space determined by the stationarity and invertibility conditions, because then $M \sim 1$ and can be treated as a constant. To determine the matrix $Q\beta\beta$, say with elements $\partial^2 Q/\partial\beta_i\partial\beta_j$, we use the corresponding second derivative matrix of $S(w_L,\beta)$ partitioned according to the two parameter groups w_L and β as

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} .$$

Provided H is evaluated at $\hat{w}_L(\beta)$, it is true that $Q\beta\beta = H_{22} - H_{21}H_{11}^{-1}H_{12}$. It is not necessary to evaluate this explicitly if it is desired to use a quasi-Newton method to calculate a parameter correction $\delta\beta$. Setting $\delta' = (\delta w_L', \delta\beta')$ and $G' = (0', G(\beta)')$, it is sufficient to solve $H\delta = -G$. The zero in G comes from the derivative of S wrt W_L. The correction term δw_L need not be saved, or even evaluated, because after the new parameter set $\beta + \delta\beta$ is constructed, $\hat{w}_L$ will be redetermined at this new set at the start of the next iteration.

A further approximation is in the evaluation of

$$H = 2\{(\Sigma\, a_{t,i}a_{t,j} + \Sigma\, a_t a_{t,ij}) - (\Sigma\, b_{t,i}b_{t,j} + \Sigma\, b_t b_{t,ij})\} . \qquad (4.10)$$

The second sum in each bracket is discarded in the approximation, where e.g. $a_{t,i}$

represents the derivative of a_t with respect to the i'th parameter in the set (w_L, β). If a_t is linear in both parameters, $a_{t,ij}$ is in any case 0, and close to the origin any nonlinearity in the parameters β is not great. Furthermore, at the true parameter values the second sum has expectation zero and magnitude $O(\sqrt{n})$ compared with a magnitude of $O(n)$ for the first term.

It is convenient if H is positive definite. The above approximation ensures this provided only that the model contains no AR parameters. If AR parameters are present an adequate remedy is to start the summation of the 'a' terms from $t = (1 - q + p)$ and to drop the summation in the 'b' terms whenever the derivative is wrt to at least one AR parameter (but not otherwise).

Constraints on the magnitude of the parameter correction $\delta\beta$ may be obtained by adding a diagonal matrix Λ to H_{22}, and adjusting Λ using strategies similar to those of Marquardt (1963) to ensure a decrease in the deviance at successive iterates. The parameters may be kept within their constraints by this device. As the iterations converge the matrix H will also converge, and it may not be worthwhile re-evaluating it. From the Cholesky factorization of H the matrix $L = Q\beta\beta^{-1}$ may then be evaluated and used unaltered to obtain $\delta\beta = -LG(\beta)$ in following iterations. A further possibility from this point is to update L using a variable metric scheme which incorporates information from the values of $G(\beta)$ at successive iterations. A library optimization routine might be used for this purpose. This may be particularly useful when the minimum deviance estimates $\hat{\beta}$ lie close to the boundary of the parameter space. Experience shows that the deviance can be remarkably flat in this region for some MA models, and iterates can be slow to converge if high precision estimates are required.

It should be noted that all the terms $a_t, a_{t,i}$ etc. used to construct the approximation to H have already been calculated to construct the deviance and its first derivative. Indeed, $H_{11} = 2A$ as used to obtain $\hat{w}_L$. Approximately $n(4q + 2p)$ further mults are required to complete H. To solve the equations for $\delta\beta$ will require

approximately $(1/6)(s + p + q)^3$ mults. The computational burden is therefore once more comparable with the calculation of the deviance itself.

Upon convergence, the matrix L may be used to obtain an approximation to the dispersion matrix C of the parameter estimates, as $C = 2L\hat{\sigma}^2$, where $\hat{\sigma}^2 = Q/(n - 1 - p - q)$. The factor 2 occurs explicitly in all the derivative formula and because it cancels, can in practice be omitted throughout, as can the factor $M^{1/n}$.

5. SOME FURTHER CONSIDERATIONS

Multiplicative Operators. The whole of the foregoing algorithm is easily modified to allow for multiplicative operators such as $(1 - \theta B)(1 - \Theta B^{12})$ which appear in seasonal models. The main point here is that there is *no* necessity to multiply out the operators in order to apply them, e.g. to derive $a_t = \{(1 - \theta B)(1 - \Theta B^{12})\}^{-1} w_t$ it is sufficient to use

$$\alpha_t = w_t + \Theta \alpha_{t-12}, \qquad a_t = \alpha_t + \theta a_{t-1} .$$

Furthermore, the derivatives of a_t wrt just one of these parameters is again simple to obtain, e.g. $\partial a_t/\partial\theta = -1/\theta(B)a_t$ as before. The use of multiplicative operators is also useful in non-seasonal contexts, particularly where quasi-cyclic patterns in data are represented by AR operators of order two with complex roots. Where more than one such cycle is present, a second order operator can be associated with each, as an explicit factor in the full AR operator. Suppose we have two such factors, so that *e.g.* $a_t = \phi_1(B)\phi_2(B)/\theta(B)w_t$, then we use simply the series $\partial a_t/\partial\phi_1 = -\phi_2(B)/\theta(B)w_t$ for the calculation of derivatives. In order to obtain det F in this case, the coefficients $f = (f_0, f_1, \ldots, f_{p_1+p_2})$ can be generated by applying $f = \phi_1(B)\phi_2(B)I$ where $I = (1,0,\ldots)$. The derivatives of log det F wrt the coefficients $\phi_{1,1} \cdots \phi_{1,p_1}$ of $\phi_1(B)$ are $h_1 = (h_{1,1}, \ldots, h_{1,p_1}) = \phi_2(B^{-1})h$, where $h = (h_1, h_2, \ldots, h_{p_1+p_2})$ are as calculated in (4.9), with $p' = p_1 + p_2 - 1$ and ρ_k derived from f.

Missing data. Methods based on the Kalman filter are well suited to handling missing data. Provided there is not a very large number of missing values, the algorithm presented here can be used with reasonable efficiency, and certainly with convenience. A missing value at time τ, say w_τ is appended to the set w_L and treated in exactly the same way. The derivatives are extracted from the same v sequence, and used in the minimization of S wrt the extended set, to yield Q and det A. Note that A is also extended in size by these missing values, and this treatment supplies both the estimates of the missing values and the exact likelihood of the set of observed data points. An

alternative is to treat the missing values as nuisance parameters which are estimated by least squares. This would correspond to the intervention analysis approach of estimating missing or corrupted data. The procedure would be exactly the same except for the definition of the likelihood in which $\det A_{11}$ would be used in place of $\det A$, where A_{11} is the submatrix of A corresponding to the backforecasts w_L above.

If there is a large consecutive set of missing values then it is again possible to make economies in the accumulation of the elements of A, but this is not worth while for irregularly scattered missing data.

Box-Cox transformation of data. Following Box and Cox (1964) define

$$w = \begin{cases} (x^{\lambda} - 1)/\lambda - c & \text{for } \lambda \neq 0 \\ \log x - c & \text{for } \lambda = 0 . \end{cases}$$

In this case the likelihood is multiplied by $\Pi x_t^{\lambda-1}$ and the deviance is redefined as $D = M^{1/n}QG^{-2(\lambda-1)}$ where $G = \{\Pi x_t\}^{1/n}$ is the geometric mean of the data. It was pointed out by Box and Cox (1964) that if the data is first scaled by G, i.e. x_t/G is used in place of x_t, then the geometric mean of the scaled data is one, and the Jacobian disappears from the likelihood. In our case the deviance would only then depend on λ via the factor Q. We do not use this device, but follow its implications. Proceeding directly, the factor $G^{-2(\lambda-1)}$ is formally retained in the definition of D and all its derivatives, including those wrt λ, provided we use $\partial Q/\partial\lambda = 2(\Sigma\, a_t a\lambda_t - \Sigma\, b_t b\lambda_t)$ where e.g. $a\lambda_t = G^{\lambda}\partial(a_t/G^{\lambda})/\partial\lambda = \partial a_t/\partial\lambda - a_t \log G$. We derive $\partial a_t/\partial\lambda$ as $\pi(B)w\lambda_t$ where $w\lambda = \partial w/\partial\lambda$, and using $L = \log x$, $r = \lambda L$, $e = \exp(r) = x^{\lambda}$ and $w = (e - 1)/\lambda$, we have

$$w\lambda = (Le - w)/\lambda \quad \text{for } \lambda \neq 0, \quad \text{and } \tfrac{1}{2}L^2 \quad \text{for } \lambda = 0.$$

A similar formula is used in constructing the approximation to the Hessian, so that an element $H\beta\lambda$, where β is any other parameter is formed as $2(\Sigma\, a\beta_t a\lambda_t - \Sigma\, b\beta_t b\lambda_t)$. However, when the element $H\lambda\lambda$ is formed the corresponding expression $\Sigma\, a\lambda_t^2 - \Sigma\, b\lambda_t^2$ must be augmented by a term

$$vL = \Sigma\, a_t a\lambda\lambda_t - \Sigma\, b_t b\lambda\lambda_t$$

where e.g. $a\lambda\lambda_t = \partial^2 a_t/\partial\lambda^2 - a_t(\log G)^2$. This supplies the exact second derivative at the minimum of Q wrt λ.

The point here is that vL does not become negligible for large n. Indeed Brubacher (1976) shows that provided w_t is Gaussian, $vL/n \to \mathrm{Var}(\log x_t) > 0$ at the true value of λ. It is therefore recommended that whenever vL is positive it is incorporated in $H\lambda\lambda$ throughout the iterations. Besides possibly slowing the convergence of the iterates, its absence will lead to an over estimate of $\mathrm{var}(\hat{\lambda})$ from the inverse Hessian.

Again we derive $\partial^2 a_t/\partial\lambda^2$ as $\pi(B)w\lambda\lambda$ where now $w\lambda\lambda = \partial^2 w_t/\partial\lambda^2 = (L^2 e - 2w\lambda)/\lambda$ for $\lambda \neq 0$ and $1/3L^3$ for $\lambda = 0$. It is important to avoid numerical instability in the region close to $\lambda = 0$, and it is recommended that if $|r| = |\lambda \log x|$ is small, the transformation and its derivatives should be evaluated using the first two or three terms in the series

$$w = L(1 + 1/2r + 1/6r^2 + \cdots)$$

$$w\lambda = L^2(1/2 + 1/3r + 1/8r^2 + \cdots)$$

$$w\lambda\lambda = L^3(1/3 + 1/4r + 1/10r^2 + \cdots) .$$

It will be noted that the factor $G^{-2(\lambda-1)}$ appears in both the first derivatives of D and in the approximations to the second derivatives, so that it may be omitted from both of these (in the same way as $M^{1/n}$ was omitted) when solving the equations for the parameter corrections.

6. IMPLEMENTATION

The foregoing algorithm has been implemented and the numerical values obtained for the likelihood in many examples have been found to agree to machine accuracy with the values obtained from the algorithm of Mélard (1983), which have themselves been checked with values from the algorithms of Gardner et al. (1980) and Ansley (1979). No comparison of actual computation times has been made, partly because not all the possible economies in computation have yet been implemented, but mostly because the algorithms have been embedded in a library routine and statistical package, both designed to derive the maximum likelihood (minimum deviance) estimates. From the point of view of the user it is the numerical accuracy and efficiency in locating these estimates which is of prime importance. Practical experience suggests that for nonseasonal models and for seasonal models with period 12, such as used for the airline data in Box and Jenkins, the computational requirements are most reasonable, being of the order 5 seconds on a modern machine, for the 10 to 15 iterations required.

7. REFERENCES

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. Biometrika 66, 59-65.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. J. R. Statist. Soc. B 26, 211-243.

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis, Forecasting and Control, 2nd edition, San Francisco, Holden-Day.

Brubacher, S. R. (1976). Time series modelling with instantaneous nonlinear transformations. Ph.D. Thesis, University of Lancaster, U.K.

Dickinson, B. W. (1978). Two recursive estimates of autoregressive models based on maximum likelihood. J. Statist. Comput. Simul. 7, 85-92.

Duffin, E. J. (1969). Algorithms for classical stability problems. S.I.A.M. Rev. 11, 196-213.

Gardner, G., Harvey, A. C. and Phillips, G. D. A. (1980). Algorithm AS154, An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. Appl. Statist. 29, 311-322.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. J. Amer. Statist. Assoc. 74, 652-660.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. Biometrika 66, 265-70.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. Biometrika 64, 531-34.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. Jour. Soc. Ind. Appl. Math. 11, 431-441.

Mélard, G. (1983). A fast algorithm for the exact likelihood of autoregressive-moving average models. Appl. Statistics - to appear.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. Biometrika 61, 423-6.

Tunnicliffe-Wilson, G. (1979). Some efficient computational procedures for high order ARMA models. J. Statist. Comput. Simul. 8, 301-309.

GTW:scr

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 2528	2. GOVT ACCESSION NO.: AD-A130500	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): THE ESTIMATION OF TIME SERIES MODELS PART I. YET ANOTHER ALGORITHM FOR THE EXACT LIKELIHOOD OF ARMA MODELS	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): G. Tunnicliffe-Wilson	8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709	12. REPORT DATE: June 1983
	13. NUMBER OF PAGES: 20
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office):	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Autoregressive-moving average model
Time series estimation
Exact likelihood

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper presents a method for calculating the likelihood function of autoregressive-moving average (ARMA) models for time series data. Model estimation requires maximization of the likelihood, and to assist in this, a method for calculating derivatives of the function is also presented. The computational efficiency is competitive with that of other algorithms for this purpose. Extensions which allow for seasonal models, missing data, and the estimation of a data transformation are also described.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DATE
ILME