Volume Title: Annals of Economic and Social Measurement, Volume 5, number 1

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm76-1

Publication Date: 1976

Chapter Title: Maximum Likelihood Estimation of Moving Average Processes

Chapter Author: Denise R. Osborn

Chapter URL: http://www.nber.org/chapters/c12707

Chapter pages in book: (p. 75 - 87)

Annals of Economic and Social Measurement, 5/1, 1976

MAXIMUM LIKELIHOOD ESTIMATION OF MOVING AVERAGE PROCESSES

BY DENISE R. OSBORN[1]

This article demonstrates the computational practicality of the maximum likelihood estimation of moving average processes by non-linear optimization. Because of rounding errors and identification problems, the roots of the process are restricted. Aspects of inference and some examples are also discussed.

1. INTRODUCTION

A q-th order pure moving average process is defined by

$$w_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{1.1}$$

where the $\{\varepsilon_t\}$ are assumed to be NID$(0, \sigma^2)$. The process (1.1) is written more compactly in terms of the lag operator, L, as

$$w_t = \theta(L)\varepsilon_t \tag{1.2}$$

where $Lx_t = x_{t-1}$, and $\theta(L)$ is defined as the polynomial $(1 + \theta_1 L + \ldots + \theta_q L^q)$.

If the roots of

$$\theta(L) = 0 \tag{1.3}$$

lie outside the unit circle, then (1.1) is said to be invertible (Box and Jenkins, 1971). An equivalent invertibility condition is that the representation

$$\theta(L) = \prod_{i=1}^{q} (1 - \alpha_i L) \tag{1.4}$$

has all $|\alpha_i| < 1$. It is well-known that there may be as many as 2^q moving average processes, obtained by inverting the roots of a given MA(q) process subject to the requirement that the coefficients be real, which have identical autocorrelation properties. As a result, it is customary to impose the invertibility condition in order to identify the parameters.

To date, the estimation of the vector of parameters $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_q)$ by a full maximum likelihood procedure has generally been regarded as computationally impractical, and a number of alternative approaches have been suggested. Early contributors, such as Durbin (1959) and Walker (1961), were forced to look to methods other than maximum likelihood because the maximization of the likelihood function could not be carried out by linear methods. With the development of non-linear optimization routines, a class of non-linear least squares estimators has come into common use: these have sometimes been claimed as maximum likelihood, but they are in fact only asymptotically so. The least squares estimators are usually derived by assuming that the pre-sample period disturbances $\boldsymbol{\varepsilon}^{*\prime}$

[1] I am greatly indebted to Dr. K. F. Wallis for his advice and to Professor J. D. Sargan who suggested the constrained estimation procedure to me. I would also like to thank members of the Econometric Methodology Workshop Group at LSE for helpful discussion, and the referee for his comments on an earlier draft.

$= (\varepsilon_{1-q}, \ldots \varepsilon_0)$ are fixed numbers; then the logarithm of the likelihood function is, for a sample of n observations,

$$L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}S(\boldsymbol{\theta}) \tag{1.5}$$

where $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}' = (\varepsilon_{1-q}, \ldots, \varepsilon_n)$. Differentiating (1.5) with respect to σ^2 and equating this to zero, we obtain the maximum likelihood estimator of σ^2 as

$$\hat{\sigma}^2 = \frac{S(\boldsymbol{\theta})}{n} \tag{1.6}$$

Substituting for σ^2 in (1.5), the concentrated log likelihood function is

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\frac{S(\boldsymbol{\theta})}{n} \tag{1.7}$$

so that minimizing the residual sum of squares, $S(\boldsymbol{\theta})$, is equivalent to maximizing (1.5).

The estimators within this least squares class differ in their treatment of the "starting residuals", $\mathbf{e}^{*\prime} = (e_{1-q}, \ldots, e_0)$: Åström and Bohlin (1966) set $\mathbf{e}^* = \mathbf{0}$; the Phillips method (detailed by Trivedi, 1970) is to estimate the $\mathbf{e}^*$ as nuisance parameters; Box and Jenkins (1971) suggest either equating them to their expected value of zero (the Åström and Bohlin method) or "back forecasting" them. Once the $\mathbf{e}^*$ are given, all other residuals can be computed recursively from (1.1) for given parameter values. The applications of these methods to time series and econometric estimation are numerous. Although the effect of these different treatments is not clear, Nelson (1974) compares the two Box–Jenkins estimators for the first-order moving average case, and Hendry and Trivedi (1972) study the small sample properties of the Phillips estimator.

The derivation of (1.5), and hence the least squares estimators, depends on the assumption that the pre-sample period disturbances are constants: if these are recognized as random variables, then the "full" likelihood function is obtained, and its logarithm is given by

$$L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2\sigma^2}(\mathbf{w}'\mathbf{V}^{-1}\mathbf{w}) \tag{1.8}$$

where $\sigma^2\mathbf{V}$ is the variance-covariance matrix of $\mathbf{w}$, and $\mathbf{w}' = (w_1, \ldots, w_n)$. Pesaran (1973) has shown that the maximization of (1.8) is feasible in the first-order moving average case, because $\mathbf{V}$ can be reduced to a diagonal matrix by an orthogonal transformation. However, the method does not generalize to higher order cases. Anderson (1973) has suggested a rather different procedure based on the (exact) likelihood equations for the autocovariances of the process.

Box and Jenkins have clarified the relationship between (1.5) and (1.8) by showing that the latter may be written as

$$L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|\mathbf{X}'\mathbf{X}| - \frac{1}{2\sigma^2}S(\boldsymbol{\theta}) \tag{1.9}$$

where $|\mathbf{X}'\mathbf{X}| = |\mathbf{V}|$ and the $\mathbf{e}^*$ used in computing $S(\boldsymbol{\theta})$ are obtained by least squares (see below). Therefore, the non-linear least squares estimators of $\boldsymbol{\theta}$ omit the term involving $|\mathbf{V}|$, and may introduce further levels of approximation, depending on the treatment of starting residuals. But clearly they are asymptotically maximum likelihood.[2]

Kang (1975) has recently examined the properties of the likelihood and sum of squares surfaces. She has shown that the sum of squares function has the undesirable properties that it is decreasing as the boundary of the invertibility region is crossed (so that the minimum may be at $|\alpha_i| = 1$), and approaches its minimum value of zero as $|\alpha_i| \to \infty$.[3] On the other hand, the likelihood function is stationary on the boundary of the invertibility region, and takes the same value for roots α_i and $1/\alpha_i$. Therefore, the assumption that the process is invertible is necessary before any estimate can be obtained by a least squares procedure, but in maximum likelihood estimation it is simply a condition imposed for identification.

Using small samples (12 and 25 observations) generated by invertible MA(1) processes, Kang went on to compare the relative performance of the maximum likelihood estimator, the least squares estimator with computed starting residuals and the conditional least squares estimator with zero starting residuals. The first two sets of coefficient estimates were obtained by grid searches, while the Marquardt algorithm (recommended by Box and Jenkins) was used for the third. Although the "full" least squares estimator was the poorest of the three, for given θ_1 the conditional least squares estimator had similar mean value to the maximum likelihood estimator and often had lower mean square error. However, near the boundary of the invertibility region the conditional estimator was biased towards zero and, for the smaller sample size, had much larger mean square error than the maximum likelihood estimator.

The purpose of this paper is to show that full maximum likelihood estimation of moving average processes by non-linear optimization is a computationally practical procedure. The likelihood function and its evaluation are considered in more detail in the next section. However, because of rounding errors and the question of identification, it is found necessary to restrict the roots of the process (Section 3), and a method of ensuring that the estimated process lies within the invertibility region is suggested in Section 4. Finally, we look at some aspects of inference (Section 5) and some examples (Section 6). The estimation procedure may be readily extended to the case of a regression equation with moving average errors.

2. The Likelihood Function

Box and Jenkins (Appendix A7.4) derive the log likelihood function, equation (1.9), where the matrix $\mathbf{X}$ is defined, and the starting residuals computed, as

[2] The frequency domain estimators of Hannan (1970) are also asymptotically maximum likelihood, but we restrict our interest to time domain methods.

[3] Box and Jenkins (Appendix A7.6, contained only in later printings) also note that outside the invertibility region the sum of squares has no meaningful minimum.

follows. Writing (1.1) as $(n+q)$ equations in ε_t,

$$
\begin{aligned}
\varepsilon_{1-q} &= \varepsilon_{1-q} \\
&\vdots \quad \vdots \\
\varepsilon_0 &= \varepsilon_0 \\
\varepsilon_1 &= w_1 - \theta_1\varepsilon_0 - \theta_2\varepsilon_{-1} - \ldots - \theta_q\varepsilon_{1-q} \\
\varepsilon_2 &= w_2 - \theta_1\varepsilon_1 - \theta_2\varepsilon_0 - \ldots - \theta_q\varepsilon_{2-q} \\
&\vdots \quad \vdots \\
\varepsilon_n &= w_n - \theta_1\varepsilon_{n-1} - \theta_2\varepsilon_{n-2} - \ldots - \theta_q\varepsilon_{n-q}
\end{aligned}
\tag{2.1}
$$

then substituting for $\varepsilon_1, \ldots, \varepsilon_{n-1}$ in terms of $w_1, \ldots, w_{n-1}$ and $\varepsilon_{1-q}, \ldots, \varepsilon_0$, these equations may be written such that the right-hand side does not contain the sample period disturbances, $\varepsilon_1, \ldots, \varepsilon_n$. The system of equations is then

$$\boldsymbol{\varepsilon} = \mathbf{M}\mathbf{w} + \mathbf{X}\boldsymbol{\varepsilon}^* \tag{2.2}$$

where the vectors $\boldsymbol{\varepsilon}$, $\mathbf{w}$ and $\boldsymbol{\varepsilon}^*$ have been defined previously. The coefficient matrices $\mathbf{M}$ and $\mathbf{X}$ are of dimensions $(n+q)\times n$ and $(n+q)\times q$ respectively, and their elements are functions of the elements of $\boldsymbol{\theta}$ only. The maximum likelihood estimate, $\mathbf{e}^*$, of $\boldsymbol{\varepsilon}^*$, is obtained from (2.2) by ordinary least squares; that is,

$$\mathbf{e}^* = -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{w} \tag{2.3}$$

Differentiating (1.9) with respect to σ^2, we see that the maximum likelihood estimator of σ^2 is still given by (1.6), and the concentrated log likelihood function is:

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{S(\boldsymbol{\theta})}{n} - \frac{1}{2}\log|\mathbf{X}'\mathbf{X}| - \frac{n}{2} \tag{2.4}$$

The maximization of the full likelihood function is therefore equivalent to the minimization of

$$L^*(\boldsymbol{\theta}) = n\log(S(\boldsymbol{\theta})) + \log|\mathbf{X}'\mathbf{X}| \tag{2.5}$$

Consider once again the matrices $\mathbf{M}$ and $\mathbf{X}$. The (i, j)-th element of $\mathbf{M}$ is given by

$$
\begin{aligned}
m_{ij} &= 0 && i = 1, \ldots, q, && j = 1, \ldots, n \\
m_{ij} &= 0 && j > i - q, && i = q+1, \ldots, q+n \\
m_{ij} &= 1 && j = i - q, && i = q+1, \ldots, q+n \\
m_{ij} &= -\theta_1 m_{i-1,j} - \theta_2 m_{i-2,j} - \ldots - \theta_q m_{i-q,j} && j < i - q, && i = q+1, \ldots, q+n
\end{aligned}
\tag{2.6}
$$

and the (i, j)-th element of $\mathbf{X}$ is

$$
\begin{aligned}
&x_{ii} = 1 && i = 1, \ldots, q \\
&x_{ij} = 0 && i \neq j, \qquad i, j = 1, \ldots, q \\
&x_{ij} = -\theta_1 x_{i-1,j} - \theta_2 x_{i-2,j} - \ldots - \theta_q x_{i-q,j} && i = q+1, \ldots, q+n, \quad j = 1, \ldots, q
\end{aligned}
\tag{2.7}
$$

Exploiting the recursive relationship in each column of $\mathbf{M}$ and $\mathbf{X}$, these matrices can be easily formed.

The dimensions of $\mathbf{M}$, $(n+q) \times n$, are large, but there is no need to store this matrix. Use of the relationship

$$
\begin{aligned}
&m_{ij} = m_{i-1,j-1} && i = 2, \ldots, n+q \qquad j = 2, \ldots, n \\
&m_{1j} = 0 && j = 2, \ldots, n
\end{aligned}
\tag{2.8}
$$

renders unnecessary the computation and storage of any columns beyond the first, and as

$$
\begin{aligned}
&m_{i1} = x_{i-1,q} && i = 2, \ldots, n+q \\
&m_{11} = 0
\end{aligned}
\tag{2.9}
$$

even this is not required. Combining (2.8) and (2.9), the non-zero elements of the j-th column of $\mathbf{M}$ may be obtained from the last column of $\mathbf{X}$:

$$
m_{ij} = x_{i-j,q} \qquad j \leq i - q, \qquad i = q+1, \ldots, q+n \tag{2.10}
$$

Therefore, (2.5) may be evaluated as follows: $\mathbf{X}$ is formed, and $\mathbf{e}^*$ obtained from (2.3) using (2.10). The remaining residuals, $e_1, \ldots, e_n$ are then computed recursively using (2.1). Finally, the sum of squares

$$
S(\boldsymbol{\theta}) = \sum_{t=1-q}^{n} e_t^2 \tag{2.11}
$$

and $L^*(\boldsymbol{\theta})$ are evaluated. The required minimum may be achieved via a general numerical optimization procedure, and we have chosen the Powell conjugate direction method (see Powell, 1964). The Powell method has been found to perform satisfactorily in a number of non-linear optimization problems, including those involving transformations (Box, 1966) and regression estimation in the presence of autoregressive errors (Hendry, 1971).

3. Difficulties with Direct Evaluation

In practice, the computation of the maximum likelihood estimates by minimizing (2.5) with respect to $\boldsymbol{\theta}$ cannot be treated as an unconstrained optimization problem. For when the elements of $\boldsymbol{\theta}$ are unrestricted, there may be 2^q points at which the function takes its minimum value: that is, there is a problem of identification. Also, rounding errors become important when some roots of the moving average process lie inside the unit circle and some outside it. The origin of

these rounding error problems can be seen by considering a second-order moving average. Then

$$\theta_1 = -(\alpha_1 + \alpha_2) \tag{3.1}$$

and

$$\theta_2 = \alpha_1 \alpha_2 \tag{3.2}$$

where α_1 and α_2 are defined by (1.4). The matrix $\mathbf{X}$ may be expressed in terms of (real) α_1 and α_2 as:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -\alpha_1\alpha_2 & \alpha_1+\alpha_2 \\ -\alpha_1\alpha_2(\alpha_1+\alpha_2) & \alpha_1+\alpha_1\alpha_2+\alpha_2^2 \\ \vdots & \vdots \\ -\alpha_1\alpha_2(\alpha_1^{n-1}+\alpha_1^{n-2}\alpha_2+\ldots & \alpha_1^n+\alpha_1^{n-1}\alpha_2+\ldots \\ \quad +\alpha_1\alpha_2^{n-2}+\alpha_2^{n-1}) & \quad +\alpha_1\alpha_2^{n-1}+\alpha_2^n \end{bmatrix} \tag{3.3}$$

If, say, $|\alpha_1| \geq 1$ and $|\alpha_2| < 1$, then $|\alpha_1^n| \to \infty$ and $|\alpha_2^n| \to 0$ as $n \to \infty$. For any sample size contemplated in empirical work, the elements of $\mathbf{X}$ involve powers, products, and sums of numbers of very different orders of magnitude, so that the estimates obtained under these circumstances will be doubtful. In particular, $(\mathbf{X}'\mathbf{X})$ will be dominated by high order powers of α_1, and the inversion may break down. It is clear that both the identification and rounding error problems are avoided if the roots of the moving average process are restricted so that the estimated process lies within the invertibility region.

It may appear unlikely that these problems will arise if all moving average coefficients are given initial values of zero in the estimation procedure. However, this has been found not to be the case: the estimation often "missed" the function minimum contained within the invertibility region when one or more roots lay even moderately close to the boundary. Box and Jenkins suggest estimating the initial values of the moving average coefficients from the sample autocorrelations of the process, but this generally involves iterative techniques (Wilson, 1969), and in any case does not guarantee that the roots will remain within the invertibility region.

To overcome these problems it may be feasible to test the roots at the end of each iteration, and to re-set any root which has moved outside the invertibility region to a value inside it. This is the type of procedure adopted by Nicholls (1972). An alternative is to set up the estimation as a problem in constrained optimization, as in the next section.

4. Ensuring Invertibility

Consider the α_i of (1.4): these may be divided into pairs, which are either complex conjugates or both are real, and each pair can be regarded as the roots of

a quadratic

$$\lambda^2 + a_i\lambda + b_i = 0 \tag{4.1}$$

Now, both these roots lie within the unit circle if and only if

$$|b_i| < 1 \quad \text{and} \quad |a_i| < 1 + b_i \tag{4.2}$$

and these conditions will be fulfilled if we define a_i and b_i by

$$b_i = \frac{\delta_i}{1 + |\delta_i|} \tag{4.3}$$

$$a_i = \frac{\gamma_i(1 + b_i)}{1 + |\gamma_i|} \tag{4.4}$$

where the parameters δ_i and γ_i are unrestricted.[4] That is, with b_i and a_i defined by (4.3) and (4.4), the quadratic may be solved to obtain a pair of roots satisfying the invertibility condition. If the order of the moving average is an odd number, then the remaining root after taking pairs is necessarily real and can be specified in terms of one parameter. In this case it is convenient to define the root directly as

$$d = -\frac{\pi}{1 + |\pi|} \tag{4.5}$$

The negative of π is used because if $\delta_i = 0$, then the non-zero root of the quadratic (4.1) is

$$-a_i = -\frac{\gamma_i}{1 + |\gamma_i|} \tag{4.6}$$

so that d may be regarded as the non-zero root of (4.1) when $b_i = 0$.

The q parameters used to define the roots of the MA(q) process uniquely give the moving average coefficients. By adopting this parameterization, we estimate the roots α_i (constrained to lie within the unit circle), and obtain the moving average coefficients from these via (1.4). Of course, once the corresponding coefficients have been computed, the function (2.5) can be evaluated as in Section 2.

If a moving average of order greater than 1 is to be estimated, the procedure may be commenced at the order required with zero initial values for all parameters. Alternatively, a step-wise procedure may be adopted, beginning with MA(1), and increasing the order by 1 at each step. Ordering the unrestricted parameters within each pair as (γ_i, δ_i), with π as the last parameter if q is odd, then the estimates at the previous step are used as initial values for the first q parameters when the order is increased from q to $(q+1)$. With zero as the initial value for the additional parameter, the first q elements of $\boldsymbol{\theta}$ are the coefficient estimates of the last step, and the $(q+1)$-th coefficient is initially zero. Note that if q is odd, then the estimated value of π becomes the initial value of the last γ_i for the $(q+1)$ process. In terms of computer time, there appears to be little additional cost in carrying out this stepwise estimation.

[4] a_i and b_i could have been defined in terms of other functions of δ_i and γ_i, such as trigonometric functions, but (4.3) and (4.4) require fewer operations for their evaluation.

While a root of the moving average process on the unit circle causes no identification problem, it is not permitted by the invertibility condition and cannot be obtained with this estimation procedure. However, roots can come arbitrarily close to the unit circle if parameters γ_i, δ_i and π are unrestricted. It is usual to test the convergence of the Powell routine in terms of the parameters of the function, in this case the γ_i, δ_i, and π. But as a root approaches the unit circle, large changes in the corresponding parameter have little effect on the moving average coefficients, and hence on the value of the likelihood function, and the estimation may fail to converge. Therefore, it is advisable either to specify some arbitrary maximum absolute value that the parameters may take, or to test convergence in terms of the roots of the process.

5. Inference

The asymptotic properties of the least squares estimates of the parameters for an invertible moving average process with finite fourth moment have been derived by Whittle (1954, 1961) and Walker (1964). These properties are shared by our exact maximum likelihood estimates for this case. In particular, Whittle and Walker show that the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is, as $n\to\infty$, $N(\mathbf{O}, \mathbf{U}^{-1})$ where

$$\mathbf{U}=\lim_{n\to\infty} E\left[-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \tag{5.1}$$

Further, $n^{1/2}(\hat{\sigma}^2-\sigma^2)$ has, as $n\to\infty$, a limiting normal distribution independent of that for $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$.

As a consequence of (5.1), a sample estimate of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$V(\hat{\boldsymbol{\theta}})=\left[-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{5.2}$$

$$=2\left[\frac{\partial^2 L^*(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{5.3}$$

which can be computed numerically, the elements of $\hat{\boldsymbol{\theta}}$ being used directly in the evaluation of L^*.

Neither Whittle nor Walker is concerned with the influence of the starting residuals. However, Pierce (1971) considers the effect of a change in the set of starting residuals by $\Delta\mathbf{e}^*$. For given coefficients, $\dot{\theta}_1, \dot{\theta}_2, \ldots, \dot{\theta}_q$, he shows that

$$\Delta\epsilon_t=\sum_{i=1}^{q} c_i\dot{\alpha}_i^t \qquad t>0 \tag{5.4}$$

where the c_i depend on $\Delta\mathbf{e}^*$ and, for $\dot{\theta}(L)$, the $\dot{\alpha}_i$ are defined by (1.4). Pierce uses (5.4) to show that the asymptotic properties of the least squares estimates are independent of the starting residuals if the process is invertible. However, it is clear that the effect of a change in $\mathbf{e}^*$ on the residual series is not asymptotically negligible for a process outside the invertibility region: in these circumstances,

therefore, we would not expect the starting residuals and coefficient estimates (obtained by maximum likelihood) to be asymptotically independent.

Since the likelihood function $L(\boldsymbol{\theta})$ has been concentrated with respect to $\boldsymbol{\varepsilon}^*$, its use in (5.2) or (5.3) is valid only when the starting values and coefficient estimates are asymptotically independent. Therefore, if these formulae are to be used for inference, the estimated process should be restricted to the invertibility region.

A significance test not affected by the imposition of the invertibility condition is the likelihood ratio test, which may be used to test the significance of an increase (decrease) in the likelihood when the order of the moving average is increased (decreased). If L_q^* is the minimum value of (2.5) for q-th order process, and L_{q-r}^* is the minimum for a process of order $(q-r)$, then $(L_{q-r}^*-L_q^*)$ is asymptotically distributed as a χ^2 variable with r degrees of freedom under the null hypothesis that the lower order model is the correct one (Kendall and Stuart, 1967, pp. 230–231). Incidentally, the use of the step-wise procedure for estimating an MA(q) process, as outlined above, makes the application of this test particularly simple.

6. Examples

Box and Jenkins specify and estimate autoregressive-moving average models for six series, with between 70 and 369 observations. The two shortest series for which they "identify" pure moving average models are A and C; these provide convenient "test" cases and in Table 1 we compare the estimates given by Box and Jenkins with those obtained by maximum likelihood. The estimated standard errors, obtained from (5.3) for the maximum likelihood values, are given in brackets; $\hat{\sigma}^2$ is the maximum likelihood estimate of σ^2. The operator ∇ is the first difference operator.

TABLE 1
Estimation of Models for Series A and C

Series	n	Method	Fitted Model	$\hat{\sigma}^2$
A	197	BJ	$\nabla w_t = e_t - \underset{(0.05)}{0.70} e_{t-1}$	0·101
		ML	$\nabla w_t = e_t - \underset{(0.06)}{0.70} e_{t-1}$	0.101
C	226	BJ	$\nabla^2 w_t = e_t - \underset{(0.07)}{0.13} e_{t-1} - \underset{(0.07)}{0.12} e_{t-2}$	0.019
		ML	$\nabla^2 w_t = e_t - \underset{(0.07)}{0.13} e_{t-1} - \underset{(0.08)}{0.12} e_{t-2}$	0.019

For each series, the two sets of estimates are virtually the same. This is not surprising, as series containing about 200 observations may be expected to follow large sample theory. Nevertheless, the Box–Jenkins series are useful as a reference point because the data are "real"; in order to obtain shorter series for purposes of comparison, we sub-divided Series C. Of course, if the total series can

be represented by an ARIMA(0, 2, 2) model with constant parameters, each sub-series can be represented by this same model. We emphasize that the maximum likelihood estimation of this model for the complete series presented no problems and yielded coefficients well inside the invertibility region (with $\hat{\alpha}_1 = 0.41$ and $\hat{\alpha}_2 = -0.29$). We would expect the sub-series to be similarly well-behaved.

The first 224 observations of Series C were used to provide eight consecutive sub-series of 28 observations; after second differences were taken within each sub-series, 26 observations were available for estimation. The models of Table 2 were obtained by least squares (setting the starting residuals to zero) and maximum likelihood as indicated, with all estimations constrained to the invertibility region (see Section 4). The roots given in this table are the estimated α's; for a complex pair of roots the absolute value $|\hat{\alpha}|$ is also shown. For comparison of performance and computational cost, an unconstrained estimation was also carried out in each case.

With the exception of the maximum likelihood estimations for sub-series 7, the Powell minimization procedure was commenced from initial parameter values of zero in all estimations. For sub-series 7, this initialization led to convergence of the constrained and unconstrained maximum likelihood procedures at (different) local minimum points; in fact, the first parameter of the constrained procedure did not move away from zero. This problem of local minima may possibly have been avoided had a tighter convergence criterion been used. However, our maximum likelihood estimates for this sub-series were obtained by commencing the constrained estimation at the parameter values given by the constrained least squares procedure. The unconstrained procedure was also re-commenced at the least squares estimates. There is no reason obvious to us why estimation using the full likelihood function should be more troublesome than that using the sum of squares only, and we attribute the difficulties in this case to the sub-series not being well-behaved.

The unconstrained least squares procedure for sub-series 2 converged to a point outside the invertibility region, with a root equal to -1.03. Except for this case (and the initial maximum likelihood estimation of sub-series 7), the constrained and unconstrained procedures for a given sub-series and given (ML or LS) estimation method both converged to the same point. However, the fact that unconstrained least squares yielded invalid estimates in one case out of eight should not be overlooked.

There are two types of comparisons which may be made on the basis of Table 2: one regarding the computational cost of constraining an estimation to the invertibility region, and the other regarding the relative performance and cost of the maximum likelihood and least squares procedures. With respect to the first of these, the number of function evaluations required for convergence does not fully reflect the relative cost of a constrained and unconstrained estimation, because the former involves more operations per function evaluation than the latter. However, the number of additional operations required is relatively small and depends only on the order of the moving average.

From Table 2 it may be seen that constraining an estimation to the invertibility region (as outlined in Section 4) has resulted in many more function evalua-

TABLE 2
ESTIMATED ARIMA (0, 2, 2) MODELS FOR SUB-SERIES OF C

	Method	Function Evaluations		Constrained Parameter Estimates				
		Unconstd.	Constd.	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}_2$	Roots	$\lvert\hat{\alpha}\rvert$
1	ML	42	47	−0.18	−0.16	0.016	0.50, −0.31	
				(0.34)	(0.32)			
	LS	31	36	−0.12	−0.14	0.016	0.44, −0.32	
				(0.27)	(0.26)			
2	ML	39	44	0.22	−0.37	0.011	0.51, −0.73	
				(0.20)	(0.23)			
	LS	73	165	0.28	−0.60	0.011	0.65, −0.92	
				(0.06)	(0.06)			
3	ML	42	44	0.67	0.56	0.033	−0.34±0.67i	0.75
				(0.21)	(0.17)			
	LS	32	35	0.64	0.55	0.039	−0.32±0.67i	0.74
				(0.20)	(0.21)			
4	ML	33	31	−0.59	0.09	0.0069	0.29±0.04i	0.30
				(0.22)	(0.22)			
	LS	33	33	−0.61	0.10	0.0069	0.31±0.09i	0.32
				(0.23)	(0.24)			
5	ML	32	38	−0.05	−0.27	0.0094	0.54, −0.50	
				(0.21)	(0.21)			
	LS	36	38	−0.04	−0.27	0.0094	0.54, −0.50	
				(0.20)	(0.20)			
6	ML	28	40	−0.39	0.26	0.017	0.20±0.47i	0.51
				(0.19)	(0.24)			
	LS	31	40	−0.38	0.27	0.017	0.19±0.49i	0.52
				(0.19)	(0.25)			
7	ML	30	105	−0.98	0.99	0.0031	0.49±0.87i	0.99
				(0.16)	(0.25)			
	LS	43	137	−0.92	0.80	0.0039	0.46±0.77i	0.89
				(0.14)	(0.12)			
8	ML	18	25	−0.04	−0.01	0.011	0.10, −0.06	
				(0.21)	(0.29)			
	LS	23	21	−0.04	−0.01	0.011	0.12, −0.08	
				(0.22)	(0.29)			

tions being required only when the estimated model is near the boundary of the invertibility region. But this is precisely the case where the constraint is likely to be of importance, as demonstrated by the least squares estimation for sub-series 2. Because the computational cost is relatively small in other cases, we conclude that constrained estimation may as well be used also for models well inside the invertibility region.

In comparing the maximum likelihood and least squares coefficient estimates, we are again led to consider sub-series 2 and 7. For sub-series 2, both roots estimated by least squares are closer to the unit circle than those estimated by maximum likelihood. The position is reversed for sub-series 7.

The two sets of estimated roots have led to rather different estimates of θ_2 in sub-series 2. But perhaps more important are the implications of the two sets of standard errors. If we had carried out only the constrained least squares estimation, we would have been very confident about the significance of the second moving average coefficient ($t = 10$). However, the standard errors have been

estimated numerically, and the small values in this case appear to have been caused by the sum of squares function not being well-behaved near the boundary of the invertibility region.

Another maximum likelihood/least squares comparison of interest is the relative computational cost. For our sub-series, the two methods have generally required about the same number of function evaluations for convergence, although the estimations for sub-series 7 are not comparable in these terms because they were commenced from different initial parameter values. For sub-series 2, the maximum likelihood estimation is computationally more efficient because it is not close to the boundary of the invertibility region. However, it is of course true that evaluation of the full likelihood function is computationally more expensive than evaluation of the sum of squares only. The additional cost (which depends on the sample size) is largely incurred in the estimation of the starting residuals.

Finally, we note that the estimated models of Table 2 suggest that the moving average coefficients may not be stable over time. It has not, however, been possible to test for stability: although the sub-samples are independent, we have "lost" observations by differencing after sub-dividing the series, and an F test cannot be carried out.

7. Conclusion

We believe that the maximum likelihood estimation of moving average processes is not computationally impractical for small to moderate size samples. For example, on the University of London's CDC 6600 machine, the central processor time required to obtain the four sets of estimates (by constrained and unconstrained least squares and maximum likelihood) for each sub-series in Table 2 was, on average, approximately $1\frac{1}{4}$ seconds.

A thorough investigation of the relative performance of the maximum likelihood and least squares procedures would require a large-scale simulation study, which we have not attempted to carry out here. However, we have found evidence of the superiority of the maximum likelihood procedure, especially in the numerical estimation of standard errors for the coefficients of a model near the boundary of the invertibility region.

The method suggested for constraining the estimated process to the invertibility region may be used in either maximum likelihood or least squares estimation. It requires few additional operations in any evaluation of the function being minimized and, in most cases, few additional function evaluations for convergence.

London School of Economics
and Political Science

References

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structures. *Annals of Statistics*, 1, 135–141.

Åström, K. J. and T. Bohlin (1966). Numerical identification of linear dynamic systems from normal operating records, in P. H. Hammond (ed), *Theory of Self-Adaptive Control Systems*. New York: Plenum Press.

Box, G. E. P. and G. M. Jenkins (1971). *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day, second printing.
Box, M. J. (1966). A comparison of several current optimization methods, and the use of transformations in constrained problems. *Computer Journal*, 9, 67–77.
Durbin, J. (1959). Efficient estimation of parameters in moving-average models. *Biometrika*, 46, 306–316.
Hannan, E. J. (1970). *Multiple Time Series.* New York: John Wiley.
Hendry, D. F. (1971). Maximum likelihood estimation of systems of simultaneous regression equations with errors generated by a vector autoregressive process. *International Economic Review*, 12, 257–272.
Hendry, D. F. and P. K. Trivedi (1972). Maximum likelihood estimation of difference equations with moving average errors: A simulation study. *Review of Economic Studies*, 39, 117–145.
Kang, K. M. (1975). A comparison of estimators for moving average processes. Unpublished paper, Australian Bureau of Statistics.
Kendall, M. G. and A. Stuart (1967). *The Advanced Theory of Statistics*, Vol. 2. London: Charles Griffin, second edition.
Nelson, C. R. (1974). The first-order moving average process. *Journal of Econometrics*, 2, 121–141.
Nicholls, D. F. (1972). On Hannan's estimation of ARMA models. *Australian Journal of Statistics*, 14, 262–269.
Pesaran, M. H. (1973). Exact maximum likelihood estimation of a regression equation with first-order moving-average error. *Review of Economic Studies*, 40, 529–535.
Pierce, D. A. (1971). Least squares estimation in the regression model with autoregressive-moving average errors. *Biometrika*, 58, 299–312.
Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, 7, 155–162.
Trivedi, P. K. (1970). Inventory behaviour in U.K. manufacturing, 1956–67. *Review of Economic Studies*, 37, 517–536.
Walker, A. M. (1961). Large sample estimation of parameters for moving-average models. *Biometrika*, 48, 343–357.
Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *Journal of the Australian Mathematical Society*, 4, 363–384.
Whittle, P. (1954). Some recent contributions to the theory of stationary processes, Appendix 2 to H. Wold, *A Study in the Analysis of Stationary Time Series.* Stockholm: Almqvist and Wiksell.
Whittle, P. (1961). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, 39(2), 105–129.
Wilson, G. (1969). Factorization of the covariance generating function of a pure moving average process. *SIAM Journal of Numerical Analysis*, 6, 1–7.

received December 1974
revised July 1975