*The Annals of Statistics*
1977, Vol. 5, No. 5, 842-865

# ESTIMATION FOR AUTOREGRESSIVE MOVING AVERAGE MODELS IN THE TIME AND FREQUENCY DOMAINS[1]

BY T. W. ANDERSON

*Stanford University and London School of Economics*

The autoregressive moving average model is a stationary stochastic process $\{y_t\}$ satisfying $\sum_{k=0}^{p} \beta_k y_{t-k} = \sum_{g=0}^{q} \alpha_g v_{t-g}$, where the (unobservable) process $\{v_t\}$ consists of independently identically distributed random variables. The coefficients in this equation and the variance of $v_t$ are to be estimated from an observed sequence $y_1, \cdots, y_T$. To apply the method of maximum likelihood under normality the model is modified (i) by setting $y_0 = y_{-1} = \cdots = y_{1-p} = 0$ and $v_0 = v_{-1} = \cdots = v_{1-q} = 0$ and alternatively (ii) by setting $y_0 \equiv y_T, \cdots, y_{1-p} \equiv y_{T+1-p}$ and $v_0 \equiv v_T, \cdots, v_{1-q} \equiv v_{T+1-q}$; the former lead to procedures in the time domain and the latter to procedures in the frequency domain. Matrix methods are used for a unified development of the Newton-Raphson and scoring iterative procedures; most of the procedures have been obtained previously by different methods. Estimation of the covariances of the moving average part is also treated.

**1. Introduction.** Several methods have been proposed for estimating the parameters of an autoregressive moving average model based on approximating or modifying the normal likelihood function. The main purpose of this paper is to develop these methods and some new procedures by a unified approach.

The autoregressive moving average process $\{y_t\}$ satisfies

$$\sum_{k=0}^{p} \beta_k y_{t-k} = \sum_{g=0}^{q} \alpha_g v_{t-g}\,, \tag{1.1}$$

$t = \cdots, -1, 0, 1, \cdots,$ where the sequence $\{v_t\}$ consists of (unobservable) independently identically distributed random variables. (See Section 5.8 of Anderson (1971a) and Box and Jenkins (1970).) To avoid indeterminancy we require $\beta_0 = \alpha_0 = 1$. Because attention will be concentrated on the coefficients in (1.1) we shall assume the means of $\{y_t\}$ are known and 0; the specification, then, is $\mathscr{E} v_t = 0$, $\mathscr{E} v_t^2 = \sigma^2 > 0$.

Received June 1975; revised March 1977.

[1] Research supported by Office of Naval Research under Contract Number N00014-75-C-0442. Reproduction in whole or in part permitted for any purpose of the United States Government. This paper was started when the author was at the Center for Advanced Study in the Behavioral Sciences. It was presented as a Special Invited Paper to the 1973 Spring Meeting of the Institute of Mathematical Statistics at Albuquerque, New Mexico. The author thanks Paul Shaman for helpful comments.

*AMS* 1970 *subject classifications.* Primary 62M10; Secondary 62H99.

*Key words and phrases.* Maximum likelihood estimation, autoregressive moving average models, Newton-Raphson and scoring iterative procedures, time and frequency domains, time series analysis.

Let

$$A(z) = \textstyle\sum_{g=0}^{q} \alpha_g z^g \,, \qquad \alpha_q \neq 0 \,, \tag{1.2}$$

$$B(z) = \textstyle\sum_{k=0}^{p} \beta_k z^k \,, \qquad \beta_p \neq 0 \,. \tag{1.3}$$

We shall asuume the roots of $A(z) = 0$ and of $B(z) = 0$ are greater than 1 in absolute value and that the two sets have no roots in common. The process $\{y_t\}$ is stationary.

When the $v_t$'s are normally distributed, that is, the process is Gaussian, the model is completely specified by $\beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q$, and $\sigma^2$. We study the estimation of these parameters on the basis of a set of observations at $T$ successive time points, $y_1, \cdots, y_T$ $(T > p + q)$. An alternative set of parameters consists of $\beta_1, \cdots, \beta_p$ and the variance and first $q$ covariances of $u_t = \sum_{g=0}^{q} \alpha_g v_{t-g}$, namely

$$\sigma_h = \sigma^2 \textstyle\sum_{f=0}^{q-h} \alpha_f \alpha_{f+h} \,, \qquad h = 0, 1, \cdots, q \,; \tag{1.4}$$

the estimation of this set of parameters is also treated. For the observed stationary process the spectral density is

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|A(e^{i\lambda})|^2}{|B(e^{i\lambda})|^2} = \frac{\sum_{h=-q}^{q} \sigma_h \cos \lambda h}{2\pi |B(e^{i\lambda})|^2} \,, \tag{1.5}$$

where $\sigma_{-h} = \sigma_h$, $h = 1, \cdots, q$.

Since the set of observable variables $y_1, \cdots, y_T$ has a multivariate normal distribution (Anderson (1958), for example), the method of maximum likelihood is appealing. However, even in the simplest cases—the purely autoregressive model—the equations obtained by setting the derivatives of the likelihood function equal to 0 are nonlinear. Mann and Wald (1943) modified the autoregressive model by considering the first $p$ observations as fixed; then maximum likelihood was least squares, resulting in linear equations. These equations may be further adjusted by adding some end terms to sums of squares and cross-products. The asymptotic properties (as $T \rightarrow \infty$) are not affected by the alterations.

The moving average part of the model introduces considerably greater complications, even in the absence of the autoregressive part. The inverse of the covariance matrix (which is proportional to the matrix of the quadratic form in the exponent of the likelihood function) consists of $T$ polynomials in the coefficients, and the number of sufficient statistics is the number of observations. The derivative equations are highly nonlinear.

To obtain feasible methods in the time domain we modify the model by setting all variables with nonpositive indices equal to zero. When the parameters include $\alpha_1, \cdots, \alpha_q$, we set $v_0 = v_{-1} = \cdots = v_{1-q} = 0$ as well as $y_0 = y_{-1} = \cdots = y_{1-p} = 0$; when the parameters include $\sigma_0, \sigma_1, \cdots, \sigma_q$, the variables $v_0, v_{-1}, \cdots, v_{1-q}$ do not appear.

To obtain corresponding methods in the frequency domain we replace the stationary model by the corresponding circular model; that is, let (1.1) for $t = 1, \cdots, T$ define $y_1, \cdots, y_T$ on the basis of $v_1, \cdots, v_T$ using the notational convention $y_{-k} \equiv y_{T-k}$, $k = 0, 1, \cdots, p-1$, and $v_{-g} \equiv v_{T-g}$, $g = 0, 1, \cdots, q-1$. Then the Fourier unitary transformation of $\mathbf{y} = (y_1, \cdots, y_T)'$ diagonalizes the covariance matrix, and the likelihood depends on $\mathbf{y}$ through the sample spectral density (periodogram). On a large sample basis (that is, large $T$) the modification has a negligible effect. (See, for example, Wahba (1968) and Hannan (1960), Chapter 1.)

These models are then special cases of

$$\sum_{k=0}^{p} \beta_k \mathbf{K}_k \mathbf{y} = \sum_{g=0}^{q} \alpha_g \mathbf{J}_g \mathbf{v} , \tag{1.6}$$

where $\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ are $p+1$ known linearly independent $T \times T$ matrices, $\mathbf{J}_0, \mathbf{J}_1, \cdots, \mathbf{J}_q$ are $q+1$ known linearly independent $T \times T$ matrices, and $\mathbf{v} = (v_1, \cdots, v_T)'$ has a multivariate normal distribution with mean vector $\mathscr{E}\mathbf{v} = \mathbf{0}$ and covariance matrix $\mathscr{C}(\mathbf{v}) = \sigma^2\mathbf{I}$. The parametrization with $\sigma_0, \sigma_1, \cdots, \sigma_q$ is a special case of the model

$$\sum_{k=0}^{p} \beta_k \mathbf{K}_k \mathbf{y} = \mathbf{u} , \tag{1.7}$$

where $\mathbf{u}$ has a multivariate normal distribution with mean vector $\mathscr{E}\mathbf{u} = \mathbf{0}$ and covariance matrix

$$\mathscr{C}(\mathbf{u}) = \mathscr{E}\mathbf{u}\mathbf{u}' = \sum_{g=0}^{q} \sigma_g \mathbf{G}_g , \tag{1.8}$$

and $\mathbf{G}_0, \mathbf{G}_1, \cdots, \mathbf{G}_q$ are $q+1$ known symmetric, linearly independent $T \times T$ matrices such that the linear combination is positive definite. Anderson (1975a) developed the maximum likelihood equations for these general models. Because they are nonlinear in most cases, iterative procedures based on the method of scoring were proposed. In the present paper the Newton–Raphson procedures are also given.

The main purpose of this paper is to give a unified development of the scoring and the Newton–Raphson methods for the maximum likelihood estimation of the alternative sets of parameters using the observations in the time sequence and using the sample spectral density. The Newton–Raphson method for the $\beta_k$'s, $\alpha_g$'s and $\sigma^2$ based on setting some unobserved variables 0 corresponds to the method of Åström and Bohlin (1966); the scoring method was given by Anderson (1975a). The Newton–Raphson method for these parameters based on the circular model is approximatey that proposed by Hannan (1969) and (1970). (The latter fact has been shown by Akaike (1973) in a different manner.) The scoring procedure has been proposed by Dzhaparidze and Yaglom (1974). The Newton–Raphson method for the $\beta_k$'s and $\sigma_h$'s based on setting some unobserved variables 0 is given here for the first time; the scoring method was presented by Anderson (1975a). The Newton–Raphson method for these parameters based on the circular model seems to be new; the scoring method is that proposed by Clevenson (1970) and Parzen (1971).

The modifications of the model are discussed in Section 2. The equations for eight cases are given and discussed in the next two sections. Some mathematical details are considered later. (A fuller mathematical treatment is available in Anderson (1975 b), from which this present paper was condensed.)

**2. Two modifications of the autoregressive moving average process.**

2.1. *Use of a matrix lag operator.* Let the $T \times T$ matrix $\mathbf{L}$ be

$$\mathbf{L} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{2.1}$$

where $\mathbf{I}$ is of order $T-1$. Then $\mathbf{L}^t$ is of the same form, but $\mathbf{I}$ is of order $T-t$, $t = 0, 1, \cdots, T-1$, and $\mathbf{L}^t = \mathbf{0}$ for $t = T, T+1, \cdots$. For any $T$-component vector $\mathbf{x}$ the vector $\mathbf{L}\mathbf{x}$ has 0 as its first component and the $(t-1)$st component of $\mathbf{x}$ as its $t$th component, $t = 2, \cdots, T$; the vector $\mathbf{L}^s\mathbf{x}$ has 0 as its $t$th component, $t = 1, \cdots, s$, and the $(t-s)$th component of $\mathbf{x}$ as its $t$th component, $t = s+1, \cdots, T$. The modified model for $\mathbf{y}$

$$\textstyle\sum_{k=0}^{p} \beta_k \mathbf{L}^k \mathbf{y} = \sum_{g=0}^{q} \alpha_g \mathbf{L}^g \mathbf{v} , \tag{2.2}$$

where $\mathbf{v}$ has the distribution $N(\mathbf{0}, \sigma^2\mathbf{I})$, corresponds to (1.1) with $y_0 = y_{-1} = \cdots = y_{1-p} = 0$ and $v_0 = v_{-1} = \cdots = v_{1-q} = 0$. This is a special case of (1.6).

To estimate the process with $\alpha_g$'s as parameters we define $\mathbf{u} = (u_1, \cdots, u_T)'$, where $u_t = \sum_{g=0}^{q} \alpha_g v_{t-g}$. Then $\mathscr{E}\mathbf{u} = \mathbf{0}$ and the covariance matrix of $\mathbf{u}$ is (1.8), where $\mathbf{G}_0 = \mathbf{I}$ and

$$\mathbf{G}_g = \mathbf{L}^g + \mathbf{L}'^g , \qquad g = 1, \cdots, q . \tag{2.3}$$

We modify the model (1.1) by defining $\mathbf{y}$ by

$$\textstyle\sum_{k=0}^{p} \beta_k \mathbf{L}^k \mathbf{y} = \mathbf{u} , \tag{2.4}$$

which is equivalent to setting $y_0 = y_{-1} = \cdots = y_{1-p} = 0$. This is a special case of (1.7).

2.2. *The circular model.* Let

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{2.5}$$

where $\mathbf{I}$ in the lower left-hand corner is of order $T-1$ and $\mathbf{I}$ in the upper right-hand corner is 1. Then $\mathbf{M}^t$ is of the same form, but $\mathbf{I}$ in the lower left-hand corner is of order $T-t$ and $\mathbf{I}$ in the upper right-hand corner is of order $t$, $t = 0, 1, \cdots, T-1$. Note $\mathbf{M}^T = \mathbf{I}$, and $\mathbf{M}' = \mathbf{M}^{T-1}$; thus $\mathbf{M}'^k = \mathbf{M}^{T-k} = \mathbf{M}^{-k}$. The circular model for $\mathbf{y}$ is

$$\textstyle\sum_{k=0}^{p} \beta_k \mathbf{M}^k \mathbf{y} = \sum_{g=0}^{q} \alpha_g \mathbf{M}^g \mathbf{v} , \tag{2.6}$$

corresponding to (1.1) with $y_{-k} \equiv y_{T-k}$, $k = 0, 1, \cdots, p-1$, and $v_{-g} \equiv v_{T-g}$, $g = 1, \cdots, q$. The covariance of $\mathbf{y}$ is

$$\begin{aligned} \mathscr{E}(\mathbf{y}) = \mathscr{E}\mathbf{y}\mathbf{y}' &= \sigma^2\left(\textstyle\sum_{k=0}^{p} \beta_k \mathbf{M}^k\right)^{-1} \textstyle\sum_{g,h=0}^{q} \alpha_g \alpha_h \mathbf{M}^g \mathbf{M}'^h \left(\textstyle\sum_{l=0}^{p} \beta_l \mathbf{M}'^l\right)^{-1} \\ &= \sigma^2\left(\textstyle\sum_{k=0}^{p} \beta_k \mathbf{M}^k\right)^{-1} \textstyle\sum_{g,h=0}^{q} \alpha_g \alpha_h \mathbf{M}^{g-h} \left(\textstyle\sum_{l=0}^{p} \beta_l \mathbf{M}'^l\right)^{-1} . \end{aligned} \tag{2.7}$$

Define the $T \times T$ Fourier unitary matrix as

$$\mathbf{U} = \left(\frac{1}{T^{\frac{1}{2}}} e^{i2\pi ts/T}\right). \tag{2.8}$$

Then

$$\begin{aligned} \textstyle\sum_{t=1}^{T} m_{rt} u_{ts} &= u_{r-1,s} = \frac{1}{T^{\frac{1}{2}}} e^{i2\pi(r-1)s/T} \\ &= e^{-i2\pi s/T} u_{rs}\,, \qquad r, s = 1, \cdots, T\,. \end{aligned} \tag{2.9}$$

Let $\mathbf{D}$ be a diagonal matrix with $e^{-i2\pi s/T}$ as the $s$th diagonal element. Then (2.9) can be written $\mathbf{MU} = \mathbf{UD}$, from which we obtain $\mathbf{M} = \mathbf{U}\mathbf{D}\bar{\mathbf{U}}'$, where

$$\bar{\mathbf{U}} = \left(\frac{1}{T^{\frac{1}{2}}} e^{-i2\pi ts/T}\right) \tag{2.10}$$

and $\bar{\mathbf{U}}'\mathbf{U} = \mathbf{U}\bar{\mathbf{U}}' = \mathbf{I}$. Then

$$\mathbf{M}^k = \mathbf{U}\mathbf{D}^k\bar{\mathbf{U}}'\,, \tag{2.11}$$

$$\mathbf{A} = A(\mathbf{M}) = \textstyle\sum_{g=0}^{q} \alpha_g \mathbf{M}^g = \mathbf{U}A(\mathbf{D})\bar{\mathbf{U}}'\,, \tag{2.12}$$

$$\mathbf{B} = B(\mathbf{M}) = \textstyle\sum_{k=0}^{p} \beta_k \mathbf{M}^k = \mathbf{U}B(\mathbf{D})\bar{\mathbf{U}}'\,. \tag{2.13}$$

Since $\mathbf{M}$ is real, $\mathbf{M}' = \bar{\mathbf{M}}' = \mathbf{U}\bar{\mathbf{D}}\bar{\mathbf{U}}'$, $\mathbf{A}' = \mathbf{U}A(\bar{\mathbf{D}})\bar{\mathbf{U}}'$, $\mathbf{B}' = \mathbf{U}B(\bar{\mathbf{D}})\bar{\mathbf{U}}'$, $\mathbf{A}^{-1} = \mathbf{U}A^{-1}(\mathbf{D})\bar{\mathbf{U}}'$, and $\mathbf{B}^{-1} = \mathbf{U}B^{-1}(\mathbf{D})\bar{\mathbf{U}}'$. Then

$$\mathscr{E}(\mathbf{y}) = \sigma^2 \mathbf{U}B^{-1}(\mathbf{D})A(\mathbf{D})A(\bar{\mathbf{D}})B^{-1}(\bar{\mathbf{D}})\bar{\mathbf{U}}'\,. \tag{2.14}$$

The matrix $B^{-1}(\mathbf{D})A(\mathbf{D})A(\bar{\mathbf{D}})B^{-1}(\bar{\mathbf{D}})$ is diagonal, and the $t$th diagonal element is

$$\frac{|A(e^{i\lambda_t})|^2}{|B(e^{i\lambda_t})|^2} = \frac{2\pi}{\sigma^2} f(\lambda_t)\,, \tag{2.15}$$

where $\lambda_t = 2\pi t/T$. The quadratic form in the density of $\mathbf{y}$ is $-\frac{1}{2}$ times

$$\frac{1}{\sigma^2} \mathbf{y}'\mathbf{U}B(\bar{\mathbf{D}})A^{-1}(\bar{\mathbf{D}})A^{-1}(\mathbf{D})B(\mathbf{D})\bar{\mathbf{U}}'\mathbf{y}\,. \tag{2.16}$$

If we define $\mathbf{z} = \mathbf{U}'\mathbf{y}$, that is,

$$z_t = \frac{1}{T^{\frac{1}{2}}} \textstyle\sum_{s=1}^{T} e^{i2\pi st/T} y_s = \frac{1}{T^{\frac{1}{2}}} \sum_{s=1}^{T} e^{i\lambda_t s} y_s\,, \qquad t = 1, \cdots, T\,, \tag{2.17}$$

is a component of a Fourier transform of $\mathbf{y}$, the sample spectral density (often called the "periodogram") at $\lambda = \lambda_t$ is

$$I(\lambda_t) = \frac{1}{2\pi} |z_t|^2 = \frac{1}{2\pi T} \textstyle\sum_{s,r=1}^{T} e^{i\lambda_t(s-r)} y_s y_r\,. \tag{2.18}$$

Then the exponent in the density (or likelihood function) is $-\frac{1}{2}$ times

$$\begin{aligned} &\frac{1}{\sigma^2} \mathbf{z}'B(\bar{\mathbf{D}})A^{-1}(\bar{\mathbf{D}})A^{-1}(\mathbf{D})B(\mathbf{D})\bar{\mathbf{z}} \\ &\quad = \frac{1}{\sigma^2} \textstyle\sum_{t=1}^{T} |z_t|^2 \frac{|B(e^{i\lambda_t})|^2}{|A(e^{i\lambda_t})|^2} = \sum_{t=1}^{T} \frac{I(\lambda_t)}{f(\lambda_t)}\,. \end{aligned} \tag{2.19}$$

Whittle (1953) proposed an integral analogous to this sum as an approximation to $-2$ times the exponent in the likelihood function of a stationary process; Walker (1964) made further study.

For $t = T$ and $t = \frac{1}{2}T$ if $T$ is even, $z_t$ is real:

$$(2.20)\qquad z_T = \frac{1}{T^{\frac{1}{2}}} \sum_{t=1}^{T} y_t = T^{\frac{1}{2}}\bar{y}\,, \qquad z_{\frac{1}{2}T} = \frac{1}{T^{\frac{1}{2}}} \sum_{t=1}^{T} (-1)^t y_t\,.$$

In general

$$(2.21)\qquad z_t = \frac{1}{T^{\frac{1}{2}}} \sum_{s=1}^{T} y_s \cos \lambda_t s + i\frac{1}{T^{\frac{1}{2}}} \sum_{s=1}^{T} y_s \sin \lambda_t s\,, \qquad t = 1, \cdots, T\,.$$

The real part of $z_t$ is equal to the real part of $z_{T-t}$, the imaginary part of $z_t$ is the negative of the imaginary part of $z_{T-t}$, and hence $|z_t|^2 = |z_{T-t}|^2$ and $I(\lambda_t) = I(\lambda_{T-t})$. The sum (2.19) can be written

$$(2.22)\qquad 2\sum_{t=1}^{\frac{1}{2}T-1} \frac{I(\lambda_t)}{f(\lambda_t)} + \frac{I(0)}{f(0)} + \frac{I(\pi)}{f(\pi)}$$

if $T$ is even $[I(2\pi) = I(0)$ and $f(2\pi) = f(0)]$; the first sum goes to $\frac{1}{2}(T-1)$ and the last term is dropped if $T$ is odd. In the circular model $2I(\lambda_t)/f(\lambda_t)$, $t = 1, \cdots, \frac{1}{2}T - 1$ or $\frac{1}{2}(T-1)$, are independently distributed, each with a $\chi_2{}^2$-distribution, and independently of $I(0)$ and $I(\pi)$, if $T$ is even.

When we define $\mathbf{y}$ by

$$(2.23)\qquad \sum_{k=0}^{p} \beta_k \mathbf{M}^k \mathbf{y} = \mathbf{u}\,,$$

and $\mathbf{u}$ by

$$(2.24)\qquad \mathbf{u} = \sum_{g=0}^{q} \alpha_g \mathbf{M}^g \mathbf{v}\,,$$

then the covariance matrix of $\mathbf{u}$ is (1.8) where $\mathbf{G}_0 = \mathbf{I} = \mathbf{U}\bar{\mathbf{U}}'$ and

$$(2.25)\qquad \begin{aligned} \mathbf{G}_g &= \mathbf{M}^g + \mathbf{M}'^g \\ &= \mathbf{U}(\mathbf{D}^g + \mathbf{D}^{-g})\bar{U}' = \mathbf{U}(\mathbf{D}^g + \bar{\mathbf{D}}^g)\bar{\mathbf{U}}' \\ &= \mathbf{U}\boldsymbol{\Gamma}_g\bar{\mathbf{U}}'\,, \qquad\qquad g = 1, \cdots, q\,, \end{aligned}$$

where $\boldsymbol{\Gamma}_g$ is diagonal and the $t$th diagonal element is $\gamma_{tg} = e^{i\lambda_t g} + e^{-i\lambda_t g} = 2\cos\lambda_t g$. Then

$$(2.26)\qquad \mathscr{E}(\mathbf{u}) = \boldsymbol{\Sigma}^u = \mathbf{U} \sum_{g=0}^{q} \sigma_g \boldsymbol{\Gamma}_g \bar{\mathbf{U}}'\,,$$

where $\boldsymbol{\Gamma}_0 = \mathbf{I}$. The quadratic form in the density of $\mathbf{y}$ is $-\frac{1}{2}$ times

$$(2.27)\qquad \mathbf{y}'\mathbf{U}B(\bar{\mathbf{D}})\Big(\sum_{g=0}^{q} \sigma_g \boldsymbol{\Gamma}_g\Big)^{-1} B(\mathbf{D})\bar{\mathbf{U}}'\mathbf{y} = \sum_{t=1}^{T} |z_t|^2 \frac{|B(e^{i\lambda_t})|^2}{\sum_{g=0}^{q} \sigma_g \gamma_{tg}} = \sum_{t=1}^{T} \frac{I(\lambda_t)}{f(\lambda_t)}$$

since

$$(2.28)\qquad \begin{aligned} \sum_{g=0}^{q} \sigma_g \gamma_{tg} &= \sum_{g=-q}^{q} \sigma_g \cos\lambda_t g = \sigma^2 \sum_{f,h=0}^{q} \alpha_f \alpha_h e^{i\lambda_t(h-f)} \\ &= \sigma^2 |A(e^{i\lambda_t})|^2\,. \end{aligned}$$

2.3. *Iterative procedures.* Except in the case of the purely autoregressive process, the derivatives of the likelihood function (as modified in Sections 2.1 or 2.2) lead to nonlinear equations, which cannot be solved algebraically. Several methods of solving the likelihood equations numerically are based on a Taylor's expansion, say

$$\text{(2.29)} \qquad \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\mathbf{y} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \log L(\mathbf{y} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) + \mathbf{R}(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_0) \,.$$

The Newton–Raphson method is based on setting the right-hand side equal to $\mathbf{0}$ with $\mathbf{R}(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ replaced by $\mathbf{0}$. At the $i$th step of the iteration $\boldsymbol{\theta}_0$ is the result of the previous iteration, and the equations are solved for $\boldsymbol{\theta}^*$; for $i = 1$, $\boldsymbol{\theta}_0$ is an initial estimate of $\boldsymbol{\theta}$.

In the method of scoring the matrix of second derivatives is replaced by

$$\text{(2.30)} \qquad \left[\mathscr{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \log L(\mathbf{y} \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} ,$$

where here $\mathbf{y}$ is considered as a random vector with distribution having the parameter $\boldsymbol{\theta}$. The negative of (2.30) is the information matrix (evaluated at $\boldsymbol{\theta}_0$). The iteration is based on solving for $\boldsymbol{\theta}^*$ the equations

$$\text{(2.31)} \qquad -\left[\mathscr{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \log L(\mathbf{y} \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\mathbf{y} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} .$$

In either case the iterative procedure will converge to the maximum likelihood estimate if the initial values are close enough. If the initial estimates are consistent, usually the estimates obtained at the first stage are consistent, asymptotically normal, and asymptotically efficient. The matrix of second derivatives in (2.29) or (2.30), properly normalized, is a consistent estimate of the negative of the average information matrix; the inverse of such a matrix is proportional to a consistent estimate of the covariance matrix of the limiting normal distribution.

## 3. Estimation of the coefficients of an autoregressive moving average process.

3.1. *The iterative procedures in general.* The logarithm of the (modified) likelihood function is

$$\text{(3.1)} \qquad \log L = -\tfrac{1}{2} T \log 2\pi - \tfrac{1}{2} T \log \sigma^2 + \log |\mathbf{B}| - \log |\mathbf{A}| - \frac{1}{2\sigma^2} \mathbf{y}'\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{y} \,,$$

where $\mathbf{A} = \sum_{g=0}^{q} \alpha_g \mathbf{J}_g$ and $\mathbf{B} = \sum_{k=0}^{p} \beta_k \mathbf{K}_k$ have positive determinants and $\mathbf{J}_g = \mathbf{K}_g$ is $\mathbf{L}^g$ or $\mathbf{M}^g$. If we use (Dwyer (1967) or Appendix A of Anderson (1958), for example) $\partial \log |\mathbf{A}| / \partial \alpha_h = \operatorname{tr} \mathbf{A}^{-1}\, \partial \mathbf{A} / \partial \alpha_h$, $\partial \mathbf{A}^{-1} / \partial \alpha_h = -\mathbf{A}^{-1}(\partial \mathbf{A} / \partial \alpha_h)\mathbf{A}^{-1}$, $\partial\, \mathbf{A} / \partial\, \alpha_h = \mathbf{J}_h$, etc., we obtain the derivatives ((5.4), (5.5), and (5.6) of Anderson

(1975a))

$$(3.2)\qquad \frac{\partial}{\partial \alpha_g} \log L = -\operatorname{tr} \mathbf{A}^{-1}\mathbf{J}_g + \frac{1}{\sigma^2} \mathbf{y}'\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{J}_g\mathbf{A}^{-1}\mathbf{B}\mathbf{y}\,, \qquad g = 1, \cdots, q\,,$$

$$(3.3)\qquad \frac{\partial}{\partial \beta_k} \log L = \operatorname{tr} \mathbf{B}^{-1}\mathbf{K}_k - \frac{1}{\sigma^2} \mathbf{y}'\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{K}_k\mathbf{y}\,, \qquad k = 1, \cdots, p\,,$$

$$(3.4)\qquad \frac{\partial}{\partial \sigma^2} \log L = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \mathbf{y}'\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{y}\,.$$

Setting these derivatives equal to 0 gives nonlinear equations which cannot be solved explicitly except in very special cases. Hence, an iterative procedure will be needed.

Define $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_q)'$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$. Let $\boldsymbol{\theta}'$ be the vector of parameters $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \sigma^2)$ and let $\hat{\boldsymbol{\theta}}_i' = (\hat{\boldsymbol{\alpha}}_i', \hat{\boldsymbol{\beta}}_i', \hat{\sigma}_i^2)$ be the vector of estimates to be determined at the $i$th iteration, $i = 1, 2, \cdots$, with $\hat{\boldsymbol{\theta}}_0' = (\hat{\boldsymbol{\alpha}}_0', \hat{\boldsymbol{\beta}}_0', \hat{\sigma}_0^2)$ as the vector of initial estimates. Then the equations for $\hat{\boldsymbol{\alpha}}_i$, $\hat{\boldsymbol{\beta}}_i$ at the $i$th stage are

$$(3.5)\qquad \begin{bmatrix} \hat{\boldsymbol{\Phi}}_{i-1} & \hat{\boldsymbol{\Omega}}_{i-1} \\ \hat{\boldsymbol{\Omega}}'_{i-1} & \hat{\boldsymbol{\Psi}}_{i-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\alpha}}_{i-1} \\ \hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i-1} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{q}}_{i-1} \\ \hat{\mathbf{p}}_{i-1} \end{bmatrix},$$

where the matrix on the left-hand side is an estimate of the information matrix for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ based on the $(i-1)$st estimates and the vector on the right-hand side is composed of estimates of $\partial \log L/\partial \alpha_g$, $g = 1, \cdots, q$, and $\partial \log L/\partial \beta_k$, $k = 1, \cdots, p$. In each case the estimate $\hat{\sigma}_i^2$ is a function of $\mathbf{y}$ and $\hat{\boldsymbol{\theta}}_{i-1}$. Throughout the presentation we shall add a subscript $i - 1$ and a carat ^ to any function of $\boldsymbol{\theta}$ to denote the same function of $\hat{\boldsymbol{\theta}}_{i-1}$.

3.2. *Iterative procedures in the time domain.* In the time domain we approximate the likelihood function of (1.1) by (2.2) using the matrix lag operator $\mathbf{L}$. It will be convenient to define $\hat{\mathbf{v}}_{i-1} = \hat{\mathbf{A}}_{i-1}^{-1}\hat{\mathbf{B}}_{i-1}\mathbf{y} = \hat{\mathbf{B}}_{i-1}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}$, which can be interpreted as an "estimate" of $\mathbf{v}$. In both the Newton–Raphson and scoring methods the components of the right-hand side of (3.5) are the quadratic forms

$$(3.6)\qquad [\hat{\mathbf{q}}_{i-1}]_g = \frac{\hat{\mathbf{v}}'_{i-1}\mathbf{L}^g\hat{\mathbf{A}}_{i-1}^{-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}}\,,$$

$$(3.7)\qquad [\hat{\mathbf{p}}_{i-1}]_k = -\frac{\hat{\mathbf{v}}'_{i-1}\mathbf{L}^k\hat{\mathbf{B}}_{i-1}^{-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}} = -\frac{\hat{\mathbf{v}}'_{i-1}\mathbf{L}^k\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}}{\hat{\sigma}^2_{i-1}}\,,$$

for $g = 1, \cdots, q$ and $k = 1, \cdots, p$. In the Newton–Raphson method the components of $\hat{\boldsymbol{\Phi}}_{i-1}$, $\hat{\boldsymbol{\Omega}}_{i-1}$, and $\hat{\boldsymbol{\Psi}}_{i-1}$ are the quadratic forms

$$(3.8)\qquad [\hat{\boldsymbol{\Phi}}_{i-1}]_{gf} = \frac{\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^g\mathbf{L}^f\hat{\mathbf{A}}_{i-1}^{-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}}\,,$$

$$(3.9)\qquad [\hat{\boldsymbol{\Omega}}_{i-1}]_{gl} = -\frac{\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^g\mathbf{L}^l\hat{\mathbf{B}}_{i-1}^{-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}} = -\frac{\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^g\mathbf{L}^l\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}}{\hat{\sigma}^2_{i-1}}\,,$$

$$(3.10)\qquad [\hat{\boldsymbol{\Psi}}_{i-1}]_{kl} = \frac{\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{B}}'^{-1}_{i-1}\mathbf{L}'^{k}\mathbf{L}^{l}\hat{\mathbf{B}}^{-1}_{i-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}} = \frac{\mathbf{y}'\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^{k}\mathbf{L}^{l}\hat{\mathbf{A}}^{-1}_{i-1}\mathbf{y}}{\hat{\sigma}^2_{i-1}},$$

for $g, f = 1, \cdots, q$ and $k, l = 1, \cdots, p$. In these forms of equations (3.5) the factor $\hat{\sigma}^2_{i-1}$ can be deleted, and the iteration for $\hat{\boldsymbol{\alpha}}_i$, $\hat{\boldsymbol{\beta}}_i$ can be carried out without computation of $\hat{\sigma}^2_{i-1}$.

The Newton–Raphson equation for $\hat{\sigma}_i^2$ is

$$(3.11)\qquad \left[2\,\frac{\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{v}}_{i-1}}{\hat{\sigma}^2_{i-1}} - T\right]\hat{\sigma}_i^2 = 3\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{v}}_{i-1} - 2T\hat{\sigma}^2_{i-1}\,.$$

However, instead of calculating this sequence for $i = 1, 2, \cdots$, an alternative is to use the last estimate $\hat{\mathbf{v}}_i$ in $\hat{\mathbf{v}}_i'\hat{\mathbf{v}}_i/T$; this estimate of $\sigma^2$`is obtained by replacing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_i$ in the equation obtained by setting the derivative (3.4) equal to 0.

The equations (3.5) are formally equivalent to least squares equations with $\hat{\mathbf{v}}_{i-1}$ as the vector of dependent variables and $\mathbf{L}^g\hat{\mathbf{A}}^{-1}_{i-1}\hat{\mathbf{v}}_{i-1}$ and $-\mathbf{L}^k\hat{\mathbf{B}}^{-1}_{i-1}\hat{\mathbf{v}}_{i-1}$ as the vectors of independent variables. An intuitive interpretation is that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ tend to be estimated so that the constructed vector $\hat{\mathbf{v}}$ has components uncorrelated with the components of certain linear combinations of the vectors consisting of lagged components of $\hat{\mathbf{v}}$. These sample properties reflect the process properties that $v_t$ is uncorrelated with $v_{t-s}$ and $y_{t-s}$ for $s = 1, 2, \cdots$.

Each quadratic form has the nature of $\mathbf{x}'\mathbf{z} = \sum_{t=1}^T x_t z_t$, where one or both of the vectors is of the nature $\mathbf{A}^{-1}\mathbf{w}$ or $\mathbf{B}^{-1}\mathbf{w}$. Because $\mathbf{A}$ and $\mathbf{B}$ are triangular, the components of the vector can be computed recursively. In particular, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{w}$ is found by solving $\mathbf{A}\mathbf{x} = \mathbf{w}$, yielding $x_1 = w_1$, $x_t = w_t - \sum_{s=1}^{t-1}\alpha_s x_{t-s}$, $t = 2, \cdots, q$, and $x_t = w_t - \sum_{s=1}^{q}\alpha_s x_{t-s}$, $t = q+1, \cdots, T$.

In the method of scoring the components of $\hat{\boldsymbol{\Phi}}_{i-1}$, $\hat{\boldsymbol{\Omega}}_{i-1}$, and $\hat{\boldsymbol{\Psi}}_{i-1}$ are

$$(3.12)\qquad [\hat{\boldsymbol{\Phi}}_{i-1}]_{gf} = \operatorname{tr}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^{g}\mathbf{L}^{f}\hat{\mathbf{A}}^{-1}_{i-1}\,,$$

$$(3.13)\qquad [\hat{\boldsymbol{\Omega}}_{i-1}]_{gl} = -\operatorname{tr}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^{g}\mathbf{L}^{l}\hat{\mathbf{B}}^{-1}_{i-1}\,,$$

$$(3.14)\qquad [\hat{\boldsymbol{\Psi}}_{i-1}]_{kl} = \operatorname{tr}\hat{\mathbf{B}}'^{-1}_{i-1}\mathbf{L}'^{k}\mathbf{L}^{l}\hat{\mathbf{B}}^{-1}_{i-1}\,,$$

for $g, f = 1, \cdots, q$ and $k, l = 1, \cdots, p$. The equation for $i$th estimate of the variance is $\hat{\mathbf{v}}'_{i-1}\hat{\mathbf{v}}_{i-1}/T$ or $\hat{\mathbf{v}}_i'\hat{\mathbf{v}}_i/T$. In effect, the scoring equations result from replacing $\mathbf{v}'\mathbf{A}'^{-1}\mathbf{L}'^{g}\mathbf{L}^{f}\mathbf{A}^{-1}\mathbf{v}$ by its expectation which is $\mathscr{E}\operatorname{tr}\mathbf{A}'^{-1}\mathbf{L}'^{g}\mathbf{L}^{f}\mathbf{A}^{-1}\mathbf{v}\mathbf{v}'$, etc.

If $A^{-1}(z) = \sum_{j=0}^\infty \delta_j z^j$, then $\mathbf{A}^{-1} = \sum_{t=0}^{T-1}\delta_t\mathbf{L}^t$. The coefficients are found by solving $\mathbf{A}\boldsymbol{\delta} = (1, 0, \cdots, 0)'$, where $\boldsymbol{\delta} = (\delta_0, \delta_1, \cdots, \delta_{T-1})'$. The explicit solution is $\delta_0 = 1$, $\delta_t = -\sum_{s=1}^t \alpha_s\delta_{t-s}$, $t = 1, \cdots, q$, $\delta_t = -\sum_{s=1}^q\alpha_s\delta_{t-s}$, $t = q+1, \cdots, T-1$. Then a term $\operatorname{tr}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}'^{g}\mathbf{L}^{f}\hat{\mathbf{A}}^{-1}_{i-1}$ is composed of a weighted sum of squares or products of $\hat{\delta}_j^{(i-1)}$, $j = 0, \cdots, T-1$, where $\hat{\mathbf{A}}_{i-1} = \sum_{j=0}^{T-1}\hat{\delta}_j^{(i-1)}\mathbf{L}^j$.

3.3. *Iterative procedures in the frequency domain.* To obtain procedures in the frequency domain we approximate the likelihood function of (1.1) by (2.6); this amounts to replacing $\mathbf{L}$ in Section 3.2 by $\mathbf{M} = \mathbf{U}\mathbf{D}\bar{\mathbf{U}}'$. In both the Newton–

Raphson and scoring methods the components of the right-hand sides of (3.5) are

$$(3.15)\qquad [\hat{\mathbf{q}}_{i-1}]_g = \sum_{t=1}^T \frac{I(\lambda_t)e^{i\lambda_t g}}{\hat f_{i-1}(\lambda_t)\hat A_{i-1}(e^{i\lambda_t})} = \sum_{f=0}^q \sum_{t=1}^T \frac{I(\lambda_t)\cos\lambda_t(f-g)}{\hat f_{i-1}(\lambda_t)|\hat A_{i-1}(e^{i\lambda_t})|^2}\hat\alpha_f^{(i-1)},$$

$$(3.16)\qquad [\hat{\mathbf{p}}_{i-1}]_k = -\sum_{t=1}^T \frac{I(\lambda_t)e^{i\lambda_t k}}{\hat f_{i-1}(\lambda_t)\hat B_{i-1}(e^{i\lambda_t})}$$
$$= -\sum_{l=0}^p \sum_{t=1}^T \frac{I(\lambda_t)\cos\lambda_t(l-k)}{\hat f_{i-1}(\lambda_t)|\hat B_{i-1}(e^{i\lambda_t})|^2}\hat\beta_l^{(i-1)},$$

for $g = 1, \cdots, q$ and $k = 1, \cdots, p$. In the Newton–Raphson procedure the components of the matrix on the left-hand side of (3.5) are

$$(3.17)\qquad [\mathbf{\Phi}_{i-1}]_{gf} = \sum_{t=1}^T \frac{I(\lambda_i)e^{i\lambda_t g}e^{-i\lambda_t f}}{\hat f_{i-1}(\lambda_t)|\hat A_{i-1}(e^{i\lambda_t})|^2} = \sum_{t=1}^T \frac{I(\lambda_t)\cos\lambda_t(g-f)}{\hat f_{i-1}(\lambda_t)|\hat A_{i-1}(e^{i\lambda_t})|^2},$$

$$(3.18)\qquad [\hat{\mathbf{\Omega}}_{i-1}]_{gl} = -\sum_{t=1}^T \frac{I(\lambda_t)e^{i\lambda_t g}e^{-i\lambda_t l}}{\hat f_{i-1}(\lambda_t)\hat A_{i-1}(e^{i\lambda_t})\hat B_{i-1}(e^{-i\lambda_t})}$$
$$= -\sum_{h=0}^q \sum_{j=0}^p \sum_{t=1}^T \frac{'I(\lambda_t)\cos\lambda_t(g-l+j-h)}{\hat f_{i-1}(\lambda_t)|\hat A_{i-1}(e^{i\lambda_t})|^2|\hat B_{i-1}(e^{i\lambda_t})|^2}\hat\alpha_h^{(i-1)}\hat\beta_j^{(i-1)},$$

$$(3.19)\qquad [\hat{\mathbf{\Psi}}_{i-1}]_{kl} = \sum_{t=1}^T \frac{I(\lambda_t)e^{i\lambda_t k}e^{-i\lambda_t l}}{\hat f_{i-1}(\lambda_t)|\hat B_{i-1}(e^{i\lambda_t})|^2} = \sum_{t=1}^T \frac{I(\lambda_t)\cos\lambda_t(k-l)}{\hat f_{i-1}(\lambda_t)|\hat B_{i-1}(e^{i\lambda_t})|^2},$$

for $g, f = 1, \cdots, q$ and $k, l = 1, \cdots, p$. The equation for $\hat\sigma_i^2$ is

$$(3.20)\qquad \left[2\sum_{t=1}^T \frac{I(\lambda_t)}{\hat f_{i-1}(\lambda_t)} - T\right]\hat\sigma_i^2 = \left[3\sum_{t=1}^T \frac{I(\lambda_t)}{\hat f_{i-1}(\lambda_t)} - 2T\right]\hat\sigma_{i-1}^2 .$$

In this case equations (3.5) have the nature of least squares equations with dependent variable $\hat B_{i-1}(e^{i\lambda_t})z_t/\hat A_{i-1}(e^{i\lambda_t}) = \hat w_t^{(i-1)}$, say, and independent variables $e^{i\lambda_t g}\hat w_t^{(i-1)}/\hat A_{i-1}(e^{i\lambda_t})$ and $-e^{i\lambda_t k}\hat w_t^{(i-1)}/\hat B_{i-1}(e^{i\lambda_t})$. An intuitive interpretation is that $\boldsymbol\alpha$ and $\boldsymbol\beta$ tend to be estimated so that the constructs $\hat w_t$ are uncorrelated with the "relative exponentials" $e^{i\lambda_t g}\hat w_t/\hat A(e^{i\lambda_t})$ and $-e^{i\lambda_t k}\hat w_t/\hat B(e^{i\lambda_t})$. The corresponding random variable $w_t = B(e^{i\lambda_t})z_t/A(e^{i\lambda_t})$ is proportional to a signed square root of $I(\lambda_t)/f(\lambda_t)$; the latter are independently distributed ($t = 1, \cdots, \frac{1}{2}[T]$) as $\frac{1}{2}\chi_2^2$. An alternative view is that the equations correspond to weighted least squares with weights $1/\hat f_{i-1}(\lambda_t)$, dependent variable $z_t$ and independent variables $e^{i\lambda_t g}z_t/\hat A_{i-1}(e^{i\lambda_t})$ and $-e^{i\lambda_t k}z_t/\hat B_{i-1}(e^{i\lambda_t})$.

In the scoring method the components of the left-hand side of (3.5) are

$$(3.21)\qquad [\hat{\mathbf{\Phi}}_{i-1}]_{gf} = \sum_{t=1}^T \frac{e^{i\lambda_t g}e^{-i\lambda_t f}}{|\hat A_{i-1}(e^{i\lambda_t})|^2} = \sum_{t=1}^T \frac{\cos\lambda_t(g-f)}{|\hat A_{i-1}(e^{i\lambda_t})|^2},$$

$$(3.22)\qquad [\hat{\mathbf{\Omega}}_{i-1}]_{gl} = -\sum_{t=1}^T \frac{e^{i\lambda_t g}e^{-i\lambda_t l}}{\hat A_{i-1}(e^{i\lambda_t})\hat B_{i-1}(e^{-i\lambda_t})}$$
$$= -\sum_{h=0}^q \sum_{j=0}^p \sum_{t=1}^T \frac{\cos\lambda_t(g-l+j-h)}{|\hat A_{i-1}(e^{i\lambda_t})|^2|\hat B_{i-1}(e^{i\lambda_t})|^2}\hat\alpha_h^{(i-1)}\hat\beta_j^{(i-1)},$$

$$(3.23)\qquad [\hat{\mathbf{\Psi}}_{i-1}]_{kl} = \textstyle\sum_{t=1}^{T} \dfrac{e^{i\lambda_t k} e^{-i\lambda_t l}}{|\hat{B}_{i-1}(e^{i\lambda_t})|^2} = \textstyle\sum_{t=1}^{T} \dfrac{\cos \lambda_t(k-l)}{|\hat{B}_{i-1}(e^{i\lambda_t})|^2}\,,$$

for $g, f = 1, \cdots, q$ and $k, l = 1, \cdots, p$. The iteration for the estimate of the variance is

$$(3.24)\qquad \hat{\sigma}_i^{\,2} = \frac{1}{T} \textstyle\sum_{t=1}^{T} \dfrac{I(\lambda_t)}{\hat{f}_{i-1}(\lambda_t)}\, \hat{\sigma}_{i-1}^2\,.$$

The coefficients on the left-hand side of (3.5) for the method of scoring are obtained by replacing $I(\lambda_t)/\hat{f}_{i-1}(\lambda_t)$ for the Newton–Raphson method by 1 since in the circular model $\mathscr{E}I(\lambda_t) = f(\lambda_t)$.

3.4. *Further discussion.* Advantage can be taken of the fact that $\partial \log L/\partial\beta_k$ [as in (3.3)] is linear in the $\beta_l$'s. In (3.5) if $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i-1}$ is eliminated, the equation for $\hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\alpha}}_{i-1}$ is

$$(3.25)\qquad (\hat{\mathbf{\Phi}}_{i-1} - \hat{\mathbf{\Omega}}_{i-1}\hat{\mathbf{\Psi}}_{i-1}^{-1}\hat{\mathbf{\Omega}}'_{i-1})(\hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\alpha}}_{i-1}) = \hat{\mathbf{q}}_{i-1} - \hat{\mathbf{\Omega}}_{i-1}\hat{\mathbf{\Psi}}_{i-1}^{-1}\hat{\mathbf{p}}_{i-1}\,.$$

If partial derivatives of the logarithm of the likelihood function with respect to the $\beta_k$'s in the direct lag model are set equal to 0 and if $\mathbf{A}$ is replaced by $\hat{\mathbf{A}}_{i-1}$ and $\beta_l$ by $\hat{\beta}_l$, $l = 1, \cdots, p$, the resulting equations are

$$(3.26)\qquad \textstyle\sum_{l=1}^{p} \mathbf{y}'\hat{\mathbf{A}}_{i-1}'^{-1}\mathbf{L}'^{l}\mathbf{L}^{k}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}\hat{\beta}_l = -\mathbf{y}'\hat{\mathbf{A}}_{i-1}'^{-1}\mathbf{L}^{k}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}\,, \qquad k = 1, \cdots, p\,.$$

When these estimates of $\beta_1, \cdots, \beta_p$ are used in place of $\hat{\beta}_1^{\,(i-1)}, \cdots, \hat{\beta}_p^{\,(i-1)}$ in (3.25) the replacement of $\hat{\mathbf{p}}_{i-1}$ is $\mathbf{0}$ and the right-hand side of (3.25) is simply $\hat{\mathbf{q}}_{i-1}$ with these estimates of $\beta_1, \cdots, \beta_p$. In the case of the circular model the equations corresponding to (3.26) are

$$(3.27)\qquad \textstyle\sum_{l=1}^{p}\sum_{t=1}^{T} \dfrac{I(\lambda_t)\cos\lambda_t(l-k)}{\hat{f}_{i-1}(\lambda_t)|\hat{B}_{i-1}(e^{i\lambda_t})|^2}\,\hat{\beta}_l = -\textstyle\sum_{t=1}^{T} \dfrac{I(\lambda_t)\cos\lambda_t k}{\hat{f}_{i-1}(\lambda_t)|\hat{B}_{i-1}(e^{i\lambda_t})|^2}\,, \qquad k = 1, \cdots, p\,.$$

In any method if the estimate $\hat{\boldsymbol{\theta}}_{i-1}$ is consistent for any $i = 1, 2, \cdots$, the coefficients in the left-hand side of (3.5) divided by $T$ are consistent estimates of corresponding elements of the limiting average information matrix. More precisely

$$(3.28)\qquad \begin{aligned} &\operatorname{p\,lim}_{T\to\infty}\left[\frac{1}{T}\hat{\mathbf{\Phi}}_{i-1}\right]_{gf} \\ &\quad = \lim_{T\to\infty}\frac{1}{T}\textstyle\sum_{t=1}^{T}\dfrac{\cos\lambda_t(f-g)}{|A(e^{i\lambda_t})|^2} = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi}\dfrac{\cos\lambda(g-f)}{|A(e^{i\lambda})|^2}\,d\lambda \\ &\quad = \textstyle\sum_{t=\max(g,f)}^{\infty}\delta_{t-g}\delta_{t-f} = \phi_{gf}\,, \end{aligned}$$

$$(3.29)\qquad \begin{aligned} \operatorname{p\,lim}_{T\to\infty}\left[\frac{1}{T}\hat{\mathbf{\Omega}}_{i-1}\right]_{gl} &= -\lim_{T\to\infty}\frac{1}{T}\textstyle\sum_{t=1}^{T}\dfrac{e^{i\lambda_t(g-l)}}{A(e^{i\lambda_t})B(e^{-i\lambda_t})} \\ &= -\dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi}\dfrac{e^{i\lambda(g-l)}}{A(e^{i\lambda})B(e^{-i\lambda})}\,d\lambda \\ &= -\textstyle\sum_{t=\max(g,l)}^{\infty}\delta_{t-g}\gamma_{t-l} = \omega_{gl}\,, \end{aligned}$$

$$(3.30)\qquad \begin{aligned} \operatorname{p\,lim}_{T\to\infty}\left[\frac{1}{T}\hat{\mathbf{\Psi}}_{i-1}\right]_{kl} &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\cos\lambda_t(k-l)}{|B(e^{i\lambda_t})|^2} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\cos\lambda(k-l)}{|B(e^{i\lambda})|^2}\,d\lambda \\ &= \sum_{t=\max(k,l)}^{\infty}\gamma_{t-k}\gamma_{t-l} = \psi_{kl} \end{aligned}$$

for $g, f = 1, \cdots, q$ and $k, l = 1, \cdots, p$, where $A^{-1}(z) = \sum_{s=0}^{\infty}\delta_s z^s$ and $B^{-1}(z) = \sum_{s=0}^{\infty}\gamma_s z^s$. The components of the limits of the average information matrix referring to components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vs. $\sigma^2$ are 0. The limit of the average information for $\sigma^2$ is $1/(2\sigma^4)$.

In any method if $\hat{\boldsymbol{\theta}}_{i-1}$ is a consistent estimate of $\boldsymbol{\theta}$ of order $T^{-\frac{1}{2}}$ in probability, then as $T \to \infty$ $T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix equal to the inverse of the limit of the average information matrix (as detailed above). The estimates are asymptotically efficient.

In the model using the matrix $\mathbf{L}$ the matrices $\mathbf{A}$ and $\mathbf{B}$ are triangular with 1's on the main diagonal; hence $|\mathbf{A}| = |\mathbf{B}| = 1$, and their derivatives with respect to elements of $\boldsymbol{\theta}$ are 0. The equations for the scoring method are based exactly on $\mathscr{E}\,\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$, but the equations for the Newton–Raphson method involve dropping from $\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ terms which are asymptotically negligible as $T \to \infty$. In the circular model $\mathbf{A}$ and $\mathbf{B}$ are circulants, and the determinants depend on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. To obtain the equations (3.5), terms which are asymptotically negligible have been deleted from $\partial \log L/\partial\boldsymbol{\theta}$ and from $\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ in the Newton–Raphson method and from $\mathscr{E}\,\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ in the scoring method. A more detailed discussion is provided in Section 6.

## 4. Estimation of the moving average covariances and the autoregressive coefficients.

4.1. *The iterative procedures in general.* When the parameters are the $\beta_k$'s and $\sigma_k$'s, the logarithm of the (modified) likelihood function is

$$(4.1)\qquad \log L = -\tfrac{1}{2}T\log 2\pi - \tfrac{1}{2}\log|\mathbf{\Sigma}^u| + \log|\mathbf{B}| - \tfrac{1}{2}\mathbf{y}'\mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{B}\mathbf{y}\,,$$

where $\mathbf{\Sigma}^u$ is given by (1.8) and $\mathbf{B} = \sum_{k=0}^{p}\beta_k\mathbf{K}_k$. In the model with unobserved variables set equal to 0, $\mathbf{G}_g = \mathbf{L}^g + \mathbf{L}'^g$, $g = 1, \cdots, q$, and $\mathbf{K}_k = \mathbf{L}^k$, $k = 0, 1, \cdots, p$, and in the circular model $\mathbf{G}_g = \mathbf{M}^g + \mathbf{M}'^g$, $g = 1, \cdots, q$, and $\mathbf{K}_k = \mathbf{M}^k$, $k = 0, 1, \cdots, p$. The derivatives are

$$(4.2)\qquad \frac{\partial}{\partial\sigma_g}\log L = -\tfrac{1}{2}\operatorname{tr}(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_g + \tfrac{1}{2}\mathbf{y}'\mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_g(\mathbf{\Sigma}^u)^{-1}\mathbf{B}\mathbf{y}\,, \qquad g = 0, 1, \cdots, q\,,$$

$$(4.3)\qquad \frac{\partial}{\partial\beta_k}\log L = \operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_k - \mathbf{y}'\mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_k\mathbf{y}\,, \qquad k = 1, \cdots, p\,.$$

Let $\boldsymbol{\sigma}' = (\sigma_0, \sigma_1, \cdots, \sigma_q)$, $\boldsymbol{\theta}' = (\boldsymbol{\sigma}', \boldsymbol{\beta}')$, $\hat{\boldsymbol{\theta}}_i' = (\hat{\boldsymbol{\sigma}}_i', \hat{\boldsymbol{\beta}}_i')$, the vector of estimates at the $i$th stage, $i = 1, 2, \cdots$, and $\hat{\boldsymbol{\theta}}_0$ the vector of initial estimates. The equations for $\hat{\boldsymbol{\theta}}_i$ at the $i$th stage are

$$\begin{bmatrix} \hat{\mathbf{\Lambda}}_{i-1} & \hat{\mathbf{N}}_{i-1} \\ \hat{\mathbf{N}}'_{i-1} & \hat{\mathbf{\Psi}}_{i-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\sigma}}_i - \hat{\boldsymbol{\sigma}}_{i-1} \\ \hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i-1} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{s}}_{i-1} \\ \hat{\mathbf{p}}_{i-1} \end{bmatrix}, \tag{4.4}$$

where the matrix on the left-hand side is an estimate of the information matrix for $\boldsymbol{\theta}$ and the vector on the right-hand side is composed of estimates of $\partial \log L/\partial\sigma_h$ and $\partial \log L/\partial\beta_k$.

4.2. *Iterative procedures in the time domain.* Let $\hat{\mathbf{u}}_{i-1} = \hat{\mathbf{B}}_{i-1}\mathbf{y} = \hat{\mathbf{A}}_{i-1}\hat{\mathbf{v}}_{i-1}$, where $\hat{\mathbf{B}}_{i-1} = \hat{B}_{i-1}(\mathbf{L})$ and $\hat{\mathbf{A}}_{i-1} = \hat{A}_{i-1}(\mathbf{L})$. Then

$$[\hat{\mathbf{s}}_{i-1}]_g = \tfrac{1}{2}\hat{\mathbf{u}}'_{i-1}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})\hat{\mathbf{u}}_{i-1} - \tfrac{1}{2}\operatorname{tr}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g\,, \tag{4.5}$$

$$[\hat{\mathbf{p}}_{i-1}]_k = -\hat{\mathbf{u}}'_{i-1}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{L}^k\mathbf{y}\,, \tag{4.6}$$

for $g = 0, 1, \cdots, q$ and $k = 1, \cdots, p$. In the Newton–Raphson method

$$\begin{aligned}[\hat{\mathbf{\Lambda}}_{i-1}]_{gf} = {} & \hat{\mathbf{u}}'_{i-1}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_f(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\hat{\mathbf{u}}_{i-1} \\ & - \tfrac{1}{2}\operatorname{tr}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_f\,,\end{aligned} \tag{4.7}$$

$$[\hat{\mathbf{N}}_{i-1}]_{gl} = -\hat{\mathbf{u}}'_{i-1}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{L}^l\mathbf{y}\,, \tag{4.8}$$

$$[\hat{\mathbf{\Psi}}_{i-1}]_{kl} = \mathbf{y}'\mathbf{L}'^k(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{L}^l\mathbf{y}\,, \tag{4.9}$$

for $g, f = 0, 1, \cdots, q$ and $k, l = 1, \cdots, p$. The scoring procedure is developed from the fact that $\mathscr{E}\mathbf{u}\mathbf{u}' = \mathbf{\Sigma}^u$ and $\mathbf{y} = \mathbf{B}^{-1}\mathbf{u}$. The coefficients are

$$[\hat{\mathbf{\Lambda}}_{i-1}]_{gf} = \tfrac{1}{2}\operatorname{tr}(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{G}_f\,, \tag{4.10}$$

$$[\hat{\mathbf{N}}_{i-1}]_{gl} = -\operatorname{tr}\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}\mathbf{L}^l\hat{\mathbf{B}}^{-1}_{i-1}\,, \tag{4.11}$$

and (3.14) for $g, f = 0, 1, \cdots, q$ and $k, l = 1, \cdots, p$. The equations (4.4) can be simplified somewhat by rewriting them as equations for $\hat{\boldsymbol{\sigma}}_i$ and $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i-1}$.

In the case of the pure moving average process the equations for $\hat{\boldsymbol{\sigma}}^{(i)}$ can be written (with $u$ dropped from $\mathbf{\Sigma}^u$)

$$\textstyle\sum_{f=0}^{q}\operatorname{tr}\hat{\mathbf{\Sigma}}^{-1}_{i-1}\mathbf{G}_g\hat{\mathbf{\Sigma}}^{-1}_{i-1}\mathbf{G}_f\hat{\sigma}^{(i)}_f = \mathbf{y}'\hat{\mathbf{\Sigma}}^{-1}_{i-1}\mathbf{G}_g\hat{\mathbf{\Sigma}}^{-1}_{i-1}\mathbf{y}\,, \qquad g = 0, 1, \cdots, q\,. \tag{4.12}$$

These equations constitute a weighted least squares (as shown by Anderson (1969) and (1973)). If the different components of $\mathbf{y}\mathbf{y}'$ are arranged in a vector, say $\mathbf{c}$, and those of $\mathbf{G}_g$ similarly, say $\mathbf{g}_g$, then $\mathscr{E}\mathbf{c} = \sum_{g=0}^{q}\sigma_g\mathbf{g}_g$, and (4.12) with $\hat{\mathbf{\Sigma}}_{i-1}$ replaced by $\mathbf{\Sigma}$ defines the weighted least squares (or Markov) estimates of $\sigma_0, \sigma_1, \cdots, \sigma_q$.

The calculation of vectors in the form $\mathbf{x} = \mathbf{\Sigma}^{-1}\mathbf{w}$ amounts to obtaining the solution of $\mathbf{\Sigma}\mathbf{x} = \mathbf{w}$ for $\mathbf{x}$. The forward solution is recursive as in calculating $\mathbf{A}^{-1}\mathbf{z}$; the backward solution for $\mathbf{x}$ is also recursive. (See Anderson (1971 b) for details.) The computation of coefficients involving $\mathbf{G}_g(\hat{\mathbf{\Sigma}}^u_{i-1})^{-1}$ is more complicated. Approximations to $[\hat{\mathbf{\Lambda}}_{i-1}]_{gf}$ in the scoring method and exact computations for $q = 1$ were given in Anderson (1971 b).

4.3. *Iterative procedures in the frequency domain.* The components on the right-hand side of (4.4) are, with $n_0 = 1$ and $n_g = 2$, $g = 1, \cdots, q$,

$$[\hat{\mathbf{s}}_{i-1}]_g = \tfrac{1}{2} n_g \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t g}{\hat{f}_{i-1}(\lambda_t) 2\pi \hat{f}_{i-1}^u(\lambda_t)} - \tfrac{1}{2} n_g \textstyle\sum_{t=1}^T \dfrac{\cos \lambda_t g}{2\pi \hat{f}_{i-1}^u(\lambda_t)}, \tag{4.13}$$

$$\begin{aligned}[\hat{\mathbf{p}}_{i-1}]_k &= -\textstyle\sum_{l=0}^p \sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t(l-k)}{\hat{f}_{i-1}^u(\lambda_t)} \hat{\beta}_l^{(i-1)} \\ &= -\textstyle\sum_{l=0}^p \sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t(l-k)}{\hat{f}_{i-1}(\lambda_t)|\hat{B}_{i-1}(e^{i\lambda_t})|^2} \hat{\beta}_l^{(i-1)}\end{aligned} \tag{4.14}$$

for $g = 0, 1, \cdots, q$ and $k = 1, \cdots, p$. In the Newton-Raphson method

$$[\hat{\mathbf{\Lambda}}_{i-1}]_{gf} = n_g n_f \left[ \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t g \cos \lambda_t f}{\hat{f}_{i-1}(\lambda_t)[2\pi \hat{f}_{i-1}^u(\lambda_t)]^2} - \tfrac{1}{2} \textstyle\sum_{t=1}^T \dfrac{\cos \lambda_t g \cos \lambda_t f}{[2\pi \hat{f}_{i-1}^u(\lambda_t)]^2} \right], \tag{4.15}$$

$$\begin{aligned}[\hat{\mathbf{N}}_{i-1}]_{gl} &= -n_g \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t g\, e^{-i\lambda_t l} \hat{B}_{i-1}(e^{i\lambda_t})}{2\pi[\hat{f}_{i-1}^u(\lambda_t)]^2} \\ &= -n_g \textstyle\sum_{t=1}^T \sum_{m=0}^p \dfrac{I(\lambda_t) \cos \lambda_t g \cos \lambda_t(m-l)}{2\pi[\hat{f}_{i-1}^u(\lambda_t)]^2} \hat{\beta}_m^{(i-1)} \\ &= -n_g \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t g\, e^{-i\lambda_t l}}{2\pi \hat{f}_{i-1}(\lambda_t) \hat{f}_{i-1}^u(\lambda_t) \hat{B}_{i-1}(e^{-i\lambda_t})},\end{aligned} \tag{4.16}$$

$$[\hat{\mathbf{\Psi}}_{i-1}]_{kl} = \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) e^{i\lambda_t k} e^{-i\lambda_t l}}{\hat{f}_{i-1}^u(\lambda_t)} = \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t(k-l)}{\hat{f}_{i-1}(\lambda_t)|\hat{B}_{i-1}(e^{i\lambda_t})|^2}, \tag{4.17}$$

for $g, f = 0, 1, \cdots, q$ and $k, l = 1, \cdots, p$. Note that $\hat{\mathbf{p}}_{i-1}$ and $\hat{\mathbf{\Psi}}_{i-1}$ are the same as in Section 3.3. Here $\hat{f}_{i-1}^u(\lambda_t) = \sum_{g=1}^q n_g \hat{\gamma}_g^{(i-1)} \cos \lambda_t g/(2\pi)$.

The scoring procedure is obtained from the Newton-Raphson procedure by replacing $I(\lambda_t)/\hat{f}_{i-1}(\lambda_t)$ by 1 on the left-hand side because in the circular model $\mathscr{E} I(\lambda_t) = f(\lambda_t)$. The components of the left-hand side of (4.4) are

$$[\hat{\mathbf{\Lambda}}_{i-1}]_{gf} = \tfrac{1}{2} n_g n_f \textstyle\sum_{t=1}^T \dfrac{\cos \lambda_t g \cos \lambda_t f}{[2\pi \hat{f}_{i-1}^u(\lambda_t)]^2}, \tag{4.18}$$

$$\begin{aligned}[\hat{\mathbf{N}}_{i-1}]_{gl} &= -n_g \textstyle\sum_{t=1}^T \dfrac{\cos \lambda_t g\, e^{-i\lambda_t l}}{2\pi \hat{f}_{i-1}^u(\lambda_t) \hat{B}_{i-1}(e^{-i\lambda_t})} \\ &= -n_g \textstyle\sum_{t=1}^T \sum_{m=0}^p \dfrac{\cos \lambda_t g \cos \lambda_t(m-l)}{2\pi \hat{f}_{i-1}^u(\lambda_t)|\hat{B}_{i-1}(\lambda_t)|^2} \hat{\beta}_m^{(i-1)},\end{aligned} \tag{4.19}$$

and $[\hat{\mathbf{\Psi}}_{i-1}]_{kl}$ given by (3.23) for $g, f = 0, 1, \cdots, q$ and $k, l = 1, \cdots, p$.

In the case of the pure moving average process the scoring equations can be written (with $u$ dropped from $f^u(\lambda)$) as

$$n_g \textstyle\sum_{f=0}^q n_f \sum_{t=1}^T \dfrac{\cos \lambda_t g \cos \lambda_t f}{[2\pi \hat{f}_{i-1}(\lambda_t)]^2} \hat{\sigma}_f^{(i)} = n_g \textstyle\sum_{t=1}^T \dfrac{I(\lambda_t) \cos \lambda_t g}{2\pi \hat{f}_{i-1}^2(\lambda_t)}, \qquad g = 0, 1, \cdots, q. \tag{4.20}$$

These equations can be considered as least squares with $I(\lambda_t)/\hat{f}_{i-1}(\lambda_t)$ as the dependent variable and $\cos \lambda_t g/[2\pi \hat{f}_{i-1}(\lambda_t)]$ as the independent variables or as weighted least squares with $1/[2\pi \hat{f}_{i-1}(\lambda_t)]^2$ as weights, $2\pi I(\lambda_t)$ as the dependent variable, and $\cos \lambda_t g$ as the independent variables.

4.4. *Further discussion.* As in the other parametrization, we can exploit the fact that $\partial \log L/\partial \beta_k$ is linear in the $\beta_l$'s. If $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{i-1}$ is eliminated in (4.4), the equation for $\hat{\boldsymbol{\sigma}}_i - \hat{\boldsymbol{\sigma}}_{i-1}$ is

$$(4.21) \qquad (\hat{\mathbf{\Lambda}}_{i-1} - \hat{\mathbf{N}}_{i-1}\hat{\mathbf{\Psi}}_{i-1}^{-1}\hat{\mathbf{N}}'_{i-1})(\hat{\boldsymbol{\sigma}}_i - \hat{\boldsymbol{\sigma}}_{i-1}) = \hat{\mathbf{s}}_{i-1} - \hat{\mathbf{N}}_{i-1}\hat{\mathbf{\Psi}}_{i-1}^{-1}\hat{\mathbf{p}}_{i-1} \,.$$

One can estimate $\beta_1, \cdots, \beta_p$ in the matrix lag operator model from

$$(4.22) \qquad \textstyle\sum_{l=1}^{p} \mathbf{y}'\mathbf{L}'^k(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^l\mathbf{y}\hat{\beta}_l = -\mathbf{y}'\mathbf{L}'^k(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{y} \,, \qquad k = 1, \cdots, p \,,$$

and in the circular model from

$$(4.23) \qquad \textstyle\sum_{l=1}^{p} \sum_{t=1}^{T} \dfrac{I(\lambda_t)\cos \lambda_t(k-l)}{\hat{f}_{i-1}^u(\lambda_t)}\hat{\beta}_l = -\sum_{t=1}^{T} \dfrac{I(\lambda_t)\cos \lambda_t k}{\hat{f}_{i-1}^u(\lambda_t)} \,, \qquad k = 1, \cdots, p \,.$$

When $\hat{\mathbf{p}}_{i-1}$ is computed on the basis of these estimates, it is $\mathbf{0}$.

If the estimate $\hat{\boldsymbol{\theta}}_{i-1}$ is consistent for any $i = 1, 2, \cdots$, the coefficients on the left-hand side of (4.4) yield consistent estimates of corresponding elements of the limiting average information matrix. That is,

$$(4.24) \qquad \operatorname{p\,lim}_{T\to\infty}\left[\frac{1}{T}\hat{\mathbf{\Lambda}}_{i-1}\right]_{gf} = \tfrac{1}{2}n_g n_f \lim_{T\to\infty} \frac{1}{T}\textstyle\sum_{t=1}^{T} \dfrac{\cos \lambda_t g \cos \lambda_t f}{[2\pi f^u(\lambda_t)]^2}$$
$$= \frac{1}{2}\frac{n_g n_f}{(2\pi)^3}\int_{-\pi}^{\pi} \frac{\cos \lambda_g \cos \lambda_f}{[f^u(\lambda)]^2}\,d\lambda = \lambda_{gf} \,,$$

$$\operatorname{p\,lim}_{T\to\infty}\left[\frac{1}{T}\hat{\mathbf{N}}_{i-1}\right]_{gl}$$
$$(4.25) \qquad = -n_g \lim_{T\to\infty} \textstyle\sum_{m=0}^{p} \dfrac{1}{T} \sum_{t=1}^{T} \dfrac{\cos \lambda_t g \cos \lambda_t(m-l)}{2\pi f^u(\lambda_t)|B(e^{i\lambda_t})|^2}\beta_m$$
$$= -\frac{n_g}{(2\pi)^2}\textstyle\sum_{m=0}^{p} \displaystyle\int_{-\pi}^{\pi} \frac{\cos \lambda g \cos \lambda(m-l)}{f^u(\lambda)|B(e^{i\lambda})|^2}\,d\lambda\,\beta_m = \nu_{gl}$$

for $g, f = 0, 1, \cdots, q$ and $l = 1, \cdots, p$. If $\hat{\boldsymbol{\theta}}_{i-1}$ is a consistent estimate of $\boldsymbol{\theta}$ of order $T^{-\frac{1}{2}}$ in probability as $T \to \infty$, then $T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix equal to the inverse of the limit of the average information matrix. The estimates are asymptotically efficient.

The coefficients for the lag operator model are exactly the partial derivatives of $\log L$ and their expectations. In the circular model a term has been dropped from $\hat{\mathbf{p}}_{i-1}$ and a term from $\hat{\mathbf{\Psi}}_{i-1}$; each of these is asymptotically negligible as $T \to \infty$. (See Section 6.) The matrix $\hat{\mathbf{\Sigma}}_{i-1}$ should be positive definite, and since $\hat{f}_i^u(\lambda)$ is an estimate of a spectral density it should be positive for $-\pi \leqq \lambda \leqq \pi$. A solution of (4.4) for $\hat{\boldsymbol{\sigma}}_i$ may lead to estimates of $\mathbf{\Sigma}^u$ and $f^u(\lambda)$ that do not satisfy the necessary conditions. A remedy for this difficulty is to replace $\hat{\sigma}_0^{(i)}$

by a number large enough to insure that the estimated matrix and spectral density satisfy the necessary conditions. This procedure amounts to adding possibly a multiple of $\mathbf{I}$ to $\hat{\mathbf{\Sigma}}_i^u$ or a constant to $\hat{f}_i^u(\lambda)$. (The suggestion is similar to the use of ridge regression.)

Initial consistent estimates of $\beta_1, \cdots, \beta_p$ can be obtained by solving the Yule–Walker type equations

$$\sum_{l=1}^{p} c_{k-l}\hat{\beta}_l^{(0)} = -c_k\,, \qquad k = q+1, \cdots, q+p\,, \tag{4.26}$$

where

$$c_k = c_{-k} = \frac{1}{T}\sum_{t=1}^{T-k} y_t y_{t+k} = \frac{1}{T}\,\mathbf{y}'\mathbf{L}^k\mathbf{y}\,, \qquad k = 0, 1, \cdots, q+p\,. \tag{4.27}$$

Then construct $\hat{\mathbf{u}}_0 = \hat{\mathbf{B}}_0\mathbf{y}$ and estimate $\sigma_0, \sigma_1, \cdots, \sigma_q$ by

$$\hat{\sigma}_g^{(0)} = \frac{1}{T}\,\hat{\mathbf{u}}_0'\mathbf{L}_g\hat{\mathbf{u}}_0\,, \qquad g = 0, 1, \cdots, q\,. \tag{4.28}$$

When initial estimates of $\alpha_1, \cdots, \alpha_q$ are needed (Section 3), estimate $f^u(\lambda)$ by

$$\hat{f}_0^u(\lambda) = \frac{1}{2\pi}\sum_{g=0}^{q} n_g\hat{\sigma}_g^{(0)}\cos\lambda g \tag{4.29}$$

and factor $\hat{f}_0^u(\lambda)$ into $\hat{\sigma}_0^2|\hat{A}_0(e^{i\lambda})|^2/(2\pi)^2$, where $\hat{A}_0(z) = 1 + \hat{\alpha}_1^{(0)}z + \cdots + \hat{\alpha}_q^{(0)}z^q$ has real coefficients and zeros outside the unit circle. ($\hat{\sigma}_0^{(0)}$ may have to be adjusted to assure $\hat{f}_0^u(\lambda) \geqq 0$.) The (implied) equations $\hat{\sigma}_g^{(0)} = \hat{\sigma}_0^2\sum_{f=0}^{q-g}\hat{\alpha}_f^{(0)}\hat{\alpha}_{f+g}^{(0)}$, $g = 0, 1, \cdots, q$, can be solved by an algorithm of Wilson (1969).

## 5. Comparison of procedures.

5.1. *Selection of parameters.* For the moving average part there is a choice of parameters, either the covariances $\sigma_0, \sigma_1, \cdots, \sigma_q$ of $u_t = \sum_{g=0}^{q}\alpha_g v_{t-g}$ or the coefficients $\alpha_1, \cdots, \alpha_q$ and the variance $\sigma^2$ of $v_t$. (We shall not consider the use of $\alpha_0, \alpha_1, \cdots, \alpha_q$ with $\sigma^2 = 1$.) An advantage of the use of the covariances is that no modification of the moving average part is made in the direct lag model; $\mathbf{\Sigma}^u = \sum_{g=0}^{q}\sigma_g\mathbf{G}_g$ is exactly the covariance matrix of $u_1, \cdots, u_T$. If the process is purely moving average, the iterative procedures (4.12) and (4.20) are based on the exact (modified) likelihoods. The lag operator model is exactly the density of the stationary process.

An advantage of estimating the coefficients directly is that often they are the parameters of interest. If the covariances are estimated, to obtain estimates of the coefficients the resulting estimated spectral density $\hat{f}_i^u(\lambda) = \sum_{g=0}^{q}\hat{\sigma}_g^{(i)}n_g\cos\lambda g/(2\pi)$ must be "factored," that is, must be expressed as $\hat{\sigma}_i^2\sum_{f,g=0}^{q}\hat{\alpha}_f^{(i)}\hat{\alpha}_g^{(i)}\cos\lambda(f-g)/(2\pi)$ for real $\hat{\alpha}_1^{(i)}, \cdots, \hat{\alpha}_q^{(i)}$. This can be done if and only if $\hat{f}_i^u(\lambda) \geqq 0$, $-\pi \leqq \lambda \leqq \pi$. If the estimated spectral density is not nonnegative, it can be made so by increasing $\hat{\sigma}_0^{(i)}$ sufficiently, but this method is arbitrary.

In the time domain the computation of the coefficients in the equations with

$\mathbf{L}$ is more difficult for the covariances than the coefficients of the moving average part because $\operatorname{tr} \hat{\mathbf{\Sigma}}_{i-1}^{-1} \mathbf{G}_g \hat{\mathbf{\Sigma}}_{i-1}^{-1} \mathbf{G}_f$ involves the inverse of $\hat{\mathbf{\Sigma}}_{i-1}$ rather than of $\hat{\mathbf{A}}_{i-1}$. (See Anderson (1971 b), (1973), (1977) for further discussion of the computation.)

To make the definition of $\alpha_1, \cdots, \alpha_q$ unique we took the zeros of $A(z)$ to be outside the unit circle (on the assumption that none is on the unit circle). The corresponding conditions may be placed on the estimates to make them unique. However, there seems to be evidence that there are computational difficulties when some roots of $\hat{A}_i(z)$ are near the unit circle.

5.2. *Direct lag and circular models.* The model based on $\mathbf{L}$ involves approximating the stationary process by a distribution based on $y_0 = y_{-1} = \cdots = y_{1-p} = 0$ and (in case of moving average coefficients) $v_0 = v_{-1} = \cdots = v_{1-q} = 0$. For large $T$ this effect washes out. The circular model involves a distribution based on $y_0 \equiv y_T, y_{-1} \equiv y_{T-1}, \cdots, y_{1-p} \equiv y_{T-p+1}$ and $v_0 \equiv v_T, v_{-1} \equiv v_{T-1}, \cdots, v_{1-q} \equiv v_{T-q+1}$. This would appear to be a rougher approximation to a stationary process model than the model based on $\mathbf{L}$. To obtain the estimation equations more negligible terms must be dropped from the circular model than the other model. Of course, the asymptotic distributions (as $T \to \infty$) are the same.

The circular model requires computation of $I(\lambda_t) = I(2\pi t/T)$, $t = 1, \cdots, T$. By use of the fast Fourier transform the number of calculations (multiplications, for example) is roughly proportional to $T \log_2 T$. The lag operator model for the coefficients involves (at the $i$th stage) $\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{w}$, where $\mathbf{w}$ is a $T$-component vector. The number of calculations is proportional to $qT$. In the time domain updating may be easier.

5.3. *Scoring versus Newton–Raphson.* In the case of the matrix lag operator model for the right-hand side of the iteration equations one must compute $\hat{\mathbf{A}}_{i-1}^{-2}\mathbf{y} = \hat{\mathbf{A}}_{i-1}^{-1}(\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y})$ and similar vectors; the number of calculations is roughly proportional to $qT$. Given such vectors, the computation of the coefficients of the equations in the Newton–Raphson method only involves obtaining $\frac{1}{2}(q + p)(q + p + 1)$ sums of squares and cross-products. The scoring method involves rather similar calculations when $\alpha_1, \cdots, \alpha_q$ and $\sigma^2$ are among the parameters; the first column of $\hat{\mathbf{A}}_{i-1}^{-1}$ is the solution of the linear equations with the first column of $\mathbf{I}$ on the right-hand side, and each other column is the first column shifted down. In the circular model the terms on the right-hand sides of the iterative equations are also used as terms on the left-hand sides in the Newton–Raphson procedure; however, the elements of $\hat{\mathbf{\Omega}}_{i-1}$ have to be computed in addition.

Maximum likelihood methods in the time domain for the coefficients (proposed by Åström and Bohlin (1966)) have been used and studied by control and electrical engineers. Numerical aspects of the likelihood maximization have been extensively discussed in papers which appeared in a special issue on system identification and time-series analysis of the *IEEE Transactions on Automatic*

*Control* (Number 6 of Volume **AC-19** (1974)). In particular, Gupta and Mehra (1974) have indicated some of the computational problems of implementing the Newton–Raphson method and the scoring method (termed Gauss–Newton) as well as variable metric methods and suggest modifications for singular or nearly singular estimated information matrices. A comprehensive summary of the developments in computation, which should include the more general area of maximization and optimization of nonlinear functions, is beyond the scope of this paper. Mention might be made of a computer program by Akaike, Arahata and Ozaki (1975) (referred to by Tong (1975)).

5.4. *Monte Carlo studies.* The eight sets of estimates of the coefficients (four obtained from the estimates of $\sigma_g$'s and $\beta_k$'s) have the same asymptotic distribution as $T \rightarrow \infty$. However, for any finite value of $T$ the sampling distributions of the eight sets will differ. Because the exact distribution of a set of estimates obtained by an iterative procedure is usually intractable, the distributions of these estimates for small or moderate lengths of series have not been obtained. To compare procedures on the basis of sampling characteristics, resort to Monte Carlo studies must be made. Since it is the moving average aspect that requires the use of iteration, the relevant simulation studies have mainly been done for the pure moving average process, in fact, in the simplest case: $q = 1$.

McClave (1974) has investigated the Newton–Raphson method for coefficients and the scoring method for covariances in the frequency domain and the scoring method for covariances in the time domain (as well as methods due to Durbin and Walker, which are not considered in this paper because they are based on principles different from approximating the likelihood of the observations). In this study $T = 100$, the number of replications is 100, and the values of $\alpha_1$ are .5 and .9. One apparent conclusion is that although the initial estimate is consistent, the estimate from the first iteration has a mean and standard deviation quite different from $\alpha_1$ and the asymptotic standard deviation, respectively. Several successive iterations improve the performance of each procedure. At $\alpha_1 = .5$ the behavior of the scoring method for covariances seems to stabilize on the third iteration and does not differ from the asymptotic theory; the behavior of the other two methods seems to stabilize at the fifth or sixth iteration and also does not differ from the asymptotic theory. After a sufficient number of iterations (at most ten) at $\alpha_1 = .5$ the estimates by each of the three methods agree with asymptotic theory and there is no statistically sound evidence on which to prefer one method.[2]

At $\alpha_1 = .9$ the mean of the estimates by each procedure (about .84) is not significantly different from the parameter value, but the variance of the observed estimates was 2 or 3 times the asymptotic value (which is also the Cramér–Rao

[2] At a customary significance level one would not reject the hypothesis that the mean or standard deviation of the estimate is equal to that of the asymptotic theory. The investigator did not study how close the empirical distributions were to the normal.

lower bound). A conclusion is that for $\alpha_1$ as large as .9 the length of the series must be greater than 100 for the asymptotic theory to be a good approximation. The behavior of the two methods in the frequency domain stabilizes after 2 to 4 iterations, but the variance of the estimate based on scoring in the time domain does not stabilize by the tenth iteration. At $\alpha_1 = .9$ and $T = 100$ there is a nonnegligible probability that the estimates are such that $\hat{f}(\lambda)$ is not nonnegative for all $\lambda$ in $[-\pi \leqq \lambda \leqq \pi]$ or that $|\hat{\alpha}_1| < 1$. Thus there is some arbitrary treatment of estimates failing to meet these conditions; that treatment affects the reported statistical characteristics.

Nelson (1974) has also investigated the first-order pure moving average process for $\alpha_1 = -.9, -.5, -.2, 0, .2, .5$, and .9 on the basis of 500 replications for $T = 30$ and 200 replications for $T = 100$. In addition to studying the initial estimate (described in Section 4.4) and Durbin's estimate, he studied a method that is approximately the Newton-Raphson method in the time domain (of Section 3.2) and a method to minimize $\mathbf{y}'\mathbf{\Sigma}^{-1}\mathbf{y}$ due to Box and Jenkins. Iterations were carried to convergence (or stopped after 70 steps).

The means and variances of these two estimates at $T = 100$ agree with asymptotic theory except at $\alpha_1 = \pm.9$. In the latter cases the mean is significantly different from the parameter value, biased towards 0; the observed variances are reported so ambiguously that it is impossible to test the hypothesis that the observed variance agrees with the theoretical. (Reporting a variance of .002 only implies the observed variance is in the interval (.0015, .0025).) For $T = 30$ the conclusions roughly are that the means of the estimates are within sampling error of $\alpha_1$ except at $\alpha_1 = \pm.5$ and $\pm.9$, where the estimates seem biased towards 0, and each observed variance is significantly greater than the asymptotic value.

Kashyap and Nasburg (1974) used the Newton-Raphson methods in the time and frequency domains at $T = 100$ with 20 replications for the model with $\alpha_1 = .5$ and for the model with $\beta_1 = -.8$, $\alpha_1 = .5$. The means and variances are not significantly different from the parameter values and the asymptotic variances.

To summarize: the asymptotic theory seems adequate at $T = 100$ and $|\alpha_1|$ less than some value between .5 and .9; there is insufficient evidence to compare procedures for short series.

## 6. Mathematical details.

6.1. *Estimation of the coefficients.* The first and second partial derivatives of the logarithm of the likelihood function for the general model (1.6) were given in Section 5 of Anderson (1975a) for arbitrary matrices $\mathbf{J}_g$ and $\mathbf{K}_k$. Newton-Raphson procedures are developed in the time domain by substituting $\mathbf{J}_g = \mathbf{L}^g$ and $\mathbf{K}_k = \mathbf{L}^k$ and in the frequency domain by substituting $\mathbf{J}_g = \mathbf{M}^g$ and $\mathbf{K}_k = \mathbf{M}^k$, using the algebra of Section 2.2. (For $\mathbf{C} = \mathbf{yy}'$ one uses $\operatorname{tr} \mathbf{PCQ} = \operatorname{tr} \mathbf{Pyy}'\mathbf{Q} = \operatorname{tr} \mathbf{y}'\mathbf{QPy} = \mathbf{y}'\mathbf{QPy}$.) However, some terms are asymptotically negligible and are dropped to obtain the procedures presented in Section 3.

The Newton–Raphson equations obtained by setting the right-hand side of (2.29) equal to $\mathbf{0}$ and replacing $\mathbf{R}(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ by $\mathbf{0}$ can be rewritten with $\boldsymbol{\theta}^*$ replaced by $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_0$ replaced by $\hat{\boldsymbol{\theta}}_{i-1}$ as

$$(6.1)\qquad -\frac{1}{T}\frac{\partial^2 \log L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}_{i-1}} T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) = T^{-\frac{1}{2}}\frac{\partial \log L}{\partial\boldsymbol{\theta}}\bigg|_{\hat{\boldsymbol{\theta}}_{i-1}} - \frac{1}{T}\frac{\partial^2 \log L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}_{i-1}} T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta})\,.$$

We are interested in cases where $T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})$ has a limiting normal distribution, its coefficient in (6.1) is a consistent estimate of the average information matrix, and

$$(6.2)\qquad T^{-\frac{1}{2}}\frac{\partial \log L}{\partial\boldsymbol{\theta}}\bigg|_{\hat{\boldsymbol{\theta}}_{i-1}} - \frac{1}{T}\frac{\partial^2 \log L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}_{i-1}} T^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}) \simeq T^{-\frac{1}{2}}\frac{\partial \log L}{\partial\boldsymbol{\theta}}$$

(evaluated at the "true" parameter vector $\boldsymbol{\theta}$) has approximately a normal distribution. Thus we shall drop a term from $\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'|_{\hat{\boldsymbol{\theta}}_{i-1}}$ if such a term divided by $T$ converges to $\mathbf{0}$ in probability, and we shall drop a term from $\partial \log L/\partial\boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}_{i-1}}$ if such a term divided by $T^{\frac{1}{2}}$ converges to $\mathbf{0}$ in probability (as $T \to \infty$).

A term in $\partial^2 \log L/\partial\alpha_g\,\partial\alpha_f|_{\hat{\boldsymbol{\theta}}_{i-1}}$ based on use of $\mathbf{L}$ is $2T/\hat{\sigma}^2_{i-1}$ times

$$\frac{1}{T}\,\mathbf{y}'\hat{\mathbf{B}}'_{i-1}\hat{\mathbf{A}}'^{-1}_{i-1}\mathbf{L}^{g+f}\hat{\mathbf{A}}^{-3}_{i-1}\hat{\mathbf{B}}_{i-1}\mathbf{y}$$

$$(6.3)\qquad = \frac{1}{T}\sum\nolimits_{k,l=0}^{p}\hat{\beta}_k^{(i-1)}\hat{\beta}_l^{(i-1)} \times \sum\nolimits_{h,j,m,n=0}^{\infty}\hat{\delta}_h^{(i-1)}\hat{\delta}_j^{(i-1)}\hat{\delta}_m^{(i-1)}\hat{\delta}_n^{(i-1)}\mathbf{y}'\mathbf{L}'^{h+k}\mathbf{L}^{g+f+j+l+m+n}\mathbf{y}\,,$$

where $\hat{A}^{-1}_{i-1}(z) = \sum_{h=0}^{\infty}\hat{\delta}_h^{(i-1)}z^h$. (For simplicity of notation it is convenient to ignore the fact that $\mathbf{L}^m = \mathbf{0}$ for $m \geqq T$.) Since

$$(6.4)\qquad \left|\frac{1}{T}\,\mathbf{y}'\mathbf{L}'^h\mathbf{L}^m\mathbf{y}\right| \leqq \frac{1}{T}\,\mathbf{y}'\mathbf{y}$$

has expected value uniformly bounded and $|\hat{\delta}_m^{(i-1)}| < K(\hat{\rho}^{(i-1)})^m$ for some $0 < \hat{\rho}^{(i-1)} < 1$ with arbitrarily high probability for suitably large $T$, (6.3) differs by an arbitrarily small amount from

$$(6.5)\qquad \frac{1}{T}\sum\nolimits_{k,l=0}^{p}\hat{\beta}_k^{(i-1)}\hat{\beta}_l^{(i-1)} \times \sum\nolimits_{h,j,m,n=0}^{N}\hat{\delta}_h^{(i-1)}\hat{\delta}_j^{(i-1)}\hat{\delta}_m^{(i-1)}\hat{\delta}_n^{(i-1)}\mathbf{y}'\mathbf{L}'^{h+k}\mathbf{L}^{g+f+j+l+m+n}\mathbf{y}$$

for sufficiently large $N$, which in turn (by consistency of estimates) differs by an arbitrarily small amount from

$$(6.6)\qquad \frac{1}{T}\sum\nolimits_{k,l=0}^{p}\beta_k\beta_l\sum\nolimits_{h,j,m,n=0}^{N}\delta_h\delta_j\delta_m\delta_n\,\mathbf{y}'\mathbf{L}'^{h+k}\mathbf{L}^{g+f+j+l+m+n}\mathbf{y}\,,$$

which in turn is approximated by

$$(6.7)\qquad \frac{1}{T}\sum_{k,l=0}^{p}\beta_k\beta_l\sum_{h,m=0}^{\infty}\sum_{j,n=0}^{N}\delta_h\delta_j\delta_m\delta_n\mathbf{y}'\mathbf{L}'^{h+k}\mathbf{L}^{g+f+j+l+m+n}\mathbf{y}$$
$$= \frac{1}{T}\sum_{j,n=0}^{N}\delta_j\delta_n\mathbf{v}'\mathbf{L}^{g+f+j+n}\mathbf{v}\ .$$

Since

$$(6.8)\qquad \operatorname{p\,lim}_{T\to\infty}\frac{1}{T}\mathbf{v}'\mathbf{L}^{g+f+j+n}\mathbf{v} = \mathscr{E}v_t v_{t-g-f-j-n} = 0\,,$$

the probability limit of (6.7), and hence of (6.3) is 0. Other terms in second partial derivatives such that when divided by $T$ have probability limits of 0 are dropped; the terms are those that have explicitly powers of $\mathbf{L}$ but not also of $\mathbf{L}'$ or powers of $\mathbf{L}'$ but not also of $\mathbf{L}$. Thus (3.8) to (3.10) are obtained.

In the circular model the terms $-\operatorname{tr}\mathbf{A}^{-1}\mathbf{M}^g$ and $\operatorname{tr}\mathbf{B}^{-1}\mathbf{M}^k$ in the first derivatives are not automatically 0, but are asymptotically 0. One such term, for example, in $\partial\log L/\partial\alpha_g|_{\theta=\hat{\theta}_{i-1}}/T^{\frac{1}{2}}$ is $-1/T^{\frac{1}{2}}$ times

$$(6.9)\qquad \operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{M}^g = \sum_{t=1}^{T}\frac{e^{-i\lambda_t g}}{\hat{A}_{i-1}(e^{-i\lambda_t})}$$
$$= \sum_{t=1}^{T}\sum_{j=0}^{\infty}\hat{\delta}_j^{(i-1)}e^{-i\lambda_t(j+g)} = T\sum_{k=1}^{\infty}\hat{\delta}_{kT-g}^{(i-1)}$$

because $\sum_{t=1}^{T}e^{i\lambda_t tf} = T$ if $f$ is an integral multiple of $T$ and 0 otherwise. If the zeros of $\hat{A}_{i-1}(z)$ are different, $|\hat{\delta}_j^{(i-1)}| \leqq \hat{K}_{i-1}\hat{\rho}_{i-1}^j$, where $\hat{K}_{i-1}$ is a positive quantity and $\hat{\rho}_{i-1}$ is the reciprocal of the smallest absolute value of the zeros. Because $\hat{\boldsymbol{\alpha}}_{i-1}$ is a consistent estimate of $\boldsymbol{\alpha}$, for sufficiently large $T$ the probability is arbitrarily great that $\hat{K}_{i-1}$ is not greater than some positive constant $K$ and $\hat{\rho}_{i-1} \leqq \rho$, where $\rho$ is greater than the reciprocal of the smallest absolute value of the zeros of $A(z)$ and less than 1. Then for sufficiently large $T$ and arbitrarily high probability (6.9) in absolute value is less than or equal to

$$(6.10)\qquad T\sum_{k=1}^{\infty}K\rho^{kT-g} = TK\rho^{T-g}\sum_{l=0}^{\infty}(\rho^T)^l = \frac{TK\rho^{T-g}}{1-\rho^T}\,,$$

which converges to 0 as $T\to\infty$.

Terms of $\partial^2\log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'/T$ which converge to 0 in the direct lag model also do so for the circular model. In (6.3) to (6.7), for instance, $\mathbf{L}$ is replaced by $\mathbf{M}$. However, in (6.8) if $g+f+j+n = kT$ for some integer $k$ the expectation is $\sigma^2$, but the sum of the coefficients of such a term is in absolute value

$$(6.11)\qquad \left|\sum_{k=1}^{\infty}\sum_{j=0}^{kT-g-f}\delta_{kT-j-g-f}\delta_j\right|$$
$$\leqq \sum_{k=1}^{\infty}\sum_{j=0}^{kT-f-g}K\rho^{kT-g-f} = K\rho^{T-g-f}\sum_{l=0}^{\infty}(lT+T-g-f)\rho^{lT}\,,$$

which converges to 0 as $T\to\infty$. Also one can drop such asymptotically negligible terms such as

$$(6.12)\qquad \operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{M}^g\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{M}^f = \sum_{t=1}^{T}\frac{e^{-i\lambda_t(g+f)}}{\hat{A}_{i-1}^2(e^{-i\lambda_t})}$$
$$= \sum_{t=1}^{T}\sum_{h,j=0}^{\infty}\hat{\delta}_h^{(i-1)}\hat{\delta}_j^{(i-1)}e^{-i\lambda_t(h+j+g+f)}\,,$$

which converges to 0 in probability by an argument similar to those used above. Thus (3.15) to (3.20) are obtained (by use of the algebra of Section 2.2).

To develop the scoring equations components of $-\mathscr{E}\,\partial^2 \log L/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ are needed. These have been given for the general model (1.6) in Section 5 of Anderson (1975a), and the scoring procedure for the lag operator model was presented. No terms needed to be dropped.

For the circular model $\mathbf{J}_g$ is replaced by $\mathbf{M}^g$ and $\mathbf{K}_k$ by $\mathbf{M}^k$. Again terms are dropped which are asymptotically negligible. More details of these developments are presented in Section 8 and Appendix A of Anderson (1975b).

6.2. *Estimation of the covariances and coefficients*. The first and second partial derivatives of the logarithm of the likelihood function for the general model (1.7) and (1.8) were displayed in Section 4 of Anderson (1975a). To develop Newton–Raphson procedures in the time domain $\mathbf{G}_g$ is defined by $\mathbf{L}^g + \mathbf{L}'^g$ and $\mathbf{K}_k = \mathbf{L}^k$; the asymptotically negligible term $\operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\mathbf{L}^k\hat{\mathbf{B}}_{i-1}^{-1}\mathbf{L}^l$ is dropped to obtain (4.5) to (4.9). In the frequency domain $\mathbf{G}_g = \mathbf{M}^g + \mathbf{M}^{-g} = \mathbf{U}\boldsymbol{\Gamma}_g\bar{\mathbf{U}}'$ and $\mathbf{K}_k = \mathbf{M}^k = \mathbf{U}\mathbf{D}^k\bar{\mathbf{U}}'$. Dropping asymptotically negligible terms yields (4.13) to (4.17).

In Section 4 of Anderson (1975a) were given the scoring equations for the general model and for the model in the time domain. The term $\operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\hat{\boldsymbol{\Sigma}}_{i-1}^u\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^k(\hat{\boldsymbol{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^l$ has been replaced by (3.14) here which is asymptotically equivalent. More precisely

$$\operatorname{p\,lim}_{T\to\infty} \frac{1}{T}\,[\operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\hat{\boldsymbol{\Sigma}}_{i-1}^u\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^k(\hat{\boldsymbol{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^l - \operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^k\mathbf{L}^l] = 0\,. \tag{6.13}$$

The argument is that if $\sum_{g=-q}^{q} \hat{\sigma}_g^{(i-1)}z^g \equiv A^*(z)A^*(1/z)$ then $\hat{\boldsymbol{\Sigma}}_{i-1}^u$ is approximated by $A^*(\mathbf{L})A^*(\mathbf{L}')$ and $(\hat{\boldsymbol{\Sigma}}_{i-1}^u)^{-1}$ is approximated by $A^*(\mathbf{L}')^{-1}A^*(\mathbf{L})^{-1}$.

To obtain the scoring equations in the frequency domain $\mathbf{G}_g = \mathbf{M}^g + \mathbf{M}^{-g}$ and $\mathbf{K}_k = \mathbf{M}^k$; asymptotically negligible terms are dropped. More details are given in Section 9 and Appendix A of Anderson (1975b).

## APPENDIX

*Derivation of Hannan's estimates in the notation of this paper.* Hannan (1970) specified an estimation procedure in Subsection (c) of Section 5 of Chapter VI (and in (1969)). The initial estimates $\hat{\boldsymbol{\beta}}_0$ and $\hat{f}_0^u(\lambda) = \hat{\sigma}_0^2|\hat{A}_0(e^{i\lambda})|^2/(2\pi)$ are those described in Section 4.4. Then (3.27) or (4.23) for $i = 0$ is solved for $\hat{\boldsymbol{\beta}}$. Next solve for $\mathbf{t}$

$$\hat{\boldsymbol{\Phi}}_0\mathbf{t} = -\mathbf{h}\,, \tag{A.1}$$

where

$$[\mathbf{h}]_g = \sum_{t=1}^{T} \frac{I(\lambda_t)\cos\lambda_t g}{\hat{f}(\lambda_t)|\hat{A}_0(e^{i\lambda_t})|^2}\,, \qquad g = 1, \cdots, q\,, \tag{A.2}$$

and $\hat{\boldsymbol{\Phi}}_0$ and $\hat{f}(\lambda_t)$ are based on $\hat{f}_0^u(\lambda)$ and $\hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{\Omega}}_0$ and $\hat{\boldsymbol{\Psi}}_0$ be defined as the coefficients in Section 3.3 for the Newton–Raphson method in the circular model. Then Hannan proposed the estimate

$$
\begin{aligned}
\hat{\boldsymbol{\alpha}}_1 &= [\mathbf{I} - \hat{\boldsymbol{\Phi}}_0^{-1}\hat{\boldsymbol{\Omega}}_0\hat{\boldsymbol{\Psi}}_0^{-1}\hat{\boldsymbol{\Omega}}_0']^{-1}(\hat{\boldsymbol{\alpha}}_0 - \mathbf{t}) + \hat{\boldsymbol{\alpha}}_0 \\
&= [\hat{\boldsymbol{\Phi}}_0 - \hat{\boldsymbol{\Omega}}_0\hat{\boldsymbol{\Psi}}_0^{-1}\hat{\boldsymbol{\Omega}}_0']^{-1}\hat{\boldsymbol{\Phi}}_0(\hat{\boldsymbol{\alpha}}_0 - \mathbf{t}) + \hat{\boldsymbol{\alpha}}_0 \\
&= [\hat{\boldsymbol{\Phi}}_0 - \hat{\boldsymbol{\Omega}}_0\hat{\boldsymbol{\Psi}}_0^{-1}\hat{\boldsymbol{\Omega}}_0']^{-1}[\hat{\boldsymbol{\Phi}}_0\hat{\boldsymbol{\alpha}}_0 + \mathbf{h}] + \hat{\boldsymbol{\alpha}}_0
\end{aligned}
\tag{A.3}
$$

or

$$
[\hat{\boldsymbol{\Phi}}_0 - \hat{\boldsymbol{\Omega}}_0\hat{\boldsymbol{\Psi}}_0^{-1}\hat{\boldsymbol{\Omega}}_0'](\hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_0) = \hat{\mathbf{q}}_0 \, . \tag{A.4}
$$

This is (3.25) for $i = 1$ and $\hat{\mathbf{p}}_0 = \mathbf{0}$, which is the case because the estimates of $\beta_1, \cdots, \beta_p$ satisfy (3.27) or (4.23). With coefficients of (3.27) or (4.23) calculated on the basis of $\hat{\boldsymbol{\alpha}}_1$ another solution for an estimate of $\boldsymbol{\beta}$ can be computed. Note that $\hat{f}_{i-1}(\lambda_t)|\hat{B}_{i-1}(e^{i\lambda_t})|^2 = \hat{\sigma}_{i-1}^2|\hat{A}_{i-1}(e^{i\lambda_t})|^2/(2\pi)$. The estimates at the end of the first stage may be used to iterate to a second stage. (In Hannan (1969) $\hat{\boldsymbol{\Omega}}_0$ is based on Newton–Raphson, and in Hannan (1970) on scoring.)

## REFERENCES

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60** 255–265.

Akaike, H., Arahata, E. and Ozaki, T. (1975). TIMSAC-74. A time series analysis and control program package (1). *Computer Science Monographs* No. 5, The Institute of Statistical Mathematics, Tokyo.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. *Multivariate Analysis-II* (P. R. Krishnaiah, ed.) 55–66. Academic Press, New York.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, eds.) 1–24. Univ. of North Carolina Press.

Anderson, T. W. (1971 a). *The Statistical Analysis of Time Series.* Wiley, New York.

Anderson, T. W. (1971 b). Estimation of covariance matrices with linear structure and moving average processes of finite order. Technical Report No. 6, Contract N00014-67-A-0112-0030, Stanford Univ.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141.

Anderson, T. W. (1975 a). Maximum likelihood estimation of parameters of autoregressive processes with moving average residuals and other covariance matrices with linear structure. *Ann. Statist.* **3** 1283–1304.

Anderson, T. W. (1975 b). Estimation by maximum likelihood in autoregressive moving average models in the time and frequency domains. Technical Report No. 20, Contract N00014-75-C-0442, Stanford Univ.

Anderson, T. W. (1977). On maximum likelihood estimation of parameters of autoregressive moving average processes. *Transactions of the Seventh Prague Conference and 1974 EMS*, in press.

Åström, Karl-Johann and Bohlin, Torsten (1966). Numerical identification of linear dynamic systems from normal operating records. *Theory of Self Adaptive Control Systems* (P. H. Hammond, ed.) 96–111. Plenum Press, New York.

Box, George E. P. and Jenkins, Gwilym M. (1970). *Time Series Analysis Forecasting and Control.* Holden-Day, San Francisco.

CLEVENSON, M. LAWRENCE (1970). Asymptotically efficient estimates of the parameters of a moving average time series. Technical Report No. 15, Contract NONR-225 (80), Stanford Univ.

DWYER, P. S. (1967). Some application of matrix derivatives in multivariate analysis. *J. Amer. Statist. Assoc.* **62** 607–625.

DZHAPARIDZE, K. O. and YAGLOM, A. M. (1974). Application of the modified Fisher's "method of scoring" to spectrum parameter estimation for stochastic processes. *Doklady Akad. Nauk SSSR* **217** 512–515.

GUPTA, N. K. and MEHRA, R. K. (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity calculations. *IEEE Transactions on Automatic Control* **AC-19** 774–783.

HANNAN, E. J. (1960). *Time Series Analysis*. Methuen, London, and Wiley, New York.

HANNAN, E. J. (1969). The estimation of mixed moving average autoregressive systems. *Biometrika* **56** 579–594.

HANNAN, E. J. (1970). *Multiple Time Series*. Wiley, New York.

KASHYAP, R. L. and NASBURG, ROBERT E. (1974). Parameter estimation in multivariate stochastic difference equations. *IEEE Transactions on Automatic Control* **AC-19** 784–797.

MANN, H. B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* **11** 173–220.

MCCLAVE, JAMES T. (1974). A comparison of moving average estimation procedures. *Comm. Statist.* **3** 865–883.

NELSON, CHARLES R. (1974). The first-order moving average process. *J. Econometrics* **2** 121–141.

PARZEN, EMANUEL (1971). Efficient estimation of stationary time series mixed schemes. *Bull. Inst. Internat. Statist.* **44** 315–319.

TONG, H. (1975). Autoregressive model fitting with noisy data by Akaike's information criterion. *IEEE Transactions on Information Theory* **IT-21** 476–480.

WAHBA, GRACE (1968). On the distribution of some statistics useful in the analysis of jointly stationary time series. *Ann. Math. Statist.* **39** 1849–1862.

WALKER, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *J. Australian Math. Soc.* **4** 363–384.

WHITTLE, P. (1953). Estimation and information in stationary time series. *Ark. Mat.* **2** 423–434.

WILSON, G. J. (1969). Factorization of the generating function of a pure moving average process. *SIAM J. Numerical Analysis* **6** 1–7.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305