Exact Maximum Likelihood Estimation of Stationary Vector ARMA Models

José Alberto MAURICIO*

The problems of evaluating and subsequently maximizing the exact likelihood function of vector autoregressive moving average (ARMA) models are considered separately. A new and efficient procedure for evaluating the exact likelihood function is presented. This method puts together a set of useful features that can only be found separately in currently available algorithms. A procedure for maximizing the exact likelihood function, which takes full advantage of the properties offered by the evaluation algorithm, is also considered. Combining these two procedures, a new algorithm for exact maximum likelihood estimation of vector ARMA models is obtained. Comparisons with existing procedures, in terms of both analytical arguments and a numerical example, are given to show that the new estimation algorithm performs at least as well as existing ones, and that relevant real situations occur in which it does better.

KEY WORDS: Cholesky decomposition; Invertibility; Multiple autoregressive moving average model; Quasi-Newton method; Residuals; Stationarity.

1. INTRODUCTION

This article proposes a new procedure for exact maximum likelihood estimation of vector autoregressive moving average (ARMA) models. A sharp distinction between evaluating and maximizing the likelihood function is made. This permits a detailed analysis of all problems that arise in the estimation process. The solutions obtained through this analysis can then be integrated into a complete estimation procedure that takes advantage of some properties of the likelihood function that have not been fully exploited in previous papers.

Although there has been abundant research on evaluating the likelihood function of vector ARMA models (see, for example, Hall and Nicholls 1980; Hillmer and Tiao 1979; Nicholls and Hall 1979; and Shea 1987), only a few authors have paid attention to the problem of its subsequent maximization. Furthermore, such attention has usually been restricted to suggesting, in a few lines, the use of a standard optimization algorithm to maximize the likelihood function, evaluated as extensively described in the just-cited papers. An interesting exception was provided by Shea (1984, pp. 99–100).

With regard to the computation of the likelihood function, none of the existing methods can be considered fully satisfactory. This is due to the fact that many of the necessary properties of a method for evaluating the likelihood function are scattered among the existing procedures. Thus, although each of many existing algorithms has some useful properties, it also lacks other properties that can be found in alternative procedures.

For instance, the algorithm of Shea (1984, 1987) can be considered, from a computational viewpoint, to be the most efficient of the existing procedures. But its use does not permit the automatic detection of noninvertible models. This task is easily handled with the algorithm of Hall and Nicholls (1980) and with an extension to the multivariate context of the algorithm of Ljung and Box (1979). But the former involves a high computational cost (in many cases), whereas the latter involves both some computational inefficiency and a loss of numerical precision, due to the requirement for an explicit matrix inversion. Finally, the algorithm of Hillmer and Tiao (1979) does not permit the computation of either the exact likelihood function or an appropriate residual vector, except in the case of pure moving average (MA) models; this fact may become an important drawback when the model considered has an autoregressive (AR) part and the sample contains extreme values among the initial observations.

A thorough analysis of currently available procedures allows one to discover and fully exploit new possibilities ignored in the previously cited papers. Thus, in Section 2 a new method of evaluating the exact likelihood function of vector ARMA models is described in detail. The new algorithm puts together the advantages that can be found separately in existing procedures and does not suffer from any of their drawbacks. In Section 3 computational techniques for maximizing the likelihood function are considered. These techniques take full advantage of the properties offered by the evaluation algorithm. An illustrative example of an actual situation in which the new estimation procedure performs better than one of the most frequently used procedures is given in Section 4. Finally, in Section 5 conclusions are summarized.

2. EVALUATION OF THE EXACT LIKELIHOOD FUNCTION

Let $\mathbf{w}_t$ be an m-dimensional vector-valued time series. It is assumed that $\mathbf{w}_t$ follows the vector ARMA(p, q) model

* José Alberto Mauricio is Professor A.E.U., Departamento de Fundamentos del Análisis Económico II (Economía Cuantitativa), Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, 28223, Spain. This article summarizes the author's doctoral dissertation *Evaluación y Maximización de la Función de Verosimilitud de Procesos ARMA Multivariantes* (1993, Madrid: Universidad Complutense de Madrid), and was partially financed under the project for a *Forecast and Monitor Service for the Spanish Economy*, sponsored by Caja de Madrid (Spain), and by the DGICYT Project PB89-0129 (Spain). The author is grateful to Arthur B. Treadway, under whose supervision this research was carried out, and to Emilio Cerdá, José Luis Gallego, Víctor M. Gonzalo, Miguel Jerez, Alfonso Novales, Daniel Peña, Fernando Tussel, two anonymous referees, and an associate editor for helpful suggestions.

$$\Phi(B)\tilde{\mathbf{w}}_t = \Theta(B)\mathbf{a}_t, \tag{1}$$

where $\Phi(B) = \mathbf{I} - \Phi_1 B - \cdots - \Phi_p B^p$, $\Theta(B) = \mathbf{I} - \Theta_1 B - \cdots - \Theta_q B^q$; B is the back shift operator; $\tilde{\mathbf{w}}_t = \mathbf{w}_t - \mu$, Φ_i $(i = 1, \ldots, p)$, Θ_i $(i = 1, \ldots, q)$, and μ are $m \times m$, $m \times m$, and $m \times 1$ parameter matrices; and the $\mathbf{a}_t$'s are $m \times 1$ random vectors identically and independently distributed as $N(\mathbf{0}, \sigma^2\mathbf{Q})$, with $\sigma^2 > 0$ and $\mathbf{Q}(m \times m)$ symmetric and positive definite. This decomposition of $E[\mathbf{a}_t\mathbf{a}_t^T]$, although not unique, is useful for obtaining maximum likelihood estimates by maximizing a concentrated log-likelihood as a function of Φ_i $(i = 1, \ldots, p)$, Θ_i $(i = 1, \ldots, q)$, and $\mathbf{Q}$ only (Sec. 3). For stationarity, it is required that the zeros of $|\Phi(B)|$ lie outside the unit circle. Furthermore, (1) is assumed to satisfy the conditions derived by Hannan (1969) for the model to be identified.

Consider a sample of size n and let $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_1^T, \ldots, \tilde{\mathbf{w}}_n^T)^T$ (mean-corrected observations), $\mathbf{a} = (\mathbf{a}_1^T, \ldots, \mathbf{a}_n^T)^T$ (white noise perturbations), and $\mathbf{u}_* = (\tilde{\mathbf{w}}_{1-p}^T, \ldots, \tilde{\mathbf{w}}_0^T, \mathbf{a}_{1-q}^T, \ldots, \mathbf{a}_0^T)^T$ (unknown presample values). Then (1) may be written as

$$\mathbf{D}_{\Phi,n}\tilde{\mathbf{w}} = \mathbf{D}_{\Theta,n}\mathbf{a} + \mathbf{V}\mathbf{u}_*, \tag{2}$$

where $\mathbf{D}_{\Phi,n}$ and $\mathbf{D}_{\Theta,n}$ are $nm \times nm$ block matrices with identity matrices on the main diagonal and $-\Phi_k$ and $-\Theta_k$ down the kth subdiagonal. Further, $\mathbf{V}$ is the $nm \times (p+q)m$ block matrix $\mathbf{V} = (\mathbf{G}_{\Phi,n}, \mathbf{G}_{\Theta,n})$, where $\mathbf{G}_{\Phi,n}$ and $\mathbf{G}_{\Theta,n}$ are the following $nm \times pm$ and $nm \times qm$ block matrices:

$$\mathbf{G}_{\Phi,n} = \begin{bmatrix} \Phi_p & \Phi_{p-1} & \cdots & \Phi_1 \\ \mathbf{0} & \Phi_p & \cdots & \Phi_2 \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Phi_p \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix};$$

$$\mathbf{G}_{\Theta,n} = \begin{bmatrix} -\Theta_q & -\Theta_{q-1} & \cdots & -\Theta_1 \\ \mathbf{0} & -\Theta_q & \cdots & -\Theta_2 \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & -\Theta_q \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}.$$

On the basis of the previous definitions, Nicholls and Hall (1979) have shown that the exact likelihood function of the parameters $\Phi = (\Phi_1, \ldots, \Phi_p)$, $\Theta = (\Theta_1, \ldots, \Theta_q)$, μ, σ^2, and $\mathbf{Q}$ is given by

$$L(\Phi, \Theta, \mu, \sigma^2, \mathbf{Q}|\mathbf{w}) = (2\pi\sigma^2)^{-(nm/2)}|\mathbf{Q}|^{-(n/2)} \times |\mathbf{\Lambda}^T\mathbf{\Lambda}|^{-(1/2)}\exp\left(-\frac{1}{2\sigma^2}S(\Phi, \Theta, \mu, \mathbf{Q}|\mathbf{w})\right). \tag{3}$$

The quadratic form in the exponential is given by $S(\Phi, \Theta, \mu, \mathbf{Q}|\mathbf{w}) = (\mathbf{\Upsilon}\tilde{\mathbf{w}} + \mathbf{\Lambda}\hat{\mathbf{e}}_*)^T(\mathbf{\Upsilon}\tilde{\mathbf{w}} + \mathbf{\Lambda}\hat{\mathbf{e}}_*)$, where $\mathbf{\Upsilon}$ and $\mathbf{\Lambda}$ are the following $(p+q+n)m \times nm$ and $(p+q+n)m \times (p+q)m$ matrices:

$$\mathbf{\Upsilon} = \begin{bmatrix} \mathbf{0} \\ (\mathbf{I} \otimes \mathbf{R})\mathbf{K} \end{bmatrix}, \qquad \mathbf{\Lambda} = \begin{bmatrix} \mathbf{I} \\ (\mathbf{I} \otimes \mathbf{R})\mathbf{Z}\mathbf{T}^{-1} \end{bmatrix}, \tag{4}$$

and

$$\hat{\mathbf{e}}_* = \mathbf{T}\hat{\mathbf{u}}_* = -(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{\Upsilon}\tilde{\mathbf{w}}. \tag{5}$$

In (4), the $nm \times nm$ matrix $\mathbf{K}$ is given by $\mathbf{K} = \mathbf{D}_{\Theta,n}^{-1}\mathbf{D}_{\Phi,n}$, the $nm \times (p+q)m$ matrix $\mathbf{Z}$ is given by $\mathbf{Z} = -\mathbf{D}_{\Theta,n}^{-1}\mathbf{V}$, and, if $E[\mathbf{a}_t\mathbf{a}_t^T] = \sigma^2\mathbf{Q}$ and $E[\mathbf{u}_*\mathbf{u}_*^T] = \sigma^2\mathbf{\Omega}$, then the $m \times m$ and $(p+q)m \times (p+q)m$ matrices $\mathbf{R}$ and $\mathbf{T}$ are such that $\mathbf{R}\mathbf{Q}\mathbf{R}^T = \mathbf{I}$ (i.e., $\mathbf{Q}^{-1} = \mathbf{R}^T\mathbf{R}$) and $\mathbf{T}\mathbf{\Omega}\mathbf{T}^T = \mathbf{I}$ (i.e., $\mathbf{\Omega}^{-1} = \mathbf{T}^T\mathbf{T}$). Further, the matrix $\mathbf{\Omega}$ can be partitioned as follows:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}.$$

The (i, j)th block of $\mathbf{A}$ is given by $\mathbf{A}_{ij} = \sigma^{-2}E[\tilde{\mathbf{w}}_{i-p}\tilde{\mathbf{w}}_{j-p}^T] = \mathbf{\Gamma}(j-i)$ $(i, j = 1, \ldots, p)$, and the (i, j)th block of matrix $\mathbf{B}$ is $\mathbf{B}_{ij} = \sigma^{-2}E[\tilde{\mathbf{w}}_{i-p}\mathbf{a}_{j-q}^T] = \mathbf{\Gamma}_{\mathbf{wa}}(j-i-q+p)$ $(i = 1, \ldots, p; j = 1, \ldots, q)$. Because $\mathbf{A}$ is symmetric and $E[\tilde{\mathbf{w}}_{t-i}\mathbf{a}_t^T] = \mathbf{0}$ for $i > 0$, to compute $\mathbf{A}$ and $\mathbf{B}$ only the theoretical autocovariance and cross-covariance matrices $\mathbf{\Gamma}(k)$, $k = 0, \ldots, p-1$, and $\mathbf{\Gamma}_{\mathbf{wa}}(k)$, $k = -q+1, \ldots, 0$, are needed. Finally, $\mathbf{C}$ is a block-diagonal $qm \times qm$ matrix with $\mathbf{Q}$'s along the main diagonal.

Thus, to evaluate (3), one must compute the determinant $|\mathbf{\Lambda}^T\mathbf{\Lambda}|$ and the quadratic form $S(\Phi, \Theta, \mu, \mathbf{Q}|\mathbf{w})$. Hall and Nicholls (1980) have suggested computing the latter as follows:

$$S(\Phi, \Theta, \mu, \mathbf{Q}|\mathbf{w}) = (\mathbf{\Upsilon}\tilde{\mathbf{w}})^T[\mathbf{I} - \mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T](\mathbf{\Upsilon}\tilde{\mathbf{w}}), \tag{6}$$

which may be regarded as the residual sum of squares of the regression of $(\mathbf{\Upsilon}\tilde{\mathbf{w}})$ on $\mathbf{\Lambda}$. To evaluate $(\mathbf{\Upsilon}\tilde{\mathbf{w}})$ and $\mathbf{\Lambda}$, they computed $(\mathbf{I} \otimes \mathbf{R})\mathbf{K}\tilde{\mathbf{w}}$ and $(\mathbf{I} \otimes \mathbf{R})\mathbf{Z}$ [see (4)] recursively.

It is shown next, by exploring in further detail the elements of (3), how to compute the determinant $|\mathbf{\Lambda}^T\mathbf{\Lambda}|$ and the quadratic form (6) in a computationally more efficient manner. It is also shown (1) how to compute an approximation to the exact likelihood function to any desired degree of accuracy, (2) how to detect noninvertible and/or nonstationary models, and (3) how to calculate the residual vector for a given set of data and parameter values.

2.1 A New Method of Computing the Exact Likelihood Function

First, it may be noted from (4) that

$$\mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T = \begin{bmatrix} (\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1} & (\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{T}^{-1T}\mathbf{Z}^T(\mathbf{I} \otimes \mathbf{R}^T) \\ (\mathbf{I} \otimes \mathbf{R})\mathbf{Z}\mathbf{T}^{-1}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1} & (\mathbf{I} \otimes \mathbf{R})\mathbf{Z}\mathbf{T}^{-1}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{T}^{-1T}\mathbf{Z}^T(\mathbf{I} \otimes \mathbf{R}^T) \end{bmatrix},$$

and, because $\mathbf{\Upsilon}\tilde{\mathbf{w}} = (\mathbf{0}^T, \eta^T)^T$, where

$$\eta = (\mathbf{I} \otimes \mathbf{R})\hat{\mathbf{a}}_0 \tag{7}$$

and

$$\hat{\mathbf{a}}_0 = E[\mathbf{a}|\mathbf{w}, \mathbf{u}_* = \mathbf{0}] = \mathbf{K}\tilde{\mathbf{w}}, \tag{8}$$

the quadratic form (6) can be written as

$$S(\Phi, \Theta, \mu, \mathbf{Q}|\mathbf{w}) = \eta^T\eta - \eta^T(\mathbf{I} \otimes \mathbf{R}) \times \mathbf{Z}\mathbf{T}^{-1}(\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{T}^{-1T}\mathbf{Z}^T(\mathbf{I} \otimes \mathbf{R}^T)\eta. \tag{9}$$

Also from (4), it may be noted that $\mathbf{\Lambda}^T\mathbf{\Lambda} = \mathbf{I} + \mathbf{T}^{-1T}\mathbf{X}\mathbf{T}^{-1}$, where the $(p+q)m \times (p+q)m$ matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{V}_1^T\mathbf{H}^T\mathbf{H}\mathbf{V}_1 \tag{10}$$

and, if $g = \max(p, q)$, then the $gm \times (p+q)m$ matrix $\mathbf{V}_1$ consists of the first gm rows of $\mathbf{V}$ and the $nm \times gm$ matrix $\mathbf{H}$ consists of the first gm columns of $(\mathbf{I} \otimes \mathbf{R})\mathbf{D}_{\boldsymbol{\theta},n}^{-1}$. Thus equation (9) can be rewritten as

$$S(\mathbf{\Phi}, \boldsymbol{\Theta}, \mu, \mathbf{Q}|\mathbf{w}) = \boldsymbol{\eta}^T\boldsymbol{\eta} - \boldsymbol{\eta}^T(\mathbf{I} \otimes \mathbf{R}) \times \mathbf{Z}(\mathbf{\Omega}^{-1} + \mathbf{V}_1^T\mathbf{H}^T\mathbf{H}\mathbf{V}_1)^{-1}\mathbf{Z}^T(\mathbf{I} \otimes \mathbf{R}^T)\boldsymbol{\eta}. \tag{11}$$

Now define the following $nm \times 1$ vector:

$$\mathbf{h} = \mathbf{D}_{\boldsymbol{\theta},n}^{-1T}(\mathbf{I} \otimes \mathbf{R}^T)\boldsymbol{\eta}, \tag{12}$$

and let $\tilde{\mathbf{h}}$ contain the first gm elements of $\mathbf{h}$. Then equation (11) can be rewritten as

$$S(\mathbf{\Phi}, \boldsymbol{\Theta}, \mu, \mathbf{Q}|\mathbf{w}) = \boldsymbol{\eta}^T\boldsymbol{\eta} - \tilde{\mathbf{h}}^T\mathbf{V}_1(\mathbf{\Omega}^{-1} + \mathbf{V}_1^T\mathbf{H}^T\mathbf{H}\mathbf{V}_1)^{-1}\mathbf{V}_1^T\tilde{\mathbf{h}},$$

where it may be verified that $\mathbf{V}_1(\mathbf{\Omega}^{-1} + \mathbf{V}_1^T\mathbf{H}^T\mathbf{H}\mathbf{V}_1)^{-1}\mathbf{V}_1^T = [(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)^{-1} + \mathbf{H}^T\mathbf{H}]^{-1}$. Thus the quadratic form (11) is

$$S(\mathbf{\Phi}, \boldsymbol{\Theta}, \mu, \mathbf{Q}|\mathbf{w}) = \boldsymbol{\eta}^T\boldsymbol{\eta} - \tilde{\mathbf{h}}^T[(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)^{-1} + \mathbf{H}^T\mathbf{H}]^{-1}\tilde{\mathbf{h}}. \tag{13}$$

Then, noting from (4) that $\mathbf{\Lambda}^T\mathbf{\Lambda} = \mathbf{I} + \mathbf{T}^{-1T}\mathbf{X}\mathbf{T}^{-1}$, it is clear that $\mathbf{T}^T\mathbf{\Lambda}^T\mathbf{\Lambda}\mathbf{T} = \mathbf{\Omega}^{-1} + \mathbf{X}$, so [see (10)] the determinant $|\mathbf{\Lambda}^T\mathbf{\Lambda}|$ can be calculated as

$$|\mathbf{\Lambda}^T\mathbf{\Lambda}| = |\mathbf{\Omega}|\,|\mathbf{\Omega}^{-1} + \mathbf{V}_1^T\mathbf{H}^T\mathbf{H}\mathbf{V}_1| = |\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T|\,|(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)^{-1} + \mathbf{H}^T\mathbf{H}|, \tag{14}$$

which is readily available as a by-product of the evaluation of the second term on the right side of (13). Note that when $m = 1$ (i.e., when dealing with univariate models), expressions (11) and (14) reduce to equations (2.6) and (2.4) in the work of Ljung and Box (1979). Also, note that evaluation of these expressions requires the explicit inversion of the matrix $\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T$, which results in a loss of computational efficiency and occasionally in a loss of numerical precision as well. This matrix inversion can be avoided as follows.

Let $\mathbf{M}$ denote the Cholesky factor of $\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T$, so that $\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T = \mathbf{M}\mathbf{M}^T$ and $\mathbf{M}^T(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)^{-1}\mathbf{M} = \mathbf{I}$. Then $[(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)^{-1} + \mathbf{H}^T\mathbf{H}]^{-1} = \mathbf{M}(\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})^{-1}\mathbf{M}^T$, so that the quadratic form (13) can be finally expressed as

$$S(\mathbf{\Phi}, \boldsymbol{\Theta}, \mu, \mathbf{Q}|\mathbf{w}) = \boldsymbol{\eta}^T\boldsymbol{\eta} - (\mathbf{M}^T\tilde{\mathbf{h}})^T(\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})^{-1}(\mathbf{M}^T\tilde{\mathbf{h}}). \tag{15}$$

Computation of the second term on the right side of (15) gives as a by-product the components of the determinant (14), which can be written in the form

$$|\mathbf{\Lambda}^T\mathbf{\Lambda}| = |\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}|. \tag{16}$$

The exact likelihood function (3) is then computed using (15) and (16). To evaluate these expressions, one needs (1) the $gm \times gm$ symmetric matrix $\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T$, (2) the $gm \times 1$ vector $\tilde{\mathbf{h}}$, (3) the $gm \times gm$ symmetric matrix $\mathbf{H}^T\mathbf{H}$, and (4) the $nm \times 1$ vector $\boldsymbol{\eta}$.

Once the components of $\mathbf{\Omega}$ are available (Hall and Nicholls 1980, pp. 254–256; Kohn and Ansley 1982), the (i, j)th block ($i = 1, \ldots, g; j = 1, \ldots, i$) of $\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T$ is given by

$$(\mathbf{V}_1\mathbf{\Omega}\mathbf{V}_1^T)_{ij} = \sum_{k=0}^{p-i} \mathbf{\Phi}_{p-k}\mathbf{E}_{k+i,j} - \sum_{k=0}^{q-i} \boldsymbol{\Theta}_{q-k}\mathbf{E}_{k+p+i,j}, \tag{17}$$

where, for $j = 1, \ldots, g$,

$$\mathbf{E}_{ij} = \sum_{k=j-i}^{p-i} \mathbf{\Gamma}(k)\mathbf{\Phi}_{p-k-i+j}^T - \sum_{k=j-i}^{q-i} \mathbf{\Gamma}_{\mathbf{wa}}(-q+p+k)\boldsymbol{\Theta}_{q-k-i+j}^T \qquad (i = 1, \ldots, p)$$

and

$$\mathbf{E}_{ij} = \sum_{k=p+j-i}^{2p-i} \mathbf{\Gamma}_{\mathbf{wa}}(-q+p-k)^T\mathbf{\Phi}_{2p-k-i+j}^T - \mathbf{Q}\boldsymbol{\Theta}_{q+p-i+j}^T \qquad (i = p+1, \ldots, p+q),$$

with $\mathbf{\Gamma}(k) = \mathbf{\Gamma}(-k)^T$ for $k < 0$, $\mathbf{\Gamma}_{\mathbf{wa}}(k) = \mathbf{0}$ for $k > 0$, and $\boldsymbol{\Theta}_i = \mathbf{0}$ for $i > q$. Now, because $\tilde{\mathbf{h}}$ contains the first gm components of $\mathbf{h}$ [see (12)] and $\mathbf{H}$ consists of the first gm columns of $(\mathbf{I} \otimes \mathbf{R})\mathbf{D}_{\boldsymbol{\theta},n}^{-1}$, to evaluate $\tilde{\mathbf{h}}$ and $\mathbf{H}^T\mathbf{H}$, the matrix $\mathbf{D}_{\boldsymbol{\theta},n}^{-1}$ is needed. It can be shown (Hillmer and Tiao 1979, pp. 652–653) that $\mathbf{D}_{\boldsymbol{\theta},n}^{-1}$ is a lower triangular block matrix with identity matrices along the main diagonal, $\mathbf{\Xi}_1$ down the subdiagonal, and so on, where the $\mathbf{\Xi}_k$'s are evaluated recursively as

$$\mathbf{\Xi}_k = \sum_{j=1}^{q} \boldsymbol{\Theta}_j\mathbf{\Xi}_{k-j} \qquad (k = 1, \ldots, n-1), \tag{18}$$

with $\mathbf{\Xi}_0 = \mathbf{I}$ and $\mathbf{\Xi}_k = \mathbf{0}$ for $k < 0$. Then the jth block of vector $\tilde{\mathbf{h}}$ is given by

$$\tilde{\mathbf{h}}_j = \sum_{i=0}^{n-j} \mathbf{\Xi}_i^T\mathbf{R}^T\boldsymbol{\eta}_{i+j} \qquad (j = 1, \ldots, g). \tag{19}$$

From the special structure of matrix $\mathbf{H}$, the first block column of $\mathbf{H}^T\mathbf{H}$ is given by

$$(\mathbf{H}^T\mathbf{H})_{i1} = \sum_{k=0}^{n-i} \mathbf{\Xi}_k^T\mathbf{R}^T\mathbf{R}\mathbf{\Xi}_{k+i-1} \qquad (i = 1, \ldots, g), \tag{20}$$

and the remaining diagonal and subdiagonal blocks of matrix $\mathbf{H}^T\mathbf{H}$ are evaluated in the following recursive manner:

$$(\mathbf{H}^T\mathbf{H})_{ij} = (\mathbf{H}^T\mathbf{H})_{i-1,j-1} - \mathbf{\Xi}_{n-i+1}^T\mathbf{R}^T\mathbf{R}\mathbf{\Xi}_{n-j+1}, \tag{21}$$

with $i = 2, \ldots, g$ and $j = 2, \ldots, i$. Finally, from (7) and (8), the n blocks that make up vector $\hat{\mathbf{a}}_0$ can be computed recursively as follows:

$$\hat{\mathbf{a}}_{0i} = \tilde{\mathbf{w}}_i - \sum_{j=1}^{p} \mathbf{\Phi}_j\tilde{\mathbf{w}}_{i-j} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j\hat{\mathbf{a}}_{0,i-j} \qquad (i = 1, \ldots, n), \tag{22}$$

with $\tilde{\mathbf{w}}_i = \mathbf{0}$ for $i < 1$ and $\hat{\mathbf{a}}_{0i} = \mathbf{0}$ for $i < 1$. Then the ith block of vector $\boldsymbol{\eta}$ is given by $\boldsymbol{\eta}_i = \mathbf{R}\hat{\mathbf{a}}_{0i}$ ($i = 1, \ldots, n$). (The calculation of $\boldsymbol{\eta}$ through (7), (8), and (22) can also be found in Hall and Nicholls 1980, p. 256 and Ljung and Box 1979, p. 267.)

In summary, the following procedure is suggested to evaluate the exact likelihood function of a vector ARMA model

(note that, except for step 1, no explicit matrix inversion is required):

1. Compute the Cholesky factor of matrix $\mathbf{Q}$ (say $\mathbf{Q}_1$), its determinant ($|\mathbf{Q}| = |\mathbf{Q}_1|^2$), and a matrix $\mathbf{R}$ such that $\mathbf{RQR}^T = \mathbf{I}$ ($\mathbf{R} = \mathbf{Q}_1^{-1}$).
2. Evaluate the theoretical autocovariance and cross-covariance matrices $\boldsymbol{\Gamma}(k)$ ($k = 0, \ldots, p-1$) and $\boldsymbol{\Gamma}_{wa}(k)$ ($k = -q+1, \ldots, 0$).
3. Compute matrix $\mathbf{V}_1\boldsymbol{\Omega}\mathbf{V}_1^T$ from (17) and compute its Cholesky factor $\mathbf{M}$.
4. Evaluate the sequence $\boldsymbol{\Xi}_k$ ($k = 1, \ldots, n-1$) from (18).
5. Calculate vector $\boldsymbol{\eta}$ using (22).
6. Compute vector $\tilde{\mathbf{h}}$ from (19) and evaluate vector $\mathbf{M}^T\tilde{\mathbf{h}}$.
7. Evaluate matrix $\mathbf{H}^T\mathbf{H}$ from (20) and (21).
8. Compute matrix $\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{HM}$, its Cholesky factor (say $\mathbf{L}$), and its determinant ($|\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{HM}| = |\mathbf{L}|^2$), which in turn is the determinant (16).
9. Use forward substitution to solve for $\boldsymbol{\lambda}$ in the triangular system $\mathbf{L}\boldsymbol{\lambda} = (\mathbf{M}^T\tilde{\mathbf{h}})$.
10. Compute the quadratic form (15) as $S(\boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{\mu}, \mathbf{Q}|\mathbf{w}) = \boldsymbol{\eta}^T\boldsymbol{\eta} - \boldsymbol{\lambda}^T\boldsymbol{\lambda}$.

Following the guidelines of Hillmer and Tiao (1979, pp. 653–654) and Ljung and Box (1979, p. 269), this procedure can also take advantage of the special structure of some specific models, such as multiplicative pure MA models and pure AR models. The details are straightforward and have been omitted.

2.2 Properties of the New Algorithm

The previously outlined procedure is basically the result of extending and taking one step beyond the method of Ljung and Box (1979) for the scalar ARMA model. Because the explicit inversion of matrix $\mathbf{V}_1\boldsymbol{\Omega}\mathbf{V}_1^T$ is avoided through the use of its Cholesky factorization, a more computationally efficient and numerically stable method is obtained. Further, because the new algorithm operates with $gm \times gm$ matrices [see (15) and (16)] instead of $(p+q)m \times (p+q)m$ matrices [see (6)], it is also preferable from a computational standpoint to that of Hall and Nicholls (1980). To illustrate this, computer programs were written for the algorithm of Hall and Nicholls (1980) and the new algorithm developed in this article. The exact likelihood function was evaluated for a variety of vector ARMA models suitable for annual, quarterly, and monthly data, under the assumption that 25 years of data were available. In Table 1 the ratio between the number of time-consuming operations (multiplications, divisions, and square roots) required by the algorithm of Hall and Nicholls (1987) and those required by the new algorithm is presented for each of the models considered. This ratio is always greater than or equal to 1 and reaches its highest value for models with both p and q large.

The comparison from a computational standpoint between the new algorithm and that of Shea (1987) is summarized in Table 2, which contains the same kind of information as Table 1. Apart from minor change through refinements in coding, it can be seen that the relative efficiency of the new algorithm increases with m (except if p is high and much larger than q), and that the new algorithm is clearly preferable for low- to medium-order models, whereas the method of Shea (1987) is more efficient for some higher-order models.

Table 1. Ratio between the Number of Time-consuming Operations Required by the Algorithm of Hall and Nicholls (1980) and Those Required by the New Algorithm, to Evaluate the Exact Likelihood Function for Various Models

Models	$m = 2$	$m = 4$
Models for annual data ($n = 25$)		
AR(1)	1.00	1.00
AR(2)	1.00	1.00
MA(1)	1.00	1.00
MA(2)	1.00	1.00
ARMA(1,1)	1.05	1.04
ARMA(2,1)	1.09	1.10
ARMA(1,2)	1.09	1.10
ARMA(2,2)	1.19	1.21
Models for quarterly data ($n = 100$)		
$AR(1)_4$	1.03	1.05
$MA(1)_4$	1.02	1.02
$ARMA(1,1)_4$	1.20	1.23
$AR(1) \times AR(1)_4$	1.04	1.07
$AR(1) \times MA(1)_4$	1.07	1.08
$AR(1) \times ARMA(1,1)_4$	1.24	1.28
$MA(1) \times AR(1)_4$	1.06	1.07
$MA(1) \times MA(1)_4$	1.03	1.04
$MA(1) \times ARMA(1,1)_4$	1.24	1.28
$ARMA(1,1) \times AR(1)_4$	1.07	1.09
$ARMA(1,1) \times MA(1)_4$	1.09	1.11
$ARMA(1,1) \times ARMA(1,1)_4$	1.30	1.34
Models for monthly data ($n = 300$)		
$AR(1)_{12}$	1.09	1.11
$MA(1)_{12}$	1.07	1.09
$ARMA(1,1)_{12}$	1.50	1.59
$AR(1) \times AR(1)_{12}$	1.10	1.12
$AR(1) \times MA(1)_{12}$	1.13	1.15
$AR(1) \times ARMA(1,1)_{12}$	1.52	1.61
$MA(1) \times AR(1)_{12}$	1.08	1.10
$MA(1) \times MA(1)_{12}$	1.09	1.10
$MA(1) \times ARMA(1,1)_{12}$	1.53	1.61
$ARMA(1,1) \times AR(1)_{12}$	1.09	1.11
$ARMA(1,1) \times MA(1)_{12}$	1.15	1.75
$ARMA(1,1) \times ARMA(1,1)_{12}$	1.57	1.65

NOTE: The operations required to compute the first $p-1$ autocovariance and the first $q-1$ cross-covariance matrices have been excluded, because they are required by both algorithms.

Finally, note that the method of Hillmer and Tiao (1979) does not allow for an exact evaluation of the likelihood function when the model contains an AR part; furthermore, in the case of pure MA models, the expression of the exact likelihood function obtained in that paper is equivalent to that of Nicholls and Hall (1979).

With regard to other interesting properties, note first that when the model considered is invertible, the matrix sequence (18) converges to $\mathbf{0}$; this convergence is more rapid, the larger the moduli of the zeros of $|\boldsymbol{\Theta}(B)|$ are (obviously, when $q = 0$, $\boldsymbol{\Xi}_k = \mathbf{0}$ for $k \geq 1$). This may be exploited in the subsequent computation of (19), (20), and (21), because if $\boldsymbol{\Xi}_k = \mathbf{0}$ for, say, $k \geq r^*$, then not all of the operations involved in those expressions need to be carried out. The sequence (18) may be considered to have converged when

Table 2. Ratio between the Number of Time-consuming Operations Required by the Algorithm of Shea (1989) and Those Required by the New Algorithm, to Evaluate the Exact Likelihood Function for Various Models

Models	$m = 2$	$m = 4$
Models for annual data ($n = 25$)		
AR(1)	1.06*	1.20*
AR(2)	1.00*	1.07*
MA(1)	2.37*	2.69*
MA(2)	1.75*	1.92*
ARMA(1,1)	2.37*	2.75*
ARMA(2,1)	1.83*	2.00*
ARMA(1,2)	1.76*	1.95*
ARMA(2,2)	1.74*	1.94*
Models for quarterly data ($n = 100$)		
$AR(1)_4$	0.84	0.77
$MA(1)_4$	1.48*	1.60*
$ARMA(1,1)_4$	1.56*	1.72*
$AR(1) \times AR(1)_4$	0.74	0.65
$AR(1) \times MA(1)_4$	1.51*	1.65*
$AR(1) \times ARMA(1,1)_4$	1.49*	1.57*
$MA(1) \times AR(1)_4$	1.53*	1.60*
$MA(1) \times MA(1)_4$	1.36*	1.46*
$MA(1) \times ARMA(1,1)_4$	1.41*	1.54*
$ARMA(1,1) \times AR(1)_4$	1.34*	1.37*
$ARMA(1,1) \times MA(1)_4$	1.38*	1.49*
$ARMA(1,1) \times ARMA(1,1)_4$	1.41*	1.53*
Models for monthly data ($n = 300$)		
$AR(1)_{12}$	0.48	0.36
$MA(1)_{12}$	1.09*	1.16*
$ARMA(1,1)_{12}$	1.12*	1.19*
$AR(1) \times AR(1)_{12}$	0.44	0.33
$AR(1) \times MA(1)_{12}$	1.10*	1.17*
$AR(1) \times ARMA(1,1)_{12}$	1.15*	1.17*
$MA(1) \times AR(1)_{12}$	0.93	0.89
$MA(1) \times MA(1)_{12}$	1.05*	1.11*
$MA(1) \times ARMA(1,1)_{12}$	1.08*	1.13*
$ARMA(1,1) \times AR(1)_{12}$	0.87	0.82
$ARMA(1,1) \times MA(1)_{12}$	1.06*	1.12*
$ARMA(1,1) \times ARMA(1,1)_{12}$	1.07*	1.12*

NOTES: An asterisk indicates that the new algorithm is preferred. The operations required to compute the first $p - 1$ autocovariance and the first $q - 1$ cross-covariance matrices have been excluded, because they are required by both algorithms.

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{m} |\Xi_{r^*}(i,j)| \right) < \delta,$$

where the parameter $\delta > 0$ can be used to control the desired degree of approximation to the exact computation of the whole sequence (18). It is possible to make the convergence criterion sufficiently rigid (i.e., δ sufficiently small) such that the error implied by considering $\Xi_k = \mathbf{0}$ for $k \geq r^*$ becomes negligible and hence so too does the difference between the exact (calculated with Ξ_k from $k = 1$ to $k = n - 1$) and the "approximate" (calculated with $\Xi_k = \mathbf{0}$ for $k \geq r^*$) likelihood. Note that this property, which may save much computing time, is analogous to the "quick recursions" property offered by the Chandrasekhar equations that form the basis of the method of Shea (1989, pp. 169–170). Furthermore, using the convergence property of (18) for invertible models, it is straightforward to detect the presence of any root of the MA operator lying inside the unit circle, because in such a case the sequence (18) will be explosive. In practice, it has been observed that the following inequality holds for strictly noninvertible models:

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{m} |\Xi_h(i,j)| \right) > \sum_{k=1}^{\min(h,q)} \left(\sum_{i=1}^{m} \sum_{j=1}^{m} |\Xi_k(i,j)| \right),$$

for at least one $h < n - 1$. In general, when the MA operator has at least one root inside the unit circle, this condition will be true for h slightly larger than q, allowing detection of strict noninvertibility at the beginning of the computation of the sequence (18). In such a case, the evaluation algorithm flags a warning and stops, to avoid overflow problems in the subsequent computation of (20) and (21). But there is no problem in evaluating these expressions when any root of the MA operator lies on the unit circle, provided that the other roots have moduli larger than unity.

Also as a by-product, the new algorithm provides a necessary (though not sufficient, except for pure AR models) check on the stationarity of the model. This is due to the fact that the Cholesky decomposition of $\mathbf{V}_1 \boldsymbol{\Omega} \mathbf{V}_1^T$ exists if and only if $\boldsymbol{\Omega}$ is positive definite, which in turn is a necessary condition for stationarity. (It is also sufficient when $q = 0$; see, for example, Ansley 1979, pp. 61–62.) Because the computation of the Cholesky factor of $\mathbf{V}_1 \boldsymbol{\Omega} \mathbf{V}_1^T$ is a key step in the new algorithm, the impossibility of carrying out this operation indicates that the model considered is not stationary. Note, however, that existence of the Cholesky decomposition does not guarantee stationarity for a mixed model.

Finally, it is shown how to calculate the residuals for a given set of observations and parameter values, using some of the computations carried so far to evaluate the exact likelihood function. From (2) and the definition of $\mathbf{K}$, $\mathbf{Z}$, it is clear that $\hat{\mathbf{a}} = \mathbf{K}\tilde{\mathbf{w}} + \mathbf{Z}\hat{\mathbf{u}}_*$, with $\mathbf{K}\tilde{\mathbf{w}} = \hat{\mathbf{a}}_0$ [see (8)], and, from (5), it follows that $\hat{\mathbf{u}}_* = -\mathbf{T}^{-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Upsilon}\tilde{\mathbf{w}}$. Then, noting (4), (7), (8), and (12), it is possible to show that

$$\hat{\mathbf{a}} = \hat{\mathbf{a}}_0 - \mathbf{D}_{\boldsymbol{\theta},n}^{-1} \begin{bmatrix} \mathbf{M}(\mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})^{-1}\mathbf{M}^T\tilde{\mathbf{h}} \\ \mathbf{0} \end{bmatrix}. \quad (23)$$

Thus, using previous computations, the exact residual vector $\hat{\mathbf{a}}$ can be evaluated by (1) using backward substitution to solve for $\mathbf{c}$ in the triangular system $\mathbf{L}^T\mathbf{c} = \boldsymbol{\lambda}$, (2) computing the $gm \times 1$ vector $\mathbf{d} = \mathbf{M}\mathbf{c}$, and (3) evaluating $\hat{\mathbf{a}} = \hat{\mathbf{a}}_0 - \mathbf{r}$, where the ith block of the $nm \times 1$ vector $\mathbf{r} = \mathbf{D}_{\boldsymbol{\theta},n}^{-1}(\mathbf{d}^T, \mathbf{0}^T)^T$ is given by

$$\mathbf{r}_i = \sum_{j=1}^{i} \Xi_{i-j}\mathbf{d}_j \qquad (i = 1, \ldots, n),$$

with $\mathbf{d}_j = \mathbf{0}$ for $j > g$. Conditional on maximum likelihood estimates being equal to the true parameter values, the residuals thus calculated can be shown to be normally distributed with $E[\hat{\mathbf{a}}_t] = \mathbf{0}$. Further, it can be shown that as t tends to n, $\hat{\mathbf{a}}_t$ converges in quadratic mean to $\mathbf{a}_t$ and the $\hat{\mathbf{a}}_t$'s tend to be uncorrelated, with $E[\hat{\mathbf{a}}_t\hat{\mathbf{a}}_t^T]$ converging to $\sigma^2\mathbf{Q}$. (When $q = 0$, this convergence occurs exactly for $t > p$.) These properties are shared with the residuals obtained by using the Kalman filter to evaluate the exact likelihood function (Shea 1984, p. 93; 1989, p. 162).

To conclude, Table 3 shows a comparative summary of the evaluation algorithms considered in this section, in terms of the following features: (1) exact evaluation of the likelihood function; (2) "approximate" evaluation of the likeli-

Table 3. Comparison of the Various Algorithms for Evaluating the Likelihood Function

	Hall and Nicholls (1980)	Ljung and Box (1979)	Hillmer and Tiao (1979)	Shea (1987, 1989)	New Algorithm
Exact evaluation	Yes	Yes	No[a]	Yes	Yes
Approximate evaluation	Yes[b]	Yes[b]	Yes[c]	Yes	Yes
Computational efficiency	No	No	Yes[d]	Yes[e]	Yes[e]
Numerical precision and stability	Yes	No	No[a]	Yes	Yes
Detection of nonstationarity	Yes	Yes	No	Yes	Yes
Detection of noninvertibility	Yes	Yes	Yes	No	Yes
Calculation of appropriate residuals	Yes[b]	Yes	No[a]	Yes	Yes

[a] Except in the case of pure moving average models ($p = 0$).
[b] Though this feature is not mentioned nor exploited in the papers cited.
[c] In fact this is the only possibility offered by this method, which is based on questionable assumptions and does not permit control of the accuracy of the approximation.
[d] Only when the model contains an autoregressive part ($p \neq 0$).
[e] See the comparison in Table 2.

hood function, as accurate as desired and in most cases significantly faster than the exact evaluation; (3) computational efficiency, in terms of the number of time-consuming operations required; (4) numerical accuracy and stability; (5) detection, as a by-product, of nonstationary and/or noninvertible models; and (6) evaluation of an appropriate residual vector using some of the computations carried to evaluate the likelihood function.

In summary, the new evaluation algorithm provides a set of useful features, not found together in any of the other existing methods, that can be put to work effectively in the context of maximum likelihood estimation of vector ARMA models.

3. MAXIMIZATION OF THE EXACT LIKELIHOOD FUNCTION

Having devised a method of computing the exact likelihood function, we now seek how to maximize it with respect to the parameters $\mathbf{\Phi} = (\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p)$, $\mathbf{\Theta} = (\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_q)$, μ, σ^2, and $\mathbf{Q}$. The parameter σ^2 may be differentiated out of Equation (3) to yield the following concentrated log-likelihood:

$$l_*(\mathbf{\Phi}, \mathbf{\Theta}, \mu, \mathbf{Q}|\mathbf{w}) = -\frac{mn}{2}\left[\log\left(\frac{2\pi}{mn}\right) + 1\right] - \frac{n}{2}\log(\Pi_1\Pi_2), \quad (24)$$

where

$$\Pi_1 = (\boldsymbol{\eta}^T\boldsymbol{\eta} - \boldsymbol{\lambda}^T\boldsymbol{\lambda})^m, \quad (25)$$

$$\Pi_2 = |\mathbf{Q}|\,|\mathbf{D}|^{1/n}, \quad (26)$$

and the $gm \times gm$ matrix $\mathbf{D}$ is given by $\mathbf{D} = \mathbf{I} + \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}$ [see (15) and (16)]. Thus maximizing (24) is equivalent to minimizing

$$\Pi = \Pi_1\Pi_2. \quad (27)$$

Let Π_0 be the value of (27) at the initial estimates of the parameters ($\Pi_0 = \Pi_{10}\Pi_{20}$). Thus if we minimize, instead of (27), the function

$$F = \frac{\Pi}{\Pi_0} = \frac{\Pi_1}{\Pi_{10}}\frac{\Pi_2}{\Pi_{20}}, \quad (28)$$

using a routine that generates descent search directions in every iteration, then the objective function F always lies in the interval $(0, 1)$. This fact has two advantages. On the one hand, it improves the overall accuracy and numerical stability of the minimization routine, especially in the computation of the gradient vector through finite differences. On the other hand, it provides a simple means of handling situations in which the algorithm generates new estimates that imply nonstationarity, noninvertibility, and/or non-positive definiteness of the matrix $\mathbf{Q}$. In such instances, which can be detected as described in the previous section, the scaled objective function (28) is set to 1. Thus the minimization routine will reject these points and continue the search for an acceptable local optimum. Note that this strategy is basically the one proposed by Shea (1984, pp. 99–100), although we do not solve the determinantal polynomials $|\mathbf{\Phi}(B)| = 0$ and $|\mathbf{\Theta}(B)| = 0$ to check for nonstationarity and noninvertibility, because the new evaluation algorithm provides simpler means for carrying out those checks.

To generate improving search directions, we use a quasi-Newton method based on the factorized version of the BFGS formula (see, for example, Dennis and Schnabel 1983, chap. 9). Besides computational efficiency, this method provides, as a by-product, a means of estimating the covariance matrix of the parameter estimates, because the relevant information on the curvature of the objective function (28) is updated at every iteration along with the computation of the search direction. From (24), the information matrix is given by

$$\mathbf{I} = E\left(-\frac{n}{2\Pi^2}\nabla\Pi\nabla\Pi^T + \frac{n}{2\Pi}\nabla^2\Pi\right),$$

where $\nabla\Pi$ is the gradient vector of (27) and $\nabla^2\Pi$ is the Hessian matrix. Because $\nabla\Pi = \mathbf{0}$ at any local optimum, a sample estimate of the covariance matrix is given by $2F(n\nabla^2 F)^{-1}$, where F and $(\nabla^2 F)^{-1}$ are evaluated at the final estimates. If we use a quasi-Newton method based on the factorized version of the BFGS formula to minimize (28), then we will have at the end of the iterative process an approximation to the Cholesky factor of $\nabla^2 F$ (see Dennis and Schnabel 1983, pp. 206–207), which makes the computation of the covariance matrix estimate straightforward.

In summary, we suggest the use of the following procedure to maximize the exact likelihood function of a vector ARMA model:

1. Choose a suitable set of initial estimates of the parameters and compute Π_{10} and Π_{20} from (25) and (26).

2. Minimize the scaled objective function (28) using a quasi-Newton method based on the factorized version of the BFGS formula.

3. On convergence, use the accumulated information on the Cholesky factor of $\nabla^2 F$ to evaluate a sample estimate of the covariance matrix as $2F(n\nabla^2 F)^{-1}$.

Each time that we compute Π_1 and Π_2 from (25) and (26), we make use of the evaluation algorithm of the previous section and set the scaled objective function (28) to 1 whenever the algorithm detects nonstationarity, noninvertibility, and/or non-positive definiteness of $\mathbf{Q}$. It must be noted that for a mixed model, the algorithm may converge to a nonstationary point. Although this has never happened in practice, the computation (on convergence) of the roots of $|\Phi(B)| = 0$ should be performed to ensure that the final estimates are admissible. The residual vector is evaluated only after the minimization routine has converged, because it is not used during the iterative process. Note also that all of these computations can be speeded up using the approximation to the exact likelihood function discussed in the previous section. Finally, initial estimates may be conditional maximum likelihood estimates or those obtained with other fast linear estimation methods (see, for example, Koreisha and Pukkila 1989 or Shea 1987), although care must be taken to ensure that they are admissible.

4. AN EXAMPLE

It is well known (see, for example, Ansley and Newbold 1980 or Hillmer and Tiao 1979) that exact maximum likelihood estimation is usually preferable to other approximate estimation criteria, especially in the case of small- to moderate-sized samples and/or parameters close to the boundaries of the admissible regions. This issue is not pursued further here. Instead, it is illustrated with an example that a set of conditions may hold under which the estimation method proposed in this article performs better than one of the most frequently used in practice, the "exact" version of the procedure of Hillmer and Tiao (1979) as implemented in *The PC SCA Statistical System,* release 4.1 (see Liu and Hudak 1992, pp. 5.15–5.16).

A series of 120 monthly observations on the Energy component of the Spanish Industrial Production Index, covering the period January 1982–December 1991, has been considered. The data have been obtained from the *Boletín Estadístico del Banco de España* (Banco de España, Madrid) and are available on request from the author.

After trying some alternative patterns of differentiation on the original series, it seems clear that $w_t = \nabla\nabla_{12} z_t$, where z_t denotes the natural logarithm of the original series, can be considered to be stationary (see Fig. 1). The autocorrelation and partial autocorrelation functions for w_t (see Fig. 1) suggest that this time series might be described by an $\mathrm{MA}(1) \times \mathrm{MA}(1)_{12}$ model.

However, nonsample information suggests the inclusion of two deterministic variables, ξ_{t1} and ξ_{t2}, representing a unit impulse effect in February 1990 and the Easter holiday. Thus the following intervention model is specified:

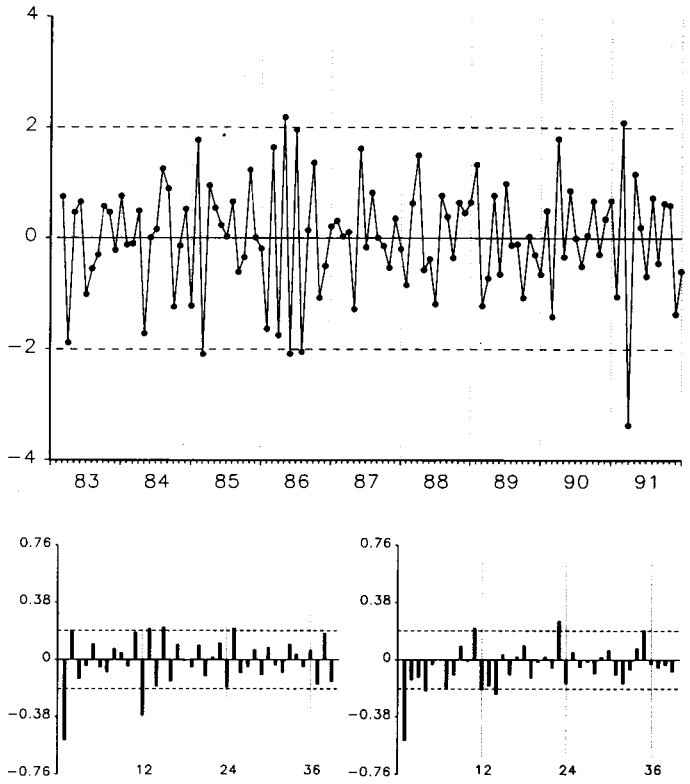

Figure 1. Series $w_t = \nabla\nabla_{12} z_t$: Standardized Series (Top), Autocorrelation (Bottom Left), and Partial Autocorrelation (Bottom Right) Functions. $\bar{w}$ = −.0011 (.0051); σ_w = .0524. Ljung–Box (1978) statistic: Q(39) = 118.0. This high value of the Q statistic merely reflects the so-far unmodeled structure observed in the autocorrelation and partial autocorrelation functions.

$$z_t = \omega_1 \xi_{t1} + \omega_2 \xi_{t2} + N_t, \tag{29}$$

$$\nabla\nabla_{12} N_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12}) a_t. \tag{30}$$

The estimates of (29) and (30) obtained with both the exact and approximate versions of the new algorithm, and those obtained with the algorithm of Hillmer and Tiao (1979), are summarized in Table 4. Clearly, there is no appreciable difference between the exact and approximate (obtained with $\delta = .01$) estimates calculated with the new algorithm. Further, these estimates are almost identical to those obtained with the procedure of Hillmer and Tiao (1979).

Table 4. Estimation of Model (29)–(30)[a,b]

	Exact[c]	*Approximate*[c]	*Hillmer and Tiao (1979)*[d]
ω_1	−.04 (.01)	−.04 (.01)	−.04 (.01)
ω_2	−.09 (.03)	−.09 (.02)	−.09 (.03)
θ_1	.60 (.14)	.60 (.14)	.61 (.08)
Θ_1	.84 (.13)	.84 (.13)	.86 (.06)
σ_a	.0298	.0298	.0296

[a] Initial estimates: $\omega_1 = -.03$, $\omega_2 = -.09$, $\theta_1 = .5$, and $\Theta_1 = .5$.
[b] Estimated standard errors in parentheses.
[c] Convergence obtained in 25 iterations.
[d] Convergence obtained in 49 iterations.

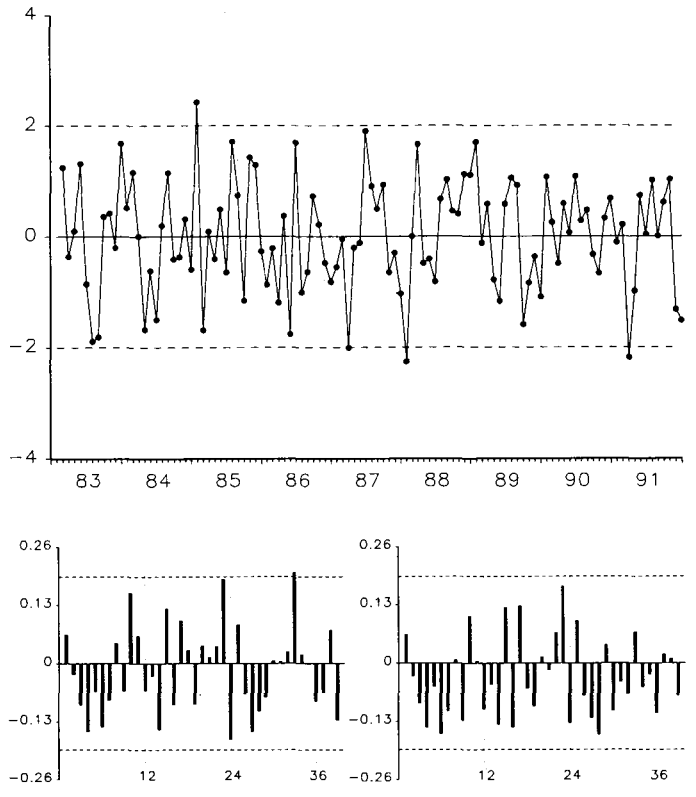

Figure 2. Estimation of (29)–(30) with the New Algorithm: Standardized Residuals (Top), Autocorrelation (Bottom Left), and Partial Autocorrelation (Bottom Right) Functions. $\bar{a} = -.0016$ (.0028); $\hat{\sigma}_a = .0298$. Ljung–Box (1978) statistic: $Q(35) = 49.4$. The autocorrelation function might suggest the adding of a second-order seasonal autoregressive operator to the model.

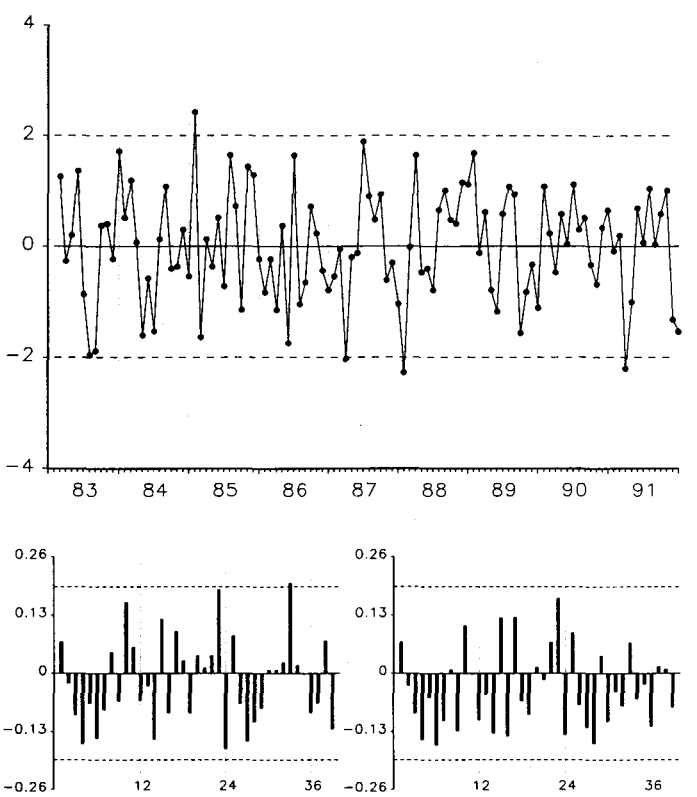

Figure 3. Estimation of (29)–(30) with the Algorithm of Hillmer and Tiao (1979): Standardized Residuals (Top), Autocorrelation (Bottom Left), and Partial Autocorrelation (Bottom Right) Functions. $\bar{a} = -.0015$ (.0027); $\hat{\sigma}_a = .0296$. Ljung–Box (1978) statistic: $Q(35) = 49.4$. The autocorrelation function might suggest the adding of a second-order seasonal autoregressive operator to the model.

Of course, there is no appreciable difference to be found in the corresponding residuals and their autocorrelation functions, presented in Figures 2 and 3.

Although the estimated model seems adequate, one may want to add a second-order seasonal AR term to the model as an overfitting exercise, because $AR(2)_{12}$ operators with imaginary roots are found frequently in Spanish Industrial Production Indices. This action will cope with the high value of the residual autocorrelation functions at lag 24 (see Figs. 2 and 3), and, although the resulting model may be overparameterized, it is a model that might be used for forecasting purposes. Thus model (29)–(30) is respecified as follows:

$$z_t = \omega_1 \xi_{t1} + \omega_2 \xi_{t2} + N_t, \tag{31}$$

$$(1 - \Phi_1 B^{12} - \Phi_2 B^{24}) \nabla \nabla_{12} N_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12}) a_t. \tag{32}$$

The estimates of (31) and (32) obtained with both the exact and approximate versions of the new algorithm, and those obtained with the algorithm of Hillmer and Tiao (1979), are summarized in Table 5.

The estimates obtained with the exact and approximate versions of the new algorithm are again almost identical. Further, these estimates are close to those obtained with the procedure of Hillmer and Tiao (1979), except for the seasonal MA parameter, which is estimated to be noninvertible using that method, whereas it is invertible if we use the new algorithm. This difference can be explained by examining the residuals depicted in Figures 4 and 5.

Figure 4 presents the (standardized) residuals, evaluated from (23), corresponding to the estimates obtained with the new algorithm, along with its autocorrelation and partial autocorrelation functions. The same information is presented in Figure 5, using the output generated by the procedure of Hillmer and Tiao (1979).

With regard to Figure 5, it may be noted that the first 24 residuals are not available, because the first p observations

Table 5. Estimation of Model (31)–(32)[a,b]

	Exact[c]	Approximate[c]	Hillmer and Tiao (1979)[d]
ω_1	−.04 (.01)	−.04 (.01)	−.04 (.01)
ω_2	−.09 (.02)	−.09 (.02)	−.09 (.02)
Φ_1	−.04 (.14)	−.04 (.14)	−.11 (.10)
Φ_2	−.23 (.12)	−.24 (.12)	−.23 (.10)
θ_1	.54 (.16)	.54 (.16)	.63 (.09)
Θ_1	.75 (.15)	.76 (.15)	1.09 (.08)
σ_a	.0293	.0293	.0240

[a] Initial estimates: $\omega_1 = -.03$, $\omega_2 = -.09$, $\Phi_1 = .1$, $\Phi_2 = -.1$, $\theta_1 = .5$, $\Theta_1 = .5$.
[b] Estimated standard errors in parentheses.
[c] Convergence obtained in 36 iterations.
[d] Convergence not obtained within 100 iterations. The procedure was restarted using as initial estimates the final ones from the first column, but it did not converge.

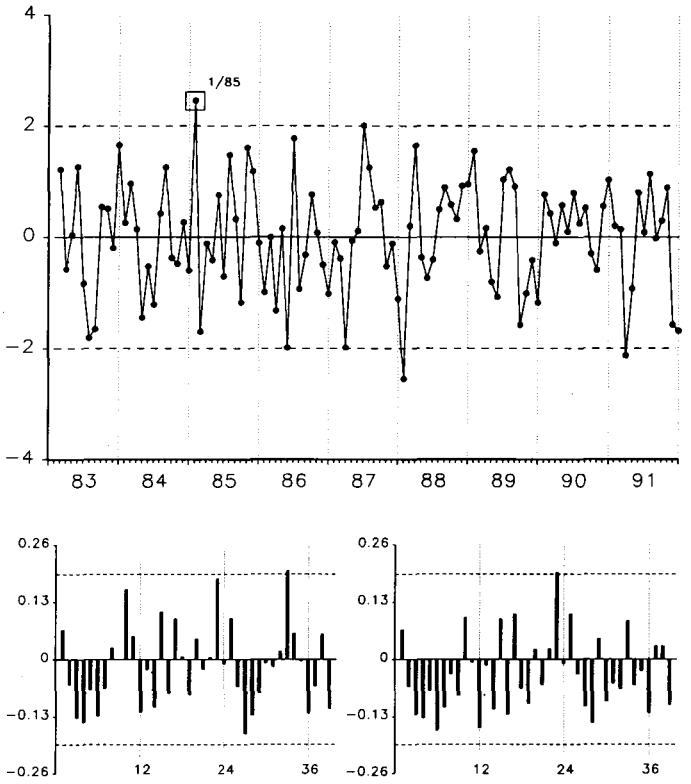

Figure 4. Estimation of (31)–(32) with the New Algorithm: Standardized Residuals (Top), Autocorrelation (Bottom Left), and Partial Autocorrelation (Bottom Right) Functions. $\bar{a} = -.0016$ (.0027); $\hat{\sigma}_a = .0293$. Ljung–Box (1978) statistic: $Q(33) = 45.8$. Although $Q(33)$ suggests misspecification, no structure is appreciated in the autocorrelation function. (Most likely, that value is due to outliers.)

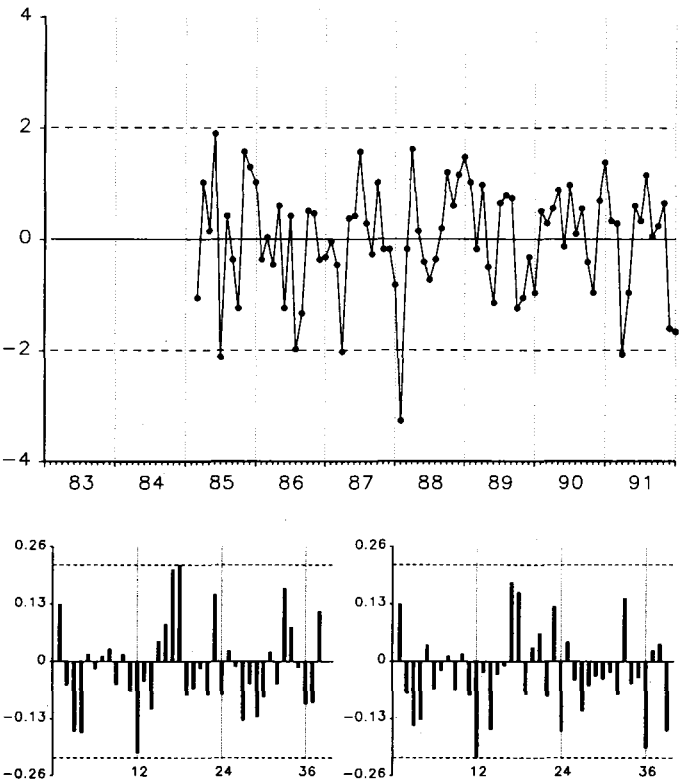

Figure 5. Estimation of (31)–(32) with the Algorithm of Hillmer and Tiao (1979): Standardized Residuals (Top), Autocorrelation (Bottom Left), and Partial Autocorrelation (Bottom Right) Functions. $\bar{a} = -.0001$ (.0024); $\hat{\sigma}_a = .0240$. Ljung–Box (1978) statistic: $Q(33) = 44.3$. The autocorrelation and partial autocorrelation functions suggest the need for a first-order seasonal moving average operator.

of w_t are used by the method of Hillmer and Tiao (1979) as starting values to compute a sequence supposedly generated by the MA part of the model. Further, the autocorrelation and partial autocorrelation functions of Figure 5 suggest the need for a first-order seasonal MA operator, which in fact is already included but is estimated to be noninvertible.

The problem lies in the presence of the observation corresponding to January 1985, which shows a residual only slightly larger than two standard deviations in Figure 4, though it pushes the seasonal MA operator out of the invertibility region when the method of Hillmer and Tiao (1979) is used. To see this, the following intervention model has been estimated:

$$z_t = \omega_1 \xi_{t1} + \omega_2 \xi_{t2} + \omega_3 \xi_{t3} + N_t, \tag{33}$$

$$(1 - \Phi_1 B^{12} - \Phi_2 B^{24}) \nabla \nabla_{12} N_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12}) a_t, \tag{34}$$

where ξ_{t3} is a unit impulse variable in January 1985. The estimates of (33) and (34) are summarized in Table 6. The seasonal MA operator now lies within the invertibility region using either the new algorithm or the procedure of Hillmer and Tiao (1979). This in fact was the case in Table 5 when the new algorithm was used. Note also from Table 6 that Φ_2 is clearly significantly different from zero.

Thus when estimating a model with an AR term, if an extreme value occurs in one of the first p observations, then the new algorithm performs robustly, whereas the algorithm of Hillmer and Tiao (1979) can be misleading. Furthermore, in the example the residual corresponding to January 1985 does not appear in Figure 5; hence a situation of this kind is hard to detect if the latter method is used.

5. CONCLUSIONS

Both the theoretical development and the illustration of the performance of the estimation algorithm proposed in this article have shown the following important points:

1. It is possible to improve, as in Section 2, existing methods of evaluating the exact likelihood function of vector

Table 6. Estimation of Model (33)–(34)[a,b]

	Exact[c]	Approximate[c]	Hillmer and Tiao (1979)[d]
ω_1	−.04 (.01)	−.04 (.01)	−.04 (.01)
ω_2	−.10 (.02)	−.10 (.02)	−.10 (.02)
ω_3	.08 (.02)	.08 (.02)	.07 (.02)
Φ_1	−.12 (.14)	−.12 (.14)	−.17 (.11)
Φ_2	−.34 (.12)	−.33 (.12)	−.31 (.10)
θ_1	.39 (.15)	.39 (.15)	.50 (.10)
Θ_1	.64 (.15)	.64 (.15)	.70 (.09)
σ_a	.0280	.0280	.0266

[a] Initial estimates: $\omega_1 = -.03$, $\omega_2 = -.09$, $\omega_3 = .08$, $\Phi_1 = .1$, $\Phi_2 = -.1$, $\theta_1 = .5$, and $\Theta_1 = .5$.
[b] Estimated standard errors in parentheses.
[c] Convergence obtained in 41 iterations.
[d] Convergence obtained in 50 iterations.

ARMA models to put together a set of useful features that can only be found separately in currently available methods.

2. The adaptation of a computationally efficient minimization routine to those features yields an estimation procedure that not only provides true maximum likelihood estimates, but also provides useful instruments for diagnostic checking of the fitted models.

3. Actual situations may occur in which the new estimation algorithm performs better than the algorithms frequently used.

The procedures outlined in this article can be taken as a starting point in the development of new methods of estimating some generalizations of the vector ARMA model, such as the joint estimation of both the ARMA parameters and the coefficients of common nonstationary factors in multivariate models with series containing such factors. Other applications of those procedures, including the joint estimation of both the ARMA structure and the deterministic components associated with a vector of time series, are straightforward.

[Received April 1993. Revised January 1994.]

REFERENCES

Ansley, C. F. (1979), "An Algorithm for the Exact Likelihood of a Mixed Autoregressive-Moving Average Process," *Biometrika,* 64, 59–65.

Ansley, C. F., and Newbold, P. (1980), "Finite Sample Properties of Estimators for Autoregressive-Moving Average Models," *Journal of Econometrics,* 13, 159–183.

Dennis, J. E., and Schnabel, R. B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations,* Englewood Cliffs, NJ: Prentice-Hall.

Hannan, E. J. (1969), "The Identification of Vector Mixed Autoregressive-Moving Average Systems," *Biometrika,* 56, 223–225.

Hall, A. D., and Nicholls, D. F. (1980), "The Evaluation of Exact Maximum Likelihood Estimates for VARMA Models," *Journal of Statistical Computation and Simulation,* 10, 251–262.

Hillmer, S. C., and Tiao, G. C. (1979), "Likelihood Function of Stationary Multiple Autoregressive Moving Average Models," *Journal of the American Statistical Association,* 74, 652–660.

Kohn, R., and Ansley, C. F. (1982), "A Note on Obtaining the Theoretical Autocovariances of an ARMA Process," *Journal of Statistical Computation and Simulation,* 15, 273–283.

Koreisha, S., and Pukkila, T. (1989), "Fast Linear Estimation Methods for Vector Autoregressive Moving Average Models," *Journal of Time Series Analysis,* 10, 325–339.

Liu, L. M., and Hudak, G. B. (1992), *Forecasting and Time Series Analysis Using the SCA Statistical System* (Vol. 1), DeKalb, IL: Scientific Computing Associates.

Ljung, G. M., and Box, G. E. P. (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika,* 65, 297–303.

——— (1979), "The Likelihood Function of Stationary Autoregressive-Moving Average Models," *Biometrika,* 66, 265–270.

Nicholls, D. F., and Hall, A. D. (1979), "The Exact Likelihood Function of Multivariate Autoregressive-Moving Average Models," *Biometrika,* 66, 259–264.

Shea, B. L. (1984), "Maximum Likelihood Estimation of Multivariate ARMA Processes via the Kalman Filter," in *Time Series Analysis: Theory and Practice* (Vol. 5), (ed. O. D. Anderson), Amsterdam: North-Holland, pp. 91–101.

——— (1987), "Estimation of Multivariate Time Series," *Journal of Time Series Analysis,* 8, 95–109.

——— (1989), "Algorithm AS 242. The Exact Likelihood of a Vector Autoregressive-Moving Average Model," *Applied Statistics,* 38, 161–204.