A Fast Estimation Method for the Vector Autoregressive Moving Average Model With Exogenous Variables

HENRIK SPLIID*

A very fast and simple algorithm for estimation of the parameters of large multivariate time series and distributed lag models is presented. An analysis of the distribution of the estimates shows that they are asymptotically normal and unbiased, and that they have a variance that decreases like $1/n$, n being the sample size. The algorithm is especially applicable for estimation of large multivariate models where it is generally many times faster than maximalization algorithms.

KEY WORDS: Fast estimation; Multivariate time series; Multivariate distributed lags; Linear regression; Vector ARMAX models; Dynamic systems.

1. INTRODUCTION

Consider a k-variate time series $\mathbf{y}_t$ generated by a mixed autoregressive moving average model of order (p, q):

$$\boldsymbol{\phi}(B)\mathbf{y}_t = \boldsymbol{\theta}(B)\mathbf{a}_t, \quad t \text{ discrete}, \tag{1.1}$$

where $\boldsymbol{\phi}(B) = \mathbf{I} - \boldsymbol{\phi}_1 B - \cdots - \boldsymbol{\phi}_p B^p$ and $\boldsymbol{\theta}(B) = \mathbf{I} - \boldsymbol{\theta}_1 B - \cdots - \boldsymbol{\theta}_q B^q$ are $k \times k$ matrix polynomials with no common left divisors. B is the usual backward shift operator and $\mathbf{I}$ is the unit matrix; $\mathbf{a}_t$ is white noise with covariance matrix $\boldsymbol{\Omega}$. Finally, the model is assumed to be stationary and invertible.

The estimation of the parameters in $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\theta}(\cdot)$ has been treated by several authors both from a theoretical and from a computational point of view. Briefly, it is well known that the maximum likelihood estimates are efficient and the estimates based on, for example, direct use of sample covariances can be rather inefficient when $q > 0$. This was first discussed by Whittle (1951,1952,1953) in three classic articles about maximum likelihood methods in both the univariate and the multivariate cases. Since then maximum likelihood procedures have received much attention in the literature, especially after the publication of the book by Box and Jenkins (1970), in which the univariate model is described. The multivariate or vector model has been treated by Hannan (1970), Akaike (1973,1974), Wilson (1973), Phadke and Kedem (1978), Nicholls and Hall (1979) and Hillmer and Tiao (1979). A recent survey of problems and methods in multivariate time series analysis is given by Jenkins and Alavi (1981).

In many situations the process $\mathbf{y}_t$ is not only a result of purely stochastic inputs, but it may also be dependent on, for example, controlled inputs. A general linear model that includes lagged regression variables is obtained by extending (1.1):

$$\boldsymbol{\phi}(B)\mathbf{y}_t = \boldsymbol{\beta}(B)\mathbf{x}_t + \boldsymbol{\theta}(B)\mathbf{a}_t, \quad t \text{ discrete}, \tag{1.2}$$

where $\boldsymbol{\beta}(B) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 B + \cdots + \boldsymbol{\beta}_{r-1} B^{r-1}$ is a $k \times m$ matrix polynomial and the series $\mathbf{x}_t$ is an m-variate series representing controlled inputs and/or observable disturbances. Models of this nature have been discussed, for example, by Hannan, Duinsmuir, and Deistler (1980), Tiao and Box (1981), and Priestley (1981). In "control theory" a number of recursive algorithms have been proposed. Some of these algorithms are essentially based on the same idea as the algorithm we present here and they are known under names such as "extended matrix method," "approximate maximum likelihood," and "recursive maximum likelihood," although they are not maximum likelihood methods. A description of these procedures is given by Goodwin and Payne (1977).

Often it is advantageous to embed the model (1.2) in a model of the form (1.1) by including the $\mathbf{x}_t$ series in the $\mathbf{y}_t$ series. Briefly, this important approach results in a more flexible model by which possible feedback loops in which $\mathbf{x}$ variables take part may be studied. Hence (1.1) creates a common framework for analyzing both open and closed loop dynamic systems (see, e.g., Tiao and Box 1981).

In this article an algorithm is presented that essentially is a method of moments, although not in a traditional form. Throughout the article the estimates of $\boldsymbol{\phi}(\cdot)$, $\boldsymbol{\theta}(\cdot)$, and $\boldsymbol{\beta}(\cdot)$ resulting from the algorithm are called $\tilde{\boldsymbol{\phi}}(\cdot)$, $\tilde{\boldsymbol{\theta}}(\cdot)$, and $\tilde{\boldsymbol{\beta}}(\cdot)$, respectively, and the corresponding residuals are called $\tilde{\mathbf{a}}_t$. These residuals are found by recursive

[*] Henrik Spliid is Associate Professor, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, DK-2800 Lyngby, Denmark. The author wishes to thank Jan Holst for valuable criticism and suggestions. The research was supported by the Danish National Research Council through Grant 511-6801 and the Danish Technical Research Council through Grant 516-15263. During preparation of the article substantial changes were made as a result of the constructive suggestions made by the referee and the Associate Editor, for which the author is very grateful. Finally the author wants to thank Eivind Damsleth at the Norwegian Computing Center for helping with the WMTS-1 program comparisons.

© Journal of the American Statistical Association
December 1983, Volume 78, Number 384
Theory and Methods Section

use of the estimated model in the usual way:

$$\tilde{\boldsymbol{\theta}}(B)\tilde{\mathbf{a}}_t = \tilde{\boldsymbol{\phi}}(B)\mathbf{y}_t - \tilde{\boldsymbol{\beta}}(B)\mathbf{x}_t, \quad t = 1, \ldots, n. \tag{1.3}$$

When the algorithm has converged we will show that the empirical autocorrelations and crosscorrelations are zero; that is,

$$\sum_t \tilde{\mathbf{a}}_{t-i}\tilde{\mathbf{a}}_t{}^T = 0, \quad i = 1, \ldots, q, \tag{1.4}$$

$$\sum_t \mathbf{y}_{t-i}\tilde{\mathbf{a}}_t{}^T = 0, \quad i = 1, \ldots, p, \tag{1.5}$$

$$\sum_t \mathbf{x}_{t-i}\tilde{\mathbf{a}}_t{}^T = 0, \quad i = 0, \ldots, r - 1. \tag{1.6}$$

From these relations it follows that $\sum(\mathbf{y}_t - \tilde{\mathbf{a}}_t)\tilde{\mathbf{a}}_t{}^T = \mathbf{0}$ such that the proposed estimates imply a separation of the $\mathbf{y}_t$ series in two orthogonal parts, namely $\tilde{\mathbf{a}}_t$ and $(\mathbf{y}_t - \tilde{\mathbf{a}}_t)$. This corresponds to the theoretical behavior of the postulated model where $\mathbf{a}_t$ and $(\mathbf{y}_t - \mathbf{a}_t)$ are independent.

2. THE ALGORITHM

The algorithm is presented in its general multivariate distributed lag version (1.2). The sample size is n.

Let $\mathbf{y}$ denote the $n \times k$ matrix of the observed time series:

$$\mathbf{y} = \begin{bmatrix} y_{l,l} & \cdots & y_{l,k} \\ \vdots & & \vdots \\ y_{n,l} & \cdots & y_{n,k} \end{bmatrix} = \{y_{ij}\},$$

$$i = 1, n, \quad j = 1, k. \tag{2.1}$$

Similarly define the matrix of (unknown) residuals $\mathbf{a} = \{a_{ij}\}$, $i = 1, n,\ j = 1, k$, and the matrix of regression variables $\mathbf{x} = \{x_{ij}\}$, $i = 1, n,\ j = 1, m$. Finally, define the following matrices of lagged data:

$$\mathbf{Y} = (B\mathbf{y}, B^2\mathbf{y}, \ldots, B^p\mathbf{y}),$$

$$\mathbf{A} = (B\mathbf{a}, B^2\mathbf{a}, \ldots, B^q\mathbf{a}),$$

$$\mathbf{X} = (\mathbf{x}, B\mathbf{x}, \ldots, B^{r-1}\mathbf{x}) \quad \text{and}$$

$$\mathbf{U} = (-\mathbf{A}, \mathbf{Y}, \mathbf{X}), \tag{2.2}$$

where, for example, $B^s\mathbf{y} = \{y_{i-s,j}\}$, $i = 1, n,\ j = 1, k$.

The matrices of coefficients are arranged in one single matrix of order $(kq + kp + mr) \times k$:

$$\boldsymbol{\delta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_q, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_p, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{r-1})^T. \tag{2.3}$$

Denote by $\boldsymbol{\delta}_{\cdot f}$ the fth column of $\boldsymbol{\delta}$. Thus $\boldsymbol{\delta}_{\cdot f}$ contains the parameters of the model for the fth variable in the series $\mathbf{y}$.

The model (1.2) may now be written in the equivalent forms:

$$\mathbf{y} = \mathbf{U}\boldsymbol{\delta} + \mathbf{a} \tag{2.4}$$

or

$$\mathbf{y}_{\cdot f} = \mathbf{U}\boldsymbol{\delta}_{\cdot f} + \mathbf{a}_{\cdot f}, \quad f = 1, \ldots, k, \tag{2.5}$$

where $\mathbf{y}_{\cdot f}$ and $\mathbf{a}_{\cdot f}$ are the fth columns of $\mathbf{y}$ and $\mathbf{a}$.

The problem of nonpositively indexed values may be dealt with by back-forecasting as described by Box and Jenkins (1970) in the pure ARMA case or by giving the first observations negative indices. Nonpositively indexed a's may be put equal to zero, at least during the first iterations. In the following it is assumed that some reasonable procedure is used and that the effect on the estimation is small.

The algorithm we initiated by estimating a model without moving average terms, but with an increased number, say s, of autoregressive terms (in practice $s = p + q$ often works well). This initialization is done by linear regression in the following Step 0. Afterwards the algorithm cycles through Steps 1–3 until convergence. It is worth noting that no initial parameter estimates are needed.

In the algorithm we denote by, say, $\hat{\mathbf{z}}(j)$ an estimate of a quantity $\mathbf{z}$ found in iteration number j.

Procedure

Step 0. Construct the matrix $\mathbf{W} = (B\mathbf{y}, B^2\mathbf{y}, \ldots, B^s\mathbf{y}, \mathbf{X})$, where $\mathbf{X}$ is defined in (2.2) and s is the chosen order of the initial autoregressive part of the model. Linear regression gives

$$\hat{\mathbf{a}}(0) = \mathbf{y} - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y},$$

where $\hat{\mathbf{a}}(0)$ denotes the estimate of the matrix $\mathbf{a}$ obtained in iteration no. 0 as mentioned earlier. Put $\hat{\boldsymbol{\delta}}(0) = 0$, put $j = 0$, and proceed with Step 2.

Step 1. Compute recursively residuals using for $t = 1, 2, \ldots, n$:

$$\hat{\mathbf{a}}_t(j) = \mathbf{y}_t - \sum_{i=1}^{p} \hat{\boldsymbol{\phi}}_i(j)\mathbf{y}_{t-i} + \sum_{i=1}^{q} \hat{\boldsymbol{\theta}}_i(j)\hat{\mathbf{a}}_{t-i}(j) - \sum_{i=0}^{r-1} \hat{\boldsymbol{\beta}}_i(j)\mathbf{x}_{t-i}.$$

Step 2. Construct $\hat{\mathbf{A}}(j)$, recall $\mathbf{Y}$ and $\mathbf{X}$, create $\hat{\mathbf{U}}(j)$, and compute new estimates by linear regression, that is, by solving

$$\hat{\mathbf{U}}^T(j)\,\hat{\mathbf{U}}(j)\,\hat{\boldsymbol{\delta}}(j + 1) = \hat{\mathbf{U}}^T(j)\,\mathbf{y}. \tag{2.6}$$

Step 3. If $\hat{\boldsymbol{\delta}}(j + 1) \neq \hat{\boldsymbol{\delta}}(j)$, increase j by 1 and repeat Steps 1 through 3. If $\hat{\boldsymbol{\delta}}(j + 1) = \hat{\boldsymbol{\delta}}(j)$, let $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(j + 1)$ and stop.

End of Procedure

We prove that the algorithm stops if and only if the relations (1.4), (1.5), and (1.6) are all satisfied.

By elementary algebra one can see that successive estimates satisfy the relationship

$$\hat{\mathbf{U}}^T(j)\,\hat{\mathbf{U}}(j)\,(\hat{\boldsymbol{\delta}}(j + 1) - \hat{\boldsymbol{\delta}}(j)) = \hat{\mathbf{U}}^T(j)\,\hat{\mathbf{a}}(j). \tag{2.7}$$

In regular cases the first product on the left side results in a nonsingular matrix. In these cases, obviously, the algorithm stops if and only if $\hat{\mathbf{U}}^T(j)\,\hat{\mathbf{a}}(j) = 0$. This completes the proof.

In some situations linear restrictions are imposed on the parameters of the model. If these restrictions are general in the sense that they relate parameters in different columns of $\boldsymbol{\delta}$, then Step 2 of the algorithm takes a more general form. Suppose the restrictions are of the form $\mathbf{C}\ \mathrm{vec}(\boldsymbol{\delta}) = \mathbf{d}$, where $\mathrm{vec}(\boldsymbol{\delta})$ is the vector organization of $\boldsymbol{\delta}$; that is, $\mathrm{vec}(\boldsymbol{\delta}) = (\boldsymbol{\delta}_{\cdot 1}{}^T, \boldsymbol{\delta}_{\cdot 2}{}^T, \ldots, \boldsymbol{\delta}_{\cdot k}{}^T)^T$, $\mathbf{C}$ is a matrix of coefficients, and $\mathbf{d}$ is a vector of constants. In this case the estimation equation of Step 2 becomes

$$(\mathbf{I} \otimes (\hat{\mathbf{U}}^T(j)\ \hat{\mathbf{U}}(j)))\ \mathrm{vec}(\hat{\boldsymbol{\delta}}(j+1)) + \mathbf{C}^T\boldsymbol{\lambda} = (\mathbf{I} \otimes \hat{\mathbf{U}}^T(j))\ \mathrm{vec}(\mathbf{y}),$$
$$\mathbf{C}\ \mathrm{vec}(\hat{\boldsymbol{\delta}}(j+1)) = \mathbf{d}, \quad (2.8)$$

where $\mathbf{I}$ is a unit matrix of order k, $\otimes$ is the tensor product, and $\boldsymbol{\lambda}$ is a Lagrange multiplier.

Successive estimates satisfy

$$\mathbf{M}(\mathrm{vec}(\hat{\boldsymbol{\delta}}(j+1) - \hat{\boldsymbol{\delta}}(j))) = (\mathbf{E} - \mathbf{H})(\mathbf{I} \otimes \hat{\mathbf{U}}(j))^T \mathrm{vec}(\hat{\mathbf{a}}(j)) - \mathbf{C}^T\mathbf{C}\mathbf{C}^T\mathbf{d}, \quad (2.9)$$

where

$$\mathbf{M} = \mathbf{I} \otimes (\hat{\mathbf{U}}^T(j)\hat{\mathbf{U}}(j)),$$
$$\mathbf{H} = \mathbf{C}^T(\mathbf{C}\mathbf{M}^{-1}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{M}^{-1},$$

and $\mathbf{E}$ is a unit matrix of order $k(kq + kp + mr)$. $\mathbf{M}^{-1} = \mathbf{I} \otimes (\hat{\mathbf{U}}^T(j)\hat{\mathbf{U}}(j))^{-1}$ is assumed to exist. The algorithm stops when the right side becomes zero.

If $\mathbf{C}$ and $\mathbf{d}$ are both zero, (2.8) and (2.9) reduce to (2.6) and (2.7), respectively. It is interesting to note that the extra amount of computing required as a result of the linear restrictions essentially relates to the computation of $\mathbf{H}$. If $\mathbf{C}$ has few rows or if $\mathbf{C}$ only restricts the parameters of the models for each of the variables (restrictions of the form $\mathbf{C}_f\ \boldsymbol{\delta}_{\cdot f} = \mathbf{d}_f$, $f = 1, k$), then the solution to (2.8) is generally easy to obtain from the solution of (2.6).

Finally some points of caution, based on practical experiences, are emphasized. It can be necessary to damp succesive parameter changes by multiplying the right side of (2.7) or (2.9) with a reduction factor of, say, .50–.75. When one estimates short series the treatment of the first observations in Step. 1 can be of great importance, but generally a simple back-forecasting seems to work well. Not surprisingly, convergence problems may occur if (a) an overparameterized model is estimated, (b) the time series is very short, (c) the moving average part of the model is nearly noninvertible, or (d) the time series exhibits nonstationary types of variations.

3. APPROXIMATE DISTRIBUTION OF PARAMETER ESTIMATES

In this section an approximate distribution of the proposed estimates is derived. We restrict attention to the autoregressive moving average model given in (1.1). Thus we consider estimates of the parameter matrices $\boldsymbol{\theta}_i$, $i = 1, q$, and $\boldsymbol{\phi}_i$, $i = 1, p$, in (1.1). The estimates obtained by our algorithm are called $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\phi}}$, respectively, and other estimates are called $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$.

The elements of the parameter matrices are organized in one vector $\boldsymbol{\alpha}$ of length $k^2(p + q)$ given by

$$\boldsymbol{\alpha} = \mathrm{vec}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_q, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_p), \quad (3.1)$$

which is the columns of $\boldsymbol{\theta}_1$ under each other followed by those of $\boldsymbol{\theta}_2$, and so on. Estimates of $\boldsymbol{\alpha}$ are called $\tilde{\boldsymbol{\alpha}}$ or $\hat{\boldsymbol{\alpha}}$, as explained.

Now consider a time series $\mathbf{y}_t$, $t = 1, n$, generated by the model (1.1), and assume an estimate $\hat{\boldsymbol{\alpha}}$ to be at hand. The corresponding series of estimated residuals is called $\hat{\mathbf{a}}_t$, $t = 1, n$, and it is found in the usual recursive way. With these residuals we construct the following $k \times k$ matrices of sample moments:

$$\mathbf{R}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = \mathbf{R}_i(\hat{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{t=1}^{n} \hat{\mathbf{a}}_{t-i}\hat{\mathbf{a}}_t{}^T, \quad i = 1, \ldots, q, \quad (3.2)$$

$$\mathbf{G}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = \mathbf{G}_i(\hat{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-i}\hat{\mathbf{a}}_t{}^T, \quad i = 1, \ldots, p, \quad (3.3)$$

which, for a given time series, depend only on the value of $\hat{\boldsymbol{\alpha}}$.

These matrices are also arranged in a $k^2(p + q)$ vector,

$$\mathbf{V}_n(\hat{\boldsymbol{\alpha}}) = \mathrm{vec}(\mathbf{R}_1(\hat{\boldsymbol{\alpha}}), \ldots, \mathbf{R}_q(\hat{\boldsymbol{\alpha}}), \mathbf{G}_1(\hat{\boldsymbol{\alpha}}), \ldots, \mathbf{G}_p(\hat{\boldsymbol{\alpha}})), \quad (3.4)$$

where n indicates the sample size.

The asymptotic value of $\mathbf{V}_n(\hat{\boldsymbol{\alpha}})$ is $\mathbf{V}(\hat{\boldsymbol{\alpha}})$, which contains the true autocovariances and crosscovariances of the model for the residuals $\hat{\mathbf{a}}_t$, that is

$$\hat{\mathbf{a}}_t = (\hat{\boldsymbol{\phi}}^{-1}(B)\hat{\boldsymbol{\theta}}(B))^{-1}(\boldsymbol{\phi}^{-1}(B)\boldsymbol{\theta}(B))\mathbf{a}_t,$$

when the estimate $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ is used.

We may now use a first-order Taylor expansion of $\mathbf{V}_n(\hat{\boldsymbol{\alpha}})$ about the point $\boldsymbol{\alpha}$, and by inversion we have the approximation

$$\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \simeq (\partial\mathbf{V}_n(\hat{\boldsymbol{\alpha}})/\partial\hat{\boldsymbol{\alpha}})^{-1}\ (\mathbf{V}_n(\hat{\boldsymbol{\alpha}}) - \mathbf{V}_n(\boldsymbol{\alpha})), \quad (3.5)$$

where the matrix of derivatives is evaluated at $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$. The expansion is valid for $\hat{\boldsymbol{\alpha}}$ in a small region around $\boldsymbol{\alpha}$, where $\mathbf{V}_n(\cdot)$ is continuously differentiable.

The expansion (3.5) is used for the particular estimate $\hat{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}$ in which case, by (1.4) and (1.5), $\mathbf{V}_n(\hat{\boldsymbol{\alpha}}) = \mathbf{V}_n(\tilde{\boldsymbol{\alpha}}) = \mathbf{0}$. Next, the matrix of derivatives is replaced by its large-sample expectation denoted by $\mathbf{D}(\boldsymbol{\alpha}) = \partial\mathbf{V}(\hat{\boldsymbol{\alpha}})/\partial(\hat{\boldsymbol{\alpha}})_{|\hat{\boldsymbol{\alpha}}=\boldsymbol{\alpha}}$. This leads to the following theorem.

Theorem. Under the conditions just mentioned, for large n, the estimator $\tilde{\boldsymbol{\alpha}}$ is approximately normally distributed with mean $\boldsymbol{\alpha}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}} = \mathbf{D}^{-1}(\boldsymbol{\alpha})\mathbf{C}(\boldsymbol{\alpha})(\mathbf{D}^{-1}(\boldsymbol{\alpha}))^T/n$, where $\mathbf{C}(\boldsymbol{\alpha})/n$ is the asymptotic covariance matrix of $\mathbf{V}_n(\boldsymbol{\alpha})$; $\tilde{\boldsymbol{\alpha}}$ converges in probability to $\boldsymbol{\alpha}$; $\mathbf{D}(\boldsymbol{\alpha})$ and $\mathbf{C}(\boldsymbol{\alpha})$ are given in (3.6)–(3.12).

In the remaining part of this section we discuss the theorem briefly and find expressions for $\mathbf{D}(\boldsymbol{\alpha})$ and $\mathbf{C}(\boldsymbol{\alpha})$. Finally a small illustration is given.

The right side of (3.5) with $\tilde{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}$ is approximated by $\mathbf{D}^{-1}(\boldsymbol{\alpha})\mathbf{V}_n(\boldsymbol{\alpha})$. $\mathbf{D}(\boldsymbol{\alpha})$ is constant while $\mathbf{V}_n(\boldsymbol{\alpha})$ is the residual autocovariance and autocrosscovariance function at the first q respectively p lags in the particular situation where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$. In this case $\hat{\mathbf{a}}_t = \mathbf{a}_t$, and it is straightforward to show that the large-sample distribution of $\mathbf{V}_n(\boldsymbol{\alpha})$ is approximately normal with zero mean.

We indicate a proof that $\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}})$ converges in probability to $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\phi})$. For brevity we consider a pure moving average model of order q: $\mathbf{y}_t = \boldsymbol{\theta}(B)\mathbf{a}_t$. Estimating $\boldsymbol{\theta}$ by $\tilde{\boldsymbol{\theta}}$ implies that the corresponding residuals $\tilde{\mathbf{a}}_t$ are generated by the mixed autoregressive moving average model of order (q, q): $\tilde{\mathbf{a}}_t = \tilde{\boldsymbol{\theta}}^{-1}(B)\boldsymbol{\theta}(B)\mathbf{a}_t$. Following Hannan (1976), the difference between the true and the sample autocovariance function of this model converges in probability to zero roughly at the rate $n^{-1/2}$. But by (1.4) the first q sample autocovariances are all zero such that the first q true autocovariances of the residual model corresponding to $\tilde{\boldsymbol{\theta}}$ converge in probability to zero. Denote by $\boldsymbol{\Gamma}^q(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \{\boldsymbol{\Gamma}_i(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta})\}$, $i = 1, q$, the autocovariance function of $\tilde{\mathbf{a}}_t$ at the first q lags. $\boldsymbol{\Gamma}^q(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta})$ is an analytical function of $\tilde{\boldsymbol{\theta}}$ and since $\boldsymbol{\Gamma}^q(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \mathbf{0}$ has one and only one solution within the invertibility region, namely $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, it clearly follows that $\tilde{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ as $\boldsymbol{\Gamma}^q(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta})$ converges to zero. This reasoning may be repeated for the mixed model, which completes the proof.

Next, we evaluate the derivatives of $\mathbf{R}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ and $\mathbf{G}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ with respect to the parameter estimates. We use the notation $\hat{\mathbf{R}}_i$ for $\mathbf{R}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ and $\hat{\mathbf{G}}_i$ for $\mathbf{G}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$. Differentiation of $\hat{\boldsymbol{\phi}}(B)\mathbf{y}_t = \hat{\boldsymbol{\theta}}(B)\hat{\mathbf{a}}_t$ gives

$$\begin{aligned} \partial\hat{\mathbf{a}}_t/\partial\hat{\theta}_{rs,j} &= \hat{\boldsymbol{\theta}}^{-1}(B)\mathbf{D}_{rs}\hat{\mathbf{a}}_{t-j}, \\ \partial\hat{\mathbf{a}}_t/\partial\hat{\phi}_{rs,j} &= -\hat{\boldsymbol{\theta}}^{-1}(B)\mathbf{D}_{rs}\mathbf{y}_{t-j} \\ &= -\hat{\boldsymbol{\theta}}^{-1}(B)\mathbf{D}_{rs}\hat{\boldsymbol{\phi}}^{-1}(B)\hat{\boldsymbol{\theta}}(B)\hat{\mathbf{a}}_{t-j}, \end{aligned}$$

where $\hat{\theta}_{rs,j}$ is element (r, s) in $\hat{\boldsymbol{\theta}}_j$, $\hat{\phi}_{rs,j}$ is element (r, s) in $\hat{\boldsymbol{\phi}}_j$, and $\mathbf{D}_{rs}$ is a $k \times k$ matrix with 1 in position (r, s) and 0 elsewhere.

From these expressions we can find asymptotic derivatives evaluated at $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = (\boldsymbol{\theta}, \boldsymbol{\phi})$. We introduce the following operator polynomials: $\boldsymbol{\lambda}(B) = \boldsymbol{\phi}^{-1}(B)\boldsymbol{\theta}(B)$; $\boldsymbol{\lambda}_0 = -\mathbf{I}$; $\boldsymbol{\eta}_{rs}(B) = \boldsymbol{\theta}^{-1}(B)\mathbf{D}_{rs}\boldsymbol{\phi}^{-1}(B)\boldsymbol{\theta}(B)$; $\boldsymbol{\eta}_{rs,0} = -\mathbf{D}_{rs}$; $\boldsymbol{\psi}_{rs}(B) = \boldsymbol{\theta}^{-1}(B)\mathbf{D}_{rs}$; $\boldsymbol{\psi}_{rs,0} = -\mathbf{D}_{rs}$; and by using the fact that $\mathrm{E}(\mathbf{a}_{t-i}\mathbf{a}_{t-j}^T) = \boldsymbol{\Omega}$, $i = j$, and zero otherwise, we find:

$$\partial\hat{\mathbf{R}}_i/\partial\hat{\theta}_{rs,j} \simeq -\boldsymbol{\Omega}\,\boldsymbol{\psi}_{rs,i-j}{}^T, \quad j \leq i, \text{ 0 otherwise,} \tag{3.6}$$

$$\partial\hat{\mathbf{R}}_i/\partial\hat{\phi}_{rs,j} \simeq \boldsymbol{\Omega}\,\boldsymbol{\eta}_{rs,i-j}{}^T, \quad j \leq i, \text{ 0 otherwise,} \tag{3.7}$$

$$\partial\hat{\mathbf{G}}_i/\partial\hat{\theta}_{rs,j} \simeq \begin{cases} \displaystyle\sum_{\tau=0}^{\infty} \boldsymbol{\lambda}_\tau\,\boldsymbol{\Omega}\,\boldsymbol{\psi}_{rs,\tau+j-i}{}^T, & j \leq i, \\ \displaystyle\sum_{\tau=0}^{\infty} \boldsymbol{\lambda}_{\tau+j-i}\,\boldsymbol{\Omega}\,\boldsymbol{\psi}_{rs,\tau}{}^T, & j > i, \end{cases} \tag{3.8}$$

$$\partial\hat{\mathbf{G}}_i/\partial\hat{\phi}_{rs,j} \simeq \begin{cases} -\displaystyle\sum_{\tau=0}^{\infty} \boldsymbol{\lambda}_\tau\,\boldsymbol{\Omega}\,\boldsymbol{\eta}_{rs,\tau+i-j}{}^T, & j < i \\ -\displaystyle\sum_{\tau=0}^{\infty} \boldsymbol{\lambda}_{\tau+j-i}\,\boldsymbol{\Omega}\,\boldsymbol{\eta}_{rs,\tau}{}^T, & j > i. \end{cases} \tag{3.9}$$

Finally, we find the asymptotic covariance matrix of the elements of $\mathbf{V}_n(\boldsymbol{\alpha})$, that is, $\mathbf{C}(\boldsymbol{\alpha})/n$. By writing out the variances and covariances of the columns of $\mathbf{R}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$, $i = 1, q$, and of $\mathbf{G}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$, $i = 1, p$, we can construct the desired matrix. Let $\mathbf{C}(\boldsymbol{\alpha}) = \{\mathbf{C}_{i,j}(\boldsymbol{\alpha})\}$, where $\mathbf{C}_{i,j}(\boldsymbol{\alpha})$, $i = 1, q + p$, $j = 1, q + p$, are $k^2 \times k^2$ matrices. We then find:

for $1 \leq i \leq q$, $1 \leq j \leq q$,

$$\begin{aligned} \mathbf{C}_{i,j}(\boldsymbol{\alpha}) &= n \cdot \mathrm{cov}(\mathrm{vec}\,\mathbf{R}_i, \mathrm{vec}\,\mathbf{R}_j) \\ &\simeq \boldsymbol{\Omega} \otimes \boldsymbol{\Omega} \quad \text{if } i = j, \text{ 0 otherwise,} \end{aligned} \tag{3.10}$$

for $1 \leq i \leq q$, $q + 1 \leq j \leq q + p$,

$$\begin{aligned} &\mathbf{C}_{i,j}(\boldsymbol{\alpha}) = n \cdot \mathrm{cov}(\mathrm{vec}\,\mathbf{R}_i, \mathrm{vec}\,\mathbf{G}_l) \\ &\simeq -\boldsymbol{\Omega} \otimes (\boldsymbol{\Omega}\,\boldsymbol{\lambda}_{i-l}{}^T), \quad i \geq l = j - q, \text{ 0 otherwise,} \end{aligned} \tag{3.11}$$

for $q + 1 \leq i \leq q + p$, $q + 1 \leq j \leq q + p$,

$$\begin{aligned} \mathbf{C}_{i,j}(\boldsymbol{\alpha}) &= n \cdot \mathrm{cov}(\mathrm{vec}\,\mathbf{G}_\nu, \mathrm{vec}\,\mathbf{G}_l) \\ &\simeq \boldsymbol{\Omega} \otimes \left(\sum_{\tau=0}^{\infty} \boldsymbol{\lambda}_\tau\,\boldsymbol{\Omega}\,\boldsymbol{\lambda}_{\tau+|\nu-l|}{}^T \right), \\ &\qquad \nu = i - q, \; l = j - q, \end{aligned} \tag{3.12}$$

where $\boldsymbol{\lambda}(B) = \boldsymbol{\phi}^{-1}(B)\boldsymbol{\theta}(B)$, as before.

Illustration. The results of this section can be illustrated by a small example. We consider a univariate mixed autoregressive moving average model of order (1, 1) and we compare the estimates with the corresponding maximum likelihood estimates. In the univariate case the formulas become very simple, in particular $\boldsymbol{\Omega} = \sigma^2$, $r = s = 1$, $\mathbf{D}_{rs} = 1$ in (3.6)–(3.8) and all the quantities in (3.6)–(3.12) are scalars.

Suppose $y_t = \phi y_{t-1} + a_t - \theta a_{t-1}$ and is univariate. By elementary calculations we find that

$$\mathbf{D}(\theta, \phi) = \sigma^2 \begin{bmatrix} 1 & -1 \\ (1 - \theta^2)/(1 - \phi\theta) & -(1 - \phi\theta)/(1 - \phi^2) \end{bmatrix},$$

$$\mathbf{C}(\theta, \phi) = \sigma^4 \begin{bmatrix} 1 & 1 \\ 1 & 1 + (\phi - \theta)^2/(1 - \phi^2) \end{bmatrix}.$$

This results in the following asymptotic covariance matrix for the estimates:

$$\mathrm{var}(\tilde{\theta}, \tilde{\phi}) \simeq \frac{1}{n(\phi - \theta)^2} \times \begin{bmatrix} (1 - \phi\theta)^2 & (1 - \phi\theta)(1 - \phi^2) \\ (1 - \phi\theta)(1 - \phi^2) & (1 - \phi^2)(1 + \theta^2 - 2\phi\theta) \end{bmatrix}.$$

In Box and Jenkins (1976, p. 246) the covariance matrix of the maximum likelihood estimates is given,

$$\mathrm{var}(\hat{\theta}, \hat{\phi}) \simeq \frac{1}{n(\phi - \theta)^2} \times \begin{bmatrix} (1 - \theta^2)(1 - \phi\theta)^2 & (1 - \phi^2)(1 - \theta^2)(1 - \phi\theta) \\ (1 - \phi^2)(1 - \theta^2)(1 - \phi\theta) & (1 - \phi^2)(1 - \phi\theta)^2 \end{bmatrix},$$

and it is interesting to note that the critical term is $(\phi - \theta)^2$ in both cases. The relative efficiency of the algorithm's estimate is

$$|\,\mathrm{var}(\hat{\theta}, \hat{\phi})\,| \,/\, |\,\mathrm{var}(\tilde{\theta}, \tilde{\phi})\,| = 1 - \theta^2.$$

It is also seen that the relative efficiency decreases near the invertibility boundary. This property has been found in other approximate estimation methods and it can cause problems when the data have been differenced or seasonally differenced. This is because a differencing operation tends to introduce one or several roots, respectively, near or on the invertibility boundary into the moving average part of the model.

The accuracy of the proposed approximate distribution has been investigated by numerical simulation. A univariate version of the algorithm, which uses "back-forecasting," was used. The study showed that the approximation is in fact very good, even for rather small samples. Some of these results are reported in Spliid (1980b).

4. COMPUTATIONAL SPEED FOR LARGE PROBLEMS

We indicate the computational advantages of the algorithm by comparing it with an algorithm that uses gradients to locate a maximum. As a measure we take the number of multiplications used by the algorithms. This choice of criterion is justified by the fact that, in the type of computations we are dealing with, the number of additions as well as loading and storing operations is largely proportional to the number of multiplications. This statement also holds for the more hidden operations, such as computation of indices. In general if a program makes use of, for example, mathematical functions or sorting, the number of multiplications is less useful as a computational indicator, but that is not the case in multivariate time series estimation.

Consider the model given in (1.2) and assume that the orders of the three polynomials in the model are p, r, and q, respectively. The dimension of $\mathbf{y}_t$ is k and for brevity the dimension of $\mathbf{x}_t$ is also k. The sample size is n. One recursive calculation of the residuals (Step 1) requires roughly $n(p + q + r)k^2$ multiplications.

An algorithm that uses gradients to locate the maximum of the likelihood function, for example, will generally need more iterations than parameters and in each iteration gradients have to be computed in as many directions as parameters. Hence an estimate of the necessary number of multiplications will be at least

$$M_{\text{grad}} = n(p + q + r)k^2 \times ((p + q + r)k^2)^2 \text{ multiplications} \quad (4.1)$$

if the likelihood is not too intricate. The realism of this figure may be judged by evaluating the number of multiplications needed to estimate the second derivative of the log-likelihood with respect to the parameter estimates, that is the Hessian matrix $\partial^2 L(\hat{\boldsymbol{\delta}})/(\partial\hat{\boldsymbol{\delta}}^T \partial\hat{\boldsymbol{\delta}})$, where $L(\cdot)$ is the log-likelihood. This matrix is used in most algorithms implemented so far. Numerical evaluation of the Hessian matrix at one point requires $1 + (p + q + r)k^2((p + q + r)k^2 + 3)/2$ values of L, each value requiring computation of new residuals. Hence an estimate of the Hessian requires more than $\frac{1}{2} \cdot M_{\text{grad}}$ multiplications. Therefore, since extra computations are needed to estimate where the optimum is, compute new values of L, and update the Hessian, it is not unrealistic to believe that the total number of multiplications required by these algorithms is of the same order of magnitude as M_{grad} in reasonably well-behaved cases. Otherwise it can be a good deal larger.

The proposed algorithm converges in most cases in less than 10 iterations regardless of the size of the model, especially if the moving average part is not overparameterized. In situations with overparameterization, maximum likelihood algorithms do not work either. In one iteration the algorithm computes new residuals once. In "Step 2" a least squares problem is solved. It consists of k subproblems with identical left sides and different right sides. Each subproblem is of order $(p + q + r)k$, and they can all be solved in approximately $(k(p + q + r))^3/3 + 2k(k(p + q + r))$ multiplications by elimination. Creation of the left side of (2.2) requires approximately $nk^2(p + 2q + r - 1)$ multiplications if the band structure is used and if products not involving $\hat{\mathbf{A}}$ are reused. The k right sides can be created in nk^2q multiplications or less. Therefore, if IT denotes the number of iterations, one concludes that the algorithm requires approximately

$$\begin{aligned} M_{\text{algor}} &= IT \cdot [nk^2(p + q + r) \\ &\quad + (k(p + q + r))^3/3 + 2k(k(p + q + r)) \\ &\quad + nk^2(p + 2q + r - 1) + nk^2q] \\ &\simeq IT \cdot nk^2(2p + 4q + 2r - 1) \quad \text{multiplications,} \end{aligned} \quad (4.2)$$

since in most cases $n \gg k(p + q + r)^2$ and $2k^2(p + q + r)$ is small.

If, for example, $IT = 10$, $k = 3$, and $p = q = r = 1$, the algorithm is probably more than 30 times faster than a maximum likelihood algorithm. If $k = 4$ and the other figures are unaltered, the algorithm can be around 100 times faster.

It should be added that all figures given in this section are only indicative. Clever programming as well as different sets of data may change the computational requirements of any algorithm. In the following section a few computational examples are presented and they do support the conjecture just given, namely that the algorithm is much faster than maximum likelihood procedures.

5. CASE STUDY

In the article by Tiao and Box (1981) an analysis of data given by Coen, Gomme, and Kendall (1969) is shown. The analysis concerns the possibility of predicting the Financial Time Ordinary Share Index $(y_{1,t})$ by the U.K. car production $(y_{2,t})$ and the Financial Time Commodity Price Index $(y_{3,t})$. For further description of the problem the reader is referred to the two articles mentioned. Here we only present the results obtained by the described algorithm.

The program documented in Spliid (1980a) was used to analyze the data and for comparison we give the result for the general ARMA(1, 1) model suggested by Tiao and Box.

The model is given by $\mathbf{z}_t = \boldsymbol{\phi}\mathbf{z}_{t-1} + \mathbf{a}_t - \boldsymbol{\theta}\mathbf{a}_{t-1}$, where $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are 3×3 matrices and $\mathbf{z}_t$ is the vector containing $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$, corrected for means and standard deviations. In six iterations the following model was obtained:

$$\tilde{\boldsymbol{\phi}} = \begin{bmatrix} .70 & .24 & -.05 \\ -.14 & 1.01 & -.08 \\ -.37 & .33 & .74 \end{bmatrix} \quad \text{and}$$

$$\tilde{\boldsymbol{\theta}} = \begin{bmatrix} -.18 & .22 & .15 \\ -.36 & .08 & -.08 \\ -.88 & .59 & -.44 \end{bmatrix}.$$

These results, as well as the corresponding residual covariance matrix, are essentially the same as those presented by Tiao and Box. Specifically, all 18 parameters have the same sign and magnitude. The analysis of Tiao and Box was performed by a program, WMTS-1, developed by Tiao and Box, et al. (1979). The program uses conditional maximum likelihood for initial estimation and exact maximum likelihood for final estimation.

A comparison was made between the WMTS-1 program and the perviously mentioned program. The cpu-time used for some typical estimation situations are given. The data were the Coen, Gomme, and Kendall data; the computer was an IBM 3033. In the three first cases no initial parameter estimates were supplied. In the two last cases the conditional likelihood method and the algorithm were used to find initial estimates, and the figure given is the total cpu time.

Estimation method	*cpu-time millisec.*
Algorithm:	200
Cond. likelihood:	6800
Exact likelihood:	16100
Cond. likelihood plus exact likel.:	13200
Algorithm plus exact likelihood:	6600

The interesting comparison is between the algorithm and the conditional likelihood method and in this particular study the algorithm is seen to be faster by a factor of 34. The close agreement with the factor 30 in Section 4 is a coincidence, but the result clearly demonstrates that the algorithm is much faster than the likelihood procedures.

In the article by Spliid, Jensen, and Pedersen (1981, p. 356) a four-dimensional ARMA(1, 1) model is described. For a series consisting of 200 observations of such data it turned out that the algorithm used only 1 cpu second while the conditional likelihood method used 50 seconds and the exact likelihood method used 3.5 minutes. The combined methods used around 2 minutes.

6. CONCLUSION

The proposed algorithm is believed to be an economic alternative for estimation of large time series models where maximum likelihood methods generally can be expensive to apply. If maximum likelihood estimates are wanted for a large model the algorithm may reduce the computations considerably by supplying good initial estimates. Last, but not least, it should be mentioned that in the model identification stage of the analysis of multivariate time series it is often desirable to estimate a number of alternative models and for this purpose the algorithm is very well suited.

The algorithm has been used successfully in a number of situations. Some of these are described by Clausen, Holst, and Spliid (1982) and by Spliid, Jensen, and Pedersen (1981).

[*Received March 1982. Revised February 1983.*]

REFERENCES

AKAIKE, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255–265.

——— (1974), "Markovian Representation of Stochastic Processes and its Applications to the Analysis of Autoregressive Moving Average Processes," *Annals of the Institute of Statistical Mathematics*, 26, 363–387.

BOX, G.E.P., and JENKINS, G.M. (1970), *Time Series Analysis, Forecasting and Control*, San Francisco: Holden-Day.

CLAUSEN, J., HOLST, J., and SPLIID, H. (1982), "Adaptive Prediction of High Tides," *Proceedings of the 1982 Nordic Symposium in Applied Statistics and Data Processing*, NEUCC Technical University of Denmark.

COEN, P.G., GOMME, E.D., and KENDALL, M.G. (1969), "Lagged Relationships in Economic Forecasting," *Journal of the Royal Statistical Society*, Ser. A, 132, 133–163.

GOODWIN, G.C., and PAYNE, R.L. (1977), *Dynamic System Identification, Experiment Design and Analysis*, New York: Academic Press.

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

——— (1976), "The Asymptotic Distribution of Serial Covariances," *Annals of Statistics*, 4, 396–399.

HANNAN, E.J., DUNSMUIR, W.T.M., and DEISTLER, M. (1980), "Estimation of Vector ARMAX Models," *Journal of Multivariate Analysis*, 10, 275–295.

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.

JENKINS, G.M., and ALAVI, A.S. (1981), "Some Aspects of Modeling and Forecasting Multivariate Time Series," *Journal of Time Series Analysis*, 2, 1–47.

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

PHADKE, M.S., and KEDEM, G. (1978), "Computation of the Exact Likelihood Function of Multivariate Moving Average Models," *Biometrika*, 65, 511–519.

PRIESTLEY, M.B. (1981), *Spectral Analysis and Time Series*, Vol. 2, New York: Academic Press.

SPLIID, H. (1980a), "MARIMA, Estimation of Multivariate Time Series and Transfer Function Models," *Computer Program Documentation*, Technical University of Denmark, Institute of Mathematical Statistics and Operations Research (IMSOR).

——— (1980b), "Estimation of Time Series Models by a Regression

Analysis Algorithm," Research Report, Technical University of Denmark, Institute of Mathematical Statistics and Operations Research (IMSOR).

SPLIID, H., JENSEN, S.M., and PEDERSEN, S.B. (1981), "Empirical Models for the Spreading of Spills in Marine Environments," in *Proceedings of the 4th International Time Series Meeting*, ed. O.D. Anderson, New York: North-Holland.

TIAO, G.C., and BOX, G.E.P. (1981), "Modeling Multiple Time Series with Applications," *Journal of the American Statistical Association*, 76, 802–816.

TIAO, G.C., BOX, G.E.P., GRUPE, M.R., HUDAK, G.B., BELL, W.R., and CHANG, I. (1979), "The Wisconsin Multiple Time Series Program (WMTS-1): A Preliminary Guide," University of Wisconsin, Dept. of Statistics.

WHITTLE, P. (1951), *Hypothesis Testing in Time Series*, Uppsala: Almquist and Wiksell.

——— (1952), "Some Results in Time Series Analysis," *Scandinavisk Aktuarie Tidsskrift*, 35, 48–60.

——— (1953), "The Analysis of Multiple Stationary Time Series," *Journal of the Royal Statistical Society*, Ser. B, 15, 125–139.

WILSON, G.T. (1973), "The Estimation of Parameters in Multivariate Time Series Models," *Journal of the Royal Statistical Society*, Ser. B, 40, 76–85.