THE ESTIMATION OF MIXED REGRESSION, AUTOREgressive, MOVING AVERAGE, AND DISTRIBUTED LAG MODELS

BY E. J. HANNA N AND D. F. NICHOLLS

In this paper a procedure is presented for the (asymptotically) efficient estimation of the parameters of autoregressive moving average models with exogenous variables. The estimation of distributed lag models is discussed as a particular case of models of this form. Finally a number of numerical examples are described.

1. INTRODUCTION

In this paper, which is closely related to Hannan [4], we shall deal with the estimation of the parameters in the model

\[ \sum_{j=0}^{q} \beta(j)y(n-j) + \sum_{j=1}^{r} \delta(j)x_j(n) = \sum_{j=0}^{p} \alpha(j)x(n-j), \quad \alpha(0) = \beta(0) = 1, \]

assuming that we have observations for \( n = 1, \ldots, N \) on \( x_j(n) \) and \( y(n) \). We also make the following assumptions:

(i) The \( \epsilon(n) \) are i.i.d. \( (0, \sigma^2) \);

(ii) all zeros of the \( z \) transforms

\[ g(z) = \sum_{j=0}^{p} \alpha(j)z^j, \quad h(z) = \sum_{j=0}^{q} \beta(j)z^j \]

lie outside of the unit circle; and

(iii) the \( x_j(n) \) come from infinite sequences which satisfy, almost surely,

\[ \lim_{N \to \infty} N^{-1} \sum_{m=1}^{N} x_j(m)x_k(m + n) = \gamma_{jk}(n) = \gamma_{jk}(-n) \quad (j, k = 1, \ldots, r; n = 0, 1, \ldots), \]

independently of the \( \epsilon(n) \) sequence.

As a consequence of (iii) we have

\[ \gamma_{jk}(n) = \int_{-\pi}^{\pi} e^{in\lambda} dF_{jk}(\lambda). \]

The matrix \( F(\lambda) \) with \( F_{jk}(\lambda) \) as the entry in row \( j \), column \( k \) is, of course, a spectral distribution matrix. We also assume that (1) is identified. This question is fairly fully discussed in [5].

The condition (ii) is reasonable in the following sense. If the condition on \( g(z) \) is not imposed, then it is possible to re-express the right hand side of (1) in terms of a new orthogonal sequence with zero mean and constant variance, which we call \( \varepsilon_4(n) \), so that the right hand side of (1) may be written

\[ \sum_{j=0}^{p} \alpha(j)\varepsilon_4(n-j), \quad p_1 \leq p, \]

1 The research described in this paper was carried out under grants from the National Science Foundation and from the Ford Foundation.
and for which \( g_1 = \sum \varepsilon_1(j)z^j \) has all of its zeros on or outside of the unit circle. Moreover the \( \varepsilon_1(n) \) will be the linear, one step, prediction errors for the right hand side of (1) so that they will be the linear prediction errors for \( y(n) \) also when the \( x_j(n) \) sequences are known. If the data is Gaussian, then the \( \varepsilon_1(n) \) sequence will be independent also, but in any case it seems as reasonable to hope that they are independent as to hope that the \( \varepsilon(n) \) are so. The exclusion of zeros on the unit circle from \( g(z) \) is required because the other case causes difficulty with the method which we present. It is felt that if there is a zero on the unit circle the location of this zero will be known beforehand, in which case the method we introduce below can be modified to allow for it. The second requirement in (ii) is necessary if there is to exist a stationary solution, \( y(n) \), to (1), which is expressible in terms of the \( \varepsilon(n - k) \), \( x_j(n - k) \), \( k \geq 0 \). (Of course the possibility of stationarity will depend on the \( x_j(n) \) sequences also.)

It seems preferable to treat the \( x_j(n) \) as fixed sequences as in (iii), rather than to prescribe them stochastically as this makes the treatment less restrictive. One could begin from a more general specification such that the \( y(n) \) and \( x_j(n) \) in (1) are residuals from a regression. For example we might have

\[
y(n) = \tilde{y}(n) + \mu z(n)
\]

where \( \mu \) is a vector of regression coefficients and \( z(n) \) is a vector of exogenous variables, so that it is \( \tilde{y}(n) \) which is observed. Under reasonable conditions on the \( x_j(n) \) in (1) it will be possible to remove the effects of \( z(n) \) first by regression and then to work with the estimates of \( y(n) \) obtained from the computed regression, and the results of this paper will continue to hold. In this way non-stationary (trend) components in \( \tilde{y}(n) \) can be allowed for. We shall not deal with this explicitly below except for the case where \( z(n) \equiv 1 \) (i.e., except for a mean correction). Of course such a \( z(n) \) could also be incorporated among the \( x_j(n) \), but it will be equally as efficient and computationally simpler to account for it by a mean correction.

The equation (1) may be regarded as an alternative form of the distributed lag model

(2) \[
y(n) = \mu(n) + \sum_{j=1}^{r} \delta_j \sum_{k=0}^{\infty} a_k x_j(n - k) + u(n)
\]

where

\[
\sum_{k=0}^{\infty} a_k z^k = h(z)^{-1}, \quad \sum_{j=0}^{q} \beta(j)u(n - j) = \sum_{j=0}^{p} \alpha(j)\varepsilon(n - j),
\]

and \( \mu(n) \) is a solution of the homogeneous equation

\[
\sum_{j=0}^{q} \beta(j)\mu(n - j) = 0.
\]

This solution decays to zero at an exponential rate as \( n \) increases. In particular we might have \( x_j(n) = x(n - j) \) in which case we are dealing with the general
rational distributed lag model (we put \( \mu(n) \equiv 0 \)),

\[
y(n) = \delta \sum_{\ell=0}^{\infty} \lambda_{\ell} x(n - \ell) + u(n),
\]

with \( u(n) \) as before. Of course the model (1) (or its equivalent form (2)) covers also the case of more than one exogenous variable, but the generating functions (3) for these different variables will all have the same denominator. (More general cases can be dealt with but lead to very complex calculations.) A special case is that for which \( p = q \) and \( \alpha(j) \equiv \beta(j) \) so that \( u(n) = \alpha(n) \). We shall discuss this special case further in Section 3. In that section we shall also discuss problems associated with another way in which there might be unrecognized restrictions on the \( \alpha(j) \) and \( \beta(j) \), namely that for which \( \alpha(p) \) or \( \beta(q) \) is zero.

Finally, in this section we shall describe the (restricted) sense in which we shall speak of (asymptotically) efficient estimates. If the \( \epsilon(n) \) and \( x(n) \) are Gaussian and the \( x(j) \) processes are suitably restricted, we may set up the likelihood function and obtain the maximum likelihood estimators. With small modifications these estimators may be shown to have a limiting normal distribution, after appropriate normalization. (This is discussed in [8].) He deals basically with the univariate case without exogenous variables, but it is clear that the results extend to our case. For example the covariance matrix for the estimators \( \hat{\alpha} \), \( \hat{\beta} \) of \( \alpha \) and \( \beta \) in the model (2) becomes

\[
\left\{ \frac{1}{\sigma^2} \left[ \begin{array}{c}
\beta e^{i\lambda} \\
1 - \alpha e^{i\lambda}
\end{array} \right] \left[ \begin{array}{cc}
1 & \beta e^{i\lambda} \\
1 - \alpha e^{i\lambda} & \beta^2
\end{array} \right]^{-1} dF(\lambda) \right\}^{-1}.
\]

The existing maximum likelihood results would cover the case of only rather special \( F(\lambda) \). We shall establish central limit theorems for our estimates, under the conditions (i), (ii), and (iii), and shall speak of them as asymptotically efficient if in their limiting distribution they have (4) as covariance matrix (or have the analogous matrix as covariance matrix if, for example, the more general model (1) is being considered).

The likelihood equations are non-linear and are difficult to solve. One procedure, described in Section 2, is based on Fourier transformation of the data. This (approximately) reduces the likelihood function to a simpler form, as described in [3]. Adopting this approximate likelihood function we obtain a sequence of iterations that, no doubt, converge in some suitable sense to the solutions of the (approximate) likelihood equations. The technique for finding these iterations is related to the method of scoring (see [6, p. 49]) and is described in [4] and further discussed at the end of this paper. In Section 2 we have, for the most part, presented these iterations in real terms near to those suitable for computing. They are more
perspicuous in complex variable form, but to reduce the amount of detail in the
paper we have not shown the complex variable formula. As we have already indi-
cated we do not show that these iterations converge to the maximum likelihood
solution. What we do show, in Theorem 1, is that the results at the jth iteration
are asymptotically efficient (for any j) in the sense described above.

2. THE ESTIMATION PROCEDURE AND THE CENTRAL LIMIT THEOREM

The basic statistics on which our computations are based are the finite Fourier
transforms

\[ w_j(\omega_t) = \frac{1}{\sqrt{2\pi N}} \sum_{n=1}^{N} y(n) e^{i\omega n}, \qquad w_j(\omega_t) = \frac{1}{\sqrt{2\pi N}} \sum_{n=1}^{N} x_j(n) e^{i\omega n}, \]

\[ \omega_t = 2\pi t/N, \quad (t = 1, \ldots, [N/2]; j = 1, \ldots, r). \]

We also put \( w(\omega_t) = w(\omega_t) \) and

\[ \sqrt{2\pi N} w_j(\omega_t) = u_j(\omega_t) + iv_j(\omega_t), \quad \sqrt{2\pi N} w(\omega_t) = u(\omega_t) + iv(\omega_t) \]

so that, for example,

\[ u_j(\omega_t) = \sum_{n=1}^{N} x_j(n) \cos n\omega_t. \]

We shall often omit the \( \omega_t \) argument, even in summations, for simplicity of printing.
It is the \( u_j, v_j, v_j, v_j \) which are the basic building blocks and the \( w_j, w_j \) are introduced
so that formulae can be written more perspicuously. We have omitted the terms for \( t = 0 \) since these are annihilated by mean correction while the others are not
affected by this. We put

\[ I_j = |w_j|^2 = (2\pi N)^{-1}(u_j + v_j), \]

\[ I_j = w_j \bar{w}_j = (2\pi N)^{-1}(c_j - iq_j), \quad c_j = u_j \mu_j + v_j \nu_j, \quad q_j = u_j \nu_j - v_j \mu_j, \]

\[ I_{jk} = w_j \bar{w}_k = (2\pi N)^{-1}(c_{jk} - iq_{jk}), \quad c_{jk} = u_j \mu_k + v_j \nu_k, \quad q_{jk} = u_j \nu_k - v_j \mu_k. \]

We arrange the \( r \) quantities \( I_j \) in a column vector \( I \) and the \( r^2 \) quantities \( I_{jk} \) in the
matrix \( I_x \). Then

\[ I = (2\pi N)^{-1}\{C - iQ\}, \quad I_x = (2\pi N)^{-1}\{C_x - iQ_x\} \]

where \( C, C_x, Q, Q_x \) are defined in terms of \( c_j, c_{jk}, q_j, q_{jk} \), as were \( I \) and \( I_x \) in terms
of \( I_j, I_{jk} \).

We also need the serial covariances

\[ \hat{c}(n) = \hat{c}(-n) = (N - n)^{-1} \sum_{m=1}^{N-n} (y(m) - \bar{y})(y(m + n) - \bar{y}) \]

\[ (n = 0, 1, \ldots, p + q), \]
\[ \hat{c}_k(n) = \hat{c}_k(-n) = (N - n)^{-1} \sum_{m=1}^{N-n} (x(m) - \bar{x})(x(m + n) - \bar{x}) \]

\[ (j, k = 1, \ldots, r; n = 0, 1, \ldots, p), \]

\[ \hat{c}_y(n) = \hat{c}_y(-n) = (N - n)^{-1} \sum_{m=1}^{N-n} (y(m) - \bar{y})(x(m + n) - \bar{x}) \]

\[ (j = 1, \ldots, r; n = 0, 1, \ldots, p + q). \]

Our estimation procedure consists of two steps, of which only the second is iterated.

**Step 1.** We may estimate \( \beta, \delta \) by means of

\[ \sum_{k=1}^{q} \hat{c}_k(p + j - k)\hat{\beta}(k) + \sum_{k=1}^{r} \hat{c}_k(p + j)\hat{\delta}(k) = -\hat{c}_j(p) \quad (j = 1, \ldots, q), \]

and

\[ \sum_{k=1}^{q} \hat{c}_y(k)\hat{\beta}(k) + \sum_{k=1}^{r} \hat{c}_y(0)\hat{\delta}(k) = -\hat{c}_y(0) \quad (j = 1, \ldots, r; \hat{\beta}(0) = 1). \]

It is convenient to introduce the partitioned (row) vectors

\[ \rho' = (\beta'; \delta'), \quad \rho' = (\hat{\beta}'; \hat{\delta}') \]

where \( \beta \) has \( \beta(j) \) in the \( j \)th place, \( j = 1, \ldots, q \) (and similarly for \( \hat{\beta} \)) and \( \delta \) has \( \delta(j) \) in the \( j \)th place, \( j = 1, \ldots, r \) (and similarly for \( \hat{\delta} \)). We may thus write the equations for \( \hat{\beta}, \hat{\delta} \) as

\[ \hat{\rho} = -\hat{C}^{-1}\hat{c}. \]

We observe that \( \hat{C} \) converges almost surely\(^2\) to the matrix

\[ \hat{C} = \begin{bmatrix} \int_{-\pi}^{\pi} e^{i(j \lambda)} \frac{\sigma^2 |g|^2}{2\pi h} d\lambda + \frac{\delta' F(\lambda)\delta}{h} & -\int_{-\pi}^{\pi} e^{i(p + j)\lambda} d\lambda \int_{-\pi}^{\pi} \delta' F(\lambda) \frac{e^{-ik\lambda}}{h} d\lambda \\
-\int_{-\pi}^{\pi} e^{-ik\lambda} F(\lambda) d\lambda & \int_{-\pi}^{\pi} d\lambda \end{bmatrix}. \]

Here \( j, k \) vary from 1 to \( q \) so that the top left hand corner part is a \( q \times q \) matrix. Since \( F(\lambda) \) is \( r \times r \), the top right hand corner is \( q \times r \) and the bottom right hand corner is \( r \times r \). We have omitted the argument \( \exp i\lambda \) from \( g \) and \( h \), for convenience.

Our final assumption is:

(iv) \( \hat{C} \) is non-singular.

\(^2\) In future we shall often omit the qualification "almost surely" when we speak of this mode of convergence.
For given \( F, g, h \) the vectors \( \delta \) making \( C \) singular will lie in an \((r - 1)\) dimensional subset so that a priori it seems most unlikely that \( C \) will be singular. In any case there is no shortage of “instrumental variables” to replace some or all of the \( y(n - p - j), j = 1, \ldots, q; x_j(n), j = 1, \ldots, p \). In particular the \( x_j(n - k), k > 0 \), could be used to replace the \( y(n - p - j) \). The use of these would have one decided virtue, namely that the equations give valid estimates independent of the size of \( p \) so that if \( p \) is to be increased in a later calculation, the initial estimate of \( \rho \) does not have to be recomputed. Thus if \( r = 1 \), for example, we might choose to replace the first \( q \) equations defining \( \tilde{\rho} \) by

\[
\sum_{j=1}^{q} \tilde{\epsilon}_x(k - j) \tilde{\rho}(k) + \tilde{\epsilon}_x(1) \tilde{\rho}(1) = -\tilde{\epsilon}_x(j) \quad (j = 0, \ldots, q; \tilde{\rho}(0) = 1).
\]

For \( r > 1 \) if one wishes to use the \( x_j(n - k) \), only, as instruments, one will have to choose which set to adopt. Presumably, the maximum lag will be kept as low as possible. We emphasize that Theorem 1 below will hold in its entirety no matter what system of equations is used out of those just discussed, provided only that (iv) is replaced by an equivalent condition ensuring that these initial equations for \( \rho \) have, asymptotically, a unique solution.

Once having obtained our first estimate \( \tilde{\rho} \), we next form

\[
\hat{h}(e^{j\omega t}) = \sum_{j=0}^{q} \tilde{\beta}(j) e^{j\omega_t} \quad (t = 1, \ldots, \lfloor \frac{1}{2} N \rfloor),
\]

where, let us say, \( \tilde{h} = \tilde{\xi}_h + i \tilde{\eta}_h \). We also form the autocovariances

\[
\tilde{\epsilon}_u(n) = \sum_{j=0}^{q} \tilde{\beta}(j) \tilde{\epsilon}(n + k - j) + \sum_{j,k=1}^{r} \tilde{\delta}(j) \tilde{\delta}(k) \tilde{\epsilon}_u(n) + \sum_{j=0}^{q} \tilde{\beta}(j) \tilde{\delta}(k) \{ c_j(n + j) + c_k(n - j) \} \quad (n = 0, 1, \ldots, p).
\]

From these we obtain

\[
\hat{f}_w(\omega_t) = \frac{1}{2\pi} \left\{ \tilde{\epsilon}_u(0) + 2 \sum_{n=1}^{p} \tilde{\epsilon}_u(n) \cos n \omega_t \right\} = \hat{f}_w(-\omega_t) \quad (t = 0, 1, \ldots, \lfloor \frac{1}{2} N \rfloor).
\]

The function \( \hat{f}_w \) estimates the spectrum of

\[
w(n) = \sum_{j=0}^{p} \alpha(j) e^{j\omega_t(n - j)}.
\]

It thus estimates a positive function. Nevertheless it may be negative, perhaps for a value other than some \( \omega_t \). If and only if it is positive may it be, uniquely, factored in the form

\[
\hat{f}_w = \frac{\hat{\alpha}^2}{2\pi} \sum_{j=0}^{p} \hat{\alpha}(j) e^{j\omega_t} = \frac{\hat{\alpha}^2}{2\pi} |\hat{\alpha}|^2, \quad \hat{\alpha}(0) = 1,
\]
where all zeros of \( \hat{g}(z) \) lie outside of the unit circle. We shall call \( \hat{a} \) the vector with \( \hat{a}(j) \) in the \( j \)th place, \( j = 1, \ldots, p \). Since the factorization may be troublesome (or impossible), we now show how to avoid it. We introduce the symbol \( \kappa_t, t = 1, \ldots, N, \) which is to be 1 unless \( t = \frac{1}{2} N \) (i.e., \( N \) even), when it is to be \( \frac{1}{2} \), and form

\[
\hat{a}(k) = \frac{2}{N} \sum_{t=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t \left( \cos k\omega_t | \hat{h}_{w_t} + \sum_{j=1}^{r} \delta(j)w_j \right)^2 \right)^{1/2} \hat{f}_w
\]

\[
= (4\pi/N^2) \sum_{t=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t \left( \cos k\omega_t | \hat{h}_{w_t} 2\pi N I_p + \delta^2 C_x \hat{\delta} + \delta' C \hat{\delta}' \right) (2\pi\hat{f}_w)^2 \quad (k = 0, \ldots, p).
\]

The second formula is written in a form that eliminates factors the introduction of which causes needless computing and which are absorbed into the factor \( 4\pi/N^2 \).

Now let \( \hat{A} \) have \( \hat{a}(k - l) \) in row \( k \), column \( l \), \( k, l = 1, \ldots, p \), and \( \hat{a} \) have \( \hat{a}(k) \) in row \( k, k = 1, \ldots, p \), and put

\[
\hat{a} = -\hat{A}^{-1} \hat{a}.
\]

We have here used \( \hat{a} \) as a name for two separate estimates of \( a \), but since only one of these will be computed and they occupy the same place in later formulæ, no confusion will result. This completes Step 1, which is not repeated in later iterations.

**Step 2.** From the \( \hat{a}(j) \) we have

\[
\hat{g} = \sum_{l=0}^{p} \hat{a}(j) e^{j\omega_l} = \hat{\xi}_g + i\hat{\eta}_g,
\]

let us say. If \( \hat{f}_w \) was not factored, we estimate \( \sigma^2 \) by

\[
\sigma^2 = \frac{4\pi}{N} \sum_{t=0}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t \left( \hat{f}_w | \hat{g} \right)^2
\]

which is, of course, in agreement with the definition when \( \hat{f}_w \) is factored.

We next form

\[
\hat{\delta}_{kl} = (2/N^2) \sum_{t=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t \left( 2\pi N I_p \cos (k - l)\omega_t \right) | \hat{g} \right)^2 \quad (k, l = 1, \ldots, q).
\]

\[
= (2/N^2) \sum_{t=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t \left( c_j \cos k\omega_t + q_j \sin k\omega_t \right) | \hat{g} \right)^2 \quad (k = 1, \ldots, q; l = q + j; j = 1, \ldots, r).
\]

\[
= (2/N^2) \sum_{t=1}^{\left\lfloor \frac{N}{2} \right\rfloor} \kappa_t c_{ij} | \hat{g} \right)^2 \quad (k = q + i; l = q + j; i, j = 1, \ldots, r).
\]
Finally $d_{k,k} = d_{k,l}$, which completes the definition (with some redundancy). We define $\hat{\sigma}^2 \hat{D}$ to have $\hat{d}_{k,l}$ in row $k$, column $l$, $l = 1, \ldots, q + r$. We define $\hat{d}^2 \hat{d}$ to have $\hat{d}_{k,l}$ in row $k$ where $\hat{d}_{k,l} = d_{k,l}, k = 1, \ldots, q + r$. Thus,

$$d_{k,l} = \frac{1}{2N^2} \sum_{i=1}^{N} \kappa_i \frac{2\pi N}{|\hat{g}|^2} \cos k o_i / |\hat{g}|^2 \quad (k = 1, \ldots, q),$$

$$d_{g+j,l} = \frac{1}{2N^2} \sum_{i=1}^{N} \kappa_i \frac{2\pi N}{|\hat{g}|^2} \quad (g = j; j = 1, \ldots, r).$$

We now put

$$\hat{d}^{11} = -\hat{D}^{-1} \hat{d},$$

so that the factor $\hat{d}^2$ does not need to be introduced (or the factors $2/N^2$). We have defined $\hat{D}$ and $\hat{d}$ in the above manner because this definition is needed later.

We now use $\hat{d}^{11}$ to compute $\hat{h}^{11} = \hat{h}_s^{11} + i\hat{h}_v^{11}$ as we used $\hat{d}$ to compute $\hat{h}$ in (5), and we then compute $\hat{A}^{11}, \hat{a}^{11}$ in the same way as we computed $\hat{A}, \hat{a}$, in and below (6), but using $\hat{h}^{11}, \hat{d}^{11}$, and $\hat{a}^2 / |\hat{g}|^2 / 2\pi$ in place of $f_n$. In computing the $\hat{d}^{11}(k)$, we may omit all constant factors, such as $\hat{d}^2 / 2\pi$, but it is convenient to retain the definition as given. We now form

$$\hat{a}^{11} = -\hat{A}^{11}^{-1} \hat{d}^{11}.$$

We next compute $\hat{\Omega}$ whose typical entry is $\hat{\sigma}(k-l)$, in row $k$, column $l$, where $k = 1, \ldots, q$; $l = 1, \ldots, p$, and

$$\hat{\sigma}(k) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2\pi} e^{-ik\omega_i}, \quad -p + 1 \leq k \leq q - 1,$$

$$= \frac{1}{N} \sum_{i=1}^{N} \kappa_i \left\{ (\hat{\xi}_s \hat{\xi}_h + \hat{\eta}_s \hat{\eta}_h) \cos k \omega_i + (\hat{\xi}_s \hat{\eta}_h - \hat{\eta}_s \hat{\xi}_h) \sin k \omega_i \right\} / |\hat{g}|^2 / |\hat{h}|^2.$$

If $p$ and $q$ are not large, it may be easier to compute $\hat{\sigma}(k)$ from the following formula for it, namely,

$$\hat{\sigma}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \hat{g}(e^{ik}) \hat{h}(e^{-ik}) \right\}^{-1} e^{-i(k \lambda)} \hat{d}_\lambda.$$

For $p \gg q$ this gives

$$\hat{\sigma}(k) = \hat{\sigma}(p)^{-1} \sum_{j=1}^{p} \left\{ \hat{h}(\xi_j^{-1})^{-1} \xi_j^{-k-1} \prod_{i=1}^{p} (\xi_i - \xi_j)^{-1} \right\},$$

where $\xi_j$ are the zeros of $\hat{g}(z)$. For $q \gg p$ we may use

$$\hat{\sigma}(k) = \hat{\sigma}(q)^{-1} \sum_{j=1}^{q} \left\{ g(z_j^{-1})^{-1} z_j^{-k-1} \prod_{i=1}^{q} (z_i - z_j)^{-1} \right\},$$

and where $\xi_j$ are the zeros of $\hat{g}(z)$. For $q \gg p$ we may use

$$\hat{\sigma}(k) = \hat{\sigma}(q)^{-1} \sum_{j=1}^{q} \left\{ g(z_j^{-1})^{-1} z_j^{-k-1} \prod_{i=1}^{q} (z_i - z_j)^{-1} \right\},$$

where $\xi_j$ are the zeros of $\hat{g}(z)$. For $q \gg p$ we may use

$$\hat{\sigma}(k) = \hat{\sigma}(q)^{-1} \sum_{j=1}^{q} \left\{ g(z_j^{-1})^{-1} z_j^{-k-1} \prod_{i=1}^{q} (z_i - z_j)^{-1} \right\},$$

where $\xi_j$ are the zeros of $\hat{g}(z)$. For $q \gg p$ we may use

$$\hat{\sigma}(k) = \hat{\sigma}(q)^{-1} \sum_{j=1}^{q} \left\{ g(z_j^{-1})^{-1} z_j^{-k-1} \prod_{i=1}^{q} (z_i - z_j)^{-1} \right\},$$

where $\xi_j$ are the zeros of $\hat{g}(z)$. For $q \gg p$ we may use
where the $z_j$ are the zeros of $\hat{h}(z)$. In particular, for $p = q = 1$ we get $\hat{\alpha}(0) = \{1 - \hat{\alpha}(1)\hat{\beta}(1)\}^{-1}$. If $p = 2, q = 1$, and $\xi$ is complex, then

$$
\hat{\alpha}(0) = \{\hat{\alpha}(2)\xi + \hat{\beta}(1)\}^{-1}
$$

while $\hat{\alpha}(1) = -2\hat{\alpha}(0)\hat{\beta}(1)x_1y_1$ where $\xi = x_1 + iy_1$. It is not difficult to derive simple explicit formulae in any case where $p$ and $q$ are small, as we have said.

Now

$$
\hat{\alpha}^{(1)} = \left\{1_p - (2\pi/\hat{\alpha}^2)\hat{A}^{-1}\left[\hat{\Omega}'0\right]\hat{B}^{-1}\left[\begin{array}{c} \hat{\Omega} \\ 0 \end{array}\right]\right\}^{-1} (\hat{\alpha} - \hat{\alpha}^{(1)}) + \hat{\alpha}.
$$

If $f_w$ is factored then $\hat{A}$ will not have been computed, in which case it may be replaced by $\hat{A}^{(1)}$. For computations we may cancel the factor $\hat{\beta}^{-2}$ occurring before $\hat{A}$ with the factor $\hat{\alpha}^2$ introduced in the definition of $\hat{B}$. Finally

$$
\hat{\beta}^{(1)} = -\hat{B}^{(1)-1}\hat{\beta}^{(1)}
$$

where the $\hat{\alpha}_{kl}^{(1)}, \hat{\alpha}_k^{(1)}$ are defined as were $\hat{\alpha}_{kl}, \hat{\alpha}_k$ but using $\hat{\beta}^{(1)} = \Sigma\hat{\beta}^{(1)}(k)\exp ik\omega = \hat{\xi}_h + \hat{\eta}_h^{(1)}$ in place of $\hat{\beta}$.

We shall shortly state a theorem which establishes the asymptotic efficiency of $\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}$ but nevertheless it will often be necessary to iterate Step 2. Thus we commence from $\hat{g}^{(1)}, \hat{\xi}_h^{(1)}, \hat{\beta}^{(1)}, \hat{\beta}^{(1)}$, this last being obtained via (5) but using $\hat{\beta}^{(1)}$ in place of $\hat{\beta}$. Of course $\hat{\xi}^{(1)} = \hat{\xi}_h^{(1)} + \hat{i}\eta_h^{(1)}$. The estimate $\hat{\alpha}^2$ in (7) is now replaced by

$$
\hat{\alpha}^{(2)} = \frac{4\pi\hat{\beta}^{(1)}}{N}\sum_{n=1}^{N}\kappa(n)\hat{f}_w^{(1)}/\hat{\beta}^{(1)}\hat{\beta}^{(1)}
$$

where we have used $\hat{\beta}^{(1)}$, $\hat{\beta}^{(1)}$ to form the $\hat{\xi}_w^{(1)}(n)$, and hence the $\hat{f}_w^{(1)}$ in exactly the same manner as that in which $\hat{\beta}, \hat{\beta}$ were used to form the $\hat{\xi}_w(n)$ and hence the $\hat{f}_w$.

We may now repeat Step 2 arriving at $\hat{\alpha}^{(2)}, \hat{\beta}^{(2)}$, which is the basis for a further iteration. This completes the description of the computations. We call the outcome of the $j$th iteration $\hat{\alpha}_{kl}^{(0)}, \hat{\beta}^{(0)}$.

We introduce the matrices

$$
\Phi = \left[\begin{array}{c} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi x} d\xi \\ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi^2} d\xi \end{array}\right] \quad (k, l = 1, \ldots, p);
$$

$$
\Psi = \left[\begin{array}{c} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi x} d\xi \\ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi^2} d\xi \end{array}\right] \quad (k, l = 1, \ldots, q);
$$

$$
\Omega = \left[\begin{array}{c} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi x} d\xi \\ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi^2} d\xi \end{array}\right] \quad (k = 1, \ldots, q; l = 1, \ldots, p);
$$

$$
A = \begin{bmatrix} A_{11} + \Psi & -A_{12} \\ -A_{21} & A_{22} \end{bmatrix}.
$$
\[ d_{11} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \frac{e^{ik_1 - i\eta}}{|g|^2|h|^2} d(\delta'F(\lambda)\delta) \]  
\[ (k, l = 1, \ldots, q); \]
\[ A_{12} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \frac{e^{ik_2}}{|g|^2} d(\delta'F(\lambda)) \]  
\[ (k = 1, \ldots, q); \]
\[ A_{22} = \frac{1}{\sigma^2} \int \frac{1}{|g|^2} dF(\lambda). \]

Since \( F(\lambda) \) is an \( r \times r \) matrix, \( A_{12} \) is \( q \times \) \( r \), and \( A_{22} \) is \( r \times r \). Now we have the following theorem:

**Theorem 1:** Under conditions (i), (ii), (iii), and (iv), the vector \((\tilde{g}^{(j)}, \tilde{p}^{(j)})\) converges almost surely to \((\alpha', \rho')\) and \(\sqrt{N((\tilde{g}^{(j)} - \alpha')', (\tilde{p}^{(j)} - \rho)'\})\) has a distribution converging to the multivariate normal distribution with zero mean and covariance matrix

\[ V = \begin{bmatrix} \Phi & -\Omega' & 0 \\ -\Omega & A & 0 \\ 0 & 0 & 0 \end{bmatrix}^{-1}. \]

The estimate is asymptotically efficient. The estimates \( \Phi^{(j)}, \tilde{B}^{(j)}, \tilde{Q}^{(j)} \) converge almost surely to \( \Phi, \Delta, \Omega \) respectively.

Before discussing the proof, two points may be made. We may choose to compute \( \Phi^{(j-1)}, \tilde{B}^{(j-1)}, \tilde{Q}^{(j-1)} \) only, as the estimates for \( j \) replacing \( (j - 1) \) are needed only if \( (j + 1) \) iterations are to be done. The final statement of the theorem applies also, of course, with \( (j - 1) \) replacing \( j \). The estimate will be asymptotically efficient in the sense that they are based on estimators of the \( \alpha, \beta, \delta \) which are asymptotically efficient in the sense defined at the end of Section 1.

The second point relates to the size of \( N \). We have in mind situations where \( N \leq 500 \). In this case the calculations are well within the capacity of modern equipment. However as \( N \) increases the labor of computing the \( w(\omega) \) becomes dominant. If \( N \) is highly composite, e.g., \( N = 2^33^5 \), then much larger values may be easily handled. If really large values of \( N \) are to be used, then it might be worthwhile replacing \( I_{\alpha}, I_{\beta}, I_{\delta} \) by smoothed estimates of spectra. Provided the smoothing is not too radical, there is no doubt that Theorem 1 will continue to hold.

**Proof of Theorem 1:** We shall give the proof in outline only. In detail it is rather complicated, though it does not differ in principle from that given in [4]. The essential differences are due to the occurrence of the \( x_j(n) \) and the dropping of the assumption of stationarity for the \( y(n) \) (which is mainly a consequence of the introduction of the \( x_j(n) \)). Details of the proof will be contained in a thesis submitted by the second author to the Australian National University. First it is
shown that $\hat{\rho} = -D^{-1}\hat{d}$ (see Step 2 of the calculations, above) converges to $\rho$, and hence $\hat{f}_n$ converges to $f_\omega$. Since the zeros of $g$ lie outside of the unit circle then

$$|\hat{g}|^{-2} = \sum_{-\infty}^{\infty} \hat{\xi}(u) e^{iu\lambda}$$

where the $\hat{\xi}(u)$ converge to $\xi(u)$, and also for $N$ sufficiently large the sequence $\hat{\xi}(u)$ converges to zero exponentially with $u$. As a result we may show that $\hat{D}$ converges to $\Delta$. For example

$$\frac{1}{N} \sum_{i=1}^{N-1} e^{i\phi_{x_i}} \frac{I_x}{|\hat{g}|^2} = \frac{1}{2\pi} \sum_{-N+1}^{N-1} \hat{\xi}(u) \left(1 - \frac{|u|}{N}\right) \sum_{j=-\infty}^{\infty} \hat{\xi}(k - v + jN)$$

which consequently converges to

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\lambda} \left\{ \frac{\sigma^2}{2\pi|h|^2} d\lambda + \frac{\delta^2 dF(\lambda) d\lambda}{|g|^2 |h|^2} \right\}.$$ 

Similar results hold for other components of $\hat{D}$ so that $\hat{D}$ converges to $\Delta$.

Now we introduce the matrix and vector $D, d$ which differ from $\hat{D}, \hat{d}$ only in replacing $\hat{g}$ by $g$. (Of course they are not computable.) Then $D$ converges to $\Delta$, and putting $\hat{\rho} = -D^{-1}\hat{d}$ we have

$$\sqrt{N}(\hat{\rho}^{(1)} - \rho) = -\sqrt{N}(D^{-1}\hat{d} - D^{-1}d)$$

which converges in probability to $-\Delta^{-1}\sqrt{N}(\hat{D} - D)\rho + (\hat{d} - d)$. In much the same way as for $\hat{D}$, but after considerable manipulations, $\sqrt{N}(\hat{\rho}^{(1)} - \rho)$ may be shown to converge in probability to

$$\sqrt{N}(\hat{\rho} - \rho) + \Delta^{-1} \begin{pmatrix} \Omega^T \\ 0 \end{pmatrix} \sqrt{N}(\hat{d} - d).$$

For example the typical entry in $\hat{D} - D$ is

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{1}{|\hat{g}|^2} - \frac{1}{|g|^2} \right\} I_x e^{i\phi - i\phi_{x_i}}$$

which converges in probability to

$$\frac{1}{N} \sum_{i=1}^{N} \sqrt{N}(\hat{g} - g) I_x e^{i\phi - i\phi_{x_i}}.$$ 

In the same way we introduce $\hat{a} = -A^{-1}a$ where $A, a$ differ from $\hat{A}^{(1)}, \hat{a}^{(1)}$ in replacing $\hat{h}^{(1)}, \hat{\xi}^{(1)}$, and $\hat{g}$ by $h, \delta$, and $g$. Then $\sqrt{N}(\hat{a}^{(1)} - a)$ converges in probability to

$$\sqrt{N}(\hat{a} - a) + 2\sqrt{N}(\hat{a} - a) - \Phi^{-1} \Omega^T \sqrt{N}(\hat{\rho}^{(1)} - \rho).$$

Now using (8) and (9) we may express $\sqrt{N}(\hat{a}^{(1)} - a)$ and $\sqrt{N}(\hat{\rho}^{(1)} - \rho)$ linearly in terms of $\sqrt{N}(\hat{a} - a)$ and $\sqrt{N}(\hat{\rho} - \rho)$ and it remains only to discover the asymptotic
joint distribution of these two vectors and convert that to a distribution for
\( \sqrt{N(\hat{\alpha}^{(t)} - \alpha)} \), \( \sqrt{N(\hat{\rho}^{(t)} - \rho)} \). Thus \( \sqrt{N(\hat{\alpha} - \alpha)} = -A^{-1} \sqrt{N(\alpha + A\alpha)} \) which we
may replace by \( -\Phi^{-1}u \) where \( u \) has typical element given by
\[
\frac{\sigma^2}{2\pi} u_k = N^{-\frac{1}{4}} \sum_{t=0}^{N-1} |g|^{-2} g^{-1} \left| h_{w_j} + \sum_{j=1}^{r} \delta(j)w_j \right|^2 e^{ik\omega t}
\]
which may be replaced by
\[
N^{-\frac{1}{4}} \sum_{t=0}^{N-1} |g|^{-2} g^{-1} I_w e^{ik\omega t}
\]
where \( I_w \) is the periodogram for the uncomputable \( w(n) \). Now exactly as in [4], we
see that \( u_k \) may be replaced by
\[
\sigma^{-2} N^{-\frac{1}{4}} \sum_{n=1}^{N} \varepsilon(n)\eta(n - k)
\]
where the \( \eta(n) \) are a stationary sequence satisfying
\[
\sum_{j=0}^{p} a(j)\eta(n - j) = \varepsilon(n).
\]
Similarly we may replace \( \sqrt{N(\hat{\rho} - \rho)} \) by \( -A^{-1}v \) where \( v \) has typical element
given by
\[
(s^2/2\pi)v_k = N^{-\frac{1}{4}} \sum_{t=0}^{N-1} |g|^{-2} \{ I_{h,h} + \sum \delta(j)I_j \} e^{ik\omega t} \quad (k = 1, \ldots, q),
\]
\[
= N^{-\frac{1}{4}} \sum_{t=0}^{N-1} |g|^{-2} \{ I_{h,h} + \sum \delta(k)I_{j,k} \} \quad (k = q + j; j = 1, \ldots, r).
\]
We may replace \( I_{h,h} + \Sigma \delta(j)I_j \) by \( I_{yw} \) where this is the cross periodogram between \( y(n) \) and \( w(n) \), and \( I_{h,h} + \Sigma \delta(k)I_{j,k} \) by \( I_{yw} \) where this is the periodogram between \( x(n) \) and \( w(n) \). We may then replace \( I_{yw} \) by \( h^{-1}\{ -\Sigma \delta(j)I_{yw} + I_w \} \). The part
\( N^{-\frac{1}{4}} \sum |g|^{-2} h^{-1} I_w (\exp ike_{0}) \) may be replaced (again [4]) by
\[
\sigma^{-2} N^{-\frac{1}{4}} \sum_{n=1}^{N} \varepsilon(n)\xi(n - k),
\]
\[
\sum_{j=0}^{q} b(j)\xi(n - j) = \varepsilon(n)
\]
while the remaining parts are of the form
\[
\sum \delta(j) \frac{1}{\sqrt{N}} \sum I_{h_k} e^{ik\omega t}, \quad \frac{1}{\sqrt{N}} \sum I_{h_k} g.
\]
The remainder of the derivation of the joint distribution of \( \sqrt{N(\hat{\alpha} - \alpha)} \), \( \sqrt{N(\hat{\rho} - \rho)} \)
reduces to the problem of obtaining that distribution for estimated autoregression coefficients or regression coefficients. (See [1 and 2].) The conversion of these
results into the required results for \( \hat{\alpha}^{(1)}, \hat{\beta}^{(1)} \) is straightforward. The covariance matrix for the asymptotic distribution of the maximum likelihood estimators, assuming the \( x_j(n) \) to be stationary, may be found by an extension of the results given in [8] and is the same as that stated in Theorem 1.

3. SOME PROPERTIES OF THE ESTIMATORS AND DISTRIBUTED LAG MODELS

The first point which we wish to make in this section relates to the possibility that in choosing a mixed autoregressive, moving-average model one may also be choosing to accept certain risks. Let us now call \( V_{p} \) the matrix \( V \) of Theorem 1, so as to emphasize the assumed order of the moving average. Now if \( \alpha(p) = 0 \), then some calculation shows that the part of \( V_{p}^{-1} \) which gives the asymptotic variances and covariances of the parameters of the model, other than \( \alpha(p) \), is

\[ V_{p}^{-1} = (1 - e'V_{p}^{-1} e)^{-1}V_{p}^{-1} e'eV_{p}^{-1} \]

where the vector \( e \) has zeros everywhere save in the \( p \)th to \((p + q)\)th places where it has, in the \((p + j)\)th place,

\[ e_{p+j} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-i(p-j)\lambda} d\lambda \quad (j = 1, \ldots, q), \]

which is, of course, zero for \( j > p \). The matrix \( V_{p}^{-1} \) gives the variances and covariances that would obtain if we had known that \( \alpha(p) = 0 \) and had fitted only \( \alpha(1), \ldots, \alpha(p-1) \). Thus the overfitting has resulted in too large a covariance matrix. The effect can be most clearly seen when \( q = 1 \), for then \( \sqrt{N(\hat{\beta}(1) - \beta(1))} \) has a limiting distribution with variance \( \nu/(1 - \nu) \) where \( \nu \) is the variance which would obtain if it were recognized that \( \alpha(p) = 0 \). For example, when \( p = q = 1, r = 0 \), then \( \nu = (1 - \beta^2(1)) \), and the efficiency of \( \hat{\beta}(1) \) is only \( \beta(1)^2 \) when \( \alpha(1) = 0 \). It is to be expected that this efficiency will be low when \( \beta(1) \) is small, as we are then near to an unidentified situation, but even for \( \beta(1) = 0.7 \) the efficiency is under 50 per cent.

Of course when \( \alpha(p) \neq 0 \) and we assume otherwise, the estimates of the remaining parameters will not be consistent. When \( \beta(1) = 0 \) and this is not recognized, similar problems emerge, though these are more complicated to describe.

A situation where the above mentioned problem cannot arise is that where \( p = q \) and \( \alpha(j) = \beta(j), j = 1, \ldots, p \). It is instructive to consider this case. We must now require that \( \delta \neq 0 \), for otherwise the model is not identified. In this case, as we have already said, (1) is equivalent to

\[ y(n) + \sum_{j=1}^{r} \hat{\delta}(j) \sum_{k=0}^{\infty} a_k x_j(n - k) = \sigma(n), \quad (10) \]

which, for \( r = 1 \), becomes the so called distributed lag model, with the \( a_k \) generated by \( g^{-1} = h^{-1} \). As we have said, this case \( r = 1 \) is probably most interesting when \( x_j(n) = x(n - j + 1) \) so that (10) is

\[ y(n) + \delta \sum_{j} \hat{x}(n - j) = \sigma(n), \quad \delta = \hat{\delta}(1). \]
where now the \( \lambda_j \) are generated by
\[
\sum_{j=0}^{r-1} \delta(j + 1)/(\delta(1)) \sum_{j=0}^{q} \beta(j)z^j,
\]
so that we are dealing with the general rational distributed lag model with independent errors. Let us return to the general model (10) and consider the best linear combination of \( \widetilde{\beta}^{(j)} \) and \( \hat{\beta}^{(j)} \) (assuming \( j \) iterations have been completed), say \( A_1\widetilde{\beta}^{(j)} + A_2\hat{\beta}^{(j)}, A_1 + A_2 = I_p \). Then it is easily established from Theorem 1 that we must take \( A_1 = 0 \) and \( A_2 = I_p \) so that \( \hat{\beta}^{(j)} \) is the best estimator. Thus \( \hat{\beta}^{(j)} \) is our best estimator of the unknown parameters, and this has covariance matrix
\[
A_{11} - A_{12}^{-1} \quad -A_{21} \quad A_{22}^{-1}
\]
(12)
This is the covariance matrix for an asymptotically efficient estimator, needless to say. This agrees also with the result obtained for the special case \( p = q = r = 1 \) in [3]. (We should mention here that formula 4.7 in [3] is incorrectly stated as is evident from the derivation which precedes it. The formula should read
\[
\int_{\pi}^{\pi} \{2\pi f_\theta(\theta)\}^{-1} c(\theta)c(\theta)^* dF_\theta(\theta)
\]
Note that \( F_\theta(\theta) \) differs from our \( F(\theta) \) by a factor \( \gamma(0)^{-1} \). In fact an examination of the proof shows that the asymptotically efficient estimator of \( \rho \) may be more easily obtained, namely as
\[
\hat{\beta}^{(1)} = -\left\{ D - \frac{\partial^2 (A)}{2\pi} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}^{-1} \left\{ d + \frac{\partial^2 (A)}{2\pi} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}
\]
Moreover the matrix (12) is estimated by the first factor in \( -\hat{\beta}^{(1)} \). Thus the computations are greatly simplified. Further iterations are then completed beginning from \( \hat{\beta}^{(1)} \), computing \( \hat{D} \) and \( \hat{A}^{(1)} \) using \( \hat{g}^{(1)} = \hat{h}^{(1)} \) and \( \hat{\beta}^{(1)} \) in place of \( g, h, \) and \( \beta^{(1)} \). We call \( \hat{\beta}^{(j)} \) the result of the \( j \)th iteration. (The need for further iterations was not emphasized in [3], but in fact these are likely to be needed.) When this is done the estimate is asymptotically equivalent to that introduced in [3]. Indeed the likelihood function being \( L \) and \( \alpha \) being a function of \( \beta \), the equations of maximum likelihood become
\[
\frac{\partial L}{\partial \beta} + \left( \frac{\partial \alpha}{\partial \beta} \right) \frac{\partial L}{\partial \alpha} = 0, \quad \frac{\partial L}{\partial \delta} = 0.
\]
If \( \alpha = \beta \), then \( (\partial \alpha/\partial \beta) = I_p \), and we are led to Equation (13). To test the hypothesis \( \alpha = \beta \), in case \( p = q \), we may form
\[
\{2\pi N(\hat{\alpha}^{(j)} - \hat{\beta}^{(j)})^{(j-1)}(\hat{\alpha}^{(j)} - \hat{\beta}^{(j)}})^{(j-1)}\}^2
\]
(14)
which is asymptotically distributed as chi-square with \( p \) degrees of freedom.
(The \((j - 1)\) superscript in \( \hat{A} \) and \( \tilde{\sigma} \) is used because, as pointed out below Theorem 1, we do not need to compute \( \hat{A}^{(j)}, \tilde{\sigma}^{(j)} \) if \( j \) iterations only are to be done.)

We summarize these results in Theorem 2.

**Theorem 2:** If \( p = q \), we may test the hypothesis \( \alpha = \beta \) for large \( N \) by using (14) as chi-square with \( p \) degrees of freedom. If \( \alpha = \beta \), then \( \rho \) may be estimated by \( \hat{\rho}^{(j)} \) or by (13). Then \( \sqrt{N}(\hat{\rho}^{(j)} - \rho) \), \( \sqrt{N}(\tilde{\rho}^{(j)} - \rho) \) are each asymptotically normal with covariance matrix

\[
\begin{bmatrix}
\hat{A}_{11} & -\hat{A}_{12} \\
-\hat{A}_{21} & \hat{A}_{22}
\end{bmatrix}^{-1}.
\]

This matrix is consistently estimated by

\[
\begin{bmatrix}
\hat{D} - \tilde{\sigma}^2 \begin{bmatrix} \hat{A} & 0 \\ 2\pi & 0 \end{bmatrix}
\end{bmatrix}^{-1}.
\]

The procedure is asymptotically efficient and \( \tilde{\rho}^{(j)} \) converges to \( \rho \) almost surely.

In case \( X_j(n) = x(n - j + 1) \), so that (11) is the relation to be estimated the computation would be simplified as follows. We would replace \( w_j \) by

\[
w(\omega_j) e^{i(j - 1)\omega_j}, \quad w(\omega_j) = (2\pi N)^{-\frac{1}{2}} \sum_{j=1}^{N} x(n) e^{i\omega_j} = u(\omega_j) + iv(\omega_j).
\]

Then all later formulae involving the \( \omega_j \) simplify. For example \( u_j \) and \( c_{jk} \) are replaced, respectively, by

\[
u \cos(j - 1)\omega_j - v \sin(j - 1)\omega_j, \quad (u^2 + v^2) \cos(j - k)\omega_j.
\]

We would also replace \( \hat{c}_{jk}(n) \) by \( \hat{c}(n + j - k) \) where

\[
c(n) = (N - n)^{-1} \sum_{m=1}^{N-n} (x(m) - \bar{x})(x(m + n) - \bar{x}) = c(-n)
\]

\((n = 0, 1, \ldots, p + r - 1)\).

We would replace \( \hat{c}_{jk}(n) \) by \( \hat{c}_{jk}(n - j + 1) \) where

\[
\hat{c}_{jk}(n) = (N - n)^{-1} \sum_{m=1}^{N-n} (y(m) - \bar{y})(x(m + n) - \bar{x}) = \hat{c}_{xj}(-n)
\]

\((n = 0, 1, \ldots, p + q + r - 1)\).

There are two ways in which the treatment in [3] is more general than in the present paper. (Of course in other ways it is much less general.) The first of these is through the introduction of a more general stochastic structure for the term on the right side in (1). This is assumed to be of the form

\[
\sum_{j=0}^{q} \beta(j)u(n - j)
\]
where \( u(n) \) is stationary and of a sufficiently regular kind (see the references cited). If \( f_{\nu}(\lambda) \) is the spectral density of \( u(n) \), then the covariance matrix of the limiting distribution of the estimates is, for \( p = q = r = 1 \),

\[
(15) \quad \left\{ \int_{-\pi}^{\pi} \frac{1}{2\pi f_{\nu}(\lambda)} \left[ e^{-i\lambda} / h \right] \left[ e^{i\lambda} / h \right] dF(\lambda) \right\}^{-1}
\]

where \( f_{\nu}(\lambda) = |h|^{2} f_{\nu}^2(\lambda) \). Of course if \( \alpha = \beta \) and \( u(n) = \varepsilon(n) \), this agrees with our earlier result. The estimation procedure leading to (15) does not involve any parameterization of the stochastic structure of \( u(n) \) but, as we have said, only rather general specifications are made. (This is a common device in time series analysis.)

An alternative procedure is to proceed as we have in (1), which (when we do not assume \( \alpha = \beta \)) takes \( u(n) \) to be of the form

\[
(16) \quad u(n) = \sum_{0}^{\infty} b_{j} \varepsilon(n - j), \quad \sum_{0}^{\infty} b_{j} \varepsilon^{2} = g / h.
\]

Of course this specification restricts the nature of the process generating \( u(n) \) (as we have already said) since it must correspond to a rational generating function having \( h \) in the denominator. However, if \( g \) is restricted only to be a polynomial, this specification is still very general. It is therefore of interest to compare the covariance matrix of the limiting distribution of the estimates of \( \rho \), obtained by the methods of Section 2, with the generalization of (15) appropriate to the case of a general rational distributed lag model and \( r \) exogenous variables (all with lag structure such that the denominator of the generating functions is \( h \)). Of course we assume that \( f_{\nu} = (\sigma^{2}/2\pi) ||\varepsilon||^{2} / h^{2} \) so that \( f_{\nu} = (\sigma^{2}/2\pi) ||\varepsilon||^{2} \), the difference in the two methods lying in the fact that the second does not take account of the special form of \( f_{\nu} \). The generalization of (15) is

\[
\left[ \begin{array}{cc}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{array} \right]^{-1}
\]

whereas \( \sqrt{N}(\hat{\rho} - \rho) \) will have for its limiting distribution the covariance matrix

\[
\left[ \begin{array}{cc}
A_{11} + \Psi - \Omega \Phi^{-1} \Omega' & A_{12} \\
A_{21} & A_{22}
\end{array} \right]^{-1}
\]

However

\[
(17) \quad \Psi - \Omega \Phi^{-1} \Omega' \geq 0
\]

so that the use of additional information has reduced the covariance matrix (as it clearly must). When \( g = h \), then (17) is null, and the two methods are equivalent as we have already seen. If the degree of \( g \) is overstated, then, as we know, we lose efficiency, but the estimates are never worse than the methods of [3] as (17) shows. Of course if the degree of \( g \) is understated, the methods of this paper become inconsistent, but it is possible to test whether \( p \) should be increased. In addition to a more accurate estimate of \( \rho \), the methods also lead to an estimate of \( g \) and hence
(for example) to a specification of the optimal linear predictor for \( y(n) \), once the \( x(n) \) are predicted. (The prediction of the \( x(n) \) requires knowledge of their stochastic structure.) The relation between the three cases is exhibited by the special example where \( p = q = r = 1, \alpha(1) = 0 \), and \( x(n) \) consists of uncorrelated random variables with unit variance, while \( \sigma^2 = 1 \). Then we take the estimates obtained using (i) all information, (ii) only the knowledge that \( g \) is of degree one, and (iii) by the method of [3]. The three estimates of \( \delta \) have the same asymptotic variance properties, but the estimate of \( \beta \) has, in its limiting distribution, the variance respectively \( (1 - \beta^2)(1 + \delta^2) \), \( (1 - \beta^2)(\beta^2 + \delta^2) \), and \( (1 - \beta^2)/\delta^2 \).

The other direction in which the approach of [3] was more general than that of the present paper was in the fact that \( x(n) \) was allowed to be more general and in particular was allowed to contain components which were trending as well as of the kind treated in the present paper. It seems that the methods of the present paper could be extended (and modified) to deal with the more general specifications, but the statement of already complex results would be further complicated. Since we can allow for trends (for example) by the procedures described in the paragraph just above (2), it has been thought best not to seek this additional generality.

4. NUMERICAL EXAMPLES AND DISCUSSION

The methods of estimation described in Section 2 were applied to data generated by the relation

\[
y(n) + \beta y(n-1) + \delta x(n) = \epsilon(n) + \alpha \epsilon(n-1)
\]

wherein \( \alpha = 0.5 \), \( \beta = -0.8 \), and \( \delta = 0.3 \). The sequence \( x(n) \) was generated by

\[
x(n) = 0.6x(n-1) + \eta(n).
\]

Both sequences \( y(n) \) and \( x(n) \) were (very nearly) stationary. The \( \epsilon(n) \) and \( \eta(n) \) sequences were independent of each other and consisted of independent random variables with zero mean and a uniform distribution. The ratio \( \theta = \sigma_\epsilon^2/\sigma_x^2 \), which determines the relative importance of the \( \epsilon(n) \) and \( x(n) \) sequences in the generation of \( y(n) \), was taken as either 1 or 9. The sample size \( N \) was 100 or 200. (We report also some experiments for \( N = 40 \) in Table II.) For each \( \theta \), \( N \), a total of 20 replications was made. In Table I, for each \( \theta \), \( N \), the row (a) gives the mean of the observed values of the relevant parameter for the 20 runs, (b) shows the theoretical variance obtained from Theorem I, and (c) shows the observed variance of the 20 numbers. The iterations were continued until successive estimates of all parameters differed by not greater than .005. For \( N = 100 \), in 10 cases the estimates were stable after the third iteration, and in 18, after the sixth. The two exceptional cases had both stabilized by the ninth iteration. For \( N = 200 \), in 15 cases the estimates were stable after the third iteration, in 19 after the fifth, and all were stable after six iterations. The total running time for the program of 20 series for the case \( N = 100 \), including generation of the data, was 14.66 minutes (using an IBM 360/50).
there is no doubt that $\hat{\beta}$ is biased towards zero. Walker [7], for the case where $\delta = 0$ and $N = 100$, from an asymptotic expansion of the mean of $\hat{\beta}$, evaluates this bias at 0.044 which is close to the bias observed above. A similar bias would probably be observed in $\tilde{\alpha}$ if $\alpha$ were nearer to unity.

As mentioned above the same model was also estimated using $N = 40$ for $\theta = 9$. The results appear in Table II for 20 replications.

In Table II, the agreement with theory is not good, especially with $\hat{\beta}$. Among the 20 replications were two which gave very bad results. One gave $\tilde{\alpha}, \tilde{\beta}, \tilde{\delta}$ as 0.051, −0.770, 0.427, while the other gave the values 0.931, 0.004, 0.269. It is possible that in these two cases the iterations led to a solution of the likelihood equations which is not the maximum likelihood solution but examination of particular cases leads us to believe that this is not so. Certainly the limiting distribution is nowhere near approached for $N = 40$.

In this connection and in connection also with the number of iterations (described above), it is perhaps worthwhile explaining what is being achieved by the method being used. This can be done best by considering an alternative procedure in which the likelihood equations might be solved by a direct application of the Newton-Raphson technique. This would involve the evaluation of second derivatives of the likelihood function and would become unwieldy with $N, p, q$, and $r$.
large. The method used herein replaces these second derivatives by corresponding expressions to which they converge almost surely, evaluated at the estimates of the true parameter point. In the case \( q = 0 \), the method becomes especially simple since these expressions are absolute constants independent of the parameter point (see [4]). However if an initial estimate is a long way from the true parameter point, the replacement of the second derivatives by these expressions may slow the convergence or might lead to divergence. The matter needs further investigation.

Australian National University

Manuscript received September, 1970; revision received February, 1971.

REFERENCES