The Annals of Statistics
1975, Vol. 3, No. 6, 1283-1304

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS OF AUTOREGRESSIVE PROCESSES WITH MOVING AVERAGE RESIDUALS AND OTHER COVARIANCE MATRICES WITH LINEAR STRUCTURE[1]

BY T. W. ANDERSON

Stanford University and London School of Economics

The autoregressive process with moving average residuals is a stationary process $\{y_t\}$ satisfying $\sum_{s=0}^{p} \beta_s y_{t-s} = \sum_{j=0}^{q} \alpha_j v_{t-j}$, where the sequence $\{v_t\}$ consists of independently identically distributed (unobservable) random variables. The distribution of $y_1, \cdots, y_T$ can be approximated by the distribution of the T-component vector $\mathbf{y}$ satisfying $\sum_{s=0}^{p} \beta_s \mathbf{K}_s \mathbf{y} = \sum_{j=0}^{q} \alpha_j \mathbf{J}_j \mathbf{v}$, where $\mathbf{v}$ has covariance matrix $\sigma^2\mathbf{I}$, $\mathbf{K}_s = \mathbf{J}_s = \mathbf{L}^s$, and $\mathbf{L}$ is the $T \times T$ matrix with 1's immediately below the main diagonal and 0's elsewhere. Maximum likelihood estimates are obtained when $\mathbf{v}$ has a normal distribution. The method of scoring is used to find estimates defined by linear equations which are consistent, asymptotically normal, and asymptotically efficient (as $T \to \infty$). Several special cases are treated. It is shown how to calculate the estimates.

1. Introduction. A stationary stochastic process that serves as a useful model for time series analysis is the autoregressive process with moving average residuals $\{y_t\}$ which satisfies

$$\sum_{s=0}^{p} \beta_s y_{t-s} = \sum_{j=0}^{q} \alpha_j v_{t-j}\,, \tag{1.1}$$

$t = \cdots, -1, 0, 1, \cdots$, where the sequence $\{v_t\}$ consists of independently identically distributed (unobservable) random variables. [See Section 5.8 of Anderson (1971a) and Box and Jenkins (1970).] To avoid indeterminacy we set $\beta_0 = \alpha_0 = 1$. The mean of v_t (assumed to exist) is independent of t and is taken to be 0 for convenience. (Modifications necessary to account for an arbitrary mean are also discussed.) When $\mathscr{E} y_t = 0$ and second-order moments exist, stationarity implies

$$\mathscr{E} y_t y_s = \sigma(t-s)\,, \tag{1.2}$$

dependent only on the difference of the indices.

We shall assume that the v_t's are normally distributed, that is, that the process is Gaussian. Then the model is completely specified by the coefficients in (1.1) and the variance of v_t, say σ^2. A statistical problem treated here is to estimate

Received May 1974; revised April 1975.

[1] This research was supported by Office of Naval Research Contract N00014-67-A-0112-0030, Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

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

Key words and phrases. Autoregressive processes with moving average residuals, covariance matrices with linear structure, maximum likelihood estimation, estimation of linear transformations, multivariate normal distribution.

$\beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q$, and σ^2 on the basis of a set of observations at T successive time points, $y_1, \cdots, y_T$ ($T \geqq p + q + 1$).

If an arbitrary random vector has the multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$, the density of $\mathbf{y}$ is

$$\frac{1}{(2\pi)^{\frac{1}{2}T}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{y}'\mathbf{\Sigma}^{-1}\mathbf{y}}, \tag{1.3}$$

where

$$\mathscr{E} y_t y_s = \sigma_{ts}, \qquad t, s = 1, \cdots, T, \tag{1.4}$$

is the t, sth element of $\mathbf{\Sigma}$. When the components of $\mathbf{y}$ are y_t, $t = 1, \cdots, T$, a segment of the process satisfying (1.1), then (1.4) is (1.2); the covariances are functions of the parameters $\beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q$, and σ^2.

The method of maximum likelihood can be considered, but in general an explicit solution cannot be found. The approach of this paper is to modify the model slightly so that the derivatives of the likelihood function set equal to 0 yield relatively simple equations. The modification is equivalent to setting $y_0 = y_{-1} = \cdots = y_{1-p} = 0$ and $v_0 = v_{-1} = \cdots = v_{1-q} = 0$. Since the likelihood equations are, nevertheless, nonlinear, an iterative procedure is proposed based on the method of scoring, which involves estimating the information matrix. The computations to be carried out are in the time domain. If the initial estimates are consistent, the first step of the iteration yields consistent, asymptotically normal, and asymptotically efficient estimates (as $T \to \infty$).

Durbin (1959), (1960) and Walker (1961), (1962) have proposed estimation procedures which are also carried out in the time domain, but are not based on maximizing the likelihood or a modest modification of it. Box and Jenkins (1970) have suggested maximizing the likelihood function (in the time domain) by numerical means; a method of searching for the maximum by computing the likelihood over a grid of trial values of the parameters is described in detail. Åström and Bohlin (1966) have approached the maximization of the likelihood of the modified model treated in this paper in a rather different way and have developed a Newton–Raphson procedure, which seems to involve more computation than the method of scoring.

The covariance sequence (1.2) of a stationary process has a spectral representation, which is in the case of an absolutely continuous spectral distribution function

$$\sigma(h) = \int_{-\pi}^{\pi} f(\lambda) \cos \lambda h \, d\lambda, \qquad h = 0, \pm 1, \cdots. \tag{1.5}$$

The spectral density $f(\lambda)$ may be determined by

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \sigma(h) \cos \lambda h \tag{1.6}$$

when the series on the right-hand side converges absolutely. In the case of model (1.1) the spectral density is

$$f(\lambda) = \frac{\sigma^2}{2\pi} \cdot \frac{|\sum_{j=0}^{q} \alpha_j e^{i\lambda j}|^2}{|\sum_{r=0}^{p} \beta_r e^{i\lambda r}|^2}. \tag{1.7}$$

Hannan (1969), (1970) has proposed estimation methods based on the sample spectral density, the so-called periodogram. Akaike (1973) has shown that Hannan's procedures are approximately Newton–Raphson in the frequency domain. The methods developed in this paper differ from Hannan's in that (i) they are based on the use of the estimated information matrix (scoring) instead of the matrix of second derivatives of the logarithm of the likelihood (Newton–Raphson), (ii) they are carried out in the time domain instead of the frequency domain, and (iii) the only approximation involved is setting equal to 0 the variables with nonpositive indices. An advantage of making the calculations in the time domain is that the sample spectral density does not need to be computed.

If we let $u_t = \sum_{j=0}^{q} \alpha_j v_{t-j}$, where $\{v_t\}$ consists of independently, identically distributed random variables with means $\mathscr{E} v_t = 0$ and variances $\mathscr{E} v_t^2 = \sigma^2$, the spectral density of the stationary process $\{u_t\}$ is

$$(1.8) \qquad f_u(\lambda) = \frac{\sigma^2}{2\pi} \sum_{j=0}^{q} \alpha_j e^{i\lambda j} \sum_{j=0}^{q} \alpha_j e^{-i\lambda j} = \frac{1}{2\pi} \sum_{h=-q}^{q} \sigma_u(h) e^{i\lambda h} = \frac{1}{2\pi} \sum_{h=-q}^{q} \sigma_u(h) \cos \lambda h \, ,$$

where

$$(1.9) \qquad \sigma_u(h) = \sigma^2 \sum_{k=0}^{q-|h|} \alpha_k \alpha_{k+|h|} \, , \qquad h = 0, \pm 1, \cdots, \pm q \, ,$$

are the possibly nonzero covariances of $\{u_t\}$. The parameters $\alpha_1, \cdots, \alpha_q$, and σ^2 can be replaced by $\sigma_u(0), \sigma_u(1), \cdots, \sigma_u(q)$. We shall assume the zeros of

$$(1.10) \qquad M(z) = \sum_{j=0}^{q} \alpha_j z^{q-j}$$

are less than 1 in absolute value. Then given $\sigma_u(0), \sigma_u(1), \cdots, \sigma_u(q) \neq 0$ $\sum_{h=-q}^{q} \sigma_u(h) z^h$ can be factored uniquely into $\sigma M(z)[\sigma M(z^{-1})]$, thus, defining $\alpha_1, \cdots, \alpha_q$, and σ^2. [See Anderson (1971a) and (1971b) for details.] An alternative parametrization of the process $\{y_t\}$ is $\sigma_u(0), \sigma_u(1), \cdots, \sigma_u(q), \beta_1, \cdots, \beta_p$. We modify the model by setting $y_0 = y_{-1} = \cdots = y_{1-p}$ (but not altering $\{v_t\}$). Another statistical problem considered in this paper is the estimation of these parameters on the basis of a set of observations.

With this parametrization the spectral density (1.7) is $f_u(\lambda)/|\sum_{r=0}^{p} \beta_r e^{i\lambda r}|^2$. Clevenson (1970) in an unpublished paper and Parzen (1971) have developed iterative estimation methods using the sample spectral density; they are approximations to the method of scoring in the frequency domain. In principle the methods of Hannan, Clevenson, and Parzen could be obtained from the approach of Whittle (1951) [developed further by Walker (1964)] who approximated the logarithm of the likelihood function by integrals or sums of the ratio of the sample spectral density to the process spectral density depending on a finite number of parameters.

Anderson (1971b), (1973) developed an iterative procedure to estimate the parameters $\sigma_u(0), \sigma_u(1), \cdots, \sigma_u(q)$ of a pure moving average process which is essentially based on scoring (as pointed out by J. N. K. Rao). The covariance matrix of a segment of length T of the process was considered as a special case of a covariance matrix with the linear structure $\mathbf{\Sigma}^u = \sum_{h=0}^q \sigma_u(h)\mathbf{G}_h$, where $\mathbf{G}_0$, $\mathbf{G}_1, \cdots, \mathbf{G}_q$ are known linearly independent symmetric matrices. The Newton–Raphson method in general terms was developed in Anderson (1969), (1970); in these papers methods were also developed for models in which the inverse of the covariance matrix was such a linear combination. In this paper the models are generalized to include covariance matrices of more general form, $\mathbf{B}^{-1}\mathbf{\Sigma}^u\mathbf{B}'^{-1}$ and $\mathbf{B}^{-1}\mathbf{A}\mathbf{A}'\mathbf{B}'^{-1}$, where $\mathbf{A} = \sum_{k=0}^q \alpha_k \mathbf{J}_k$ and $\mathbf{B} = \sum_{l=0}^p \beta_l \mathbf{K}_l$. There may be N observations on a random vector $\mathbf{y}$ with such covariance matrices. The method of scoring is developed in general terms and then specialized to the modified autoregressive moving average models. The corresponding Newton–Raphson methods can be written down similarly from the matrices of second derivatives. The pure autoregressive model (Section 2) and the pure moving average model (Section 3) are treated in these terms before the general models.

The multivariate models are treated in general terms in order that the methods be available for other applications. The model in which the covariance matrix is a linear combination of known symmetric matrices has been useful in the analysis of variance and genetics; some models in econometrics and psychometrics can be formulated in the terms of this paper. Estimation of the parameters of the autoregressive moving average process in the frequency domain can also be obtained from these general results.

2. Estimation of coefficients of linear transformations to approximate autoregressive processes.

2.1. *General linear transformations.* Suppose $\mathbf{y}$ is a T-component random vector defined by

$$\sum_{l=0}^p \beta_l \mathbf{K}_l \mathbf{y} = \mathbf{v}\,, \tag{2.1}$$

where $\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ are $p+1$ known linearly independent $T \times T$ matrices, $\beta_0 = 1$ and $\beta_1, \cdots, \beta_p$ are p parameters such that $\sum_{l=0}^p \beta_l \mathbf{K}_l$ is nonsingular; we assume that there is at least one such set. Suppose $\mathbf{v}$ is a T-component random vector with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Then

$$\mathbf{y} = \left(\sum_{l=0}^p \beta_l \mathbf{K}_l\right)^{-1}\mathbf{v} \tag{2.2}$$

has mean vector $\mathbf{0}$ and covariance matrix

$$\mathscr{E}\mathbf{y}\mathbf{y}' = \sigma^2\left(\sum_{l=0}^p \beta_l \mathbf{K}_l\right)^{-1}\left(\sum_{k=0}^p \beta_k \mathbf{K}_k'\right)^{-1} = \sigma^2\left(\sum_{k,l=0}^p \beta_k\beta_l \mathbf{K}_k'\mathbf{K}_l\right)^{-1}. \tag{2.3}$$

Let $\mathbf{y}_1, \cdots, \mathbf{y}_N$ be N independent observations on $\mathbf{y}$, and let L denote the likelihood function when $\mathbf{v}$ has a normal distribution. Then

$$\frac{2}{N}\log L = -T\log 2\pi - T\log\sigma^2 + \log\left|\sum_{l=0}^p \beta_l \mathbf{K}_l\right|^2$$
$$- \frac{1}{N\sigma^2}\sum_{\alpha=1}^N \left(\sum_{k=0}^p \beta_k \mathbf{K}_k \mathbf{y}_\alpha\right)'\left(\sum_{l=0}^p \beta_l \mathbf{K}_l \mathbf{y}_\alpha\right) \tag{2.4}$$

$$= -T \log 2\pi - T \log \sigma^2 + 2 \log |\textstyle\sum_{l=0}^p \beta_l \mathbf{K}_l| - \frac{1}{\sigma^2} \operatorname{tr} \sum_{k,l=0}^p \beta_k \beta_l \mathbf{K}_k' \mathbf{K}_l \mathbf{C} ,$$

where

$$\mathbf{C} = \frac{1}{N} \textstyle\sum_{\alpha=1}^N \mathbf{y}_\alpha \mathbf{y}_\alpha' , \tag{2.5}$$

and tr denotes the trace of the matrix that follows. [The proof of Lemma 3.2.2 of Anderson (1958) shows that (2.4) $\rightarrow -\infty$ for any parameter $\rightarrow \infty$ and for $\sigma^2 \rightarrow 0$; therefore, the maximum occurs at a set of values of $\beta_1, \cdots, \beta_p$, and σ^2 for which the partial derivatives are set equal to 0.] To find partial derivatives in this paper we use the results

$$\frac{\partial |\mathbf{A}|}{\partial \theta} = |\mathbf{A}| \operatorname{tr} \mathbf{A}^{-1} \frac{\partial}{\partial \theta} \mathbf{A} , \tag{2.6}$$

$$\frac{\partial}{\partial \theta} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \frac{\partial}{\partial \theta} \mathbf{A} \mathbf{A}^{-1} . \tag{2.7}$$

[See Dwyer (1967) or Appendix A of Anderson (1958), for example.] Then

$$\frac{\partial}{\partial \beta_l} \frac{2}{N} \log L = 2 \operatorname{tr} (\textstyle\sum_{k=0}^p \beta_k \mathbf{K}_k)^{-1} \mathbf{K}_l - \frac{2}{N\sigma^2} \sum_{\alpha=1}^N \mathbf{y}_\alpha' \sum_{k=0}^p \beta_k \mathbf{K}_k' \mathbf{K}_l \mathbf{y}_\alpha$$

$$= 2 \operatorname{tr} (\textstyle\sum_{k=0}^p \beta_k \mathbf{K}_k)^{-1} \mathbf{K}_l - \frac{2}{\sigma^2} \operatorname{tr} \sum_{k=0}^p \beta_k \mathbf{K}_k' \mathbf{K}_l \mathbf{C} , \qquad l = 1, \cdots, p , \tag{2.8}$$

$$\frac{\partial}{\partial \sigma^2} \frac{2}{N} \log L = -\frac{T}{\sigma^2} + \frac{1}{\sigma^4} \operatorname{tr} \textstyle\sum_{k,l=0}^p \beta_k \beta_l \mathbf{K}_k' \mathbf{K}_l \mathbf{C} . \tag{2.9}$$

The maximum likelihood estimates may be defined by setting the derivatives equal to 0. The derivative equations are

$$\operatorname{tr} (\textstyle\sum_{k=0}^p \hat{\beta}_k \mathbf{K}_k)^{-1} \mathbf{K}_l = \frac{1}{\hat{\sigma}^2} \sum_{k=0}^p \hat{\beta}_k \operatorname{tr} \mathbf{K}_k' \mathbf{K}_l \mathbf{C} , \qquad l = 1, \cdots, p , \tag{2.10}$$

$$\hat{\sigma}^2 = \frac{1}{T} \textstyle\sum_{k,l=0}^p \hat{\beta}_k \hat{\beta}_l \operatorname{tr} \mathbf{K}_k' \mathbf{K}_l \mathbf{C} . \tag{2.11}$$

We can develop these equations in an alternative way by letting

$$\mathbf{K}_k \mathbf{y}_\alpha = \mathbf{y}_\alpha^{(k)} , \qquad k = 0, 1, \cdots, p , \quad \alpha = 1, \cdots, N . \tag{2.12}$$

Then

$$\frac{2}{N} \log L = -T \log 2\pi - T \log \sigma^2 + \log |\textstyle\sum_{l=0}^p \beta_l \mathbf{K}_l|^2 - \frac{1}{N\sigma^2} \sum_{\alpha=1}^N (\sum_{k=0}^p \beta_k \mathbf{y}_\alpha^{(k)})' (\sum_{l=0}^p \beta_l \mathbf{y}_\alpha^{(l)}) \tag{2.13}$$

$$= -T \log 2\pi - T \log \sigma^2 + 2 \log |\textstyle\sum_{l=0}^p \beta_l \mathbf{K}_l| - \frac{1}{\sigma^2} \boldsymbol{\beta}' \mathbf{M} \boldsymbol{\beta} ,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ and

(2.14)
$$\mathbf{M} = \frac{1}{N} \sum_{\alpha=1}^{N} \begin{pmatrix} \mathbf{y}_\alpha^{(0)\prime}\mathbf{y}_\alpha^{(0)} & \mathbf{y}_\alpha^{(0)\prime}\mathbf{y}_\alpha^{(1)} & \cdots & \mathbf{y}_\alpha^{(0)\prime}\mathbf{y}_\alpha^{(p)} \\ \mathbf{y}_\alpha^{(1)\prime}\mathbf{y}_\alpha^{(0)} & \mathbf{y}_\alpha^{(1)\prime}\mathbf{y}_\alpha^{(1)} & \cdots & \mathbf{y}_\alpha^{(1)\prime}\mathbf{y}_\alpha^{(p)} \\ \vdots & \vdots & & \vdots \\ \mathbf{y}_\alpha^{(p)\prime}\mathbf{y}_\alpha^{(0)} & \mathbf{y}_\alpha^{(p)\prime}\mathbf{y}_\alpha^{(1)} & \cdots & \mathbf{y}_\alpha^{(p)\prime}\mathbf{y}_\alpha^{(p)} \end{pmatrix}.$$

The partial derivatives of $(2/N) \log L$ set equal to 0 can be written in terms of the elements of $\mathbf{M}$ as

(2.15)
$$[\mathrm{tr}\,(\textstyle\sum_{k=0}^{p} \hat{\beta}_k \mathbf{K}_k)^{-1}\mathbf{K}_l] = \frac{1}{\hat{\sigma}^2} \hat{\boldsymbol{\beta}}'\mathbf{M}\,,$$

(2.16)
$$\hat{\sigma}^2 = \frac{1}{T} \hat{\boldsymbol{\beta}}'\mathbf{M}\hat{\boldsymbol{\beta}}\,;$$

the left-hand side of (2.19) denotes a row vector with the lth component given explicitly.

If $N > 1$ and $\mathscr{E}\mathbf{y} = \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is an arbitrary vector, then the sample mean

(2.17)
$$\bar{\mathbf{y}} = \frac{1}{N} \sum_{\alpha=1}^{N} \mathbf{y}_\alpha$$

is the maximum likelihood estimate of $\boldsymbol{\mu}$, and in the likelihood equations (2.10) and (2.11), $\mathbf{C}$ should be replaced by

(2.18)
$$\hat{\mathbf{C}} = \frac{1}{N} \sum_{\alpha=1}^{N} (\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})'\,.$$

In some models one wants $\mathscr{E}y_j = \mu$; that is, $\mathscr{E}\mathbf{y} = \mu\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (1, 1, \cdots, 1)'$. Then $2/N$ times the logarithm of the likelihood function is (2.4) with $\mathbf{C}$ replaced by

(2.19)
$$\mathbf{C}^* = \frac{1}{N} \sum_{\alpha=1}^{N} (\mathbf{y}_\alpha - \mu\boldsymbol{\varepsilon})(\mathbf{y}_\alpha - \mu\boldsymbol{\varepsilon})'\,.$$

The derivative of $2/N$ times the logarithm of the likelihood with respect to μ is

(2.20)
$$\frac{\partial}{\partial\mu} \frac{2}{N} \log L = \frac{2}{N\sigma^2} \boldsymbol{\varepsilon}' \sum_{k,l=0}^{p} \beta_k \beta_l \mathbf{K}_k'\mathbf{K}_l \sum_{\alpha=1}^{N} (\mathbf{y}_\alpha - \mu\boldsymbol{\varepsilon})\,.$$

If $\boldsymbol{\varepsilon}$ is a (right) characteristic vector of $\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ and $\mathbf{K}_0', \mathbf{K}_1', \cdots, \mathbf{K}_p'$, then

(2.21)
$$\hat{\mu} = \frac{1}{NT} \sum_{\alpha=1}^{N} \boldsymbol{\varepsilon}'\mathbf{y}_\alpha\,;$$

and in the other derivative equations $\mathbf{C}$ is replaced by

(2.22)
$$\frac{1}{N} \sum_{\alpha=1}^{N} (\mathbf{y}_\alpha - \hat{\mu}\boldsymbol{\varepsilon})(\mathbf{y}_\alpha - \hat{\mu}\boldsymbol{\varepsilon})'\,.$$

If $\boldsymbol{\varepsilon}$ is not a right characteristic vector of $\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ and $\mathbf{K}_0', \mathbf{K}_1', \cdots, \mathbf{K}_p'$, then usually (2.21) will not be the maximum likelihood estimate of μ.

The elements of the information matrix are N times

$$
\begin{aligned}
-\mathscr{E}\frac{\partial^2}{\partial\beta_j\partial\beta_l}\frac{1}{N}\log L &= \operatorname{tr}\left(\sum_{k=0}^p \beta_k \mathbf{K}_k\right)^{-1}\mathbf{K}_j\left(\sum_{k=0}^p \beta_k \mathbf{K}_k\right)^{-1}\mathbf{K}_l \\
&\qquad + \frac{1}{\sigma^2}\mathscr{E}\operatorname{tr}\mathbf{K}_j'\mathbf{K}_l\mathbf{C} \\
&= \operatorname{tr}\left(\sum_{k=0}^p \beta_k \mathbf{K}_k\right)^{-1}\mathbf{K}_j\left(\sum_{k=0}^p \beta_k \mathbf{K}_k\right)^{-1}\mathbf{K}_l \\
&\qquad + \operatorname{tr}\left(\sum_{k=0}^p \beta_k \mathbf{K}_k'\right)^{-1}\mathbf{K}_j'\mathbf{K}_l\left(\sum_{k=0}^p \beta_k \mathbf{K}_k\right)^{-1}, \\
&\qquad\qquad j, l = 1, \cdots, p,
\end{aligned} \tag{2.23}
$$

$$
\begin{aligned}
-\mathscr{E}\frac{\partial^2}{\partial\beta_j\partial\sigma^2}\frac{1}{N}\log L &= -\frac{1}{\sigma^4}\mathscr{E}\operatorname{tr}\sum_{l=0}^p \beta_l\mathbf{K}_j'\mathbf{K}_l\mathbf{C} \\
&= -\frac{1}{\sigma^2}\operatorname{tr}\left(\sum_{k=0}^p \beta_k\mathbf{K}_k\right)^{-1}\mathbf{K}_j, \qquad j = 1, \cdots, p,
\end{aligned} \tag{2.24}
$$

$$
-\mathscr{E}\frac{\partial^2}{(\partial\sigma^2)^2}\frac{1}{N}\log L = -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6}\mathscr{E}\operatorname{tr}\sum_{k,l=0}^p \beta_k\beta_l\mathbf{K}_k'\mathbf{K}_l\mathbf{C} = \frac{T}{2\sigma^4}. \tag{2.25}
$$

As $N \to \infty$, the normalized maximum likelihood estimates have a limiting normal distribution with covariance matrix whose inverse has elements given by (2.23), (2.24), and (2.25).

2.2. *Autoregressive processes approximated by linear transformations.* The pure autoregressive process $\{y_t\}$ satisfies (1.1) for $\alpha_1 = \cdots = \alpha_q = 0$, that is,

$$
\sum_{s=0}^p \beta_s y_{t-s} = v_t, \tag{2.26}
$$

$t = \cdots, -1, 0, 1, \cdots$. Let $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)'$. Then the distribution of $y_1, \cdots, y_T$ is approximated by the distribution of $\mathbf{y}$ defined by (2.1) when $\mathbf{K}_g = \mathbf{L}^g$, $g = 0, 1, \cdots, p$, where $\mathbf{L}$ is

$$
\mathbf{L} = \begin{pmatrix} \mathbf{0} & 0 \\ \mathbf{I} & \mathbf{0} \end{pmatrix}; \tag{2.27}
$$

in (2.27) $\mathbf{I}$ is of order $T-1$. Then $\mathbf{L}^2$ is of the same form, but $\mathbf{I}$ is of order $T-2$ and the upper right-hand 0 is replaced by a 2×2 matrix $\mathbf{0}$. In general $\mathbf{L}^g$ has all 0's except for 1's g units below the main diagonal. Note that $\mathbf{L}^g\mathbf{L}^h = \mathbf{L}^{g+h}$, $g, h = 0, 1, \cdots$, and $\mathbf{L}^g = \mathbf{0}$, $g = T, T+1, \cdots$. In this case $\sum_{k=0}^p \beta_k\mathbf{K}_k = \sum_{k=0}^p \beta_k\mathbf{L}^k$ has i, jth component equal to β_{i-j} for $0 \leqq i-j \leqq p$ and 0 otherwise and is thus a lower triangular matrix with determinant 1. The components of (2.1) are

$$
\sum_{s=0}^{t-1} \beta_s y_{t-s} = v_t, \qquad t = 1, \cdots, p, \tag{2.28}
$$

and (2.26) for $t = p+1, \cdots, T$. The equation (2.28) is such that the sequence $y_1, \cdots, y_T$ does not start out as a stationary process. An alternative way of considering the equation (2.28) is that (2.26) holds with $y_0 = y_{-1} = \cdots = y_{-(p-1)} = 0$.

In this model we are often interested in $N = 1$, $\mathbf{y}_1 = \mathbf{y}$, and $\mathbf{C} = \mathbf{y}\mathbf{y}'$. Then

$$
\mathbf{y}^{(k)} = \mathbf{K}_k\mathbf{y} = \mathbf{L}^k\mathbf{y} = (0 \cdots 0\ y_1 \cdots y_{T-k})', \qquad k = 0, 1, \cdots, T-1, \tag{2.29}
$$

where there are k 0's, and $\mathbf{L}^k\mathbf{y} = \mathbf{0}$, $k = T, T+1, \cdots$. Since $|\sum_{k=0}^p \beta_k \mathbf{L}^k| = 1$, its derivatives vanish. The derivative equations (2.10) can be written in this case as

$$\sum_{k=1}^p \hat{\beta}_k \mathbf{y}^{(k)\prime}\mathbf{y}^{(l)} = -\mathbf{y}^{(0)\prime}\mathbf{y}^{(l)}, \qquad l = 1, \cdots, p. \tag{2.30}$$

In components these are

$$\sum_{k=1}^p \hat{\beta}_k \sum_{t=1}^T y_{t-k}y_{t-l} = -\sum_{t=1}^T y_t y_{t-l}, \qquad l = 1, \cdots, p, \tag{2.31}$$

where $y_0 = y_{-1} = \cdots = y_{-(p-1)} = 0$. These are the usual maximum likelihood estimates of $\beta_1, \cdots, \beta_p$ for initial values $y_0 = y_{-1} = \cdots = y_{-(p-1)} = 0$ or the "least squares estimates" since they minimize

$$\sum_{t=1}^T \left(\sum_{k=0}^p \beta_k y_{t-k}\right)^2. \tag{2.32}$$

[See Anderson (1971a), Sections 2.2 and 5.4, for example.]

Let

$$c_{-h} = c_h = \frac{1}{T}\sum_{i=1}^{T-h} y_i y_{i+h}, \qquad h = 0, 1, \cdots, T-1, \tag{2.33}$$

The right-hand side of (2.30) is $-Tc_l$. The sum $\sum_{t=1}^T y_{t-k}y_{t-l}$ differs from $Tc_{|k-l|}$ by omission of

$$\sum_{t=T-\max(k,l)+1}^{T-|k-l|} y_t y_{t+|k-l|}. \tag{2.34}$$

These terms can be added to the coefficients so as to make the equations agree with

$$\sum_{k=1}^p \hat{\beta}_k c_{l-k} = -c_l, \qquad l = 1, \cdots, p. \tag{2.35}$$

[See Anderson (1971a), Section 5.6, for example.] Then the estimates derived from (2.35) are the coefficients of a stationary process. [See Anderson (1971c), for example.]

In this case of $\mathbf{K}_k = \mathbf{L}^k$ the elements of the information matrix are N times

$$-\mathscr{E}\frac{\partial^2}{\partial\beta_j\partial\beta_l}\frac{1}{N}\log L = \operatorname{tr}\left(\sum_{k=0}^p \beta_k \mathbf{L}'^k\right)^{-1}\mathbf{L}'^j\mathbf{L}^l\left(\sum_{k=0}^p \beta_k\mathbf{L}^k\right)^{-1}, \qquad j, l = 1, \cdots, p, \tag{2.36}$$

$$-\mathscr{E}\frac{\partial^2}{\partial\beta_j\partial\sigma^2}\frac{1}{N}\log L = 0, \qquad j = 1, \cdots, p, \tag{2.37}$$

and (2.25).

It is of interest to compare the covariance matrix of $\mathbf{y}$ defined by (2.1) with that of T terms from the stationary process defined by (2.26). For $p = 1$ and $\beta_1 = \beta$ the covariances of the stationary process are

$$\sigma_{ts} = \sigma(t-s) = \sigma^2(-\beta)^{|t-s|}/(1-\beta^2), \qquad t, s = 1, \cdots, T, \tag{2.38}$$

and for $\mathbf{y}$ defined by (2.1) with $\mathbf{K}_k = \mathbf{L}^k$ the covariances are

$$\sigma_{ts} = [\sigma^2(-\beta)^{|t-s|}/(1-\beta^2)][1-\beta^{2\min(t,s)}], \qquad t, s = 1, \cdots, T. \tag{2.39}$$

For a stationary process $|\beta| < 1$, and hence (2.39) is close to (2.38) if t and s are large.

3. Estimation of coefficients of linear transformations to approximate moving average processes.

3.1. *General linear transformations.* Another model is defined by

$$\mathbf{y} = \textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k \mathbf{v} \,, \tag{3.1}$$

where $\mathbf{J}_0, \mathbf{J}_1, \cdots, \mathbf{J}_q$ are $q + 1$ known linearly independent $T \times T$ matrices, $\alpha_0 = 1$, and $\alpha_1, \cdots, \alpha_q$ are q parameters such that $\sum_{l=0}^{q} \alpha_l \mathbf{J}_l$ is nonsingular; we assume that there is at least one such set. Suppose $\mathbf{v}$ is a random vector with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Then the mean vector of $\mathbf{y}$ is $\mathscr{E}\mathbf{y} = \mathbf{0}$ and the covariance matrix is

$$\mathscr{E}\mathbf{y}\mathbf{y}' = \sigma^2 \textstyle\sum_{k,l=0}^{q} \alpha_k \alpha_l \mathbf{J}_k \mathbf{J}_l' = \sigma^2(\sum_{k=0}^{q} \alpha_k \mathbf{J}_k)(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l') \,. \tag{3.2}$$

If L denotes the likelihood function when $\mathbf{v}$ has a normal distribution, then

$$\begin{aligned} \frac{2}{N} \log L = & -T \log 2\pi - T \log \sigma^2 - \log |\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k|^2 \\ & - \frac{1}{\sigma^2} \operatorname{tr} (\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k')^{-1}(\sum_{k=0}^{q} \alpha_l \mathbf{J}_l)^{-1}\mathbf{C} \,, \end{aligned} \tag{3.3}$$

where $\mathbf{C}$ is defined by (2.5). The partial derivatives of $(2/N) \log L$ are

$$\begin{aligned} \frac{\partial}{\partial \alpha_j} \frac{2}{N} \log L = & -2 \operatorname{tr} (\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k)^{-1}\mathbf{J}_j \\ & + \frac{2}{\sigma^2} \operatorname{tr} (\textstyle\sum_{l=0}^{q} \alpha_l \mathbf{J}_l)^{-1}\mathbf{C}(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l')^{-1}\mathbf{J}_j'(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l')^{-1} \,, \\ & \qquad j = 1, \cdots, q \,, \end{aligned} \tag{3.4}$$

$$\frac{\partial}{\partial \sigma^2} \frac{2}{N} \log L = -\frac{T}{\sigma^2} + \frac{1}{\sigma^4} \operatorname{tr} (\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k')^{-1}(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l)^{-1}\mathbf{C} \,. \tag{3.5}$$

The likelihood equations can be written [with the second term on the right-hand side of (3.4) transposed]

$$\operatorname{tr} (\textstyle\sum_{k=0}^{q} \hat{\alpha}_k \mathbf{J}_k)^{-1}\mathbf{J}_j = \frac{1}{\hat{\sigma}^2} \operatorname{tr} (\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{J}_l)^{-1}\mathbf{J}_j(\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{J}_l)^{-1}\mathbf{C}(\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{J}_l')^{-1} \,, \qquad j = 1, \cdots, q \,, \tag{3.6}$$

$$\hat{\sigma}^2 = \frac{1}{T} \operatorname{tr} (\textstyle\sum_{k=0}^{q} \hat{\alpha}_k \mathbf{J}_k')^{-1}(\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{J}_l)^{-1}\mathbf{C} \,. \tag{3.7}$$

The second partial derivatives of $(1/N) \log L$ are evaluated by use of (2.7). Their expected values constitute the information matrix whose elements are N times

$$\begin{aligned} -\mathscr{E} \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \frac{1}{N} \log L = & \operatorname{tr} (\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k)^{-1}\mathbf{J}_i(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l)^{-1}\mathbf{J}_j \\ & + \operatorname{tr} (\textstyle\sum_{k=0}^{q} \alpha_k \mathbf{J}_k)^{-1}\mathbf{J}_i \mathbf{J}_j'(\sum_{l=0}^{q} \alpha_l \mathbf{J}_l')^{-1} \,, \\ & \qquad i, j = 1, \cdots, q \,, \end{aligned} \tag{3.8}$$

$$(3.9)\qquad -\mathscr{E}\frac{\partial^2}{\partial\alpha_j\partial\sigma^2}\frac{1}{N}\log L = \frac{1}{\sigma^2}\operatorname{tr}\mathbf{J}_j'(\textstyle\sum_{l=0}^q \alpha_l \mathbf{J}_l')^{-1}, \qquad j = 1, \cdots, q,$$

$$(3.10)\qquad -\mathscr{E}\frac{\partial^2}{\partial(\sigma^2)^2}\frac{1}{N}\log L = \frac{T}{2\sigma^4}.$$

As $N \to \infty$, the normalized maximum likelihood estimates have a limiting normal distribution with covariance matrix whose inverse has elements given by (3.8), (3.9) and (3.10).

The likelihood equations (3.6) and (3.7) cannot in general be solved explicitly since they are nonlinear in $\alpha_1, \cdots, \alpha_q$. However, iterative procedures can be based on expanding the partial derivatives in Taylor's series and using the linear terms. In general if $L(\mathbf{x} \mid \boldsymbol{\theta})$ is the likelihood function of a vector parameter $\boldsymbol{\theta}$ and $\mathbf{x}$ is a vector observation, the expansion is

$$(3.11)\qquad \frac{\partial \log L(\mathbf{x}\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left.\frac{\partial \log L(\mathbf{x}\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + \left.\frac{\partial^2 \log L(\mathbf{x}\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + \mathbf{R}(\mathbf{x}\mid\boldsymbol{\theta},\boldsymbol{\theta}^*),$$

where $\partial/\partial\boldsymbol{\theta}$ and $\partial^2/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ denote a vector and matrix of derivatives, respectively. A Newton–Raphson procedure is started by taking $\boldsymbol{\theta}^*$ as an initial estimate and solving for $\boldsymbol{\theta}$ the equation obtained by setting (3.11) equal to $\mathbf{0}$ and replacing $\mathbf{R}(\mathbf{x}\mid\boldsymbol{\theta},\boldsymbol{\theta}^*)$ by $\mathbf{0}$; the solution is an improved estimate. The next step is to set $\boldsymbol{\theta}^*$ equal to the improved estimate and proceed as before. The method of scoring is based on

$$(3.12)\qquad -\mathscr{E}\left.\frac{\partial^2 \log L(X,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\boldsymbol{\theta}-\boldsymbol{\theta}^*) = \left.\frac{\partial \log L(x,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}.$$

For both methods under certain conditions if $\boldsymbol{\theta}^*$ is a consistent estimate of the "true" value, the solution is a consistent, asymptotically efficient, and asymptotically normal estimate of $\boldsymbol{\theta}$. In suitable circumstances the sequence of iterates will converge to the maximum likelihood estimate.

In the present case let $\hat\alpha_1^{(0)}, \cdots, \hat\alpha_q^{(0)}, \hat\sigma_0^2$ be a set of initial estimates, and let $\hat\alpha_1^{(i)}, \cdots, \hat\alpha_q^{(i)}, \hat\sigma_i^2$ be the solution to the ith set of equations. It will be convenient to let

$$(3.13)\qquad \hat{\mathbf{A}}_{i-1} = \textstyle\sum_{k=0}^q \hat\alpha_k^{(i-1)}\mathbf{J}_k.$$

Then the ith iteration involves the equations

$$(3.14)\qquad \begin{aligned} &\textstyle\sum_{j=1}^q [\operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{J}_g\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{J}_j + \operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{J}_g\mathbf{J}_j'\hat{\mathbf{A}}_{i-1}'^{-1}](\hat\alpha_j^{(i)} - \hat\alpha_j^{(i-1)}) \\ &\quad + \frac{1}{\hat\sigma_{i-1}^2}\operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{J}_g\hat\sigma_i^2 \\ &\quad = \frac{1}{\hat\sigma_{i-1}^2}\operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{C}\hat{\mathbf{A}}_{i-1}'^{-1}\mathbf{J}_g'\hat{\mathbf{A}}_{i-1}'^{-1}, \qquad g = 1, \cdots, q, \end{aligned}$$

$$(3.15)\qquad \frac{1}{\hat\sigma_{i-1}^2}\textstyle\sum_{j=1}^q \operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{J}_j(\hat\alpha_j^{(i)} - \hat\alpha_j^{(i-1)}) + \dfrac{T}{2\hat\sigma_{i-1}^4}\hat\sigma_i^2 = \dfrac{1}{2\hat\sigma_{i-1}^4}\operatorname{tr}\hat{\mathbf{A}}_{i-1}'^{-1}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{C}.$$

If $\sigma^2 = 1$ and α_0 is a free parameter (not specified), the likelihood satisfies (3.3) with $\sigma^2 = 1$, the first partial derivatives are (3.4) for $j = 0, 1, \cdots, q$, the elements of the information matrix are N times (3.8) for $i, j = 0, 1, \cdots, q$, and the equations for scoring are (3.14) for $g = 0, 1, \cdots, q$.

3.2. *Moving average processes approximated by linear transformations.* The moving average process $\{y_t\}$ is (1.1) for $\beta_1 = \cdots = \beta_q = 0$, that is,

$$y_t = \sum_{j=0}^{q} \alpha_j v_{t-j} , \tag{3.16}$$

$t = \cdots, -1, 0, 1, \cdots$. Then the distribution of $y_1, \cdots, y_T$ is approximated by the distribution of $\mathbf{y}$ defined by (3.1) when $\mathbf{J}_g = \mathbf{L}^g$, $g = 0, 1, \cdots, q$. The components of (3.1) are

$$y_t = \sum_{j=0}^{t-1} \alpha_j v_{t-j} , \qquad t = 1, \cdots, q , \tag{3.17}$$

and (3.16) for $t = q + 1, \cdots, T$. The moving averages for the first q observations, represented by (3.17), are truncated.

The covariance matrix of T successive observations on the moving average process defined by (3.16) for $t = \cdots, -1, 0, 1, \cdots$ is

$$\sum_{g=0}^{q} \sigma_g \mathbf{G}_g , \tag{3.18}$$

where $\mathbf{G}_0 = \mathbf{I}$,

$$\mathbf{G}_g = \mathbf{L}^g + \mathbf{L}'^g , \qquad g = 1, \cdots, q , \tag{3.19}$$

$$\sigma_g = \sigma^2 \sum_{j=0}^{q-g} \alpha_j \alpha_{j+g} , \qquad g = 0, 1, \cdots, q . \tag{3.20}$$

[This is of the form considered in Anderson (1969), (1970), (1971b), and (1973).] The covariance matrix of $y_1, \cdots, y_T$ defined by (3.17) and (3.16) for $t = q + 1, \cdots, T$ differs from (3.18) only in the upper left-hand $q \times q$ submatrix. If T is large relative to q the difference between the two models will not be important; the model (3.1) with $\mathbf{J}_j = \mathbf{L}^j$ can be considered as an approximation to the moving average process.

When $\mathbf{J}_j = \mathbf{L}^j$, $\operatorname{tr} (\sum_{l=0}^{q} \alpha_l \mathbf{J}_l)\mathbf{J}_j = 0$, $j = 1, \cdots, q$. The likelihood equations (3.6) and (3.7) for $\hat{\alpha}_1, \cdots, \hat{\alpha}_q$ and $\hat{\sigma}^2$ (with $\hat{\alpha}_0 = 1$) are

$$\operatorname{tr} (\textstyle\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{L}^l)^{-1} \mathbf{L}^g (\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{L}^l)^{-1} \mathbf{C} (\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{L}'^l)^{-1} = 0 , \qquad g = 1, \cdots, q , \tag{3.21}$$

$$\hat{\sigma}^2 = \frac{1}{T} \operatorname{tr} (\textstyle\sum_{k=0}^{q} \hat{\alpha}_k \mathbf{L}'^k)^{-1} (\sum_{l=0}^{q} \hat{\alpha}_l \mathbf{L}^l)^{-1} \mathbf{C} . \tag{3.22}$$

The method of scoring leads to

$$\sum_{j=1}^{q} \operatorname{tr} \hat{\mathbf{A}}_{i-1}^{-1} \mathbf{L}^g \mathbf{L}'^j \hat{\mathbf{A}}_{i-1}'^{-1} (\hat{\alpha}_j^{(i)} - \hat{\alpha}_j^{(i-1)}) = \frac{1}{\hat{\sigma}_{i-1}^2} \operatorname{tr} \hat{\mathbf{A}}_{i-1}^{-1} \mathbf{C} \hat{\mathbf{A}}_{i-1}'^{-1} \mathbf{L}'^g \hat{\mathbf{A}}_{i-1}'^{-1} , \qquad g = 1, \cdots, q , \tag{3.23}$$

$$\hat{\sigma}_i^2 = \frac{1}{T} \operatorname{tr} \hat{\mathbf{A}}_{i-1}'^{-1} \hat{\mathbf{A}}_{i-1}^{-1} \mathbf{C} , \tag{3.24}$$

where

$$\hat{\mathbf{A}}_{i-1} = \sum_{l=0}^{q} \hat{\alpha}_l^{(i-1)} \mathbf{L}^l . \tag{3.25}$$

The set of linear equations (3.23) is solved for $\hat{\alpha}_1^{(i)} - \hat{\alpha}_1^{(i-1)}, \cdots, \hat{\alpha}_q^{(i)} - \hat{\alpha}_q^{(i-1)}$.

If $N = 1$ and $\mathbf{y}_1 = \mathbf{y}$, then $\mathbf{C} = \mathbf{y}\mathbf{y}'$. The equations (3.23) and (3.24) are

$$\sum_{j=1}^{q} \operatorname{tr} \hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^g(\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^j)'(\hat{\alpha}_j^{(i)} - \hat{\alpha}_j^{(i-1)}) = \frac{1}{\hat{\sigma}_{i-1}^2} (\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y})'\mathbf{L}^g\hat{\mathbf{A}}_{i-1}^{-1}(\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}) , \qquad g = 1, \cdots, q , \tag{3.26}$$

$$\hat{\sigma}_i^2 = \frac{1}{T} (\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y})'\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y} . \tag{3.27}$$

The calculation of $\mathbf{z} = \hat{\mathbf{A}}_{i-1}^{-1}\mathbf{y}$ can be done by solving

$$\sum_{l=0}^{q} \hat{\alpha}_l^{(i-1)}\mathbf{L}^l\mathbf{z} = \mathbf{y} . \tag{3.28}$$

The component equations are $z_1 = y_1$,

$$z_t + \sum_{s=1}^{t-1} \hat{\alpha}_s^{(i-1)} z_{t-s} = y_t , \qquad t = 2, \cdots, q , \tag{3.29}$$

$$z_t + \sum_{s=1}^{q} \hat{\alpha}_s^{(i-1)} z_{t-s} = y_t , \qquad t = q + 1, \cdots, T . \tag{3.30}$$

These can be solved successively for $z_2, \cdots, z_T$. Each component z_t involves at most q multiplications and the entire solution less than qT multiplications.

The first column of $\hat{\mathbf{A}}_{i-1}^{-1}$ can be obtained by solving (3.28) with $\mathbf{y}$ replaced by the first column of $\mathbf{I}$. Thus $z_1 = 1$ and the successive calculations are $z_t = -\sum_{s=1}^{t-1} \hat{\alpha}_s^{(i-1)} z_{t-s}$, $t = 2, \cdots, q$, $z_t = -\sum_{s=1}^{q} \hat{\alpha}_s^{(i-1)} z_{t-s}$, $t = q + 1, \cdots, T$. Because the $(j + 1)$th column of $\mathbf{I}$ is $\mathbf{L}^j$ times the first column of $\mathbf{I}$, the $(j + 1)$th column of $\hat{\mathbf{A}}_{i-1}^{-1}$ is simply $\mathbf{L}^j$ times the first column; that is, it is the first column displaced by j positions; the tth component of $\mathbf{L}^j\mathbf{z}$ is 0, $t = 1, \cdots, j$, and is z_{t-j}, $t = j + 1, \cdots, T$. Further $\mathbf{L}^j\mathbf{z} = \mathbf{0}$ for $j = T, T + 1, \cdots$. The calculation of $\hat{\mathbf{A}}_{i-1}^{-1}$ involves less than T_q multiplications. Note that $\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^j$ is $\hat{\mathbf{A}}_{i-1}^{-1}$ with each row displaced downwards by j rows.

If we let the elements of the first column of $\hat{\mathbf{A}}_{i-1}^{-1}$ be $\hat{\delta}_0^{(i-1)}, \hat{\delta}_1^{(i-1)}, \cdots, \hat{\delta}_{T-1}^{(i-1)}$, we can write

$$\hat{\mathbf{A}}_{i-1}^{-1} = \sum_{t=0}^{T-1} \hat{\delta}_t^{(i-1)}\mathbf{L}^t . \tag{3.31}$$

Then

$$\begin{aligned} \operatorname{tr} \hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^g\mathbf{L}'^j\hat{\mathbf{A}}_{i-1}'^{-1} &= \sum_{t,s=0}^{T-1} \hat{\delta}_t^{(i-1)}\hat{\delta}_s^{(i-1)} \operatorname{tr} \mathbf{L}^{k+t}\mathbf{L}'^{j+s} \\ &= \sum_{t=0}^{T-1-\max(g,j)} [T - t - \max(g, j)]\hat{\delta}_{t+|g-j|}^{(i-1)}\hat{\delta}_t^{(i-1)} \end{aligned} \tag{3.32}$$

because $\operatorname{tr} \mathbf{L}^h\mathbf{L}'^m = 0$, $l \neq m$, $\operatorname{tr} \mathbf{L}^h\mathbf{L}'^h = T - h$, $h = 0, 1, \cdots, T - 1$, and $\operatorname{tr} \mathbf{L}^h\mathbf{L}'^h = 0$, $h = T, T + 1, \cdots$.

Note that $\hat{\delta}_{t-1}^{(i-1)} = z_t$, $t = 1, 2, \cdots$, satisfy the homogeneous linear difference equation (3.30) with y_t replaced by 0. If the roots of the associated polynomial equation

$$\sum_{k=0}^{q} \hat{\alpha}_k^{(i-1)} x^{q-k} = 0 \tag{3.33}$$

are $x_1, \cdots, x_q$, then $\hat{\delta}_t^{(i-1)} = \sum_{l=1}^{q} k_l x_l^t$, $t = 0, 1, \cdots$, for suitable $k_1, \cdots, k_q$. If the roots of (3.33) are different and of absolute value less than 1 (which can be made the case with arbitrarily high probabilily if $\hat{\alpha}_1^{(i-1)}, \cdots, \hat{\alpha}_q^{(i-1)}$ are consistent estimates), $\hat{\delta}_t^{(i-1)}$ damps geometrically. Then (3.32) is approximately T times

$$\sum_{t=0}^{\infty} \hat{\delta}_{t+|g-j|}^{(i-1)} \hat{\delta}_t^{(i-1)} = \frac{\sigma_{AR}(g-j)}{\sigma^2}, \tag{3.34}$$

where $\sigma_{AR}(g-j)$ is the g, jth covariance of the autoregressive process corresponding to the coefficients 1, $\hat{\alpha}_1^{(i-1)}, \cdots, \hat{\alpha}_q^{(i-1)}$ and σ^2. [See Section 5.2 of Anderson (1971a).] The equations (3.26) are approximately

$$\sum_{j=1}^{q} \sigma_{AR}(g-j)(\hat{\alpha}_j^{(i)} - \hat{\alpha}_j^{(i-1)}) = d_g, \tag{3.35}$$

where d_g is the right-hand side of (3.26) divided by T. The solution to (3.35) is

$$\hat{\alpha}_j^{(i)} - \hat{\alpha}_j^{(i-1)} = \sum_{g=1}^{q} f_{jg} d_g, \qquad j = 1, \cdots, q, \tag{3.36}$$

where $(f_{jg}) = [\sigma_{AR}(g-j)]^{-1}$. The elements f_{jg} are the coefficients of the quadratic form of $w_1, \cdots, w_q$ having a normal distribution with covariance $\sigma_{AR}(g-j)$.

4. Estimation of coefficients of linear transformations when a covariance matrix has linear structure: autoregressive processes with moving average residuals. Let

$$\mathbf{B}\mathbf{y} = \mathbf{u}, \tag{4.1}$$

where

$$\mathbf{B} = \sum_{k=0}^{p} \beta_k \mathbf{K}_k, \tag{4.2}$$

$\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ are $p+1$ known linearly independent $T \times T$ matrices, $\beta_0 = 1$, and $\beta_1, \cdots, \beta_p$ are p parameters such that $\sum_{l=0}^{p} \beta_l \mathbf{K}_l$ is nonsingular; we assume there is at least one such set. Suppose $\mathbf{u}$ is a random vector with mean vector $\mathscr{E}\mathbf{u} = \mathbf{0}$ and covariance matrix

$$\mathbf{\Sigma}^u = \sum_{g=0}^{q} \sigma_g \mathbf{G}_g, \tag{4.3}$$

where $\mathbf{G}_0, \mathbf{G}_1, \cdots, \mathbf{G}_q$ are $q+1$ known linearly independent symmetric $T \times T$ matrices and $\sigma_0, \sigma_1, \cdots, \sigma_q$ are $q+1$ parameters such that $\sum_{g=0}^{q} \sigma_g \mathbf{G}_g$ is positive definite; we assume there is at least one such set. Then $\mathbf{y} = \mathbf{B}^{-1}\mathbf{u}$ has mean vector $\mathscr{E}\mathbf{y} = \mathbf{0}$ and covariance matrix

$$\mathscr{E}\mathbf{y}\mathbf{y}' = \mathbf{B}^{-1}\mathbf{\Sigma}^u\mathbf{B}'^{-1}. \tag{4.4}$$

If $\mathbf{u}$ is normally distributed, then $2/N$ times the logarithm of the likelihood is

$$\frac{2}{N} \log L = -T \log 2\pi + \log |\mathbf{B}|^2 - \log |\mathbf{\Sigma}^u| - \operatorname{tr} \mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{B}\mathbf{C}. \tag{4.5}$$

The partial derivatives are

$$\frac{\partial}{\partial \sigma_f} \frac{2}{N} \log L = -\operatorname{tr} (\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f + \operatorname{tr} \mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{B}\mathbf{C}, \qquad f = 0, 1, \cdots, q, \tag{4.6}$$

$$\frac{\partial}{\partial \beta_l} \frac{2}{N} \log L = 2 \operatorname{tr} \mathbf{B}^{-1}\mathbf{K}_l - 2 \operatorname{tr} \mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_l\mathbf{C}, \qquad l = 1, \cdots, p. \tag{4.7}$$

In case $\mathbf{K}_k = \mathbf{L}^k$ and $\mathbf{y}^{(k)} = \mathbf{L}^k\mathbf{y}_1$, $k = 0, 1, \cdots, p$, and $N = 1$, the derivative equations are

$$\operatorname{tr}\left(\textstyle\sum_{g=0}^{q} \hat{\sigma}_g \mathbf{G}_g\right)^{-1}\mathbf{G}_f = \textstyle\sum_{k,l=0}^{p} \hat{\beta}_k \hat{\beta}_l \mathbf{y}^{(k)\prime}\left(\sum_{g=0}^{q} \hat{\sigma}_g \mathbf{G}_g\right)^{-1}\mathbf{G}_f\left(\sum_{g=0}^{q} \hat{\sigma}_g \mathbf{G}_g\right)^{-1}\mathbf{y}^{(l)}, \quad f = 0, 1, \cdots, q, \tag{4.8}$$

$$\textstyle\sum_{k=1}^{p} \mathbf{y}^{(k)\prime}\left(\sum_{g=0}^{q} \hat{\sigma}_g \mathbf{G}_g\right)^{-1}\mathbf{y}^{(l)}\hat{\beta}_k = -\mathbf{y}^{(0)\prime}\left(\sum_{g=0}^{q} \hat{\sigma}_g \mathbf{G}_g\right)^{-1}\mathbf{y}^{(l)}, \quad l = 1, \cdots, p. \tag{4.9}$$

The equations are nonlinear in $\hat{\sigma}_0, \hat{\sigma}_1, \cdots, \hat{\sigma}_q$, but (4.9) is linear in $\hat{\beta}_1, \cdots, \hat{\beta}_p$.

The second partial derivatives of $(2/N) \log L$ are

$$\begin{aligned}\frac{\partial^2}{\partial\sigma_f\,\partial\sigma_h}\frac{2}{N}\log L = {} & \operatorname{tr}(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_h \\ & - 2\operatorname{tr}\mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_h(\mathbf{\Sigma}^u)^{-1}\mathbf{BC}, \\ & f, h = 0, 1, \cdots, q,\end{aligned} \tag{4.10}$$

$$\frac{\partial^2}{\partial\sigma_f\,\partial\beta_l}\frac{2}{N}\log L = 2\operatorname{tr}\mathbf{B}'(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_l\mathbf{C}, \quad f = 0, 1, \cdots, q, \quad l = 1, \cdots, p. \tag{4.11}$$

$$\frac{\partial^2}{\partial\beta_k\,\partial\beta_l}\frac{2}{N}\log L = -2\operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_k\mathbf{B}^{-1}\mathbf{K}_l - 2\operatorname{tr}\mathbf{K}_k'(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_l\mathbf{C}, \quad k, l = 1, \cdots, p. \tag{4.12}$$

The information matrix has elements that are N times

$$-\mathscr{E}\frac{\partial^2}{\partial\sigma_f\,\partial\sigma_h}\frac{1}{N}\log L = \tfrac{1}{2}\operatorname{tr}(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{G}_h, \quad f, h = 0, 1, \cdots, q, \tag{4.13}$$

$$-\mathscr{E}\frac{\partial^2}{\partial\sigma_f\,\partial\beta_l}\frac{1}{N}\log L = -\operatorname{tr}\mathbf{G}_f(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_l\mathbf{B}^{-1}, \quad f = 0, 1, \cdots, q, \quad l = 1, \cdots, p, \tag{4.14}$$

$$-\mathscr{E}\frac{\partial^2}{\partial\beta_k\,\partial\beta_l}\frac{1}{N}\log L = \operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_k\mathbf{B}^{-1}\mathbf{K}_l + \operatorname{tr}\mathbf{B}^{-1}\mathbf{\Sigma}^u\mathbf{B}'^{-1}\mathbf{K}_k'(\mathbf{\Sigma}^u)^{-1}\mathbf{K}_l, \quad k, l = 1, \cdots, p. \tag{4.15}$$

Let $\hat{\mathbf{B}}_{i-1}$ be (4.2) with β_k replaced by $\hat{\beta}_k^{(i-1)}$, and let $\hat{\mathbf{\Sigma}}_{i-1}^u$ be (4.3) with σ_g replaced by $\hat{\sigma}_g^{(i-1)}$. The method of scoring leads to the following iterative procedure:

$$\begin{aligned}&\textstyle\sum_{h=0}^{q} \operatorname{tr}(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_h\hat{\sigma}_h^{(i)} \\ &\quad - 2\textstyle\sum_{l=1}^{p} \operatorname{tr}\mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{K}_l\hat{\mathbf{B}}_{i-1}^{-1}(\hat{\beta}_l^{(i)} - \hat{\beta}_l^{(i-1)}) \\ &\quad = \operatorname{tr}\hat{\mathbf{B}}_{i-1}'(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\hat{\mathbf{B}}_{i-1}\mathbf{C}, \qquad f = 0, 1, \cdots, q,\end{aligned} \tag{4.16}$$

$$\begin{aligned}&-2\textstyle\sum_{h=0}^{q} \operatorname{tr}\mathbf{G}_h(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{K}_k\hat{\mathbf{B}}_{i-1}^{-1}\hat{\sigma}_h^{(i)} + 2\textstyle\sum_{l=1}^{p}[\operatorname{tr}\hat{\mathbf{B}}_{i-1}^{-1}\mathbf{K}_k\hat{\mathbf{B}}_{i-1}^{-1}\mathbf{K}_l \\ &\quad + \operatorname{tr}\hat{\mathbf{B}}_{i-1}^{-1}\hat{\mathbf{\Sigma}}_{i-1}^u\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{K}_k'(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{K}_l](\hat{\beta}_l^{(i)} - \hat{\beta}_l^{(i-1)}) \\ &\quad = -2\operatorname{tr}\mathbf{C}\hat{\mathbf{B}}_{i-1}'(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{K}_k, \qquad j = 1, \cdots, p.\end{aligned} \tag{4.17}$$

If $\mathbf{K}_j = \mathbf{L}^j$, $j = 0, 1, \cdots, p$, $N = 1$, and $\mathbf{y}_1 = \mathbf{y}$, the scoring equations are

$$\begin{aligned}(4.18)\qquad &\textstyle\sum_{h=0}^{q} \operatorname{tr} (\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_h \hat{\sigma}_h^{(i)} \\ &\quad - 2 \textstyle\sum_{l=1}^{p} \operatorname{tr} \mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^l\hat{\mathbf{B}}_{i-1}^{-1}(\hat{\beta}_l^{(i)} - \hat{\beta}_l^{(i-1)}) \\ &\qquad = \mathbf{y}'\hat{\mathbf{B}}_{i-1}'^{-1}(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{G}_f(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\hat{\mathbf{B}}_{i-1}\mathbf{y}\,, \qquad f = 0, 1, \cdots, q\,.\end{aligned}$$

$$\begin{aligned}(4.19)\qquad &-2 \textstyle\sum_{h=0}^{q} \operatorname{tr} \mathbf{G}_h(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^k\hat{\mathbf{B}}_{i-1}^{-1}\hat{\sigma}_h^{(i)} \\ &\quad + 2 \textstyle\sum_{l=1}^{p} \operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\hat{\mathbf{\Sigma}}_{i-1}^u\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^l(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^k(\hat{\beta}_l^{(i)} - \hat{\beta}_l^{(i-1)}) \\ &\qquad = -2\mathbf{y}'\hat{\mathbf{B}}_{i-1}'(\hat{\mathbf{\Sigma}}_{i-1}^u)^{-1}\mathbf{L}^k\mathbf{y}\,, \qquad k = 1, \cdots, p\,.\end{aligned}$$

When $\mathbf{\Sigma}^u$ represents the covariance matrix of a moving average process, $\mathbf{G}_0=\mathbf{I}$, $\mathbf{G}_g$ is given by (3.19), $g = 1, \cdots, q$, and σ_g is given by (3.20), $g = 0, 1, \cdots, q$. Since $\mathbf{L}^g\mathbf{L}'^h$ is $\mathbf{L}^{g-h}$, $h \leqq g$, except for at most h 1's being replaced by 0's, $\hat{\mathbf{\Sigma}}_{i-1}^u$ and $\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^l$ almost commute and the coefficient of $(\hat{\beta}_l^{(i)} - \hat{\beta}_l^{(i-1)})$ in (4.19) is approximately

$$(4.20)\qquad 2 \operatorname{tr} \hat{\mathbf{B}}_{i-1}^{-1}\mathbf{L}^k\mathbf{L}'^l\hat{\mathbf{B}}_{i-1}'^{-1}\,.$$

The matrix $\hat{\mathbf{B}}_{i-1}$ has the same form as $\hat{\mathbf{A}}_{i-1}$; hence, the computation of expressions involving $\hat{\mathbf{B}}_{i-1}^{-1}$ can be carried out as suggested in Section 3 for $\hat{\mathbf{A}}_{i-1}^{-1}$. The computation of expressions involving $\hat{\mathbf{\Sigma}}_{i-1}^{-1}$ was considered in Anderson (1971b), (1973).

Initial (consistent) estimates of the parameters can be obtained by equating the first $p + q + 1$ theoretical and observed covariances. From initial estimates $\hat{\sigma}_0^{(0)}, \hat{\sigma}_0^{(0)}, \cdots, \hat{\sigma}_q^{(0)}$ (4.9) can be solved for improved initial estimates $\hat{\beta}_1^{(0)}, \cdots, \hat{\beta}_p^{(0)}$. Then (4.18) and (4.19) can be solved for $\hat{\sigma}_0^{(1)}, \hat{\sigma}_1^{(1)}, \cdots, \hat{\sigma}_q^{(1)}$ [the right-hand side of (4.19) being 0], and these can be used in (4.9) again.

The second derivatives, (4.10), (4.11), and (4.12), can be used to set up a Newton–Raphson method.

5. Estimation of coefficients of linear transformations; autoregressive processes with moving average residuals. Here we combine the models of Sections 2 and 3. Let

$$(5.1)\qquad \textstyle\sum_{l=0}^{p} \beta_l\mathbf{K}_l\mathbf{y} = \sum_{k=0}^{q} \alpha_k\mathbf{J}_k\mathbf{v}\,,$$

where $\mathbf{K}_0, \mathbf{K}_1, \cdots, \mathbf{K}_p$ are $p + 1$ known linearly independent $T \times T$ matrices, $\mathbf{J}_0, \mathbf{J}_1, \cdots, \mathbf{J}_q$ are $q + 1$ known linearly independent $T \times T$ matrices, $\beta_0 = \alpha_0 = 1$, $\beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q$ are $p + q$ parameters such that $\sum_{l=0}^{p} \beta_l\mathbf{K}_l$ and $\sum_{k=0}^{p} \alpha_k\mathbf{J}_k$ are nonsingular; we assume there is at least one such set. Suppose $\mathbf{v}$ is a random vector with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. Then $\mathbf{y} = \mathbf{B}^{-1}\mathbf{A}\mathbf{v}$ has mean vector $\mathbf{0}$ and covariance matrix

$$(5.2)\qquad \mathscr{E}\mathbf{y}\mathbf{y}' = \sigma^2\mathbf{B}^{-1}\mathbf{A}\mathbf{A}'\mathbf{B}'^{-1}\,,$$

where $\mathbf{A} = \sum_{k=0}^{q} \alpha_k\mathbf{J}_k$ and $\mathbf{B} = \sum_{l=0}^{p} \beta_l\mathbf{K}_l$.

If $\mathbf{y}_1, \cdots, \mathbf{y}_N$ are N observations on $\mathbf{y}$ with a normal distribution, $2/N$ times the logarithm of the likelihood function L is

$$\frac{2}{N}\log L = -T\log 2\pi - T\log\sigma^2 + \log|\mathbf{B}|^2 - \log|\mathbf{A}|^2 - \frac{1}{\sigma^2}\operatorname{tr}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}. \tag{5.3}$$

The partial derivatives are

$$\frac{\partial}{\partial\alpha_g}\frac{2}{N}\log L = -2\operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_g + \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_g'\mathbf{A}'^{-1}, \qquad g = 1, \cdots, q, \tag{5.4}$$

$$\frac{\partial}{\partial\beta_h}\frac{2}{N}\log L = 2\operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_h - \frac{2}{\sigma^2}\operatorname{tr}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{C}, \qquad h = 1, \cdots, p, \tag{5.5}$$

$$\frac{\partial}{\partial\sigma^2}\frac{2}{N}\log L = -\frac{T}{\sigma^2} + \frac{1}{\sigma^4}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}, \tag{5.6}$$

In case $\mathbf{K}_k = \mathbf{L}^k$, $k = 0, 1, \cdots, p$, $\mathbf{J}_g = \mathbf{L}^g$, $g = 0, 1, \cdots, q$, $N = 1$, and $\mathbf{y}_1 = \mathbf{y}$, the derivative equations are

$$\textstyle\sum_{k,l=0}^{p}\hat\beta_k\hat\beta_l\mathbf{y}'\mathbf{L}'^k(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}'^j)^{-1}(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}'^j)^{-2}\mathbf{L}^{g+l}\mathbf{y}, \qquad g = 1, \cdots, q, \tag{5.7}$$

$$\textstyle\sum_{k=1}^{q}\mathbf{y}'\mathbf{L}'^k(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}'^j)^{-1}(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}^j)^{-1}\mathbf{L}^h\mathbf{y}\hat\beta_k = -\mathbf{y}'(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}'^j)^{-1}(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}^j)^{-1}\mathbf{L}^h\mathbf{y}, \qquad h = 1, \cdots, p, \tag{5.8}$$

$$\hat\sigma^2 = \frac{1}{T}\textstyle\sum_{k,l=0}^{p}\hat\beta_k\hat\beta_l\mathbf{y}'\mathbf{L}'^k(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}'^j)^{-1}(\sum_{j=0}^{q}\hat\alpha_j\mathbf{L}^j)^{-1}\mathbf{L}^l\mathbf{y}. \tag{5.9}$$

The second partial derivatives of $(2/N)\log L$ are

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_g\,\partial\alpha_f}\frac{2}{N}\log L = {} & 2\operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_g\mathbf{A}^{-1}\mathbf{J}_f - \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_f\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_g'\mathbf{A}'^{-1}\\ & - \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_f'\mathbf{A}'^{-1}\mathbf{J}_g'\mathbf{A}'^{-1}\\ & - \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_g'\mathbf{A}'^{-1}\mathbf{J}_f'\mathbf{A}'^{-1}, \qquad g, f = 1, \cdots, q,\end{aligned} \tag{5.10}$$

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_g\,\partial\beta_h}\frac{2}{N}\log L = {} & \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_g'\mathbf{A}'^{-1}\\ & + \frac{2}{\sigma^2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_g\mathbf{A}^{-1}\mathbf{K}_h\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}, \qquad g = 1, \cdots, q,\ h = 1, \cdots, p,\end{aligned} \tag{5.11}$$

$$\frac{\partial^2}{\partial\beta_h\,\partial\beta_j}\frac{2}{N}\log L = -2\operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_j\mathbf{B}^{-1}\mathbf{K}_h - \frac{2}{\sigma^2}\operatorname{tr}\mathbf{K}_j'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{C}, \qquad h, j = 1, \cdots, p, \tag{5.12}$$

$$(5.13)\qquad \frac{\partial^2}{\partial\alpha_g\,\partial\sigma^2}\frac{2}{N}\log L = -\frac{2}{\sigma^4}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{J}_g{}'\mathbf{A}'^{-1}\,,\qquad g=1,\cdots,q\,,$$

$$(5.14)\qquad \frac{\partial^2}{\partial\beta_h\,\partial\sigma^2}\frac{2}{N}\log L = \frac{2}{\sigma^4}\operatorname{tr}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\,,\qquad h=1,\cdots,p\,,$$

$$(5.15)\qquad \frac{\partial^2}{\partial(\sigma^2)^2}\frac{2}{N}\log L = \frac{T}{\sigma^4} - \frac{2}{\sigma^6}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{B}'\mathbf{A}'^{-1}\,.$$

The elements of the information matrix are N times

$$(5.16)\qquad -\mathscr{E}\frac{\partial^2}{\partial\alpha_g\,\partial\alpha_f}\frac{1}{N}\log L = \operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_g\mathbf{A}^{-1}\mathbf{J}_f + \operatorname{tr}\mathbf{A}^{-1}\mathbf{J}_g\mathbf{J}_f{}'\mathbf{A}'^{-1}\,,\qquad g,f=1,\cdots,q\,,$$

$$(5.17)\qquad -\mathscr{E}\frac{\partial^2}{\partial\alpha_g\,\partial\beta_h}\frac{1}{N}\log L = -\operatorname{tr}\mathbf{J}_g\mathbf{A}^{-1}\mathbf{K}_h\mathbf{B}^{-1} - \operatorname{tr}\mathbf{J}_g{}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{B}^{-1}\mathbf{A}\,,\qquad g=1,\cdots,q\,,\ h=1,\cdots,p\,,$$

$$(5.18)\qquad -\mathscr{E}\frac{\partial^2}{\partial\beta_h\,\partial\beta_j}\frac{1}{N}\log L = \operatorname{tr}\mathbf{B}^{-1}\mathbf{K}_j\mathbf{B}^{-1}\mathbf{K}_h + \operatorname{tr}\mathbf{K}_j{}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{K}_h\mathbf{B}^{-1}\mathbf{A}\mathbf{A}'\mathbf{B}'^{-1}\,,\qquad h,j=1,\cdots,p\,,$$

$$(5.19)\qquad -\mathscr{E}\frac{\partial^2}{\partial\alpha_g\,\partial\sigma^2}\frac{1}{N}\log L = \frac{1}{\partial^2}\operatorname{tr}\mathbf{J}_g{}'\mathbf{A}'^{-1}\,,\qquad g=1,\cdots,q\,,$$

$$(5.20)\qquad -\mathscr{E}\frac{\partial^2}{\partial\beta_h\,\partial\sigma^2}\frac{1}{N}\log L = -\frac{1}{\sigma^2}\operatorname{tr}\mathbf{K}_h\mathbf{B}^{-1}\,,\qquad h=1,\cdots,p\,,$$

$$(5.21)\qquad -\mathscr{E}\frac{\partial^2}{\partial(\sigma^2)^2}\frac{1}{N}\log L = \frac{T}{2\sigma^4}\,.$$

The method of scoring can be developed from these results.

When $\mathbf{K}_k = \mathbf{L}^k$, $k = 0, 1, \cdots, p$, and $\mathbf{J}_g = \mathbf{L}^g$, $g = 0, 1, \cdots, q$, the first term in each of (5.16), (5.17), and (5.18) is 0, and (5.19) and (5.20) are 0. Moreover, because $\mathbf{L}^g$, $g = 0, 1, \cdots$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ are polynomials in $\mathbf{L}$ they commute. Then (5.17) is $-\operatorname{tr}\mathbf{B}^{-1}\mathbf{L}^h\mathbf{L}'^g\mathbf{A}'^{-1}$ and (5.18) is $\mathbf{B}^{-1}\mathbf{L}^h\mathbf{L}'^j\mathbf{B}'^{-1}$. When $N = 1$ and $\mathbf{y}_1 = \mathbf{y}$, the scoring equations are

$$(5.22)\qquad \begin{aligned}&\textstyle\sum_{f=1}^{q}\operatorname{tr}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^g\mathbf{L}'^f\hat{\mathbf{A}}_{i-1}'^{-1}(\hat{\alpha}_f{}^{(i)} - \hat{\alpha}_f{}^{(i-1)})\\ &\qquad - \textstyle\sum_{h=1}^{p}\operatorname{tr}\mathbf{L}'^g\hat{\mathbf{A}}_{i-1}'^{-1}\mathbf{L}^h\hat{\mathbf{B}}_{i-1}^{-1}(\hat{\beta}_h{}^{(i)} - \hat{\beta}_h{}^{(i-1)})\\ &\qquad\qquad = \frac{1}{\hat{\sigma}_{i-1}^2}\mathbf{y}'\hat{\mathbf{B}}_{i-1}'\hat{\mathbf{A}}_{i-1}'^{-1}\hat{\mathbf{A}}_{i-1}^{-2}\mathbf{L}^g\mathbf{y}\,,\qquad g=1,\cdots,q\,,\end{aligned}$$

$$(5.23)\qquad \begin{aligned}&-\textstyle\sum_{f=1}^{q}\operatorname{tr}\mathbf{L}^j\hat{\mathbf{B}}_{i-1}^{-1}\mathbf{L}'^f\hat{\mathbf{A}}_{i-1}'^{-1}(\hat{\alpha}_f{}^{(i)} - \hat{\alpha}_f{}^{(i-1)})\\ &\qquad + \textstyle\sum_{h=1}^{p}\operatorname{tr}\hat{\mathbf{B}}_{i-1}'^{-1}\mathbf{L}'^j\mathbf{L}^h\hat{\mathbf{B}}_{i-1}^{-1}(\hat{\beta}_h{}^{(i)} - \hat{\beta}_h{}^{(i-1)})\\ &\qquad\qquad = -\frac{1}{\hat{\sigma}_{i-1}^2}\mathbf{y}'\hat{\mathbf{B}}_{i-1}'\hat{\mathbf{A}}_{i-1}'^{-1}\hat{\mathbf{A}}_{i-1}^{-1}\mathbf{L}^j\mathbf{y}\,,\qquad j=1,\cdots,p\,,\end{aligned}$$

and (5.9) for $\hat{\alpha}_j = \hat{\alpha}_j{}^{(i-1)}$, $\hat{\beta}_R = \hat{\beta}_R{}^{(i-1)}$, and $\hat{\sigma}^2 = \hat{\sigma}_i{}^2$.

The computations involving $\hat{\mathbf{A}}_{i-1}^{-1}$ and $\hat{\mathbf{B}}_{i-1}^{-1}$ have been discussed earlier. Initial estimates of $\alpha_1, \cdots, \alpha_q$ and σ^2 can be obtained by "factoring" the initial estimate of $f_u(\lambda)$. Then (5.8) can be used to obtain (improved) initial estimates of $\beta_1, \cdots, \beta_p$. Then (5.22) and (5.23) are solved for $\hat{\alpha}_1^{(1)} - \hat{\alpha}_1^{(0)}, \cdots, \hat{\alpha}_q^{(1)} - \hat{\alpha}_q^{(0)}$, and these can be used in (5.8) again.

The second derivatives, (5.10) to (5.15), can be used to develop a Newton–Raphson method.

6. Asymptotic theory. The exact distributions of the maximum likelihood estimates developed in this paper cannot be obtained in closed form in general. However, asymptotic distributions can be found. If $N \to \infty$ we have the case of repeated observations on the random vector $\mathbf{y}$; in the case of time series, however, N may be 1 and $T \to \infty$. In either case when consistent estimates of the parameters are used as initial estimates (with either the coefficients or variance of order $1/T^{\frac{1}{2}}$ in probability), the estimates obtained in the first step of the iteration procedure are consistent, asymptotically normal, and asymptotically efficient (when normalized by $N^{\frac{1}{2}}$ or $T^{\frac{1}{2}}$, as the case may be).

In the model of Section 2.1 no iteration is involved and the asymptotic properties are the usual ones as the number of observations N increases. The model of Section 2.2 is the autoregressive model with the first p observations treated as fixed ($y_{-p+1} = \cdots = y_0 = 0$); the asymptotic theory as $T \to \infty$ is well known. [See Anderson (1971a), Section 5.5, for example.]

For each of the models in the other sections an iterative procedure was proposed. If the initial estimates are consistent, the matrix of coefficients of the linear equations is a consistent estimate of the information matrix of one observation. The asymptotic distribution of the right-hand sides is normal with covariance matrix equal to this matrix. It then follows that the estimates have the stated properties. We shall carry out the details of the proof only for the model of Section 3.2, which shows the pattern.

Let $\mathbf{y} = (y_1, \cdots, y_T)'$ be defined by

$$\mathbf{y} = \sum_{k=0}^{q} \alpha_k \mathbf{L}^k \mathbf{v} = \mathbf{A}\mathbf{v} . \tag{6.1}$$

We shall let $T \to \infty$. We assume that the zeros of (1.10) are less than 1 in absolute value. For $i = 1$ (3.26) and (3.27) are

$$\sum_{j=1}^{q} \operatorname{tr} \hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\mathbf{L}'^j\hat{\mathbf{A}}_0'^{-1}\hat{\alpha}_j^{(1)} = \frac{1}{\hat{\sigma}_0^2}\mathbf{y}'\hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}\mathbf{y} - \operatorname{tr} \hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0'^{-1} , \quad g = 1, \cdots, q , \tag{6.2}$$

$$\hat{\sigma}_1^2 = \frac{1}{T}\mathbf{y}'\hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}\mathbf{y} . \tag{6.3}$$

We shall show that

$$\operatorname{plim}_{T\to\infty} \frac{1}{T} \operatorname{tr} \hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\mathbf{L}'^j\hat{\mathbf{A}}_0'^{-1} = \lim_{T\to\infty} \frac{1}{T} \operatorname{tr} \mathbf{A}^{-1}\mathbf{L}^g\mathbf{L}'^j\mathbf{A}'^{-1} . \tag{6.4}$$

The right-hand side is

$$\sum_{i=0}^{\infty} \delta_{i+|g-j|} \delta_i . \tag{6.5}$$

The left-hand side is the probability limit of

$$\sum_{i=0}^{T-1-\max(g,j)} \left[1 - \frac{i + \max(g, j)}{T}\right] \hat{\delta}_{i+|g-j|}^{(0)} \hat{\delta}_i^{(0)} . \tag{6.6}$$

With arbitrarily high probability for T sufficiently large $\hat{\alpha}_1^{(0)}, \cdots, \hat{\alpha}_q^{(0)}$ are such that the roots of the polynomial equation with these coefficients are less than 1 in absolute value, in fact, are less than $\rho < 1$ for some ρ [greater than the largest zero of (1.10)]. Then (6.6) converges in probability to (6.5).

We can write (6.2) as

$$\begin{aligned} \sum_{j=1}^{q} \frac{1}{T} \operatorname{tr} \hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\mathbf{L}'^j\hat{\mathbf{A}}_0'^{-1}T^{\frac{1}{2}}(\hat{\alpha}_j^{(1)} - \alpha_j) \\ = \frac{1}{T^{\frac{1}{2}}}\left[\frac{1}{\hat{\sigma}_0^2} \mathbf{y}' \hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}\mathbf{y} - \operatorname{tr} \hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\mathbf{A}'\hat{\mathbf{A}}_0'^{-1}\right], \\ g = 1, \cdots, q . \end{aligned} \tag{6.7}$$

We want to show that the right-hand sides have a limiting normal distribution with means 0 and covariance matrix with elements (6.4).

Consider

$$\begin{aligned} \frac{1}{T^{\frac{1}{2}}\sigma^2} \mathbf{y}'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{L}^g\mathbf{A}^{-1}\mathbf{y} &= \frac{1}{T^{\frac{1}{2}}\sigma^2} \mathbf{v}'\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} \\ &= \sum_{i=1}^{\infty} \delta_i \frac{1}{T^{\frac{1}{2}}\sigma^2} \mathbf{v}'\mathbf{L}^{i+g}\mathbf{v} \\ &= \sum_{i=0}^{T-g-1} \delta_i \frac{1}{T^{\frac{1}{2}}\sigma^2} \sum_{t=1}^{T-(i+g)} v_t v_{t+i+g} . \end{aligned} \tag{6.8}$$

For any n the set $(1/T^{\frac{1}{2}}) \sum_{t=1}^{T} v_t v_{t+1}, \cdots, (1/T^{\frac{1}{2}}) \sum_{t=1}^{T} v_t v_{t+n}$ has a limiting normal distribution [Theorem 7.7.6 of Anderson (1971a), for example] with means 0 and covariances

$$\begin{aligned} \frac{1}{T} \mathscr{E} \sum_{t,s=1}^{T} v_t v_{t+j} v_s v_{s+h} = \frac{1}{T} \mathscr{E} \sum_{t=1}^{T} v_t^2 v_{t+j} v_{t+h} \\ = \sigma^4 , \qquad j = h = 1, \cdots, \\ = 0 , \qquad j \neq h . \end{aligned} \tag{6.9}$$

Then the set

$$\sum_{i=0}^{n-q} \delta_i \frac{1}{T^{\frac{1}{2}}\sigma^2} \sum_{t=1}^{T} v_t v_{t+i+g} , \qquad g = 1, \cdots, q , \tag{6.10}$$

has a limiting normal distribution with means 0 and covariances

$$\frac{1}{\sigma^4} \sum_{i,h=0}^{n-q} \delta_i \delta_h \frac{1}{T} \mathscr{E} \sum_{t,s=1}^{T} v_t v_{t+i+g} v_s v_{s+j+h} = \sum_{i=0}^{n-q-|g-j|} \delta_i \delta_{i+|g-j|} , \tag{6.11}$$

which has the limit as $n \to \infty$ of (6.5). That the limiting distribution of (6.8) is the limit as $n \to \infty$ of the limiting distribution of (6.10) is justified by Corollary 7.7.1 of T. W. Anderson (1971a), for example. Note that

$$(6.12)\qquad \mathscr{E}\left(\sum_{i=n-q+1}^{T-g-1} \delta_i \frac{1}{T^{\frac{1}{2}}\sigma^2} \sum_{t=1}^{T-(i+g)} v_t v_{t+i+g}\right)^2 \leqq \sum_{i=n-q+1}^{T-g-1} \delta_i^2 \leqq \sum_{i=n-q+1}^{\infty} \delta_i^2 .$$

Now consider the difference of (6.8) and (6.7), which is

$$(6.13)\qquad \frac{1}{T^{\frac{1}{2}}}\left[\frac{1}{\sigma^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} - \frac{1}{\hat\sigma_0^2}\mathbf{v}'\mathbf{A}'\hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}\mathbf{A}\mathbf{v} + \operatorname{tr}\hat{\mathbf{A}}_0^{-1}\mathbf{L}^g\mathbf{A}'\hat{\mathbf{A}}_0'^{-1}\right].$$

We write

$$(6.14)\qquad \hat{\mathbf{A}}_0^{-1} = \mathbf{A}^{-1} - \hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}.$$

Then (6.13) is

$$\begin{aligned}
&\frac{1}{T^{\frac{1}{2}}}\Big\{\frac{1}{\sigma^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} - \frac{1}{\hat\sigma_0^2}[\mathbf{v}'\mathbf{A}'(\mathbf{A}'^{-1} - \mathbf{A}'^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1})\\
&\quad\times(\mathbf{A}^{-1} - \hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1})\mathbf{L}^g(\mathbf{A}^{-1} - \hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1})\mathbf{A}\mathbf{v}]\\
&\quad+\operatorname{tr}(\mathbf{A}^{-1} - \hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1})\mathbf{L}^g\mathbf{A}'(\mathbf{A}'^{-1} - \mathbf{A}'^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1})\Big\}\\
&= \frac{1}{T^{\frac{1}{2}}}\Big\{\Big(\frac{1}{\sigma^2} - \frac{1}{\hat\sigma_0^2}\Big)\mathbf{v}'\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} + \frac{1}{\hat\sigma_0^2}\mathbf{v}'(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v}\\
(6.15)\qquad&\quad+\frac{1}{\hat\sigma_0^2}\mathbf{v}'\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} + \frac{1}{\hat\sigma_0^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{v}\\
&\quad-\operatorname{tr}\mathbf{A}^{-1}\mathbf{L}^g(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1} - \frac{1}{\hat\sigma_0^2}\mathbf{v}'(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\\
&\quad\times\mathbf{A}^{-1}\mathbf{L}^g\mathbf{v} - \frac{1}{\hat\sigma_0^2}\mathbf{v}'(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\mathbf{A}^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{v}\\
&\quad-\frac{1}{\hat\sigma_0^2}\mathbf{v}'\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{v}\\
&\quad+\frac{1}{\hat\sigma_0^2}\mathbf{v}'(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}\mathbf{L}^g\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{v}\\
&\quad+\operatorname{tr}\hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}\mathbf{L}^g(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\Big\}.
\end{aligned}$$

The first term on the right-hand side of (6.15) has probability limit 0 because (6.8) has a limiting normal distribution and $\operatorname{plim}_{T\to\infty}\hat\sigma_0^2 = \sigma^2 > 0$. Each of the third and fourth terms is

$$\begin{aligned}
(6.16)\qquad &\frac{1}{\hat\sigma_0^2}\frac{1}{T^{\frac{1}{2}}}\mathbf{v}'\hat{\mathbf{A}}_0^{-1}\mathbf{L}^g(\hat{\mathbf{A}}_0 - \mathbf{A})\mathbf{A}^{-1}\mathbf{v}\\
&\quad= \frac{1}{\hat\sigma_0^2}\sum_{k=1}^{q}(\hat\alpha_k^{(0)} - \alpha_k)\sum_{i,j=0}^{\infty}\hat\delta_i^0\delta_j\frac{1}{T^{\frac{1}{2}}}\mathbf{v}'\mathbf{L}^{g+i+j+k}\mathbf{v}.
\end{aligned}$$

Let

$$(6.17)\qquad W_{hT} = \sum_{j=0}^{\infty}\delta_j\frac{1}{T^{\frac{1}{2}}}\mathbf{v}'\mathbf{L}^{h+j}\mathbf{v}.$$

Then

$$\mathscr{E} W_{hT}^2 \leqq \sigma^4 \sum_{j=0}^{\infty} \delta_j^2 . \tag{6.18}$$

We can write

$$\sum_{i,j=0}^{\infty} \hat{\delta}_i^0 \delta_j \frac{1}{T^{\frac{1}{2}}} \mathbf{v}' \mathbf{L}^{g+k+i+j} \mathbf{v} = \sum_{i=0}^{\infty} \hat{\delta}_i^0 W_{g+k+i,T} . \tag{6.19}$$

With arbitrarily high probability $|\hat{\delta}_i^{(0)}| < \rho_0^i$ for some ρ_0 such that $0 < \rho_0 < \rho_1 < 1$. Then the square of (6.19) is less than

$$\sum_{i=0}^{\infty} \left(\frac{\hat{\delta}_i^{(0)}}{\rho_1^i}\right)^2 \sum_{i=0}^{\infty} \rho_1^{2i} W_{g+k+i,T}^2 . \tag{6.20}$$

Since the expected value of the second sum (which is nonnegative) is less than $\sigma^4 \sum_{j=0}^{\infty} \delta_j^2/(1-\rho_1^2)$, (6.20) is bounded in probability. Since $\text{plim}_{T\to\infty} \hat{\alpha}_k^{(0)} = \alpha_k$, (6.16) has probability limit 0. The second term and fifth terms give

$$\begin{aligned} &\frac{1}{\hat{\sigma}_0^2} \frac{1}{T^{\frac{1}{2}}} \mathbf{v}'(\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\mathbf{A}^{-1}\mathbf{L}^g \mathbf{v} - \frac{1}{T^{\frac{1}{2}}} \operatorname{tr} (\hat{\mathbf{A}}_0 - \mathbf{A})'\hat{\mathbf{A}}_0'^{-1}\mathbf{A}^{-1}\mathbf{L}^g \\ &\quad = \frac{1}{\hat{\sigma}_0^2} \sum_{k=1}^{q} (\hat{\alpha}_k^{(0)} - \alpha_k) \\ &\qquad \times \left[\sum_{i,j=0}^{\infty} \hat{\delta}_i^0 \delta_j \frac{1}{T^{\frac{1}{2}}} (\mathbf{v}'\mathbf{L}'^{k+i}\mathbf{L}^{g+j}\mathbf{v} - \sigma^2 \operatorname{tr} \mathbf{L}'^{k+i}\mathbf{L}^{g+j}) \right. \\ &\qquad \left. + \frac{1}{T^{\frac{1}{2}}} \sum_{i,j=0}^{\infty} \hat{\delta}_i^0 \delta_j (\sigma^2 - \hat{\sigma}_0^2) \operatorname{tr} \mathbf{L}'^{k+i}\mathbf{L}^{g+j} \right]. \end{aligned} \tag{6.21}$$

The sum of δ_j times the first parenthesis in the brackets is treated like (6.17); note that the parenthesis has mean 0 and the right-hand side of (6.18) is a bound on the expected value of its square. The same argument carries through. If $T^{\frac{1}{2}}(\hat{\sigma}^2 - \hat{\sigma}_0^2)$ is bounded in probability [or $T^{\frac{1}{2}}(\hat{\alpha}_k^{(0)} - \alpha_k)$ is], then the second term converges to 0 in probability. The other terms in (6.15) are treated similarly.

It follows from these results that the solutions to (6.7), namely $T^{\frac{1}{2}}(\hat{\alpha}_1^{(1)} - \alpha_1), \cdots, T^{\frac{1}{2}}(\hat{\alpha}_q^{(1)} - \alpha_q)$, have a limiting normal distribution with means 0 and a covariance matrix that is the inverse of the information matrix.

The sample covariances c_h defined by (2.33) are consistent estimates of $\sigma(h)$, $h = 0, 1, \cdots, p+q$. From these can be obtained consistent estimates of $\beta_1, \cdots, \beta_p, \sigma_u(0), \cdots, \sigma_u(q)$ and of $\beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q$, and σ^2 as described in Section 5.8.1 of Anderson (1971a).

Acknowledgments. This paper was begun while the author was a visiting scholar at the Center for Advanced Study in the Behavioral Sciences. The author is indebted to Paul Shaman for helpful suggestions on exposition.

REFERENCES

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

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

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

ANDERSON, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in Probability and Statistics* 1–24. Univ. of North Carolina Press.

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

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

ANDERSON, T. W. (1971c). The stationarity of an estimated autoregressive process. Technical Report No. 7, Contract N00014-67-A-0112-0030, Stanford Univ.

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

ÅSTRÖM, KARL-JOHANN, and BOHLIN, TORSTEN (1966). Numerical identification of linear dynamic systems from normal operating records. *Theory of Self Adaptive Control Systems* (P. H. Hammond, ed.) 96–111. Plenum, New York.

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

CLEVENSON, M. LAWRENCE (1970). Asymptotically efficient estimates of the parameters of a moving average time series. Statistics Department, Stanford Univ.

DURBIN, J. (1959). Efficient estimation of parameters in moving-average models. *Biometrika* **46** 306–316.

DURBIN, J. (1960). The fitting of time-series models. *Rev. Inst. Internat. Statist.* **28** 233–244.

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

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

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

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

WALKER, A. M. (1961). Large-sample estimation for moving-average models. *Biometrika* **48** 343–357.

WALKER, A. M. (1962). Large sample estimation of parameters for autoregressive processes with moving-average residuals. *Biometrika* **49** 117–131.

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

WHITTLE, PETER (1951). *Hypothesis Testing in Time Series Analysis*. Almqvist and Wicksell, Uppsala.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305