The Annals of Statistics
1979, Vol. 7, No. 2, 297–315

THE STRONG CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS FOR ARMA PROCESSES[1]

BY J. RISSANEN AND P. E. CAINES

IBM Research Laboratory and Harvard University

The strong consistency of the maximum likelihood parameter estimation method is established for multivariate Gaussian stochastic processes possessing autoregressive moving average (ARMA) representations. The demonstration in this paper exploits the ergodic theorem together with results from linear prediction theory.

1. Introduction. In this paper we consider the problem of estimating the parameters $A_1, \cdots, A_n, B_0, \cdots, B_n$ in the autoregressive moving average (ARMA) scheme

$$y_t + A_1 y_{t-1} + \cdots + A_n y_{t-n} = B_0 u_t + \cdots + B_n u_{t-n} \tag{1.1}$$

by the maximum likelihood technique. We take y and u to be real p component zero mean Gaussian stationary processes, and, in addition, we take u to be an orthonormal process, i.e., $Eu_t u_s^T = I\delta_{t,s}$ where $\delta_{t,s}$ takes the value 1 for $t = s$ and 0 otherwise. Consequently, the parameters $\{A_1, \cdots, A_n, B_0, \cdots, B_n\}$ are real $(p \times p)$matrices. Without loss of generality, y is assumed to be of full rank, i.e., $E\varepsilon_t \varepsilon_t^T > 0$ for all $t \in Z$, where $\varepsilon_t := y_t - y_t^*$, when y_t^* denotes $E(y_t | y_i, i \leqslant t-1)$, the conditional expectation of y_t with respect to $\{y_i, i \leqslant t-1\}$. Here Z denotes the integers.

Since it is assumed that the observed process y is Gaussian, the distribution of a sample $\{y_t; t \in Z_+\}$, where Z_+ denotes the nonnegative integers, is determined by the second order statistics of y. But these statistics are functions only of $(A_1, \cdots, A_n, B_0, \cdots, B_n)$. As a result, the likelihood function depends only upon the data $\{y_0, \cdots, y_t; t \in Z_+\}$ and the values assigned to $(A_1, \cdots, A_n, B_0, \cdots, B_n)$.

We shall denote the list of matrix coefficients $(A_1, \cdots, A_n, B_0, \cdots, B_n)$ by the $p \times (2n+1)$ matrix expression (A, B). Further, the parameter vector $\psi = \psi(A, B)$ will denote a list of the entries of (A, B) taken in some arbitrary order. Clearly the vectors ψ are in one to one correspondence with the pairs of matrices (A, B). To each vector ψ there corresponds a rational matrix $\Phi(z)$ associated with the system (1.1). This is given by

$$\Phi(z) = A^{-1}(z)B(z) = \Phi_0 + \Phi_1 z + \Phi_2 z^2 + \cdots, \tag{1.2}$$

Received March 1974; revised January 1978.

[1]Research sponsored in part by the National Science Foundation Grants NSF-GK 10656x3, NSF-GK 31627, National Aeronautics and Space Agency Grant NGL 05-020-007 and the Joint Services Electronics Program under Contract N00014-75-C-0648.

AMS 1970 *subject classifications.* Primary F10; Secondary F99, M10, M20, N15.

Key words and phrases. Estimation, parametric case, asymptotic theory, time series.

where

(1.3a) $$A(z) := I + A_1 z + \cdots + A_n z^n$$

and

(1.3b) $$B(z) := B_0 + B_1 z + \cdots + B_n z^n.$$

The representation

$$y_t = \Phi_0 u_t + \Phi_1 u_{t-1} + \Phi_2 u_{t-2} + \cdots$$

is called the innovations representation or Wold decomposition of y (see e.g., [19], [22]), and in the engineering literature the sequence $\{\Phi_0, \Phi_1, \cdots\}$ is described as the impulse response of the transfer function $A^{-1}(z)B(z)$.

Since y is a stationary process the roots of the polynomial $\det A(z)$ must lie outside the closed unit disk in the complex plane. Now suppose the matrix polynomial $B(z)$ in (1.3b) is restricted so that $\det B(z)$ has no roots inside the unit circle. Then, modulo right multiplication by orthogonal matrices, the transfer function $\Phi(z)$ is in one to one correspondence with the second order statistics of the process y (see [19]). Furthermore, the second order statistics of y are given by the coefficients of the spectral function $\Phi(z)\Phi^T(z^{-1})$. As a result, we see that knowledge of (i) a set of parameters (A, B) for (1.1), (ii) the entries of the rational transfer function $A^{-1}(z)B(z)$, (iii) the impulse response $\{\Phi_0, \Phi_1, \cdots\}$, and (iv) the spectral function $\Phi(z)\Phi^T(z^{-1})$ are equivalent, in the sense that knowledge of any one of these quantities allows, in principle, for the computation of the other three quantities. In Section 3 we discuss further the question of the parameterization of (1.1).

The study of the weak consistency and asymptotic normality of maximum likelihood estimators for (1.1) in the scalar case was initiated by Whittle [21]. In this paper we prove the strong consistency of such estimators by using an idea suggested by Kendall and Stuart [12, pages 41–43]. Our result was outlined in [5]. In [8] (see also [7]) Dunsmuir and Hannan also establish the strong consistency of m.l. estimators for (1.1). They use weaker conditions than those used here, but employ more elaborate proof techniques.

The samples $\{y_t; t \in Z_+\}$ generated by (1.1) are not statistically independent, consequently the present case differs from the situation analyzed both in [12] and in the classic paper on maximum likelihood estimators by Wald [20]. We deal with this by employing the ergodic theorem as suggested in an earlier study by Åström and Bohlin [1]. The basic idea of the proof is straightforward in that it consists of a comparison of the asymptotic value of the likelihood function at various points in the parameter space.

We construct the likelihood function via the innovations process $\{e_t; t \in Z_+\}$, where e_t denotes $y_t - \hat{y}_t = y_t - E(y_t | y_{t-1}, \cdots, y_0)$. This construction is different from that used in [7, 8, 21] and we believe it to be of independent interest. Notice that by assumption the observed process y, generated by (1.1), is stationary; in fact, one may interpret y as the output process of the system (1.1) after it has been

initialized in the remote past and has then been permitted to attain steady state behavior. On the other hand, the innovations process e is nonstationary and is generated by the time varying filter (prediction algorithm) given below in (2.3–2.7). We remark that the stationary form of the likelihood function is commonly used in practice because its evaluation requires substantially less computation than the exact form, see e.g., [2].

A version of the techniques employed in this paper may also be used to establish the strong consistency of the so-called prediction error estimators [3, 4, 13, 14] for processes that are not necessarily Gaussian. Furthermore, in [4] the asymptotic normality of a large class of prediction error estimators is established and an asymptotically efficient class of prediction error estimators is characterized. However, for theoretical simplicity it is assumed in [3, 4] that all prediction algorithms involved in the analysis have attained steady state. Hence these results are not immediately applicable to the analysis of the behavior of the maximum likelihood estimator for the parameters of (1.1).

We point out that the strong consistency of m.l. estimators for Markov processes has been established by Roussas [18]. He uses the ergodic theorem in a proof along the same lines as that of Wald [20]. The process y in (1.1) is not a Markov process and consequently the results of [18] cannot be applied directly to the m.l. estimation of ψ. However, it might be possible to extend these results to the present case by constructing a Markovian model whose states are estimates of some particular state sequence of (1.1).

The organization of this paper is as follows: in Section 2 the likelihood function is constructed and it is shown how this function may be computed in terms of the prediction errors of the observed process y. In Section 3 the main consistency theorem is established, while the lemmas required in its demonstration are proved in the Appendix.

2. The likelihood function. We obtain the likelihood function of the sequence of observations $\{y_0, \cdots, y_t;\ t \in Z_+\}$ by the method of orthogonalization. To describe this procedure, let $\mathbf{H}$ denote the Hilbert space spanned by $\{y_t^i;\ 1 \leqslant i \leqslant p,\ t \in Z\}$, where y_t^i denotes the ith component of y_t, and let $\mathbf{Y}_{0,t}$ denote the subspace of H spanned by $\{y_s^i;\ 1 \leqslant i \leqslant p,\ 0 \leqslant s \leqslant t\}$. We shall denote the orthogonal projection of y_t^i on $\mathbf{Y}_{0,t-1}$ by $\hat{y}_t^i$, and write $\hat{y}_t$ for the vector $(\hat{y}_t^1, \cdots, \hat{y}_t^p)^T$. The process $\{\hat{y}_t;\ t \in Z_+\}$ constitutes the sequence of linear least squares estimates of the process $\{y_t;\ t \in Z_+\}$ and, since y is Gaussian, $\hat{y}_t = E(y_t|y_{t-1}, \cdots, y_0)$. As described in the introduction, we define the innovations process e by

$$e_t = y_t - \hat{y}_t, \qquad t \in Z_+; \tag{2.1a}$$

further we set

$$\Sigma_t = Ee_te_t^T, \qquad t \in Z_+. \tag{2.1b}$$

Notice that $\hat{y}_0 = 0$, by the zero mean assumption on the y process. Further, since $Ee_te_t^T \geqslant E\varepsilon_t\varepsilon_t^T > 0$ we see that Σ_t is invertible for all $t \in Z_+$.

The Gaussian density function $f_N(y_0, \cdots, y_N; \psi)$ for $(y_0, \cdots, y_N)$ parameterized by $\psi = \psi(A, B)$ is obtained by iterating the formula

$$f_N(y_0, \cdots, y_N; \psi) = f_N(y_N|y_0, \cdots, y_{N-1}; \psi) f_{N-1}(y_0, \cdots, y_{N-1}; \psi),$$

where $f_N(y_N|y_0, \cdots, y_{N-1}; \psi)$ denotes the conditional Gaussian density function. This yields

$$(2.2)\quad f_N(y_0, \cdots, y_N; \psi) = (2\pi)^{-((N+1)p)/2} \Pi_{t=0}^N (\det \Sigma_t)^{-\frac{1}{2}} \exp -\frac{1}{2} e_t^T \Sigma_t^{-1} e_t.$$

In the expression (2.2) the parameter ψ enters implicitly via the prediction errors $\{e_t;\ t \in Z_+\}$ and their covariances $\{\Sigma_t;\ t \in Z_+\}$ in a manner explained below.

It has been shown by Rissanen [16] and Rissanen and Barbosa [17] that the least squares prediction errors $\{e_t;\ t \in Z_+\}$ for the process (1.1) are generated by the time dependent ARMA scheme

$$(2.3)\quad e_t + C_{t,t-1}e_{t-1} + \cdots + C_{t,t-n}e_{t-n} = y_t + A_1 y_{t-1} + \cdots + A_n y_{t-n}, \qquad t \geqslant n,$$

where

$$(2.4)\qquad C_{t,j} = B_{t,j}B_{j,j}^{-1}, \qquad t \geqslant n,\ t-1 \geqslant j \geqslant t-n,$$

and where the matrices $B_{t,j}$, $t \geqslant 0$, $t \geqslant j \geqslant \max(t-n, 0)$, are obtained recursively from the equations:

$$(2.5)\qquad \begin{aligned} B_{t,t-n} &= R_n\big(B_{t-n,t-n}^T\big)^{-1}, \\ B_{t,t-n+1} &= \big(R_{n-1} - B_{t,t-n}B_{t-n+1,t-n}^T\big)\big(B_{t-n+1,t-n+1}^T\big)^{-1}, \\ &\cdots \qquad\qquad \cdots \\ B_{t,t}B_{t,t}^T &= \big(R_0 - B_{t,t-1}B_{t,t-1}^T - \cdots - B_{t,t-n}B_{t,t-n}^T\big), \qquad t \geqslant 2n, \end{aligned}$$

where all $B_{t,t}$, $t \in Z_+$, are taken to be upper triangular with positive elements in the diagonal. It remains to describe the matrices R_i, $0 \leqslant i \leqslant n$, which appear in (2.5), and the initial conditions for the schemes (2.3) and (2.5).

The $(p \times p)$ matrices R_i are given in terms of the parameter vector ψ as follows:

$$R_i = B_0B_i^T + B_1B_{i+1}^T + \cdots + B_{n-i}B_n^T, \qquad 0 \leqslant i \leqslant n.$$

Let the $(2np \times 2np)$ covariance matrix R be specified by

$$(2.6)\quad R = E\begin{bmatrix} B_0u_{2n-1} + \cdots + B_nu_{n-1} \\ \vdots \\ B_0u_n + \cdots + B_nu_0 \\ y_{n-1} \\ \vdots \\ y_0 \end{bmatrix} \Big[(B_0u_{2n-1} + \cdots + B_nu_{n-1})^T, \cdots, (B_0u_n + \cdots + B_nu_0)^T, y_{n-1}^T, \cdots, y_0^T \Big] > 0$$

Then the initial conditions $B_{t,j}$, $0 \leqslant t \leqslant 2n$, $t \geqslant j \geqslant \max(t-n, 0)$, for (2.5) are given as the $(p \times p)$ matrix elements of the unique $(2np \times 2np)$ upper triangular factor D of R with positive elements on the diagonal:

$$R = DD^T.$$

R can be calculated in terms of ψ from (1.1),but explicit expressions for these terms are not needed here.

The initial conditions for (2.3) are now given in terms of the $B_{t,j}$ $0 \leqslant t$, $t \geqslant j \geqslant \max(t-n, 0)$ via

$$\begin{bmatrix} e_{n-1} \\ \vdots \\ e_1 \end{bmatrix} = \begin{bmatrix} y_{n-1} \\ \vdots \\ y_1 \end{bmatrix} - \begin{bmatrix} B_{n-1,n-2} & \cdots & B_{n-1,0} \\ \vdots & & \vdots \\ 0 & & B_{1,0} \end{bmatrix} \begin{bmatrix} B_{n-2,n-2} & \cdots & B_{n-2,0} \\ \vdots & & \vdots \\ 0 & & B_{0,0} \end{bmatrix}^{-1} \begin{bmatrix} y_{n-2} \\ \vdots \\ y_0 \end{bmatrix} \tag{2.7}$$

and $e_0 = y_0$.

Having completed the description of the algorithm (2.3) we observe from (2.1) and (2.3) that the estimates $\{\hat{y}_t;\ t \in Z_+\}$ are generated by the scheme

$$\hat{y}_t + C_{t,t-1}\hat{y}_{t-1} + \cdots + C_{t,t-n}\hat{y}_{t-n} = (C_{t,t-1} - A_1)y_{t-1} + \cdots + (C_{t,t-n} - A_n)y_{t-n}, \qquad t \geqslant n, \tag{2.8}$$

with initial conditions $\hat{y}_0, \cdots, \hat{y}_{n-1}$ computed using $e_0, \cdots, e_{n-1}$ given by (2.7) above.

REMARK. There exist procedures for the prediction of the process y in (1.1) which use Markov state models (see [9]). We use the ARMA prediction algorithm (2.3)–(2.7) because it is better suited to proving the technical results in the appendices. For practical computation of the estimates $\{\hat{y}_t;\ t \in Z_+\}$ there exist new algorithms (see e.g., [15, 16]) which require an order fewer arithmetic operations than (2.3)–(2.7) or the procedure in [9].

For a given observation sequence $\{y_0, \cdots, y_N\}$, and parameter ψ, the function $f_N(y_0, \cdots, y_N; \psi)$ is called the likelihood function on the observations at $\psi \in \mathbb{R}^{(2n+1)p}$. We now define the scaled log-likelihood function

$$L_N(y_0, \cdots, y_N; \psi) = p \log 2\pi - \frac{2}{N+1} \log f_N(y_0, \cdots, y_N; \psi). \tag{2.9}$$

From the construction of the process $\{e_t;\ t \in Z_+\}$ it is known that

$$Ee_te_t^T = \Sigma_t = B_{t,t}B_{t,t}^T, \qquad t \in Z, \tag{2.10}$$

where the sequence $B_{t,t}$; $t \in Z_+$ is generated by (2.5) and (2.6). Consequently (2.2) and (2.10) yield

$$L_N(y_0, \cdots, y_N; \psi) = \frac{1}{N+1} \Sigma_{i=0}^{N} (\log \det \Sigma_i(\psi) + e_i^T(\psi)\Sigma_i^{-1}(\psi)e_i(\psi)), \tag{2.11}$$

where the argument ψ is now shown explicitly in the various quantities to emphasize the dependence upon $\psi = \psi(A, B)$.

We have defined $L_N(y_0, \cdots, y_N; \psi)$ because the expression (2.11) is especially convenient to work with in the remainder of the paper. We observe that $f_N(y_0, \cdots, y_N; \psi)$ is maximized on a compact set S at some point ψ^* if and only if $L_N(y_0, \cdots, y_N; \psi)$ is minimized over S at ψ^*. It follows that minimization of $L_N(y_0, \cdots, y_N; \psi)$ generates the maximum likelihood estimate of the true parameter for (1.1).

3. Main result. There is an inherent nonuniqueness in the parameterization (A, B) for the process (1.1). An infinite set of such matrix pairs will yield via (1.1) a process y with the same innovations representation impulse response and the same spectral function. This is true even when the McMillan degree (see e.g. [11], page 286) of $\Phi(z)$ has been fixed. In this case the multiplicity of the representations is equivalent to the nonuniqueness of observable and controllable realizations of $\Phi(z)$, in other words, to the nonuniqueness of Markovian representations for y of minimal state dimension. Such minimal Markovian representations are characterized by the well-known state space isomorphism theorem (see e.g. [11], page 317).

The questions introduced above lead to the problem of the description of a set of canonical forms for a set of transfer function matrices $\{\Phi(z)\}$, i.e., a bijective relation between a parameter space and a set of impulse responses. Such canonical forms are sought in either matrix fraction form, i.e., as a pair of polynomial matrices $\{A(z), B(z)\}$ such that $A^{-1}(z)B(z) = \Phi(z)$, or in Markov state space form, i.e., as a quartet of matrices $\{F, G, H, J\}$ such that $H(Iz^{-1} - F)^{-1}G + J = \Phi(z)$. Now let $\{\Phi(z)\}_\delta$ denote the set of transfer functions with McMillan degree (= minimal state space dimension) δ. Then it has been established in [6] (see also [7], [8]) that there exists a family of local canonical forms for $\{\Phi(z)\}_\delta$ in state space form, yielding the charts of an analytical manifold whose dimension depends upon p and the fixed McMillan degree δ. We shall not go into this important topic any further here, but we shall obtain a suitable compact parameter set S in the following way: let the true impulse response $\mathring{\Phi} := \{\mathring{\Phi}_0, \mathring{\Phi}_1, \cdots\}$ have a rational transfer function $\mathring{\Phi}(z)$ with McMillan degree δ, i.e., $\mathring{\Phi}(z) \in \{\Phi(z)\}_\delta$. Assume the poles of $\mathring{\Phi}(z)$ and the zeroes of $\det \mathring{\Phi}(z)$ lie outside the closed unit disk in the complex plane. Further, assume that the matrix $\mathring{\Phi}_0$ is upper triangular with positive elements on the diagonal. This latter assumption merely avoids the nonuniqueness of normalized innovations representations and spectral factors that arises from right multiplication of $\Phi_0(z) := A_\theta^{-1}(z)B_\theta(z)$ by orthogonal matrices (see [24]). Let a typical element of the appropriate manifold of canonical forms be denoted by θ and a specified resulting matrix fraction representation by $(A(\theta), B(\theta))$. Next, let S_0 denote the open subset of this manifold for which the zeroes of $\det A_\theta(z)$ and $\det B_\theta(z)$ lie outside the closed unit disk in the complex plane, where

$$A_\theta(z) := I + \Sigma_{i=1}^n A_i(\theta)z^i \qquad \text{and} \qquad B_\theta(z) := \Sigma_{i=0}^n B_i(\theta)z^i,$$

for n sufficiently large. Then S is taken to be any compact subset of S_0 which contains the parameter $\mathring{\theta}$ corresponding to $\mathring{\Phi}(z)$.

Now all formulae of Sections 1 and 2, which were given for the parameterization $\psi = \psi(A, B)$, are valid, without any modification, for the matrix fraction representations $(A(\theta), B(\theta))$, $\theta \in S_0$. Because this is the case we have, as before, that the maximum likelihood estimators θ^N of $\mathring{\theta}$ will be generated by minimizing $L_N(y_0, \cdots, y_N; \theta)$ over S.

We point out that Dunsmuir and Hannan [7, 8] maximize the likelihood function over a not necessarily compact parameter set and, in addition, do not require the poles of $\mathring{\Phi}(z)$ to lie in a compact set which is known a priori nor the zeroes of $\det \mathring{\Phi}(z)$ to lie outside the closed unit disk. However, as remarked earlier, the stronger conditions of this paper permit the use of simpler proof techniques.

The main result of this paper may now be stated.

THEOREM. *Let the stationary Gaussian stochastic process y be generated by* (1.1) *and hence have the representation*

$$y_t = \mathring{\Phi}_0 u_t + \mathring{\Phi}_1 u_{t-1} + \cdots, \qquad t \in Z,$$

where u is a stationary Gaussian orthonormal process. Let $\mathring{\Phi}(z) = A_{\mathring{\theta}}^{-1}(z) B_{\mathring{\theta}}(z)$ with $\mathring{\theta} \in S$ as described above, and let θ^N minimize

$$(3.1)\quad L_N(y_0, \cdots, y_N; \theta) = \frac{1}{N+1} \Sigma_{i=0}^{N} \left(\log \det \Sigma_i(\theta) + e_i^T(\theta) \Sigma_i^{-1}(\theta) e_i(\theta)\right)$$

over the compact set S. Then $\theta^N \to \mathring{\theta}$ in the manifold topology of S_0 a.s. as $N \to \infty$. In other words, the maximum likelihood estimator for the parameter of the process y is strongly consistent at $\mathring{\theta}$.

PROOF. The scheme (2.3)–(2.7) generates the nonstationary process $\{e_t = e_t(\theta), \theta \in S, t \in Z_+\}$ when the y process is taken as input. We define a closely related stationary process $\{\varepsilon_t(\theta); t \in Z\}$ given by

$$(3.2)\qquad \varepsilon_t + C_1 \varepsilon_{t-1} + \cdots + C_n \varepsilon_{t-n} = y_t + A_1 y_{t-1} + \cdots + A_n y_{t-n}, \qquad t \in Z,$$

where $C_j := B_j B_0^{-1}$, $1 \leqslant j \leqslant n$, and B_0 is invertible by the definition of S. Now it may be shown (see e.g. [17]) that $\varepsilon(\mathring{\theta})$ is the orthogonal process of least squares prediction errors, i.e.,

$$\varepsilon_t(\mathring{\theta}) = y_t - y_t^* = \varepsilon_t, \qquad t \in Z,$$

where, as before, y_t^* denotes the orthogonal projection of y_t on the subspace $\mathbf{Y}_{-\infty, t-1}$ of $\mathbf{H}$ spanned by $\{y_i^j;\ 1 \leqslant j \leqslant p,\ -\infty < i \leqslant t-1\}$. From (3.2) it is clear that the process y^* is generated by the scheme

$$(3.3)\quad y_t^* + C_1 y_{t-1}^* + \cdots + C_n y_{t-n}^* = (C_1 - A_1) y_{t-1} + \cdots + (C_n - A_n) y_{t-n}, \qquad t \in Z.$$

This is the stationary version of (2.8) as is demonstrated in Lemma 1 in the Appendix.

Notice that unless $\theta = \mathring{\theta}$, i.e., unless $A_\theta^{-1}(z)B_\theta(z) = A_{\mathring{\theta}}^{-1}(z)B_{\mathring{\theta}}(z) = \mathring{\Phi}(z)$ the sequence $\{\varepsilon_t(\theta), t \in Z\}$ is not orthogonal to the sequence of spaces $\{\mathbf{Y}_{-\infty, t-1}; t \in Z\}$.

Since y is a Gaussian process y_t^* is identical to $E(y_t|\mathbf{Y}_{-\infty, t-1})$. It follows that

$$\lambda^T(E\varepsilon_0(\theta)\varepsilon_0^T(\theta))\lambda = E(\lambda^T\varepsilon_0(\theta))^2, \qquad \lambda \in \mathbb{R}^p,$$

is minimized with respect to $\theta \in S$ at $\varepsilon_0(\mathring{\theta}) = y_0 - y_0^*$ for all $\lambda \in \mathbb{R}^p$. Consequently,

$$E\varepsilon_0(\theta)\varepsilon_0^T(\theta) \geqslant E\varepsilon_0(\mathring{\theta})\varepsilon_0^T(\mathring{\theta}), \tag{3.4}$$

for all $\theta \in S$. Further, the impulse response of the transfer function (3.3) is in one-to-one correspondence with the impulse response Φ of (1.1) when $\theta \in S$. As a result, equality holds in (3.4) if and only if $\theta = \mathring{\theta}$.

In the rest of the proof a fixed sample $\{y_t; t \in Z\}$ is selected and the sequence of m.l. estimates is computed in terms of the observations $\{y_t; t \in Z_+\}$. This sample will be taken to be a member of the set $\mathcal{Y} \subset \{y; y \text{ generated by (1.1)}\}$ such that

$$L_N(y_0, \cdots, y_N; \theta) \to L(\theta) \qquad \text{as} \quad N \to \infty,$$

uniformly for $\theta \in S$, where

$$L(\theta) = \log \det \Sigma(\theta) + E\varepsilon_0^T(\theta)\Sigma^{-1}(\theta)\varepsilon_0(\theta),$$

and where the matrix $\Sigma(\theta)$ is defined to be the limit as $t \to \infty$, of $\Sigma_t(\theta) = B_{t,t}(\theta)B_{t,t}^T(\theta)$. Lemma 1 shows this latter limit exists and equals $B_0(\theta)B_0^T(\theta)$. The fact that the set $\mathcal{Y}$ is of measure 1 is proved in Lemma 3. This implies that the convergence of $L_N(y_0, \cdots, y_N; \theta)$ takes place almost surely as $N \to \infty$ uniformly in $\theta \in S$.

We shall now show that $L(\theta) \geqslant L(\mathring{\theta})$ for all θ in S. To do this, consider the function

$$\log \det X + \operatorname{trace}(QX^{-1})$$

of the positive definite $(p \times p)$ matrices X and Q. It may be verified that the minimum of this function is obtained for $X = Q$ and the minimum value is clearly $\log \det Q + p$. Now set $X = \Sigma(\theta)$ and $Q = E\varepsilon_0(\theta)\varepsilon_0^T(\theta)$. Then $\operatorname{trace}(QX^{-1}) = E\varepsilon_0^T(\theta)\Sigma^{-1}(\theta)\varepsilon_0(\theta)$. It follows that

$$L(\theta) \geqslant \log \det E\varepsilon_0^T(\theta)\varepsilon_0^T(\theta) + p \geqslant L(\mathring{\theta}) \tag{3.5}$$

where the second inequality follows from (3.4). Equality between $L(\theta)$ and $L(\mathring{\theta})$ holds only if $\theta = \mathring{\theta}$, since, for any two positive definite matrices V and W, $V \geqslant W$ and $\det V = \det W$ implies $V = W$.

We observe that since S is compact, and since an analytical manifold is by definition second countable, the set S is sequentially compact. Now consider the sequence of m.l. estimates $\{\theta^N; N \in Z_+\}$ calculated along the given sample $\{y_N; N \in Z_+\}$. Let θ^* be a limit point of this sequence in the sequentially compact set S and let $\{\theta^M; M = i_1, i_2, \cdots, i_j, \cdots \subset Z_+\}$ be a subsequence such

that $\theta^M \to \theta^*$ as $M \to \infty$. By the minimizing property of θ^M

$$L_M(\mathring{\theta}) \geqslant L_M(\theta^M) \qquad \text{for all} \quad M \in Z_+,$$

where $L_M(\theta)$ denotes $L_M(y_0, \cdots, y_M; \theta)$. Further, by Lemma 3, for every $\varepsilon > 0$, and for some fixed y in $\mathcal{Y}$, there exists N_ε, depending on y but not on θ^M, such that

$$L_M(\theta^M) \geqslant L(\theta^M) - \varepsilon \qquad \text{for all} \quad M > N_\varepsilon.$$

The two inequalities above immediately yield

$$L_M(\mathring{\theta}) \geqslant L(\theta^M) - \varepsilon,$$

for all $M > N_\varepsilon$. Hence, using (3.5), we obtain

$$\begin{aligned} (3.6) \qquad L(\theta^*) &\geqslant L(\mathring{\theta}) = \lim_{M\to\infty} L_M(\mathring{\theta}) \geqslant \lim_{M\to\infty} L(\theta^M) - \varepsilon \\ &= L(\theta^*) - \varepsilon, \end{aligned}$$

for all $\varepsilon > 0$. It follows that $L(\theta) = L(\theta^*)$ and so $\theta^* = \mathring{\theta}$. Consequently the sequence $\{\theta^N; N \in Z_+\}$ converges to $\mathring{\theta}$ for all $y \in \mathcal{Y}$. This proves the strong consistency of the sequence of the m.l. estimators $\{\theta^N; N \in Z_+\}$ and concludes the proof of the theorem.

APPENDIX

LEMMA 1. (1) *There exist* $n + 1 (p \times p)$ *matrices* $\bar{B}_0(\theta), \cdots, \bar{B}_n(\theta)$ *and two positive numbers* K *and* α, $0 < \alpha < 1$, *such that, for all* $\theta \in S$,

$$(A.1.1) \qquad \|B_{t,t-i}(\theta) - \bar{B}_i(\theta)\| < K\alpha^t, \qquad i = 0, \cdots, n, \quad t \in Z_+.$$

In other words, the matrices $B_{t,t-i}(\theta)$ *generated by the algorithm* (2.4)–(2.6) *converge geometrically as* $t \to \infty$ *and uniformly with respect to* $\theta \in S$.

(2) *All the roots of the* det $\bar{B}(z)$ *lie outside the closed unit disk when* $\bar{B}(z) := \bar{B}_0 + \bar{B}_1 z + \cdots + \bar{B}_n z^n$.

PROOF. (1) We begin by demonstrating that the output of the scheme (2.8) asymptotically approaches the output of the scheme (3.3). This enables us to show that $\|B_{t,t}(\theta) - \bar{B}_0(\theta)\| < K\alpha^t$, $t \in Z_+$, and then, further, to establish (A1.1) for $i = 1, \cdots, n$. The second part of the lemma is proven by giving representations of the left-hand side of (1.1) in terms of (2.3) and (3.2), respectively.

For each $\theta \in S$ assume that the process $y(\theta)$ is generated by (1.1) with the matrices $(A_1, \cdots, A_n, B_0, \cdots, B_n)$ corresponding to θ. As in Section 2 we define

$$e_t(\theta) = y_t(\theta) - E(y_t(\theta)|y_{t-1}(\theta), \cdots, y_0(\theta)) = y_t(\theta) - \hat{y}_t(\theta), \quad t \in Z_+,$$

$$\Sigma_t(\theta) = Ee_t(\theta)e_t^T(\theta), \qquad t \in Z_+,$$

where $\hat{y}_0(\theta) = 0$. Then for each $\theta \in S$ we may define an orthonormal process

$\{w_t, t \in Z_+\}$ by

$$y_t = e_t + \hat{y}_t = B_{t,t}w_t + \hat{y}_t, \qquad t \in Z_+, \tag{A1.2}$$

since $Ee_te_t^T = B_{t,t}B_{t,t}^T$, $t \in Z_+$. Also, for y_t^* denoting $y_t^*(\theta) = E(y_t(\theta)|y_i(\theta)$, $i \leqslant t-1), t \in Z$, we have

$$y_t = \varepsilon_t + y_t^*, \qquad t \in Z, \tag{A1.3}$$

and we may define

$$\Sigma = E\varepsilon_t\varepsilon_t^T, \qquad t \in Z.$$

However, inspection of (3.3) shows immediately that $B_0B_0^T = \Sigma$. Now the scheme (3.3) generating y^* may also be represented by

$$y_t^* = \Sigma_{i=-\infty}^{t-1}\Gamma_{t-i-1}y_i, \qquad t \in Z, \tag{A1.4}$$

where the sequence $\{\Gamma_t, t \in Z_+\}$ is a function of θ. Since the zeros of $\det B(z)$ for all $\theta \in S$ lie in a compact set which does not intersect the closed unit disk there exist positive numbers K_1 and α, $\alpha < 1$, independent of $\theta \in S$, such that

$$\|\Gamma_t\| \leqslant K_1\alpha^t, \qquad t \in Z_+.$$

The space $\mathbf{Y}_{0,t-1}$ is a subspace of $\mathbf{Y}_{-\infty,t-1}$. It follows that ε_t is orthogonal to $\mathbf{Y}_{0,t-1}$ and consequently the orthogonal projections of the components of y_t^i and y_t^{*i} on $\mathbf{Y}_{0,t-1}$ are both equal to $\hat{y}_t^i$ for $1 \leqslant i \leqslant p$. Let us write

$$y_t^* = \delta_t + \gamma_t, \qquad t \in Z_+,$$

where

$$\delta_t := \Sigma_{i=0}^{t-1}\Gamma_{t-i-1}y_t, \qquad t \in Z_+,$$

and

$$\gamma_t := \Sigma_{i=-\infty}^{-1}\Gamma_{t-i-1}y_i, \qquad t \in Z_+. \tag{A1.5}$$

Then

$$\|y_t^* - \hat{y}_t\|_H \leqslant \|y_t^* - \delta_t\|_H = \|\gamma_t\|_H, \qquad t \in Z_+, \tag{A1.6}$$

when $\|x\|_H := (Ex^Tx)^{\frac{1}{2}}$. (The situation in (A1.6) is made clear by drawing the appropriate simple diagram; this shows $\hat{y}_t^i$ is closer to y_t^{*i} than the point $\delta_t^i \in \mathbf{Y}_{0,t-1}$ for $i = 1, \cdots, p$.) By (A1.5) and the stationarity of y,

$$\|\gamma_t\|_H \leqslant K_1\Sigma_{i=-\infty}^{-1}\alpha^{t-i-1}\|y_i\|_H = K_1\frac{\|y_0\|_H}{(1-\alpha)}\alpha^{t+1}, \qquad t \in Z_+.$$

Observe now that by (1.1) and (1.2), $\|y_0\|_H = \Sigma_{i=0}^{\infty}$ trace $\Phi_i\Phi_i^T$, and so $\|y_0\|_H$ is a continuous function of θ over S. Consequently, for some positive K_2 independent of θ,

$$\|\gamma_t\|_H \leqslant K_2\alpha^t, \qquad t \in Z_+, \tag{A1.7}$$

for all $\theta \in S$. Finally, from $B_{t,t}w_t - \varepsilon_t = y_t^* - \hat{y}_t$, (A1.6) and (A1.7) we conclude

that

$$\|B_{t,t}w_t - \varepsilon_t\|_H \leqslant K_2\alpha^t, \qquad t \in Z_+. \tag{A1.8}$$

But we have the identity

$$E(\varepsilon_t - B_{t,t}w_t)(\varepsilon_t - B_{t,t}w_t)^T = \Sigma - B_{t,t}B_{t,t}^T - E(\hat{y}_t - y_t^*)w_t^T B_{t,t}^T - EB_{t,t}w_t(\hat{y}_t - y_t^*)^T, \qquad t \in Z_+,$$

and so

$$\operatorname{trace}(B_{t,t}B_{t,t}^T - \Sigma) \leqslant \|\varepsilon_t - B_{t,t}w_t\|_H^2 + 2\|\hat{y}_t - y_t^*\|_H\|e_t\|_H, \qquad t \in Z_+.$$

Clearly $\|e_t\|_H \leqslant \|y_0\|_H$, $t \in Z_+$. Hence, using (A1.6), (A1.7), (A1.8) and the boundedness of $\|y_0\|_H$ over S, we obtain

$$\operatorname{trace}(B_{t,t}B_{t,t}^T - \Sigma) \leqslant K_3\alpha^{2t}, \qquad t \in Z_+,$$

for some positive number K_3. Now for any matrix A, $(\operatorname{trace} AA^T)^{\frac{1}{2}}$ is a norm on A. But all norms on any given finite dimensional vector space are equivalent, hence for some positive K_4

$$\|B_{t,t}B_{t,t}^T - \Sigma\| \leqslant K_4\alpha^{2t}, \qquad t \in Z_+,$$

where

$$\|X\|^2 := \Sigma_{1 \leqslant i,j \leqslant p} X_{ij}^2 \qquad \text{for} \quad X = (X_{ij}),$$

Next, let $\bar{B}_0 = B_0$, the unique upper triangular factor of Σ with positive elements on the diagonal. Now the computation of such a (Cholesky) factor yields a continuous function of the elements of Σ; and the same is true of the Cholesky factor $B_{t,t}$ of $B_{t,t}B_{t,t}^T = \Sigma_t$. Hence

$$\|B_{t,t} - \bar{B}_0\| \leqslant K\alpha^t, \qquad t \in Z_+,$$

for some positive K and some new α, $0 < \alpha < 1$. But this is the claim in (A1.1) for $i = 0$.

To prove the first part of the lemma for $i = 1, \cdots, n$ we write $\varepsilon_t = \bar{B}_0\bar{u}_t$, $t \in Z$, where the process $\bar{u}$ is orthonormal. Then from (A1.8) and the inequality above we obtain

$$\|\bar{u}_t - w_t\|_H \leqslant K_5\alpha^t, \qquad t \in Z_+, \tag{A1.9}$$

for some positive constant K_5 independent of $\theta \in S$. However, from (1.1), (2.3) and (A1.2):

$$v_t := B_0u_t + B_1u_{t-1} + \cdots + B_nu_{t-n} = B_{t,t}w_t + \cdots + B_{t,t-n}w_{t-n}, \qquad t \geqslant n. \tag{A1.10}$$

Since

$$v_t = y_t + A_1y_{t-1} + \cdots + A_ny_{t-n} = \varepsilon_t + C_1\varepsilon_{t-1} + \cdots + C_n\varepsilon_{t-n}, \qquad t \in Z,$$

we may also express v_t relative to the orthonormal process $\bar{u}$ which gives

$$v_t = \bar{B}_0\bar{u}_t + \bar{B}_1\bar{u}_{t-1} + \cdots + \bar{B}_n\bar{u}_{t-n}, \qquad t \in Z, \tag{A1.11}$$

where $\bar{B}_i = Ev_t\bar{u}_{t-i}$, $0 \leqslant i \leqslant n$.

Equating the expressions for v_t from (A1.10) and (A1.11) yields

$$B_{t,t}w_t + \cdots + B_{t,t-n}w_{t-n} = \bar{B}_0\bar{u}_t + \cdots + \bar{B}_n\bar{u}_{t-n}, \qquad t \geqslant n.$$

Now (2.5) shows that the $B_{t,t-i}$ are uniformly bounded for $t \in Z_+$, $0 \leqslant i \leqslant n$, and $\theta \in S$. Consequently, rearranging the equation above as

$$B_{t,t}(w_t - \bar{u}_t)$$

$$+ \cdots + B_{t,t-n}(w_{t-n} - \bar{u}_{t-n}) = (\bar{B}_0 - B_{t,t})\bar{u}_t + \cdots + (\bar{B}_n - B_{t,t-n})\bar{u}_{t-n},$$

and using (A1.9) we obtain part (1) of the lemma.

(2) To prove part (2) it is convenient to use the infinite matrix R_∞ defined by

$$R_\infty = E\begin{bmatrix} \vdots \\ v_{n+1} \\ v_n \\ y_{n-1} \\ \vdots \\ y_0 \end{bmatrix}\left(\cdots v_{n+1}^T v_n^T y_{n-1}^T \cdots y_0^T\right).$$

The initial $(2np \times 2np)$ block of R_∞ is the matrix R given in (2.6) and the remaining nonzero $(p \times p)$ matrix terms are given by $(R_\infty)_{i,j} = R_{j-i}$, for $n \geqslant j - i \geqslant 0$, and $(R_\infty)_{i,j} = R_{i-j}^T$ for $n \geqslant i - j \geqslant 0$, where $R_0, \cdots, R_n$ were defined in Section 2. Since the process y is of full rank the matrix R_∞ is positive definite, i.e., all its initial $tp \times tp$ sections are greater than $kI_{tp\times tp}$, $t \in Z_+$, where $I_{tp\times tp}$ is the identity matrix of the indicated size and $k > 0$.

Let B_∞ denote the unique upper triangular infinite matrix with positive elements on the diagonal such that $R_\infty = B_\infty B_\infty^T$. The positive definiteness of R_∞ implies that the inverse of R_∞ and, thus, of B_∞ exist as bounded operators; in particular this means the rows of B_∞^{-1} are square summable.

Consider the dynamical system

$$w_t + B_{t,t}^{-1}(B_{t,t-1}w_{t-1} + \cdots + B_{t,t-n}w_{t-n}) = B_{t,t}^{-1}v_t, \qquad t \geqslant n, \tag{A1.12}$$

obtained from (A1.10). The block elements $\{H_{i,j};\ i,j \in Z\}$ of the block rows of B_∞^{-1} define the impulse response of (A1.12). Consequently, the solution to (A1.12) is given by

$$w_t = H_{t,t}v_t + H_{t,t-1}v_{t-1} + \cdots + H_{t,t-n}v_n + \eta_{t,n}(w_{n-1}, \cdots, w_0), \qquad t \geqslant n, \tag{A1.13}$$

where $\eta_{t,n}(w_{n-1}, \cdots, w_0)$ is the homogeneous solution to (A1.12) which is also given in terms of $\{H_{i,j};\ i,j \in Z_+\}$. The square summability of the rows of B_∞^{-1} then implies that $\|\eta_{t,n}(w_{n-1}, \cdots, w_0)\|_H \to 0$ as $t \to \infty$.

From (A1.11) we get in the same manner

$$\bar{u}_t = H_0 v_t + \cdots + H_{t-n} v_n + \eta_{t-n}(\bar{u}_{n-1}, \cdots, \bar{u}_0), \qquad t \geqslant n, \tag{A1.14}$$

where $\{H_0, H_1, \cdots\}$ is the impulse response of the system

$$\bar{u}_t + \bar{B}_0^{-1}\left(\bar{B}_1 \bar{u}_{t-1} + \cdots + \bar{B}_n \bar{u}_{t-n}\right) = \bar{B}_0^{-1} v_t, \qquad t \in Z_+. \tag{A1.15}$$

We shall prove that this system is asymptotically stable.

Substituting for v_t in terms of u_t from (A1.10) in (A1.13) and (A1.14) yields

$$\begin{aligned} w_t &= M_{t,t} u_t + \cdots + M_{t,0} u_0 + \eta_{t,n}(w_{n-1}, \cdots, w_0), \qquad t \in Z_+, \\ \bar{u}_t &= N_0 u_t + \cdots + N_t u_0 + \eta_{t-n}(\bar{u}_{n-1}, \cdots, \bar{u}_0), \qquad t \in Z_+, \end{aligned} \tag{A1.16}$$

for certain matrix coefficients $M_{i,j}$, N_i, $i, j \in Z_+$. The components $\eta^i_{t,n}$ and η^i_{t-n}, of $\eta_{t,n}$ and η_{t-n}, respectively, are linear functions of $w_{n-1}, \cdots, w_0$ and $\bar{u}_{n-1}, \cdots, \bar{u}_0$, respectively. Consequently they belong to $\mathbf{Y}_{-\infty, n-1}$. Since $u_{n+1}, u_{n+2}, \cdots$ are orthogonal to $\mathbf{Y}_{-\infty, n}$ it follows from (A1.9) and (A1.16) that

$$\begin{aligned} &\| M_{t,n-1} u_{n-1} + \cdots + M_{t,0} u_0 + \eta_{t,n}(w_{n-1}, \cdots, w_0) \\ &- \left(N_{t-n+1} u_{n-1} + \cdots + N_t u_0 + \eta_{t-n}(\bar{u}_{n-1}, \cdots, \bar{u}_0)\right) \|_H \to 0, \qquad \text{as} \quad t \to \infty. \end{aligned}$$

But $M_{t,k} \to 0$ and $\eta_{t,k} \to 0$ as $t \to \infty$ for fixed k. Consequently,

$$\| N_{t-n+1} u_{n-1} + \cdots + N_t u_0 + \eta_{t-n}(\bar{u}_{n-1}, \cdots, \bar{u}_0) \|_H \to 0 \qquad \text{as} \quad t \to \infty.$$

It then follows from (A1.16) that $\{H_0, H_1, \cdots\}$ is square summmable. This implies, in turn, that $\det(\bar{B}_0 + \bar{B}_1 z + \cdots + \bar{B}_n z^n)$ has all its roots outside the closed unit disk and so the proof of Lemma 1 is complete.

REMARK. Lemma 1 establishes that $\bar{B}(z)$ is asymptotically stable independent of the stability of $B(z)$. However, in this paper, we have assumed $B(z)$ is asymptotically stable. Further, both $\bar{B}_0$ and B_0 are upper triangular with positive elements on the diagonal. Hence $\bar{B}(z)$ and $B(z)$ are identical to the unique stable factor of $B(z)B^T(z^{-1})$ subject to this constraint. It follows that $\bar{B}_i = B_i$; $0 \leqslant i \leqslant n$.

REMARK. For ARMA systems such as (1.1) the predictor (2.3)–(2.7) is equivalent to the Kalman filter [9, 11] and its associated Riccati equation. Lemma 1 establishes results for ARMA systems which correspond to the known convergence results for the Markov state prediction problem (see [23]). However, in Lemma 1 we establish the additional property that the convergence of the matrices appearing in the predictor (2.3)–(2.7) is uniform over the set S. To our knowledge an equivalent result has not so far been proved for the Riccati equation directly.

LEMMA 2. *When $\varepsilon(\theta)$ is generated by* (3.2) *and* $\Sigma(\theta) = B_0(\theta)B_0^T(\theta)$

$$\frac{1}{N+1} \Sigma_{t=0}^{N} \varepsilon_t^T(\theta) \Sigma^{-1}(\theta) \varepsilon_t(\theta) \to E \varepsilon_0^T(\theta) \Sigma^{-1}(\theta) \varepsilon_0(\theta) \quad \text{a.s.} \qquad \textit{as} \quad N \to \infty \tag{A2.1}$$

uniformly in $\theta \in S$, where the expectation is taken with respect to the distribution indexed by $\mathring{\theta} \in S$.

PROOF. For each $\theta \in S$ the indicated convergence results from the ergodic theorem since the process $\varepsilon(\theta)$ is stationary and Gaussian and the covariance sequence of $\varepsilon(\theta)$ is summable. Let

$$W_N(\theta) := \frac{1}{N+1}\Sigma_{t=0}^{N}\varepsilon_t^T(\theta)\Sigma^{-1}(\theta)\varepsilon_t(\theta), \qquad n \in Z_+.$$

We shall demonstrate that $\{\|\partial W_N/\partial\theta_i(\theta)\|;\ N \in Z_+\}$ is almost surely bounded uniformly with respect to $\theta \in S$ and $N \in Z_+$. Then, by the mean value theorem and the compactness of S, we obtain Lemma 2. (The interested reader is referred to [3] for a detailed presentation of this latter step.).

Let θ_i, $1 \leqslant i \leqslant \nu$, denote a component of $\theta \in S$ in a given set of local coordinates and write $\Sigma^{-1}(\theta) = \Pi^T(\theta)\Pi(\theta)$ where $\Pi^T(\theta)$ is uniquely defined by being upper triangular with positive elements on the diagonal. Then, for each i, $1 \leqslant i \leqslant \nu$,

$$\text{(A2.2)}\quad \frac{\partial W_N(\theta)}{\partial\theta_i} = \frac{1}{N+1}\Sigma_{t=0}^{N}\varepsilon_t^T(\theta)\left(\Pi^T(\theta)\frac{\partial\Pi(\theta)}{\partial\theta_i} + \frac{\partial\Pi^T(\theta)}{\partial\theta_i}\Pi(\theta)\right)\varepsilon_t(\theta)$$

$$+ \frac{1}{N+1}\Sigma_{t=0}^{N}\left(\varepsilon_t^T(\theta)\Pi^T(\theta)\Pi(\theta)\frac{\partial\varepsilon_t^T(\theta)}{\partial\theta_i} + \frac{\partial\varepsilon_t^T(\theta)}{\partial\theta_i}\Pi^T(\theta)\Pi(\theta)\varepsilon_t(\theta)\right).$$

The process of derivatives $\{\eta_t(\theta) := \partial\varepsilon_t(\theta)/\partial\theta_i;\ t \in Z\}$, obtained by differentiation of (3.2), is a stationary stochastic process because it satisfies

$$\text{(A2.3)}\qquad \eta_t + C_1\eta_{t-1} + \cdots + C_n\eta_{t-n} = \xi_t \qquad t \in Z,$$

where ξ is the stationary process,

$$\text{(A2.4)}\quad \xi_t = -\frac{\partial C_1}{\partial\theta_i}\varepsilon_{t-1}(\theta) - \cdots - \frac{\partial C_n}{\partial\theta_i}\varepsilon_{t-n}(\theta) + \frac{\partial A_1}{\partial\theta_i}y_{t-1} + \cdots + \frac{\partial A_n}{\partial\theta_i}y_{t-n}, \qquad t \in Z.$$

The derivatives $\{\partial A_j/\partial\theta_i,\ 1 \leqslant j \leqslant n\}$ exist and are continuous over S. Since $\partial C_j/\partial\theta_i = \partial(B_jB_0^{-1})/\partial\theta_i$, $i \leqslant j \leqslant n$, it is clear that $\{\partial C_j/\partial\theta_i,\ 1 \leqslant j \leqslant n\}$ also exist and are continuous over S. We further note that since $\Pi(\theta)$ and $\partial\Pi(\theta)/\partial\theta_i$ are continuous in $\theta \in S$, the maximum and minimum eigenvalues of $(\Pi^T(\theta)\partial\Pi(\theta)/\partial\theta_i + \partial\Pi^T(\theta)/\partial\theta_i\Pi(\theta))$ are bounded over S.

Observe that $A_\theta(z)$ and $B_\theta(z)$ are asymptotically stable for all $\theta \in S$, that the roots of a polynomial are continuous functions of its coefficients, and that S is compact. Consequently, from (3.2), it follows that for some $\rho_1 > 1$ and some $K_1 > 0$ independent of θ

$$\text{(A2.5)}\qquad \|\varepsilon_t(\theta)\| \leqslant K_1\Sigma_{i=0}^{\infty}\rho_1^{-i}\|y_{t-i}\|, \qquad t \in Z.$$

Further, by (A2.3), (A2.4) and the continuity in $\theta \in S$ of

$$\left\{\frac{\partial C_j}{\partial\theta_i}, \frac{\partial A_j}{\partial\theta_i};\ 1 \leqslant j \leqslant n\right\},$$

we also have

$$\|\eta_t(\theta)\| \leqslant K_2 \Sigma_{i=0}^{\infty} \rho_2^{-1} \|y_{t-i}\| \tag{A2.6}$$

for some $\rho_2 > 1$ and some $K_2 > 0$ independent of θ.

The right-hand sides of (A2.5) and (A2.6) constitute bounding ergodic stochastic processes, independent of θ, and we obtain

$$\left\| \frac{\partial W_N(\theta)}{\partial \theta_i} \right\| \leqslant K_3 \frac{1}{N+1} \Sigma_{t=0}^{N} \Big[\big(\Sigma_{i=0}^{\infty} \rho_1^{-i} \|y_{t-i}\|\big)^2 + \big(\Sigma_{i=0}^{\infty} \rho_1^{-i} \|y_{t-i}\|\big)\big(\Sigma_{i=0}^{\infty} \rho_2^{-i} \|y_{t-i}\|\big) \Big] \tag{A2.7}$$

for some $K_3 > 0$. By the ergodic theorem the right-hand side of (A2.7) converges a.s. to

$$K_3 E\Big[\big(\Sigma_{i=0}^{\infty} \rho_1^{-i} \|y_{-i}\|\big)^2 + \big(\Sigma_{i=0}^{\infty} \rho_1^{-i} \|y_{-i}\|\big)\big(\Sigma_{i=0}^{\infty} \rho_2^{-i} \|y_{-i}\|\big) \Big] \tag{A2.8}$$

if this quantity is finite. However it is straightforward to verify that (A2.8) is bounded. As a result we obtain the desired almost sure uniform boundedness of $\{\|\partial W_N(\theta)/\partial \theta_i\|;\ N \in Z_+\}$.

LEMMA 3. *Assume the observed sample* $\{y_t;\ t \in Z_+\}$ *of the process* y *is generated by* (1.1) *with* $\theta = \mathring{\theta}$ *and, as in Section* 2, *define*

$$L_N(y_0, \cdots, y_N; \theta) = \frac{1}{N+1} \Sigma_{t=0}^{N} \big(\log \det \Sigma_t(\theta) + e_t^T(\theta) \Sigma_t^{-1}(\theta) e_t(\theta)\big), \qquad N \in Z_+,$$

where $e(\theta)$ *and* $\Sigma_t(\theta) = B_{t,t}(\theta) B_{t,t}^T(\theta)$, $t \in Z_+$, *are computed via* (2.3)–(2.7) *for all* $\theta \in S$. *Then*

$$L_N(y_0, \cdots, y_N; \theta) \to L(\theta) = \log \det \Sigma(\theta) + E\varepsilon_0^T(\theta) \Sigma^{-1}(\theta) \varepsilon_0(\theta) \quad \text{a.s.} \quad \text{as} \quad N \to \infty,$$

uniformly in $\theta \in S$, *where the expectation is taken with respect to the distribution indexed by* $\mathring{\theta} \in S$.

PROOF. We let

$$\Big(\Delta_N'(\theta) := \frac{1}{N+1} \Sigma_{t=0}^{N} \big(\log \det \Sigma_t(\theta) - \log \det \Sigma(\theta)\big)\Big), \qquad N \in Z_+, \tag{A3.1a}$$

and

$$\Delta_N''(\theta) := \frac{1}{N+1} \Sigma_{t=0}^{N} \big(e_t^T(\theta) \Sigma_t^{-1}(\theta) e_t(\theta) - \varepsilon_t^T(\theta) \Sigma^{-1}(\theta) \varepsilon_t(\theta)\big), \qquad N \in Z_+. \tag{A3.1b}$$

Further, by writing $\Sigma_t(\theta) = \Sigma(\theta) + \Delta\Sigma_t(\theta) = \Sigma(\theta)(I + \Sigma_t^{-1}(\theta) \Delta\Sigma_t(\theta))$, $t \in Z_+$ we have

$$\Delta_N'(\theta) = \frac{1}{N+1} \Sigma_{t=0}^{N} \log \det\big(I + \Sigma^{-1}(\theta) \Delta\Sigma_t(\theta)\big).$$

By Lemma 1, $\|\Sigma^{-1}(\theta)\Delta\Sigma_t(\theta)\| \to 0$ uniformly and geometrically in $\theta \in S$ as $t \to \infty$, and hence the product of the eigenvalues of $I + \Sigma^{-1}(\theta)\Delta\Sigma_t(\theta)$ converges to 1 uniformly in $\theta \in S$ and geometrically in t as $t \to \infty$. This implies that $\Delta_N'(\theta) \to 0$ as $N \to \infty$ uniformly in $\theta \in S$. We now consider $\Delta_N''(\theta)$. Let

$$d_t(\theta) = e_t(\theta) - \varepsilon_t(\theta) = y_t^*(\theta) - \hat{y}_t(\theta), \qquad t \in Z_+, \tag{A3.2}$$

$$Q_t(\theta) = \Sigma_t^{-1}(\theta) - \Sigma^{-1}(\theta), \qquad t \in Z_+.$$

Recall that $\hat{y}(\theta)$ and $y^*(\theta)$ are computed using (2.8) and (3.3), respectively, for all $\theta \in S$, but the input process y for these schemes is generated by (1.1) with $\theta = \mathring{\theta}$.

Using (A3.2) $\Delta_N''(\theta)$ may be written as

(A3.3)

$$\Delta_N''(\theta) = \frac{1}{N+1}\Sigma_{t=0}^{N}\left(\varepsilon_t^T(\theta)Q_t\varepsilon_t(\theta) + 2d_t^T(\theta)\Sigma_t^{-1}(\theta)\varepsilon_t(\theta) - d_t^T(\theta)\Sigma_t^{-1}(\theta)d_t(\theta)\right), \qquad N \in Z_+.$$

By Lemma 1, $\|Q_t(\theta)\| \to 0$ geometrically as $t \to \infty$, uniformly in $\theta \in S$, and by the ergodic theorem

$$\frac{1}{N+1}\Sigma_{t=0}^{N}\varepsilon_t^T(\theta)\varepsilon_t(\theta) \to E\varepsilon_0^T(\theta)\varepsilon_0(\theta) \quad \text{a.s.} \qquad \text{as} \quad N \to \infty.$$

Moreover, by (3.2) $E\varepsilon_0^T(\theta)\varepsilon_0(\theta)$ is uniformly bounded with respect to $\theta \in S$. Hence, the first sum in (A3.3) converges to zero a.s. uniformly in $\theta \in S$. Since by Lemma 1 $\Sigma_t^{-1}(\theta) < K_1 I$, for some positive K_1, we see that the absolute value of the second sum in (A3.3) is majorized by

$$2K_1\left(\frac{1}{N+1}\Sigma_{t=0}^{N} d_t^T(\theta)\, d_t(\theta)\right)^{\frac{1}{2}}\left(\frac{1}{N+1}\Sigma_{t=0}^{N}\varepsilon_t^T(\theta)\varepsilon_t(\theta)\right)^{\frac{1}{2}}$$

and the third sum by

$$\frac{K_1}{N+1}\Sigma_{t=0}^{N} d_t^T(\theta)\, d_t(\theta).$$

Consequently, to prove the second and third sums in (A3.3) converge to zero it is sufficient to prove that

$$\frac{1}{N+1}\Sigma_{t=0}^{N} d_t^T(\theta)\, d_t(\theta) \to 0 \quad \text{a.s.} \qquad \text{as} \quad N \to \infty, \tag{A3.4}$$

uniformly in $\theta \in S$.

Manipulating equations (2.8) and (A3.2) yields:

$$\begin{aligned} d_t(\theta) + C_{t,t-1}d_{t-1}(\theta) + \cdots + C_{t,t-n}d_{t-n}(\theta) \\ = (C_1 - C_{t,t-1})\varepsilon_{t-1}(\theta) + \cdots + (C_n - C_{t,t-n})\varepsilon_{t-n}(\theta), \qquad t \geqslant n, \end{aligned} \tag{A3.5}$$

where, for each $\theta \in S$, the initial conditions $\varepsilon_0(\theta), \cdots, \varepsilon_{n-1}(\theta)$ are given by (2.7) and we recall that the matrix coefficients in (A3.5) are functions of θ.

The solution to the scheme (A3.5) is given by:

$$d_t(\theta) = \Sigma_{i=0}^{t-1}\Phi_{t,i}\varepsilon_i(\theta) + \Sigma_{i=0}^{n-1}\Psi_{t,i}d_i(\theta), \qquad t \geqslant n,$$

where the second term is the homogeneous solution. By Lemma 1, for some $K > 0$ and some α, $0 < \alpha < 1$, $\|C_j - C_{t,t-j}\| \leqslant K\alpha^t$, $j = 0, \cdots, n$ and the limits C_j define a stable system. It follows by standard stability analysis [10] that the system (A3.5) is uniformly asymptotically stable. Moreover, if $1/\alpha'$ is the modulus of the smallest of the roots of $\det(I + C_1 z + \cdots + C_n z^n)$, then for some new α, $0 < \alpha' < \alpha < 1$,

$$\text{(A3.6)} \qquad \|\Psi_{t,i}\| < K_2\alpha^{t-i}, \qquad \|\Phi_{t,i}\| < K_3\alpha^{t-i}, \qquad t \geqslant i \geqslant 0,$$

for some positive constants K_2, K_3.

Next, by writing $\omega_t(\theta) = \Sigma_{i=0}^{n-1}\Psi_{t,i}d_i(\theta)$, $t \in Z_+$, we calculate:

$$\text{(A3.7)} \qquad \frac{1}{N+1}\Sigma_{t=0}^{N}d_t^T(\theta)d_t(\theta) = \frac{1}{N+1}\Sigma_{t=0}^{N}\big(\omega_t^T(\theta)\omega_t(\theta) + 2\omega_t^T(\theta)\Sigma_{i=0}^{t-1}\Phi_{t,i}\varepsilon_i(\theta) + \Sigma_{i,j=0}^{t-1}\varepsilon_i^T(\theta)\Phi_{t,i}^T\Phi_{t,j}\varepsilon_j(\theta)\big).$$

The first sum in (A3.7) is bounded by:

$$\frac{1}{N+1}\Sigma_{t=0}^{N}\omega_t^T(\theta)\omega_t(\theta) < \frac{K_2^2}{N+1}\Sigma_{t=0}^{N}\alpha^{2(t-n)}\Sigma_{i,j=0}^{n-1}|d_i^T(\theta)d_j(\theta)|$$

so that this term converges to zero a.s. as $N \to \infty$. Moreover, since K_1, K_2 and α may be chosen so that (A3.6) holds uniformly in $\theta \in S$, and, by (A3.2), the finite sequence $\{d_i(\theta), 0 \leqslant i \leqslant n-1\}$ is uniformly bounded in $\theta \in S$, the a.s. convergence is uniform in $\theta \in S$. For the second sum in (A3.7) it may be verified that

$$\text{(A3.8)} \qquad \frac{2}{N+1}\Sigma_{t=0}^{N}|\omega_t^T\Sigma_{i=0}^{t-1}\Phi_{t,i}\varepsilon_i(\theta)| < \frac{K_4}{N+1}\big(\Sigma_{i,j=0}^{n-1}|d_i^T(\theta)d_j(\theta)|\big)^{\frac{1}{2}} \times\Big(\Sigma_{t=0}^{N-1}t\alpha^{2(t-n)}\frac{1}{t}\Sigma_{i=0}^{t-1}\|\varepsilon_i(\theta)\|\Big)$$

for some $K_4 > 0$. Recall that the process $\varepsilon(\theta)$ is ergodic. Hence, $1/t\Sigma_{i=0}^{t-1}\|\varepsilon_i(\theta)\|$ converges a.s. as $t \to \infty$. The convergence, moreover, as in Lemma 2, is uniform in $\theta \in S$. Hence, for each sample of the process y, in a set with probability 1, the sums $1/t\Sigma_{i=0}^{t-1}\|\varepsilon_i(\theta)\|$, $t \in Z_+$, are bounded uniformly in t and $\theta \in S$. This with (A3.8) implies that the second sum in (A3.7) converges to zero a.s., uniformly in $\theta \in S$.

For the last term in (A3.7) we have, with (A3.6),

$$\frac{1}{N+1}\Sigma_{t=0}^{N}\Sigma_{i,j=0}^{t-1}|\varepsilon_i^T(\theta)\Phi_{t,i}^T\Phi_{t,j}\varepsilon_j(\theta)| < \frac{K_3^2}{N+1}\Sigma_{t=0}^{N}t^2\alpha^{2t}\frac{1}{t^2}\Sigma_{i,j=0}^{t}\|\varepsilon_i(\theta)\|\,\|\varepsilon_j(\theta)\|,$$

which, by the same arguments as in the preceding case, is seen to converge to zero a.s. as $N \to \infty$, uniformly in $\theta \in S$. Consequently, we have proved (A3.4). It then follows that $\Delta_N''(\theta) \to 0$ a.s. uniformly in $\theta \in S$, and with Lemma 2 and (A3.1) we conclude that Lemma 3 holds.

REFERENCES

[1] ÅSTRÖM, K. J., BOHLIN, T. and WENSMARK, S. (1965). Automatic construction of linear stochastic dynamic models for stationary industrial processes with random disturbances using operating records. Rep. TP 18. 150, IBM Nordic Laboratories, Lidingö, Sweden.

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

[3] CAINES, P. E. (1976). Prediction error identification methods for stationary stochastic processes. *IEEE Trans. Automatic Control*, AC-**21**, 500–505.

[4] CAINES, P. E. and LJUNG, L. (1976). Asymptotic normality and accuracy of prediction error estimators. *Joint Automatic Control Conference*, Purdue, Lafayette, Indiana, July 1976. An extended version is available as Research Report 7602, 1976, Dept. Electrical Engineering, Univ. Toronto.

[5] CAINES, P. E. and RISSANEN, J. (1974). Maximum liklihood estimation of parameters in multivariate Gaussian stochastic processes. *IEEE Trans. Information Theory*. (Corresp.) IT-**20** 102–104.

[6] CLARK, J. M. C. (1976). The consistent selection of local coordinates in linear system identification. *Joint Automatic Control Conference*, Purdue University.

[7] DEISTLER, M., DUNSMUIR, W. and HANNAN, E. J. (1978). Vector linear time series models—corrections and extensions. *Advances in Appl. Probability* **10** 360–372.

[8] DUNSMUIR, W. and HANNAN, E. J. (1976). Vector linear time series models. *Advances in Appl. Probability* **8** 2 339–364.

[9] KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D, J. Basic Eng.* **82** 35–45.

[10] KALMAN, R. E. and BERTRAM, J. (1960). Control system analysis and design via the "second method" of Lyapunov. *Trans. ASME Ser. D, J. Basic Eng.* **82** 371–393.

[11] KALMAN, R. E., FALB, P. L., and ARBIB, M. A. (1969). *Topics in Mathematical System Theory*. McGraw-Hill, New York.

[12] KENDALL, M. G. and STUART, A. (1973). *The Advanced Theory of Statistics*, **2**. Hafner, New York.

[13] LJUNG, L. (1976). On consistency for prediction error identification methods. In *System Identification: Advances and Case Studies* (eds. R. K. Mehra and D. G. Lainiotis). Academic Press, New York.

[14] LJUNG, L. (1976). On consistency and identifiability. In *Mathematical Programming Studies 5* North Holland, New York. 169–190.

[15] MORF, M., SIDHU, S. and KAILATH, T. (1974). Some new algorithms for recursive estimation in constant, linear, discrete-time systems. *IEEE Trans. Automatic Control* AC-**19** 4 315–323.

[16] RISSANEN, J. (1973). A fast algorithm for optimum linear prediction. *IEEE Trans. Automatic Control*, AC-**18** 555.

[17] RISSANEN, J. and BARBOSA, L. (1969). Properties of infinite covariance matrices and stability of optimum predictors. *Information Sci.* **1** 221–236.

[18] ROUSSAS, G. G. (1967). Extension to Markov processes of a result by A. Wald about the consistency of the maximum likelihood estimate. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4** 69–73.

[19] ROZANOV, YU. (1967). *Stationary Stochastic Processes*. Holden Day, San Francisco.

[20] WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

[21] WHITTLE, P. (1956). Estimation and information in stationary time series. *Ark. Mat.* **2** 23.

[22] WIENER, N. and MASANI, P. (1957). The prediction theory of multivariate stochastic processes, part I. *Acta Math.* **98** 111–150.

[23] WONHAM, W. M. (1968). On a matrix Riccati equation of stochastic control. *SIAM J. Control* **6** 4 681–697.

[24] YOULA, D. C. (1961). On the factorization of rational matrices. *IRE Trans. Information Theory*. IT-7 172–189.

IBM RESEARCH LABORATORY
SAN JOSE, CALIFORNIA 95168

DIVISION OF APPLIED SCIENCES
HARVARD UNIVERSITY
PIERCE HALL
CAMBRIDGE, MASSACHUSETTS 02138