CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS FOR MILTIVARIATE GAUSSIAN PROCESSES WITH RATIONAL SPECTRUM

by

J. Rissanen and P. E. Caines

Memorandum No. ERL-M428

March 20, 1974

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS FOR MULTIVARIATE GAUSSIAN PROCESSES WITH RATIONAL SPECTRUM

J. Rissanen

Department of Electrical Engineering
Linköping University
S-581 83 Linköping, Sweden

P. E. Caines

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

Abstract

We give a proof of the strong consistency of the maximum likelihood estimates of the parameters of gaussian random processes possessing linear autoregressive moving average representations.

Research sponsored in part by the National Science Foundation Grants NSF-GK 10656x3, NSF-GK 31627 and the National Aeronautics and Space Agency Grant NGL 05-020-007.

1. Introduction

Consider a p-component stationary full rank random process $y = \{y_t | t \text{ integer}\}$ with a one-sided moving average representation, or Wold-decomposition

$$y_t = \phi_0 u_t + \phi_1 u_{t-1} + \cdots \tag{1.1}$$

see e.g. [1]. Then ϕ_0 is invertible and the subspaces spanned by $\{u_t^i | t \leq n,\ i=1,\ldots,p\}$ and $\{y_t^i | t \leq n,\ i=1,\ldots,p\}$ are the same. For a given process y the orthonormal process u is uniquely determined and $\phi = (\phi_0, \phi_1, \ldots)$ is unique up to right multiplication by orthogonal matrices. By choosing all sequences ϕ such that ϕ_0 is upper triangular with positive elements on the diagonal we obtain a one-to-one relation between ϕ and the covariance function of y. In addition, suppose that y is finitely generated in the sense that there exist two sets of (real) matrices $A = (A_1, \ldots, A_n)$ and $B = (B_0, \ldots, B_n)$ such that

$$y_t + A_1 y_{t-1} + \cdots + A_n y_{t-n} = B_0 u_t + \cdots + B_n u_{t-n}. \tag{1.2}$$

As is well known, the covariance function of y is then rational.

We are interested in the problem of estimating the matrices A and B from increasing sequences $y_0, \ldots, y_n$ of the observations of y by the maximum likelihood technique. Our aim, in particular, is to study the consistency of the resulting estimator.

In the general case with vector processes y there is an immedaite difficulty to be dealt with before a consistency proof can be attempted. This comes from the fact that there are in general several matrix pairs (A,B) which generate y from the orthonormal process u by (1.2); i.e., such that

$$A^{-1}(z)\ B(z) = \phi(z) = \phi_0 + \phi_1 z + \dots , \tag{1.3}$$

where

$$\begin{aligned} A(z) &= I + A_1 z + \dots + A_n z^n \\ B(z) &= B_0 + B_1 z + \dots + B_n z^n . \end{aligned} \tag{1.4}$$

And this is true even if all the cancellations of common factors in the elements of $B(z)$ and det $A(z)$ have been made.

The source of difficulty lies in the fact that the parameters (A,B) are not independent. As studied in [2] and [3] an independent set of such parameters can be found but a part of this set consists of p non-negative integers determined by $\phi(z)$, the so-called Kronecker indices, which may be said to define the structure of the system. Once these indices are known $A(z)$ and $B(z)$ can be expressed in suitable canonical forms with independent real-valued parameters.

Unfortunately, the dependence of the Kronecker indices on $\phi(z)$ is such that two sequences $\phi = (\phi_0, \phi_1, \dots)$ and $\bar{\phi} = (\bar{\phi}_0, \bar{\phi}_1, \dots)$ can be arbitrarily close in, say, the l_1-metric for such sequences and yet give rise to two different sets of Kronecker indices. This means that a sequence of m.l estimates (A^N, B^N), even if expressed in canonical forms, cannot possibly converge unless they all have the same Kronecker indices.

The estimation of the Kronecker indices while certainly possible seems to be a fairly complex matter which we shall not go into in this paper; for a recent study of these and related matters, see [4,5]. Instead, we shall prove the result that if (A^N, B^N) denotes a pair (not necessarily unique) which maximizes a likelihood function constructed from the observed data $y_0,\ldots,y_N$, and the degree n of $A^N(z)$ is not less than that of the true polynomial $\overset{o}{A}(z)$ in (1.3), then $\phi^N \to \overset{o}{\phi}$ in l_1 a.s., where $\phi^N(z) = (A^N(z))^{-1}B^N(z)$. This means that if the Kronecker indices have been consistently estimated then $(A^N(z), B^N(z))$, when expressed in the corresponding canonical forms, also converge to $(\overset{o}{A}(z), \overset{o}{B}(z))$ a.s. .

We should also add that an analogous estimation problem results when (1.2) is written in terms of a Markovian state process; see [4]-[8].

The classical result of Wald [9] established the consistency of the m. l. estimate for independent identically distributed observations and formed the basis for the analysis of Kendall and Stuart [10]. Subsequently, consistency results were studied in [11] - [13] and [8], amongst others; reference [12], in particular, shows how the assumption of independent observations in the classical approach can be relaxed by using the ergodic theorem. This idea, suitably complemented and made precise, also allows us to establish the main theorem, for which to our knowledge no correct proof has appeared before.[1]

2. Recursive Prediction and the Likelihood Function

As is well known [15] the way to obtain the likelihood function for the observations $y_0,\ldots,y_N$ and the parameters to be estimated is to orthogonalize this sequence. This amounts to finding the least squares predictions of y_t given $\{y_0,\ldots,y_{t-1}\}$ for all t. To describe this let $Y_{0,t}$ denote the Euclidean space spanned by the components of the vector

[1] The proof given here was outlined in [14].

variables $y_0,\ldots,y_t$. Let $\hat{y}_t' = (\hat{y}_t^1,\ldots,\hat{y}_t^p)$,[2] where $\hat{y}_t^i$ denotes the orthogonal projection of y_t^i on $Y_{0,t-1}$, and write

$$e_t = y_t - \hat{y}_t \quad , \quad E\, e_t e_t' = \Sigma_t, \quad t = 0,1,\ldots \tag{2.1}$$

Since the process y is full rank Σ_t is invertible.

We now regard the y-process as being gaussian. Then the variables e_t are independent, and with Bayes' formula the probability density function for the joint vector random variables $y_0,\ldots,y_N$ can be written as

$$l_N(y_0,\ldots,y_N) = (2\pi)^{\frac{-(N+1)p}{2}} \prod_{t=0}^{N} \left(|\Sigma_t|^{-\frac{1}{2}} \exp - \tfrac{1}{2} e_t' \Sigma_t^{-1} e_t\right), \tag{2.2}$$

where the y_i's appear implicitly in the e_t's via (2.1); we write $|A|$ for the determinant of A.

It remains to describe the projection $\hat{y}_t$ in terms of the parameters A and B. This could be done by constructing the Kalman predictor for the process y via another state-process where the associated Riccati-equations, [16], are to be supplied with certain initial conditions determined by A and B. For the present situation a more direct approach is to construct the predictor equations in another form as intorduced in [17]; in particular, the Riccati-equations are replaced by an algorithm resulting from the Cholesky-factorization of a covariance matrix. The algorithm generalizes an old algorithm due to Bauer [18].

---

[2] We regard the vector random variables as column arrays of their components, and a prime "'" indicates the transposition. For typographical reasons arrays are always written as rows.

This solution to the predictor problem is an extension of the classical approach of Wiener's based on factoring the spectrum of a stationary process. As the outlined method for obtaining recursively the predictions of the y-process is not, perhaps, well known we shall give the relevant results in the theorem: (For more details we refer to [17]).

Theorem. For the process (1.2) the least squares predictions

$$y_{t+1} = E(y_{t+1}/y_0,\ldots,y_t) \tag{2.3}$$

are given by the equations:

$$\begin{aligned}\hat{y}_t + C_{t,t-1}\hat{y}_{t-1} + \ldots + C_{t,t-n}\hat{y}_{t-n} &= \\ = (C_{t,t-1} - A_1)\, y_{t-1} + \ldots + (C_{t,t-n} - A_n)\, y_{t-n} \qquad t \geq n\end{aligned} \tag{2.4}$$

where

$$C_{t,j} = B_{tj}B_{jj}^{-1}, \quad t \geq n,\ t-1 \leq j \leq t-n; \tag{2.5}$$

and the $B_{t,t-i}$'s are defined as the pxp-block elements of the upper triangular factor B of a positive definite band covariance matrix R,

$$R = BB' \tag{2.6}$$

defined as follows:

$$R = E\begin{pmatrix} \vdots \\ v_{n+1} \\ v_n \\ y_{n-1} \\ \vdots \\ y_0 \end{pmatrix} (\dots v'_{n+1}\ v'_n\ y'_{n-1}\ \cdots\ y'_0), \tag{2.7}$$

with $v_t = B_0u_t + B_1u_{t-1} + \dots + B_nu_{t-n}$. The factors B are, moreover, made unique by requiring B_{jj} to be upper triangular with positive elements on the diagonal. Finally, the initial conditions $\hat{y}_0,\dots,\hat{y}_{n-1}$ are given by:

$$\begin{pmatrix} \hat{y}_{n-1} \\ \vdots \\ \hat{y}_1 \end{pmatrix} = \begin{bmatrix} B_{n-1,n-2} & \cdots & B_{n-1,0} \\ & \ddots & \vdots \\ 0 & & B_{10} \end{bmatrix} \begin{bmatrix} B_{n-2,n-2} & \cdots & B_{n-2,0} \\ & \ddots & \\ B & & B_{00} \end{bmatrix}^{-1} \begin{pmatrix} y_{n-2} \\ \vdots \\ y_0 \end{pmatrix}$$

$$\hat{y}_0 = 0. \tag{2.8}$$

<u>Remark</u>. A comparison of the elements in (2.6), starting with right-bottom corner, gives easily recurrence relations for the B_{ij}'s, see [17] or [14], which we omit since they are not needed here.

In order to give the reader an impression of the appearance of R and the way its elements are determined by A and B we describe the factorization (2.6) for n = 2:

$$\begin{bmatrix} \ddots & \ddots & \ddots & \ddots & 0 \\ \ddots & R_0 & R_1' & R_2' & \\ \ddots & R_1 & R_0 & Q' & R_2' \\ & R_2 & Q & P_0 & P_1' \\ 0 & & R_2 & P_1 & P_0 \end{bmatrix} =$$

$$\begin{bmatrix} \ddots & \ddots & \ddots & \ddots & 0 \\ & & B_{22} & B_{21} & B_{20} \\ 0 & & & B_{11} & B_{10} \\ & & & & B_{00} \end{bmatrix} \times \begin{bmatrix} \ddots & & & 0 \\ \ddots & B_{22}' & & \\ \ddots & B_{21}' & B_{11}' & \\ 0 & B_{20}' & B_{10}' & B_{00}' \end{bmatrix}$$

where

$$R_0 = B_0B_0' + B_1B_1' + B_2B_2'$$

$$R_1 = B_0B_1' + B_1B_2'$$

$$R_2 = B_0B_2'$$

$$Q = R_1' - A_1B_0$$

$$P_0 = E\ y_0y_0'$$

$$P_1 = E\ y_1y_0'$$

We omit the well-known equations which determine P_0 and P_1 as continuous functions of A and B.

Observe, in particular, that apart from the initial $2np \times 2np$-portion of R it is a band block Toeplitz-matrix. We add that if we are willing to consider that near-maximum likelihood problem obtained by putting $y_0 = \dots = y_{n-1} = 0$ then the factorization R in (2.6) can be done in an order of magnitude fewer arithmetic operations than the straight-forward comparison algorithm above, which is basically just a Gauss-elimination scheme, see [19].

The scaled log-likelihood function, $L_N = \frac{(N+1)p}{2} \log (2\pi) - \frac{2}{N+1} \log \ell_N$, is now completely described in terms of the parameters $\theta = (A,B)$ and the data $y_0,\dots,y_N$:

$$L_N(y_0,\dots,y_N,\theta) = \frac{1}{N+1} \sum_{i=0}^{N} (\log|\Sigma_i(\theta)| + e_i'(\theta)\Sigma_i^{-1}(\theta)e_i(\theta)) \tag{2.9}$$

where

$$e_i(\theta) = y_i - \hat{y}_i(\theta); \quad \Sigma_i(\theta) = B_{ii}(\theta)\, B_{ii}'(\theta); \tag{2.10}$$

and $\hat{y}_i(\theta)$ is the solution to (2.4);[3] we also wrote B_{ii} as $B_{ii}(\theta)$ to emphasize the dependency on θ.

3. Auxiliary Lemmas

We begin by defining the space of the parameters as the subset of the euclidean space $R^{(2n+1)p^2}$ consisting of all sequences $\theta = (A_1,\dots, A_n, B_0,\dots,B_n)$ for which the roots of $|A(z)|$ and $|B(z)|$, see (1.4), are outside the unit circle. Then, calearly, $B_0 = B(\theta)$ is invertible, and (1.2) describes a full rank stationary process y which has a 1-sided moving average representation. Furthermore we shall restrict

[3] In the literature the initial conditions $\hat{y}_0,\dots,\hat{y}_{n-1}$ are frequently considered as independent parameters, which clearly is not the case.

the set of admissible parameters so that B_0 is upper triangular with positive elements on the diagonal.

Let S denote any closed and bounded subset of the parameter space containing a "true" parameter $\overset{o}{\theta} = \langle \overset{o}{A}, \overset{o}{B} \rangle$ for (1.2); i.e. any parameters which produces the given gaussian process from which the data y_0, $y_1, \ldots$ is being drawn.

Consider now a process (1.2) for a given $\theta \in S$ and for all integer values of t. Let $\bar{y}_t' = (\bar{y}_t^1, \ldots, \bar{y}_t^p)$ where $\bar{y}_t^i$ is the orthogonal projection of y_t^i on $Y_{-\infty, t-1}$, the Hilbert space spanned by $\{y_k^i | \ i=1,\ldots,p, k \leq t-1\}$. Then,

$$y_t = \varepsilon_t + \bar{y}_t \quad , \quad E\, \varepsilon_t \, \varepsilon_t' = \Sigma. \tag{3.1}$$

As is well known the stationary process $\{\bar{y}_t\}$ is generated by the difference equation:

$$\bar{y}_t + C_1 \bar{y}_{t-1} + \cdots + C_n \bar{y}_{t-n} = (C_1 - A_1)\, y_{t-1} + \cdots + (C_n - A_n)\, y_{t-n}. \tag{3.2}$$

where $C_i = B_i B_0^{-1}$. This equation, indeed, defines a stationary process, since for $\theta \in S$ the system (3.2) is asymptotically stable. We have the solution

$$\bar{y}_t = \sum_{i=-\infty}^{t-1} \Gamma_{t-1}\, y_i \tag{3.3}$$

for which, moreover, two positive numbers K_1 and α, $\alpha < 1$, exist such that

$$\|\Gamma_t\| < K_1 \alpha^t \text{ for all } \theta \in S \text{ uniformly.} \tag{3.4}$$

Here, of course, $\Gamma_t = \Gamma_t(\theta)$ is regarded as a function of θ and $\|A\|$ denotes any of the norms of a matrix A.

We shall give a detailed proof of the following fundamental result, which sharpens the results of [17] and [20]. Analogous results have been proved for the Riccati equation in the recent theory of positive real functions, [7], [22].

<u>Lemma</u>. (1) There exist p×p-matrices $\bar{B}_0(\theta), \ldots, \bar{B}_n(\theta)$ and two positive numbers K and α, $\alpha < 1$, such that for all $\theta \in S$,

$$\|B_{t,t-1}(\theta) - \bar{B}_i(\theta)\| < K\alpha^t, \quad i = 0,\ldots,n. \tag{3.5}$$

(2) All the roots of $\det(\bar{B}_0 + \bar{B}_1 z + \ldots + \bar{B}_n z^n)$ are outside the unit circle.

Proof. By the Theorem in section 2 we have

$$y_t = e_t + \hat{y}_t = B_{tt}\, w_t + \hat{y}_t, \qquad t \geq 0 \tag{3.6}$$

where $\{w_t\}$ is an orthonormal process. Write $\|x\| = (Ex'x)^{1/2}$ for $x' = (x^1,\ldots,x^p)$ and $x^i \in Y_{-\infty,t}$. As $Y_{0,t-1} \subset Y_{-\infty,t-1}$, ε_t in (3.1) is orthogonal to $Y_{0,t-1}$, and the orthogonal projections of y_t^i and $\bar{y}_t^i$ on $Y_{0,t-1}$ are equal, namely, $\hat{y}_t^i$. Therefore, if we write (3.3) as

$$\bar{y}_t = \delta_t + \gamma_t,$$

where

$$\delta_t = \sum_{i=0}^{t-1} \Gamma_{t-1}\, y_i \text{ and } \gamma_t = \sum_{i=-\infty}^{0} \Gamma_{t-1}\, y_i$$

we have

$$\|\bar{y}_t - \hat{y}_t\| \leq \|\bar{y}_t - \delta_t\| = \|\gamma_t\| \tag{3.7}$$

which merely states that $\hat{y}_t^i$ is closer to $\bar{y}_t^i$ than the point $\delta_t^i \in Y_{0,t-1}$ for $i=1,\ldots,p$.

Further, by (3.3) - (3.4)

$$\|\gamma_t\| \leq K_1 \sum_{i=-\infty}^{0} \alpha^{t-i} \|y_i\| = K_1 \frac{\|y_0\|}{1-\alpha} \alpha^t.$$

By (1.2) $\|y_0\|$ is a continuous function of θ over the compactum S and consequently has a maximum for some positive K_2, $\|\gamma_t\| < K_2\alpha^t$ for all $\theta \in S$. From (3.1), (3.6), and (3.7):

$$\|\varepsilon_t - B_{tt}w_t\| < K_2\alpha^t. \tag{3.8}$$

This gives in turn for some positive K_3,

$$\|\Sigma - B_{tt}B_{tt}'\| < K_3\alpha^t \qquad \text{for all } \theta \in S.$$

Let $\bar{B}_0$ be the upper triangular factor in $\Sigma = \bar{B}_0\bar{B}_0'$ with positive elements on the diagonal. Then the last inequality implies the claim (3.5) for $i = 0$.

To prove (3.5) for the other values for i write,

$$\varepsilon_t = \bar{B}_0\bar{u}_t \tag{3.9}$$

where the process $\{\bar{u}_t\}$ is orthonormal. From what was just proven, we have for some positive K_4:

$$\|\bar{u}_t - w_t\| < K_4\alpha^t \qquad \text{for all } \theta \in S. \tag{3.10}$$

From the theorem in section 2 the stationary process $\{v_t | t \geq n\}$ is given by:

$$v_t \triangleq B_0u_t + B_1u_{t-1} + \ldots + B_nu_{t-n} = B_{tt}w_t + \ldots + B_{t,t-n}w_{t-n} \tag{3.11}$$

If we express v_t relative to the orthonormal process $\bar{u}_t$ we get

$$v_t = \bar{B}_0\bar{u}_t + \bar{B}_1\bar{u}_{t-1} + \dots + \bar{B}_n\bar{u}_{t-n}, \tag{3.12}$$

where

$$\bar{B}_i = E\, v_t\bar{u}'_{t-1}.$$

From (3.10) - (3.12) we conclude (3.5), or part (1) of the lemma.

As the process $\{y_i | i=0,1,\dots\}$ is of full rank the matrix R in (2.7) is positive definite; i.e., all its initial tp × tp-sections are greater than $kI_{tp\times tp}$, where $I_{tp\times tp}$ is the identity matrix of the indicated size and $k > 0$. This means that the inverse of R and, thus, of B exist as bounded operators; in particular, the rows of B^{-1} are square summable. Then the block-elements H_{ij} of the block rows of B^{-1} define the impulse response of the following dynamic system obtained from (3.11):

$$w_t + B_{tt}^{-1}(B_{t,t-1}w_{t-1} + \dots + B_{t-t-n}w_{t-n}) = B_{tt}^{-1} v_t. \tag{3.13}$$

Its solution is then given by

$$\begin{aligned} w_t &= H_{tt}v_t + H_{t,t-1}v_{t-1} + \dots + H_{t,t-\tau}v_{t-\tau} + \\ &\quad + \eta_{t,\tau}(w_{\tau-1},\dots,w_{\tau-n}) \end{aligned} \tag{3.14}$$

where $\eta_{t,\tau}(w_{\tau-1},\dots,w_{\tau-n})$ is the homogeneous solution to (3.13), also given in terms of H_{ij}. The square summability of these implies $\eta_{t,\tau}(w_{\tau-1},\dots,w_{\tau-n}) \to 0$ as $t \to \infty$.

From (3.12) we get in the same manner

$$\bar{u}_t = H_0v_t + \dots + H_\tau v_{t-\tau} + \eta_{t-\tau}(\bar{u}_{\tau-1},\dots,\bar{u}_{\tau-n}) \tag{3.15}$$

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

$$\bar{u}_t + \bar{B}_0^{-1}(\bar{B}_1\bar{u}_{t-1} + \ldots + \bar{B}_n\bar{u}_{t-n}) = \bar{B}_o^{-1}v_t. \tag{3.16}$$

We shall prove that this system is asymptotically stable.

By substituting v_t in terms of u_t from (3.11) in (3.14) and (3.15) we get:

$$\begin{aligned} w_t &= M_{tt}u_t + \ldots + M_{t,t-\tau-n}u_{t-\tau-n} + \eta_{t,\tau}(w_{\tau-1},\ldots,w_{\tau-n}) \\ \bar{u}_t &= N_{tt}u_t + \ldots + N_{t,t-\tau-n}u_{t-\tau-n} + \eta_{t-\tau}(\bar{u}_{\tau-1},\ldots,\bar{u}_{\tau-n}) \end{aligned} \tag{3.17}$$

for certain matrix coefficients M_{ij} and N_{ij}. The components $\eta^i_{t,\tau}$ and $\eta^i_{t-\tau}$ being linear functions of $w_{\tau-1},\ldots,w_{\tau-n}$ and $\bar{u}_{\tau-1},\ldots,\bar{u}_{\tau-n}$, respectively, belong to $Y_{-\infty,\tau-1}$. Since $u_t,\ldots,u_{t-\tau-n}$ (for $t > n+2\tau-1$) are orthogonal to this subspace we get with (3.10) from (3.17):

$$\eta_{t,\tau}(w_{\tau-1},\ldots,w_{\tau-n}) - \eta_{t-\tau}(\bar{u}_{\tau-1},\ldots,\bar{u}_{\tau-n}) \to 0,$$

as $t \to \infty$. This with $\eta_{t,\tau} \to 0$ implies $\eta_{t-\tau} \to 0$ as $t \to \infty$. Further, (3.15) then implies that $\{H_0, H_1, \ldots\}$ is square summable, which implies that $\det(\bar{B}_0 + \bar{B}_1 z + \ldots + \bar{B}_n z^n)$ has all the roots outside the unit circle. The proof of the lemma is complete.

In the next results we regard the process $y_t = y_t(\overset{o}{\theta})$ as fixed and given by (1.2) for $\theta = \overset{o}{\theta}$, a "true" parameter. With this process as the input the two predictor equations (2.4) and (3.2) still describe the corresponding predictions for each $\theta \in S$, say, $\tilde{y}_t = \tilde{y}_t(\theta,\overset{o}{\theta})$ and $y^*_t = y^*_t(\theta,\overset{o}{\theta})$, respectively, which, of course, no longer conicide with the orthogonal projections of y_t on $Y_{0,t-1}$ and $Y_{-\infty,t-1}$, respectively, unless $\theta = \overset{o}{\theta}$. For later reference we rewrite these equations here:

$$\tilde{y}_t + C_{t,t-1}\tilde{y}_{t-1} + \dots + C_{t,t-n}\tilde{y}_{t-n} = (C_{t,t-1}-A_1)y_{t-1} + \dots$$

$$\dots + (C_{t,t-n}-A_n)y_{t-n}, \qquad t \geq n$$

$$\tilde{y}_t = \hat{y}_t \text{ for } t = 0,\dots,n-1 \tag{3.18}$$

$$y^*_t + C_1 y^*_{t-1} + \dots + C_n y^*_{t-n} = (C_1-A_1)y_{t-1} + \dots$$

$$\dots + (C_n-A_n)y_{t-n}, \text{ all } t.$$

<u>Lemma 2</u>. With $\varepsilon_i(\theta,\overset{o}{\theta}) = y_t(\overset{o}{\theta}) - y^*_t(\theta,\overset{o}{\theta})$ (defined above) and $\Sigma(\theta) = B_0(\theta)\, B_0'(\theta)$,

$$\frac{1}{N+1}\sum_{t=0}^{N} \varepsilon_t'(\theta,\overset{o}{\theta})\, \Sigma^{-1}(\theta)\, \varepsilon_t(\theta,\overset{o}{\theta}) \rightarrow E\, \varepsilon_0'(\theta,\overset{o}{\theta})\, \Sigma^{-1}(\theta)\, \varepsilon_0(\theta,\overset{o}{\theta})$$

a.s. uniformly in $\theta \in S$.

Proof. For each $\theta \in S$ the indicated convergence results from the ergodic theorem since the covariance sequence of $\varepsilon_i(\theta,\overset{o}{\theta})$ is summable. The limit function is by (3.18) a uniformly continuous function over the compactum S. Since, similarly, each of the converging functions is uniformly continuous over S the result follows.

<u>Lemma 3</u>. With $e_t(\theta,\overset{o}{\theta}) = y_t(\overset{o}{\theta}) - \tilde{y}_t(\theta,\overset{o}{\theta})$ and $\varepsilon_t(\theta,\overset{o}{\theta}) = y_t(\overset{o}{\theta}) - y^*_t(\theta,\overset{o}{\theta})$, $L_N(y_0,\dots,y_N,\theta,\overset{o}{\theta}) \rightarrow L(\theta,\overset{o}{\theta})$ a.s. uniformly in $\theta \in S$, where

$$L_N(y_0,\dots,y_N,\theta,\overset{o}{\theta}) = \frac{1}{N+1}\sum_{t=0}^{N} [\log|\Sigma_t(\theta)| + e_t'(\theta,\overset{o}{\theta})\, \Sigma_t^{-1}(\theta)\, e_t(\theta,\overset{o}{\theta})] \text{ and}$$

$$L(\theta,\overset{o}{\theta}) = \log|\Sigma(\theta)| + E\varepsilon_0'(\theta,\overset{o}{\theta})\, \Sigma^{-1}(\theta)\varepsilon_0(\theta,\overset{o}{\theta}).$$

Proof. Write

$$\Delta_N'(\theta) = \frac{1}{N+1} \sum_{t=0}^{N} (\log|\Sigma_t(\theta)| - \log|\Sigma(\theta)|)$$

$$\Delta_N''(\theta) = \frac{1}{N+1} \sum_{t=0}^{N} [e_t'(\theta,\overset{o}{\theta})\, \Sigma_t^{-1}(\theta) e_t(\theta,\overset{o}{\theta}) - \tag{3.19}$$

$$- \varepsilon_t'(\theta,\overset{o}{\theta})\, \Sigma^{-1}(\theta)\, \varepsilon_t(\theta,\overset{o}{\theta})]$$

By further writing $\Sigma_t(\theta) = \Sigma(\theta) + \Delta\Sigma_t(\theta) = \Sigma(\theta)(I+\Sigma^{-1}(\theta)\Delta\Sigma_t(\theta))$ we have

$$\Delta_N'(\theta) = \frac{1}{N+1} \underset{t=0}{\overset{N}{\Delta}} \log|\, I + \Sigma^{-1}(\theta)\, \Delta\Sigma_t(\theta)|.$$

By Lemma 1 $\Sigma^{-1}(\theta)\, \Delta\Sigma_t(\theta) \to 0$ and hence the product of the eignevalues of $I + \Sigma^{-1}(\theta)\Delta\Sigma_t(\theta)$ converges to 1, all exponentially uniformly in $\theta \in S$. This implies that $\Delta_N'(\theta) \to 0$ uniformly in $\theta \in S$.

Let

$$d_t(\theta,\overset{o}{\theta}) = e_t(\theta,\overset{o}{\theta}) - \varepsilon_t(\theta,\overset{o}{\theta}) = y_t^*(\theta,\overset{o}{\theta}) - \tilde{y}_t(\theta,\overset{o}{\theta})$$

$$Q_t(\theta) = \Sigma_t^{-1}(\theta) - \Sigma^{-1}(\theta). \tag{3.20}$$

To abbreviate the following expressions we drop the arguments θ and $\overset{o}{\theta}$. Then

$$\Delta_N'' = \frac{1}{N+1} \sum_{t=0}^{N} [\varepsilon_t' Q_t \varepsilon_t + 2d_t'\Sigma_t^{-1} e_t - d_t'\Sigma_t^{-1} d_t]. \tag{3.21}$$

The first sum is majorized by

$$(\frac{1}{N+1} \sum_{t=0}^{N} \varepsilon_t'\varepsilon_t)\, \|Q_t\|.$$

By Lemma 1 $\|Q_t\| \to 0$ exponentially uniformly in $\theta \in S$, and by the ergodic theorem

$$\frac{1}{N+1} \sum_{t=0}^{N} \varepsilon_t'\varepsilon_t \to E\, \varepsilon_0'\varepsilon_0 \qquad \text{a.s.}\ .$$

Moreover, by (3.18) $E\varepsilon_0'\varepsilon_0 < M$ for all $\theta \in S$. Hence, the first sum in (3.21) converges to zero a.s. uniformly in $\theta \in S$. Since by Lemma 1 $\Sigma_t^{-1}(\theta) < k \cdot I$ we see that the absolute value of the second sum in (3.21) is majorized by

$$2k \left(\frac{1}{N+1} \sum_{t=0}^{N} d_t'\, d_t\right)^{1/2} \left(\frac{1}{N+1} \sum_{t=0}^{N} \varepsilon_t'\, \varepsilon_t\right)^{1/2} \tag{3.22}$$

and the third sum by

$$\frac{k}{N+1} \sum_{t=0}^{N} d_t'\, d_t \tag{3.23}$$

Consequently to prove the second and third sums in (3.21) converge to zero it is sufficient to prove that

$$\frac{1}{N+1} \sum_{t=0}^{N} d_t'\, d_t \to 0 \tag{3.24}$$

a.s. uniformly in $\theta \in S$.

Manipulating the equations (3.18) and (3.20) yields:

$$d_t + C_{t,t-1} d_{t-1} + \cdots + C_{t,t-n} d_{t-n} =$$

$$= (C_1 - C_{t,t-1})\varepsilon_{t-1} + \cdots + (C_n - C_{t,t-n})\, \varepsilon_{t-n}. \tag{3.25}$$

Its solution is given by:

$$d_t = \sum_{i=0}^{t-1} \phi_{ti}\varepsilon_i + \sum_{i=0}^{n-1} \psi_{ti}\, d_i \quad , \quad t \geq n \tag{3.26}$$

where the second term is the homogeneous solution. By Lemma 1, $\|C_j - C_{t,t-j}\| < K\alpha^t$. $j = 0,\ldots,n$. It follows by standard stability analysis, [21], that the system (3.25) is uniformly asymptotically stable since the constant system defined by the limits C_i is such. Moreover, if $\frac{1}{\beta}$ is the modulus of the smallest of the roots of $\det(I + C_1 z + \ldots + C_n z^n)$ then $\beta < 1$ and

$$\begin{aligned} \|\psi_{ti}\| &< K_1 \beta^{t-1} \\ \|\phi_{ti}\| &< K_2 \beta^{t-1} \alpha^{i}. \end{aligned} \tag{3.27}$$

for some positive constants K_1, K_2.

By picking α so that $\beta < \alpha < 1$ and putting $\beta = \gamma \cdot \alpha$ we get the following estimates:

$$\begin{aligned} \|\psi_{ti}\| &< K_1 \alpha^{t-i} \\ \|\phi_{ti}\| &< K_2 \alpha^{t} \gamma^{t-i}. \end{aligned} \tag{3.28}$$

Next, by writing $\omega_t = \sum_{i=0}^{n-1} \psi_{ti} d_i$, we calculate:

$$\frac{1}{N+1} \sum_{t=0}^{N} d_t' d_t = \frac{1}{N+1} \sum_{t=0}^{N} [\omega_t'\omega_t + 2\omega_t' \sum_{i=0}^{t-1} \phi_{ti}\varepsilon_i + \sum_{i,j=0}^{t-1} \varepsilon_i' \phi_{ti}' \phi_{tj} \varepsilon_j] \tag{3.29}$$

The first sum is bounded by:

$$\frac{1}{N+1}\sum_{t=0}^{N}\omega_t'\omega_t < \frac{K_1}{N+1}\sum_{t=0}^{N}\alpha^{2\,(t-n)}\sum_{i,j=0}^{n-1}|d_i'd_j|,$$

so that this term converges to zero a.s. . Moreover, as the K_1, K_2, α, and β may be chosen so that (3.28) hold uniformly in $\theta \in S$, the a.s. convergence is uniform in $\theta \in S$. For the second sum in (3.29) we get

$$\frac{2}{N+1}\sum_{t=0}^{N}\Big|w_t'\sum_{i=0}^{t-1}\phi_{ti}\,\varepsilon_i\Big|$$

$$< \frac{2K_2}{N+1}\sum_{t=0}^{N-1}\|w_t\|\sum_{i=0}^{t-1}\alpha^t\,\gamma^{t-1}\|\varepsilon_i\|$$

$$< \frac{2K_1K_2}{N+1}\cdot\Big(\sum_{i,j=0}^{n-1}|d_i'd_j\Big)^{1/2}\cdot\sum_{t=0}^{N-1}t\,\alpha^{2(t-n)}\cdot\frac{1}{t}\sum_{i=0}^{t-1}\|\varepsilon_i\| \quad (3.30)$$

The process $\{\varepsilon_i\}$ is ergodic. Hence, $\frac{1}{t}\sum_{i=0}^{t}\|\varepsilon_i\|$ converges a.s. . The convergence, moreover, as in Lemma 2, is uniform in $\theta \in S$. Hence, for each sample of the process $\{y_0(\overset{o}{\theta}),y_1(\overset{o}{\theta}),\ldots\}$ in a set with probability 1 the sums $\frac{1}{t}\sum_{i=0}^{t}\|\varepsilon_i(\theta,\overset{o}{\theta})\|$ are bounded uniformly in t and $\theta \in S$. This with (3.30) implies that the second sum in (3.29) converges to zero a.s., uniformly in $\theta \in S$.

For the last term in (3.29) we have with (3.28):

$$\frac{1}{N+1}\sum_{t=0}^{N}\sum_{i,j=0}^{t-1}|\varepsilon_i'\phi_{ti}'\phi_{tj}\varepsilon_j| < \frac{K_2^2}{N+1}\sum_{t=0}^{N}t^2\alpha^{2t}\frac{1}{t^2}\sum_{i,j=0}^{t}\|\varepsilon_i\|\,\|\varepsilon_j\|$$

which by the same arguments as in the preceding case is seen to converge to zero a.s., uniformly in $\theta \in S$. Consequently, we have proved (3.24). It then follows that $\Delta_N'' \to 0$ a.s. uniformly in $\theta \in S$, and with Lemma 2 and (3.19) we conclude Lemma 3.

4. Main Theorem

As the maximum likelihood estimates θ^N of the true parameter $\overset{\circ}{\theta} = (\overset{\circ}{A},\overset{\circ}{B})$ themselves do not necessarily converge we pass over to the parameters $\phi^N = (\phi_0^N,\phi_1^N,\ldots)$ obtained from the matrix coefficients of the expansion

$$A^N(z)^{-1}B^N(z) = \phi_0^N + \phi_1^N z + \ldots + \quad \triangleq \phi^N(z), \tag{4.1}$$

where

$$B^N(z) = B_0^N + B_1^N z + \ldots + B_n^N z^n$$

$$A^N(z) = I + A_1^N z + \ldots + A_n^N z^n.$$

Theorem. If $\theta^N = (A^N,B^N)$ minimizes $L_N(y_0,\ldots,y_N,\theta,\overset{\circ}{\theta})$ over S, then

$$\sum_{i=0}^{\infty} \|\phi_i^N - \overset{\circ}{\phi}_i\| \to 0 \text{ a.s.} \qquad \text{as } N \to \infty, \tag{4.2}$$

where $\overset{\circ}{\phi}(z) = \overset{\circ}{A}(z)^{-1}\,\overset{\circ}{B}(z)$.

Proof. First, for $\theta = \overset{\circ}{\theta}$ in (3.18) $y_t^*(\overset{\circ}{\theta},\overset{\circ}{\theta})$ is the orthogonal projection of y_t on $Y_{-\infty,t-1}$. Hence, for all $\theta \in S$,

$$E\varepsilon_0(\theta,\overset{\circ}{\theta})\ \varepsilon_0'(\theta,\overset{\circ}{\theta}) \geq E\varepsilon_0(\overset{\circ}{\theta},\overset{\circ}{\theta})\ \varepsilon_0'(\overset{\circ}{\theta},\overset{\circ}{\theta}) = \Sigma(\overset{\circ}{\theta}). \tag{4.3}$$

As $y = y(\overset{\circ}{\theta})$ is a full rank process we have,

$$y_t^*(\overset{o}{\theta},\overset{o}{\theta}) = \Gamma_1 y_{t-1}(\overset{o}{\theta}) + \Gamma_2 y_{t-2}(\overset{o}{\theta}) + \ldots$$

for a unique set of p×p-matrix coefficients. By (1.2) another parameter θ generates also the process $\{y_t(\overset{o}{\theta})\}$ if and only if $\theta \cong \overset{o}{\theta}$ in the usual sense that $A^{-1}(z)\ B(z) = \overset{o}{A}{}^{-1}(z)\ \overset{o}{B}(z)$. Therefore, the equality in (4.3) holds if and only if $\theta \cong \overset{o}{\theta}$.

Next, recall the inequality:

$$\log|X| + \text{trace } (AX^{-1}) \geq \log|A| + \text{trace } I,$$

which with $X = \Sigma(\theta)$, $A = E\varepsilon_0(\theta,\overset{o}{\theta})\ \varepsilon_0'(\theta,\overset{o}{\theta})$, and the fact that trace (E yy') B = E y' By gives

$$L(\theta,\overset{o}{\theta}) \geq \log|E\varepsilon_0(\theta,\overset{o}{\theta})\ \varepsilon_0'(\theta,\overset{o}{\theta})| + \text{trace } I.$$

By (4.3) $\log|E\varepsilon_0(\theta,\overset{o}{\theta})\ \varepsilon_0'(\theta,\overset{o}{\theta})| \geq \log|\Sigma(\overset{o}{\theta})|$, and therefore

$$L(\theta,\overset{o}{\theta}) \geq \log|\Sigma(\overset{o}{\theta})| + \text{tr } I = L(\overset{o}{\theta},\overset{o}{\theta}). \tag{4.4}$$

Here, equality again holds if and only if $\theta \cong \overset{o}{\theta}$.

Let $\phi^N = (\phi_0^N,\phi_1^N,\ldots)$ denote the impulse response sequence (4.1). Since the function $(A(z),B(z)) \to A^{-1}(z)\ B(z) = \phi_0+\phi_1 z+\ldots$ is continuous when the space of the image sequences is given the l_1-metric in (4.2), the set $T = \{\phi(z) = A_N^{-1}(z)\ B(z)|(A,B) \in S\}$ is compact. Consider the subset $\{\phi^N\}$ of T corresponding to the m.ℓ. estimates $\theta^N \in S$ for N=1,2,... We shall construct accumulation points of $\{\phi^N\}$ in T which are measurable random variables with respect to a σ-field generated by the y process. The following measurable selection rule achieves this: Let $\phi_{n_{ij}}$ denote the i, j entry of the matrix ϕ_n in the sequence defining ϕ. To begin, we define $\bar{\phi}_{0_{11}} = \limsup_{N\to\infty} \phi^N_{0_{11}}$ and $\underline{\phi}_{o_{11}} = \liminf_{N\to\infty} \phi^N_{o_{11}}$.

These random variables are measurable with respect to the tail σ-field of the y process. We may now select subsequences $\{\bar{N}_{1,k}\} \subset \{N_{0,k}\}$ and $\{\underline{N}_{1,k}\} \subset \{N_{0,k}\}$ where $\{N_{0,k}\} \triangleq \{N_k\}$ for which these accumulation points are limit points.

At stage ν we define $\bar{\phi}_{n_{ij}} = \limsup_{N\to\infty} \phi^N_{n_{ij}}$, $N \in \{\bar{N}_{\nu-1,k}\}$, and $\underline{\phi}_{n_{ij}} = \liminf_{N\to\infty} \phi^N_{n_{ij}}$, $N \in \{\underline{N}_{\nu-1,k}\}$, where $\nu = np^2 + (i-1)\,p + j$. The subsequences $\{\bar{N}_{\nu,k}\} \subset \{\bar{N}_{\nu-1,k}\}$ and $\{\underline{N}_{\nu,k}\} \subset \{\underline{N}_{\nu-1,k}\}$ are then extracted and the process repeated. Continuing in this way we clearly obtain accumulation points $\bar{\phi}$ and $\underline{\phi}$ of $\{\phi^N\}$ which are measurable with respect to the tail σ-field of the y process.

Let ϕ^* denote either $\bar{\phi}$ or $\underline{\phi}$ and let θ be any parameter (A,B) chosen by some specified rule such that $A(z)^{-1}B(z) = \phi_0 + \phi_1 z + \dots$. Finally, let M denote any member of the appropriate subsequence for which $\lim_{M\to\infty} \theta^M = \theta^*$.

By writing $L_M(y_0,\dots,y_M,\theta,\mathring{\theta})$ as $L_M(\theta,\mathring{\theta})$ we have $L_M(\mathring{\theta},\mathring{\theta}) \geq L_M(\theta^M,\mathring{\theta})$ by the minimizing property of θ^M. By Lemma 3, for each $\varepsilon > 0$, there exists N_ε, not dependent on $\theta^M \in S$, such that

$$L_M(\theta^M,\mathring{\theta}) \geq L(\theta^M,\mathring{\theta}) - \varepsilon \qquad \text{a.s. for all } M > N_\varepsilon$$

But then from these two inequalities we get:

$$L(\mathring{\theta},\mathring{\theta}) = \lim_{M\to\infty} L_M(\mathring{\theta},\mathring{\theta}) \geq \lim_{M\to\infty} L(\theta^M,\mathring{\theta}) = L(\theta^*,\mathring{\theta}) \quad \text{a.s.} \tag{4.5}$$

The inequalities (4.4) and (4.5) with $\theta = \theta^*$ give

$$L(\mathring{\theta},\mathring{\theta}) = L(\theta^*,\mathring{\theta}) \text{ a.s.}$$

which implies $\theta^* \cong \overset{o}{\theta}$ and $\phi^* = \overset{o}{\theta}$ a.s. . It follows that $\bar{\phi} = \phi^0$ a.s. and $\underline{\phi} = \phi^0$ a.s. By the construction of these limits this implies that $\lim_{N\to\infty} \phi^N$ exists with $\lim_{N\to\infty} \phi^N = \phi^0$ a.s. and this gives the theorem.

Concluding Remarks

If from each estimated sequence ϕ^N one also estimates the Kronecker indices in a consistent manner then θ^N can be picked in a suitable canonical form, as described in [2] and [3], and the sequence θ^N itself converges a.s. .

Often in a system of type (1.2) B_0 is not invertible, e.g. the u-process has fewer components than the y-process. Such systems are pathological in the sense that the y-process is not of a full rank and the maximum likelihood function (2.9) is not valid for such processes. What one can do in such a case is to pick a largest full rank process out of the components of the y-process, and estimate the parameters as above for that subprocess.

References

[1] N. Wiener, P. Masani, "The Prediction Theory of Multivariate Stochastic Processes," I, Acta Mathematica 98, pp. 111-150, 1957.

[2] V. M. Popov, "Some Properties of the Control Systems with Irreducible Matrix-Transfre Functions," Lecture Notes in Math., No. 144, Springer-Verlag, Berlin, 1969.

[3] J. Rissanen, "Basis of Invariants and Canonical Forms for Linear Dynamic Systems," Automatica, March 1974 (to appear).

[4] H. Akaike, "Markovian Representation of Stochastic Processes and Its Application to the Analysis of Autoregressive Moving Average Processes," TR No. 41, Dept. of Mathematics, The Univ. of Manchester Institute of Science and Technology, May 1973.

[5] D. Q. Mayne, "Parameterization and Identification of Linear Multivariable Systems," Stability of Stochastic Dynamical Systems, Springer Lecture Notes in Math., No. 294, pp. 56-61, 1972.

[6] D. Q. Mayne, "A Canonical Model for Identification of Multivariable Linear Systems," IEEE Trans. Automatic Control, Vol. AC-17, No. 5, pp. 728-729, 1972.

[7] P. Faurre, "Identification par Minimization d'une Representation Markovienne de Processus Aleatoire," Symp. on Optimization, Springer Lecture Notes in Math., Vol. 132, 1970.

[8] P. E. Caines, "The Parameter Estimation of State Variable Models of Multivariable Linear Systems," Proc. Fourth UKAC Control Convention, Manchester, September 1971, IEE Conference Publication No. 78.

[9] A. Wald, "Note on the Consistency of the Maximum Likelihood Estimate," Ann. Math. Stat., Vol. 20, pp. 595-601, 1949.

[10] M. G. Kendall, A. Stuart, The Advanced Theory of Statistics, Vol. 2, Hafner Publishing Co., New York, 1967.

[11] C. A. Galtieri, "The Problem of Identification of Discrete Time Processes, " ERL Report No. 63-19, University of California, Berkeley, 1963.

[12] K. J. Åström, T. Bohlin, S. Wensmark, "Automatic Construction of Linear Stochastic Dynamic Models for Stationary Industrial Processes with Random Distrubances Using Operating Records," Rep. TP 18.150, IBM Nordic Laboratories, Lidingö, Sweden, 1965.

[13] G. C. Roussas, "Extension to Markov Processes of a Result by A. Wald About the Consistency of the Maximum Likelihood Estimate," Z. Wahrscheinlichkeitsteorie, Vol. 4, pp. 69-73, 1967.

[14] P. E. Caines, J. Rissanen, "Maximum Likelihood Estimation of Parameters in Multivariate Gaussian Stochastic Processes," IEEE Trans. Information Theory, Vol. IT 20, No. 1, January 1974.

[15] T. Kailath, "The Innovations Approach to Detection and Estimation Theory," Proc. IEEE, Vol. 58, No. 5, May, 1970.

[16] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," J. Basic Eng., ASME, Vol. 82, pp. 35-45, 1960.

[17] J. Rissanen, L. Barbosa, "Properties of Infinite Covariance Matrices and STability of Optimum Predictors," Inform. Sci., 1, pp. 221-236, 1969.

[18] F. Bauer, "Ein Direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms," Mitt aus dem Mat. Inst. der Tech. Hochschule München, AEU, 285-290, 1955.

[19] J. Rissanen, "Algorithms for Triangular Decomposition of Block Hankel and Toeplitz Matrices with Application to Factoring Positive Matrix Polynomials," Mathematics of Computation, Vol. 27, No. 121, January 1973.

[20] J. Rissanen, T. Kailath, "Partial Realization of Random Systems," Automatica, Vol. 8, 1972, pp. 389-396.

[21] R. E. Kalman, J. Bertram, "Control System Analysis and Design via the "second method" of Lyapunov," J. Basic Eng., 82, June 1960, pp. 371-393.

[22] B. D. O. Anderson, K. L. Hitz, N. D. Diem, "Recursive Algorithm for Spectral Factorization," (to appear).