JOURNAL OF MULTIVARIATE ANALYSIS 11, 459–473 (1981)

Estimating the Dimension of a Linear System

E. J. HANNAN

The Australian National University

Communicated by the Editors

The problem considered is that of estimating the integer or integers that prescribe the dimension of a linear system. These could be the Kronecker indices. Though attention is concentrated on the order or McMillan degree, which specifies the dimension of a minimal state vector, the same results are available for other cases. A fairly complete theorem is proved relating to conditions under which strong or weak convergence will hold for an estimate of the McMillan degree when the estimation is based on minimisation of a criterion of the form $\log\det(\hat{\Omega}_n) + nC(T)/T$, where $\hat{\Omega}_n$ is the estimate of the prediction error covariance matrix and the McMillan degree is assumed to be n. The conditions relate to the prescribed sequence $C(T)$.

1. INTRODUCTION

Let $y(t)$, $t = 1,..., T$, be r-vectors generated by an ergodic stationary process and satisfying

$$\sum_0^p A(j)\, y(t-j) = \sum_0^q B(j)\, \varepsilon(t-j), \qquad A(0) = B(0), \tag{1.1}$$

$$\mathscr{E}\{\varepsilon(t) \mid \mathscr{F}_{t-1}\} = 0, \qquad \mathscr{E}\{\varepsilon(t)\,\varepsilon(t)' \mid \mathscr{F}_{t-1}\} = \Omega > 0. \tag{1.2}$$

Here $\mathscr{F}_t$ is the σ-algebra of events determined by the $\varepsilon(s)$, $s \leqslant t$, Ω is a constant matrix and $A(0)$ is nonsingular. Put $\tilde{g}(z) = \Sigma A(j) z^j$, $\tilde{h}(z) = \Sigma B(j) z^j$. (The tilde has been used to reserve g, h, without such an addition, for another purpose.) It may be assumed that

$$\tilde{g}, \tilde{h} \text{ are relatively left prime and } \det(\tilde{g}) \neq 0,\ (z) \leqslant 1,\ \det(\tilde{h}) \neq 0,\ |z| < 1. \tag{1.3}$$

(The definition of prime polynomial matrices is given in [18].) When (1.3)

Received January 22, 1980; revised March 13, 1981.
AMS 1979 subject classifications: Primary 62M99; Secondary 93B30.
Key words and phrases: Ergodic, stationary process. ARMA, McMillan degree, Kronecker indices, manifold, AIC, strong convergence, weak convergence, law of the iterated logarithm.

0047-259X/81/040459–15$02.00/0

holds the $\varepsilon(t)$ are the linear innovations. As is well known [5, 18] $y(t)$ may also be represented in state space form

$$y(t) = Hx(t) + \varepsilon(t), \qquad x(t+1) = Fx(t) + K\varepsilon(t),$$

where $x(t)$ is not observed and is made up of predictions of the $y_j(t+k)$, for various k, from $y(s)$, $s \leqslant t-1$. F may be chosen to be of minimal dimension, n let us say, and this integer is called the McMillan degree. It is shown in [4] that the set of all structures (1.1), for given n, Ω, may be topologised as an analytic manifold, of dimension $2nr$, which will be called $M(n)$, if $\det(\tilde{h}) \neq 0$, $|z| \leqslant 1$. (This last condition is of no importance of itself but reflects only the fact that a manifold is open. In any case a stricter condition will shortly be imposed.) There are $r(r+1)/2$ additional parameters needed to describe Ω. The manifold, $M(n)$, may be covered by $\binom{n+r-1}{r-1}$ coordinate neighbourhoods, each dense in $M(n)$. For details concerning these neighbourhoods the reader may consult [7] and the references therein. Here we give only a brief description.

Put $k(z) = \tilde{g}(z^{-1})^{-1}\tilde{h}(z^{-1})$. This is evidently a matrix of rational functions of z. We may find polynomial matrices g, h so that $k(z) = g(z)^{-1}h(z)$, g, h left prime, so that g has diagonal elements that are monic polynomials of degree n_i, for the ith row, $\Sigma n_i = n$, and $\delta(g_{ij}) < n_j$, $i \neq j$, where $\delta(\cdot)$ is the degree of the indicated polynomial. The set of all k having such a representation for given n_i constitutes one of the coordinate neighbourhoods and there is one such neighbourhood for each of the $\binom{n+r-1}{r-1}$ partitions, $n = \Sigma n_i$. It is shown in [7] that coordinates mapping the neighbourhood into $2nr$ dimensional Euclidean space may be chosen to be the elements of the coefficient matrices of g that are not identically 0 or 1, and, if they lie in the ith row, are coefficients of z^m, $m \leqslant n_i - 1$. We shall index the coordinate neighbourhoods by α and call U_α a typical neighbourhood. The points of $M(n)$ are transfer functions, $k(z)$, but it will also be convenient to refer to them by the symbol k_θ and we shall also write k_θ for a representative transfer function. If ϕ_θ is the mapping that maps U_α into Euclidean space via the coordinates in U_α then we shall write $\theta_\alpha = \phi_\alpha(\theta)$. Each U_α contains a subset V_α consisting of those k_θ for which g, h may be chosen so that, in addition to $\delta(g_{ij}) < n_j$, $i \neq j$, it may be further required that $\delta(g_{ij}) < n_i$, $j > i$; $\delta(g_{ij}) \leqslant n_i$, $j \leqslant i$; $\delta(h_{ij}) \leqslant n_i$, $j = 1,..., r$. If $n_1 = n_2 = \cdots = n_s = n_{j+1} + 1 = \cdots = n_r + 1$ (so that $n = n_1 r + s$, $s < r$) then $V_\alpha = U_\alpha$ but otherwise V_α is of lower dimension, $d(n_j) = n(r+1) + \sum\sum_{j<k}\{\min(n_j, n_k) + \min(n_j, n_k + 1)\}$. The V_α are disjoint and their union is $M(n)$. The canonical forms just described are called the echelon forms (see [5, 10]). Each V_α is a submanifold of $M(n)$ and V_α may be mapped diffeomorphically into Euclidean space of dimension $d(n_j)$, for the appropriate n_j, by means of the coordinates which are the elements, not identically zero or unity, of the coef-

ficient matrices of the echelon form. The n_j for the echelon form, for a given $k(z)$, when arranged (say) in non-decreasing order of magnitude, are called the Kronecker invariants. When taken in the (arbitrary) order determined by the order in which the components of $y(t)$ were arranged down the vector, we shall call the n_j, for this echelon form, the Kronecker indices. If $k = g^{-1}h$ is in echelon form and we write g_i, $\tilde{g}_i$, etc., for the ith row of $g, \tilde{g}$, etc., then putting

$$\tilde{g}_i = z^{n_i} g_i(z^{-1}), \qquad \tilde{h}_i = z^{n_i} h_i(z^{-1}),$$

the matrices $\tilde{g}, \tilde{h}$ will satisfy (1.1), (1.3).

We here consider maximum likelihood estimation (ML) based on the likelihood constructed as if the data were Gaussian, though that hypothesis is not maintained. Let ν be the true McMillan degree. Except for ν we shall use a zero subscript for a true value. We shall use a "hat" for an estimate. When ν is known then it is shown in [14] that $\hat{\theta} \to \theta_0$, a.s., and, if $\theta \in U_\alpha$, then eventually (i.e., for T large enough) $\hat{\theta} \varepsilon U_\alpha$ and $T^{1/2}(\hat{\theta}_\alpha - \theta_\alpha)$ obeys the central limit theorem. Thus the problem of estimating ν needs consideration. This is given in [1, 9, 17] for special cases or with an incomplete treatment and, in the second reference, by different methods from those used here. Let $\hat{\Omega}(n)$ be the ML estimator of Ω_0 if it is assumed that $\nu = n$. Then, following [1], put

$$A(n) = \log \det\{\hat{\Omega}(n)\} + nC(T)/T,\ C(T) \geqslant 0, \qquad n \leqslant N. \tag{1.4}$$

Here $C(T)$ and N are prescribed a priori. If $C(T) = 4r$ then $A(n)$ is what would usually be called AIC (see [1]). The integer $\hat{n}$ is chosen to minimise $A(n)$. The theorem proved below requires additional restrictions.

$$\begin{gathered} \det(k_0) \neq 0, |z| > 1 - \delta, \delta > 0; \qquad \|\phi_\alpha(\theta_0)\| \leqslant \rho < \infty; \\ \{|\varepsilon_j(t)|^\gamma\} < \infty, j = 1,..., r. \end{gathered} \tag{1.5}$$

By the second of these conditions we mean that in some suitable coordinate system the Euclidean length of $\theta_{0\alpha}$ is known to be bounded, a priori. Correspondingly we shall, for any n, consider only points θ such that, for some α, $\|\theta_\alpha\| \leqslant \rho$. In practice this means that the coordinates used in an iterative solution of the equations of ML will be restricted in magnitude. Similarly only points $\theta \in M(n)$ will be considered for which $\det(k_\theta) \neq 0$, $|z| > 1 - \delta$. These first two parts of (1.5) will be very difficult to avoid and it may be that they are impossible to avoid. It would seem incautious, to say the least, to commence a calculation without imposing such conditions. The third part of (1.5) is minor. The value of γ will be prescribed in the theorem below.

THEOREM. (i) *If* (1.2), (1.5) *hold for* $\gamma = 4$, $\nu \leqslant N$, *the* $\varepsilon(t)$ *are independent random vectors and*

$$\liminf_{T\to\infty} C(T)/\{2 \log\log T\} > 1, \qquad C(T)/T \to 0, \tag{1.6}$$

then $\hat{n} \to \nu$, a.s. *If* $\limsup C(T)/\{2\log\log T\} < 1$ *and* $\nu < N$, *then a.s. convergence does not hold.*

(ii) *If* (1.2), (1.5) *hold for* $\gamma > 4$, $\nu \leqslant N$, *and*

$$\liminf_{T\to\infty} C(T)/\{\log T\} > 0, \qquad C(T)/T \to 0, \tag{1.7}$$

then $\hat{n} \to \nu$ *a.s.*

(iii) *If* $C(T)/T \to 0$, $C(T) \uparrow \infty$, *under* (1.2), (1.5) *for* $\gamma > 4$ *and* $\nu \leqslant N$, *then* $\hat{n} \to \nu$ *in probability. If* (1.2), (1.5) *hold,* $\gamma \geqslant 4$, $\nu < N$, *and*

$$\limsup_{T\to\infty} C(T) < \infty, \tag{1.8}$$

then

$$\lim_{\delta\to 0} \lim_{T\to\infty} P\{\hat{n} > \nu\} = 1. \tag{1.9}$$

This theorem gves a fairly complete account of the asymptotic behaviour of $\hat{n}$. The requirement of independence in relation to (1.6) is almost certainly unnecessary and the result will hold under the same conditions as for (1.7). However, this has not been proved. If $\liminf C(T)/T = c > 0$ then it is still possible that $\hat{n} \to \nu$, a.s. Indeed this will be so if and only if

$$\max_{n<\nu} (\nu - n)^{-1}\{\log\det \hat{\Omega}_n - \log\det \Omega_0\} > c;$$

$$\hat{\Omega}_n = \inf_{\theta\in M(n)} \int_{-\pi}^{\pi} k_\theta(e^{i\omega})^{-1} f_0(\omega)\, k_\theta^*(e^{i\omega})^{-1}\, d\omega; \tag{1.10}$$

$$2\pi f_0(\omega) = k_0(e^{i\omega})\, \Omega_0 k_0^*(e^{i\omega}) \ldots .$$

However for any ν there will be k_0 for which (1.10) does not hold. This result, which is relatively easy to establish, shows that $C(T)/T \to 0$ is an unavoidable restriction. In any case one will wish to take $C(T)$ as small as is consistent with convergence since underestimation of ν is more serious than overestimation. The result (1.9) is noteworthy for, while *AIC* is known not to be consistent (see [19]), in the present context it is shown to be sure to overestimate ν, in a certain sense. For several reasons this result may not be as important as might appear. In the first place the result is asymptotic and T may need to be very large before (1.9) is relevant. In the second place it may not be true that $\nu < \infty$. The result of overestimating ν may also not be very serious for prediction because even if $n > \nu$ still $\hat{k} \to k_0$, a.s. and uniformly on the closed unit disc. Of course if $n > \nu$ too many parameters will be estimated and some efficiency will be lost. For some purposes, also, the actual representation in coordinates may be important. For $n > \nu$, $\hat{\theta}_\alpha$ may not converge, and in any case not to a vector having much meaning in relation to the physical origins of the problem. This may mislead the investigator. For example, in the scalar case for $n = \nu + 1$, an additional zero and pole, nearly cancelling, will appear in $\hat{k}$ and as shown in [13]; if δ is small these will be near to the unit circle. This could give rise to a false impression of the nature of k_0 in relation to stability questions.

The techniques of the present paper apply to other cases. For example one might set out to estimate the Kronecker indices, forming $A(n_j) = \log\det \Omega(n_j) + d(n_j)\, C(T)/T$, $n_j \leqslant N_j$, $\nu_j \leqslant N_j$, where the ν_j are the true indices and the likelihood is maximised over V_α for α corresponding to these indices. A result fully equivalent to the theorem is then true. Again we could consider (1.1) subject to (1.3), and $[A(p){:}\, B(q)]$ of full rank (see [6, 8]). This set of structures, $M(p, q)$ let us say, may again be topologised as a manifold, of dimension $(p+q)r^2$, in the trivial sense that it may be represented as an open set in Euclidean space of that dimension. This kind of case relates closely to the situation where the actual coordinates (i.e., the elements of the $A(j)$, $B(j)$) have meaning, so that overestimation of ν may be important. The theorem may not fully generalise in this case because it is possible that a given $k(z)$ may have representation in $M(p, q)$ for different p, q. Nevertheless essentially the same results will hold. However, we confine the proof to the case stated in the theorem to avoid a complicated profusion.

2. Proof of the Theorem

Let $g_0^{-1} h_0 = k_0$ be the, unique, echelon form representation of k_0 with ν_j, $j = 1,\ldots, r$, $\sum \nu_j = \nu$, being the Kronecker indices. Under very general conditions, including (1.1), (1.2), (1.3), $\hat{k}(z) \to k_0(z)$, a.s. and uniformly on the closed unit disc and $\hat{\Omega} \to \Omega_0$ if $n \geqslant \nu$ (see [6, 8]). Let $\tilde{M}(n)$ be the closure of the part of $M(n)$ for which the middle part of (1.5) holds for some α.

The key idea is the introduction of new coordinates (See [12, 13] for $r = 1$, which case is much simpler). Unproved statements in the present paragraph are established in [7], and repeated reference will not be made to that. Let $\bar{M}(n)$ be the closure of $M(n)$. Points in $\bar{M}(n)$ not in $M(n)$ are either points for which $\det(k) = 0$, for some z with $|z| = 1$, or points for which $k(z)$ has a pole on the unit circle or they correspond to $k(z)$ for McMillan degree less than n. Each point of $\bar{M}(n)$ is a limit point of each U_α since U_α is dense in $M(n)$. However, $\phi_\alpha(U_\alpha)$ might not contain, in its closure, any point corresponding to some point in $\bar{M}(n)$ because that point is mapped onto infinity by ϕ_α. Indeed if $\nu < n$ then k_0 will be represented as a (finite) limit point of $\phi_\alpha(U_\alpha)$ if and only if $n_i \geqslant \nu_i$, $i = 1,..., r$. *Indices α for which this last relation holds will be called "proper,"* for brevity. Since $\hat{k} \to k_0$ then eventually $\hat{k}$ can be represented by a point, $\hat{\theta}_\alpha$, in the closure of $\phi_\alpha(U_\alpha)$, only for proper α. Along an infinite sequence of T values for which $\hat{\theta}_\alpha$ has meaning then $\hat{h}_\alpha - \hat{g}_\alpha k_0 \to 0$, since $\hat{g}_\alpha$ is uniformly bounded. Of course if we optimised only over $\phi_\alpha(U_\alpha)$, for proper α, then the optimal $\hat{k}$ would again converge to k_0. Here $\hat{k} = \hat{g}_\alpha^{-1}\hat{h}_\alpha$ is the representation of $\hat{k}$ corresponding to the αth coordinate system. Thus for any α we consider

$$h_\alpha - g_\alpha k_0 = \sum_{-\infty}^{m} \psi^{(\alpha)}(j)\, z^j, \qquad m \leqslant \max_j (n_j) - 1. \tag{2.1}$$

We consider elements of the coefficient matrices on the right side of (2.1), for $j \leqslant n_i - 1$ for the ith row. These are then linear (affine) expressions in θ_α and the infinite set of these has rank $\leqslant (n+\nu)r$ and this rank is actually $(n+\nu)r$ if and only if α is proper. We choose a maximal linearly independent set for each α and arrange these as the elements of a vector $\psi_\alpha = \psi(\theta_\alpha)$, $\theta \in U_\alpha$. Hence, for α proper, ψ_α has $(n+\nu)r$ components. If $\psi_\alpha = 0$ and α is proper then $k = k_0$ and conversely when $k = k_0$, $\psi_\alpha = 0$. When $\psi_\alpha = 0$ then $g_\alpha = d_\alpha g_0$, $h_\alpha = d_\alpha h_0$. We may, when α is proper, find precisely $(n - \nu_0)r$ further linear expressions in θ_α which together with the elements of ψ_α make up a linearly independent set of $2nr$ such linear expressions. Call χ_α the vector made up of these $(n-\nu_0)r$ further functions. Then the elements of ψ_α, χ_α may equally well serve as coordinates in $\phi_\alpha(U_\alpha)$. It is evident that d_α, above, is a function only of χ_α. There is a further technical point with which we now deal. It is conceivable that $\hat{\psi}_\alpha, \hat{\chi}_\alpha$ is a boundary point of the closure of $\phi_\alpha(U_\alpha)$. Eventually this can only be so because $\det\{\hat{h}_\alpha(z)\}$ has a zero on the boundary of the unit disc or $\hat{\chi}_\alpha$ is so large that θ_α is at the boundary of the region defined by the middle part of (1.5). Any other point $\hat{\psi}_\alpha, \hat{\chi}_\alpha$ will eventually be interior. (Points for which $\hat{k}$ has a pole on the unit circle eventually cannot occur because $\hat{g}_\alpha, \hat{h}_\alpha$ must eventually be near to a point $d_\alpha g_0$, $d_\alpha h_0$ and the first part of (1.5) prevents $\det(d_\alpha)$ from having a zero too near the unit circle. For the same reason points for which $\hat{\psi}_\alpha$ are large need not be considered.) We can always either slightly enlarge $\phi_\alpha(U_\alpha)$ or slightly reduce it, by an amount that can be made to depend on δ so as to converge to zero as δ does so, so that at "edge" points of the section, $\psi_\alpha = 0$, the boundary of $\phi_\alpha(U_\alpha)$ cuts $\psi_\alpha = 0$ orthogonally. Since $\hat{\psi}_\alpha \to 0$, then at any point $(\hat{\psi}_\alpha, \hat{\chi}_\alpha)$ it must be true that the derivative of any function (for example, the likelihood) being optimised, with respect to the component of ψ_α, must be zero. In the three parts of the theorem where a sufficient condition for convergence is being proved (which we call the sufficiency parts), if, for $n > \nu$, it is shown that ν is preferred to that n, asymptotically, then, a fortiori, the result will be established for optimisation over $\tilde{M}(n)$. For the first and third parts, which effectively establish a necessary condition for convergence, the situation is reversed so that we can proceed by optimising over a slightly smaller region. We shall therefore, in the developments below, without comment assume that the derivatives with respect to the components of ψ_α are zero at the optimal points.

If $\mathscr{L}$ is the likelihood let $L = -2T^{-1} \log \mathscr{L}$. Then, omitting a constant,

$$L(\theta, \Omega) = T^{-1} \log \det \{\Gamma_T(\theta, \Omega)\} + T^{-1} y_T' \Gamma_T(\theta, \Omega)^{-1} y_T.$$

Here y_T has $y_j(t)$, $j = 1,..., r$, $t = 1,..., T$, as components in dictionary order, first according to t and then j, $\Gamma_T = \mathscr{E}\{y_T y_T'\}$. For the αth coordinate neighborhood let $A_\alpha(j)$, $B_\alpha(j)$ be the coefficient matrices in g_α, h_α and let A_α have T rows and columns of blocks of $r \times r$ matrices, and be lower triangular with $A_\alpha(j)$ in the jth diagonal of blocks below the main diagonal. Let B_α be similarly defined in terms of the $B_\alpha(j)$. Then, for T sufficiently large, $K_\theta = A_\alpha^{-1} B_\alpha$ is lower triangular with $K_\theta(j)$, in the jth diagonal of blocks, where

$$k_\theta(z) = \sum_0^\infty K_\theta(j) z^{-j}, \qquad K_\theta(0) = I_r.$$

Thus K_θ is independent of the coordinate system. Then it is easily shown that

$$\Gamma_T = K_\theta(I_T \otimes \Omega) K_\theta' + A_\alpha^{-1} R_\alpha A_\alpha^{-1\prime}, \tag{2.2}$$

where R_α is null outside of a block in the top left-hand corner with a fixed number of rows and columns. Evidently the last term in (2.2) is also independent of the coordinate system. Let $\tilde{\Omega}(\theta)$ be the sum of the diagonal blocks of $T^{-1} y_T'(K_\theta')^{-1}(I_T \otimes \Omega^{-1}) K_\theta^{-1} y_T$. Then

$$\tilde{\Omega}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} k_\theta^{-1} I(\omega) k_\theta^{*-1} \, d\omega. \tag{2.3}$$

Here we have omitted the argument variable, exp $i\omega$, in k_θ, for brevity. Here also

$$I(\omega) = w(\omega) w(\omega)^*, \qquad w(\omega) = \frac{1}{\sqrt{T}} \sum_1^T y(t) e^{-it\omega}.$$

Thus after a little manipulation it may be checked that

$$L(\theta, \Omega) = \log \det \Omega + \operatorname{tr} \{\Omega^{-1} \tilde{\Omega}(\theta)\} + T^{-1} \rho_T(\theta, \Omega),$$

where $\rho_T(\theta, \Omega)$, and its derivatives with respect to any coordinates, are a.s. uniformly bounded. Let $\tilde{\theta}$ minimise $\log\det\tilde{\Omega}(\theta)$. Now $\tilde{\Omega}(\theta)$ converges, uniformly in θ in $\tilde{M}(n)$, and almost surely (see [6, 8, 14]) to

$$\int_{-\pi}^{\pi} k_\theta^{-1} f_0(\omega)\, k_\theta^{*-1}\, d\omega. \tag{2.4}$$

For $n \geqslant \nu$, $\tilde{\Omega}(\tilde{\theta})$ converges to Ω_0 and k_θ, evaluated at $\tilde{\theta}$, converges to k_0 (see [6, 8]). Thus

$$L(\hat{\theta}, \hat{\Omega}) = \log\det\{\tilde{\Omega}(\tilde{\theta})\} + r + O(T^{-1})$$

$$\log\det(\hat{\Omega}) = \log\det\{\tilde{\Omega}(\tilde{\theta})\} + O(T^{-1}),$$

where the terms that are $O(T^{-1})$ are of that order a.s., and uniformly in $\hat{\theta}, \tilde{\theta}$. (When the notation $O(T^{-1})$, or $o(1)$ or $o(T^{-1/2})$, and so on, is used it will always be meant that the neglected terms are of the indicated order a.s. and uniformly in $\tilde{\theta}$ or $\hat{\theta}$, etc.) Thus we may replace $A(n)$ by

$$\tilde{A}(n) = \log\det\{\tilde{\Omega}(\tilde{\theta})\} + nC(T)/T.$$

Since $\tilde{\Omega}(\tilde{\theta})$ depends on n we shall write this as $\tilde{\Omega}_n(\tilde{\theta})$. If $n < \nu$ then the lower bound to the determinant of (2.4) is strictly greater than $\det(\Omega_0)$. Thus for $n < \nu$

$$\lim_{T\to\infty} \{\log\det\tilde{\Omega}_n(\tilde{\theta}) - \log\det\tilde{\Omega}_\nu(\tilde{\theta})\} > 0, \text{ a.s.},$$

and hence $\tilde{A}(n) - \tilde{A}(\nu) > 0$ for all T sufficiently large, a.s., $n < \nu$, provided $C(T)/T \to 0$. Thus we cannot have $\hat{n} < \nu$, at least for T sufficiently large and we henceforth consider only $n \geqslant \nu$. We need to show, for the first parts of the theorem, that, eventually,

$$\log\det\tilde{\Omega}_n(\tilde{\theta}) - \log\det\tilde{\Omega}_\nu(\tilde{\theta}) + (n-\nu)\, C(T)/T > 0, \qquad n > \nu, \tag{2.5}$$

either a.s. or in probability as the case may be, so that ν is preferred to n. Let $\tilde{\theta}$ have the coordinates $\tilde{\psi}_\alpha, \tilde{\chi}_\alpha$ in a suitable, proper, coordinate system. Now $\det\tilde{\Omega}_n(0, \tilde{\chi}_\alpha)$ is $\det\tilde{\Omega}_\nu(\theta)$ evaluated at θ_0, since $\tilde{\Omega}(\theta)$ depends only on k_θ. Since $\det\tilde{\Omega}_\nu(\tilde{\theta})$ is the minimised value of $\tilde{\Omega}_\nu(\theta)$ over $\tilde{M}(\nu)$ then $\det\tilde{\Omega}_n(0, \tilde{\chi}_\alpha) \geqslant \det\tilde{\Omega}_\nu(\tilde{\theta})$. Also $T[\log\det\tilde{\Omega}_n(0, \tilde{\chi}_\alpha) - \log\det\tilde{\Omega}_\nu(\tilde{\theta})]$ is asymptotically distributed as chi-square with $2\nu r$ degrees of freedom [6, 8]). Thus since $C(T) \uparrow \infty$ we may replace (2.5), for the three sufficiency parts of the theorem by

$$\log\det\{\tilde{\Omega}_n(\tilde{\theta})\} - \log\det\{\tilde{\Omega}_n(0, \tilde{\chi}_\alpha)\} + (n-\nu)\, C(T)/T, \qquad \nu < n. \tag{2.6}$$

We deal with the second part of the theorem last in the proof.

For the final part of the theorem if we show that for some $n > \nu$ and $\limsup C(T) < \infty$ the probability, that T by (2.6) is less than $-C$ for any finite positive constant C, will, as first $T \to \infty$ and then $\delta \to 0$, converge to unity, then evidently the same will be true for (2.5) because of the nature of the asymptotic distribution of $T\,[\log\det \tilde{\Omega}_n(0, \tilde{\chi}_\alpha) - \log\det \Omega_\nu(\tilde{\theta})]$. Thus we henceforth need consider only (2.6). Moreover if we prove the theorem when $\tilde{\theta}, \tilde{\chi}_\alpha$ are retricted to a given proper coordinate system, for the three suffiency parts, we shall evidently have proved the theorem as stated. Thus we now confine ourselves to a fixed proper α. Since $\partial \log\det \tilde{\Omega}(\tilde{\theta})/\partial\psi_\alpha$ is null then

$$\log\det \tilde{\Omega}(\tilde{\theta}) - \log\det \tilde{\Omega}(0, \tilde{\chi}_\alpha)$$
$$= \frac{\partial \log\det \tilde{\Omega}(0, \tilde{\chi}_\alpha)}{\partial \psi_\alpha}\tilde{\psi}_\alpha + \tilde{\psi}'_\alpha \frac{\partial^2 \log\det \tilde{\Omega}(\bar{\psi}, \tilde{\chi}_\alpha)}{\partial\psi_\alpha\,\partial\psi'_\alpha}\tilde{\psi}_\alpha, \tag{2.7}$$

where $\bar{\psi}$ is intermediate between 0 and $\tilde{\psi}_\alpha$ and hence also converges to zero. Also

$$\frac{\partial \log\det \tilde{\Omega}(0, \tilde{\chi}_\alpha)}{\partial\psi_\alpha} = \frac{\partial^2 \log\det \tilde{\Omega}(\bar{\bar{\psi}}, \tilde{\chi}_\alpha)}{\partial\psi_\alpha\partial\psi'_\alpha}\tilde{\psi}_\alpha,$$

where $\bar{\bar{\psi}}$ is again intermediate between 0 and $\tilde{\psi}_\alpha$. Now

$$\frac{\partial \log\det \tilde{\Omega}(0, \tilde{\chi}_\alpha)}{\partial\psi_\alpha} = \frac{\partial}{\partial\psi_\alpha}\log\det\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} k_\theta^{-1} I(\omega)\, k_\theta^{*-1}\, d\omega\right\}_{0,\tilde{\chi}_\alpha},$$

where the notation indicates evaluation at $0, \tilde{\chi}_\alpha$ after differentiation. To evaluate this gradient let $\Psi_{\alpha,u}$ be the derivative of k_θ with respect to the uth element of ψ_α, evaluated at $(0, \tilde{\chi}_\alpha)$. Recall also that for any non singular matrix function, $X(x)$, $\partial \log\det X/\partial x = \operatorname{tr}(X^{-1}\,\partial X/\partial x)$, $\partial X^{-1}/\partial x = -X^{-1}\,\partial X/\partial x X^{-1}$. We then obtain for the uth component of the gradient

$$-2\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\{\tilde{\Omega}(0, \tilde{\chi}_\alpha)^{-1} k_0^{-1}\Psi_{\alpha,u} k_0^{-1} I(\omega)\, k_0^{*-1}\}\, d\omega.$$

Now, by precisely the same argument as in [12, pp. 1075, 1076] we may, neglecting a term that is $O(T^{-1/2})$, replace $I(\omega)$ by $k_0 I_\epsilon(\omega) k_0^*$, where $I_\epsilon = w_\epsilon w_\epsilon^{-*}$, $w_\epsilon = T^{-1/2}\sum \varepsilon(t)\exp(-it\omega)$, so that we arrive at

$$-\frac{1}{\pi}\int_{-\pi}^{\pi} \operatorname{tr}\{\tilde{\Omega}(0, \tilde{\chi}_\alpha)^{-1} k_0^{-1}\Psi_{\alpha,u} I_\epsilon(\omega)\}\, d\omega. \tag{2.8}$$

Now, recalling (2.3), $\tilde{\Omega}(0, \tilde{\chi}_\alpha)$ converges to Ω_0 so that (2.8) is

$$-2\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\{\Omega_0^{-1} k_0^{-1}\Psi_{\alpha,u}(\tilde{\chi}_\alpha) I_\epsilon(\omega)\}\, d\omega\{1 + o(1)\}. \tag{2.9}$$

Call $b_u(\tilde{\chi}_\alpha)$ the expression (2.9) ignoring $-2\{1+o(1)\}$ and call $b(\tilde{\chi}_\alpha)$ the vector of these $(n+v)r$ quantities. If χ_α were fixed then $T\mathscr{E}\{b(\chi_\alpha)\,b(\chi_\alpha)'\} \to B(\chi_\alpha)$, where it is easily checked that

$$2\partial^2 \log \det \tilde{\Omega}(\bar{\psi}, \tilde{\chi}_\alpha)/\partial\psi_\alpha \partial\psi'_\alpha \to B(\tilde{\chi}_\alpha) \text{ a.s.}$$

and uniformly in $\tilde{\chi}_\alpha$. Thus (2.6) may be replaced by

$$-b(\tilde{\chi}_\alpha)' B(\tilde{\chi}_\alpha)^{-1} b(\tilde{\chi}_\alpha) + (n-v)\, C(T)/T. \tag{2.10}$$

Put

$$c_{ij}(k) = \frac{1}{T} \sum_{1}^{T-k} \varepsilon_i(t)\, \varepsilon_j(t+k), \qquad C(k) = [c_{ij}(k)]_{i,j=1,\ldots,r}.$$

Then

$$b_u(\tilde{\chi}_\alpha) = \operatorname{tr} \left\{ \sum_{1}^{T-1} C(k)\, \Lambda_u(k) \right\},$$

where $\Lambda_u(k)$ depends on $\tilde{\chi}_\alpha$ and is the typical coefficient in the expansion of $\Omega_0^{-1} k_0^{-1} \Psi_{\alpha,u}(\tilde{\chi}_\alpha)$, and which therefore converges to zero at some geometric rate independent of $\tilde{\chi}_\alpha$. (Note that since $k_\theta(\infty) = I_r$, $\Lambda_u(0) = 0$.) Again as in [12] we may also express $b_u(\tilde{\chi}_\alpha)$ as

$$b_u(\tilde{\chi}_\alpha) = \operatorname{tr} \left\{ \frac{1}{T} \sum_{1}^{T} \varepsilon(t)\, \xi'_u(t) \right\} + O(T^{-1}), \qquad \xi_u(t) = \sum_{1}^{\infty} \Lambda_u(k)' \varepsilon(t-k).$$

Of course $\xi_u(t)$ depends on $\tilde{\chi}_\alpha$. Now the quantities $\xi_u(t)'\varepsilon(t)$ are, for $\tilde{\chi}_\alpha$ fixed, stationary ergodic square integrable martingale differences. The proof of the three sufficiency parts of the Theorem are now essentially the same as for the theorem in [12].

We give some details here to help the reader translate the brief proof in [12] into the present context. Considering the first part of the theorem we observe (see [11]) that the law of the iterated logarithm holds for stationary, ergodic, square integrable martingale differences and that the components of $\varepsilon(t)\, \xi'_u(t)$ are such, for fixed χ_α. We may write

$$b(\chi_\alpha)' B(\chi_\alpha)^{-1} b(\chi_\alpha) = T^{-2} \sum_j b^2_{T,j}(\chi_\alpha),$$

where there are $r(n-v)$ terms in the sum, by reducing B to diagonal form by an orthogonal transformation. Then each $b^2_{T,j}$ is the square of a sum of stationary, square integrable, ergodic martingale differences with unit variance. Thus, for χ_α fixed,

$$\limsup_{T\to\infty}\left\{(2T\log\log T)^{-1}\sum_j b_{T,j}^2(\chi_\alpha)\right\}=1.$$

(See the argument on pp. 1076, 1077 of [12].) It will be sufficient to prove that this convergence is uniform in χ_α to establish the sufficiency part of the first part of the theorem and for this it is sufficient to prove that the expression whose lim sup is being taken is equicontinuous, i.e., continuous in χ_α, uniformly in T, because χ_α varies over a conditionally compact set. It is thus sufficient to prove equicontinuity for $(T/\log\log T)^{1/2}b_u(\chi_\alpha)$ since $B(\chi_\alpha)$ is independent of T and is continuous. To prove this, precisely as in [12], we first show that we may, for T large enough, replace $(T/\log\log T)^{1/2}b_u(\chi_\alpha)$ by

$$(T/\log\log T)^{1/2}\operatorname{tr}\left\{\sum_1^{d(T)} C(j)\,A_u(j)\right\},\qquad d(T)=d\log\log T,\ d<\infty. \tag{2.11}$$

To complete the proof it is only necessary to show that, for each a, b,

$$\limsup_{T\to\infty}\max_{j\leqslant d(T)}|c_{ab}(j)j^{-1}(T/\log\log T)^{1/2}|\leqslant C<\infty, \tag{2.12}$$

where $c_{ab}(j)$ is the typical element of $C(j)$. This is because $\|jA(j)\|$ certainly converges uniformly in χ_α. The proof of (2.12) is indicated in [12], the independence of the $\varepsilon(t)$ being used. The technique is to decompose $Tc_{ab}(j)$ into $(j+1)$ sums $S_T(j,v)$, $v=1,\ldots,j+1$,

$$S_T(j,v)=\sum_m \varepsilon_a\{(j+1)m+v\}\,\varepsilon_b\{(j+1)m+v+j\},\qquad v=1,\ldots,j+1,$$

where m runs over $0, 1,\ldots, [(T-j-v)/(j+1)]$. Each of these is a sum of independent random variables. The proof, following the classic proof of the law of the iterated logarithm, is now completed as at the bottom of p. 1078 of [12]. (It is not sufficiently clearly indicated there that the trancation procedure referred to requires that $\varepsilon_a(t)$ have a moment of order $\gamma>4$, so that $\varepsilon_a(t)\,\varepsilon_b(t+j)$ has a moment of order higher than 2. However, the more delicate transcation method of the proof of the law of the iterated logarithm due to Hartman and Wintner [16] may be used to avoid the higher moment requirement. Since this is of little consequence we omit details here.)

The proof of the sufficiency point of (ii) of the Theorem depends on the inequality

$$\mathscr{E}\{\max_{j+1\leqslant T\leqslant\tau}|Tc_{ab}(j)|^\gamma\}\leqslant C\tau^{\gamma/2}. \tag{2.13}$$

This follows from Doob's inequality followed by Burkolder's inequality (see [11]). Since independence was used only at the end of the proof in (i) we need consider only (2.11) with $d(T)$ as before but with the factor $(T/\log\log T)^{1/2}$ replaced by $(T/\log T)^{1/2}$ and hence it is sufficient to show that

$$\lim_{T\to\infty} (T/\log T)^{1/2} \max_{j\leqslant d(T)} C_{ab}(j) = 0, \text{ a.s.}$$

Using (2.13) this is proved by the method of subsequences as in [12]. (In [12], p. 1079, the right side of (18) should be $CM^{\alpha/2}$; $c(k+j)$ should be replaced by $Nc(k+j)$ two lines down; and $N^{-1/2}$ should be $N^{1/2}$ two lines lower again.)

The proof of the sufficiency part of (iii) is simple and is as in [12].

We turn finally to the necessity parts of the theorem. It will be sufficient to prove (1.9) for $N=\nu+1$, for if (1.9) then holds it will certainly hold for larger N. It will also be sufficient to prove the theorem when optimisation is effected over a smaller set than $\tilde{M}(n)$, for the same reason, and we consider only those parts of $\phi_\alpha(U_\alpha)$, for proper α, whose elements are in echelon form (and satisfy (1.5) of course). There are now r proper α, namely, those for $n_j=\nu_j$, $j\neq k$, $n_k=\nu_k+1$, $k=1,\dots,r$. We shall write $b(k)$, $B(k)$ for the value of b, B for the kth proper α. We must show that, *for any positive constant c*

$$\lim_{\delta\to 0}\lim_{T\to\infty} P\{T\max_k \{b(k)'B(k)^{-1}b(k)\} > c\} = 1. \tag{2.14}$$

Here each $b(k)$ has been itself optimised over the echelon forms in the kth proper coordinate neighbourhood (satisfying (1.5)).

It is now easy to calculate $\partial k_\theta/\partial\psi_\alpha = \partial\{k_\theta - k_0\}/\partial\psi_\alpha = \partial[g^{-1}\{h_\alpha - g_\alpha h_0\}]/\partial\psi_\alpha$ and since $h_\alpha - g_\alpha k_0 = 0$ at $\psi_\alpha = 0$, this has, at $\psi_\alpha = 0$, $(d_\alpha g_0)^{-1}E_{ij}z^u$ as the component corresponding to differentiation with respect to the (i,j)th element of $\psi^{(\alpha)}(u)$, where E_{ij} has zero elements save for a unit in the (i,j)th place. Here u runs over a set of values that certainly includes $0\leqslant u\leqslant n_i - 1$. Note also that $d_\alpha g_0$ is g evaluated at $(0,\psi_\alpha)$ (see below (11)). Then, omitting the factor $-2\{1+o(1)\}$, (2.9) becomes

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\{\Omega_0^{-1}(d_\alpha h_0)^{-1}E_{ij}e^{iu\omega}I_\epsilon(\omega)\}\,d\omega. \tag{2.15}$$

Also it is easily checked that B becomes a matrix with element in the (i,j,u)th and (p,q,v)th place which is

$$2\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\{\Omega_0^{-1}(d_\alpha h_0)^{-1}E_{ij}\,\Omega_0 E_{q,p}(d_\alpha h_0)^{*-1}\,e^{i(u-v)\omega}\}\,d\omega.$$

Now we further restrict ourselves to the case where d_α, for the jth proper index α, has zeros everywhere save in the jth row and units, in the main diagonal, and for the jth row has $(z-a)\,(b_{j1}, b_{j2},..., b_{jr})$. Here $b_{jj}=1$ and for $i \neq j$ b_{ji} is zero if $\nu_j < \nu_i$. If $\nu_j \geqslant \nu_i$ then b_{ji} is the coefficient of the $(\nu_j - 1)$th power of z in the (j, i)th place of g_0. (This is the highest power that, then, occurs.) This ensures that $d_\alpha g_0$, $d_\alpha h_0$ are in echelon form (though, of course, not prime). Now we call $b(j, a)$ the vector b for the jth proper index and this d_α. Now $(1-a^2)B$ has elements that are uniformly bounded since

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1-a^2}{|1-ae^{i\omega}|^2}\,d\omega \equiv 1.$$

Thus it will be sufficient to replace the expression under the probability in (2.14) by

$$\max_k \sup_{-1+\delta \leqslant a \leqslant 1-\delta} T(1-a^2)\,b(k,a)'\,v_k v_k' b(k,a) \tag{2.16}$$

for a set of fixed vectors v_k with $v_k' v_k > 0$. Now choose v_k so that its components are null except at the components $(k, 1, u)$ and so that these components, $v_k(u)$, satisfy

$$\sum_0^{n-1} v_k(u) z^u = h_{0k1}(z), \tag{2.17}$$

where, of course, $h_{0k1}(z)$ is the element in the $(k, 1)$th place in h_0. It is possible to attain (2.17) since $n_k = \nu_k + 1$ so that $n_k - 1 = \nu_k$ and that is the degree of $h_{0k1}(z)$. Now

$$\max_k \sup_a T(1-a^2)(b(k,a)'v_k)^2 \geqslant r^{-2} T \sup_a (1-a^2)\{b(k,a)'v_k\}^2 \tag{2.18}$$

and we may replace (2.16) by the right side of (2.18). However, from (2.15),

$$b(k,a)'v_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{-i\omega}}{1-ae^{-i\omega}} \operatorname{tr}\{h_{0k1}(e^{i\omega})\,E_{k,1} I_\epsilon(\omega)\,\Omega_0^{-1} h_0^{-1}\}\,d\omega \tag{2.19}$$

since for α, corresponding to the kth proper index,

$$d_\alpha^{-1} E_{k,1} = \frac{1}{z-a} E_{k,1}.$$

Now (2.18) is of the same basic form as was studied in [13]. Namely,

$$\sum b(k,a')v_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{-i\omega}}{1-ae^{-i\omega}} \operatorname{tr}\{E_{11} I_\epsilon(\omega)\,\Omega_0^{-1}\}\,d\omega \tag{2.20}$$

since $\sum h_{0k1}(z)E_{k,1}$ is just the first column, h_{01}, of $h_0(z)$ and $h_0^{-1}h_{01}$ has unity in the first place and zeros elsewhere. Now (2.20) is of the form

$$\frac{1}{T}\sum_{1}^{T}\varepsilon_1(t)\sum_{0}^{\infty}a^j\xi(t-1-j),$$

where $\xi(t)$ is the first component in $\varepsilon(t)\Omega_0^{-1}$. Now as in [13] it may be shown, using the methods of [3], that

$$(1-a^2)\,T^{-1/2}\sum_{1}^{T}\varepsilon_1(t)\sum_{0}^{\infty}a^j\xi(t-1-j)$$

converges weakly, as $T\to\infty$, to a Gaussian process on $-1+\delta\leqslant a\leqslant 1-\delta$ with correlation function, between a_1, a_2, that is $(1-a_1^2)^{1/2}(1-a_2^2)^{1/2}/\{1-a_1a_2\}$. The transformation $s=\log\{(1+a)/(1-a)\}$ maps $-1+\delta\leqslant a\leqslant 1-\delta$ onto $-\log\{(2-\delta)/\delta\}\leqslant s\leqslant\log\{(2-\delta)/\delta\}$ and transforms the correlation function into that of a stationary process with spectral density $\{\cosh\pi\omega\}^{-1}$. The theorem then follows from [2] precisely as in [13].

The necessity part constituting the second statement of the theorem is established fairly easily as follows. Consider (2.5) for $n=\nu+1$. It will be sufficient to show that this is negative infinitely often if the condition on $C(T)$ is satisfied. It will therefore be sufficient to show that when $\tilde{\Omega}_n(\tilde{\theta})$ is replaced by a larger quantity. Consider a coordinate neighbourhood in $\tilde{M}(\nu)$ containing k_0. Consider the set S obtained by taking all g, h corresponding to a point in that neighbourhood but with $g_{11}(z)$ monic and of degree one higher than that allowed for a point in the neighbourhood. Replacing $\tilde{\Omega}_n(\tilde{\theta})$ by the optimised value of $\tilde{\Omega}_n(\theta)$ over S then asymptotically (2.5) is of the form $x_T^2+C(T)/T$, where $T^{1/2}x_T$ is a martingale with stationary ergodic martingale differences having variance unity. This follows from standard likelihood ratio arguments, via the results in [15], since there is now no lack of identification and the space allong which χ_α varies is null. It also follows from the expression for b_u given below (2.10), remembering that now b_u is not a function of χ_α. Thus the necessity part of (i) follows from the law of the iterated logarithm.

References

[1] Akaike, H., (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–247.

[2] Berman, S., (1971). Asymptotic independence of the number of high and low level crossings of stationary Gaussian processes. *Ann. Math. Statist.* **42**, 927–945.

[3] Billingsley, P., (1968). *Convergence of Probability Measures*, Wiley, New York.

[4] Clark, J. M. C., (1976). The consistent selection of parameterizations in system identification. Paper presented at JACC, Purdue University.

[5] DENHAM, M. J., (1974). Canonical forms for the identification of multivariable linear systems. *IEEE Trans. Automat. Control* **AC-19**, 646–656.

[6] DEISTLER, M., DUNSMUIR, W., AND HANNAN, E. J., (1978). Vector linear time series models: Corrections and extensions. *Adv. in Appl. Probab.* **10** , 360–372.

[7] DEISTLER, M., AND HANNAN, E. J., (1981). Some properties of the parameterization of ARMA systems with unknown order. *J. Multivariate Anal.* **11**, 474–484.

[8] DUNSMUIR, W., AND HANNAN, E. J., (1976). Vector linear time series models. *Adv. in Appl. Probab.* **8**, 339–364.

[9] FINE, T. L., AND HWANG, W. G., (1979). Consistent estimation of system order. *IEEE Trans. Automat. Control* **AC-24**, 387–402.

[10] HAJDASINSKI, A. K., AND NIEDERLINSKI, A. (1979). Multivariable system identification. In *Identification and System Parameter Estimation. Fifth IFAC Symposium* (R. Isermann, Ed.), pp. 43–76. Pergamon, Elmsford, N. Y.

[11] HALL, P., AND HEYDE, C. C. (1980). *Martingale Limit Theory and its Application.* Academic Press, New York.

[12] HANNAN, E. J., (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071–1081.

[13] HANNAN, E. J., (1982). Testing for autocorrelation and Akaike's criterion. In *Essays in Statistical Science* (*The Moran Festschrift*) (J. M. Gani and E. J. Hannan, Eds.). Applied Probability Trust, Sheffield.

[14] HANNAN, E. J., AND ROBINSON, P. M., (1973). Lagged regression with unknown lags. *J. Roy. Statist. Soc.* **35**, 252–267.

[15] HANNAN, E. J., DUNSMUIR, W., AND DEISTLER, M. (1980). Estimation of vector ARMAX models. *J. Multivariate Anal.* **10**, 275–295.

[16] HARTMAN P., AND WINTNER, A. (1941). On the law of the iterated logarithm. *Amer. J. Math.* **63**, 169–176.

[17] RISSANEN, J., (1979). Shortest data description and consistency of order estimates in ARMA processes. *International Symposium on Systems Optimization and Analysis* (A. Bensoussan and J. L. Lions, Ed.), pp. 92–98.

[18] ROSENBROCK, H. H., (1970). *State Space and Multivariate Theory.* Wiley, New York.

[19] SHIBATA, R., (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.