Automatica, Vol. 14, pp. 615–622
Pergamon Press Ltd. 1978. Printed in Great Britain

Brief Paper

Model Approximations Via Prediction Error Identification*†

B. D. O. ANDERSON,‡ J. B. MOORE‡ and R. M. HAWKES§

Key Word Index—Identification; modelling; approximation theory; system order reduction; least squares approximation.

Summary—Identification is considered of a dynamic system by a model in a model set of which the system is not a member. This is achieved by defining a performance index related to prediction error performance indices, and taking that model minimizing the performance index as that which is closest to the system. The index has an intuitively pleasing spectral interpretation in the stationary case for large measurement intervals. The length of measurement interval needed for identification is discussed by studying the limiting behaviour of the performance indices, as is also the relation of the index to the Kullback information measure. The communication theoretic issue of convergence of *a posteriori* densities when Bayesian estimation is being undertaken with a finite model set is examined.

1. *Introduction*

IN THIS paper, we address the following kind of problem which arises in connection with modelling physical systems. There is given a dynamic system $\mathscr{S}$ and measurements of some of the variables associated with $\mathscr{S}$. We have to 'identify' $\mathscr{S}$, and for this purpose there is given a collection of models, such that either $\mathscr{S}$ lies within the model set, or else there is a particular model within the model set which in some way is closest to $\mathscr{S}$. By using the available measurements, we have to decide which model in the model set is identical with or is closest to $\mathscr{S}$.

In [1], Liporace gives convergence results for *Bayesian estimation* of identically and independently distributed discrete-time processes for the cases when the system $\mathscr{S}$ (nondynamic) is contained in the finite model set (also nondynamic) and when it is not. Building on these results Moore and Hawkes first showed in [2] consistency results and performance bounds for the case when the system and models are dynamic but with a linear Gaussian restriction and $\mathscr{S}$ belongs to the model set. Also Hawkes and Moore in [3] give continuous-time versions of [1] for nonlinear signal models with a case study application giving insight into the use of the Kullback information function for studying situations when $\mathscr{S}$ does not belong to the model set, and in [4] develop a theory using the Kullback information function for the case of linear gaussian signal models when $\mathscr{S}$ does not belong to the model set. More recently Baram in [5] has introduced a related information function which gives the same results as the Kullback information measure but has the properties of a true metric.

For the case when $\mathscr{S}$ is in the model set, consistency results for autoregressive moving average models and *maximum likelihood* (ML) estimation are studied by Caines and Rissanen in [6]. Of related interest are the results on Tse in [7], and also Caines in [8] where the submartingale property of sequences of maximized ML ratios on finite parameter sets is used to prove the consistency of ML estimates on such sets for a general class of observation sequences, satisfying a certain probabilistic condition.

For the case when $\mathscr{S}$ is not necessarily in the model set, Ljung[9, 10] has given strong consistency results for *prediction error* schemes and very general process models such as when feedback is allowed, nonlinearities are included and the signals are nonstationary. Results for the stationary case are given by Caines in [11] using [6, 9]. In the Gaussian case, of course prediction error schemes lead to the Bayesian and maximum likelihood case. In the non-Gaussian case they are more readily implemented since the second order statistics alone are used for identification which may or may not give useful identification.

In this paper, we work with many of the same ideas of our earlier studies[2–4], but in the prediction error context of [9–11]. The work on [9–11] is extended by examining more closely convergence rates, by setting up a useful distance measure for the case when the system does not belong to the model set, relating this measure to the Kullback information measure, and by giving spectral interpretations which give further insights into the behaviour of the prediction error algorithms. A number of the results of [4], with tedious derivations as yet unpublished, are achieved here more directly than in [4] as a special case of the general prediction error identification theory.

2. *Systems, models and identification criteria*

This section is definitional in character. Consider a causal discrete time stochastic system $\mathscr{S}$ with 'known' output sequence $\{y_i\}$ and 'known' input sequence $\{u_i\}$ and such that a one-step ahead prediction estimate $\hat{y}_{t|\mathscr{S}}$ is available given $y_1 \ldots y_{t-1}$ and $u_1 \ldots u_{t-1}$ and *a priori* data x_0. We write

$$y_t = \hat{y}_{t|\mathscr{S}} + e_{t|\mathscr{S}}. \tag{2.1}$$

Consider a collection $\mathscr{M} = \{\mathscr{M}_\theta\}$ of stochastic models, each defined by a parameter vector θ, and to keep the theoretical development as simple as possible assume that θ belongs to a finite set $\{\theta_1, \ldots, \theta_p\}$. The parameter vector θ may be involved in determining both the probability densities governing the stochastic signals driving $\mathscr{M}_\theta$, as well as the equation set defining how the output signals arise from the known and unknown input signal. For each model, we assume that it is possible to construct an innovations representation

$$y_t = \hat{y}_{t|\theta} + e_{t|\theta}. \tag{2.2}$$

Here $\hat{y}_{t|\theta}$ is defined in the same manner as $\hat{y}_{t|\mathscr{S}}$, save that the $\{y_t\}$ process is assumed to result from the model, rather than the system. We further assume that there is a known function $\mathscr{P}_\theta$ for the conditional one step ahead predictor such that

$$\hat{y}_{t|\theta} = \mathscr{P}_\theta(t, y_{t-1}, y_{t-2}, \ldots, y_1, u_t, \ldots, u_1, x_0). \tag{2.3}$$

*Received March 21 1977; revised December 12 1977; revised 8 May 1978. The original version of this paper was presented at the 7th IFAC World Congress on A Link between Science and Applications of Automatic Control which was held in Helsinki, Finland during June 1978. The published Proceedings of this IFAC Meeting may be ordered from: Pergamon Press Limited, Headington Hill Hall, Oxford, OX3 0BW, England. This paper was recommended for publication in revised form by associate editor K. J. Åström.

†Work supported by the Australian Research Grants Committee.

‡Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia.

§Weapons Research Establishment, Adelaide, South Australia, 5001, Australia.

Now note that even if the $\{y_t\}$ process is not generated by the model, one can still form the right hand side of (2.3), and in this sense formally obtain $\hat{y}_{t|\theta}$ and $e_{t|\theta}$. These quantities will however no longer have the same statistics as if they were generated by model.

Mostly, we conceive of the models as being derived from finite-dimensional linear systems excited by white noise. The models may possibly be nonstationary; ARMA processes are naturally included in the discussion. Techniques for obtaining representations of the form of (2.2), including the Kalman filter, are then well known when (i) the noise is Gaussian and $\hat{y}_{t|\mathscr{S}}$ is the conditional mean estimate (ii) the noise is non-Gaussian and $\hat{y}_{t|\mathscr{S}}$ is the best linear conditional minimum variance estimate (iii) $\hat{y}_{t|\mathscr{S}}$ is the standard least squares estimate when it is assumed that

$$y_t = \theta' x_{t-1} + e_t$$

for x_{t-1} measurable.

Suppose now that the sequence $\{\tilde{y}_{t|\theta}\}$, where $\tilde{y}_t = y_t - \hat{y}_{t|\theta}$, is *conditionally white and Gaussian* with covariance $P_{t|\theta}$ and of course $\hat{y}_{t|\theta}$ a conditional mean estimate. Then the log likelihood function given $y_1, \ldots, y_N$ would be, to within a constant and a sign change,

$$V_N(\theta) = \sum_{t=1}^{N} (\log \det P_{t|\theta} + \tilde{y}'_{t|\theta} P_{t|\theta}^{-1} \tilde{y}_{t|\theta}). \tag{2.4}$$

One might then define that model which was the best approximation to $\mathscr{S}$ as that obtained by choosing the value of θ, call it $\bar{\theta}(N)$, minimizing (2.4). The applicability of this prediction error index can be extended in a number of ways.

We could conceive as in [10, 11] of minimizing (2.4) even with $\tilde{y}_{t|\theta}$ not conditionally Gaussian including the case when $\hat{y}_{t|\theta}$ is a best linear estimate rather than a conditional mean estimate. The index still has some intuitive content, especially if $P_{t|\theta}$ is either $E\{\tilde{y}_{t|\theta}\tilde{y}'_{t|\theta}\}$ or

$$E\{\tilde{y}_{t|\theta}\tilde{y}'_{t|\theta} \mid y_{t-1}, \ldots, y_1, u_y, \ldots, u_1, x_0\}$$

One may abstract $P_{t|\theta}$ from a covariance interpretation altogether as in Ljung[9].

The right side of (2.4) can be rewritten as

$$\sum_{t=1}^{N} \{\log \det P_{t|\theta} + \tilde{y}'_{t|\mathscr{S}} P_{t|\theta}^{-1} \tilde{y}_{t|\mathscr{S}} + (\hat{y}_{t|\mathscr{S}} - \hat{y}_{t|\theta})' P_{t|\theta}^{-1} (\hat{y}_{t|\mathscr{S}} - \hat{y}_{t|\theta}) + 2\tilde{y}'_{t|\mathscr{S}} P_{t|\theta}^{-1} (\hat{y}_{t|\mathscr{S}} - \hat{y}_{t|\theta})\}.$$

In the next section it is shown that in some situations the last term contributes in a negligible way to the sum as compared to the earlier terms, so that the third term in this index weights the error between using the system $\mathscr{S}$ and the model $\mathscr{M}_\theta$ as the basis for a prediction design.

3. *Limiting behaviour of the identification index*

In order to obtain results on the limiting behaviour of the index used for identification, we shall impose, as is normal, e.g. [9], stability constraints on $\mathscr{S}$ and $\mathscr{P}_\theta$. These stability constraints achieve three distinct goals: they ensure that some moments are bounded functions of time, they ensure that there is some mixing, or forgetting of the past, so that the covariance of the values of a stochastic process at two different instants of time dies away exponentially fast with increase in the interval between the two instants, and they eliminate deterministic processes, or processes which might be viewed as having this tendency, such as where there is perfect prediction, or prediction performance approaching perfect prediction. Specifically, we shall adopt the following assumption.

Assumption A: For all t, τ, some $\alpha > 0$ and $0 < \beta < 1$,

$$\operatorname{cov}[\tilde{y}'_{t|\theta} P_{t|\theta}^{-1} \tilde{y}_{t|\theta}, \tilde{y}'_{\tau|\theta} P_{\tau|\theta}^{-1} \tilde{y}_{\tau|\theta}] \leq \alpha \beta^{|t-\tau|}.$$

When

$$P_{t|\theta} = E[\tilde{y}_{t|\theta} \tilde{y}'_{t|\theta}],$$

all that is required is that

$$\|P_{t|\theta}^{-1}\| < \bar{\alpha}$$

for some $\bar{\alpha}$ and

$$\operatorname{cov}[\tilde{y}'_{t|\theta} \tilde{y}_{t|\theta}, \tilde{y}'_{\tau|\theta} \tilde{y}_{\tau|\theta}] \leq \alpha \beta^{|t-\tau|}.$$

Relaxed versions will be noted later. One obvious and important situation when the exponential bound on the covariance will hold is when $\mathscr{S}$ and $\mathscr{P}_\theta$ are linear systems with impulse responses $w(\cdot, \cdot)$ satisfying

$$\|w(t, \tau)\| \leq \alpha \beta^{(t-\tau)} 1(t-\tau),$$

and with no hidden (uncontrollable or unobservable) modes, and the input to $\mathscr{S}$ comprises a sequence of independent random variables with bounded fourth moment. Note that for these conditions to hold, it is not strictly necessary that either $\mathscr{M}_\theta$ or the predictor associated with $\mathscr{S}$ have a stability property.

We now strengthen a result which is obvious in the ergodic situation. For the proof, see the Appendix.

Lemma 3.1: Suppose $V_N(\theta)$ is as in (2.4), with $P_{t|\theta}$ deterministic. With Assumption A in force,

$$\lim_{N \to \infty} N^{\gamma/2} \left\{ \frac{1}{N} V_N(\theta) - E\left[\frac{1}{N} V_N(\theta)\right] \right\} = 0 \quad \text{w.p.1} \tag{3.1}$$

for all $\gamma \in [0, 1)$ and

$$\Pr\left[\frac{1}{N} V_N(\theta) - \frac{1}{N} E[V_N(\theta)] \mid > \varepsilon \right] < \frac{2\alpha}{\varepsilon^2 N (1-\beta)}. \tag{3.2}$$

Remarks 1. The proof of the above lemma makes use of a discrete time version of an ergodic theorem in [Chapter 5.5, 12], applying it to the special quantity $N^{-1} V_N(\theta)$, and allowing a generalization to the case $\gamma \neq 0$. The result (3.1) for the case when $\gamma = 0$ also appears in [9], as do the methods which allow extension of this lemma to infinite model sets if desired. Note too the important fact that $V_N(\theta)$ can be used to generate strongly consistent estimators in the stationary case is established in [11]. Finally, the theory of [12] allows a more relaxed version of Assumption A but there seems no point in exploring this fact here.

2. Inequality (3.2) is helpful in defining an appropriate length of an identification experiment, since from (3.2) one can find the value of N which guarantees that the error between $(1/N)V_N$ and $(1/N)E[V_N]$ will be less than a prescribed quantity in a given proportion of experiments. This seems the best type of result which can be obtained in the circumstances. We could proceed as follows. Suppose for simplicity that $(1/N)E(V_N)$ is constant for all N as would be the case under stationarity assumptions. With θ^* minimizing

$$\lim_{N \to \infty} \frac{1}{N} E[V_N(\theta)]$$

over a finite set and assumed here to be unique, choose

$$\bar{\varepsilon} = \tfrac{1}{2} \min_{\theta_i \neq \theta^*} \left\{ E\left[\frac{V_N(\theta_i)}{N}\right] - E\left[\frac{V_N(\theta^*)}{N}\right] \right\}.$$

Then if

$$|(1/N)V_N(\theta_i)-(1/N)E[V_N(\theta_i)]|<\bar{\varepsilon}$$

for all θ_i including θ^*, it follows that $\theta(N)$ minimizes θ^*. Consequently if the experiment length N is such that

$$\frac{2\alpha}{\bar{\varepsilon}^2 N(1-\beta)}<1-\rho^{1/p}$$

where p is the number of distinct θ_i, then the probability that $\theta(N)=\theta^*$ is at least ρ.

The above result has implications for the problem of defining a model set to achieve effective identification of a system. One certainly wants the model set to be such that for some θ^* in the set, $E[V_N(\theta^*)/N]$ is as small as possible, i.e. the model is as little different from the system as possible. One also *apparently* wants $\bar{\varepsilon}$ as large as possible. To ensure fulfillment of these conditions, one would need to know in advance what the system is. To the extent that the system might be roughly known, the first condition can be fulfilled, while the need to fulfill the second is somewhat illusory, as we now argue. If $E[V_N(\theta_i)/N]$ and $E[V_N(\theta_j)/N]$ are close for some i, j, it will require more measurements to decide whether $\mathcal{M}_{\theta_i}$ or $\mathcal{M}_{\theta_j}$ is closer to the system, but at the same time the difference in distances of $\mathcal{M}_{\theta_i}$ from the system and $\mathcal{M}_{\theta_j}$ from the system is not large, so that if a large number of measurements are not used, and one of $\mathcal{M}_{\theta_i}$ and $\mathcal{M}_{\theta_j}$ is wrongly identified, the error will not be great.

Predictor matching. In the remainder of this section, we develop two implications of Lemma 3.1. We need a slightly strengthened version of Assumption A which requires that the cascade of the system $\mathcal{S}$ and the true predictor $\mathcal{P}_{\mathcal{S}}$ have an exponential stability property.

Strengthened Assumption A: Assumption A holds for all θ in the model set, and with θ replaced by $\mathcal{S}$.

Next we note a preliminary lemma which will have applications not merely in this section, but in the rest of the paper. The result is equivalent to one proved in [13].

Lemma 3.2. Let A, B be positive definite symmetric matrices. Then

$$\operatorname{tr} AB^{-1}-\ln\det AB^{-1}-\operatorname{tr} I\geqq 0$$

with equality if and only if $A=B$.

Evidently, the quantity $\operatorname{tr} AB^{-1}-\ln\det AB^{-1}-\operatorname{tr} I$ provides a measure of the error in approximating B by A; if $\|AB^{-1}-I\|$ is small, one can show that

$$\operatorname{tr} AB^{-1}-\ln\det AB^{-1}-\operatorname{tr} I\doteq\tfrac{1}{2}\operatorname{tr}\{[(A-B)B^{-1}]^2\}.$$

Consider now the following index, which, it should be noted, is *not* likely to be computed in practice prior to identification, since it involves the quantity $\hat{y}_{t|\mathcal{S}}$.

$$W_N(\theta)=\sum_1^N[\operatorname{tr} P_{t|\mathcal{S}}P_{t|\theta}^{-1}-\ln\det P_{t|\mathcal{S}}P_{t|\theta}^{-1}-\operatorname{tr} I+(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})'P_{t|\theta}^{-1}(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})]. \tag{3.3}$$

Such indices arise in estimating the mean and covariance of a Gaussian random variable, given independent identically distributed observations, see e.g. [14]. Of course, here there is no assumption of normality. The index reflects two types of errors. From Lemma 3.2 we see that it reflects the error between $P_{t|\mathcal{S}}$ and $P_{t|\theta}$, or the error covariance associated with use of the correct predictor $\mathcal{P}_{\mathcal{S}}$ and the error covariance associated with use of the incorrect predictor $\mathcal{P}_\theta$. More obviously the index also reflects the error between $\hat{y}_{t|\mathcal{S}}$ and $\hat{y}_{t|\theta}$. The index obviously then has intuitive content. The substance of the next lemma is that the index for large N is like $V_N(\theta)$.

Lemma 3.3. With notation as above,

$$\frac{1}{N}E[V_N(\theta)-W_N(\theta)]=\frac{1}{N}\sum_{t=1}^N\ln\det P_{t|\mathcal{S}}+\operatorname{tr} I \tag{3.4}$$

and with the strengthened Assumption A in force

$$\lim_{N\to\infty}\frac{1}{N}\{W_N(\theta)-E[W_N(\theta)]\}=\lim_{N\to\infty}\frac{1}{N}\{V_N(\theta)-E[V_N(\theta)]\}=0. \tag{3.5}$$

Proof:

$$\tilde{y}_{t|\theta}'P_{t|\theta}^{-1}\tilde{y}_{t|\theta}=\tilde{y}_{t|\mathcal{S}}'P_{t|\theta}^{-1}\tilde{y}_{t|\mathcal{S}}+2\tilde{y}_{t|\mathcal{S}}'P_{t|\theta}^{-1}(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})+(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})'P_{t|\theta}^{-1}(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})$$

$$E[\tilde{y}_{t|\mathcal{S}}'P_{t|\theta}^{-1}(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})]=E\{E[\tilde{y}_{t|\mathcal{S}}'P_{t|\theta}^{-1}(\hat{y}_{t|\mathcal{S}}-\hat{y}_{t|\theta})\,|\,y_{t-1}\cdots y_1,u_t\cdots u_1,x_0]\}=0$$

Also,

$$E[\tilde{y}_{t|\mathcal{S}}'P_{t|\theta}^{-1}\tilde{y}_{t|\mathcal{S}}]=\operatorname{tr} P_{t|\mathcal{S}}P_{t|\theta}^{-1}$$

and (3.4) follows by direct substitution. An argument as in the proof of Lemma 3.1 yields (3.5) and the proof is then trivial.

As an immediate corollary, we can obtain a consistency result which appears in one form or another in [2, 10, 9].

Corollary. Under the assumptions of Lemma 3.3, and with the assumption that $\mathcal{S}$ is contained in the model set which is finite, $\mathcal{S}=\mathcal{M}_{\theta^*}$ where θ^* minimizes

$$\lim_{N\to\infty}\frac{1}{N}E[V_N(\theta)].$$

Proof. From (3.3), we see that if $\mathcal{S}=\mathcal{M}_{\theta_i}$ then $W_N(\theta_i)=0$ while for $\theta_j\neq\theta_i$, (3.3) and Lemma 3.2 show that $W_N(\theta_n)\geqq 0$. Combining this with Lemmas 3.1 and 3.3 yields the desired result.

We close this subsection with three further comments.

First, if two models, or a system and a model, differ only in respect of initial conditions, then a modification of Assumption A or its strengthened version (to ensure the exponential decay of the effect of initial conditions) will cause these conditions to be forgotten in the computation of $N^{-1}E[V_N(\theta)]$.

Second, though we have used the fact that $P_{t|\mathcal{S}}$ is the covariance of $\tilde{y}_{t|\mathcal{S}}$, we have not in *proving* Lemma 3.3 used the corresponding interpretation of $P_{t|\theta}$, (though the intuitive content of $W_N(\theta)$ at least in part depends on this interpretation). Thus $P_{t|\theta}$ could be replaced in most of the above by say, I. However, in the next subsection, the conditional error covariance property of $P_{t|\theta}$ will be used.

Kullback information measure connection. Here we further motivate the use of an index such as $V_N(\theta)$ in identification.

As already observed, in case $\tilde{y}_{t|\theta}$ is conditionally white and Gaussian, $V_N(\theta)$ is $-\ln p(y_1,\ldots,y_N|\theta)$. Consequently, [14], the task of minimising $E[V_N(\theta)]$ is then one of finding the model which is closest to the system in the sense of minimising the Kullback information measure:

$$J_N[p_s;p_\theta)=E\left[\ln\frac{p(y_1,\ldots,y_N|\mathcal{S})}{p(y_1,\ldots,y_N|\theta)}\right].$$

(This quantity is nonnegative and equal to zero if and only if

$$p(y_1,\ldots,y_N|\mathcal{S})=p(y_1,\ldots,y_N|\theta)$$

almost everywhere).

When $\mathscr{S}$ has the property that $p(y_1,\ldots,y_N)$ is also Gaussian, it is easy [14] to get a precise expression for $J_N(\mathscr{S},\theta)$:

$$J_N(\mathscr{S},\theta)=\sum_1^N[\ln P_{t|\theta}P_{t|\mathscr{S}}^{-1}+E[\tilde{y}'_{t|\theta}P_{t|\theta}^{-1}\tilde{y}_{t|\theta}]-\operatorname{tr} I]. \quad (3.6)$$

Of course, $J_N(\mathscr{S},\theta)\geqq 0$. In fact, one can check that $J_N(\mathscr{S},\theta)=E[W_N(\theta)]$, with $W_N(\theta)$ defined as in (3.3).

4. *Specialised properties of the limiting identification index*

We shall consider a specialization of the material of Section 3 to the stationary case.

Spectrum interpretations. Suppose that the system $\mathscr{S}$ is obtained by driving a linear, time-invariant, minimum phase system by a sequence of independent, identically distributed Gaussian random variables. Let us further assume that the transfer function matrix, $W_{\mathscr{S}}(z)$, has $W_{\mathscr{S}}(\infty)=I$, and the driving sequence has zero mean, bounded 4th order moments, and covariance $P_{\mathscr{S}}$. The covariance of $y_t-\hat{y}_{t|\mathscr{S}}$ is also $P_{\mathscr{S}}$. The spectrum of $\{y_t\}$ is $W_{\mathscr{S}}(z)P_{\mathscr{S}}W'_{\mathscr{S}}(z^{-1})$.

Suppose also that the model $\mathscr{M}_\theta$ is obtained in the same way, mutatis mutandis. For example, the conditional covariance of the input noise signal is P_θ. Suppose further that $W_\theta(z)$ is free of zeros on $|z|=1$, so that via a system with transfer function matrix $W_\theta^{-1}(z)$, driven by $\{y_t\}$, the stationary sequence $\{y_t-\hat{y}_{t|\theta}\}$ may be obtained. With $\{y_t\}$ actually the system output, the actual spectrum of $y_t-\hat{y}_{t|\theta}$ is evidently

$$[W_\theta^{-1}(z)W_{\mathscr{S}}(z)]P_{\mathscr{S}}[W_\theta^{-1}(z^{-1})W_{\mathscr{S}}(z^{-1})]'.$$

Further, with integrations anticlockwise round the unit circle,

$$E[(y_t-\hat{y}_{t|\theta})'P_{t|\theta}^{-1}(y_t-\hat{y}_{t|\theta})]$$

$$=\frac{1}{2\pi j}\oint\operatorname{tr}[W_\theta^{-1}(z)W_{\mathscr{S}}(z)P_{\mathscr{S}}[W_\theta^{-1}(z^{-1})W_{\mathscr{S}}(z^{-1})]'P_\theta^{-1}z^{-1}\,dz$$

$$=\frac{1}{2\pi j}\oint\operatorname{tr}[\Phi_{\mathscr{S}}(z)\Phi_\theta^{-1}(z)]z^{-1}\,dz \quad (4.1)$$

where

$$\Phi_{\mathscr{S}}(z)=W_{\mathscr{S}}(z)P_{\mathscr{S}}W'(z^{-1})$$

is the spectrum of $\{y_t\}$ and

$$\Phi_\theta(z)=W_\theta(z)P_\theta W'_\theta(z^{-1})$$

is the spectrum of $\{y_t\}$ conditioned on $\mathscr{M}_\theta=\mathscr{S}$.

The assumptions on the system and model ensure that Assumption A holds. Then a comparison of (2.4) with the above identity and use of Lemma 3.1 shows that in identification, we are seeking that θ which minimizes

$$\ln\det P_\theta+\frac{1}{2\pi j}\oint\operatorname{tr}[\Phi_{\mathscr{S}}(z)\Phi_\theta^{-1}(z)]z^{-1}\,dz$$

or equivalently

$$\ln\det P_\theta P_{\mathscr{S}}^{-1}+\frac{1}{2\pi j}\oint\operatorname{tr}[\Phi_{\mathscr{S}}(z)\Phi_{\mathscr{S}}(z)]z^{-1}\,dz.$$

Now we use an identity that is proved easily by complex variable methods, at least when $\Phi_\theta(z)$ and $\Phi_{\mathscr{S}}(z)$ are rational, and which is closely linked to the Kolmogorov–Weiner formula for the prediction error in terms of the spectrum, see e.g. [1] pp. 71–76.

$$\ln\det P_\theta P_{\mathscr{S}}^{-1}=\frac{1}{2\pi j}\oint\ln\det[\Phi_\theta(z)\Phi_{\mathscr{S}}(z)]z^{-1}\,dz. \quad (4.2)$$

From the identity, we obtain the following result:

Lemma 5.1. With the specializations of systems and models as described above and with a finite model set, the θ which minimizes $(1/N)V_N^*(\theta)$ as $N\to\infty$ is also the θ which minimizes

$$\bar{J}(S,\theta)=\frac{1}{4\pi j}\oint[\operatorname{tr}\Phi_{\mathscr{S}}(z)\Phi_\theta^{-1}(z)-\ln\det\Phi_{\mathscr{S}}(z)\Phi_\theta^{-1}(z)-\operatorname{tr} I]z^{-1}\,dz. \quad (4.3)$$

Of course, the integrand is nonnegative for all $z=e^{j\omega}$, ω real and positive, unless $\Phi_{\mathscr{S}}(z)=\Phi_\theta(z)$. It clearly shows that *identification using* (2.4) *is equivalent to picking a model with spectrum closest to that of the system, with* $\bar{J}(S,\theta)$ *measuring the difference.*

It is perhaps of interest to note that if $\|\Phi_{\mathscr{S}}(z)-\Phi_\theta(z)\|$ is small with respect to $\Phi_\theta(z)$, then (4.2) can be approximated by

$$\bar{J}(S,\theta)\doteq\frac{1}{2\pi j}\oint\tfrac{1}{2}\{\operatorname{tr}[(\Delta\Phi)\Phi_\theta^{-1}(z)]^2\}z^{-1}\,dz$$

(where $\Delta\Phi=\Phi_{\mathscr{S}}-\Phi_\theta$), a fact which shows that it is the percentage error in the spectra rather than the absolute error which is important.

In the Gaussian case, (4.3) does have an additional interpretation which we now develop in terms of the Kullback information function.

From (4.1) and (3.7), we see that

$$\ln\det P_\theta P_{\mathscr{S}}^{-1}+\frac{1}{2\pi j}\oint\operatorname{tr}[\Phi_{\mathscr{S}}(z)\Phi_\theta^{-1}(z)]z^{-1}\,dz$$

is

$$\frac{2J_N(S,\theta)}{N}+\operatorname{tr} I,$$

and thus using (4.2) and (4.3), we have

$$\bar{J}(\mathscr{S},\theta)=\frac{J_N(\mathscr{S},\theta)}{N}.$$

In the event that there are initial conditions transients so that, for example, (4.1) is actually valid only for large t, we have instead

$$\bar{J}(S,\theta)=\lim_{N\to\infty}\frac{J_N(S,\theta)}{N}.$$

Thus $\bar{J}(S,\theta)$ may be regarded as an asymptotic per sample Kullback information function. Nevertheless, we reiterate that, even though this significance is lost in the non-Gaussian case, the θ minimizing $\bar{J}(\mathscr{S},\theta)$ still identifies the model closest to the system.

One advantage that the index $\bar{J}(\mathscr{S},\theta)$ has over those used in, say, [9] is that in a quite trivial way, it allows us to distinguish models with the same spectral shape and different spectral magnitude, i.e. the spectra $\Phi_\theta(z)$ and $2\Phi_\theta(z)$ lead to different values of $\bar{J}$.

Example. Consider models with unit intensity white noise input to a transfer function

$$V_\theta(z)=\frac{z^2+\beta_1 z+\beta_2}{z^2+\alpha_1 z+\alpha_2}.$$

The parameter vector $\theta=[\beta_1,\beta_2,\alpha_1,\alpha_2]$.

(a) Suppose that the true system is defined by the parameter vector $[\bar{\beta}_1, 0, 0, 0]$. Let $[\beta_1, 0, 0, 0]$ define a model. Then one can compute

$$\bar{J}(\mathscr{S},\theta)=\tfrac{1}{2}\frac{(\bar{\beta}_1-\beta_1)^2}{(1-\beta_1)}.$$

For β_1 taking the values 0.5 and 0.9, this is plotted as a function of $\bar{\beta}_1$ in Fig. 1.

(b) Suppose the true system is defined by the parameter vector $[0, 0, \bar{\alpha}_1, 0]$ and the model is defined by $[0, 0, \alpha_1, 0]$. Then

$$\bar{J}(\mathscr{S},\theta)=\tfrac{1}{2}\frac{(\alpha_1-\bar{\alpha}_1)^2}{(1-\bar{\alpha}_1)^2}.$$

For α_1 taking the values 0.5 and -0.9, this is plotted as a function of $\bar{\alpha}_1$ in Fig. 2.

(c) Suppose the true system is defined by the parameter vector $[0, 0, \overline{\alpha_1}, 0]$ and the model is defined by $[\beta_1, \beta_2, 0, 0]$. (Thus a second order MA model is being used to approximate a first order AR system). Then

$$\bar{J}(\mathscr{S},\theta)=\tfrac{1}{2}\frac{(1+\beta_2)(1-\bar{\alpha}_1^2\beta_2)+\beta_1\bar{\alpha}_1(1-\beta_2)}{(1-\beta_1\bar{\alpha}_1+\beta_2\bar{\alpha}_1^2)(1-\bar{\alpha}_1^2)}\times(1-\beta_2)[(1+\beta_2)^2-\beta_1^2].$$

For $\beta_1=0.3$, $\beta_2=0.9$, this is plotted against $\overline{\alpha_1}$ in Fig. 3.

Figures 1 and 2 illustrate the fact that an error in zero or

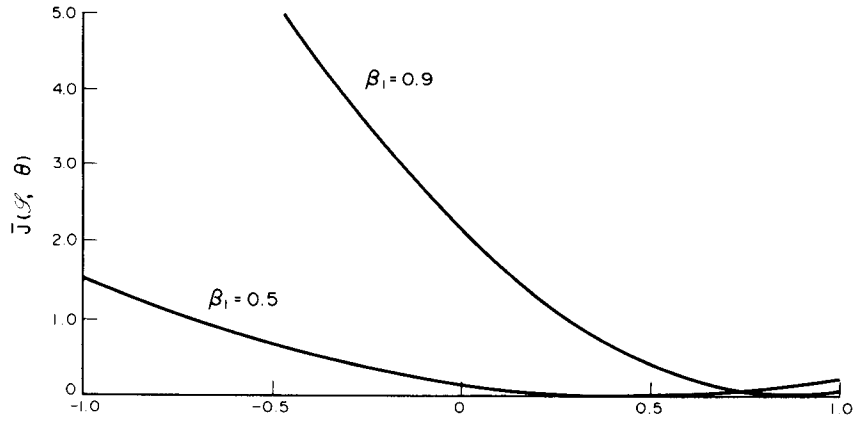

Fig. 1

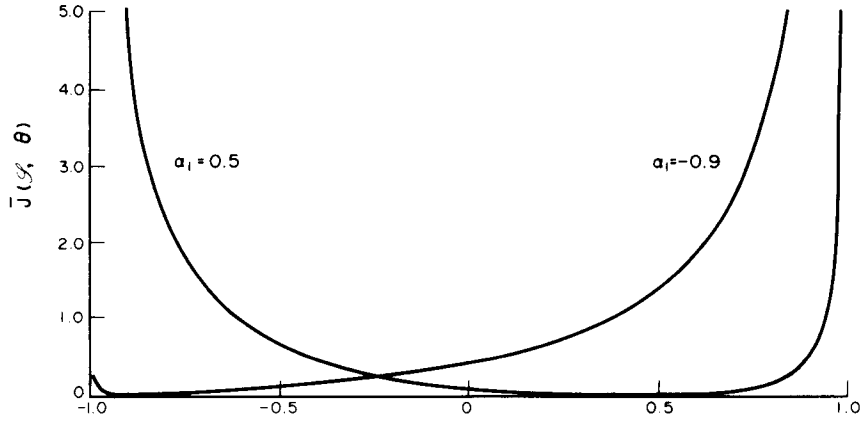

Fig. 2

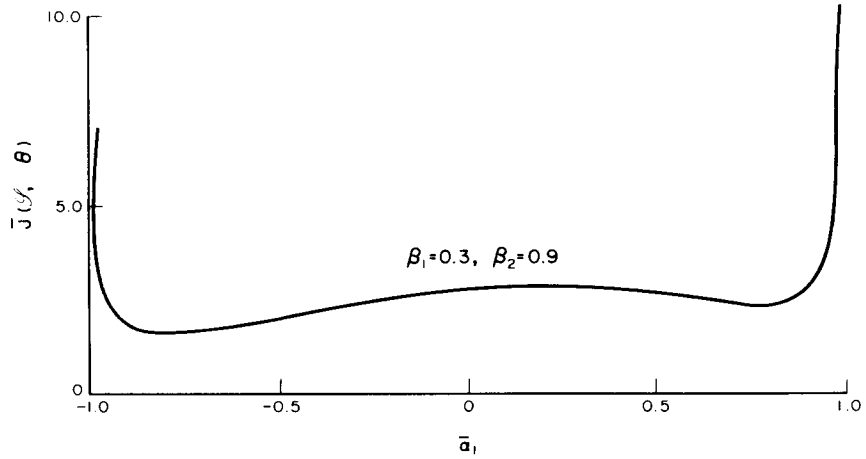

Fig. 3

pole positions will be more critical the closer the pole or zero is to the unit circle. Figure 3 illustrates there is a substantial error in approximating a first order AR system by a second order MA model: the figure also illustrates that the function $\bar{J}(\mathscr{S},\theta)$ with the system parameters variable and model parameters fixed can have more than one local minimum. In the converse situation (with system parameters fixed, the model parameters variable), the same phenomenon is to be expected. This shows the potential difficulty in carrying out a search over a continuum of parameter values to find a parameter value minimizing $\bar{J}$.

Let us mention two other applications of the formula (4.3). The first is to the low order modelling of high order systems. For formula (4.3) offers one way of evaluating of the error involved in treating a high order system, with output spectrum $\Phi_{\mathscr{S}}(z)$, as a lower order object, with output spectrum $\Phi_\theta(z)$. It has long been accepted that it is often good engineering practice in identifying a high order system to act as if it has lower dimension: that an approximation was involved has not been disputed, but it has not often been pointed out what precisely that approximation is.

The second application is to fault detection. Suppose that the system $\mathscr{S}$ is prone to a number of faults: take the model set to comprise the system $\mathscr{S}$, in unfaulted condition, and the system $\mathscr{S}$ in each one of its faulted conditions. Then computation for each θ of $V_N(\theta)$, modified with an exponential forgetting factor, will allow identification of the occurrence of a fault. The value of $J(\mathscr{S},\theta)$ provides a measure of how much the presence of a fault affects the output of $\mathscr{S}$. Note also [15] where more sophisticated computations are employed. Again, a more sophisticated approach is, for example, to examine the innovations sequence for whiteness, and tie a fault/no fault decision to a whiteness/coloured decision.

Bayesian estimation. A common practical problem often tackled by methods developed in statistical communication theory is that of deciding, given a measurement, which of a finite number of prescribed random processes gave rise to the measurement. Signal processors—termed detectors—can be constructed and where possible, supplied with *a priori* probabilities for the various alternatives[16].

In this subsection, we explain how a class of such processors will perform, and we note a robustness type of property that is essential for engineering purposes.

Suppose that the true system belongs to the model set, and that if the model set is defined by $\{\theta_1,\ldots,\theta_p\}$, we are given *a priori* probabilities that $p(\theta_{\mathscr{S}}=\theta_i)$. In this subsection, we study the evolution of

$$p(\theta_{\mathscr{S}}=\theta_i \mid y_1, y_2, \ldots, y_t).$$

With Gaussian and stationarity assumptions on $\{y_t\}$, the principal conclusion is that normally all but one of these quantities converge to zero *exponentially* fast, with the remaining quantity approaching unity. The exponent is determined by

$$\min_{\theta_i \neq \theta_{\mathscr{S}}} \bar{J}(\theta_{\mathscr{S}},\theta_i).$$

To see this, we proceed as follows.

Let $Y_t=(y_1, y_2, \ldots, y_t)$. Then it is easily checked that

$$p(\theta_i \mid Y_t)=\frac{p(y_t \mid Y_{t-1},\theta_i)p(\theta_i \mid Y_{t-1})}{\sum_{i=1}^{p} p(y_t \mid Y_{t-1},\theta_i)p(\theta_i \mid Y_{t-1})}. \tag{4.4}$$

Now if the $\{y_t\}$ process is conditionally Gaussian for each θ_i, we have

$$p(y_t \mid Y_{t-1},\theta_i)=|2\pi P_{t|\theta_i}|^{-1/2}\exp[-\tfrac{1}{2}\tilde{y}'_{t|\theta_i}P^{-1}_{t|\theta_i}\tilde{y}_{t|\theta_i}].$$

Now fix i, and set

$$L_t=p(\theta_i \mid Y_t)/p(\theta_{\mathscr{S}} \mid Y_t).$$

Then, without yet invoking stationarity,

$$\ln L_t=\tfrac{1}{2}\{\ln\det(P_{t|\theta_{\mathscr{S}}}P^{-1}_{t|\theta_i})-\tfrac{1}{2}\tilde{y}'_{t|\theta_i}P_{t|\theta_i}\tilde{y}_{t|\theta_i}+\tfrac{1}{2}\tilde{y}'_{t|\theta_{\mathscr{S}}}P^{-1}_{t|\theta_{\mathscr{S}}}\tilde{y}_{t|\theta_{\mathscr{S}}}\}+\ln L_{t-1}$$

and thus, with the exponential stability assumptions and Lemma 3.1

$$\frac{N'}{N}\ln\frac{L_N}{L_0}-\frac{N'}{2N}\sum_{t=1}^{N}\{\ln\det P_{t|\theta_{\mathscr{S}}}P^{-1}_{t|\theta_i}-E(\tilde{y}'_{t|\theta_i}P^{-1}_{t|\theta_i}\tilde{y}_{t|\theta_i})+\ln L_1\}=\frac{N'}{N}\left[\ln\frac{L_N}{L_0}+\tfrac{1}{2}J_N(\theta_{\mathscr{S}},\theta_i)\right]\to 0 \tag{4.5}$$

and in the stationary case when

$$\bar{J}(\theta_{\mathscr{S}},\theta_i)=\lim_{N\to\infty}\frac{1}{N}J_N(\theta_{\mathscr{S}},\theta_i)$$

$$\frac{N'}{N}\left[\ln\frac{L_N}{L_0}+N\bar{J}(\theta_{\mathscr{S}},\theta_i)\right]\to 0$$

from which we can conclude that for some constant ρ

$$\frac{p(\theta_i \mid Y_N)}{p(\theta_{\mathscr{S}} \mid Y_N)}\leqq\frac{p(\theta_i)}{p(\theta_{\mathscr{S}})}\exp\{-\tfrac{1}{2}N[\bar{J}(\theta_{\mathscr{S}},\theta_i)-\rho N^{-\gamma}]\}.$$

It is trivial to conclude that $p(\theta_i \mid Y_N)\to 0$ for $\theta_i\neq\theta_{\mathscr{S}}$, and $p(\theta_{\mathscr{S}} \mid Y_N)\to 1$. The above exponential bounds have an exponent $[\bar{J}(\theta_{\mathscr{S}},\theta_i)-\rho N^{-\gamma}]$, i.e. effectively $\bar{J}(\theta_{\mathscr{S}},\theta_i)$.

As one might expect, the more separated the spectra $\Phi_{\theta_i}(z)$ and $\Phi_{\theta_{\mathscr{S}}}(z)$ are, the more rapid is the convergence here.

Now suppose that the true system is not contained in the model set. (It may, however, be very close to a model in the set). What happens if we try to still use the above procedure? As before, there are assigned *a priori* probabilities $\tilde{p}(\theta_1),\ldots,\tilde{p}(\theta_p)$ to the models in the model set. (The superscript tilde is used to remind us that the assignments in effect are now falsely made, being dependent on an incorrect premise. We might call $\tilde{p}(\cdot)$ a pseudo-probability). Then (4.4) holds with tildes on the probabilities, and with

$$L_t=\tilde{p}(\theta_i \mid Y_t)/p(\theta_{\mathscr{S}} \mid Y_t),$$

we obtain (4.5), whence in the stationary case,

$$\frac{\tilde{p}(\theta_i \mid Y_N)}{\tilde{p}(\theta_j \mid Y_N)}\leqq\frac{\tilde{p}(\theta_i)}{\tilde{p}(\theta_j)}\exp[-\tfrac{1}{2}N[\bar{J}(\theta_{\mathscr{S}},\theta_i)-\bar{J}(\theta_{\mathscr{S}},\theta_j)]-\rho N^{-\gamma}].$$

Now we see that if one model is closer to the true system than all the others, i.e. $\bar{J}(\theta_{\mathscr{S}},\theta_J)<\bar{J}(\theta_{\mathscr{S}},\theta_i)$ for all $i\neq J$, we will have $\tilde{p}(\theta_i \mid Y_N)\to 0$, $i\neq J$, and $\tilde{p}(\theta_J \mid Y_N)\to 1$. The exponential bounds have an exponent now defined by

$$\min_{i\neq J}[\bar{J}(\theta_{\mathscr{S}},\theta_i)-\bar{J}(\theta_{\mathscr{S}},\theta_J)].$$

In the nonstationary case, (4.5) is the crucial equation. Provided one can compute the quantities $J_N(\theta_{\mathscr{S}},\theta_i)$ and provided that

$$\liminf_{N\to\infty} N^{-1}J_N(\theta_{\mathscr{S}},\theta_i)>0$$

for $i\neq\mathscr{S}$, one gets again exponential convergence of $p(\theta_{\mathscr{S}} \mid Y_N)$ to 1 and $p(\theta_i \mid Y_N)$ to zero for $i\neq\mathscr{S}$, with obvious adjustments in case the system is not in the model set, and pseudo probabilities are used.

5. *Conclusions*

In this section, we aim simply to summarise the main results of the paper. First, we have described a convergence result (including a rate) which applies in nonstationary si-

tuations to model identification. Second, we have given spectral interpretations of the prediction error index which exhibit parallels which correspond to asymptotic per sample Kullback information measure in a Gaussian situation. Third, we have indicated how certain probabilities, or probability-like quantities, of importance in a communication theoretic version of the model problem, behave as time evolves.

Appendix

Proof of Lemma 3.1. We shall establish (4.2) first. Observe that

$$\begin{aligned}
&E\left\{\frac{1}{N}V_N(\theta)-E\left[\frac{1}{N}V_N(\theta)\right]\right\}^2\\
&\quad\leqq\frac{1}{N^2}E\left\{\sum_1^N[\tilde{y}'_{t|\theta}P_{t|\theta}^{-1}\tilde{y}_{t|\theta}-E(\tilde{y}'_{t|\theta}P_{t|\theta}^{-1}\tilde{y}_{t|\theta})]\right.\\
&\quad=\frac{1}{N^2}\sum_{t=1}^N\sum_{\tau=1}^N\operatorname{cov}\{\tilde{y}'_{t|\theta}P_{t|\theta}^{-1}\tilde{y}_{t|\theta},\tilde{y}'_{\tau|\theta}P_{\tau|\theta}^{-1}\tilde{y}_{\tau|\theta}\}\\
&\quad=\frac{1}{N^2}\sum_{t=1}^N\sum_{\tau=1}^N\alpha\beta^{|t-\tau|}\\
&\quad\leqq\frac{1}{N^2}\sum_{t=1}^N\sum_{\tau=-\infty}^{\infty}\alpha\beta^{|t-\tau|}\\
&\quad=2\frac{2\alpha}{N(1-\beta)}.
\end{aligned}$$

Then (3.2) follows by the Markov inequality.

To obtain (3.1) set

$$\eta_M=M^{-2}V_{M^2}(\theta)-E[M^{-2}V_{M^2}(\theta)],$$

so that $E[\eta_M]=0$ and $E[\eta_M^2]<\alpha/M^2$ for some $\alpha>0$. It follows that for any $\gamma\in[0,1)$,

$$E[M^{\gamma}\eta_M^2]<\frac{\alpha^3}{M^{2-\gamma}}.$$

By the Markov inequality

$$\Pr\{|M^{\gamma-2}\eta_M|>\delta\}\leqq\frac{\alpha_4(\delta)}{M^{2-\gamma}}$$

and so

$$\sum_{M=1}^{\infty}\Pr\{|M^{\gamma-2}\eta_M|>\delta\}<\infty$$

for arbitrary $\delta>0$. By the Borel Cantelli lemma $M^{\gamma-2}\eta_M\to 0$ w.p.1 as $M\to\infty$. This shows that (3.1) holds for those N which are squares of integers.

To obtain (3.1) for all N, we use an argument of Cramer and Leadbetter, [9]. Let

$$\xi_j=\|y_j-\hat{y}_{j|\theta}\|^2-E\|y_j-\hat{y}_{j|\theta}\|^2$$

and

$$\lambda_M=\sup_{T\in[M^2,(M+1)^2]}M^{\gamma-2}\left|\frac{1}{T}\sum_{j=0}^{T}\xi_j-\frac{1}{M^2}\sum_{j=0}^{M^2}\xi_j\right|.$$

We shall first show that $\lambda_M\to 0$ w.p.1 as $M\to\infty$ and then conclude (3.1). Now

$$\frac{1}{T}\sum_{j=0}^{T}\xi_j-\frac{1}{M^2}\sum_{j=0}^{M^2}\xi_j=\left(\frac{1}{T}\sum_{j=0}^{T}\xi_j-\frac{1}{T}\sum_{j=0}^{M^2}\xi_j\right)+\left(\frac{1}{T}-\frac{1}{M^2}\right)\sum_{j=0}^{M^2}\xi_j$$

so that

$$\begin{aligned}
&\sup_{T\in[M^2,(M+1)^2]}\left|\frac{1}{T}\sum_{j=0}^{T}\xi_j-\frac{1}{M^2}\sum_{j=0}^{M^2}\xi_j\right|\\
&\qquad\leqq\frac{1}{M^2}\sum_{M}^{(M+1)^2}\left|\xi_j\right|+\frac{(M+1)^2-M^2}{M^4}\sum_{0}^{M^2}\left|\xi_j\right|.
\end{aligned}$$

Now consider $E[\lambda_M^2]$. We have, using the Schwarz inequality and Assumption A to obtain the first and second inequality respectively,

$$\begin{aligned}
E[\lambda_M^2]&\leqq M^{\gamma-2}\frac{2}{M^4}\sum_{j=M^2}^{(M+1)^2}\sum_{k=M^2}^{(M+1)^2}E|\xi_j\xi_k|\\
&\quad+M^{\gamma-2}2\left[\frac{(M+1)^2-M^2}{M^4}\right]^2\sum_{j=0}^{M^2}\sum_{k=0}^{M^2}E|\xi_j\xi_k|\\
&\leqq\frac{2\alpha_2M^{\gamma}}{M^4}(2M+1)^2+\frac{4\alpha_2}{1-\beta}M^{\gamma}\left[\frac{(M+1)^2-M^2}{M^4}\right]^2M^2\\
&\leqq\frac{\alpha_5}{M^{2-\gamma}}
\end{aligned}$$

for some $\alpha_5>0$. Again the Markov inequality and Borel Cantelli Lemma show that $\lambda_M\to 0$ w.p.1 as $M\to\infty$.

Suppose N is arbitrary, and let M be such that $M^2\leqq N\leqq(M+1)^2$. Then

$$\begin{aligned}
&\left|N^{\gamma-2}\left\{\frac{1}{N}V_N(\theta)-E\left[\frac{1}{N}V_N(\theta)\right]\right\}\right|\\
&\qquad\leqq 2M^{\gamma-2}\left[\frac{1}{M^2}V_M(\theta)-E\left[\frac{1}{M^2}V_M(\theta)\right]\right]\\
&\qquad\quad+\sup_{T\in[M^2,(M+1)^2]}M^{\gamma-2}\left|\frac{1}{T}\sum_{j=0}^{T}\xi_j-\frac{1}{M^2}\sum_{j=0}^{M^2}\xi_j\right|\\
&\qquad=2M^{\gamma-2}\eta_M+\lambda_M.
\end{aligned}$$

Now let $N\to\infty$. Then $M\to\infty$ and the convergence of $M^{\gamma-2}\eta_M$ and λ_M establishes (3.1).

References

[1] L. A. Liporace: Variance of Bayes estimate. *IEEE Trans. Inform. Theory* **IT-17** (6), 665–669 (1971).

[2] R. M. Hawkes and J. B. Moore: Performance of Bayesian parameter estimators for linear signal models. *IEEE Trans. Control* **AC-21** (4), August 1976. Also University of Newcastle Technical Report EE7410, July 1974.

[3] R. M. Hawkes and J. B. Moore: An upper bound on mean square error for Bayesian parameter estimators. *IEEE Trans. Inform. Theory* **IT-22** (5), September 1976. Also University of Newcastle Technical Report EE7514, May 1975.

[4] R. M. Hawkes and J. B. Moore: Performance bounds for adaptive estimation. *IEEE Proc.* **64** (8), August 1976. Also, Decision methods in dynamic system identification. *Proc. IEEE Decision Control Conf.*, (Special session on Adaptive Processes), December 1975, pp. 645–649. Also, Performance analysis of adaptive-estimators. *Proc. Sixth Nonlinear Estimation Conf.*, San Diego, September 1975. Also, Performance analysis of Bayesian parameter estimators. Ph.D. Thesis of R. M. Hawkes, University of Newcastle, December 1975.

[5] Y. Baram: Information, consistent estimation and dynamic system identification. Ph.D. Thesis, M.I.T., November, 1976, ERL Report No. 718.

[6] P. E. CAINES and J. RISSANEN: Maximum likelihood estimation for multivariable and Gaussian stochastic processes. *IEEE Trans. Inform. Theory* **IT-20** (1), 102 104 (1974). Also J. RISSANEN and P. E. CAINES: Consistency of maximum likelihood estimators for ARMA Processes. Control Systems Report No. 7424, Toronto, Canada, December, 1974.

[7] E. TSE: Bounds for identification error and a quantitative measure of identifiability. *Proc. 1974. IEEE Dec. and Contr.* August, Phoenix, Arizona, pp. 429 435.

[8] P. E. CAINES: A note on the consistency of maximum-likelihood estimates for finite families of stochastic processes. *Annls Statistics* **3** (2), 539 546 (1975).

[9] L. LJUNG: On consistency for prediction error identification methods. Report 7405, Division of Automatic Control, Lund Institute of Technology, Lund, Sweden, March 1974. See also *System Identification: Advances and Case Studies*, D. Lainiotis and R. Mehra, (eds.) Marcel Dekker Inc., to appear. See also, On consistency and identifiability. *Math. Programm. Stud.* **5**, 169 190 (1976).

[10] L. LJUNG: Prediction error methods. University of Linkoping Technical Report, LiTH-ISY-0139 (1977).

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

[12] H. CRAMER and M. R. LEADBETTER: *Stationary and Related Stochastic Processes*. New York, Wiley (1967).

[13] G. S. WATSON: A note on maximum likelihood. *Sankhya* **26A**, 303 304 (1964).

[14] S. KULLBACK: *Information Theory and Statistics*. New York, Wiley (1959).

[15] A. WILLSKY: A generalized likelihood ratio approach to state estimation in linear systems subject to abrupt changes. *Proc. IEEE 1974 Dec. and Contr. Conf.*, Phoenix Arizona, pp. 846 853.

[16] H. L. VAN TREES: *Detection, Estimation and Modulation Theory*. (Vol. I in particular). John Wiley, New York (1971).