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 property of a true metric.

For the case when \( \eta \) 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 \( \eta \) 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 A number of the results of [4], with tedious derivations as yet unpublished, are achieved here more directly than in [4] as a special case of the general prediction error identification theory.

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 \( \eta \) and measurements of some of the variables associated with \( \eta \). We have to "identify" \( \eta \), and for this purpose there is given a collection of models, such that either \( \eta \) lies within the model set, or else there is a particular model within the model set which in some way is closest to \( \eta \). By using the available measurements, we have to decide which model in the model set is identical with or is closest to \( \eta \).

In [1], Liporace gives convergence results for Bayesian estimation of identically and independently distributed discrete-time processes for the cases when the system \( \eta \) (non-linear) is contained in the finite model set (also non-linear) 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 \( \eta \) 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 \( \eta \) 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 \( \eta \) does not belong to the model set. More recently Baran 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.

2. Systems, models and identification criteria

This section is definitional in character. Consider a causal discrete time stochastic system \( \eta \) with 'known' output sequence \( \{y_i\} \) and 'known' input sequence \( \{u_i\} \), and such that a one-step ahead prediction estimate \( \hat{y}_{i|0} \) is available given \( y_{1}, \ldots, y_{i-1} \) and \( u_{1}, \ldots, u_{i-1} \) and \( \text{a priori} \) data \( y_{0} \). We write

\[
\hat{y}_{i|0} = \hat{y}_{i|0} + \epsilon_{i|0}
\]

Consider a collection \( \mathcal{H} = \{ \eta \} \) of stochastic models, each defined by a parameter vector \( \theta \), and to keep the theoretical development as simple as possible assume that \( \theta \) belongs to a finite set \( \{ \theta_1, \ldots, \theta_k \} \). The parameter vector \( \theta \) may be involved in determining both the probability densities governing the stochastic signals driving \( \mathcal{H}_\theta \), 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

\[
y_i = \hat{y}_{i|0} + \epsilon_i\eta_0 \]

Here \( \epsilon_i\eta_0 \) is defined in the same manner as \( \hat{y}_{i|0} \), save that the \( \{y_i\} \) process is assumed to result from the model, rather than the system. We further assume that there is a known function \( \gamma_\eta \) for the conditional one step ahead predictor such that

\[
\hat{y}_{i|0} = \gamma_\eta(\theta, y_1, \ldots, y_i, u_0, \ldots, u_i, x_0) \]
Now note that even if the \( \hat{\theta}_n \) process is not generated by the model, one can still form the right hand side of (2.3), and in this sense formally obtain \( \hat{\theta}_n \) and \( e_n \). 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 \( \hat{\theta}_n \) is the conditional mean estimate (ii) the noise is non-Gaussian and \( \hat{\theta}_n \) is the best linear conditional minimum variance estimate (iii) \( \hat{\theta}_n \) is the standard least squares estimate when it is assumed that

\[
y_t = \theta_n x_{t-1} + \epsilon_t
\]

for \( x_{t-1} \) measurable.

Suppose now that the sequence \( \{\hat{\theta}_n\}_n \), where \( \hat{\theta}_n = \theta_n - \bar{\theta}_n \), is conditionally white and Gaussian with covariance \( P_n \) and of \( \hat{\theta}_n \) 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,

\[
V_n(\theta) = N \sum_{i=1}^{n} \log \det \frac{1}{N} \mathbf{I} + \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T = \frac{1}{N} \sum_{i=1}^{n} \log \det \frac{1}{N} \mathbf{I} + \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T
\]

One might then define that model which was the best approximation to \( \mathcal{N} \) as that obtained by choosing the value of \( \theta \) which minimizes (3.1). 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 \( \bar{\theta}_n \) not conditionally Gaussian including the case when \( \bar{\theta}_n \) is a linear estimate rather than a conditional mean estimate. The index still has some intuitive content, especially if \( P_n \) is either \( E[\hat{\theta}_n \hat{\theta}_n^T] \) or

\[
E[\hat{\theta}_n y_n]|y_1, \ldots, y_n, x_1, \ldots, x_n
\]

One might abstract \( P_n \) from a covariance interpretation altogether as in Ljung [9].

The right side of (2.4) can be rewritten as

\[
\sum_{i=1}^{n} \log \det \frac{1}{N} \mathbf{I} + \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T + \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T - \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T + 2 \hat{\theta}_n \hat{\theta}_n^T - \frac{1}{N} P_n \hat{\theta}_n \hat{\theta}_n^T
\]

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 \( \mathcal{N} \) and the model \( \theta_n \) 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 \( \mathcal{N} \) and \( \theta_n \). 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 \( \epsilon > 0 \) and \( \beta \),

\[
\text{cov}(\epsilon_{n}\hat{\theta}_n, \epsilon_{n}\hat{\theta}_n) - \beta \leq \epsilon
\]

When

\[
P_{\epsilon} = E[\epsilon_{n}\hat{\theta}_n \hat{\theta}_n^T]
\]

all that is required is that

\[
\|P_{\epsilon}\| < \beta
\]

for some \( \beta \) and

\[
\text{cov}(\epsilon_{n}\hat{\theta}_n, \epsilon_{n}\hat{\theta}_n) - \beta \leq \epsilon
\]

Relaxed versions will be noted later. One obvious and important situation when the exponential bound on the covariance will hold is when \( \mathcal{N} \) and \( \theta_n \) are linear systems with impulse responses \( w(\cdot) \) satisfying

\[
\|w(t, \cdot)\| \leq \beta \epsilon^{\beta}
\]

and with no hidden (uncontrollable or unobservable) modes, and the input to \( \mathcal{N} \) 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 \( \theta_n \) or the predictor associated with \( \mathcal{N} \) 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 \( P_n \) deterministic. With Assumption A in force,

\[
\lim_{N \to \infty} \mathbb{E} \left[ \frac{1}{N} \sum_{n=1}^{N} V_n(\theta) - \frac{1}{N} E[V_n(\theta)] \right] = 0 \quad \text{w.p.1}
\]

for all \( \gamma \in [0,1] \) and

\[
\mathbb{P} \left( \left| \frac{1}{N} \sum_{n=1}^{N} V_n(\theta) - \frac{1}{N} E[V_n(\theta)] \right| > \epsilon \right) < \frac{2 \epsilon}{N(1 - \beta)}
\]

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 class 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 \( \beta \) minimizing

\[
\lim_{N \to \infty} \mathbb{E} \left[ \frac{1}{N} \sum_{n=1}^{N} V_n(\theta) \right]
\]

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

\[
\epsilon = \min_{\gamma > 0} \left\{ \mathbb{E} \left[ \frac{1}{N} \sum_{n=1}^{N} V_n(\theta) \right] - \mathbb{E} \left[ \frac{1}{N} E[V_n(\theta)] \right] \right\}
\]
Then if
\[
\left|\frac{1}{N}W_\delta(\theta) - \frac{1}{N}E[V_y(\theta)]\right| < \epsilon
\]
for all \(\theta_i\) including \(\theta^*\), it follows that \(\theta(N)\) minimizes \(\theta^*\).
Consequently if the experiment length \(N\) is such that
\[
\frac{2c}{\epsilon^2 N(1 - \beta)} < 1 - \rho^4
\]
where \(N\) is the number of distinct \(\theta_i\), then the probability that \(\theta(N) = \theta^*\) is at least \(\rho\).

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 \(\theta^*\) in the set, \(E[V_y(\theta^*)]/N\) is as small as possible, i.e., the model is as little different from the system as possible. One also apparently wants \(\epsilon\) 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 \(E[V_y(\theta_i)/N] = E[V_y(0)/N]\) for some \(i\) and \(j\), it will require more measurements to decide whether \(\theta_i\) or \(\theta_j\) is closer to the system, and at the same time the difference in distances of \(\theta_i\) from the system and \(\theta_j\) from the system is not large, so that if a large number of measurements are not used, and one of \(\theta_i\) and \(\theta_j\) is wrongly identified, the error will not be great.

**Strengthened Assumption A**: Assumption A holds for all \(\theta\) in the model set, and with \(Y\) replaced by \(\mathcal{Y}\).

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
\[
\text{tr} AB^{-1} - \ln \det AB^{-1} - \text{tr} I \leq 0
\]
with equality if and only if \(A = B\).

Evidently, the quantity \(\text{tr} AB^{-1} - \ln \det AB^{-1} - \text{tr} I\) provides a measure of the error in approximating \(A\) by \(B\); if \(\|AB^{-1} - I\|\) is small, one can show that
\[
\text{tr} AB^{-1} - \ln \det AB^{-1} - \text{tr} I \geq \frac{1}{2} \| (A - B)B^{-1} \|^2.
\]

Consider now the following index, which, it should be noted, is not likely to be computed prior to identification, since it involves the quantities \(\bar{\gamma}_{ij}^\star\).
\[
W_\delta(\theta) = \frac{N}{2} \left[ \text{tr} P_{\theta}^\delta P_{\theta}^\delta - \ln \det P_{\theta}^\delta P_{\theta}^\delta - \text{tr} I \right]
\]
\[
+ \left( \bar{\gamma}_{ij}^\star \right)_{\delta} \sum_{i,j} P_{\theta}^\delta \left( \bar{\gamma}_{ij}^\star - \bar{\gamma}_{ij} \right).
\]  
(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 \(P_{\theta}^\delta\) and \(P_{\theta}^\delta\), or the error covariance associated with use of the correct predictor \(\mathcal{Y}\), and the error covariance associated with use of the incorrect predictor \(\mathcal{Y}\). More obviously the index also reflects the error between \(\bar{\gamma}_{ij}^\star\) and \(\bar{\gamma}_{ij}\). The index obviously then has intuitive content. The substance of the next lemma is that the index for large \(N\) is like \(V_y(\theta)\).

**Lemma 3.3.** With notation as above,
\[
\frac{1}{N} E[V_y(\theta) - W_\delta(\theta)] = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N \ln \det P_{\theta\eta} + \text{tr} I
\]
and with the strengthened Assumption A in force
\[
\lim_{N \to \infty} \frac{1}{N} \left( W_\delta(\theta) - E[W_\delta(\theta)] \right) = 0.
\]  
(3.5)

**Proof:** \(\bar{\gamma}_{ij}^\star = \bar{\gamma}_{ij}^\star P_{\theta}^\delta\) for all \(\theta\), and with the strengthened Assumption A in force
\[
\frac{1}{N} \left( W_\delta(\theta) - E[W_\delta(\theta)] \right) = 0.
\]

**Kullback information measure connection.** Here we further motivate the use of an index such as \(V_y(\theta)\) in identification. As already observed, in case \(\theta_i\) is conditionally white and Gaussian, \(V_y(\theta) = -\ln p(y_1, \ldots, y_N | \theta)\). Consequently, \([14]\), the task of minimizing \(E[V_y(\theta)]\) is then one of finding the model which is closest to the system in the sense of minimizing the Kullback information measure:
\[
J_\delta(P_{\theta}; P_{\theta_0}) = \ln \frac{p(y_1, \ldots, y_N | \mathcal{Y})}{p(y_1, \ldots, y_N | \theta)}
\]

This quantity is nonnegative and equal to zero if and only if
\[
p(y_1, \ldots, y_N | \mathcal{Y}) = p(y_1, \ldots, y_N | \theta)
\]
almost everywhere.
When $\mathcal{F}$ has the property that $p(x_1, \ldots, x_n)$ is also Gaussian, it is easy [14] to get a precise expression for $J_{\mathcal{F}}(\mathcal{F}, \theta)$:

$$J_{\mathcal{F}}(\mathcal{F}, \theta) = \sum_{i=1}^{N} \left[ \ln P_{0}(i) + E\left[ P_{0}(i)^{-1} \tilde{y}_{ni}(i) - \ln I \right] \right].$$  (3.6)

Of course, $J_{\mathcal{F}}(\mathcal{F}, \theta) \geq 0$. In fact, one can check that $J_{\mathcal{F}}(\mathcal{F}, \theta) = E[W_{0}(0)\theta]$ with $W_{0}(0)$ defined as in (3.3).

4. Specialized properties of the limiting identification index

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

Spectrum interpretation. Suppose that the system $\mathcal{F}$ 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, $W_{0}(z)$, has $W_{0}(0) = I$, and the driving sequence has zero mean, bounded 4th order moments, and covariance $P_{\varphi}$. The covariance of $\tilde{y}_{ni}(i)$ is $P_{\varphi}$. The spectrum of $\tilde{y}_{ni}(i)$ may be obtained. With $\tilde{y}_{ni}$ actually the system output, the actual spectrum of $\tilde{y}_{ni}$ is evidently

$$C_{\tilde{y}_{ni}(i)} = W_{0}(z) P_{\varphi} W_{0}^{*}(z^{-1}).$$

Further, with integrations anticlockwise round the unit circle,

$$E[\tilde{y}_{ni}(i) P_{0}(i)^{-1} \tilde{y}_{ni}(i)] = \frac{1}{2\pi i} \int_{|z| = 1} \left[ \text{tr} [W_{0}(z) W_{0}^{*}(z^{-1})] P_{\varphi} - 1 \right] dz;$$

where

$$\Phi_{\mathcal{F}}(z) = W_{0}(z) P_{\varphi} W_{0}^{*}(z^{-1})$$

is the spectrum of $\tilde{y}_{ni}(i)$ and

$$\Phi_{\varphi}(z) = W_{0}(z) P_{\varphi} W_{0}^{*}(z^{-1})$$

is the spectrum of $\tilde{y}_{ni}(i)$ conditioned on $\mathcal{F}_{0} = \mathcal{F}$.

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 $\theta$ which minimizes

$$\ln \text{det} P_{\varphi} + \frac{1}{2\pi i} \int_{|z| = 1} \left[ \text{tr} [\Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z)] \right] dz;$$

or equivalently

$$\ln \text{det} P_{\varphi} - \frac{1}{2\pi i} \int_{|z| = 1} \left[ \ln \text{det} [\Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z)] \right] dz.$$  (4.1)

Now we use an identity that is proved easily by complex variable methods, at least when $\Phi_{\mathcal{F}}(z)$ and $\Phi_{\varphi}(z)$ 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,

$$\ln \text{det} P_{\varphi} - \frac{1}{2\pi i} \int_{|z| = 1} \left[ \ln \text{det} \Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z) \right] dz^{-1} dz;$$  (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 $\theta$ which minimizes $(1/N) \bar{J}(\mathcal{F}, \theta)$ as $N \to \infty$ is also the $\theta$ which minimizes

$$J(\mathcal{F}, \theta) = \frac{1}{4\pi i} \int_{|z| = 1} \left[ \text{tr} \Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z) \right] dz;$$

and

$$\ln \text{det} \Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z) - \ln I \right] dz.$$  (4.3)

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

It is perhaps of interest to note that if $\Phi_{\mathcal{F}}(z) = \Phi_{\varphi}(z)$ is small with respect to $\Phi_{\mathcal{F}}(z)$, then (4.2) can be approximated by

$$J(\mathcal{F}, \theta) = \frac{1}{4\pi i} \int_{|z| = 1} \left[ \text{tr} [\Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z)] \right] z^{-1} dz;$$

where $\Phi_{\mathcal{F}}(z) = \Phi_{\varphi}(z) = \Phi_{\mathcal{F}}(0 \theta)$, 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

$$\ln \text{det} P_{\varphi} - \frac{1}{2\pi i} \int_{|z| = 1} \left[ \text{tr} [\Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z)] \right] dz,$$

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

$$J(\mathcal{F}, \theta) = J_{\mathcal{F}}(\mathcal{F}, \theta).$$

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

$$J(\mathcal{F}, \theta) = \lim_{N \to \infty} J_{\mathcal{F}}(\mathcal{F}, \theta).$$

Thus $J(\mathcal{F}, \theta)$ 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 $\theta$ minimizing $J(\mathcal{F}, \theta)$ still identifies the model closest to the system.

One advantage that the index $J(\mathcal{F}, \theta)$ 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 $\Phi_{\mathcal{F}}(z)$ and $2 \Phi_{\mathcal{F}}(z)$ lead to different values of $J$.

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

$$V_{0}(z) = \frac{1}{z^{2} + \beta_{1}^{2} z + \beta_{2}^{2}}.$$

The parameter vector $[\beta_{1}, \beta_{2}, \beta_{3}, \beta_{4}]$ define a model. Then one can compute

$$J(\mathcal{F}, \theta) = \frac{1}{4\pi i} \int_{|z| = 1} \left[ \text{tr} \Phi_{\mathcal{F}}(z) \Phi_{\varphi}(z) \right] dz;$$

For $\beta_{1}$ taking the values 0.5 and 0.9, this is plotted as a function of $\beta_{1}$ in Fig. 1.
(b) Suppose the true system is defined by the parameter vector \([0, 0, \bar{x}, 0]\) and the model is defined by \([0, 0, x, 0]\). Then

\[
J(\theta) = \frac{(x - \bar{x})^2}{(1 - \bar{x})^2}.
\]

For \(x\) taking the values 0.5 and -0.9, this is plotted as a function of \(x\) in Fig. 2.

(c) Suppose the true system is defined by the parameter vector \([0, 0, \bar{x}, 0]\) 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

\[
J(\theta) = \frac{1}{2} \frac{1 + \beta_2 (1 - \beta_2^2)}{1 - \beta_1 (1 - \beta_2^2)(1 - \beta_1^2)}
\]

\[
\times (1 - \beta_2^2)(1 + \beta_1^2 - \beta_2^2)
\]

For \(\beta_1 = 0.3, \beta_2 = 0.9\), this is plotted against \(\bar{x}\) in Fig. 3.

Figures 1 and 2 illustrate the fact that an error in zero or
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 $J(\beta, 0)$ 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 $J$.

Let us mention two other applications of the formula (4.3). The first is low order modelling of high order systems. Let $Y = (y_1, y_2, \ldots, y_T)$. Then it is easily checked that

$$p(0_1, \ldots, 0_p | Y) = p(0_1 | Y) \cdot p(0_2 | Y, 0_1) \cdot \ldots$$

whence in the stationary case,

$$p(0_1, \ldots, 0_p | Y) = \exp[-N J(0_1, \ldots, 0_p, Y)]$$

and thus, with the exponential stability assumptions and Lemma 3.1

$$N^2 \sum \ln \det P_{\theta_\beta} - 2N \ln \sum \ln \det P_{\theta_\beta} p_{\theta_\beta} = -E \ln p_{\theta_\beta} p_{\theta_\beta} + \ln 1 \Rightarrow \lim \sup_{N \to \infty} \ln \frac{L_N}{L_{N-1}} = \frac{1}{N}$$

and in the stationary case when

$$J(0_1, \ldots, 0_p) = \lim_{N \to \infty} \frac{1}{N} \sum_{\omega} \ln \left[ \frac{L_N}{L_{N-1}} \right] \Rightarrow 0$$

from which we can conclude that for some constant $\rho$

$$p(0_1, \ldots, 0_p | Y) \leq \exp[-N J(0_1, \ldots, 0_p, Y) - \rho N \sup]$$

This is trivial to conclude that $p(0_1, \ldots, 0_p | Y) \leq p(0_1, \ldots, 0_p | Y)$. The above exponential bounds have an exponent $J(0_1, \ldots, 0_p, Y) - \rho N \sup$, i.e. effectively $J(0_1, \ldots, 0_p, Y)$.

As one might expect, the more separated the spectra $\Phi_{\theta_\beta}$ and $\Phi_{\theta_\beta}$ 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 $\rho(0_1, \ldots, 0_p) \rho(0_1, \ldots, 0_p, Y)$ to the models in the model set. (The super- minus tilde is used to remind us that the assignments in effect are now falsely made, being dependent on an incorrect premise. We might call $\rho(0_1, \ldots, 0_p)$ a pseudo-probability). Then (4.4) holds with tildes on the probabilities, and with

$$J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta) = \tilde{p}(0_1, \ldots, 0_p, \tilde{\theta}_\beta)$$

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

$$N^{1/2} J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta) \leq \rho(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta) \exp[-N J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta) - \rho N \sup]$$

Now we see that if one model is closer to the true system than all the others, i.e. $J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta)$ for all $i \neq j$, we will have $p(0_1, \ldots, 0_p | Y, \tilde{\theta}_\beta) \leq p(0_1, \ldots, 0_p | Y, \tilde{\theta}_\beta)$ for all $i \neq j$.

In the nonstationary case, (4.5) is the crucial equation. Provided one can compute the quantities $J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta)$ and provided that

$$\lim \inf_{N \to \infty} N^{1/2} J(0_1, \ldots, 0_p, Y, \tilde{\theta}_\beta) > 0$$

for $i \neq j$, one gets again exponential convergence of $p(0_1, \ldots, 0_p | Y, \tilde{\theta}_\beta)$ to 1 and $p(0_1, \ldots, 0_p | Y, \tilde{\theta}_\beta)$ to zero for $i \neq j$, 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-
tations 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

\[
E\left[ \frac{1}{N} V_{i}(\theta) - E\left[ \frac{1}{N} V_{i}(\hat{\theta}) \right] \right]^{2}
\]

so that

\[
\sup_{r \geq M^{2}} \left| \frac{1}{r} \sum_{i=1}^{r} \frac{1}{M^{2}} \sum_{j=0}^{r} \frac{y_{j}^{2}}{M^{2}} \right| \leq \frac{1}{M^{2}} \sum_{i=0}^{r} \frac{y_{i}^{2}}{M^{2}} \leq M^{2} \sum_{i=0}^{r} \frac{y_{i}^{2}}{M^{2}}
\]

Now consider \( E[\tilde{\eta}_{M}] \). We have, using the Schwarz inequality and Assumption A to obtain the first and second inequality respectively.

\[
E[\tilde{\eta}_{M}] \leq \frac{1}{N} \sum_{j=0}^{N} \sum_{i=0}^{r} E[|z_{i,j}|]
\]

for some \( z_{i,j} > 0 \). Again the Markov inequality and Borel Cantelli Lemma show that \( \tilde{\eta}_{M} \rightarrow 0 \) w.p.1 as \( M \rightarrow \infty \).

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

\[
\left| \frac{1}{N} V_{i}(\theta) - E\left[ \frac{1}{N} V_{i}(\hat{\theta}) \right] \right| \leq 2M^{2} \sum_{i=0}^{r} \frac{y_{i}^{2}}{M^{2}} \leq M^{2} \sum_{i=0}^{r} \frac{y_{i}^{2}}{M^{2}}
\]

Now let \( N \rightarrow \infty \). Then \( M \rightarrow \infty \) and the convergence of \( M^{2} \tilde{\eta}_{M} \) and \( \tilde{\eta}_{M} \) establishes (3.1).

References


