true distributions. Therefore, the values of AIC can always be understood as an approximation to the relative distance from the model to the reality. For ARMA models, the distance is the error of the estimated predictor, which is equivalent to the distance $m_h(T)$ in the frequency domain, as is mentioned in the paper. Roughly speaking, the use of the minimum AIC procedure is recommended if such distance suits for the purpose of the analysis. Otherwise, for example, if the purpose is to know the correct order or to do classification rather than to get a good approximation to the reality in terms of prediction error, a criterion like BIC is recommended provided that the true order is finite and falls into the range of selection. In any case, a plot of both criteria will be more helpful in understanding the situation. The analyst is not restricted to only selecting the minimizer of either criterion.

ADDITIONAL REFERENCES

KEMPTHORNE, P. J. (1984). Admissible variable-selection procedures when fitting regression models by least squares for prediction. *Biometrika* **71** 593–597.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHIBATA, R. (1986a). Consistency of model selection and parameter estimation. In *Essays in Time Series and Allied Processes* (J. M. Gani and M. B. Priestley, eds.) 127–141. Applied Probability Trust, Sheffield.

SHIBATA, R. (1986b). Selection of the number of regression variables; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38A** 459–474.

STONE, C. J. (1981). Admissible selection of an accurate and parsimonious normal linear regression model. *Ann. Statist.* **9** 475–485.

STONE, C. J. (1982). Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Ann. Inst. Statist. Math.* **34** 123–133.

TAKADA, Y. (1982). Admissibility of some variable selection rules in linear regression model. *J. Japan Statist. Soc.* **12** 45–49.

Comment

V. Solo

As usual Ted Hannan has provided a comprehensive discussion of a number of important and difficult topics in the statistical theory of linear systems. Some readers will find the presentation fast paced so I would like to expand on some topics and make various other comments.

1. HANKEL NORMS

If you look at the state of time series in the 1950s, particularly Whittle's work and the book by Quenouille (1957), it is quite sobering to see how well developed the field was. One big problem though was how to tackle the lag structure of multivariate time series. Ted was the first in the statistical and econometric literature to see how to handle the problem through the theory of matrices of polynomials (Hannan, 1969). At about the same time, but independently, control engineers were on to the same idea.

The next step was from Akaike (1976) who gave Kronecker indices (a control engineering development) a statistical interpretation. An exposé of the ideas is available in Solo (1982/1986).

To see the need for the Hankel norm theory that Ted relates, it is useful to look at the univariate version of Akaike's ideas. Very briefly it goes like this. The generalized Yule-Walker equations for estimating autoregressive parameters in ARMA models yields a set of equations of the form

$$H_m a_m = b_m$$

where a_m are the AR parameters; H_m is an $m \times m$ Hankel matrix; and b_m is a vector of autocovariances. The order of the ARMA model is the rank of H_m. By computing the singular values of H_m for increasing m and using an AIC criterion based on canonical correlation ideas, one can choose the order.

There are two problems. First, the procedure has very poor statistical efficiency. Second, if a_r (r is the order) is found from the above equations, there is no guarantee it gives a stable polynomial. There are two ways out of this problem, one is to use the Hankel norm approach and the other is to use a properly constituted maximum likelihood method.

If data

$$(y_1 \cdots y_n) = y$$

are available, the likelihood may be specified as

$$\log \text{lik} \propto \ln |\Sigma| - \tfrac{1}{2} y^T \Sigma^{-1} y,$$

where Σ is the Toeplitz matrix of autocovariances

$$\gamma_s = \int e^{jws} F(w \mid \theta)\, dw/2\pi$$

V. Solo is Associate Professor, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218.

and

$$\theta = (A_1 \cdots A_p B_0 \cdots B_q)^T,$$

$$F(w \mid \theta) = \frac{|\sum_0^q B_u e^{jwu}|^2}{|\sum_0^p A_u e^{jwu}|^2}.$$

This method of calculating the likelihood means it is only evaluated for stable A polynomials, and hence, if it is maximized over a compact set, it has a maximum whose A polynomial is stable.

There is a weakness though; computation (1) must be done numerically. It can be done algebraically but the description of the necessary procedures is a little long and out of place here.

2. MOVING AVERAGES

Ted comments in the second paragraph that moving averages lack realism. Many time series workers have made the same comment about ARMA models. I do not agree with this for the following simple reason. A sinusoid plus white noise can be treated as an ARMA (2.2) model with ARMA factors equal, both having roots on the unit circle. Such data can be fitted in this way—my time series students have been doing this for years. I have observed that the AR parameters seem to be estimated much better than the MA ones. It can be shown they have standard error of order $n^{-3/2}$, whereas the MA ones have standard error of order $n^{-1/2}$. This shows that deterministic models can be fitted within the stochastic framework. A priori one does not usually know which case one is dealing with, so this comprehensiveness of the stochastic framework is invaluable.

3. ALGORITHMS

It is important to distinguish between two types of on-line or adaptive parameter estimators. Those without a forgetting factor, namely long memory schemes, and those with a forgetting factor, namely short memory schemes. Only a short memory algorithm can hope to track varying parameters. Short memory algorithms have been in wide use in electrical engineering for the past twenty years (see Widrow and Stearns, 1985). The book by Ljung and Söderström (1983) deals almost exclusively with long memory algorithms.

4. ORDER ESTIMATION

It is true that Akaike was early to deal with the issue of order estimation. However, Mallows C_p criteria was also developed independently in the 1960s. It was not, of course, applied initially to time series.

It seems to me that the specification of approximation error is an issue separate from sampling considerations. Thus in fitting a stationary model for prediction one might insist that the approximated prediction variance be within a fraction δ of the lower limit. This specifies an order d as

$$\min_d \, (\sigma_{\varepsilon d}^2 - (1 + \delta)\sigma_\varepsilon^2)^2$$

where $\sigma_{\varepsilon d}^2$ is the prediction error of the approximation and σ_ε^2 is the Kolmogorov prediction variance. Now one can consider what sort of order criteria are consistent with the above value of d. More ingenuity or knowledge of a particular problem should allow one to define a deterministic sequence $d(n)$ if one believes the order should increase with sample size.

I do not understand Ted's italicized comment, a little after equation (6.2), that Akaike's canonical correlation method is hard to extend to the case of observed inputs. I have done this myself for the single input, single output case (Solo, 1983). One investigates by singular value decomposition the rank of matrices whose block elements are of the form

$$E\begin{pmatrix} y_{k-s} \\ u_{k-s} \end{pmatrix}(u_{k-2s+1} u_{k-2s}).$$

The multiple input, multiple output case is only more difficult because of the care needed with Kronecker indices. It seems to me that the approximate method described by Ted after the italicized comment could be made exact by embedding it in an EM algorithm calculation. There is not room here to go into details, but basically one treats the residuals as unobserved data in the EM calculation.

ADDITIONAL REFERENCES

QUENNOUILLE, M. H. (1957). *The Analysis of Multiple Time Series.* Hafner, New York.

SOLO, V. (1982/1986). Topics in advanced time series. *Lecture Notes in Math.* **1215**. Springer, New York.

SOLO, V. (1983). Order estimation for single input, single output transfer function models by singular value decomposition. Technical Report, Dept. Statistics, Harvard Univ.

WIDROW, B. and STEARNS, S. D. (1985). *Adaptive Signal Processing.* Prentice Hall, Englewood Cliffs, N.J.