Sections 4 and 5 made by Professor Hannan about my minimum description length (MDL) principle. Although some of the main analytical results of the predictive and the semipredictive versions of the criterion do presently require Gaussian assumptions, the same is not true of the general criterion nor by any means of the applicability of the principle itself. Furthermore, the MDL principle has more recently been expressed in a new and more satisfactory form (Rissanen, 1987), where the several earlier versions appear as computable approximations of the central notion, the stochastic complexity and which certainly is not restricted to Gaussian likelihoods nor any other ad hoc choices. In fact, an application of the principle amounts to searching for a model class among any that we can think of which permits the largest assignment of a density or probability to the actually observed set of data. The classes may, if desired, be restricted by constraints not determined by the observed data, such as "prior knowledge" or considerations involving the intended application of the model. Hence, it represents a sort of "global" maximum likelihood principle, which is free from any choices with the possible exception of the desired extraneous constraints. The principle is equally well applicable to the selection of models, regardless of the number of parameters in them, as to hypothesis testing, and consequently it is difficult for me to imagine a statistical problem which could not be dealt with in such a manner. In this my thinking appears to be bolder than Professor Hannan's more cautious view, according to which the existence of any generally applicable principle is in doubt.

ADDITIONAL REFERENCE

RISSANEN, J. (1987). Stochastic complexity (with discussion). To appear in *J. Roy. Statist. Soc. Ser B* **49**.

Comment

Ritei Shibata

It is my great pleasure to comment on Professor Ted Hannan's excellent review paper. This paper covers a wide range of topics in stationary multiple time series analysis. My comment is only on a part, "order estimation procedure" for the case $n = 1$. I strongly agree with him that there is no means by which it can be established that AIC is always to be preferred to BIC or the reverse. The admissibility result that any choice of C_T implies admissible order estimation (Stone, 1981, 1982; Takada, 1982; Kempthorne, 1984) supports us.

The results by Shibata (1986a, 1986b) suggest that consistency of order estimation and uniform order of consistency, in terms of mean squared error, of the resulting parameter estimates are not compatible. I therefore also agree with the author that the choice of procedure should be related to the purpose of the analysis. In this respect, I could not understand the derivation of BIC by Rissanen, particularly the relevance of quantizing and coding both observations and parameters. I prefer the original derivation of BIC by Schwarz (1978) from the Bayesian point of view. For a Koopman-Darmois family, the log of the marginal likelihood,

$$\log \int e^{T(y'\theta - b(\theta))}\, d\mu(\theta)$$

is approximated by

$$\sup_{\theta \in \Theta} T(y'\theta - b(\theta)) - \frac{d}{2} \log T,$$

for large T, provided that $\mu(\theta)$ has a density with respect to Lebesgue measure, which is bounded and locally bounded away from zero. The penalty term $-d/2 \log T = \log T^{-d/2}$ follows from the boundedness assumption on $\mu(\theta)$ and the fact that the integration of $\exp(-T \| \theta \|^2)$ over d-dimensional Euclidean space Θ is $(2\pi T)^{-d/2}$. However, if $\mu(\theta)$ is chosen as a measure whose density becomes peaky as T increases, then the penalty is not necessarily of the order of $\log T$. For example, if $\mu(\theta)$ is concentrated on a $1/\sqrt{T}$ neighborhood, the penalty is of the order of constant like as in AIC (Takada, 1982).

One significant difference of AIC from other criteria is in the derivation based on a distance, the Kullback-Leibler information number for the model and the

Ritei Shibata is Associate Professor in Statistics, Department of Mathematics, Keio University, 3-14-1 Hiyoshi Kohoku, Yokohama 223, Japan.

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 Quennouille (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.