Hannan has written. With his retirement, I expect to see papers being generated even more rapidly.

## ADDITIONAL REFERENCES

Beltrão, K. I. and Bloomfield, P. (1987). Determining the bandwidth of a kernel spectrum estimate. *J. Time Ser. Anal.* **8** 21–38.

Brillinger, D. R. (1975). Statistical inference for stationary point processes. In *Stochastic Processes and Related Topics* (M. L. Puri, ed.) 55–100. Academic, New York.

Brillinger, D. R. (1985). Fourier inference: some methods for the analysis of array and non-Gaussian series data. *Water Res. Bull.* **21** 743–756.

Gray, H. L., Houston, A. G. and Morgan, F. W. (1978). On G-spectral estimation. In *Applied Time Series Analysis* (D. F. Findley, ed.) 39–138. Academic, New York.

Lii, K.-S. (1985). Transfer function model order and parameter estimation. *J. Time Ser. Anal.* **6** 153–170.

Mehra, R. K. (1971). Identification of stochastic linear dynamic systems using Kalman filter representation. *AIAA J.* **9** 28–31.

Schweppe, F. C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory* IT-**11** 61–70.

Wahba, G. and Wold, S. (1975). Periodic splines for spectral density estimation: the use of cross validation for determining the degree of smoothing. *Comm. Statist.* **4** 125–141.

# Comment

**R. Dahlhaus**

This paper by Hannan is an excellent review of an important topic in time series analysis: the approximation of a nonparametric time series by a parametric rational one. The paper gives insight into the problems which arise and offers a variety of methods to tackle these problems.

To regard a fitted parametric model as an approximation to a nonparametric time series is clearly the correct point of view when dealing with parametric time series analysis. Many interesting papers dealing with related problems have been published in recent years and there is great need for further results to develop the theory sufficiently. The present paper is an important contribution to this goal.

It was therefore a pleasure for me to read this stimulating paper and to have been asked to comment on it. I will restrict my comments to the problem of estimation and in particular to the case of a one-dimensional process which is approximated by an autoregressive process.

## 1. THE APPROXIMATION CRITERION

Since the goal of the paper is the approximation of the transfer function, it seems to be natural to take a criterion which measures the quality of the approximation directly. Suppose the original series has an infinite autoregressive representation

$$\sum_{s=0}^{\infty} a_s Y_{t-s} = \varepsilon_t \quad \text{with } \varepsilon_t \text{ i.i.d.,}$$
$$E\varepsilon_t = 0, \quad \operatorname{var}(\varepsilon_t) = \sigma^2,$$

and $Y_t$ is approximated by an AR($k$)-process whose coefficients are estimated from the data by $\hat{a}_1(k), \cdots, \hat{a}_k(k), \hat{\sigma}_k^2$. An appropriate approximation criterion then would be, for example,

$$(1) \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \frac{\sigma \hat{A}_k(\lambda)}{\hat{\sigma}_k A(\lambda)} - 1 \right|^2 d\lambda,$$

where

$$A(\lambda) = \sum_{s=0}^{\infty} a_s \exp(-i\lambda s)$$

and

$$\hat{A}_k(\lambda) = \sum_{s=0}^{k} \hat{a}_s(k) \exp(-i\lambda s).$$

Considering the relative difference between $\hat{A}_k(\lambda)$ and $A(\lambda)$ is natural, since for Yule-Walker estimates (1) is approximately equal to $\sigma^{-2}T(\hat{a}(k) - a(k))$ $R(\hat{a}(k) - a(k))$ with $R = \{\operatorname{cov}(Y_i, Y_j)\}_{i,j}$, which tends weakly to a $\chi_k^2$ distribution (if the true process $Y_t$ is also an AR($k$)-process), while the limit behavior of the absolute difference would depend on $A(\lambda)$. The choice of the $\mathscr{L}_2$ norm seems to be mainly for calculational convenience. However, by using the approximation $\log(\sigma/\hat{\sigma}_k) \approx (\sigma/\hat{\sigma}_k) - 1$ (or by adding the penalty term $2[(\sigma/\hat{\sigma}_k) - 1 - \log(\sigma/\hat{\sigma}_k)]$ for the innovation variance estimate to the criterion (1)) one

*R. Dahlhaus is Privatdozent, Department of Mathematics, University of Essen, Postfach 103 764, D-4300 Essen 1, West Germany.*

concludes that (1) is the same as

$$I(f, f_{\hat\theta}) := \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log \mathrm{f}_{\hat\theta}(\lambda) + \frac{f(\lambda)}{f_{\hat\theta}(\lambda)}\right\} d\lambda - \frac{1}{4\pi}\int_{-\pi}^{\pi}\{\log f(\lambda) + 1\}\, d\lambda, \tag{2}$$

where $f$ and $f_{\hat\theta}$ are the spectral densities of the original and the fitted process. $I(f, g)$ is equal to the limit of $(1/T)$ multiplied by the information divergence $\int_{-\infty}^{\infty}\{-\log[h_{g,T}(x)/h_{f,T}(x)]\}h_{f,T}(x)\,dx$, where $h_{f,T}$ is the probability density of $T$ observations from a Gaussian process with spectral density $f$ (cf. Parzen, 1983, Section 3). Thus, choosing the $\mathscr{L}_2$ norm is implicitly a Gaussian point of view.

Once an approximation criterion like (1) has been chosen, this criterion should be used to judge the quality of all estimation procedures including the selection of the order. In order to obtain optimal estimates in this sense, one could minimize the first part of (2) directly with the unknown $f$ replaced by a suitable nonparametric estimate, e.g., the periodogram $I_T(\lambda)$ with a suitable data taper applied. If $\hat\theta$ is obtained by minimizing this criterion with the selected order $k$, it follows from Sections 3 and 4 of Findley (1985) that the "expected loss" $EI(f, f_{\hat\theta})$ is equal to

$$E\left[\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log f_{\hat\theta}(\lambda) + \frac{I_T(\lambda)}{f_{\hat\theta}(\lambda)}\right\} d\lambda + \frac{k+1}{T}\right] + o(1) - \frac{1}{4\pi}\int_{-\pi}^{\pi}\{\log f(\lambda) + 1\}\, d\lambda$$

$$= \frac{1}{2}\, E\left[\log \hat\sigma^2 + 2\,\frac{k+1}{T}\right] + o(1) - \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\{2\pi f(\lambda)\}\, d\lambda.$$

(Note that for the derivation of the above result, $R(\eta)$ in (4.2) of Findley takes the value 0 if a data taper is applied.)

Thus, the expectation of the approximation criterion (1) is minimized by selecting the order $k$ from AIC and by using the Whittle estimate (which in the AR($k$)-case is equal to the Yule-Walker estimate). The same arguments, applied to the information divergence directly (instead of its limit), would lead to selecting the order from AIC and estimating the parameters from the exact Gaussian likelihood function.

As already pointed out by Hannan, I strongly recommend the use of a data taper since this leads to improved estimates in the finite sample situation (Dahlhaus, 1986).

In the light of the above remarks I would now like to make the following comments on Hannan's paper.

Judging the quality of the different order selection procedures by the above criterion (1) favors $C_T = 2$, i.e., AIC. Shibata (1980) starts with another approximation criterion, but he ends with the same result. It would be interesting to know whether there exists any criterion like (1) depending on the difference between the transfer function which favors other values of $C_T$. Although the theorem below (5.6) is important from the mathematical point of view, I doubt its relevance for practical situations. For me it doesn't make sense to consider the distance of orders (which is done by proving convergence in probability), since this distance of the orders doesn't say very much about the "nearness" of the processes themselves (an AR(2) process with moderate roots may be much "nearer" to, e.g., an AR(1000) process, than to an AR(3) process with roots close to the unit circle).

Approximation criteria of the form (1) are attractive because the variance of the parameter estimates turns out to be an implicit natural penalty term for selecting an order which is too high. Is it possible to incorporate this idea into the considerations of Section 3 by comparing the Hankel matrix with estimated parameters with the true one? Furthermore, the Hankel norm considered (the greatest singular value) corresponds to the absolute difference of the transfer functions. In view of the above discussion, I would prefer a criterion which corresponds to the relative difference of the transfer functions. Could this be achieved by considering the relative differences of the singular values? Furthermore, it would be interesting to know whether the use of the $l_2$ space in the singular value decomposition is related in some way to a Gaussian point of view (as the $\mathscr{L}_2$ norm in (1) is).

## 2. THE ESTIMATION PROCEDURE

In the practical situation one has many possible estimation procedures, each of which may be preferable under different aspects (exact Gaussian likelihood procedure, approximations to it, nonGaussian procedures, robust procedures, etc.). From a technical point of view it is often too difficult to compare them with an approximation criterion like (1), and I therefore want to look at the procedures from a different point of view.

Suppose the parameter estimate $\hat\theta_T$ is obtained by minimizing a function $L_T(\theta)$ (likelihood function, residual sum of squares, etc.). What parameter $\theta$ do we actually estimate if the underlying process is not of the fitted structure? Obviously that parameter $\theta_0$ that minimizes $EL_T(\theta)$ or, if the fitted model doesn't depend on $T$, lim $EL_T(\theta)$, provided this limit exists. It should be noted that we have an "approximation effect" in the sense that different $L_T(\theta)$ may lead to different $\theta_0$ if the underlying process is not of the fitted structure, although they lead to the same $\theta_0$ if the process is of the fitted structure. Suppose we fit an AR($k$) model. Consider, for example, the Gaussian likelihood

$$L_T(\theta) = \frac{1}{T}\log\det\Sigma_\theta + \frac{1}{T} Y'\,\Sigma_\theta^{-1}\,Y$$

and alternatively an $M$ estimate

$$L_T^*(\theta) = \frac{1}{T}\sum_{t=1}^{T} \rho\left(\sum_{s=0}^{k} a_s Y_{t-s}\right)$$

$$\text{with}\quad Y_t = 0 \quad\text{if}\quad t \le 1$$

(if $\sigma^2$ is unknown the estimate has to be modified, cf. Martin and Yohai (1985)). Then,

$$EL_T(\theta) = \frac{1}{T}\log\det\Sigma_\theta + \frac{1}{T}\operatorname{tr}\{\Sigma\Sigma_\theta^{-1}\}$$

and

$$EL_T^*(\theta) \approx E\rho\left(\sum_{s=0}^{k} a_s Y_{t-s}\right).$$

If $Y_t$ is also an AR($k$) process then both $EL_T(\theta)$ and $EL_T^*(\theta)$ are minimized by the true parameter value, while in the case where $Y_t$ is not an AR($k$) process, $EL_T(\theta)$ and $EL_T^*(\theta)$ are minimized by different values. This means that one has not only to consider the quality of the estimation procedure, but also the "quality" of the estimated parameter.

In the formula below (5.8), Hannan should not compare the estimate $\hat{\Phi}_h(j)$ with $\Phi(j)$ but with the estimated parameter $\Phi_h(j)$ (in the above sense), obtained as a solution of the theoretical counterpart of equation (5.8), and then ask in a second step how good the $\Phi_h(j)$ represent the structure of the series (in fact, the finitely many $\Phi_h(j)$, $j = 1, \cdots, h$, describe the structure of the process "better" than the finitely many $\Phi(j)$, $j = 1, \cdots, h$).

It is obvious that the choice of an estimation procedure doesn't only imply an estimated parameter $\theta_0$ but also an optimal order. The results of Shibata (1980) favoring AIC are only for the case where the parameters are estimated by the Yule-Walker equations. It would be interesting to know whether using other estimation procedures (e.g., robust ones) leads to other order criteria.

## ADDITIONAL REFERENCES

DAHLHAUS, R. (1986). Small sample effects in time series analysis. I. Preprint. University of Essen.

MARTIN, R. D. and YOHAI, V. J. (1985). Robustness in time series and estimating ARMA models. In *Handbook of Statistics* (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) **5** 119–155. North Holland, Amsterdam.

PARZEN, E. (1983). Autoregressive spectral estimation. In *Handbook of Statistics* (D. R. Brillinger and P. R. Krishnaiah, eds.) **3** 221–247. North Holland, Amsterdam.

# Comment

**Jorma Rissanen**

In this exceptionally lucid and comprehensive survey, Professor Hannan covers essentially all the important ideas in the theory of linear dynamic systems, both deterministic and stochastic, developed during the past twenty years or so. In addition, he describes the more recently introduced new statistical ideas for selecting such models for time series. I was particularly impressed by the apparent ease and elegance with which Professor Hannan managed to explain the rather intricate notions without any undue sacrifice in precision.

I would like to comment on two issues of a general nature raised by Professor Hannan. There have been several attempts to apply the beautiful and deep approximation theory of Adamyan, Arov and Krein in a statistical context for the purpose of obtaining an optimal low order model reduction. As explained in the paper, such a procedure begins with a high order dynamic system, arrived at, perhaps, by applying physical or chemical laws to a process, or by other means. This is then, in the second stage, reduced to a desired complexity, optimally in the sense of minimum distance in a certain norm. The point I wish to make is that because the initial system, which necessarily has the status of a model rather than any "true" system, is nonunique, the end result cannot be assigned any meaningful optimality property. Instead, it is just an optimal approximation of an arbitrary model of the data.

My remaining comments aim to amplify and, perhaps, modify some of the concluding remarks in

*Jorma Rissanen is a Member of the Research Staff, IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120.*