The Annals of Statistics
1986, Vol. 14, No. 2, 502–516

A BAYES PROCEDURE FOR THE IDENTIFICATION OF UNIVARIATE TIME SERIES MODELS

BY D. S. POSKITT

Australian National University

This paper is concerned with model selection in time series analysis. An identification criterion is presented that is asymptotically equivalent to a Bayes decision rule. The discussion is conducted in the context of a general class of parametric time series models and consideration is given to the special case of order determination in autoregressive moving-average representations. Consistency of the criterion is proved.

1. Introduction. Recently, attention has been directed in the time series literature to the problems associated with choosing a finite parameter model for an observed process. Much of the discussion has been concerned with the development of model selection criteria and although alternative principles and methods have been employed in the derivations the criteria are usually expressed as functions of the estimated innovation variance. See, for example, Akaike (1969), Parzen (1974), Hannan and Quinn (1979), and Shibata (1980), where the argument is conducted in the context of autoregressive processes, and Hannan (1980), Taniguchi (1980), and Hannan and Rissanen (1982), where the analysis is extended to autoregressive moving-average models. Alternatively Hosking (1980), Poskitt and Tremayne (1981), and Pötsher (1983), amongst others, have investigated conventional hypothesis testing procedures designed to test the adequacy of a chosen fitted model. The diagnostic checking devices considered may be represented as functions of the residual autocorrelations and are related to the portmanteau statistic discussed by Box and Pierce (1970). In the present paper results from decision theory are used to determine a Bayes decision rule for time series model identification. A tenuous link between the two procedures previously mentioned is thereby obtained as the selection criterion so derived can be expressed in terms of both the one-step ahead prediction error variance and the autocorrelations of the residual process. For an alternative approach to modifying existing model selection criteria see Rissanen (1983).

Autoregressive moving-average, ARMA(p, q), representations of the form

$$(1.1)\quad x(t) + \sum_{i=1}^{p} \alpha_i x(t-i) = \xi(t) + \sum_{i=1}^{q} \mu_i \xi(t-i), \qquad t = 0, \pm 1, \ldots,$$

where $\{\Xi(t)\}$ is a white noise process, provide a general and widely applied class of models for stationary time series but the availability of finite parameter models where the power spectrum is not rational, as in Bloomfield (1973), leads to a consideration of a more general problem.

Received September 1983; revised September 1985.

AMS 1980 *subject classifications.* Primary 62M10; secondary 62C10.

Key words and phrases. Time series model, power spectrum, autoregressive moving-average representation, Bayes decision rule, model selection criterion, consistency.

ASSUMPTION P. Let $\{X(t)\}$ be a discrete time nondeterministic zero mean Gaussian stochastic process with power spectrum $f(\omega) \in L^2$, the class of functions square integrable with respect to Lebesgue measure ν on $[-\pi, \pi]$.

Suppose that a model for an observed process assumed to satisfy Assumption P is characterised by a particular functional form for the power spectrum and is specified by a parametric family $M = \{g(\theta, \omega) \in L^2, \theta \in \Theta\}$ satisfying the following assumptions.

ASSUMPTION M1. The parameter space Θ is a nonempty open subset of R^d where the positive integer d is referred to as the model dimensionality. The closure of Θ, $\bar{\Theta}$, is convex and bounded.

ASSUMPTION M2. The function g is continuous on $\bar{\Theta} \times [-\pi, \pi]$, $g > 0$, and if $\theta_1 \neq \theta_2$ then $g(\theta_1, \omega) \neq g(\theta_2, \omega)$ on a set of positive Lebesgue measure.

ASSUMPTION M3. The partial derivatives $\partial g(\theta, \omega)/\partial \theta_i$, $\partial^2 g(\theta, \omega)/\partial \theta_i \, \partial \theta_j$, and $\partial^3 g(\theta, \omega)/\partial \theta_i \, \partial \theta_j \, \partial \theta_k$, $i, j, k = 1, \ldots, d$, are continuous on $\bar{\Theta} \times [-\pi, \pi]$.

The above model requirements will be referred to as Assumptions M and, together with Assumption P on the process, will be maintained throughout the paper.

In order to quantify the adequacy of a model a suitable measure of the consequences of employing a particular parametric specification is required. In the following section of the paper a frequency domain utility function that provides a measure appropriate for the observational decision problem of discriminating among a given set of alternative models on the basis of a finite realisation is discussed. The likelihood function and certain asymptotic properties of the likelihood and the Gaussian estimator of theta are also considered. In Section 3 the principle of precise measurement, Savage (1962), is pursued. Starting from a position of prior ignorance a Bayes decision rule, which for a given realisation maximises the average utility with respect to the posterior distribution of the model and its parameters, is derived and the results are specialised to the ARMA case. Prior ignorance, that is, the notion that little is known a priori relative to the information provided by the data, is represented using an invariant prior distribution as suggested by Jeffreys (1961). Such a choice is noninformative but does not result in an improper prior distribution here due, essentially, to Assumptions M. To this extent the present formulation may be thought unconventional, although it is related to the standard Bayesian procedure for model discrimination, Pericchi (1984). This procedure has been criticised in the literature as in some situations it is asymptotically inconsistent, Pericchi op. cit., and therefore the large sample behaviour of the posterior expected utility is investigated further in Section 4. Consistency of the model selection criterion is established and the implications for the practical implementation of the identification procedure discussed. The proofs of the main lemmas contained within the paper are assembled in the final section.

2. Model utility and likelihood. In the present problem an action involves choosing a model and selecting from within the parametric family a particular member. The action and state coincide with L^2. Given a set of m models, $M_i = \{g_i(\theta_i, \omega) \in L^2, \theta_i \in \Theta_i\}$, $i = 1, \ldots, m$, what constitutes a best action depends upon the extent of available knowledge concerning the true state of nature. For any $\theta \in \Theta$ let $m(\theta)$ denote the action of choosing the particular member $g(\theta, \omega)$ from the family M. For convenience, here and throughout the paper, the model subscript i, $i = 1, \ldots, m$, is omitted and generic notation employed where this raises no ambiguity. In the extreme but unrealistic situation that the true power spectrum $f(\omega)$ is known, a natural measure of the regret or loss involved in taking action $m(\theta)$ is given by the integrated squared relative error

$$\eta\{m(\theta)\} = \int_{-\pi}^{\pi} \left(\frac{f(\omega)}{g(\theta, \omega)} - 1 \right)^2 d\omega. \tag{2.1}$$

The associated utility may be taken as $U\{m(\theta)\} = \exp[-\eta\{m(\theta)\}]$. For theoretical and practical purposes, however, it is necessary to consider a rather more complicated specification of the utility function and in particular it is necessary to allow it to depend on and be modified by the observations.

Given a realisation $\mathbf{x}_T = (x(1), \ldots, x(T))'$ of T observations on the process $\{X(t)\}$ set the sample power spectrum, or periodogram,

$$I_T(\omega) = (2\pi T)^{-1} |Z_T(\omega)|^2,$$

where

$$Z_T(\omega) = \sum_{t=1}^{T} x(t) \exp(-i\omega t).$$

Evaluating $I_T(\omega)$ at the frequencies $\omega_j = 2\pi j/N$, $-(T-1) \leq j \leq (T-1)$, $N = 2T - 1$ the loss of action $m(\theta)$ can be approximated using the numeraire

$$\eta_T\{m(\theta)\} = \frac{\pi}{N} \sum_j \left\{ \frac{I_T(\omega_j)}{g(\theta, \omega_j)} \right\}^2 - \frac{4\pi}{N} \sum_j \frac{I_T(\omega_j)}{g(\theta, \omega_j)} + 2\pi. \tag{2.2}$$

LEMMA 1. *The numeraire $\eta_T\{m(\theta)\}$ converges to $\eta\{m(\theta)\}$ with probability one, uniformly for all θ in $\bar{\Theta}$.*

The implication of Lemma 1 is that as the sample size increases any departure of $\eta_T\{m(\theta)\}$ from zero reflects more a loss from using an inappropriate model than a departure of the numeraire from the theoretical but unknown regret given in (2.1). The utility associated with $m(\theta)$ will therefore be taken as

$$U_T\{m(\theta)\} = \exp[-\eta_T\{m(\theta)\}].$$

The likelihood of a model M and its associated parameter vector θ, denoted $pr(\mathbf{x}_T | M, \theta)$, is

$$\exp\left[-\tfrac{1}{2} \left(\ln \det \mathbf{\Sigma}_T(\theta) + T \ln 2\pi + \mathbf{x}_T' \mathbf{\Sigma}_T(\theta)^{-1} \mathbf{x}_T \right) \right], \tag{2.3}$$

where the variance–covariance matrix of the vector $\mathbf{x}_T$

$$\Sigma_T(\theta) = \left[\int_{-\pi}^{\pi} g(\theta, \omega) e^{i\omega(r-s)}\, d\omega\right], \qquad r, s = 1, \ldots, T.$$

In order to re-express $pr(\mathbf{x}_T|M, \theta)$ in the frequency domain consider the function

$$l_T(\theta) = \frac{1}{N}\sum_j \left(\ln 2\pi g(\theta, \omega_j) + \frac{I_T(\omega_j)}{g(\theta, \omega_j)}\right).$$

LEMMA 2. *For all* θ *in* $\bar{\Theta}$

(i) $$\left|T^{-1}\ln\det \Sigma_T(\theta) - \ln 2\pi - N^{-1}\sum_j \ln g(\theta, \omega_j)\right| \to 0$$

and

(ii) $$\left|T^{-1}\mathbf{x}_T'\Sigma_T(\theta)^{-1}\mathbf{x}_T - N^{-1}\sum_j I_T(\omega_j)/g(\theta, \omega_j)\right| \to 0$$

almost surely (a.s.) and uniformly.

An immediate corollary of Lemma 2 is that $T^{-1}|\ln pr(\mathbf{x}_T|M, \theta) + \frac{1}{2}T\ln 2\pi + \frac{1}{2}Tl_T(\theta)| \to 0$ a.s. and the asymptotic behaviour of the likelihood can be investigated via the limiting behaviour of $l_T(\theta)$.

A second consequence of Lemma 2, and its proof, is that $l_T(\theta)$ converges uniformly in θ and with probability one to the corroborant function

$$l(\theta) = (2\pi)^{-1}\int_{-\pi}^{\pi}\left(\ln 2\pi g(\theta, \omega) + \frac{f(\omega)}{g(\theta, \omega)}\right) d\omega,$$

a continuous function of θ on the compact set $\bar{\Theta}$. Let θ_0 denote a value of theta at which the infimum of $l(\theta)$ is achieved, that is $l(\theta_0) \leq l(\theta)$ for all $\theta \in \bar{\Theta}$. The vector θ_0 yields the best fitting member of the family M for the process $\{X(t)\}$. However, since $\ln y + d/y$, $d > 0$, is minimised at $y = d$,

$$l(\theta_0) \geq 1 + (2\pi)^{-1}\int_{-\pi}^{\pi} \ln 2\pi f(\omega)\, d\omega, \tag{2.4}$$

with equality if and only if $g(\theta_0, \omega) = f(\omega)$ almost everywhere (a.e.). This equality cannot be assumed to hold for any specification. For this reason θ_0 will be referred to as the pseudo true parameter for the model. Now let $\hat{\theta}_T$ be a value minimising $l_T(\theta)$; such a value exists because $l_T(\theta)$ is a continuous function defined on a compact set. Employing the nomenclature of Whittle (1962) $\hat{\theta}_T$ will be called the Gaussian estimator. The relationship between the Gaussian estimator and the pseudo true parameter and the properties of $l_T(\theta)$ described immediately below are germane to the subsequent analysis of model expected utility.

LEMMA 3. *If* θ_0 *is unique and lies in the interior of* $\bar{\Theta}$ *then* $\hat{\theta}_T$ *is a strongly consistent estimator of* θ_0.

LEMMA 4. *Let* $\mathbf{H}_T(\theta)$ *denote the hessian matrix* $\partial^2 l_T(\theta)/\partial\theta\,\partial\theta'$. *Then* $\mathbf{H}_T(\theta)$ *converges to*

$$2\mathbf{I}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{\partial^2 \ln g(\theta,\omega)}{\partial\theta\,\partial\theta'} + f(\omega)\frac{\partial^2 g(\theta,\omega)^{-1}}{\partial\theta\,\partial\theta'}\right)d\omega$$

almost surely and uniformly in $\theta \in \bar{\Theta}$.

The following addition to Lemma 4 supplementing Assumptions M proves to be necessary in Section 3.

ASSUMPTION M4. The information matrix $\mathbf{I}(\theta_0)$ is positive definite.

By Assumption M3 $\mathbf{I}(\theta)$ and $\mathbf{H}_T(\theta)$ are continuous in θ and it follows that for T sufficiently large $\mathbf{H}_T(\hat{\theta}_T)$ will also be positive definite a.s. in a neighbourhood of θ_0 because the uniform convergence of $\frac{1}{2}\mathbf{H}_T(\theta)$ to $\mathbf{I}(\theta)$ and the convergence of $\hat{\theta}_T$ to θ_0 a.s. ensure that $\frac{1}{2}\mathbf{H}_T(\hat{\theta}_T)$ is a strongly consistent estimator of $\mathbf{I}(\theta_0)$.

REMARK. If the model obtains then the pseudo true parameter point θ_0 coincides with the true value of the parameter and $f(\omega) = g(\theta_0, \omega)$ a.e. In this case $\mathbf{I}(\theta_0)$ simplifies to

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{\partial \ln g(\theta_0,\omega)}{\partial\theta}\frac{\partial \ln g(\theta_0,\omega)}{\partial\theta'}\,d\omega.$$

This corresponds to the usual expression given for the information matrix. See Hannan (1973, page 137) and the references contained therein.

3. An asymptotic Bayes decision rule. In order to proceed a specification for the prior distribution of the model and its associated parameters $\mathrm{pr}(M, \theta)$ is required. As indicated above, the prior distribution of θ given M is chosen noninformatively following Jeffreys (1961, Chapter 3) as

$$\mathrm{pr}(\theta|M)\alpha\{\det \mathbf{I}(\theta)\}^{1/2}.$$

As it is common practice to espouse the principle of parsimony when modelling time series, the model prior $\mathrm{pr}(M)$ is assumed to be proportional to $(2\pi)^{-d/2}$. This gives a prior odds ratio of approximately 2 : 5 and, reinterpreting Jeffreys (1961, Appendix B), indicates an indecisive preference for every unit decrease in model dimensionality. For some discussion of alternative model priors and their interpretation see Poskitt and Tremayne (1983). By virtue of Assumptions M

$$0 < \int_{\Theta}\{\det \mathbf{I}(\theta)\}^{1/2}\,d\theta \leq \sup_{\Theta}\{\det \mathbf{I}(\theta)\}^{1/2} v_d(\bar{\Theta}) < \infty,$$

and given that $\Sigma_i(2\pi)^{-d_i/2} < \infty$ the prior distribution,

$$\mathrm{pr}(M,\theta) = \mathrm{pr}(M)\mathrm{pr}(\theta|M)\alpha\left\{(2\pi)^{-d}\det \mathbf{I}(\theta)\right\}^{1/2},$$

may be normalised to give a proper mixed mass-density function.

The Bayes decision rule can now be constructed using the extensive method of analysis, Raiffa and Schlaiffer (1961, Chapter 1). The action $m(\theta)$ is sought which, for a given data set, maximises the posterior expected utility

$$E\left[U\{m(\theta)\}\right] = \int_{\Theta} U\{m(\theta)\}\mathrm{pr}(M,\theta)\mathrm{pr}(\mathbf{x}_T|M,\theta)\,d\theta/\mathrm{pr}(\mathbf{x}_T),$$

where

$$\mathrm{pr}(\mathbf{x}_T) = \sum_{i=1}^{m} \mathrm{pr}(M_i)\int_{\Theta_i} \mathrm{pr}(\theta|M_i)\mathrm{pr}(\mathbf{x}_T|M_i,\theta)\,d\theta.$$

This provides a principle for determining the best action in relation to the current realisation. Let

$$E_T\left[U_T\{m(\theta)\}\right]$$

$$= K\int_{\Theta} \exp\left[-\eta_T\{m(\theta)\}\right]\left\{(2\pi)^{-d}\det \mathbf{I}(\theta)\right\}^{1/2}\exp\left[-\tfrac{1}{2}Tl_T(\theta)\right]d\theta, \tag{3.1}$$

where the constant K is the reciprocal of

$$\mathrm{pr}(\mathbf{x}_T)\sum_{i=1}^{m}\int_{\Theta_i}\left\{(2\pi)^{-d_i}\det \mathbf{I}_i(\theta)\right\}^{1/2}d\theta.$$

Employing the approach of Lindley (1960) and letting $T \to \infty$, Lemmas 1 and 2 and the Arzela–Ascoli theorem imply that $|E_T[U_T\{m(\theta)\}] - E[U\{m(\theta)\}]| \to 0$ a.s. and the limiting behaviour of the posterior expected utility can therefore be ascertained from that of the integral in (3.1). Using Lemmas 3 and 4 it is possible to establish the next lemma involving the second two factors of the integrand.

LEMMA 5. *For all values of T and each $\theta \in \overline{\Theta}$ set*

$$\phi_T(\theta) = \left\{(T/2\pi)^d\det \mathbf{I}(\theta)\right\}^{1/2}\exp\left[-\tfrac{1}{2}T\{l_T(\theta) - l_T(\hat{\theta}_T)\}\right].$$

Then $\{\phi_T(\theta)\}$ forms a sequence of regular generalised functions, translated to $\hat{\theta}_T$, converging to a Dirac delta function.

Lemma 5 is based on the result that asymptotically the likelihood, which is of order T, behaves like the kernel of a d variate Gaussian density function with mean vector $\hat{\theta}_T$ and variance–covariance matrix $2\mathbf{H}_T(\hat{\theta}_T)^{-1}/T$. Consequently, as the sample size increases $\phi_T(\theta)$, which is proportional to $\exp[\frac{1}{2}Tl_T(\hat{\theta}_T)]$ times the posterior density, approximates an impulse function centred at $\hat{\theta}_T$. Therefore when the expectation

$$E_T\left[U_T\{m(\theta)\}\right] = K\exp\left[-\tfrac{1}{2}Tl_T(\hat{\theta}_T)\right]T^{-d/2}\int_{\Theta}\exp\left[-\eta_T\{m(\theta)\}\right]\phi_T(\theta)\,d\theta$$

is evaluated any values outside of an arbitrarily small neighbourhood of $\hat{\theta}_T$ make a negligible contribution to the integral. Taking logarithms and neglecting common factors then gives rise to the following theorem.

THEOREM 1. *The posterior expected utility is maximised asymptotically by selecting the model which minimises the criterion function*

$$\Delta\{m(\hat{\theta}_T)\} = \tfrac{1}{2}Tl_T(\hat{\theta}_T) + \tfrac{1}{2}d \ln T + \eta_T\{m(\hat{\theta}_T)\}.$$

It is, perhaps, worth pointing out that the first two terms of Δ coincide with the BIC criterion function associated with Rissanen (1978) and Schwarz (1978). These terms may be thought of as determining the posterior probability of a model, Poskitt and Tremayne (1983), and the final model selection is based upon a trade-off between the estimated posterior odds of the models and their relative utilities as represented by the last term.

Consider now the ARMA(p, q) model of (1.1). In order to satisfy Assumptions M the structural parameters $\alpha_1, \ldots, \alpha_p$ and $\mu_1, \ldots, \mu_q$ are assumed to belong to the subset of $\mathscr{R}^{p+q}$ defined by the requirements that the roots of $\alpha(z)$ and $\mu(z)$ lie outside the unit circle, $\alpha(z)$ and $\mu(z)$ have no common factors and α_p and μ_q are not both zero. For this model

$$g(\theta, \omega) = \frac{\sigma^2}{2\pi}\left|K(e^{-\iota\omega})\right|^2,$$

where $K(z) = \sum k_j z^j = \mu(z)/\alpha(z)$, and the scale parameter σ^2 is assumed to lie in the interval $(\delta, 1/\delta)$ for arbitrarily small $\delta > 0$. Substituting in $l_T(\theta)$, $\theta = (\sigma^2, \alpha_1, \ldots, \alpha_p, \mu_1, \ldots, \mu_q)'$ and concentrating with respect to σ^2 it is easily shown that

$$l_T(\theta) \geq \ln s^2 + N^{-1}\sum_j \ln\left|K(e^{-\iota\omega_j})\right|^2 + 1,$$

where

$$s^2 = \frac{2\pi}{N}\sum_j \frac{I_T(\omega_j)}{\left|K(e^{-\iota\omega_j})\right|^2}.$$

Of course, K and s^2 are functions of θ and although for convenience this is not shown explicitly in the notation the evaluation of these quantities at the point corresponding to the Gaussian estimator will be indicated by the use of a circumflex. The following result now follows from the above theorem after some straightforward manipulations.

COROLLARY. *The Bayes decision rule is asymptotically equivalent to choosing the* ARMA(p, q) *model that minimises the criterion function*

$$\Delta(p, q) = \tfrac{1}{2}T\left[\ln \hat{s}^2 + N^{-1}\sum_j \ln\left|\hat{K}(e^{-\iota\omega_j})\right|^2\right] + \tfrac{1}{2}(p + q)\ln T$$

$$+ \pi/N\sum_j \left\{\frac{I_T(\omega_j)}{\hat{s}^2\left|\hat{K}(e^{-\iota\omega_j})\right|^2}\right\}^2.$$

To provide a heuristic interpretation of this corollary assume that an ARMA(p_0, q_0) model obtains so that, employing an obvious notation,

$$f(\omega) = g_0(\theta_0, \omega) = \frac{\sigma_0^2}{2\pi}\left|K_0(e^{-i\omega})\right|^2 \quad \text{a.e.}$$

and the pseudo true parameter value θ_0 now corresponds to the true parameter point in the usual sense. It is then well known that $\{\Xi(t)\}$ is the innovation process associated with Wold's decomposition theorem and that

$$\sigma_0^2 = \exp\left[(2\pi)^{-1}\int_{-\pi}^{\pi} \ln 2\pi f(\omega)\, d\omega\right],$$

the variance of the minimum mean squared error one-step ahead prediction error. This implies that

$$(2\pi)^{-1}\int_{-\pi}^{\pi} \ln\left|K_0(e^{-i\omega})\right|^2 d\omega = 0.$$

Furthermore, (1.1) defines for any p and q a residual process which may not be white noise unless the model is correct, that is $p = p_0$ and $q = q_0$, and the true parameter point θ_0 is employed. Estimating the uth autocovariance of the residual process in the frequency domain by

$$\hat{C}_T(u) = \frac{2\pi}{N}\sum_j \frac{I_T(\omega_j)e^{i\omega_j u}}{\left|\hat{K}(e^{-i\omega_j})\right|^2},$$

it follows from Parseval's theorem that the last term of $\Delta(p, q)$ may be written as

$$2\pi\sum_{u=0}^{T-1}\{\hat{r}_T(u)\}^2,$$

where $\hat{r}_T(u) = \hat{C}_T(u)/\hat{C}_T(0) = \hat{C}_T(u)/\hat{s}^2$ is the uth residual autocorrelation. The components of $\Delta(p, q)$ may therefore be regarded as assessing the extent to which the theoretical relationships of the true data generating mechanism are satisfied by the best fitting member in the family, or model, under consideration.

REMARK. The last term of $\Delta(p, q)$ is equivalent to the portmanteau statistic of Box and Pierce (1970) except that multiplication by T in the usual statistic is replaced by multiplication by the constant 2π and the range of summation is extended to include autocorrelations at high lags. The statistical properties of this statistic and some evidence on its performance relative to portmanteau tests are presented in Milhøj (1981).

To summarise, in the context of ARMA models, the criterion function leads to the selection of the model that appears to provide the best compromise between predictive characteristics, the autocorrelation structure of the residuals and model dimensionality.

4. Consistency of the delta criterion. The analysis of the previous section indicates that, given a set or range of models $\{M_i,\ i = 1,\ldots, m\}$, the adoption of Bayes rule leads to the selection of the kth model M_k if

$$\Delta\{m_k(\hat{\theta}_{kT})\} = \min_{i\in\{1,\ldots,m\}} [\Delta\{m_i(\hat{\theta}_{iT})\}].$$

Should the minimum not be unique the practitioner is thought to be indifferent between those models $M_{k_j},\ j = 1,\ldots, l \leq m$ that yield the minimising value. Taking this decision rule as given its performance can be analysed by examining the sampling behaviour of the expected utility ratio

$$R_T(i, j) = \exp[-\Delta\{m_i(\hat{\theta}_{iT})\}]/\exp[-\Delta\{m_j(\hat{\theta}_{jT})\}],$$

$i, j = 1,\ldots, m,\ i \neq j$, as $T \to \infty$.

Following the previous discussion a model is said to be true, or to obtain, if the pseudo true parameter θ_0 associated with the model is such that $g(\theta_0, \omega) = f(\omega)$ a.e. and a range of models is said to encompass the data generating mechanism if there exists a specification $M_k \in \{M_i,\ i = 1,\ldots, m\}$ that is true. By virtue of transitivity, the behaviour of the decision rule when comparing a range of models is characterised by the following theorem, in which the notation $a_T \sim b_T$ is used to mean that $(a_T/b_T) \to 1$ as $T \to \infty$.

THEOREM 2. *Let $R_T(2,1)$ denote the expected utility ratio between two models $M_i = \{g_i(\theta_i, \omega) \in L^2,\ \theta_i \in \Theta_i\}$, $i = 1, 2$, with pseudo true parameter values θ_{10} and θ_{20}, respectively. If*

(i) *M_1 and M_2 do not encompass the data generating mechanism and*

$$0 < \left|\frac{f(\omega)}{g_1(\theta_{10}, \omega)} - 1\right| < \left|\frac{f(\omega)}{g_2(\theta_{20}, \omega)} - 1\right| < \delta,$$

for $\delta < 1$, uniformly in ω, or

(ii) *M_1 is true and $g_2(\theta_{20}, \omega) \neq f(\omega)$ on a set of nonzero ν measure with relative error bounded above as in* (i),

then there exists a constant $C > 0$ and a decreasing sequence $C_T \searrow C$ such that

$$R_T(2,1) \sim \exp[-C_T T] \sim \{\exp[-C]\}^T \quad a.s.$$

If

(iii) *both M_1 and M_2 obtain,*

then

$$\{R_T(2,1)\}^{1/T} \sim \left(\sqrt[T]{T}\right)^{(d_1 - d_2)/2} \sim 1 \quad a.s.$$

To prove Theorem 2 observe from Lemmas 1, 2, 3, and 6 that

$$2\Delta\{m(\hat{\theta}_T)\} = Tl(\theta_0) + d \ln T + 2\eta\{m(\theta_0)\} + o(T) \quad \text{a.s.}$$

Assuming that condition (i) or (ii) holds, substituting the expansion

$$\ln\frac{(f\omega)}{g(\theta, \omega)} = \left(\frac{f(\omega)}{g(\theta, \omega)} - 1\right) - \left(\frac{f(\omega)}{g(\theta, \omega)} - 1\right)^2 + \cdots$$

in the definition of the corroborant function and simplifying the resulting expression for the difference $2[\Delta\{m_1(\hat{\theta}_{1T})\} - \Delta\{m_2(\hat{\theta}_{2T})\}$ gives the equation

$$\frac{2 \ln R_T(2,1)}{T} = \left(1 + \frac{2}{T}\right) \frac{1}{2\pi} \int_{\pi}^{\pi} \left(\frac{f(\omega)}{g_1(\theta_{10}, \omega)} - 1 \right)^2 - \left(\frac{f(\omega)}{g_2(\theta_{20}, \omega)} - 1 \right)^2 d\omega + O(\delta^3) + o(1) \quad \text{a.s.}$$

for the logarithm of the expected utility ratio. From the properties of the regrets $\eta\{m_1(\theta_{10})\}$ and $\eta\{m_2(\theta_{20})\}$ implied by statements (i) and (ii) the first term above is negative definite and, since δ is arbitrary, there therefore exists a constant $C > 0$ and a decreasing sequence $C_T \searrow C$ such that

$$\left| T^{-1} \ln R_T(2,1) + C_T \right| \to 0 \quad \text{a.s.}$$

as $T \to \infty$, giving the required result. Similar arguments applied under condition (iii) show that

$$\left| T^{-1}(2 \ln R_T(2,1) - (d_1 - d_2) \ln T) \right| \to 0 \quad \text{a.s.},$$

which is equivalent to the statement of the theorem as $\lim \sqrt[T]{T}$ is unity.

The structure of Theorem 2 shows that if $\{X(t)\}$ emanates from an unknown specification within $\{M_i,\ i = 1, \ldots, m\}$ then, as the sample size increases, Δ will select the true model with probability one. Furthermore, since for any $a, b > 0$, $T^a = o\{\exp(bT)\}$, the theorem also lends support to the intuitively appealing notion that it is possible to distinguish a model that is true from one that is not more readily than it is to reconcile alternative parametric families both of which obtain. In the latter situation the decision rule will, for large finite T, resolve the dilemma posed by having to select one model by choosing the most parsimonious, although any preference may be blurred. Taking dimensionality d as an index of model complexity this amounts to an implementation of the principle of simplicity, Rosenkrantz (1983, Chapter 5). See also Rissanen (1983) and, for some discussion of the philosophical point that consistency is to be equated with choosing the most parsimonious true model, Atkinson (1980).

The existence of a definitive true model of finite dimensionality can of course be called into question. Shibata (1980), for example, who is concerned with mean squared prediction error and whose results have been generalised by Taniguchi (1980), explicitly assumes infinite true order, and Stone (1979) has stressed the need to consider more complex, profligate parameterisations as $T \to \infty$. Recognition of this motivates consideration of the behaviour of $R_T(2,1)$ when the model set does not encompass the data generating mechanism as in Theorem 2(i). When δ is small both M_1 and M_2 provide reasonable guides to the data generating mechanism but the first model gives a more accurate approximation to the distribution of power induced by $f(\omega)$ and, hence, to the actual structure of the process. Asymptotically, the magnitude of $R_T(2,1)$ will reflect the relative proximity of the two models and will lead to the identification of the best fitting parametric family M_1.

The implication of the foregoing discussion is that model posterior expected utilities provide a potentially useful basis for making comparisons between

alternative parametric specifications. If, when considering a range of models, the numerical value of the criterion Δ for one model is small in relation to that of another then this will indicate that the model is more appropriate for the process in hand and is to be preferred. There may be circumstances, however, where a mechanistic application of the decision rule to select a single preferred model will be undesirable as little consideration will thereby be given to the relative merits of other specifications. If the expected utility ratio is close to one for two or more models then any preference between them may be indistinct. This raises the possibility of the practitioner simultaneously entertaining a few parametric families between which she/he is essentially indifferent or employing an average model, as suggested by Akaike (1978) for example, in order to better understand the process. The question of the possible efficacy of using the criterion Δ to identify univariate time series models can only really be answered by reference to experience and experimentation however.

5. Proofs. In the sequel $\|\mathbf{A}\|$ will be used to denote the operator norm $\sup_{\|\mathbf{z}\|=1}\|\mathbf{Az}\|$ of a matrix $\mathbf{A}$ where, for any vector $\mathbf{z}$, $\|\mathbf{z}\|$ is the Euclidean norm. The letter C denotes a universal constant. Consider first the following preliminary result.

LEMMA 6. *Let*

$$\psi_{1,T}(\boldsymbol{\theta}) = (2\pi/N)\sum_{j=-T+1}^{T-1} I_T(\omega_j)h(\boldsymbol{\theta},\omega_j)$$

and

$$\psi_{2,T}(\boldsymbol{\theta}) = (2\pi/N)\sum_{j=-T+1}^{T-1}\left\{I_T(\omega_j)h(\boldsymbol{\theta},\omega_j)\right\}^2,$$

where $h(\boldsymbol{\theta},\omega)$ is a continuous real valued function on $\bar{\Theta}\times[-\pi,\pi]$ with continuous partial derivatives $\partial h(\boldsymbol{\theta},\omega)/\partial\theta_i$, $i=1,\ldots,d$.

Then

$$\psi_{1,T}(\boldsymbol{\theta}) \to \int_{-\pi}^{\pi} f(\omega)h(\boldsymbol{\theta},\omega)\,d\omega \quad a.s.$$

and

$$\psi_{2,T}(\boldsymbol{\theta}) \to 2\int_{-\pi}^{\pi}\{f(\omega)h(\boldsymbol{\theta},\omega)\}^2\,d\omega \quad a.s.$$

uniformly for all $\boldsymbol{\theta}$ in $\bar{\Theta}$.

PROOF. The limiting behaviour of smoothed periodogram values has been discussed elsewhere in the literature under weaker regularity conditions than those assumed at present. See Anderson (1971, Chapters 8 and 9) and Brillinger (1975, Chapters 4 and 5). The methods employed by these authors can be applied here to show that for any $\theta\in\bar{\Theta}$

$$\psi_{1,T}(\boldsymbol{\theta}) \to \int_{-\pi}^{\pi} f(\omega)h(\boldsymbol{\theta},\omega)\,d\omega \quad \text{a.s.}$$

as $T \to \infty$. To prove that the convergence is uniform, note that for any $\delta > 0$ and $\theta_1, \theta_2 \in \Theta$ such that $\|\theta_1 - \theta_2\| < \delta$

$$h(\theta_1, \omega) - h(\theta_2, \omega) = (\theta_1 - \theta_2)' \partial h(\bar{\theta}, \omega)/\partial\theta,$$

where $\bar{\theta} = \theta_2 + \lambda(\theta_1 - \theta_2)$, $0 < \lambda < 1$, by the mean value theorem. Hence

$$|h(\theta_1, \omega) - h(\theta_2, \omega)| < \delta \sum_{i=1}^{d} \sup \partial h(\theta, \omega)/\partial\theta_i,$$

where the supremum is taken over $[-\pi, \pi]$ and $\theta \in N_\delta(\theta_1) = \{\theta: \|\theta_1 - \theta\| < \delta\}$, δ being chosen so that $N_\delta(\theta_1) \subset \bar{\Theta}$. Consequently there exists a constant C such that

$$\psi_{1,T}(\theta_1) - \psi_{1,T}(\theta_2) < \delta \cdot C \cdot (2\pi/N) \sum_{j=-T+1}^{T-1} I_T(\omega_j)$$

and $(2\pi/N)\Sigma_j I_T(\omega_j)$ converges to $\int f(\omega)\, d\omega$ a.s. Thus, $\psi_{1,T}(\theta)$ is equicontinuous and converges uniformly to $\int f(\omega) h(\theta, \omega)\, d\omega$.

The proof that $\psi_{2,T}(\theta)$ converges as indicated proceeds along almost identical lines, c.f. Milhøj (1981, Lemma 1).

PROOF OF LEMMA 1. This is obtained at once from Lemma 6 on setting $h(\theta, \omega) = 1/g(\theta, \omega)$.

PROOF OF LEMMA 2. Part (i) of the lemma follows almost immediately from a theorem due to Grenander and Szegö (1958, Chapter 5) that concludes that

$$\lim_{T \to \infty} T^{-1} \ln \det \Sigma_T(\theta) = (2\pi)^{-1} \int_{-\pi}^{\pi} \ln 2\pi g(\theta, \omega)\, d\omega;$$

it is only necessary to observe that

$$N^{-1} \sum_{j=-T+1}^{T-1} \ln g(\theta, \omega_j) = \sum_{k=-\infty}^{\infty} \rho(\theta, kN) \to (2\pi)^{-1} \int_{-\pi}^{\pi} \ln g(\theta, \omega)\, d\omega,$$

as the coefficients

$$\rho(\theta, n) = \int_{-\pi}^{\pi} \ln g(\theta, \omega) e^{i\omega n}\, d\omega$$

decline at a geometric rate, to show convergence for a given θ. The fact that the limit is uniform in θ is not stated explicitly in Grenander and Szegö although it follows directly from the uniformity of the order relations used in their proof. The proof depends upon the approximation of $g(\theta, \omega)$ by trigonometric polynomials. However, $g(\theta, \omega)$ is a continuous function of θ and ω on $\bar{\Theta} \times [-\pi, \pi]$ and is, by assumption, bounded away from zero. Hence $g(\theta, \omega)$ may be approximated by a polynomial uniformly in θ and ω on $\bar{\Theta} \times [-\pi, \pi]$ and this completes the proof of (i).

Part (ii) may be derived similarly from the preliminary lemma, see also Hannan (1973, Lemma 4).

PROOF OF LEMMA 3. By a direct application of Lemma 6 $l_T(\theta)$ converges uniformly to $l(\theta)$ a.s. Let $\theta^\dagger$ be a limit point of the sequence $\hat{\theta}_T$. By uniform convergence and continuity, for any $\delta > 0$, $|l_T(\theta^\dagger) - l(\theta^\dagger)| < \delta/2$ and $|l_T(\hat{\theta}_T) - l_T(\theta^\dagger)| < \delta/2$ a.s. for $T > T_\delta$ and hence $l_T(\hat{\theta}_T) \to l(\theta^\dagger)$ a.s. By definition, $l(\theta_0) \leq l(\theta^\dagger)$ and $l_T(\hat{\theta}_T) \leq l_T(\theta_0)$ for all T, which implies that $l(\theta^\dagger) = l(\theta_0)$. Since l has a unique minimum at θ_0 it follows that $\theta^\dagger = \theta_0$.

PROOF OF LEMMA 4. The proof of this lemma involves arguments similar to those already employed in proving Lemmas 1 and 2. The details are omitted.

PROOF OF LEMMA 5. Consider

$$\int_\Theta \phi_T(\theta)\,d\theta = \int_{E_\varepsilon(\hat{\theta}_T)} \phi_T(\theta)\,d\theta + \int_{\Theta \setminus E_\varepsilon(\hat{\theta}_T)} \phi_T(\theta)\,d\theta$$
$$= I_1 + I_2,$$

where, for $\varepsilon > 0$ arbitrary, $E_\varepsilon(\hat{\theta}_T)$ is the elliptical neighbourhood $\{\theta\colon (\theta - \hat{\theta}_T)'\mathbf{H}_T(\hat{\theta}_T)(\theta - \hat{\theta}_T) < \varepsilon\}$. By Lemma 3 there exists a T_{δ_1} such that for arbitrary δ_1, $\hat{\theta}_T$ lies in the spherical neighbourhood $N_{\delta_1}(\theta_0)$ for all $T > T_{\delta_1}$ and for $\theta \in \Theta \setminus N_{\delta_1}(\theta_0)$ there exists a positive constant C such that $l_T(\theta) - l_T(\theta_0) > C$. Set $\delta_1 = \varepsilon/(2\|\mathbf{H}_T(\hat{\theta}_T)\|)$. Then $N_{\delta_1}(\theta_0) \subseteq N_{2\delta_1}(\hat{\theta}_T) \subseteq E_\varepsilon(\hat{\theta}_T)$ which implies that $\Theta \setminus E_\varepsilon(\hat{\theta}_T) \subseteq \Theta \setminus N_{\delta_1}(\theta_0)$ and hence

$$I_2 < \left\{\sup_\Theta \det \mathbf{I}(\theta)\right\}^{1/2} T^{d/2}\exp\left[-\tfrac{1}{2}TC\right]\nu_d(\bar{\Theta}) \to 0 \quad \text{a.s.}$$

as $T \to \infty$.

Since $\mathbf{I}(\theta)$ is a uniformly continuous function of θ, for any $\zeta > 0$ there exists a $\delta_2 > 0$ such that

$$\det \mathbf{I}(\hat{\theta}_T)(1 - \zeta) \leq \det \mathbf{I}(\theta) \leq \det \mathbf{I}(\hat{\theta}_T)(1 + \zeta)$$

for all $\theta \in N_{\delta_2}(\hat{\theta}_T)$ and by Taylor's theorem

$$Q_T(\theta)(1 - \zeta) \leq l_T(\theta) - l_T(\hat{\theta}_T) \leq Q_T(\theta)(1 + \zeta),$$

where

$$Q_T(\theta) = \tfrac{1}{2}(\theta - \hat{\theta}_T)'\mathbf{H}_T(\hat{\theta}_T)(\theta - \hat{\theta}_T).$$

If $\varepsilon = \delta_2/\|\mathbf{H}_T(\hat{\theta}_T)^{-1}\|$ then $E_\varepsilon(\hat{\theta}_T) \subseteq N_{\delta_2}(\hat{\theta}_T)$. Substituting in $\phi_T(\theta)$ and employing Lemmas 3 and 4 in conjunction with the properties of the multivariate normal density function and the incomplete gamma ratio yields

$$I_1 = (1 \pm \zeta)^{-d/2}\{1 - \gamma_T(\varepsilon(1 \pm \zeta)/2)\},$$

where, using conventional notation,

$$\gamma_T(y) = \exp(-Ty/2)\sum_{j=0}^{n-1} (Ty/2)^j/\Gamma(j + 1)$$

when $d = 2n$ and

$$\gamma_T(y) = \exp(-Ty/2) \sum_{j=0}^{n-2} (Ty/2)^{j+1/2}/\Gamma(j + 3/2) + 2\{1 - \Phi(\sqrt{T}y)\}$$

when $d = 2n - 1$, $n \leq (d + 1)/2 < n + 1$. As ζ is arbitrary it follows that $I_1 \to 1$ a.s. as $T \to \infty$. The proof is completed using standard results from the theory of generalised functions, Zemanian (1965, Theorem 2.3.2) and Gel'fand and Shilov, (1964, pages 34–39).

Acknowledgments. I would like to thank the referees for their comments. I am also grateful to an associate editor for helpful remarks and constructive suggestions on the content of the paper and for reference to the article by M. Taniguchi. These have led to corrections, alterations and, I believe, improvements in the presentation of the paper.

REFERENCES

AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247.

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30** 9–14.

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

ATKINSON, A. C. (1980). A note on the generalised information criterion for choice of a model. *Biometrika* **67** 413–418.

BLOOMFIELD, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226.

BOX, G. E. P. and PIERCE, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Statist. Assoc.* **65** 1509–1526.

BRILLINGER, D. R. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

GEL'FAND, I. M. and SHILOV, G. E. (1964). *Generalized Functions* **1**. Academic, New York.

GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and their Applications.* Univ. California Press.

HANNAN, E. J. (1973). The asymptotic theory of linear time series models. *J. Appl. Probab.* **10** 130–145.

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

HANNAN, E. J. and RISSANEN, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69** 81–94.

HOSKING, J. R. M. (1980). Lagrange-multiplier tests of time series models. *J. Roy. Statist. Soc. Ser. B* **42** 170–181.

JEFFREYS, H. (1961). *Theory of Probability.* 3rd ed. Clarendon, Oxford.

LINDLEY, D. V. (1960). The use of prior probability distributions in statistical inference and decisions. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 453–468. Univ. California Press.

MILHØJ, A. (1981). A test of fit in time series models. *Biometrika* **68** 177–187.

PARZEN, E. (1974). Some recent advances in time series modelling. *IEEE Trans. Automat. Control* **AC-19** 723–730.

PERICCHI, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika* **71** 575–586.

POSKITT, D. S. and TREMAYNE, A. R. (1981). An approach to testing linear time series models. *Ann. Statist.* **9** 974–986.

POSKITT, D. S. and TREMAYNE, A. R. (1983). On the posterior odds of time series models. *Biometrika* **70** 157–162.

PÖTSCHER, B. M. (1983). Order estimation in Arma-models by Lagrangian multiplier tests. *Ann. Statist.* **11** 872–885.

RAIFFA, H. and SCHLAIFFER, R. (1961). *Applied Statistical Decision Theory.* M.I.T. Press, Cambridge, Mass.

RISSANEN, J. (1978). Modelling by shortest data description. *Automatica—J. IFAC* **14** 467–71.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

ROSENKRANTZ, R. D. (1983). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science.* Reidel, Boston.

SAVAGE, L. J., ET AL. (1962). *The Foundations of Statistical Inference.* Methuen, London.

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

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

STONE, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc. Ser. B* **41** 276–278.

TANIGUCHI, M. (1980). On selection of the order of the spectral density model for a stationary process. *Ann. Inst. Statist. Math.* **32** 401–419.

WHITTLE, P. (1962). Gaussian estimation in stationary time series. *Bull. Inst. Internat. Statist.* **39** 105–129.

ZEMANIAN, A. H. (1965). *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions, with Applications.* McGraw-Hill, New York.

DEPARTMENT OF STATISTICS, I.A.S.,
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 2601
AUSTRALIA