J. R. Statist. Soc. B (1980),
42, *No.* 2, *pp.* 213–220

Bayes Factors and Choice Criteria for Linear Models

By A. F. M. SMITH and D. J. SPIEGELHALTER

University of Nottingham, England

[Received August 1979. Revised December 1979]

SUMMARY

Global and local Bayes factors are defined and their respective roles examined as choice criteria among alternative linear models. The global Bayes factor is seen to function, in appropriate contexts, as a fully automatic Occam's razor and to be closely related to the Schwarz model choice criterion. The local Bayes factor is shown to have a close relationship with the Akaike Information Criterion.

Keywords: BAYES FACTORS; POSTERIOR PROBABILITIES; LINEAR MODELS; INFORMATION CRITERION; MODEL CHOICE

1. INTRODUCTION

THE problem of choosing between alternative models continues to attract a good deal of theoretical attention, much of which has been stimulated by the appearance of the Akaike Information Criterion (Akaike, 1973). The AIC and its variants (see, for example, Bhansali and Downham, 1977) essentially adjust the likelihood ratio test statistic by a constant multiple of the difference in dimensionalities of the two models under consideration. Schwarz (1978) has recently proposed a fundamentally different criterion which replaces the constant multiplier of the AIC by the logarithm of the sample size. Stone (1979) has compared and contrasted these two approaches in terms of certain of their asymptotic properties, and a further contribution in this area is provided by Hannan and Quinn (1979).

In this paper, which deals with the important special case of choice between alternative nested linear models, we shall also be concerned with comparing and contrasting these two forms of model choice criteria, but from a rather different, non-asymptotic, perspective. Our starting point is not a consideration of the criteria themselves, but, instead, a discussion of the Bayesian approach to comparing alternative nested linear models on the basis of their posterior probabilities, or, equivalently, on the basis of ratios of posterior to prior odds. Stone has, in effect, argued that the comparison of choice criteria for linear models on the basis of their asymptotic properties is rather arbitrarily dependent on the assumptions made about the embedded sequence of design matrices. Our approach avoids the arbitrary asymptotics and concentrates, instead, on the way in which the dependence of the prior specification on the design matrix influences the forms of model choice criteria which arise from consideration of posterior probabilities.

It will be shown that, depending on the nature of the prior specification adopted for model parameters, two fundamentally different forms of odds ratio, or *Bayes factor*, arise. The first of these, which we shall call the *global* Bayes factor, will be discussed in Section 2, and will be shown to lead, essentially, to the Schwarz-type of criterion. The relationship of the global Bayes factor to the so-called *Lindley Paradox* (Lindley, 1957) will also be examined and, motivated by this, in Section 3 we derive what we shall call the *local* Bayes factor. This will be shown to lead to a variant of the Akaike criterion, and a comparison will then be made with various Akaike-type procedures, including those of Bhansali and Downham (1977).

Our purpose in this paper is to put forward a unified approach to the two types of criteria by studying the forms of prior specification with respect to which they emerge as, essentially, Bayes procedures. We are not concerned here with advocating any particular procedure, nor with

emphasizing aspects of the Bayes/non-Bayes controversy. The development given here for nested models can be suitably extended to the case of non-nested models, and to more general forms of prior specification than those we shall consider. We have concentrated here on the simplest possible presentation of the basic ideas.

2. Global Bayes Factors

2.1. *Global Bayes Factors and the Schwarz Criterion*

We shall assume that $M_i, i = 0, 1$, are two nested normal-linear models, $M_0 \subset M_1$, defined by

$$\mathbf{y} \sim N(\mathbf{A}_i\, \boldsymbol{\theta}_i, \sigma^2 \mathbf{I}_n), \tag{1}$$

where $\mathbf{A}_i$ is known and of full rank p_i, $\boldsymbol{\theta}_i$ is a p_i-vector of unknown parameters and σ^2 is, in general, assumed unknown. Without loss of generality, we shall further assume that $\boldsymbol{\theta}_1^{\mathrm{T}} = [\boldsymbol{\theta}_0^{\mathrm{T}} : \boldsymbol{\theta}^{\mathrm{T}}]$ and $\mathbf{A}_1 = [\mathbf{A}_0 : \mathbf{A}]$, with the columns of $\mathbf{A}$ orthogonal to those of $\mathbf{A}_0$. The Bayes factor for M_0 against M_1 is then defined by

$$B_{01} = p(\mathbf{y} \mid M_0)/p(\mathbf{y} \mid M_1), \tag{2}$$

where

$$p(\mathbf{y} \mid M_i) = \int\int p(\mathbf{y} \mid \mathbf{A}_i, \boldsymbol{\theta}_i, \sigma)\, p(\boldsymbol{\theta}_i, \sigma \mid \mathbf{A}_i)\, d\boldsymbol{\theta}_i\, d\sigma, \tag{3}$$

and the first term in the integrand is defined by (1). So far as the second term is concerned, we shall assume that, under M_1, the prior specification has the structure

$$p(\boldsymbol{\theta}_1, \sigma \mid \mathbf{A}_1) = p(\boldsymbol{\theta} \mid \mathbf{A}, \sigma)\, p(\boldsymbol{\theta}_0 \mid \mathbf{A}_0, \sigma)\, p(\sigma), \tag{4}$$

whereas, under M_0, $p(\boldsymbol{\theta}_0, \sigma \mid \mathbf{A}_0)$ is specified by the final two terms in (4). The assumption of conditional independence in (4) is in no way crucial to the general argument here, but serves to keep algebraic complications to a minimum. Some comments on the general case are included at the end of this section.

If $p(\boldsymbol{\theta} \mid \mathbf{A}, \sigma)$, $p(\boldsymbol{\theta}_0 \mid \mathbf{A}_0, \sigma)$ are now assumed to be normal densities with covariance matrices $\sigma^2 \mathbf{V}, \sigma^2 \mathbf{V}_0$, respectively, such that $\mathbf{V}^{-1}, \mathbf{V}_0^{-1}$ are small compared with $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0$, and if $p(\sigma)$ is degenerate, so that σ^2 is effectively assumed known, then, after standard integration in (3) and substitution into (2), we have

$$B_{01}(\sigma) \approx |\mathbf{V}|^{\frac{1}{2}} |\mathbf{A}^{\mathrm{T}}\mathbf{A}|^{\frac{1}{2}} \exp\{-\tfrac{1}{2}\chi^2\}, \tag{5}$$

where $\chi^2 = \hat{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} \hat{\boldsymbol{\theta}}\, \sigma^2$ is the standard test statistic for testing M_0 against M_1 for this known variance case, and $\hat{\boldsymbol{\theta}}$ is the least squares estimate of $\boldsymbol{\theta}$.

If $p(\sigma)$ is, instead, taken to be the improper limit, σ^{-1}, then we obtain

$$B_{01} \approx |\mathbf{V}|^{\frac{1}{2}} |\mathbf{A}^{\mathrm{T}}\mathbf{A}|^{\frac{1}{2}} \left[1 + \frac{(p_1 - p_0)}{(n - p_1)} F\right]^{\frac{1}{2}n}, \tag{6}$$

where F is the standard F-statistic for testing M_0 against M_1. Detailed derivation of (5) and (6) follows along the same lines as Lempers (1971, p. 35), but with additional simplification, resulting from our specific assumptions.

Again, we note that the same conclusions will hold for more general, proper, choices of $p(\sigma)$. See, for example, the forms obtained by Lempers using a gamma-type density.

We cannot, of course, say anything completely general about the behaviour of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ as the sample size, n, increases, but in many cases it might be reasonable to regard $\mathbf{A}$ as behaving like an (n/n_0)-fold replicate of an initial design matrix $\mathbf{A}^*$ corresponding to the first n_0 observations. In such cases, for fixed p_0, p_1 and values of the test statistic, both (5) and (6) have a form proportional to $(n/n_0)^{\frac{1}{2}(p_1 - p_0)}$. The idea of replication is not, of course, essential; the same conclusion follows immediately from the more general assumption that the non-zero entries of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are $O(n)$—an assumption which would seem to cover most cases of practical interest. In the

case of (6),

$$-2\log B_{01} = \lambda - \log(n)(p_1 - p_0) - \log(a), \tag{7}$$

where $a = |\mathbf{V}|\,|(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)|$, and λ is the standard likelihood ratio test statistic. The form given in (7) is, apart from the term $\log(a)$, the Schwarz model choice criterion and exemplifies Stone's (1979) comment on the relationship between what we have called global Bayes factors and Schwarz-type criteria.

In fact, if the information in the prior and the initial sample are approximately equal, then $a \approx 1$ and (7) reduces to precisely the Schwarz criterion.

As we have remarked above, the same qualitative conclusions can be reached under more general assumptions. For example, if we consider (1) with $\sigma^2 = \sigma_i^2$, and M_0 and M_1 not necessarily nested, then, with $v_i\,\phi_i/\sigma_i^2$ assumed to have a χ^2-distribution with v_i degrees of freedom, and, given σ_i^2, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_{i0}, \sigma_i^2\,\mathbf{V}_i)$, $i = 0, 1$, the densities $p(\mathbf{y} \mid M_i)$ required for (2) are given by $C(n, v_i, \phi_i)$ times

$$|\mathbf{V}_i|^{-\frac{1}{2}}\,|\mathbf{V}_i^{-1} + \mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i|^{-\frac{1}{2}}\,[v_i\,\phi_i + R_i + (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i0})^{\mathrm{T}}\mathbf{V}_{(i)}^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i0})]^{-\frac{1}{2}(n+v_i)}$$

where

$$C(n, v_i, \phi_i) = \left[(v_i\,\phi_i)^{\frac{1}{2}v_i}\,\Gamma\left(\frac{n+v_i}{2}\right)\right] \Big/ \left[\pi^{\frac{1}{2}n}\,\Gamma\left(\frac{v_i}{2}\right)\right],$$

$$\mathbf{V}_{(i)} = \mathbf{V}_i + (\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i)^{-1},$$

and R_i is the usual residual sum of squares from a least squares fit of M_i. (See Lempers, 1971, for details.)

If $\mathbf{V}_i^{-1}$ is small compared with $\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i$, and $v_0 \approx v_1 \approx 0$, so that weak prior information is being conveyed within each model, (2) is approximately equal to

$$\left(\frac{|\mathbf{V}_1|}{|\mathbf{V}_0|}\right)^{\frac{1}{2}} \left(\frac{|\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1|}{|\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0|}\right)^{\frac{1}{2}} \left(\frac{R_1}{R_0}\right)^{\frac{1}{2}n},$$

which reduces to (6) under the simplifying assumptions considered earlier in this section.

2.2. *The Lindley Paradox*

Returning again to consideration of (6), we note that, in the case of an (n/n_0)-fold replicate, the Lindley Paradox (Lindley, 1957) consists in observing that, for fixed values of the significance test statistic, the Bayes factor increases with n. For values of the test statistic which would imply rejection of M_0, we could, for suitably large n, simultaneously arrive at high support for M_0 on the basis of the Bayes factor.

To examine this situation more closely, we shall study the behaviour of the expected value of the logarithm of the Bayes factor assuming M_1 to be true, with $\boldsymbol{\theta} \neq \mathbf{0}$. In order to keep the algebra to a minimum, we shall consider (5), with $\mathbf{A}$ assumed to be an (n/n_0)-fold replicate of an initial matrix $\mathbf{A}^*$. It follows that

$$\log B_{01}(\sigma) = C + \tfrac{1}{2}(p_1 - p_0)\log(n/n_0) - \tfrac{1}{2}(n/n_0)\,\hat{\boldsymbol{\theta}}^{\mathrm{T}}(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)\,\hat{\boldsymbol{\theta}}/\sigma^2, \tag{8}$$

and, if M_1 is assumed true,

$$E(\log B_{01}(\sigma) \mid \boldsymbol{\theta}) = C + \tfrac{1}{2}(p_1 - p_0)\log(n/n_0) - \tfrac{1}{2}[(p_1 - p_0) + (n/n_0)\,\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)\,\boldsymbol{\theta}/\sigma^2], \tag{9}$$

where C is independent of n.

The formal derivative of (9) with respect to n is equal to

$$(p_1 - p_0)/(2n) - \boldsymbol{\theta}^{\mathrm{T}}(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)\,\boldsymbol{\theta}/(2n_0\,\sigma^2), \tag{10}$$

which is positive if

$$(n_0^{-1}\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)\,\boldsymbol{\theta})^{\frac{1}{2}} < (p_1 - p_0)^{\frac{1}{2}}\frac{\sigma}{\sqrt{n}}. \tag{11}$$

Regarding the left-hand side of this inequality as a measure of the "distance" between M_0 and M_1, we see that if the models are sufficiently "close", as defined by (11), then there will be a range of values of n for which we expect evidence to be increasingly supporting M_0 even though M_1 is true.

We note first that, in terms of the prior for $\boldsymbol{\theta}$ which led to this form of Bayes factor, the probability of (11) obtaining becomes very small for large n. A second comment, which should perhaps carry more weight for a non-Bayesian than would the first, is that, if we are concerned with model choice for a purpose such as prediction, then the Bayes factor is expected to choose the wrong model only in situations where it does not matter anyway. To see this, note that if we choose M_0 instead of the correct model M_1, then the fractional increase in average mean square error over a replicated initial design $\mathbf{A}_1^* = [\mathbf{A}_0^* : \mathbf{A}^*]$ is given by

$$\frac{n_0^{-1} E\{(\mathbf{y}-\mathbf{A}_0^*\,\boldsymbol{\theta}_0)^{\mathrm{T}}(\mathbf{y}-\mathbf{A}_0^*\,\boldsymbol{\theta}_0)-(\mathbf{y}-\mathbf{A}_1^*\,\boldsymbol{\theta}_1)^{\mathrm{T}}(\mathbf{y}-\mathbf{A}_1^*\,\boldsymbol{\theta}_1)\,|\,\boldsymbol{\theta}_1\}}{n_0^{-1} E\{(\mathbf{y}-\mathbf{A}_1^*\,\boldsymbol{\theta}_1)^{\mathrm{T}}(\mathbf{y}-\mathbf{A}_1^*\,\boldsymbol{\theta}_1)\,|\,\boldsymbol{\theta}_1\}} = \frac{\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{A}^*)^{\mathrm{T}}(\mathbf{A}^*)\boldsymbol{\theta}}{n_0\,\sigma^2}, \tag{12}$$

which is less than $n^{-1}(p_1 - p_0)$ if (11) holds; a negligible increase if n is large.

Atkinson (1978) has noted some particular instances of posterior probabilities favouring the wrong model and, on the basis of his simulations, has been tempted to draw strong conclusions about the "dangers" of using posterior probabilities for model comparisons. The preceding analysis makes clear that such conclusions are not justified in situations where models are being chosen for use in prediction (and the conclusion can be extended in an obvious way to other "practical" model uses, such as control). Indeed, in such cases, although based on posterior probabilities, rather than being specifically derived with respect to a suitable loss function, the Bayes factor is seen to function as a *fully automatic Occam's Razor*—cutting back to the simpler model whenever there is nothing to be lost by so doing. This property would seem particularly desirable in the rather common situation where the model M_0 is actually being used as a proxy for a neighbourhood of models defined by a small neighbourhood of $\boldsymbol{\theta} = \mathbf{0}$.

3. Local Bayes Factors

3.1. *Local Bayes Factors and the Akaike Criterion*

In response to the conclusions of the previous section, it could be argued that interest in models is not always directly related to practical usage and assessment of model choice through loss functions of the predictive type. There may be occasions when we are concerned with the "truth"; or, less metaphysically, where we have implicit loss functions which are essentially of the zero–one type. In such situations we might regard the Lindley Paradox behaviour as unfortunate and seek reassurance that it is not simply an artefact induced by a somewhat cavalier approach to prior specification.

The approach we adopt in examining an alternative prior specification may be viewed either as a genuine subjective Bayesian analysis, with respect to a particular form of prior belief, or simply as a formal analysis, intended as a theoretical *ad hoc* device for comparing M_0 with a "local" subset of the models contained within M_1.

In any case, we have shown that when $\mathbf{A}$ behaves like an (n/n_0)-fold replicate of an initial matrix $\mathbf{A}^*$, the Lindley Paradox can only arise if $\boldsymbol{\theta}$ is sufficiently close to $\mathbf{0}$, in the sense of (11); that is, if

$$\sigma^{-2}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}) < (p_1 - p_0). \tag{13}$$

From a subjective Bayesian standpoint, *if* we regard values of $\boldsymbol{\theta}$ satisfying (13) as *a priori* plausible, we should wish, under M_1, to specify a prior which assigned a non-negligible probability to the range of $\boldsymbol{\theta}$ implied by (13). Recalling the equivalence of (13) and (11), we see that a normal prior for $\boldsymbol{\theta}$ with a covariance matrix that is *fixed*, independent of $\mathbf{A}$, will assign ever decreasing probability to the range in question as n increases due to the dependence of this range on n through A. Since we wish to consider priors for which give high weight to a *local*

neighbourhood of $\boldsymbol{\theta} = \mathbf{0}$, as defined by (13), we shall refer to what emerges later as a *local* Bayes factor.

Considering first the case of known σ^2, and recalling that the upper 60 per cent point of a χ^2_r distribution is approximately equal to r (over a wide range of r), a particular prior specification attaching reasonable prior weight to (13) would be obtained by assuming $\boldsymbol{\theta}$ to be normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$. It is easy to verify that this leads to the local Bayes factor

$$B_{01}(\sigma) = 2^{\frac{1}{2}(p_1 - p_0)} \exp\{-\tfrac{1}{4}\chi^2\}, \tag{14}$$

for which the Lindley Paradox does not occur. See, also, Cox and Hinkley (1974, Section 10.5) for a related discussion in the univariate normal case.

It is clear that the qualitative behaviour of the local Bayes factor does not depend on the particular choice of a normal density, but rather on the fact that non-decreasing probability is assigned to an appropriate local neighbourhood of $\boldsymbol{\theta}$. In fact, in what follows we shall find it more convenient to use a uniform prior density over an appropriate volume in $R^{(p_1 - p_0)}$. In order to specify a uniform density supporting $100(1-\alpha)\%$ of the normal density used to derive (14), for small α, we require $p(\ \mid \mathbf{A}, \sigma) = \| R_\alpha \|^{-1}$, where $\| R_\alpha \|$ is the volume of the region R_α such that the normal density with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ when integrated over R_α yields $1-\alpha$.

It is straightforward to verify that $\| R_\alpha \| = |\mathbf{A}^{\mathrm{T}}\mathbf{A}|^{-\frac{1}{2}} \| S_\alpha \|$, where S_α denotes the corresponding region for a normal density with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}_{(p_1 - p_0)}$. The volume S_α is therefore the volume of a $(p_1 - p_0)$-dimensional sphere with radius $\sigma^2 \bar{\chi}^2_{p_1 - p_0}(\alpha)$, the latter denoting the upper $100(1-\alpha)\%$ point of a $\chi^2_{p_1 - p_0}$ distribution. Using standard formulae, we deduce that

$$p(\boldsymbol{\theta} \mid \mathbf{A}, \sigma) = |\mathbf{A}^{\mathrm{T}}\mathbf{A}|^{\frac{1}{2}} \Gamma\left(\frac{p_1 - p_0}{2} + 1\right) \left(\tfrac{1}{2}\bar{\chi}^2_{p_1 - p_0}(\alpha)\right)^{-\frac{1}{2}(p_1 - p_0)} (2\pi\sigma^2)^{-\frac{1}{2}(p_1 - p_0)}, \tag{15}$$

which is proportional to $|\sigma^{-2}\mathbf{A}^{\mathrm{T}}\mathbf{A}|^{\frac{1}{2}}$, the determinant of the Fisher information matrix, and may be seen as a special case, with specified constant of proportionality, of a Jeffreys invariant prior (Jeffreys, 1961).

The resulting local Bayes factor may be written in the form

$$B_{01}(\sigma) = \exp\{\tfrac{1}{2}(p_1 - p_0)\gamma(p_1 - p_0, \alpha)\} \exp\left\{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\hat{\boldsymbol{\theta}}\right\}, \tag{16}$$

where

$$\gamma(q, \alpha) = \log(\tfrac{1}{2}\bar{\chi}^2_q(\alpha)) - 2q^{-1}\log\Gamma(\tfrac{1}{2}q + 1). \tag{17}$$

For $0{\cdot}001 < \alpha < 0{\cdot}05$, and for $1 \leqslant q \leqslant 25$, numerical investigation has shown that $\gamma(q, \alpha)$ is well approximated by $\frac{3}{2}$. Using this approximation, (16) takes the simple form

$$B_{01}(\sigma) \approx \exp\{\tfrac{3}{4}(p_1 - p_0)\} \exp\{-\tfrac{1}{2}\chi^2\}. \tag{18}$$

In the case of unknown σ^2, if we use (15) together with the improper prior limit $p(\sigma) \propto \sigma^{-1}$, the resulting Bayes factor is easily seen to be

$$B_{01} \approx \exp\{\tfrac{3}{4}(p_1 - p_0)\}\left[1 + \frac{(p_1 - p_0}{(n - p_1)}F\right]^{\frac{1}{2}n} \tag{19}$$

and

$$-2\log B_{01} = \lambda - \tfrac{3}{2}(p_1 - p_0), \tag{20}$$

where λ is again the standard likelihood ratio statistic. The form given in (20) is precisely that of an Akaike-type criterion, but with $\frac{3}{2}$ in place of the multiplier 2 originally advocated by Akaike.

An alternative, entropy-based, approach proposed by Box and Kanemasu (1973) also leads to a similar form. The idea is that, for $i = 0, 1$, we adopt the prior form

$$p(\boldsymbol{\theta}_i \mid \mathbf{A}_i, \sigma) = c_i(2\pi\sigma^2)^{-\frac{1}{2}p_i}, \tag{21}$$

where c_i does not depend on $\boldsymbol{\theta}_i$ or σ (but may depend on $\mathbf{A}_i$), and then choose the c_i so that the expected change in entropy should be the same for both M_0 and M_1. It is easy to see that if

$$I_n^{(i)} = -\int \log p_n(\boldsymbol{\theta}_i) \cdot p_n(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i \tag{22}$$

denotes the entropy measure corresponding to $p_n(\boldsymbol{\theta}_i)$, the posterior distribution for $\boldsymbol{\theta}_i$ given n observations, then

$$E\{I_n^{(i)} - I_0^{(i)}\} = -\log\left\{\frac{|\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i|^{\frac{1}{2}}}{(2\pi\sigma^2 e)^{\frac{1}{2}p_i}}\right\} + \log c_i. \tag{23}$$

It follows from this approach that we should take $c_i = |\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i|^{\frac{1}{2}} \exp\{-\frac{1}{2}p_i\}$, and if we again use $p(\sigma) \propto \sigma^{-1}$ in conjunction with (21), after some algebra it can be shown that the resulting Bayes factor is identical to (20), except that the multiplier $\frac{3}{2}$ is replaced by unity.

2.2. *Comparison of Model Choice Criteria*

If we consider the general form

$$\Lambda(m) = \lambda - m(p_1 - p_0), \tag{24}$$

a number of suggested criteria for model choice can conveniently be regarded as special cases. The Akaike (1973) AIC criterion corresponds to $m = 2$ and is asymptotically equivalent to a cross-validation criterion proposed by Stone (1977). The version of the local Bayes factor given by (20) corresponds to $m = \frac{3}{2}$. The GLIM goodness-of-fit procedure based on plotting deviance against degrees of freedom corresponds to $m = 1$ (Nelder and Wedderburn, 1974, Section 2.4), as does the entropy-based criterion proposed by Box and Kanemasu (1973). Straightforward use of the likelihood ratio statistic corresponds to $m = 0$, as does, essentially, the empirically weighted Bayes factor of Atkinson (1978, equation (16)). Also, as we saw in (7), the global Bayes factor corresponds to $m = \log\{na^{1/(p_1 - p_0)}\}$, a generalization of the criterion proposed by Schwarz (1978).

To provide a unified framework within which to consider other values of m considered in the literature, we recall from the discussion preceding (14) that the development of Section 2.1 was based on taking a normal prior for $\boldsymbol{\theta}$ whose covariance matrix, given σ^2, had the form $\sigma^2(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$. In the simple case, where M_0 corresponds to the null hypothesis $\theta = 0$ in a univariate normal framework, this implies a prior variance for θ, under M_1, shrinking with n, at the rate n^{-1}. The same rate of shrinkage is implied in the general case if we assume the elements of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ to be $O(n)$.

As a formal generalization of this, we could consider a prior for $\boldsymbol{\theta}$, given σ^2, with covariance matrix $\sigma^2\,\rho(n)(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$, so that the choice of $\rho(n)$ determines the rate of "shrinkage" of the prior covariance matrix.

It is straightforward to verify, generalizing the arguments of the last section, that minus twice the logarithm of the approximate Bayes factor obtained from such a prior leads to the form (24), with $m = \frac{3}{2} + \log[\rho(n)]$. Any constant choice for $\rho(n)$ leads to an Akaike-type result, so that, for example, $\rho(n) = e^0, e^{\frac{1}{2}}, e^{3/2}, e^{5/2}$ give rise, respectively, to (20) above, the Akaike criterion and the values $m = 3$, $m = 4$ considered by Bhansali and Downham (1977). Viewed in this light, the particular choices correspond to degrees of damping down the prior ordinates for local alternatives to $\boldsymbol{\theta} = \mathbf{0}$; as $\rho(n)$ increases, therefore, the simpler (null) model receives, implicitly, higher support. If the elements of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are effectively $O(n)$ for large n, then $\rho(n) = n$ corresponds, essentially, to taking a *fixed* prior, whose variance does not shink with n. In this case, $m \approx \log(n)$ and we obtain the Schwarz-type of criterion. Other rates of "shrinkage" of the

prior intermediate between $\rho(n) = \text{constant}$ and $\rho(n) = n$ could, of course, be considered, and the choice $\rho(n) = \log(n)$, leading to $m \approx \log\log(n)$ leads to the criterion studied by Hannan and Quinn (1979).

Noting that, for large n, $\lambda \sim \chi^2_{p_1 - p_0}$ under M_0, for all $\boldsymbol{\theta}_0$, we see, from (24), that, under M_0,

$$E(\Lambda(m)) \approx (1-m)(p_1 - p_0). \tag{25}$$

It is reasonable to require (25) to be non-positive, so that we should require $m \geqslant 1$. Generally speaking, values of $m < 1$ tend to favour complex models unduly. In order to explore further the

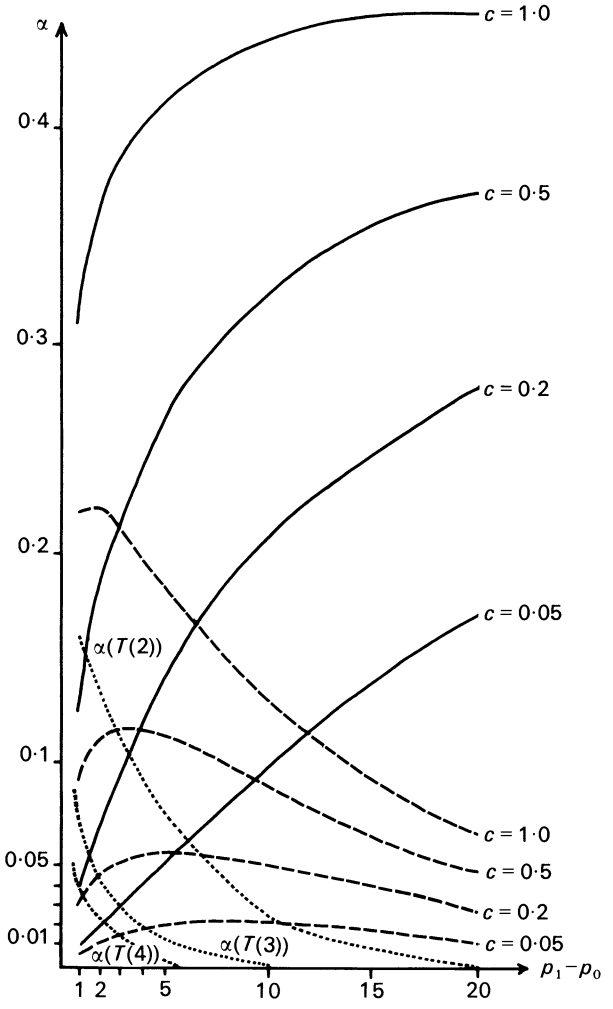

FIG. 1. Size of Bayes tests. – – –, $\alpha(T_c(\frac{3}{2}))$; ——, $\alpha(T_c(1))$; ....., $\alpha(T(r))$ $(r = 2, 3, 4)$.

behaviour of $\Lambda(1), \Lambda(\frac{3}{2})$ and $\Lambda(r)$, $r = 2, 3, 4$, we shall make use of the results of Section 2.1 showing that for $m = 1$, $m = \frac{3}{2}$ we have $\Lambda(m) = -2\log B_{01}$, for suitable local Bayes factor B_{01}.

If the Bayes factor B_{01} were to be used as a test statistic, we should reject M_0 if and only if $B_{01} < c$, where c is some function of the losses and the prior probabilities specified for M_0 and M_1. In particular, it may happen that $c \leqslant 1$, indicating a "preference" for the simple model as a result of either prior beliefs or the loss structure, or both. For values of $c \leqslant 1$, we shall illustrate

the sampling behaviour of $\Lambda(1)$ and $\Lambda(\frac{3}{2})$ when used as test statistics. We shall denote by $T_c(m)$ the test which rejects M_0 when $\Lambda(m) > -2\log c$, and denote the size of the test by

$$\alpha(T_c(m)) = P[\Lambda(m) > -2\log c \mid M_0] \approx P[\chi^2_{p_1-p_0} > m(p_1-p_0) - 2\log c] \tag{26}$$

for large n.

The test corresponding to standard use of the Akaike criterion will be denoted by $T(2)$, and rejects M_0 when $\Lambda(2) > 0$. It follows that

$$\alpha(T(2)) = P[\chi^2_{p_1-p_0} > 2(p_1-p_0)] \tag{27}$$

Fig. 1 displays the size of these tests for $c = 0{\cdot}05, 0{\cdot}2, 0{\cdot}5$ and $1{\cdot}0$.

The preference shown by $\Lambda(1)$ for over-complex models is clearly revealed. The Akaike $\Lambda(2)$ criterion is seen to have decreasing size as $(p_1 - p_0)$ increases, whereas $\Lambda(\frac{3}{2})$ is fairly stable, particularly for small c. We note, in particular, that for $(p_1 - p_0) < 20$, the Bayes test based on (20) and with $c = 0{\cdot}2$, i.e. rejecting M_0 when the Bayes factor against M_0 exceeds 5, is approximately equivalent to a 5 per cent test.

Tests based on $m = 3$, $m = 4$ have rapidly decreasing size as $(p_1 - p_0)$ increases and $\alpha(T(3))$, $\alpha(T(4))$, corresponding to (27) with $m = 3,4$ instead of $m = 2$, are also shown in Fig. 1.

In conclusion, we should like to stress, once again, that we have been concerned, in a neutral manner, with attempting to view a number of existing criteria within a unified framework, by considering the nature of the prior specification corresponding to which these criteria could be derived in an (approximately) Bayesian manner. From this point of view, it seems to us that preference for one or other of these criteria in a given context should depend on which of the prior specifications is thought most appropriate. If a utility structure is admitted, it may well be that the appropriate criterion is entirely different from *any* of the above.

ACKNOWLEDGEMENTS

Much of this material is contained in the second author's Ph.D. thesis written at University College London under the supervision of the first author, and supported by an SRC Studentship. Referees' comments on an earlier version of this paper were most helpful.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, pp. 267–281. Budapest : Akademia Kiado.

ATKINSON, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.

BHANSALI, R. J. and DOWNHAM, D. Y. (1977). Some properties of the order of an autoregression model selected by a generalization of Akaike's EPF criterion. *Biometrika*, **64**, 547–551.

BOX, G. E. P. and KANEMASU, H. (1973). Posterior probabilities of candidate models in model discrimination. Technical Report 322, University of Wisconsin.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. London : Chapman and Hall.

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

JEFFREYS, H. (1939/61). *Theory of Probability*. Oxford: Oxford University Press. (1st and 3rd editions.)

LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*. Rotterdam : University Press.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

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

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **39**, 44–47.

—— (1979). Comments on model selection criteria of Akaike and Schwarz. *J. R. Statist. Soc.* B, **41**, 276–278.