Estimation and Inference by Compact Coding

C. S. Wallace, P. R. Freeman

Journal of the Royal Statistical Society. Series B (Methodological), Volume 49, Issue 3 (1987), 240-265.

J. R. Statist. Soc. B (1987)
49, *No.* 3, *pp.* 240–265

Estimation and Inference by Compact Coding

By C. S. WALLACE† and P. R. FREEMAN‡

Monash University *University of Leicester*

Read before the Royal Statistical Society *at a meeting organised by the* Research Section *on Wednesday, March 4th, 1987,* Professor A. F. M. Smith *in the Chair*]

SUMMARY

The systematic variation within a set of data, as represented by a usual statistical model, may be used to encode the data in a more compact form than would be possible if they were considered to be purely random. The encoded form has two parts. The first states the inferred estimates of the unknown parameters in the model, the second states the data using an optimal code based on the data probability distribution implied by those parameter estimates. Choosing the model and the estimates that give the most compact coding leads to an interesting general inference procedure. In its strict form it has great generality and several nice properties but is computationally infeasible. An approximate form is developed and its relation to other methods is explored.

Keywords: OPTIMAL ENCODING; MODEL DISCRIMINATION; ESTIMATION; DATA TRANSMISSION; MINIMUM MESSAGE LENGTH; BAYESIAN INFERENCE; PRIOR DISTRIBUTIONS; INFORMATION THEORY; PRECISION

1. INTRODUCTION

Over the past twenty years or so several authors have, apparently independently, suggested that the ideas of optimal coding theory offer a general approach to the problem of inductive statistical inference. Their work has, however, tended to appear in journals of computing, engineering or information theory. These papers have either been of a very general nature or have described particular applications rather outside the normal concerns of statisticians. These factors, combined with the sheer unfamiliarity of the language in which coding theory is naturally expressed, have served to inhibit discussion of the concepts involved within the statistical community. The purpose of this paper is to bridge this communication gap, to present the concepts and the practical method of Bayesian inference to which they lead and to begin to explore their general properties and relations to other approaches.

We need to start by thinking of a set of data as a string of symbols taken from a finite alphabet (usually the digits 0 to 9, with the decimal point, plus and minus signs, etc.). If we know *a priori* that the data possess some systematic pattern, then standard techniques derived from Shannon's theory of information may be used to encode the data more briefly than if they were thought to be purely random. Similarly, if we know only the form of the pattern as specified, for example, by a conventional statistical model involving unknown parameters, briefer encoding may still be possible. We may first estimate the parameters and then encode the data under the assumption that these are the true values. The encoded string must now, however, contain a specification of the estimated values. Any model is, therefore, only worth considering if the shortening of the encoded data string achieved by adopting it more than compensates for the lengthening caused by having to quote estimated parameter values. We thus naturally arrive at a very simple trade-off between the complexity of a model and its goodness of fit. A more complicated model will usually fit a data-set better than a simpler one, enabling a briefer encoding of the data, but this must be paid for by the cost of the greater number of parameter estimates. Within a given model, the preferred parameter estimates will be those that lead to the shortest total encoded length. Similarly, the preferred model amongst a class of competing models will be the one with the shortest total length, minimised with respect to its parameter estimates.

† *Present address*: Department of Computer Science, Monash University.
‡ *Address for correspondence*: Dept of Mathematics, The University, Leicester LE1 7RH, UK.

 0035–9246/87/49240 $2.00

The concept extends easily to cases where no potential model is contemplated. Inspection of the data may allow a pattern to be conjectured and exploited to encode the data more efficiently. The code must then, however, contain a specification of the inferred pattern.

To the best of our knowledge, the first exposition of this concept as a general principle of inductive inference is due to Solomonoff (1964). The work of Kolmogorov (1965) and Chaitin (1966) on the algorithmic complexity of strings hints at the idea of measuring the "goodness" of a hypothesis by the length of the string which encodes the data according to that hypothesis. In particular, they give a theoretical basis for supposing that brief encoding of data entails the discovery and description of non-random features of the data.

Other independent expositions of the same, or very similar, ideas are in a series of applications by Wallace and co-workers, starting in 1968, in a series of papers by Rissanen since 1976 and in a non-statistical context by Maciejowski (1978, 1979, 1980).

Since the generality of the concept applies far outside the usual realms of statistical concern, most applications have been to problems not usually thought amenable to statistical analysis. Wallace's initial concern was with the problem of classification, and papers by Wallace and Boulton (1968), Boulton and Wallace (1970, 1973), Boulton (1975), Wallace (1984) develop the minimum message length solution, embodied in a computer program called SNOB. The geometry of megalithic stone circles was considered by Patrick (1978) and Patrick and Wallace (1982). Georgeff and Wallace (1983) have applied the technique to the analysis of grammars and Patrick (1986) develops alternatives to Halstead's measures for the length of a computer program. Rissanen's applications (1978, 1980, 1982, 1983) have mainly been to time series and system identification using the concept of a "universal" prior distribution over the set of positive integers to represent complete prior ignorance.

Throughout this paper we take the orthodox Bayesian view of the existence of known, proper prior distributions for unknown quantities. It has been known since Jeffreys (1939) that using improper priors for model discrimination leads to indeterminate or nonsensical answers, so there can be no substitute for careful specification of whatever prior knowledge is available.

2. OPTIMAL CODES

We briefly outline some elementary results in coding theory that will be needed later. Full details and background may be found in Shannon and Weaver (1959) and Kullback (1959).

Consider a discrete random variable X with distribution

$$p_i = P(X = v_i) \quad i = 1, 2, \ldots .$$

Let s_i be a binary string used as a code word for the value v_i. If the code is to be decipherable we must require that no two values share the same code word and that no code word should itself be the prefix of another.

An optimal code is achieved by encoding v_i to a word of length $-\log p_i$, giving the minimum expected word length of $-\Sigma_i\, p_i \log p_i$. Non-integral word lengths have to be rounded up but special techniques are available to limit the resulting increase in the total length of a concatenated message. The effect will be ignored.

The base of the logarithms should be the number of distinct symbols in the coding alphabet. For convenience, we will use natural logarithms throughout.

If a sequence of independent values of X is encoded by concatenating the optimal code words for the individual values, the result is an optimal code for the sequence. If such a sequence is extended to an infinite number of values then the optimal code forms a completely random process, in that each symbol of the string is equally likely to be any symbol of the alphabet independent of all other symbols. More generally, if the outcomes of any random process are encoded using a code that is optimal for that process, the resulting binary string forms a completely random process.

The values of an ordered pair of random variables (X, Y) may be encoded by concatenating two strings. The first is the optimal code for the value of X using the marginal distribution

of X, while the second is the optimal code for the value of Y using its conditional distribution given the observed value of X.

3. ESTIMATION AS A CODING PROCESS

From the outset we take the view that data are always recorded with finite accuracy, so we need only consider the countable set $\mathscr{H}$ of possible values of a random variable X. Let $f(x; \theta)$ denote a usual statistical model, with $x \in \mathscr{H}$ and $\theta \in \Theta$, the parameter space. We assume a known prior distribution $h(\theta)$ over Θ, giving the marginal prior distribution of X as

$$r(x) = \int f(x; \theta)\, h(\theta) d\theta$$

An estimator is a mapping of $\mathscr{H}$ into Θ written $\hat{\theta} = m(x)$. The countable set of possible estimate values is denoted

$$\Theta^* = \{\hat{\theta} : x \in \mathscr{H} \quad \text{and} \quad m(x) = \hat{\theta}\}.$$

The prior probability of getting an estimated value $\hat{\theta}$ for θ is

$$q(\hat{\theta}) = \sum_{x:m(x)=\hat{\theta}} r(x).$$

Now consider constructing a code for an observed value of X. The optimal way would be to use the prior marginal distribution $r(x)$, giving expected length

$$I_0 = -\sum_{x \in \mathscr{H}} r(x) \log r(x)$$

However, in practice such a code would be very difficult to construct. For example, when x comprises n independent observations $y_1, \ldots, y_n$ from some distribution, so that

$$f(x; \theta) = \prod_{i=1}^{n} g(y_i; \theta),$$

x cannot be coded by concatenating the code words for the y's since these are not independent when θ is unknown. Instead we consider a code word of two parts, the first encoding an estimate $\hat{\theta}$ of θ and the second giving the observed value of X optimally encoded using the distribution $f(x; \hat{\theta})$. Note that this second part now *can* be a concatenation, with each y_i encoded using $g(y_i; \hat{\theta})$.

Using optimal encoding, the length of the first part of the word is simply $-\log q(\hat{\theta})$ and that of the second part is $-\log f(x; \hat{\theta})$ since each value is encoded using its appropriate probability distribution. The expected message length is thus

$$\sum_{x \in \mathscr{H}} r(x)[-\log q(m(x)) - \log f(x; m(x))]$$

$$= -\sum_{\hat{\theta} \in \Theta^*} q(\hat{\theta}) \log q(\hat{\theta}) - \sum_{\hat{\theta} \in \Theta^*} \sum_{x:m(x)=\hat{\theta}} r(x) \log f(x; \hat{\theta})$$

The estimator $m_1(\cdot)$ which minimizes this is called the *strict minimum message length* (SMML) estimator, and the resulting minimum length is denoted by I_1. Note that $m_1(x)$ is that element $\hat{\theta}$ of Θ^* which maximizes $q(\hat{\theta})\, f(x; \hat{\theta})$ for fixed θ^*. This property is, however, deceptively simple and SMML estimators are by no means easy to calculate, since $q(\hat{\theta})$ itself depends on the estimator being considered. Wallace and Boulton (1975) give results for binomial and uniform distributions. As an example, Table 1 shows the SMML for 30 binomial trials with prior $h(\theta) = 2\theta$, $0 < \theta < 1$

The average word length I_1 is 3.408 as compared to $I_0 = 3.256$. Such an estimator is very different from ones that statisticians normally consider. Some general properties given in the next section may help to show that SMML estimators do have some advantages over usual ones.

TABLE 1

No. of successes	*Estimate of θ*
0–5	0.111
6–14	0.354
15–23	0.644
24–29	0.887
30	1.000

4. PROPERTIES OF SMML ESTIMATORS

4.1. *Generality*

The existence of an SMML estimator requires only that the integrals $r(x)$ exist for all $x \in \mathscr{H}$ and lead to a proper marginal distribution for X.

By taking Θ as the union of two or more spaces of different dimensionality, the whole minimum message length approach can be seen to embrace hypothesis testing and model description as well as simple parameter estimation. The prior $h(\theta)$ must, of course, be specified appropriately. For example, if Θ is the union of two hypotheses, with H_1 involving one unknown parameter α and H_2 involving two parameters β, γ then we need prior probabilities on H_1 and H_2 and prior densities $h(\alpha)$, $h(\beta, \gamma)$ within each hypothesis.

4.2. *Invariance*

If ϕ is any 1–1 function of θ, $\phi = g(\theta)$ say, then the SMML estimates $\hat{\phi}$ and $\hat{\theta}$ of ϕ and θ are related by

$$\hat{\phi} = g(\hat{\theta}),$$

provided that the prior for ϕ is derived from that for θ in the usual way. This property is not, of course, shared by the posterior mode or posterior mean.

4.3. *Sufficiency*

If $v(x)$ is a sufficient statistic, then the SMML estimator is a function of it.

Using the factorisation

$$f(x;\ \theta) = g(v(x),\ \theta) \cdot k(x)$$

it is easily seen that for each x, $m_1(x)$ is that $\hat{\theta} \in \Theta^*$ that maximises

$$q(\hat{\theta})\ g(v(x),\ \hat{\theta})\ k(x)$$

so that all $x \in \mathscr{H}$ yielding the same value of $v(x)$ lead to the same estimate.

4.4. *Precision*

The minimisation of the expected length of the code for X involves a compromise. The second part of each code word has length $-\log f(x; m(x))$ and would be minimised by choosing $m(x)$ to be the maximum likelihood estimate. This would, however, involve encoding $\hat{\theta}$ to many significant digits, the number of possible estimate values in Θ^* would be large and the first part of each code word, which encodes the estimate, would become long. If, on the other hand, Θ^* contains only a small number of widely-separated values, the first part of the code would be short but the large difference between the maximum likelihood estimate and its "nearest" value within Θ^* would lengthen the encoding of the second part of the word.

The compromise effected by the SMML estimator is such that values in Θ^* are separated by distances of the same order as the expected error in estimating θ. That is, the SMML estimate differs from the maximum likelihood estimate by about as much as the true value itself differs from the maximum likelihood estimate. This point will be elaborated in Section 5.3.

5. A USABLE ESTIMATOR

The minimisation of expected length in the SMML method requires that all possible values of X be considered, and the complete function $m_1(x)$ and its range Θ^* be constructed. The resulting estimator is a discontinuous function of x, and remains so even if the discretisation of the data is made arbitrarily fine. As we shall see, minimisation of the message length requires that the elements of Θ^* in any region of Θ be spaced at intervals depending on the local properties of $f(x;\theta)$. Maintenance of the optimum spacing throughout Θ means that the precise location of an element of Θ^* may be affected by the behaviour of $f(x;\theta)$ and $h(\theta)$ at distant points of Θ. These facts, together with the computational difficulty of constructing $m_1(\cdot)$, make the SMML method unsuitable for serious consideration as a general method of estimation.

The difficulties arise because in the SMML approach, a complete code is constructed capable of optimally encoding any data. We now discuss an estimation method which is much easier to apply but retains the essential features of SMML. The new method, which we call MML, is still concerned with the efficient encoding of the observed data and again results in a message which states an estimate of θ to only a limited precision. However, we no longer attempt to construct the complete code for all possible data, and simply abstract from the SMML method the essential notion of the precision of the estimate, i.e. the spacing of elements of Θ^*.

5.1. *A single scalar parameter*

Assume for the moment that θ is a single scalar parameter, and consider the following problem: given some observed data-value x, choose an estimate $\theta' \in \Theta$ and a precision quantum s so that if x is encoded as a message which states θ' with precision s, and then states x using a code optimised assuming θ equals the stated estimate, the length of the message is minimised.

The best choice of s and, of course, θ' depends on the given data x. By saying that the message states θ' with precision s, we mean that the message gives the estimate to a limited number of decimal (or binary) places. That is, it states a value $\hat{\theta}$ obtained from θ' by selecting a value from a quantized scale in which adjacent values differ by s. Hence, $|\hat{\theta} - \theta'| \leqslant s/2$.

Note that the problem just posed does not require or allow the detailed encoding of the estimate to be specified. We ask only for a 'target' value θ' and a quantum s. Thus the exact effect of replacing the target estimate θ' by the quantized value $\hat{\theta}$ cannot be predicted, and the message length can be minimised only in expectation. We assume that the expected effect of the quantization is that

$$E(\theta' - \hat{\theta}) = 0 \qquad \text{(The quantization is unbiased)}$$

$$E(\theta' - \hat{\theta})^2 = s^2/12 \qquad \text{(As for a uniform distribution)}$$

The encoding of the estimate cannot be based on the exact probability that each estimate value will be used, as is done in the SMML approach, since these probabilities depend on the details of the coding which are not here specified.

Rather, we note that the prior probability that θ lies within $\pm s/2$ of a quantized value $\hat{\theta}$ is approximately $sh(\hat{\theta})$. Basing the encoding of estimates on this probability, the expected length of the first part stating θ' to precision s is $-\log sh(\theta')$. The length of the second part is

$$-\log f(x;\hat{\theta}) = -\log f(x;\theta') - (\theta' - \hat{\theta})\frac{\partial}{\partial\theta}\log f(x;\theta') - \tfrac{1}{2}(\theta' - \hat{\theta})^2 \frac{\partial^2}{\partial\theta^2}\log f(x;\theta') + \text{etc.}$$

Using our expectation of the effects of quantization, the total length is expected to be (to second order)

$$-\log sh(\theta') - \log f(x;\theta') - \frac{s^2}{24}\frac{\partial^2}{\partial\theta^2}\log f(x;\theta')$$

which is minimized by choosing

$$s^2 = \frac{-12}{\frac{\partial^2}{\partial\theta^2}\log f(x;\,\theta')}$$

Writing

$$I(x,\,\theta') = -\frac{\partial^2}{\partial\theta^2}\log f(x;\,\theta'),$$

the expected message length is then

$$-\log h(\theta') + \tfrac{1}{2}\log\frac{I(x,\,\theta')}{12} - \log f(x;\,\theta') + \tfrac{1}{2}$$

The value θ' which minimises this is the MML estimate.

The message constructed by choosing θ' and s, then stating a quantized value $\hat{\theta}$, is not necessarily intelligible as it stands. In general, the receiver of the message cannot interpret a quantized estimate of θ unless he knows the precision quantum s. In general, therefore, the message must have three rather than two parts. The first states the quantum s, the second states the quantized estimate, and the third states the data. It might seem that we have an infinite regress, since we must now consider the encoding of s, and in particular the precision of that encoding. However, it is easily shown that the optimal precision of $\log s$ is $\sqrt{6}$, so that the first part, stating s, need encode it only to within a factor of 10 or so. Thus the expected length of the first part is small, of order one or so, but the actual length can be greater for exceptional values of X. Since the precision required in stating s is a function only of s itself, a code can be devised for the first part which is immediately intelligible.

In fact, it is often the case that no explicit first part is needed. If there exists a single statistic sufficient for θ, then both s and θ' are functions of it. If both functions are monotonic, s can be expressed as a function of θ' not otherwise involving x. A code can then be devised for the estimate which does not require a 'precision' preamble, since there is only one set of possible truncated estimates, and this set can be constructed without reference to the data.

More generally, even when no single sufficient statistic exists, it will often be possible to encode the estimate with a precision quantum given by $s^2 = 12/I(\theta')$ rather than $12/I(x,\,\theta')$ where $I(\theta')$ is the expectation of $I(x,\,\theta')$, that is, the Fisher information. The very broad minimum of the message length with respect to s implies that the above compromise quantum will be reasonably efficient for most data values. Since the compromise expression is a function of θ' only, no first part is needed.

Our MML estimator becomes the value of θ that maximises

$$\frac{h(\theta)f(x;\,\theta)}{\sqrt{I(\theta)}}.$$

This again has interesting similarity to the posterior mode, but with the advantage of invariance under 1–1 transformations. For all location problems in which the likelihood is asymptotically normal around its maximum, $I(\theta)$ will be constant and the estimates will be the same. In other cases, the MML looks for broader peaks of the posterior distribution and essentially chooses the local posterior mode with greatest probability content rather than simply the highest one.

Example: To estimate the probability of success θ in a Bernoulli experiment which has given m successes and n failures, $N = m + n$. Take

$$h(\theta) = \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)}\,\theta^a(1-\theta)^b.$$

Since $I(\theta) = N/\{\theta(1-\theta)\}$, $f(m;\theta) = \binom{N}{m}\theta^m(1-\theta)^n$, then $s(\theta) = \sqrt{\{12\theta(1-\theta)/N\}}$ and θ' is the value of θ that maximises $\theta^{a+m+\frac{1}{2}}(1-\theta)^{b+n+\frac{1}{2}}$ giving

$$\theta' = \frac{a+m+\frac{1}{2}}{a+b+N+1}.$$

The precision to which θ' is quoted is highest near 0 and 1, and increases as N, the total number of trials, increases.

5.2. *Two parameters*

Suppose $\theta = (\alpha, \beta)$.

In certain cases, the components α and β can be so ordered that the scalar MML approach can be applied to each component in turn. The code word for x now has three parts. The first states $\hat{\alpha}$, and the second $\hat{\beta}$ and the third x encoded according to $f(x; \hat{\alpha}, \hat{\beta})$. Once $\hat{\alpha}$ has been stated, the problem of choosing the target value β' and precision quantum s_β reduces to the single parameter case, giving

$$s_\beta^2 = \frac{12}{I(x, \beta' \mid \hat{\alpha})}$$

and β' minimises

$$-\log h(\beta' \mid \hat{\alpha}) + \tfrac{1}{2}\log\frac{I(x, \beta' \mid \hat{\alpha})}{12} - \log f(x; \hat{\alpha}, \beta') + \tfrac{1}{2},$$

the expected length of the second and third parts. The results are, of course, functions of the stated $\hat{\alpha}$. Denote by $L(x; \hat{\alpha})$ the resulting minimised expected length of the second and third parts. The single parameter approach is then applied again to the estimation of α, ie, to choose the target value α' and precision s_α with $L(x;\alpha)$ playing the role of $-\log f(x;\theta)$. We thus obtain

$$s_\alpha^2 = \frac{12}{\dfrac{\partial^2 L(x;\alpha')}{\partial \alpha^2}} = \frac{12}{I(x, \alpha')},$$

say; and α' minimises

$$-\log h(\alpha') + \tfrac{1}{2}\log\frac{I(x, \alpha')}{12} + L(x, \alpha') + \tfrac{1}{2}.$$

For this approach to succeed without elaboration to state precisions explicitly, we again require that $I(x, \beta' \mid \hat{\alpha})$ be expressible in terms of β' and $\hat{\alpha}$ only, or be sufficiently approximated by its expectation over $\mathscr{H}$. Similarly we require $I(x, \alpha')$ to be expressible in terms of α' only or be sufficiently approximated by its expectation. This latter condition essentially means that the optimum precision for $\hat{\alpha}$ must not depend on $\hat{\beta}$. Hence the approach may succeed with one ordering of the parameters but not the other, as in the example below. If the optimum precision for neither estimate depends on the other, both orderings will succeed and lead to identical results.

Example: To estimate μ and σ from a sample $x_1 \ldots x_n$ from $N(\mu, \sigma^2)$, let us take the usual conjugate prior

$$\mu \mid \sigma \sim N(\mu_0, k_0\sigma^2), \frac{\lambda_0}{\sigma^2} \sim \chi^2_{\nu_0}$$

so as to maintain comparability with the conventional Bayesian analysis (although we have strong reservations about its suitability for practical problems). We first consider the second

and third parts of the code word, with σ fixed. If μ' is our estimate of μ, rounded to $\hat{\mu}$ using spacing $s(\mu' \mid \sigma)$, the length of these two parts will be

$$-\log h(\mu' \mid \sigma)\, s(\mu' \mid \sigma) - \log f(x; \hat{\mu}, \sigma)$$

The second term has exactly quadratic Taylor expansion about μ' since

$$\frac{1}{2\sigma^2}\Sigma(x_i - \hat{\mu})^2 - \frac{1}{2\sigma^2}\Sigma(x_i - \mu')^2$$

has expectation $ns^2(\mu' \mid \sigma)/24\sigma^2$.
Minimisation with respect to s thus yields $s(\mu' \mid \sigma) = \sqrt{(12\sigma^2/n)}$ and then with respect to μ' gives the usual Bayes estimator

$$\mu' = \frac{n\bar{x} + (\mu_0/k_0)}{n + (1/k_0)}$$

The final minimum expected length, on substituting back, is

$$l_{2,3}(\sigma) = \log k_0 + \frac{(n+1)}{2}\log(2\pi) + \tfrac{1}{2}\log\frac{n}{12} + n\log\sigma + \frac{1}{2\sigma^2}\left[\Sigma(x_i - \bar{x})^2 + \frac{n}{nk_0 + 1}(\bar{x} - \mu_0)^2\right]$$

If we now add on the first part of the code word that quotes $\hat{\sigma}$, the message length becomes $-\log h(\sigma')s(\sigma') + l_{2,3}(\hat{\sigma})$ and expanding $l_{2,3}$ about σ' gives, similarly to above, $s(\sigma) = \sqrt{(12/I(\sigma))}$, where

$$I(\sigma) = E\left(\frac{\partial^2 l_{2,3}(\sigma)}{\partial\sigma^2}\right) = -\frac{n}{\sigma^2} + \frac{3}{\sigma^4}E\left[\Sigma(x_i - \bar{x})^2 + \frac{n}{nk_0 + 1}(\bar{x} - \mu_0)^2\right] = \frac{2n}{\sigma^2}$$

so that $s(\sigma) = \sqrt{(6\sigma^2/n)}$.

Note that this spacing on the σ-axis is proportional to σ, in contrast to the uniform spacing for μ. A final minimisation with respect to σ gives

$$\sigma'^2 = \left[\frac{n\Sigma(x_i - \bar{x})^2 + 1/k_0\Sigma(x_i - \mu_0)^2}{n + 1/k_0} + \lambda_0\right]/(n + \nu_0)$$

again familiar from usual Bayesian analysis. The overall minimum message length is

$$I_2 = -\frac{\nu_0}{2}\log\frac{\lambda_0}{2} + \log\Gamma\left(\frac{\nu_0}{2}\right) + \frac{n+1}{2}\log(2\pi) + \frac{n + \nu_0 + 1}{2} + \tfrac{1}{2}\log\frac{n(nk_0 + 1)}{288} + (n + \nu_0)\log\sigma'$$

It is important to note that if we let $k_0 \to \infty$, $\nu_0 \to 0$, $\lambda_0 \to 0$ with ν_0/λ_0 finite in order to approach the usual "no information" prior

$$h(\mu, \sigma) \propto \frac{1}{\sigma}$$

the estimates behave nicely:

$$\mu' \to \bar{x},\ \sigma'^2 \to n^{-1}\Sigma(x_i - \bar{x})^2$$

but the minimum message length tends to infinity. This is only to be expected since if transmitter and receiver share very little prior information about the parameters it will indeed need a very long message for one to describe the data to the other.

Note also that the MML estimate again involves considerably more rounding off than statisticians are normally used to. The spacing for μ implies a maximum discrepancy of $\sigma\sqrt{3}/\sqrt{n}$, 1.7 standard errors of the mean, between the "natural" estimate μ' and the transmitted value $\hat{\mu}$. As with the SMML, however, this discrepancy is of just the same order as the discrepancy between μ' and the true value μ.

5.3. *Multiple parameters*

When there are $p \geqslant 2$ parameters, the approach of the previous Section may be useful. In general, the message comprises $\hat{\theta}_1$ stated to precision $s_1(\theta_1)$, $\hat{\theta}_2$ to precision $s_2(\theta_2 \mid \hat{\theta}_1)$, $\hat{\theta}_3$ to precision $s_3(\theta_3 \mid \hat{\theta}_1, \hat{\theta}_2)$, *etc.*, followed by the data encoded according to $f(x; \hat{\theta}_1, \ldots, \hat{\theta}_p)$. It might happen, however, that some part of this message turns out to depend on a later part (such as the spacing for $\hat{\theta}_1$ depending on $\hat{\theta}_3$, for example), leaving the receiver of the message unable to decode it. We have so far been unable to find general conditions under which unambiguous coding will always be possible, but the links with existence of sufficient statistics are clearly strong.

The MML approach can, however, always be directly generalised to parameter spaces of p dimensions. The main alteration to the scalar case concerns the effect of quantization. In one dimension, the quantized estimate $\hat{\theta}$ is simply that member of a set of points with separation s which is closest to θ'. However, in p dimensions, the increase in the length of the second part of the message arising from the quantization may depend on the direction as well as the length of $\theta' - \hat{\theta}$. Further, the optimum arrangement of quantized values (in effect the set Θ^*) is not a simple cubical lattice of points in Θ.

The first complication can be removed by a linear transformation of the parameter space in the neighbourhood of θ', to

$$\phi = \Lambda^{1/2} A \theta$$

where A is the matrix of eigenvectors of $I(\hat{\theta})$ and Λ the diagonal matrix of eigenvalues. In the transformed space,

$$-\log f(x; \hat{\phi}) = -\log f(x; \phi') + \tfrac{1}{2}(\hat{\phi} - \phi')^2$$

and

$$h(\phi) = h(\theta) \mid I(\theta) \mid^{-1/2}$$

so the log likelihood function is (locally) spherically symmetric. Quantization of ϕ' to $\hat{\phi}$ thus is simply the selection of the closest quantized value to ϕ'. This leads to the estimate θ' which maximises $h(\theta) f(x; \theta) / \mid I(\theta) \mid^{1/2}$.

We now consider the precision of MML estimates. With minor modifications, the conclusions also apply to the spacing of SMML estimates. Let s now define the volume of p-space associated with each quantized value. Assuming the quantized values form a regular lattice, the expected value of $(\hat{\phi} - \phi')^2$ has the form $pk_p s^{2/p}$, where k_p is a constant depending on the geometry of the p-dimensional lattice.

Since the prior probability associated with each quantized point is $h(\hat{\phi})s$, the expected message length is minimized by allowing s to minimize

$$-\log s + \tfrac{1}{2} p k_p s^{2/p}$$

ie,

$$s = k_p^{-p/2}, \ \tfrac{1}{2} p k_p s^{2/p} = p/2$$

The optimum quantizing lattice is that with the smallest value of k_p. Quantizing lattices have been discussed by Conway and Sloane (1982). For arbitrary p, the best quantizing lattice is unknown. However Zador (1982) has shown that for all p, the best lattice has

$$\frac{\Gamma(p/2+1)^{2/p}\, \Gamma(1+2/p)}{p\pi} > k_p > \frac{\Gamma(p/2+1)^{2/p}}{(p+2)\pi}.$$

The upper bound, which can always be bettered, is the value of k_p for a random selection of points. The lower bound, which cannot be achieved for $p > 1$, is the value which would obtain if the Voronoi region of each point were a p-sphere. As p increases, both bounds approach $1/2\pi e$ from above.

It is interesting to consider the limiting behaviour of the estimate and message length as p increases in a very simple case. Consider the estimation of the mean θ of a p-dimensional

uncorrelated multivariate normal distribution with unit standard deviation in all dimensions, that is, a case where the transformation from θ to ϕ is the identity. We assume $h(\theta)$ to be uniform over a large volume V. Then the expected message length for data x comprising a single value is

$$-\log s/V - \log f(x;\theta') + \tfrac{1}{2}pk_p s^{2/p},$$

where

$$f(x;\theta') = \delta(2\pi)^{-p/2}\exp\{-\tfrac{1}{2}(x-\theta')^2\},$$

δ being the quantization volume of the observation x. Choosing the optimum $s = k_p^{-p/2}$ and $\theta' = x$ gives

$$p/2\log k_p + \log V + p/2\log 2\pi + p/2 - \log\delta.$$

For large p, $k_p \simeq (p/2)^{1/p}/2\pi e$ so the expected length is

$$-p/2\log 2\pi e + \tfrac{1}{2}\log\frac{p}{2} + \log V + p/2\log 2\pi + p/2 - \log\delta = \log(V/\delta) + \tfrac{1}{2}\log\frac{p}{2}.$$

The first term is precisely the length I_0 of the strict optimum message. Note that, in the simple case considered here, the 'approximate' message length obtained by the MML approach is exactly the expected length in the SMML approach, I_1. Hence, for this simple case, or indeed any case where the transformation from θ to ϕ is possible and yields a slowly-varying prior density for ϕ, the value of $I_1 - I_0$ only increases as the log of the number of scalar values estimated.

For given x, the posterior density of any one component of θ is simply normal with mean equal to the corresponding component of x and unit variance. If we consider the behaviour of one component of $\hat{\theta} - \theta'$ or equivalently $\hat{\theta} - x$, over the population of possible x values, it can be shown to be distributed approximately as $N(0, 1)$ for large p. Thus, for large p, the value of a component of $\hat{\theta}$ used in the first part of the message behaves as if it were selected at random from the posterior distribution of the corresponding component of θ. This result is not an accident of the particular example. For a wide range of problems involving the estimation of many parameters, there exists an SMML estimator which effectively estimates each parameter by a value chosen pseudo-randomly from its posterior distribution. This implies the statement made at the end of Section 4.4.

If $h(\phi)$ is slowly varying, ϕ' is very close to the maximum likelihood estimate ϕ_m, but $\hat{\phi}$ may be some distance from it. It is of interest to consider whether a conventional test of significance would be likely to reject the hypothesis that $\phi = \hat{\phi}$ in favour of ϕ_m.
The log likelihood ratio

$$\lambda = 2\log\frac{f(x;\phi_m)}{f(x;\hat{\phi})} = (\phi_m - \hat{\phi})^2 \simeq (\phi' - \hat{\phi})^2$$

which has expected value p so the test would be expected to reject $\hat{\phi}$ only about as often as it would reject the true value ϕ.

6. MODEL DISCRIMINATION

The minimum message length approach gives a natural way of choosing which model out of some class of possible models is the one to be preferred in the light of available data.

The parameter space simply has to be specified as the union of the spaces for all the models under consideration. The MML estimate will then automatically select a preferred model and estimate its parameters. It is this property that has made the MML method so useful in problems like intrinsic classification and grammatical inference where the class of possible models is very large or even infinite. In practice it may be better to visualise this process as one of entertaining each model in turn and finding the one with the shortest message length. A more complicated model will need a larger first part of the message to encode its greater

number of parameters, but will in general fit the data better and therefore be able to encode the second part of the message more briefly. Adding the lengths of the two halves therefore provides a natural trade-off between complexity and goodness-of-fit.

For models that have differing prior probabilities, we have to take into account the prefixes to their messages that describe which model is being transmitted. A prior probability π_i for the ith model will require a prefix of length $-\log \pi_i$. The difference between two message lengths will therefore contain a term equal to the logarithm of the prior odds. The remaining difference is then analogous to the logarithm of the Bayes factor.

The normal distribution results of the previous Section can be used to consider the simplest model discrimination problem of all, namely

Example: To choose between H_1: $\mu = \mu_1$ and H_2: $\mu \neq \mu_1$, given data $x_1, \ldots, x_n$.

(a) *If σ is known.* Model H_1 has no unknown parameters and so the message length is simply $-\log f(x; \mu_1, \sigma)$. For H_2 the analysis has already been done in Section 5.2, but the result is a little neater for present purposes if we use a uniform prior.

$$\mu \mid \sigma \sim \text{uniform } [L\sigma, U\sigma].$$

Assuming equal prior probabilities, H_1 is to be preferred to H_2. if

$$z = \left| \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} \right| < \sqrt{\left\{ \log \frac{ne(U-L)^2}{12} \right\}}$$

This contrasts with the usual significance test, and with the Akaike AIC criterion (Akaike, 1973, 1974) that both use a constant value on the right hand side, but has the same asymptotic behaviour as the Schwarz (1978) criterion for which the right hand side is simply $\sqrt{\log n}$. Note that vague prior knowledge about μ, letting $U - L \to \infty$ leads to increasingly strong preference for the simpler model, as Jeffreys (1939) first pointed out.

(b) *If σ is unknown.* Taking an inverse chi-squared prior for σ^2, the same under both models, leads to preferring H_1 to H_2 provided

$$\left\{ \frac{\Sigma(x_i - \mu_1)^2}{\Sigma(x_i - \bar{x})^2} \right\}^n < \frac{(U-L)^2 e}{12} (n - \tfrac{3}{2}).$$

which can be written less compactly as the usual t statistic being less than a term of order $\sqrt{\log n}$. The corresponding right hand side for the usual significance test is $\{1 + c^2/(n-1)\}^n$ where c is a tabulated value for Student's t distribution, for the Akaike criterion it is simply e^2 and for the Schwarz criterion simply n.

It is clear that this kind of asymptotic behaviour will persist in more complicated applications such as the general linear model. We resist the temptation to add to the enormous literature for and against the asymptotic properties of significance tests (see for example Atkinson, 1980, 1981, Smith and Spiegelhalter, 1980), except to say that to us a $\sqrt{\log n}$ criterion seems intuitively preferable.

7. DISCUSSION

7.1. *Interpretation of message length*

The first part of the message, encoding an estimate value, has a length that can be thought of, rather loosely, as minus the log prior probability of the estimate. The length of the second part is minus the log conditional probability of observations given the estimate. Hence the total length might, with caveats, be interpreted as minus the log joint probability of estimate and data, and minimising the length is therefore closely similar to maximising the posterior probability of the estimate. The MML approach will usually, for well-behaved likelihood functions, give results very close to those from a usual Bayesian analysis. For model

discrimination, in particular, the difference in minimum message length achieved by two competing hypotheses will be analogous to the log posterior odds ratio. There seem to be reasonable grounds for treating a length difference of more than 5 or 6 as fairly strong evidence favouring one hypothesis over the other.

The length I_0 always forms a lower bound to the total lengths I_1 (SMML) and I_2 (MML). The differences $I_1 - I_0$ and $I_2 - I_0$ are effectively minus the logarithm of the posterior probability of the estimate. As the discretisation of the data becomes increasingly fine, all three lengths tend to infinity but the differences converge to finite limits given by replacing summations over $\mathscr{H}$ by integrals. The estimates obtained by minimising these differences also converge to limiting forms.

7.2. *Relation to the work of Rissanen*

A series of papers (Rissanen, 1976, 1978, 1980, 1982, 1983; Rissanen and Langdon, 1981: Hannan and Rissanen, 1982) expound and develop the principle of minimum message length. The concept is almost identical to ours but its implementation in its current form differs in several details. Rissanen's criteria are only invariant under linear transformation. Much more important, however, are differences in views about the formulation of prior distributions. Rissanen finds it meaningful to consider complete, or nearly complete, prior ignorance about a parameter and proceeds to construct a "universal" prior distribution to be used in all such cases. This arranges the infinite set of discretised parameter estimate values in increasing order of distance from the origin, measured by a "natural" metric, and assigns a monotonic decreasing set of prior probabilities to those values. We prefer the philosophical view that prior information always exists and there can be no easy substitute for thinking what it is and formulating it as well as possible. This can, of course, only be done in the context of each particular problem by considering what physical quantity each parameter in the model actually represents. The health of Bayesian statistics can only be undermined by any return to the notions of ignorance current in the 1960's and 70's. Moreover, Rissanen's prior is not preserved under transformations.

With Rissanen's code for the integers, $-\Sigma_n p_n \ln p_n$ is infinite. Hence, either his encoding of the estimate is non-optimal, or its expected length is infinite. In the latter case, the statement of the estimate will present some problems.

We finally disagree with Rissanen's (1983) claim that, "while the basic elements in the principle of minimum description length can be found in several earlier papers ... a clear formal declaration of the principle itself as a means to this estimation does not seem to have been made prior to Rissanen (1978)".

Wallace and Boulton (1968) state that "We suggest that the best classification is that which results in the briefest recording of all the attribute information", and similar enunciations of the principle are to be found in Boulton and Wallace (1970) and Wallace and Boulton (1975).

7.3. *Relation to usual Bayesian results*

For a wide range of models, the minimum message length approach will give results close to, if not exactly equal to, those from a usual Bayesian analysis. The general linear model with normal errors, for example, yields a likelihood for which $-\log f(x; \hat{\theta})$ has an exact quadratic expansion about θ' and $I(\theta)$ is independent of all parameters except the error variance. The only marked difference comes in the precision of estimators, where the minimum message length view usefully, in our opinion, dramatises the current widespread tendency to quote estimates, whether maximum likelihood or posterior mode, to greater accuracy than can be justified by the inherent variability within the data.

More generally, all problems for which usual regularity conditions hold (Cox and Hinkley, 1974, Ch. 9) will show asymptotic agreement between the two methods. Here the dramatic difference is between Bayesian and frequentist-based views of model discrimination. As Stone (1979) points out, comparisons depend crucially on how the asymptotics are employed. Changing the usual frequentist formulation by either adjusting the significance level or refining

the model by adding extra parameters as the sample size increases profoundly changes model discrimination procedures for the kind of sample sizes usually encountered in practice. Much more work is needed on this, but it is clear that Bayesian methods are sensitive to the prior distributions used and all casual ideas of "complete prior ignorance" have to be replaced by careful specification of actual prior knowledge.

It is in completely non-regular problems that the minimum message length approach will produce most strikingly different and, we believe, practically useful results. Even if the approximations used in this paper cannot be justified, the general concepts can still be applied. Future papers on the Cauchy distribution, on outlier detection and on gaps within a uniform distribution will illustrate the techniques.

ACKNOWLEDGEMENT

The second author is grateful to Monash University for the award of a visiting appointment in 1985.

REFERENCES

Akaike, H. (1974) Information theory and an extension of the maximum likelihood principle. In the *Second international symposium of information theory*. (B. N. Petrov and F. Csaki, eds) Budapest: Akademia Kiedo.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control* AC**19**, 716–723.

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

Atkinson, A. C. (1981) Likelihood ratios, posterior odds and information criteria. *J. Econometrics*, **16**, 15–20.

Boulton, D. M. (1975) The information measure criterion for intrinsic classifications. Ph.D. Thesis, Monash University.

Boulton, D. M. and Wallace, C. S. (1970) A program for numerical classification. *Comp. J.*, **13**, 63–69.

Boulton, D. M. and Wallace, C. S. (1973) An information measure for hierarchic classification. *Comp. J.*, **16**, 254–261.

Chaitin, G. J. (1966) On the length of programs for computing finite sequences. *J. Ass. Comp. Mach.*, **13**, 547–549.

Conway, J. H. and Sloane, N. J. A. (1982) Voronoi regons on lattices, second moments of polytopes, and quantization. *IEEE Trans. Inf. Thy*, IT-**28**, 211–226.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical statistics*. London: Chapman and Hall.

Georgeff, M. P. and Wallace, C. S. (1984) A general selection criterion for inductive inference. In *ECAI-84: Advances in Artificial Intelligence*. (T. O'Shea, ed.), pp. 473–482, Amsterdam: Elsevier.

Hannan, E. J. and Rissanen, J. (1982) Recursive estimation of ARMA order. *Biometrika*, **69**, 81–94.

Jeffreys, Sir H. (1939) *Theory of probability*, Oxford: Clarendon Press.

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, **1**, 4–7.

Kullback, S. (1959) *Information theory and statistics*. New York: Wiley.

Maciejowski, J. M. (1978) *The modelling of systems with small observation sets*. Berlin: Springer-Verlag.

Maciejowski, J. M. (1979) Model discrimination using an algorithmic information criterion. *Automatica*, **15**, 579–93.

Maciejowski, J. M. (1980) A least-genericity principle for model selection. Tech. Rep. No. 91, Control Theory Centre, University of Warwick.

Patrick, J. D. (1978) *An information measure comparative analysis of megalithic geometries*. Ph.D. thesis, Monash University, Australia.

Patrick, J. D. (1986) An explication of Halstead's measures. Submitted for publication.

Patrick, J. D. and Wallace, C. S. (1982) Stone circle geometries: an information theory approach. In *Archaeoastronomy in the Old World*, (D. C. Heggie, ed.), Cambridge: Cambridge Univ. Press.

Rissanen, J. (1976) Parameter estimation by shortest description of data. *Proc. JACE Conference RSME*, 593.

Rissanen, J. (1978) Modelling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen, J. (1980) Consistent order-estimates of autoregressive processes by shortest description of data. In *Analysis and Optimisation of Stochastic systems*. (O. Jacobs *et al.*, eds) New York: Academic Press.

Rissanen, J. and Langdon, G. G. (1981) Universal modelling and coding. *IEEE Trans. Inf. Thy.*, IT-**27**, 12–23.

Rissanen, J. (1982) Estimation of structure by minimum description length. *Circuit systems signal Process*, **1**, 395–6.

Rissanen, J. (1983) A universal data-compression system. *IEEE Trans. Inf. Thy.*, IT-**29**, 656–664.

Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, **11**, 416–421.

Schwarz, G. (1979) Estimating the dimension of a model. *Ann. Statist.*, **6**, 416–46.

Shannon, C. E. and Weaver, W. (1959) *The mathematical theory of communication*. Urbana: Univ. Illinois Press.

Smith, A. F. M. and Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc.*, B, **42**, 213–220.

Solomonoff, R. (1964) A formal theory of inductive inference I, II. *Inf. and Cont.*, **7**, 1–22, 224–254.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc.* B, **42**, 276–278.

Wallace, C. S. and Boulton, D. M. (1968) An information measure for classification. *Comp. J.*, **11**, 185–195.

Wallace, C. S. and Boulton, D. M. (1975) An invariant Bayes method for point estimation. *Classification Soc. Bull.*, **3**, 11–34.

Zador, P. (1982) Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inf. Thy.*, IT-**28**, 139–149.

DISCUSSION OF THE PAPERS BY DR RISSANEN AND PROFESSORS WALLACE AND FREEMAN

Professor A. P. Dawid (University College London) This has been an exciting evening. The coding-theoretic approach to inductive inference has developed largely outside the statistical literature. After tonight's meeting there should be no doubt of the interest and importance of these ideas for Statistics.

Tonight's papers both start with the same basic idea, but move in interestingly different directions. In each, the basic task of data analysis is conceived to be the transmisson of a given sequence of data, by means of some code, as efficiently as possible. We do not need to imagine that some, unknown, probability distribution has generated the data. Instead, any proposed (prefix) code (more precisely, its length-function) can be associated with a probability distribution: that for which it provides the minimum expected code-length. This is an important philosophical point: stochasticity is not an attribute of data, but of our attempts to explain data. The same viewpoint underlies the theory of empirical probability developed by Dawid (1985).

What code should be used for transmitting the specific data at hand? Given two or more competitors, that yielding the shortest code-length for the actual data is, *prima facie*, to be preferred. However, in this case the sender must also tell the receiver which code to use, which itself requires a coded message, whose length must be added on. Before beginning transmission, sender and receiver must therefore agree on the structure of the "code-book" they will use. In the simplest case, this will contain but a single code. This would be optimal for a particular probability distribution over possible data-sequences, which should clearly be chosen to reflect prior expectations of the data. Alternatively, the codebook could contain a collection of codes, equivalent to a family $\mathscr{P} = \{P_\theta\}$ of data distributions, in which case a "key" code for transmitting a value of θ is also required. This is the problem tackled by Wallace and Freeman, who take $\mathscr{P}$ as given and try to optimise the code for θ, again necessarily in the light of prior knowledge. The search for optimal precision is particularly noteworthy here. Their analysis illuminates the standard theory of statistical estimation from a new and revealing angle, and holds the prospect of greatly advancing our understanding of such fundamental ideas as sufficiency and ancillarity. I look forward to a wealth of new insights from developments of their approach.

Rissanen, on the other hand, asks whether a code-book, representing a family $\mathscr{P}$, could itself be replaced by a single code. This code should perform as well as possible, perhaps in an asymptotic sense, for any data for which some $P_\theta \in \mathscr{P}$ provides an optimal code. He shows, essentially, that this can be done if the code corresponds to a Bayesian marginal distribution, $\int P_\theta(\cdot)\pi(\theta)d\theta$, formed by assigning a density π to the parameter of $\mathscr{P}$, and extends this one level deeper by introducing a parameter k labelling alternative families. The optimal code-length is the stochastic complexity.

At this point we have returned close to familiar statistical concerns, and can, if we like, ignore the coding-theoretic motivation: a code-length is just a negative log-likelihood. Indeed, I had been led to much the same end-point as Rissanen from a different starting point: the "prequential" approach of Dawid (1984). In that paper I used simple martingale theory to show that, if $\theta \in \mathbb{R}^k$, $R(\cdot) = \int P_\theta(\cdot)\pi(\theta)d\theta$ with $\pi > 0$, and Q is any distribution, then the likelihood-ratio in favour of Q as against R, based on the first n data-values, must be bounded above, as $n \to \infty$, with P_θ-probability 1, for almost all θ. This demonstrates the essential optimality of a Bayesian mixture code as a replacement for a parametric family of codes: no other single code can be shorter by an unbounded amount for (almost) any data-sequence whose optimal code lies in the family. This "almost sure" optimality improves on the "in exectation" results given by Rissanen, and leads to an "almost sure" variant of (4.3), as indicated by Rissanen, *viz.* $I(x^n) + \log f(x^n \mid \theta) - \frac{1}{2}k \log n$ is bounded, almost surely under P_θ, for almost all θ. Additional aspects such as consistent order estimation which arise when we allow choice among a discrete collection of parametrised models, as incorporated in Rissanen's framework, are likewise easily handled by the prequential methodology. Asymptotic optimality will hold for a Bayesian mixture $\Sigma_k Q(k) \int f(x \mid k, \theta)d\pi(\theta \mid k)$, with arbitrary densities $\pi(\cdot \mid k) > 0$ and model probabilities $Q(\cdot) > 0$. Uniform or universal priors are irrelevant, and I share Wallace and Freeman's opinion that (particularly for their own, non-asymptotic, theory) the best prior distribution is the one which truly expresses beliefs. Nevertheless, even a non-Bayesian should be impressed by the asymptotic effectiveness of the Bayesian methodology, almost irrespective of the choice of prior distributions.

Finally, I should be pleased to see the connexions between tonight's papers developed more thoroughly. What are the asymptotic properties of the SMML and MML coding methods, and how do they relate to Rissanen's results? There is a mismatch here, in that Wallace and Freeman only consider the encoding of known fixed quantities of data. (Presumably, if the sample size is not specified in advance, it too must be coded and transmitted, or else a special "end-of-data" message sent.) But if the data-sequence is extended, it seems that the parameter must be re-estimated and the entire coding redone from scratch.

This is clearly wasteful, and at odds with Rissanen's coding methods which take such extension in their stride. Nevertheless, earlier work of Rissanen indicates that the SMML method yields a code-length asymptotically equivalent to the stockastic complexity. Just what, then, are the deeper connexions between the two viewpoints we have heard tonight?

By organising tonight's important meeting, the Royal Statistical Society has injected into our subject a revitalising transfusion of stimulating new ideas. I therefore propose that all three authors of tonight's two papers be accorded an enthusiastic vote of thanks.

Professor P. Whittle (Statistical Laboratory, University of Cambridge): People have felt for many years that the ideas of information theory were relevant to statistical inference, and attempts have been made to demonstrate the connection. However, I feel that tonight's two papers make, at last, the right and fundamental approach. The first explicit statement of the approach adopted (the economic combined coding of model and of data conditional on the model) seems indeed to be that in the 1968 paper by Wallace and Boulton; this aproach is taken considerably further by tonight's papers.

In their Section 3 Professors Wallace and Freeman hold out as ideal a coding based on the distribution $r(x)$ (of sample averaged over parameter values) and then reject this on the grounds of inconvenience, observations not being independent on this distribution. One may make perhaps the same point by saying that the parametric nature of the model is buried in this distribution, and one exhibits the model structure much more evidently by proceeding as the authors do.

In Section 5.1 one encounters the expression $s^2/12$. Anybody who started their statistics in the first half of this century will with joy recognise in this the correction for grouping! It is strange that concepts which may arise in a seemingly pedestrian setting later prove to have a more permanent and fundamental character. In this case the point is, of course, that an economic coding leads to a natural discretisation of parameter space, analogous to the discretisation of grouping. I found it very appealing that the coding approach led to a natural level of precision for the quoting of parameter estimates, and also that these estimates could be regarded as generated quasi-randomly from the posterior distribution. On the other hand, it would have been interesting to have seen somewhat more explicitly the role of the dimension of parameter space in the analysis.

This dimension, k, played a very explicit role in Professor Rissanen's approach. He deduces a 'universal' distribution for k which certainly has an appealing character: it must be almost a boundary point for the set of distributions, in that it converges about as slowly as is consistent with summability. His two theorems are convincing. However, doubtless through my own fault, I did quite take in some of the points he made about the extra-statistical nature of some of his analysis. For one thing, in some sense he separates completely the ideas of coding and of a statistical specification of a model, whereas an economical coding (and all the Shannon theory) is based on the minimisation of an expected value. For another, he says (just before formula (2.2)) that the prior distribution should assign substantial weight to the neighbourhood of the maximum likelihood estimate. This seems a very posterior way of specifying one's prior, and to indicate that Professor Rissanen is indeed taking a novel view of the concept.

Both sets of authors take as criterion the efficient transmission of all the data—the least committing of aims, in that nothing is lost. However, in cases where one wishes to transmit less than everything the well-developed theory of data compression (again in the communication context) should be valuable. There is now a subjective element in the analysis, in that one must specify a distance function (between full and compressed data), but this is inevitable. Alternatively, it seems reasonable to transmit only that part of the data which would be of predictive value for a second and independent set of data. Professor Rissanen's $\hat{k}$ and $\hat{\theta}$ are presumably asymptotically sufficient in this respect.

It is a pleasure to me to second the motion for a vote of thanks to all three authors.

The vote of thanks was passed by acclamation.

Professor H. Tong (University of Kent at Canterbury): I would warmly concur with Professor Rissanen in acknowledging our indebtedness to the pioneering and inspiring contributions of Dr Hirotugu Akaike to the area. Now, I have a few questions and comments.

(1) I would like to hear our three authors' views, in the context of model selection, on the inter-relation between the roles of Boltzmann's entropy, upon which Akaike's approach is based, and Shannon's entropy, which has certainly much connection with the approaches presented here tonight. Dr Akaike has set out his views clearly in his recent paper in the ISI Centenary Volume (1985). The fact that

Schwarz's type of selection criterion may be constructed by following either one of the above approaches suggests that there is much common ground.

(2) Professor Rissanen objects to taking expectation operations relative to the "true" distribution. However, it seems to me that as soon as we get down to analysing any model selection criterion, we would often, and indeed Dr. Rissanen does, consider its '*asymptotic* optimality property' and 'optimality properties *in the mean*', etc. In such situations, can we really do without

(i) the notion of a true distribution
and
(ii) integration w.r.t. the true distribution?

(3) Dr. Rissanen's universal prior intrigues me greatly. Could it be related to the partition function of thermodynamics? Akaike's $\exp\{-\frac{1}{2}AIC(k)\}$ seems to be so related.

(4) I think the title of Professor Rissanen's paper wins on being almost *MDL*!

Dr R. J. Bhansali (University of Liverpool): The two papers presented tonight are welcome because they illustrate how an infusion of information—theoretic ideas into Statistics may help in providing new answers to some of its old methodological questions. Personally, I concur with Rissanen's point of view that for a given set of observations, there cannot be any "true" distribution. However, I do not entirely agree with the tone of his criticism of Akaike's work. The concept of a "true" distribution like that of a "true" parameter value has served statistics well because it provides a yardstick by which the usefulness of a new statistical procedure may be judged. This point is well illustrated in Section 5.2 of Rissanen's paper where the author is indeed talking about the implications of his results when the observations are generated from a "true" Gaussian distribution with a fixed number, m, of parameters. Of course, as pointed out by Akaike, there may not be a true finite m, and then Akaike's procedure may have an advantage from the point of view of a smaller mean squared error of prediction. At a deeper level, the question of whether or not there exists a "true" distribution is a metaphysical one and may be likened with the question of whether or not one accepts there is God. However, the use of the objective "dishonest" in relation to Akaike's, and related, work is to be regretted and it is perhaps inappropriate in scientific discourse.

Rissanen's approach is likely to have an obvious appeal in Time Series Analysis. However, there the implications of author's Theorems 1 and 2 have not been fully discussed. Reference to Hannan and Rissanen (1982) is unhelpful because as shown recently by Hannan and Kavalieris (1984) the original procedure suggested by these authors does not provide a consistent estimator of the order, and it needs a correction. Perhaps the author could comment on the correction suggested by Hannan and Kavalieris.

For fitting autoregressive models 'non-parametrically' Shibata (1980, 1981) has established an asymptotic optimality property for AIC from the point of view of prediction and spectral estimation. He has also shown that if the Minimum Descriptive Length criterion of Rissanen is used for order selection then the lower bound on the mean square error of prediction cannot be achieved. Perhaps, Rissanen could also comment on the relationship, if any, between his work and Shibata's

As regards the paper by Wallace and Freeman, it should be pointed out that the comparison carried out in section 6 of their paper is rather simple since the main advantage of using AIC, and related order determining procedures, occurs when discriminating between more than two models because then the conventional hypothesis-testing approach requires the use of a multiple testing procedure, and the question of how to choose an appropriate significance level each time the test is used is not easily answered—see Akaike (1978) for details.

Dr John T. Kent (University of Leeds): Tonight's papers offer some intriguing suggestions on the question of model choice. I would like to make some observations about the very simple model discussed in Section 6 of the Wallace and Freeman paper. Let $x_1, \ldots, x_n$ be independent observations from $N(\mu, 1)$ and consider the models $M_0 : \mu = 0$ and $M_1 : \mu$ unrestricted. We choose between these models using $n\bar{x}^2 \sim \chi_1^2(n\mu^2)$, choosing M_0 if $n\bar{x}^2 \leqslant D_n$, for some "cutoff value" D_n.

Several methods of model choice can be interpreted in terms of a "threshold" $\tau_n \geqslant 0$. We prefer M_0 if $\mu^2 \leqslant \tau_n$ and prefer M_1 otherwise. That is, we prefer M_0 unless μ is "substantially" different from 0. Often $\tau_n = \tau$ does not depend on n and is chosen by the experimenter.

1. *Significance testing with a fixed threshold.* A test of $H_0 : \mu^2 \leqslant \tau$ is obtained by taking $\mathbf{D}_n = \chi^2_{1;\alpha}(n\tau)$, the upper α critical value of the noncentral χ^2 distribution. If $\tau = 0$, we get the usual significance test of $H_0 : \mu = 0$, with $D_n = \chi^2_{1;\alpha}(0)$ not depending on n. On the other hand, for $\tau > 0$ we find $D_n = n\tau + 0(n^{1/2})$

as $n \to \infty$. For more complicated models the notion of "threshold" is most easily formalized using Kullback-Leibler information gain (Kent, 1983, Section 8).

2. *AIC*. Akaike suggested a threshold depending on n, $\tau_n = 1/n$, and a cutoff value of $D_n = 2$, the mean value of the $\chi_1^2(n\tau_n) = \chi_1^2(1)$ distribution. Again note that D_n does not depend on n.

3. *MML*. The Wallace-Freeman suggestion for model choice leads to $D_n = C \log n$ where C depends on the prior information. This choice of D_n seems related to a threshold τ_n proportional to $(\log n)/n$, which is intermediate between the AIC threshold $\tau_n = 1/n$ and the fixed threshold $\tau > 0$.

Professors Wallace and Freeman use the experimenter to obtain prior information about the parameters. Perhaps in some applications the experimenter should also be interrogated about his choice of threshold τ_n.

Dr A. O'Hagan (University of Warwick): I found these two papers extremely interesting. My comments are directed to Professors Wallace and Freeman, whose philosophical approach is much more familiar to me.

My main comment concerns the derivation of the *MML* estimator in Section 5.1. The authors expand $\log f(\chi; \hat{\theta})$ around $\hat{\theta} = \theta'$, but why do they not similarly expand $\log h(\hat{\theta})$? Instead they use only the first term $h(\theta')$. Taking both expansions to the terms in $(\theta' - \hat{\theta})^2$ and then taking expectations, the expectation of the total message length is

$$-\log s - \log g(x; \theta') - \frac{s^2}{24} \frac{\partial^2}{\partial \theta^2} \log g(x; \theta'),$$

where

$$g(x; \theta) = h(\theta) f(x; \theta)$$

is proportional to the posterior density of θ given x. This is minimised for s by

$$s^2 = 12/H(x; \theta'),$$

where

$$H(x; \theta) = -\frac{\partial^2}{\partial \theta^2} \log g(x; \theta),$$

and the *MML* estimator becomes that value of θ that maximises

$$g(x; \theta)/H(x; \theta)^{1/2}. \tag{A}$$

The authors' expansions give the same results except that $H(x; \theta)$ is replaced by the realised information $I(x; \theta)$. They then further replace $I(x; \theta)$ by its expectation, the Fisher information $I(\theta)$. This is explained by the argument that in the context of efficient coding the quantum s is better coded if it does not depend on x. However, the authors propose interpreting the *MML* estimator as an estimator in the statistical sense. In this context there is no reason to use $I(x)$. Indeed, there is no reason actually to quantise the estimator to the value $\hat{\theta}$. The natural estimator is θ', with the quantum s interpreted as a measure of its accuracy. This interpretation seems to be adopted in the paper, where the quantised $\hat{\theta}$ rarely appears explicitly.

The modified *MML* estimate maximising (A) has somewhat different properties from the authors' version. The paper stresses the relationship with Bayesian inference. The formula (A) provides a more natural Bayesian estimator, since it depends only on the posterior distribution. The authors' use of $I(\theta)$ violates the Likelihood Principle. Wallace and Freeman also comment that their *MML* estimator 'essentially chooses the local posterior mode with greatest probability content rather than simply the highest one'. This statement is much more true of our modified estimate, because $I(x; \theta)$ or $I(\theta)$ do not accurately measure the width of a mode. I have used (A) as a measure of the probability content of a mode in O'Hagan (1987).

However, θ' will maximise (A) over all θ, which does not necessarily occur at any mode. (A) can certainly be higher on a shoulder (O'Hagan 1985, 1987) than at the mode. Unfortunately, the maximum need not occur at any such useful or interesting place. When the posterior density is bimodal then there will necessarily be points of inflexion in the log density, and hence points at which $H(x; \theta)$ is zero. Clearly, the approximations used in Section 5.1 break down here. Another difference is worth mentioning—the modified *MML* estimator is not invariant under nonlinear transformation of θ.

In conclusion, I have pointed out that Wallace and Freeman's derivation expands log $f(x; \hat{\theta})$ but ignores the expansion of log $h(\hat{\theta})$. Why not expand log $h(\hat{\theta})$ and not log $f(x; \theta)$? Expanding both seems more reasonable, more Bayesian, and should be better in many cases. But the modification loses some of the appealing properties claimed by Wallace and Freeman. I look forward to the authors' reply, and thank them again for their stimulating paper.

Dr T. Fenner: Being a computer scientist, I found the connection between coding theory and statistical inference rather interesting.

Taking Professor Dawid's scenario in terms of minimising the length of the message, one way of viewing the message, why we should want to minimise the length of the message, might be in order to minimise the time it takes for the message to get across.

It occurred to me that it might be interesting to see what effect there might be on this approach if we took into account also the time, in some sense, that it took to encode and decode the message.

Professor Jose M. Bernardo (University of Valencia, Spain): It is a pleasure to be asked to contribute to the discussion of the very interesting, thought-provoking papers read tonight. The authors of both papers put forward the idea that an appropriate criterion for model selection and/or estimation consists of minimizing the code length which is necessary to describe the data. I must say that I find very attractive the idea of a trade-off between complexity and goodness-of-fit which those methods suggest. However, to accept the methods proposed one would surely require that they provide sensible answers in the simple special case where a single fixed model is to be estimated; unfortunately, I do not believe this to be the case. Due to editorial space limitations, I will limit my comments to some specific queries.

Wallace and Freeman produce their *useable MML* estimator as that value of θ which maximizes $h(\theta)f(X; \theta)/|I(\theta)|^{1/2}$, where $I(\theta)$ is the appropriate information matrix. This means that if the prior $h(\theta)$ is chosen to be Jeffreys' *non-informative* prior $\pi(\theta) = |I(\theta)|^{1/2}$, then the *MML* estimator reduces to the standard maximum likelihood estimator but, apparently, the minimum message length then tends to infinity. It is hard to accept minimizing code length as an estimation criterion given the knowledge that the old, usually well-behaved *ML* estimator seems to do precisely the opposite. If, on the other hand, the authors are serious in claiming that only subjective priors should be used, then it is hard to see why one should prefer *MML*, involving the information matrix in the denominator, to simply maximizing $h(\theta)f(x; \theta)$, i.e. taking the posterior mode or, for that matter, any estimator which is optimal under some loss function appropriate to the problem. This obviously suggests a final question: is *MML* optimal (and therefore admissible for some interesting utility function? If not, how could it possibly be justified from the subjectivist attitude adopted by the authors?

Rissanen's argument really focuses on the implied predictive distributions. Indeed, for a given model with k parameters, $f(x \mid k, \theta)$, with $f(x \mid m, \theta) = 0$ for all $m \neq k$, his procedure reduces to maximize the predictive density of the observed data $f(x) = \int f(x \mid k, \theta)\pi(\theta \mid k)d\theta$ among the available class of priors $\pi(\theta \mid k)$. This leads to that prior specification which, given the model, makes the data more likely *a priori*. Thus, in his Example 1, with $d = 1$ and the data consisting of one success ($x = 1$), the suggestion is to select any prior such that $p(x = 1) = E(\theta) = 1$, say a distribution for θ degenerated at $\theta = 1$, hardly an acceptable proposition! Indeed, a procedure which forces the prior to fit the data (through the implied predictive distribution in this case) is very suspect to me. If, on the other hand, the prior is to be fixed before the method is applied, then Rissanen's claim to an *automatic* inference machine would require an automatic procedure to specify the prior, and one which would always produce a proper prior, for otherwise the method degenerates. I do not believe that such a procedure exists. Finally, to anyone exposed to decision making the idea of a *universal* loss function (code length or whatever) is hardly acceptable; the optimal choice of model and/or estimation is bound to depend on the preferences structure in the proplem considered. As I mentioned to the author while visiting San José, an estimation procedure must be able to accommodate a situation where, say, overestimation might kill the patient but underestimation is pretty innocuous; the specification and use of an appropriate loss function then seems unavoidable.

Professor Seymour Geisser (University of Minnesota): Stochastic complexity as applied to statistical inference appears to require discriminating among models by a coding principle whose implementation is often in need of assuming, averaging, and approximating. When all of this is accepted, asymptotic claims for optimality are made depending on this principle. This is little different from many of the other

principles, too numerous to mention, for obtaining reference priors and discriminating among models (an enterprise to which I, too, am addicted, Geisser (1975, 1979)).

An additional feature here, however, is the stringing requirement in obtaining so called "honest" predictions. The pejorative use of this term, reflecting a moralistic posturing that is entirely unwelcome, was unfortunately introduced by Bayesians to disparage priors that were not their own. Now we have a second pejorative use, namely the presumption that all "predictions" that are not sequenced are "dishonest." First, most statistical data are not sequenced and, secondly, prediction in its proper sense means making an inference about a set of entities capable of being observed or measured whether in the past or future or simultaneously with a set of observations. Schemata for sample reuse procedures can be adapted to sequenced situations. In fact, the general principle underlying any predictive sample reuse schema is that it simulates the actual predictive process itself on the available data, Geisser (1975), either in a low or high structure framework. Further, if complete "honesty" (*mea culpa*) in prediction is the goal then obtaining the single best fitting model is irrelevant since the "best" prediction will depend on a mixture of the model distributions entertained, at least for Bayesians. To integrate model discrimination and prediction using the "best" model requires a further principle of parsimony or its equivalent or some loss function that heavily penalizes mixing more than one distribution.

Finally, one is tempted to discriminate among model discrimination procedures themselves. A study of several such procedures by Clayton, Geisser and Jennings (1986) applied to nested normal and exponential situations indicated that although different model selection procedures can and will give rise to appreciably different correct model selection rates when the "true" model is included, the predictive errors incurred tend to be remarkably similar. Of course, it would have been of interest to have included Dr. Rissanen's method in the study but the authors were unaware of his work at the time. In the aforementioned paper there is also an explanation as to why in nested situations the Akaike and sample reuse criteria can be preferable to the Schwarz criterion for prediction contrary to the intuitive preference expressed by Wallace and Freeman. I, also, take mild exception to their statement concerning the health of Bayesian statistics being undermined by notions of ignorance current in 1960's and 70's. Although their view is certainly plausible it must be weighed against the potential misuse and abuse inherent in formulating prior opinion on an artifice (parameter) of a statistical model known in many cases not to be the "true" one and the fact that if the Beyesian approach is to be used at all (except by thee and me) neutral or reference methods are necessary.

It appears that statisticians already beset with principles involving gambling, entropy, divergence, repeated sampling, conditionality, sample reuse, prediction, likelihood, invariance, admissibility are now to be encoded as well. Say what you like about statisticians (liars, damn liars, etc), but they are certainly not unprincipled.

Professor E. J. Hannan (Australian National University): There is no doubt that Professor Rissanen's theory is of considerable importance and that it elucidates many inferential problems. However its use faces the difficulty that the understanding in the researcher's mind concerning the underlying process generating the data may be so vaguely and intuitively held that it may be nearly impossible to realize this understanding in terms of a model class. Also the researcher may have a special end in view for his use of the data. Presumably the model class must reflect this end purpose appropriately but such a model class may be difficult to find. As a result it seems that there will always be developed techniques that are rather *ad hoc*, albeit powerful and rather general, for the treatment of particular problems. An example is, perhaps the use of the bootstrap in relation to the estimate of the accuracy of a mean. Another example may be the Fourier analysis of time series data. Here no very well defined end may be held in view other than that embodied in the expectation that interesting features of the data will relate to frequency, phase etc. It would be very useful to have Professor Rissanen's comments on all of this.

Professor D. V. Lindley (Somerset): The authors are to be thanked for introducing us to work on inference being carried out in another discipline. My major criticism of this work is its emphasis on point estimation, especially when the general tenor of the approach is Bayesian. The basic idea in the Bayesian view is the representation of uncertainty by probability and, in particular, the description of a parameter θ through its distribution. It is the conditional distribution of θ given the data that should be reported, not a mere point estimate which is a totally inadequate substitute. The concept of a point estimate, and the determination of a best one, are notions belonging to sampling-theory statistics where they arise solely because of the unsound formulation of the inference problem. They should have but a trivial role in Bayesian statistics as a central measure of the distribution of θ. What is our Bayesian

going to do with the point estimate of the example in Section 5.1 of Wallace and Freeman's paper? Why not cite the posterior distribution? In other words, I believe the workers on coding theory are addressing a non-existent problem.

It is said that the ideas have been used in classification problems and in choices between models. But the same point applies there: why not simply state the probabilities of the classes or models, their parameters having been eliminated by integrations? If it is a question of selecting one preferred model then we have a decision problem and that cannot satisfactorily be solved without explicit consideration of utilities.

The adoption of any standard utility, or universal prior is unfortunate because it discourages careful thinking about the problem. Scientists may like rote procedures, like significance tests, because of their routine nature but should be weaned away from them. But to end on a note of agreement: how good to see the other selection procedures abandoned in favour of the sensible $(\log n)^{1/2}$ approach. Scientists won't like it though: it makes it so hard to find "significant" effects.

Professor K. V. Mardia (University of Leeds): I find the papers very stimulating. However, I could not see how the work extends to spatial situations. Suppose we have a continuous spatial process $\{Z(\mathbf{x})\}$ sampled at $\mathbf{x} = \mathbf{x}_i$, $i = 1, \ldots, n$ in two dimensions. Assume that it is Gaussian with mean $\boldsymbol{\beta}'\mathbf{f}(\mathbf{x})$ and covariance function between $Z(\mathbf{x}_i)$ and $Z(\mathbf{x}_j)$ as $\sigma(|\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta})$, where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are parameters. (See, Mardia and Marshall, 1984).

On Rissanen's paper. Since there is no (meaningful) unilateral representation in two dimensions, how does one obtain the shortest description of length of the data? How will one select dimensions of $\boldsymbol{\beta}$ or/and $\boldsymbol{\theta}$? Of course, the *AIC* applies immediately for the selection of their dimensions.

On the paper by Wallace and Freeman. How does the presence of autocorrelations effect the compact coding procedure? Here the expression for the Fisher information is somewhat complicated. In the most general setting, what is the effect on the *MML* estimation procedure when the observed and the Fisher information are substantially different, and one is replaced by the other?

Dr Jon Patrick (Deakin University): My comments concern the paper by Professors Wallace and Freeman. The Minimum Message length (*MML*) of this paper is better known in the computer science discipline as the Wallace Information Measure (*WIM*). It has proved to be invaluable in tackling problems in the modelling of computer programs, to measure the optimality of an inductive inference and numerical taxonomy. It enables one to measure the complexity of an hypothesis and the fit of the data to the hypothesis and thereby compare on equal grounds two competing but structurally different hypotheses. Hence it is the first true quantification of Ockham's Razor.

The *WIM* is an important meaure to statistics because it enables an optimisation of the log likelihood function and its second derivative which is invariant under many types of transformations when the likelihood function is not. At the same time the *WIM* behaves functionally like a joint probability of the hypothesis and evidence. Therefore, it is in accordance with the approach adopted by Bayesians when there is only a finite number of hypotheses to consider.

The *WIM* should move statistical inference from its current regime of deductive inference to inductive inference. Hence deduced proof statements, which are generally not that useful, will be replaced by plausibility statements.

One of the great advantages provided by the *WIM* is its ability to measure the complexity of an hypothesis by encoding it as an algorithm, in an arbitrary language. The algorithm is the program needed to decode the message describing the data and must be provided to a receiver in advance of sending the data. The length of the message describing the algorithm is a measure of complexity in the context of the a priori knowledge embodied in the language common to the sender and receiver. Some *a priori* knowledge or agreed communication code must always exist between sender and receiver but at its simplest level can be reduced to some form of Turing machine. Great advances in both statistics and computer sicence should follow from this understanding of the knowledge content of communication codes.

The **authors** replied later, in writing, as follows.

Professor Rissanen: Professor Dawid's discussion raises a number of relevant issues, such as the relationship between the idea of stochastic complexity and the Bayesian approaches, and the selection

of priors, which were also of concern to some of the other discussants. Despite the fact that the formula (2.1) for the stochastic complexity involves a prior, this is by no means essential, and the two approaches are fundamentally different, the central distribution in the stochastic complexity being the distribution on the data rather than the posterior distribution on the parameters, which is basic in Bayesian thinking. Consider the negative logarithm of the posterior distribution $\log f(x) - (\log f(x \mid \theta) + \log \pi(\theta))$. It follows from Theorem 2 that if we substitute for θ the value that maximizes the posterior probability, then the first term, as a function of the number of parameters, has a maximum, and can be used as a model selection criterion. However, the second term partially cancels the first, and the posterior distribution will not be able to distinguish between models with different numbers of parameters. It seems to me that the central issue involved goes beyond the transformation of the prior distribution to the posterior one, as defined by Bayes' theorem. To further emphasize the relative lack of importance of the distribution $\pi(\theta)$ and the Bayesian interpretation, let me define the stochastic complexity relative to a model class of the form $M = \{f(x \mid \theta)\}$, which in a way extends the role of the likelihood function as ascribed to it by the classical likelihood principle. Suppose that the integral $K(x^n) = \int f(x^n \mid \theta) d\theta$ exists for $n \geqslant m$, say $n \geqslant 1$. Then $f(x_{t+1} \mid x^t) = K(x^{t+1})/K(x^t)$ is a proper density function, and we may define

$$I(x \mid M) = -\log K(x) + \log K(x_1) - \log f(x_1),$$

where $f(x_1)$ is any density function, which may be picked to reflect prior knowledge. For example, we may put $f(x_1) = K(x_1)/\int K(x_1) dx_1$ if the integral exists. This choice makes the stochastic complexity independent of any ordering of the data. Finally, if we wish to include a prior in the model class, then formula (2.1) may also be interpreted as a general device of the type $\Sigma_\theta 2^{-\varphi(x,\theta)}$ (θ taken discrete) to remove redundancies in coding of x stemming from the fact that the same data may be encoded with all the parameters.

I must disagree with Professor Dawid and Professors Wallace and Freeman regarding the choice of the prior by one's subjective belief. Such a choice appears to me to be arbitrary and, besides, is no guarantee of the usefulness of the prior. Rather, the place for prior knowledge is in the selection of the model class, where the first member, as I just described, is generally more important than the second. The resulting choice is then judged in the light of the data by the stochastic complexity. This leads to no absurdities, although I realize that the lack of a complete formalization of the code length needed to encode model classes in their greatest generality, which include the distributions $\pi(\theta)$, leaves one a bit uneasy. In the case with finitely many classes, say m, which still goes a long way because these may constitute of all the distributions ever printed, there are no difficulties, however, for each class may now be encoded with the same length, about $\log m$, which cancels out in comparison. Moreover, we may also easily handle priors $\pi(\theta \mid \alpha)$ with nuisance parameters α. The result bears a resemblance to the empirically determined Bayes' priors, except that instead of maximizing the posterior density we minimize the stochastic complexity $I(x \mid \alpha)$, and importantly, that the nuisance parameters are determined predictively. This last step makes sure that the result is a code length defining a distribution in contrast with $I(x \mid \hat{\alpha}(x))$. This, perhaps, explains the posterior view of specifying one's prior, which Prof. Whittle was wondering about in his discussion and which also worried Prof. Bernardo.

There is a strong similarity between Professor Dawid's prequential approach and the ideas of stochastic complexity, which I find quite pleasing. This similarity has its explanation in the fact that prediction, the foundation of Prof. Dawid's approach, is a special case of coding, as I have discussed in Rissanen (1986a). This, however, is true only when we consider prediction in its proper interpretation, for which I selected the technical but apparently unfortunate name "honest" to distinguish it from the incorrect and misleading usages of the term prediction, as defined for example by Professor Geisser in his discussion.

Professor Whittle raised also the issue of why in my approach the objective, in deviation of Shannon's theory as well as the approach taken by Wallace and Freeman, is not to minimize the expected code length. The reason is, as I was trying to explain in the beginning of Section 2, that the expected code length, relative to a distribution under our own choice, can be made as small as we like, and hence we would not be able to search for a best model class. Finally, the *MDL* estimates of k and θ are indeed asymptotically sufficient in the sense indicated by Professor Whittle.

Regarding Professor Tong's question (1), I'll try to clarify the relationship between Boltzmann's and Shannon's entropies and the stochastic complexity. As I understand it, Boltzmann's entropy is mathematically the logarithm of the number of elements in a finite set, which makes it a special case of Shannon's self-information $-\log P(x)$ as well as of the stochastic complexity. The axioms for information, or complexity in the newer parlance, do not leave room for anything but the choice of the distribution

$P(x)$, which then becomes the crux of the matter and ought to be done to reflect the constraints at hand. In Boltzmann's case a great deal of ingenuity is needed in the selection of the finite set whose logarithm defines the physically meaningful entropy. The same is true in selecting the model class defining the stochastic complexity. Further, Akaike's criterion does not correspond to this kind of complexity while Schwarz' criterion is just an asymptotically justified code length, given also by the *MDL* formula. Sharper criteria, however, result from the stochastic complexity, as illustrated in the paper.

Turning to the second question (2), which Dr Bhansali, too, was concerned with, there is a fundamental difference between an assumption of a "true" distribution generating the observed sample, which literally is quite meaningless, and such an assumption in mathematical analyses. It is quite remarkable that we can associate with the idea of stochastic complexity three valid data dependent interpretations, namely, a code length, a prediction error, and a maximized probability or density, which make no appeal to any expectation nor "true" distribution, and which jointly constitute a compelling basis for statistical inference. The purpose of mathematical analyses, again, is to provide auxiliary information about the behaviour of whatever procedure for statistical inference has been proposed. If in such an analysis a quantification such as "for all parameters θ" is sometimes identified with "for all true distributions $f(x \mid \theta)$", no serious harm is done, unless one draws too far reaching conclusions. This happens, for example, when an idealized model selection criterion is designed to achieve a mathematical property but which lacks an inherent data dependent interpretation. To justify the procedure one would have to demonstrate that the observed data, from which the criterion is approximated, form a typical sample of the assumed kind, which generally is an immensely difficult task far exceeding the capabilities of current hypothesis testing techniques. Finally, I do not know enough thermodynamics to be able to comment on the question (3).

I assure Dr Bhansali that the term "honest prediction" is a technical one, and should not be misinterpreted to have moral connotations. Further, although the implications of Theorems 1 and 2 to time series analysis have indeed not been fully discussed, the *MDL* order estimates for the *ARMA* family are consistent, Hannan (1980). Recently, Professor Hannan has shown the same about the predictive *MDL* estimates for the *AR* family. I was not aware of the corrections needed to prove the consistency of the Hannan-Rissanen estimates other than an early amendment. Regarding the asymptotic optimality of the prediction errors, it all depends on which sort of a mathematical situation we analyze. Since any finite amount of data surely can be predicted best by models with finitely many parameters, I find it more realistic to analyze such cases, only, and then the predictive *MDL*, rather than the *AIC* estimates, have been shown to be asymptotically optimal, at least in the analyzed cases. Although other situations may well be analyzed, it is a misconception to think that models with infinitely many parameters were more realistic. As regards spectral estimation, the distance measure corresponding to *AIC* is just another choice, and, of course, such a measure is minimized by the *AIC* estimates. This measure lacks any universality in contrast with the stochastic complexity, which incidentally is also applicable to spectral estimation.

To reply to Dr Fenner's question I can only say that there is an intrinsic meaning, in fact three, to the descriptive complexity, while I know of no way to associate a statistically meaningful interpretation with the computing time complexity.

I disagree with the conclusions drawn by Professor Bernardo in his problem of estimating a single fixed model with a prior for a sure event. The case $d = 1$ in Example 1, indeed, gives $P(1) = 1$, and $I(1) = 0$, from which we conclude that the assumed uniform prior, this time defined over a singleton set, gets placed on the point $P = 1$, and cannot be improved upon. I fail to see why this were an unacceptable proposition. Finally, regarding descision problems I see only advantages in separating the data dependent model fitting portion from the rest. In the case sketched by Professor Bernardo, where overestimation is innocuous, we certainly must try to learn from the data which doses will kill and which do not in order for us to be able to make proper decisions in the future. Although it is true that we do not wish to generate our data by giving overdoses that kill, whatever data we are given the principle of the stochastic complexity still applies, and no added preference structure will improve the findings. More generally, it seems to me that instead of optimizing the decisions against the assumed worst case data generating machinery, as often suggested, we should do better by first seeking for the model or model class with the smallest stochastic complexity and then optimizing the decisions against the result.

As far as my work is concerned, I disagree with the thrust of Professor Geisser's initial statement, for the amount of assumptions needed is the selection of the model class, and even that only provisionally. Further, it seems to me to be better to have an objective and meaningful goal for which to seek approximations in the form of code lengths, which all give upper bounds and hence are comparable,

than to compute an arbitrary and meaningless goal exactly. As to his second statement, the asymptotic claims are proved—not just made. Moreover, some of them such as consistency could not be proved for the other known criteria. The principle of stochastic complexity is indeed unique for the reasons discussed above in my replay to Professor Tong's discussion. It is not meant to be added to Professor Geisser's collection of principles, but to replace them. I already commented on the technical term "honest prediction", to which I was forced because of the abuse of the term "prediction", such as advocated by Professor Geisser, which surely involves an ordering of the data to avoid hindsight prediction. The sequencing requirement, as mistakenly thought by Professor Geisser, is not made in the stochastic complexity. In honest prediction the best predictor may, indeed, not involve a single model.

Professor Hannan brings up tough points, and I may have to concede that there are situations where the objective of the statistical inquiry is so vaguely defined without readily identified models that the techniques applied consist mostly of intuition and hunches. Without personal experience of such problems I do not know whether or not the stochastic complexity principle has anything to offer. Still, the absence of a rational principle in traditional statistics may well have contributed to the currently adopted irrational and aimless procedures in such problems. His second point, where one is merely interested in estimating one particular feature or property from the data rather than the best description, in which all the estimable features participate, is also well taken. Without a full understanding of such a situation it seems to me that the various properties that we model for the data do not compete for the total amount of information, which would imply that still the thing to do is search for the best model we can find and then compute the desired feature from it.

Professor Lindley's criticism in his first paragraph may be condensed in his question "Why not cite the posterior distribution?", which I discussed above and which led me to reject his suggestion to base statistical inquiry upon the posterior distribution. As regards the selection of one preferred model, we want the data to tell us which it is, rather than some subjective utility function, which I would not believe in, in the first place. If I understand the constraints in the data such as provided by the best fitting model, I surely know how to make intelligent decisions for any purpose, and better so than if I had to optimize an artificial utility function without knowing the models. It seems, indeed, that Professor Lindley and I have quite different opinions of what constitute the fundamental problems in statistics. Finally, in reply to his last phrase I note that any significant effect will reduce the stochastic complexity by an amount $O(n)$ while its estimation causes an increase by an amount $O(\log n)$. Hence, all and only the significant effects can be detected.

Turning finally to Professor Mardia's question, an asymptotically optimal approximation of stochastic complexity is as follows:

$$\min_{p,q,\beta,\eta} \{e'\Sigma^{-1}(x, \eta)e + \log |\Sigma(x, \eta)| + (p+1)\log n\},$$

where $e = x - \beta' f(x)$, and p and q denote the number of components in β and η, respectively, and $\Sigma(x, \eta)$ denotes the given covariance matrix. I personally would also compute the predictive code length, even though it is cumbersome, and pick the smaller of the two. For large n the dependence on ordering is insignificant.

Professors Wallace and Freeman: We agree with Professor Dawid that the data distribution implied by a Bayesian mixture of all tenable models and their parameter values leads to the shortest possible expected length. This is our I_0 of Section 3 which is exactly Rissanen's stochastic complexity (2.1). However, use of such a code, or any equivalent such as a predictive code in which each data value is coded according to its distribution conditioned by all preceding data, neither requires nor leads to any model selection or parameter estimation. There is no separation of information into pattern and noise. Consider a binary sequence of unknown probability θ of a "1" with prior uniform in [0, 1]. A predictive code encodes the $(i+1)$th symbol according to the probability $p_{i+1} = (n_i + 1)/(i+2)$ where n_i is the number of 1's in the first i symbols. This code is optimal, of length I_0, but p_{i+1} is not an estimate of θ: it is precisely the probability that, given the prior and the first i symbols, the next symbol will be 1. If the same technique is used to encode a pair of digits at a time, the length used for symbols $i+1$ and $i+2$ if both are 1's is $-\log((n_i+1)(n_i+2)/(i+2)(i+3))$, not $-2\log((n_i+1)/(i+2))$. Further, suppose it is known with certainty *a priori* that θ is equally likely to have any value between 0 and 1, save that it cannot lie between 0.45 and 0.55. If 5 of the first 10 symbols are 1's the predictive code uses $p_{11} = 0.5$ for the next symbol, but this is clearly nonsense if regarded as an estimate of θ. An estimate or model choice occurs only

when the message states it and uses it to encode the data as in done both in our method and Rissanen's model selection criteria.

The message then cannot be expected to be as short as I_0, since, besides the data, it conveys an inductive inference not logically implied by the data. When θ is the union of a set of models $f_k(x; \theta_k)$ of increasing order, construction of a code using $\int h_k(\theta_k) f(x; \theta_k) d\theta_k$ for the best single k implies a choice of model, (Rissanen's first criterion) but no estimation of its parameter. It is potentially misleading to do this and then attempt the estimation of θ_k, since the best parameter value in the model which is best with parameters integrated is not necessarily the same as the best fully-estimated model (Rissanen's third criterion and *MML*).

In his penultimate paragraph, Professor Dawid suggests a greater difference between Rissanen's and our criteria for model selection than perhaps exists. In both Rissanen's second and third criteria, as in *MML*, incorporation of an additional datum requires re-estimation of θ before the code lengths for each model can be computed. Rissanen's first criterion does not, but it requires the usually equally laborious recomputation of the predictive distribution for the next datum, and in fact never yields estimates.

For the former criteria, incorporation of additional data is not very hard: one simply uses Bayes theorem to update the posterior distributions, then applies Rissanen's or the *MML* approach to quantization according to taste. Neither approach actually requires construction of a code.

In response to Professor Whittle we don't regard the coding based on $r(x)$ as ideal in any sense except that it gives a lower bound to the length of a code based on an inference. It makes no inference, and no useful separation of the available information. The additional length of a message based on an inference behaves as $-\log$ (posteror probability of inference).

The $s^2/12$ arises, as in Shepard's correction, from the variance of a uniform distribution of range s. Things get more complicated in many dimensions (Section 5.3).

Our code attempts the separation of information into pattern and random parts. To the extent it is successful (and we have argued it does as well as possible) the first part, i.e. our estimate, will be sufficient for prediction.

We agree with Dr Bhansali that the examples of Section 6 are indeed too simple to show the real advantages of the approach. It is usable in the selection of a model from very large sets of candidates which may differ in order. One simply chooses the model which, when its parameters are estimated, gives the shortest message. For example, in the grammatical inference problem described in Georgeff and Wallace (1984), there were about 10^{10} structurally different models possible.

In Dr Kent's example our choice of D_n depends not only on n and τ but on the prior range of μ in M_1 (or, more precisely, the prior density of μ under M_1 near $\mu = 0$) so, in this sense, the experimenter is indirectly being asked about the value of the threshold.

Dr O'Hagan makes many pertinent comments. The reason *MML* uses the expected information $I(\theta)$ rather than the observed information $I(x, \theta)$ to determine the precision of rounding θ' is that we are considering the length of a real message which must be decodeable by someone knowing only the prior and $f(\cdot; \cdot)$ functions, $\mathscr{X}$ and Θ. If θ' were truncated to a precision depending on x as well as θ', its coded representaton would be unintelligible because, the receiver must decode $\hat{\theta}$ before she can decode x. $I(\theta)$ of course can be computed from the prior knowledge available to the receiver, so its use avoids this problem. In cases where $I(\theta)$ is a very bad predictor of $I(x, \theta)$, or perhaps does not even exist, one must use a precision quantum based on $I(x, \theta)$, but the statement of $\hat{\theta}$ must then be preceded by a coded statement of s, which plays the role of an ancillary statistic.

Why not expand $\log h(\hat{\theta})$? Dr O'Hagan has perceptively exposed a short-cut in our presentation. For θ a single scalar parameter, he suggests that if $h(\theta)$ has a postive second differential, the optimum s should be larger than we have indicated, since the advantage of a large s in shortening the first part of the message will be greater. However, our expression $-(s^2/24)E(\partial^2/\partial\theta^2) \log f$ is then also wrong, since the distribution of $\theta' - \hat{\theta}$ will no longer be a uniform density, instead having approximately the same second differential as $h(\theta)$. The effect is to increase the average of $(\theta' - \hat{\theta})^2$ above $s^2/12$, removing the advantage of a larger s. Rather than pursuing a higher-order analysis, remember that we are attempting to find the optimum size (and for $p > 1$, shape) of the region of Θ within which θ' may be rounded to $\hat{\theta}$.

The optimum region minimizes

$$\text{Av. over region } (D(\theta, \hat{\theta})) - \log(\text{total prior in region}),$$

where $D(\theta, \hat{\theta}) = E(\log f(x; \hat{\theta}) - \log f(x; \theta))$ and the average is taken on the measure $h(\theta)d\theta$.

The approach of our paper approximates $D(\theta, \hat{\theta})$ by $-\frac{1}{2}(\theta - \hat{\theta})^2 E\left(\frac{\partial^2}{\partial\theta^2} \log f(x; \theta)\right) = \frac{1}{2}(\theta - \hat{\theta})^2 I(\hat{\theta})$,

its average in the region by $(s^2/24)I(\hat{\theta})$, and the total prior by $sh(\hat{\theta})$. Instead, one can show that, with no approximations, the ideal region for some $\hat{\theta}$ is the smallest region such that for any θ on its boundary $D(\theta, \hat{\theta}) = 1 + \text{Av. over region}(D(\theta, \hat{\theta}))$, which condition suffices to define the ideal region.

This condition depends on $h(\theta)$ only via the average, and is thus invariant under transformatons of θ which preserve $h(\theta)d\theta$, in particular the transformation which makes $h(\cdot)$ uniform. The optimum so defined is not achievable in general for $p > 1$, since the regions around the different values of $\hat{\theta}$ have to fit together. The analysis of the paper ignores the differentials of $h(\theta)$ in order more simply to exhibit the effects of region packing.

We indeed propose θ', not $\hat{\theta}$, as the useful *MML* estimator, and the whole purpose of proposing *MML* is favour of *SMML* is to avoid having to use $\hat{\theta}$. Finally, note that *MML* is derived by approximating the strictly-optimum *SMML*, and so we require *MML* to retain the key properties of *SMML* such as invariance.

Dr O'Hagan's quantity (*A*), intended to measure the probability content of a mode of the posterior density, does not behave like a posterior probability. As he points out, it is possibly infinite for some θ and is sensitive to the coordinate system. The quantity which we maximise is not derived from the posterior density and does behave like a posterior probability. The resulting attractive features which Dr O'Hagan recognises seem sufficient excuse for an innocent little violation of the likelihood principle.

In reply to Dr Fenner, trying to incorporate a measure of coding effort into a measure of code length makes economic sense but the economic balance is technology-dependent. If, following Solomonoff we extend the coding approach to general scientific inference, we may account for the continued use of Newtonian mechanics for many applications in favour of Relativistic mechanics as arising from the much lower coding effort required by the former, which economically outweighs the minute increase in 'message length' from its lack of accuracy.

We beg to differ from Professor Bernardo's views about improper priors. If $|I(\theta)|^{1/2}$ cannot be normalized to give a proper distribution, it cannot be the representation of prior knowledge about θ, and such a case need not be considered. Where a proper prior proportional to $|I(\theta)|^{\frac{1}{2}}$ can be adopted the MML estimator does indeed coincide with maximum likelihood, but it will not in general lead to an indefinitely long message. Regarding loss functions, we agree that when the object of the game is to make some decision expressible as a choice of θ and there is a relevant loss function, then choosing θ to minimize expected loss is sensible. However, we dispute that the chosen $\hat{\theta}$ can properly be regarded as an estimate of θ, i.e. a credible assertion about θ. It can easily be that $\hat{\theta}$ has zero prior density and/or zero likelihood. We regard *MML* estimation as motivated much as is scientific inference: an attempt to derive a useful general assertion about the state of the world from finite and uncertain data. We do not expect the estimate to be true in any absolute sense, any more than any theory in science is expected to be true, but we do expect it to envapsulate a great deal of the apparently non-random behaviour of the data. We have argued that *MML* is optimum in this sense.

We agree with Professor Geisser that, if the prediction of a further observation is the sole objective, a Bayesian mixture of all tenable models is hard to beat. But is this really inferring anything about the source of the data? See our response to Professor Dawid. Also, would we be happy with a scientist who proposed a Bayesian mixture of a countably infinite set of incompatible models for electromagnetic fields? See our response to Professor Bernardo.

Our intuitive preference for log n criteria is based on their superior probabilities of selecting the correct model when there is one and on consistency. We quite agree that their predictive performance may well not be markedly better.

Incidentally, we too value plausibility more highly than honesty.

We agree with Professor Lindley that in the Bayesian framework the posterior distribution represents the limit beyond which deduction can usefully be carried no further, without using a utility function. However, science and indeed all human inquiry advances through induction beyond this point. The need to deal with a complex world requires us to adopt as working hypotheses general assertions which cannot be deduced from the available data but which capture the regularities we think we have seen in the data. We have tried to carry Bayesian inference this additional step and propose what we believe to be a sound principle for inductive inference of a specialised kind which as nearly as may be captures these patterns within the framework of a statistical model.

Professor Mardia's question about the presence of autocorrelations raises no problems in principle. The *MML* would consist of quantized estimates of all parameters followed by the data values encoded sequentially according to their distributions conditional on all preceding values. This could get extremely complicated, but we are not, after all, concerned with constructing the code, only with measuring its

length and that can be done by computing the joint probability of all the data, which is not particularly difficult.

Computing the Fisher information would be complicated, but as we point out in the paper, the observed information can be used instead provided a very small prefix is added to the message stating the resulting precision quantum. In fact, in problems where no single even roughly sufficient statistic exists, this is the only sensible procedure.

We agree, of course, with everything Dr Patrick says but his contribution gives both of us the opportunity to acknowledge the stimulating discussions we have had with him about information measures over the years.

Finally, it is a great pleasure to thank all discussants. The points they raised will keep us busy on further work for a long time to come.

REFERENCES IN THE DISCUSSION

Akaike, H. (1978) Comments on 'On model structure testing in system identification, *Int. J. Control*, **27**, 323–324.

Clayton, M. K., Geisser, S. and Jennings, D. E. (1986) A comparison of several model selection procedures. *In Bayesian Inference and Decision Techniques.* (P. Goel and A. Zellner, eds), pp. 425–439. Amsterdam: Elsevier.

Dawid, A. P. (1985) Calibration based empirical probability (with Discussion). *Ann. Statist.*, **13**, 1251–1285.

Geisser, S. (1975) The predictive sample reuse method with applications. *J. Amer. Statist. Ass.*, **70**, 320–328.

———In discussion of Bernardo's paper. *J. R. Statist. Soc.* B, **41**, 136–37.

Hannan, E. J. and Kavalieris, L. (1984) A method of autoregressive—moving average estimation. *Biometrika*, **72**, 273–280.

Kent, J. T. (1983) Information gain and a general method measure of correlation. *Biometrika*, **70**, 163–173.

Mardia, K. V. and Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135–146.

O'Hagan, A. (1985) Shoulders in hierarchical models. In *Bayesian Statistics 2* (J. M. Bernardo *et al.*, eds) pp. 697–710. Amsterdam: Elsevier.

———(1987) Exploring a high-dimensional posterior density. *Comp. Statist. Quart.*, in the press.

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

———(1981) An optimal autoregressive spectral estimate. *Ann. Statist.*, **9**, 300–306.