Bayes not Bust!

Why simplicity is no problem for Bayesians ⋆

David L. Dowe, Steve Gardner and Graham Oppy *

ABSTRACT

The advent of formal definitions of the simplicity of a theory has important implications for model selection. But what is the best way to define simplicity? Forster and Sober ([1994]) advocate the use of Akaike's Information Criterion (AIC), a non-Bayesian formalisation of the notion of simplicity. This forms an important part of their wider attack on Bayesianism in the philosophy of science. We defend a Bayesian alternative: the simplicity of a theory is to be characterised in terms of Wallace's Minimum Message Length (MML). We show that AIC is inadequate for many statistical problems where MML performs well. Whereas MML is always defined, AIC can be undefined. Whereas MML is not known ever to be statistically inconsistent, AIC can be. Even when defined and consistent, AIC performs worse than MML on small sample sizes. MML is statistically invariant under 1-to-1 re-parametrisation, thus avoiding a common criticism of Bayesian approaches. We also show that MML provides answers to many of Forster's objections to Bayesianism. Hence an important part of the attack on Bayesianism fails.

Key words: Minimum Message Length, MML, Bayesianism, simplicity, inference, prediction, induction, statistical invariance, statistical consistency, efficiency, model selection, point estimation, information theory, Akaike Information Criterion, AIC, predictive accuracy

---

⋆ The title is the third in a sequence: the title of (Earman [1992]) asked 'Bayes or Bust?'; in the title of his review of Earman's book in the pages of this journal, Forster ([1995]) affirmed that 'Bayes and Bust'. Now comes our dissent: 'Bayes not Bust!'

* Corresponding author. The authors are listed alphabetically.

1 Introduction

'Pluralitas non est ponenda sine necessitate,' said William of Occam in the 14th century, 'We should not posit plurality without necessity.' In modern times, Albert Einstein is said to have expressed much the same thought this way: 'Our theories should be as simple as possible, but no simpler.'

But what is simplicity in a theory? Historically, most attempts to understand simplicity have tried to connect it up with aesthetic concepts such as beauty and elegance. This approach has not been entirely satisfactory because these concepts appear to be of a familiar type that bedevils philosophers: even though we think we know them when we see them, they seem hard, if not impossible, to define.

Still, the aesthetic appreciation of scientific theories has remained widespread. This has had two important effects in the philosophy of science: firstly, it has made it seem as though the question of what makes one theory simpler than another does not have an objective answer. The argument runs roughly: simplicity is a form of beauty; beauty is a matter of taste, and there's no accounting for taste. Secondly, it has driven some philosophers of science to search elsewhere for ways to distinguish good theories from bad. Popper and Hempel, both of whom tried to distinguish among theories on the basis of their logical entailments, come especially to mind in this connection.

It often happens, however, that discoveries in other fields of enquiry can dramatically change the way we look at certain philosophical questions, and here we encounter a rather startling example of this. For it turns out

that the relatively new discipline of information theory allows us to say, with mathematical precision, exactly how simple a theory is. Not only that, it tells us exactly when we should prefer more complex theories to their simpler alternatives. To gloss the theory very briefly at the outset, here's what it says: the best theory to infer from the data is the one that can be stated with the data in a two-part message of the shortest length. This is called the Minimum Message Length (MML) principle, and we owe it largely to the work of Chris Wallace. One aim of this paper is to explain this result, and its implications for philosophy of science.

The implication we are particularly interested in is the bearing this result has on the debate about Bayesianism in the philosophy of science. This complex and ongoing debate is difficult to summarise, but we'll try to characterise the opposing positions briefly here.

Bayesians hold that all of the important beliefs, attitudes and intuitions that we have about scientific theories can be expressed in terms of probabilities that certain propositions are true; that Bayes's Rule of Conditionalisation (for hypothesis H and evidence E)

$$\text{posterior}(H) = p(H|E) = \frac{p(H)p(E|H)}{p(E)}$$

tells us how we should update our beliefs in light of new evidence; and that (therefore) some knowledge of, or reasonable assumptions about, our prior knowledge of a situation ($p(H)$ in the above) is indispensable in the calculation of what we should believe about it in the light of evidence E.

Anti-Bayesians deny all of this: they hold that important aspects of our attitudes towards scientific theories cannot adequately be captured by

any statement expressed in terms of the probability that a proposition is true; it follows that there are situations in which Bayes's Rule is of no use to us, since the Rule can only be applied if our prior knowledge of a situation can be stated in terms of probabilities; but that is no matter, since we have other tools that do not involve probabilities that tell us what we should believe about scientific theories—the use of priors is therefore not indispensable.

The connection between these two issues—on the one hand, the existence of formalised notions of theoretical simplicity, and on the other, the arguments about Bayesianism—isn't obvious. To understand it, it helps to go back to a paper written in 1994 by Malcolm Forster and Elliott Sober. In that paper, they described and defended a different formalisation of the notion of simplicity, based on the work of Akaike, than the one we wish to defend. A detailed explanation of the difference between Akaike's Information Criterion (AIC) and the Minimum Message Length (MML) principle will have to wait until later. The key point to make here in the Introduction is that AIC is a non-Bayesian technique, making no essential use of conditionalisation or of priors; by contrast, MML is a Bayesian technique that does make essential use of conditionalisation and of priors.

With that in mind, Forster and Sober's argument can be summarised as follows: a formalised notion of the simplicity of a theory would be a great breakthrough in the philosophy of science. Akaike's Information Criterion provides the best way of formalising the notion of the simplicity of a theory. But AIC is a non-Bayesian technique. We should conclude, therefore, that the best philosophy of science is non-Bayesian.

The counter-argument presented in this paper goes like this: we agree completely about the significance for philosophy of science of formalised notions of simplicity. But we shall argue that AIC is a demonstrably inadequate way of formalising that notion, and that the Minimum Message Length principle provides a much superior formalisation, one that performs better than AIC in every empirical test that we have tried. Since MML is a Bayesian technique, we should conclude that the best philosophy of science is Bayesian.

The theoretical arguments mostly come first. In section 2, we define the curve-fitting problem, the method of maximum likelihood, and the problem of over-fitting. In section 3, we describe Akaike's Information Criterion (AIC) for doing model selection, and in section 4, the broader Predictive Accuracy framework into which Forster places AIC. Section 5 introduces the Minimum Message Length (MML) Principle, and describes the construction of the Strict MML estimator, a language-invariant Bayesian estimator which can be used for both model selection and parameter estimation.

In section 6, we exhibit direct comparisons between AIC and MML on several different kinds of statistical inference problem. The comparisons are all in favour of MML over AIC. We include examples of cases which MML handles well, but for which AIC gives statistically inconsistent answers. Of special interest is the Neyman-Scott problem (section 6.1.3), for on this problem it turns out that aiming for predictive accuracy leads us to give statistically inconsistent estimates. We also give examples where it appears that AIC gives no answer at all.

Finally, in section 7 we show how the MML Principle meets two of Forster's oft-repeated objections to Bayesianism, the sub-family problem, and the

problem of approximation. Our conclusion is that for philosophers of science, Bayesianism remains the best—and perhaps the only—game in town.

2 The Curve Fitting Problem

The best way to engage with Forster and Sober's 1994 argument is to follow their exposition as far as possible, and then show where we diverge from them. Since they begin their paper by describing what they call the 'curve fitting problem', we shall do likewise.

The most general form of the curve fitting problem arises in many experimental contexts. We have some data, which we can plot on a Cartesian plane, with x- and y-axes. We represent a hypothesis about how the data was produced by some function, which maps x-values onto unique y-values. For example, the hypothesis that there is some specific linear relationship between the x and y values is represented by the function

$$y = a_1 x + a_0$$

where a_1 and a_0 are constants (or co-efficients), giving respectively the gradient of the line, and the point where the line intersects the y-axis.

2.1 Curves and families of curves

Without specifying the value of the co-efficients, we haven't picked out a *specific* function, but rather a *family* of functions, here, the family of straight lines.

Likewise, the hypothesis that there is quadratic relationship between the x and y values is represented by the function

$$y = a_2x^2 + a_1x + a_0$$

which picks out the family of parabolas.

2.2 Noise

In a perfect world, experiments would be free of noise. In such a world, if the true curve were a straight line, the data would fall exactly on that straight line. But data is usually noisy, so even if the linear hypothesis is correct, it doesn't follow that our data will fall on a straight line. Rather, while the data will tend to be close to a straight line, they will be distributed above and below it.

It is typical to assume that noise is random with a Gaussian distribution of unknown variance, σ^2. We can therefore represent our hypotheses about the observed relationship between the x and y values by adding another term to our functions, e.g.:

$$y = a_1x + a_0 + N(0, \sigma^2) \qquad \text{or} \qquad y = a_2x^2 + a_1x + a_0 + N(0, \sigma^2)$$

2.3 The Method of Maximum Likelihood

Let's imagine just for the moment that we have some noisy data, and that we know that the linear hypothesis is correct. We should like to know exactly which straight line from the family of straight lines should be our

best guess. We would also like to estimate how noisy the data is. This is called *parameter estimation*.

One answer is that we should choose the line to which the data is closest. We can measure the distance of some data from a line (or curve) by using the least squares method.

This least squares method is an instance of a more general method—the method of Maximum Likelihood (ML). The ML method says that if you want to know which curve is the most likely to be true, choose the curve which would have made the observed data most likely.

Note that it is important not to get confused between $Pr(H|E)$, the probability that a hypothesis is true given the evidence, and $Pr(E|H)$, the probability of observing the evidence, given the hypothesis, which we (following statistical usage) call the *likelihood*.

2.4 Maximum Likelihood and Over-fitting

However, there's a problem with the ML method. In the above, we assumed that we knew that the linear hypothesis was correct. What if we don't know that? What if we have to do *model selection*, as well as parameter estimation?

We can use the ML method to find the maximum likelihood straight line from the family of straight lines. But equally, we can use the ML method to find the maximum likelihood parabola from the family of parabolas, the maximum likelihood cubic from the family of cubics, etc.

Moreover, the best parabola will always have at least the likelihood, and usually greater likelihood, than the best straight line, and the best cubic at least the likelihood of the best parabola, and so on.

In fact, if we are prepared to choose a polynomial of sufficiently high degree, we can choose a curve whose distance from the data is zero. If we have N data points, a polynomial of degree $N - 1$ is sufficient for this.[1]

Yet we don't want to say in general that our best guess at the true curve is a polynomial of degree $N - 1$. To say that is to confuse signal with noise. This is the *problem of over-fitting*: by concentrating on minimising the squared distance between the true curve and the actual data observed, the ML method gives too much weight to the data; it is, in general, insufficiently cautious, willing to spuriously over-fit weak or non-existent patterns.

3 Akaike's Information Criterion (AIC)

How do we justify our preference for a curve of lower likelihood? We appeal to *simplicity*. As we suggested in our introduction, before the middle of the 20th century, this seemed to be an appeal to aesthetic criteria, an appeal to something beyond the data.

However, during the last 50 years, several different proposals have been advanced for defining the simplicity of a theory in precise mathematical terms.

Forster and Sober describe and defend one of these proposals, Akaike's Information Criterion (AIC) (Akaike [1973]).

AIC is a proposal for doing *model selection*, i.e., picking the right family of curves. If F is a family of curves, $L(F)$ the maximum likelihood member of that family, and k the number of free parameters (co-efficients) in the family, then according to AIC we should aim to minimise this quantity:

$$-2 * \text{log-likelihood}[L(F)] + 2k$$

The proposal can be described as a *penalised maximum likelihood* function. The first term says that the likelihood of a family of curves goes with its maximum likelihood member. The second term then corrects this with a penalty proportional to the complexity of the family. Akaike's proposal is only one of many such proposals, which differ in how they propose to penalise the ML estimate of the goodness-of-fit of a family of curves.

4 The Predictive Accuracy Framework

Forster ([2002], p.S160) defines a *framework* for philosophy of science by asking three questions:

(1) What goal, or goals, can be achieved in science?

(2) What possible means, method, or criterion, can achieve the goal?

(3) What *explanation* is provided of how the means tends to achieve the goal? Is there any account of the means → goal connection?

Forster places Akaike's work within what he calls the *predictive accuracy framework*: the postulation of the goal of predictive accuracy as the goal of science. 'Predictive accuracy' is the term coined by Forster and Sober ([1994]) to describe the goal of maximising the expected log-likelihood of re-sampled data (that is, future data sampled from the same source as the data

we have already). This is equivalent to minimising the expected Kullback-Leibler distance, a concept we explain in section 6.1.4.

Before proceeding any further, it's worth commenting on an interesting slip between questions (1) and (2) in Forster's framework. In the first question, Forster asks what goal or goals can be achieved in science? We think this way of putting it, which allows for the possibility of multiple goals, is right. But in the second question, the reference is to 'the goal', singular, and Forster's subsequent arguments defend only predictive accuracy as the single goal of science. So it's not clear whether Forster really believes that there could be multiple legitimate goals for science. We certainly do.

Here's how the Bayesian/MML approach to statistics looks in terms of Forster's framework: whereas Forster postulates predictive accuracy as the single goal of science, and AIC as the means of achieving it, we offer inference to the most probable theory as a goal of at least comparable importance, and the method of minimising message length as the means of achieving it. The arguments we give in this paper are intended to serve as an account of how minimising message length achieves the goal of inference to the most probable theory, a goal which Forster claims Bayesians cannot achieve. In our view, prediction remains an important goal. But we claim further that knowledge of the most probable theory gained from inference using MML can be used to make excellent predictions, much better than those made by AIC.

Two further points should be mentioned briefly here: Forster claims that Akaike's criterion provides a general means of achieving the goal of predictive accuracy. But he is mistaken about this, as we show in section 6.1.4. In

an interesting class of cases where the amount of data per parameter is bounded above, you cannot achieve predictive accuracy by using Akaike's criterion.[2]Secondly, these cases cast doubt on the goal of predictive accuracy as the single overarching goal of science, for they show that maximising predictive accuracy can lead to statistical inconsistency in inference. This demonstrates an importance difference between the goals of inference and prediction in science.

5 The Minimum Message Length (MML) Principle

Before proceeding with our criticisms of AIC, it would be well to have before us the alternative proposal we are defending, which is derived from the principle of Minimum Message Length.[3]According to the principle, we should infer the theory that allows the data to be stated in the shortest two-part message, where the first part of the message asserts the theory, and the second part of the message encodes the data under the assumption that the asserted theory is true.

The fundamental idea is that *compact coding theory* provides the right framework in which to think about inference and prediction (Wallace and Boulton [1968]; Wallace and Freeman [1987]; Wallace and Dowe [1999a]; Wallace [2005]). Begin by thinking of the data as a string of symbols in a finite alphabet. Given an estimate of parameters, we may be able to get a briefer encoding of our data under the assumption that the estimated parameters are the true values. A given model is only worth considering if the shortening of the encoded data string achieved by adopting it more than compensates for the lengthening caused by the quotation of the estimated

parameters. Within a given model, the preferred parameter estimates are those that lead to the shortest total encoded length. And the preferred model amongst a class of models is the one with the shortest total two-part message length (minimised with respect to its parameter estimates). The method is Bayesian because it assumes known proper prior distributions for unknown quantities. The method comes in various varieties: we shall describe Strict MML, even though this is not computationally tractable, except in special cases. The computationally tractable MML (Wallace and Freeman [1987]) is derived from a quadratic Taylor series approximation of Strict MML, and shares many of the desirable features of Strict MML.

5.1 The Strict MML estimator

A point estimation problem is a quadruple $\{H, X, f, p\}$:

H is a parameter space (assumed to be endowed with a σ-field of subsets).
X is a set of possible observations $\{x_i : i \in N\}$.
f is a given prior probability density function with respect to a measure dh on the parameter space $H : \int_H f(h)dh = 1$.
p is the known conditional probability function $p : (X, H) \rightarrow [0, 1]$: $p(x; h) = p(x|h)$, where $\sum_i p(x_i|h) = 1$, for all $h \in H$.

A solution to a point estimation problem is a function $m : X \rightarrow H : m(x) = h$, which given some possible observation, tells you which theory to infer from it.

A Bayesian solution to a point estimation problem makes essential use of Bayes's Theorem:

$$f(h|x) = \frac{p(x|h).f(h)}{\int_H p(x|h).f(h)dh}$$

Conditionalising on observations, we can obtain $f^*(h)$, the posterior probability density function, from $f(h)$, the prior probability density function.

If a cost-function is known which expresses the cost of making an estimate h' when the true value of the parameter is h'', then standard decision theory allows us to calculate a minimum expected cost estimate (and so we would have a solution to our point estimation problem).

In the absence of a cost-function, it is not clear how to proceed. Given that the parameter space is continuous, the probability of any individual hypothesis is 0. So we can't use Bayes's Theorem in order to calculate point estimates. We might think that we can derive a point estimate from $f^*(h)$ by choosing that value of h which maximises the posterior density. Alas, however, any such "estimate" is dependent upon parametrisation.

Forster ([1995]) calls this the 'problem of language variance' and suggests that, in light of considerations such as those just given, the game is more or less up: there is no satisfactory Bayesian statistics, or, at least, none that he knows of.[4]

However, all is not lost! Wallace and Boulton ([1975], 'An Invariant Bayes Method for Point Estimation') describe a language invariant Bayesian solution to the point estimation problem. Wallace, with the cooperation of various co-workers, has since gone on to develop this Bayesian method into an enormously powerful and well-justified approach to inference and

prediction. Here, we shall just outline a version of the argument from (Wallace and Boulton [1975]). We give a more technical (although still brief) exposition in Appendix A, while a thorough exposition of the argument can be found in (Wallace [2005], Ch. 3).

Strict MML gives us a way of dealing with any continuous or discrete prior and still ending up with a Bayesian point estimator which is statistically invariant and *in some sense* (see below) maximises the posterior *probability*.[5]

What we would like to be able to do is to choose the hypothesis with the highest posterior probability. While we cannot do this in the continuous case—for the reasons given above—we can do it in the discrete case. So, the guiding idea is that we should consider a discrete problem that is a close enough approximation to our initial point estimation problem.

To outline Strict MML, our initial objective is to construct a *codebook* with the shortest expected length of a two-part message. The first part of the message asserts a theory, while the second part of the message encodes the data under the assumption that the theory is true. The codebook will tell us, for any possible observation, which estimate allows the briefest encoding of theory and data.

By virtue of the fact that all data is recorded to finite accuracy, each of the countably many observable data has a *probability* (rather than a density) of occurring. So, given the relevant likelihood functions and Bayesian prior distributions on the parameters, we can calculate a marginal probability, $r(x_i)$ of each possibly observable datum, x_i. We note that the sum over all data of the marginal probabilities $r(x_i)$ equals 1.

We partition the data into groups, always balancing the expected lengths of the first and second parts of the message. The number of possible data is countable, and clearly so is the number of groups. Every time a new datum joins a group the prior probability of that group's being chosen goes up and the message length to encode the parameters of that group correspondingly goes down. On the other hand, the new datum will almost certainly cause a change in the group's parameter estimates, not decreasing and almost certainly increasing the expected length of encoding the values of the previous group members using the parameter estimates.

Strict MML chooses the partition which results in the shortest expected length over all possible two-part messages. The expectation is calculated using the Bayesian prior probability and the marginal probability, $r(x_i)$, of each possibly observable datum, x_i. A codebook is made from this with code-lengths, l_i, of events of probability, p_i, given by $l_i \approx -\log p_i$.[6]

The data is partitioned so that each possible datum appears in exactly one group, and each group is assigned a point estimate tailor-made to best fit (on weighted average) its group members. More explicitly, if the codebook partitions the data into J groups $\{c_j : j = 1, ..., J\}$, then the point estimate h_j for each group c_j is chosen to maximise

$$\sum_{i \in c_j} r(x_i) f(x_i|h_j). \tag{1}$$

Each group can be thought of as having a prior probability equal to the sum of the marginal probabilities of all the data in the group—and, as such, the prior probabilities of all the groups must sum to 1. We shall refer to this prior probability of the groups (and, in turn, their parameter estimates) as the *coding prior*. The partition of the possible data into groups, and the

use of the the coding prior, together constitute an acceptable approximate discretisation of the original continuous point estimation problem.

It is important to note that the SMML codebook described above is constructed prior to the observation of any data, and depends only on the likelihood functions and on Bayesian priors.

For every group and for every possible datum, x_i, two-part messages exist which encode the parameter estimates of the group in the first part of the message, followed in the second part of the message by the x_i given those parameter estimates.[7]Once we have observed a datum, choosing the estimate is simply a matter of finding out to which group the observed datum is assigned by the codebook, and choosing the estimate for that group. This is equivalent to using the coding prior and taking the Maximum A Posteriori (MAP) estimate.

5.2 An example: the Binomial distribution

We will very briefly discuss an example to make clearer how the Strict MML method works in practice.[8] The problem is that of the Binomial distribution: given a sequence of N independent trials each giving success or failure with unknown probability of success, p, and a prior distribution $h(p)$, we are to estimate p. Let $N = 100$ and assume a uniform prior $h(p) = 1$ over the possible values of p. Then the possible observations are just the set of the binary strings of length 100. The partitioning of the possible data into groups can be carried out according to an algorithm due to Farr (Farr and Wallace [2002]), and results in the formation of ten groups, each represented by a single estimate. One of the two possible mirror-image solutions for the

groups and estimates is shown in Table 1.

[Table 1 about here.]

We can see that the initial continuous, uniform prior $h(p) = 1$ has been transformed into a discrete coding prior containing just ten possible estimates. The number of groups may seem surprisingly small and the widths of the groups surprisingly wide. For example, any experiment which yields between 33 and 49 successes will result in the same estimate, $h_5 = 0.41$. But in fact $h_5 = 0.41$ is a plausible value to infer for p if the number of successes lies in that range. As Wallace ([2005], p.160) notes, the probability that 100 trials with $p = 0.41$ would yield exactly 41 successes is 0.0809, whereas the probabilities of 33 and 49 successes are respectively 0.0218 and 0.0216, over a quarter of the most probable value. The spacing of the estimates is consistent with the expected error in their estimation, an important point to which we return in section 7.2.

Table 1 also shows the difference between Strict MML and Maximum A Posteriori (MAP) estimation. For the Binomial distribution with N trials and s successes, and a uniform prior, the MAP estimate in this parametrisation is equal to the Maximum Likelihood estimate, and is given by s/N. The SMML estimate can differ significantly from this value, because of the way the SMML procedure maps different possible observations to the same estimate.

5.3 Properties of the SMML estimator

5.3.1 Bayesianism

That the method described above is Bayesian is perhaps apparent enough from the essential use it makes of prior probabilities. The length of the first part of the message, asserting the theory, is determined by our prior probabilities.

However, there is a more fundamental connection between Bayesianism and the principle of minimum message length. Recall from the Introduction our summary of Bayesian commitments: (1) the indispensability of probabilities for characterising the degree of belief in a theory; (2) use of the Rule of Conditionalisation to update our beliefs about a theory in the light of new evidence; (3) the indispensability of priors. It is possible to view the principle of minimum message length as providing *independent support* for these Bayesian principles.

The connection is made via Shannon's theory of information, in which information is equated with the negative log of a probability. At first it might seem that this objective measure of the information content of a theory has little to do with our subjective degree of belief in the theory. However, as Wallace ([2005], p.79) points out, the information content of a message (or theory) is a subjective notion: a message that tells us something we already knew conveys no information, while a message that tells us something we thought improbable tells us a great deal. This provides strong support for the idea that what we believe about theories is best characterised in terms of (subjective) probabilities. This is the first Bayesian

principle.

Next, let's look at Bayes's Rule. Suppose that you accept that the best theory is the one that can be stated in the shortest two-part message, where the first part of the message asserts the theory, and the second part of the message encodes the data under the assumption that the theory is true. For given data x, let $h = m(x)$ be the hypothesis chosen by the SMML estimator as best explaining that data, and $q(h)$ be the coding prior probability assigned to this hypothesis. Then the length of the message asserting the theory and encoding the data is given by $-\log(q(h)f(x|h))$, the negative log of the joint probability of the estimate and the data.

On the other hand, the length of a message optimally encoding the data, but without making any attempt to explain it, would be given by $-\log r(x)$, the negative log of the marginal probability of observing the data. The difference between these two quantities

$$-\log \frac{q(h)f(x|h)}{r(x)}, \tag{2}$$

is formally identical to Bayes's Rule of Conditionalisation.[9]In other words, choosing the SMML estimate leads to a degree of belief in the estimate chosen which is exactly described by Bayes's Rule.

Finally, the indispensability of priors is manifested in the choice of encoding of the first part of the message asserting the theory. So far, we have emphasised that the choice of prior determines the optimal encoding of the assertion. But in Shannon information theory, message lengths and negative log probabilities are interchangeable. So one can equally say that a choice of encoding (i.e., the choice of message length for the first part of the message)

implicitly asserts a prior probability distribution over the theories being considered, with Pr(asserted theory) = $2^{-(\text{length of encoded assertion})}$.

Imagine that you want to send a two-part message to a receiver. The question naturally arises as to how to encode the first part of the message, asserting the theory. You and the receiver of the message must agree to use some language which you both regard as reasonably efficient for the encoding of the theory, in the sense that theories which are more likely to be asserted will be given shorter encodings in the language. In adopting such a language, you and the receiver implicitly assert a prior probability distribution over the space of possible theories.

It is true, and worth noting, that the prior probability implicitly asserted by the choice of encoding may differ in a number of ways from the kinds of prior probability distributions usually considered in traditional Bayesian statistics. In the first place, such a prior may not be proper: the different possible theories may not be assigned probabilities that collectively sum to 1. In the second place, the prior may not have an easily expressible mathematical form. For these reasons, Wallace ([2005], pp.148–49) draws a distinction between prior probabilities and *coding probabilities*, where the latter are probabilities implicitly asserted by a choice of encoding for the first part of the message. Wallace notes that careful thought must be given to the choice of encoding, with regard to what coding probabilities are being implicitly asserted. He also shows that, if a code can be constructed that assigns longer strings to implausible theories and shorter strings to plausible theories, then such a code may well be an acceptable summary of vague prior beliefs. Even if the coding probability distribution, considered as a prior, is strictly speaking improper, its use in practice will often lead to

acceptable, even excellent results.[10]

5.3.2 Language invariance

The SMML estimator is language invariant. That is, both the estimator and the message length are unchanged by any one-to-one measure-preserving transformation in the parametric representation of the model. It is thus immune to the 'problem of language variance' often raised by Forster (see for example [1999]; [1995], sec. 5) as a general objection to Bayesian statistics. The model invariance of the SMML estimator follows from the fact that transformations of the model space do not affect the model distribution, and that the prior enters into the calculation of the message length only via the marginal probability of the data. Hence, if we use an appropriately transformed prior, a change in the representation of the model space has no effect on message length.

5.3.3 Generality

The SMML method (or approximations to it, see section 5.5 below), can be used for a wide variety of problems. In the first place, it is applicable equally to problems of parameter estimation and model selection. This unified treatment can be regarded not only as a strong theoretical virtue, but one which gives demonstrably better results in practice, as we show below. Many other methods are restricted in the classes of models to which they can be applied. The Maximum Likelihood method requires the set of possible models to be either countable or a continuum of fixed dimension. That is, it cannot directly be used to choose among models with different

numbers of parameters. Akaike's method cannot be applied to models with non-real valued parameters.[11]By contrast, the SMML method requires only that (a) the data can be represented by a finite binary string; (b) there exists a language for describing models of the data which is agreed to be efficient, i.e., there exists a prior density $f(h)$; (c) the integrals $r(x)$ exist for all possible data values, and satisfy $r(x) > 0$, $\sum_X r(x) = 1$.

5.3.4 Consistency and efficiency

An estimator is statistically consistent if it converges on the true distribution given enough data. It is efficient if the rate of convergence is as fast as possible. Starting from the basic results of information theory it can be shown that because the SMML estimator chooses the shortest encoding of the model and data, it must be both consistent and efficient.[12]These results are (a) that the expected message length is minimised when the asserted probability distribution agrees with the distribution of the source from which the data actually comes, and (b) that when data from some source is optimally encoded, the encoded string has the statistical properties of a random sequence.

Here's the argument. The SMML method separates data into pattern and noise. The pattern, which describes all the information relevant to the quantities we are estimating, is encoded in the first part of the message. Anything that cannot be deduced from the pattern is encoded in the second part.

Suppose on the one hand that the asserted pattern does not contain all of the pattern information which is present in the second part. Then the

second part of the message will contain some pattern information and so cannot be a random sequence. In that case, there must exist some shorter encoding of the second part, violating the assumption that the SMML estimator chooses the shortest encoding.

Now suppose, on the other hand, that the asserted pattern contains some noise, i.e., information not relevant to the quantities we are estimating. Since the estimator is a deterministic function of the data, this information must be recoverable from the data. It follows that the noise information in the first part of the message is redundant, since it can be deduced from the second part of the message. Once again the assumption that the SMML estimator chooses the shortest encoding of the data is violated.

The above argument shows that the SMML assertion contains all and only the information relevant to knowledge of the true model that can be extracted from the data.[13]

5.4 Similarity to false oracles

Wallace ([1996]) defines an *oracle* to be an estimator which, regardless of the data, always gives the true parameter value (or selects the correct model). A *false oracle* is an estimator such that no fair criterion can be expected to distinguish between it and an oracle. While we cannot expect to have access to (true) oracles, Wallace shows that we can construct false oracles, which is the next best thing. He shows firstly that *sampling from the posterior distribution* is a false oracle. Wallace ([1996], p.307) notes, 'This may seem a strange rule compared with, say, choosing the mean, median or mode of the posterior, but it has the advantage of being invariant

under arbitrary measure-preserving transformations of the parameter space.' He then shows that the Strict MML estimator closely approximates the behaviour of a false oracle.

5.5 Approximations to SMML

The basic idea of the Strict MML estimator developed above is that one partitions the set of possible observations into an exhaustive set of disjoint regions. Each region is represented by a single estimate, chosen as in Eqn. 1 of section 5.1 to maximise the weighted marginal likelihood of the observations in the region it is representing.

This procedure can be carried out if we know how to construct the correct partition of the possible observations, can calculate the marginal probability of observing any datum, and know how to choose the estimate to represent each region. However, in the general case this is far from easy to do. Farr and Wallace ([2002]) exhibit a polynomial-time algorithm for constructing an SMML estimator for the binomial distribution, and they show that this algorithm can be applied more generally to any estimation problem which is one-dimensional in character. Wallace ([2005], chapter 3) has an example of a one-dimensional problem of this type, the estimation of the mean of a Normal distribution of known variance. More generally still, as Farr and Wallace ([2002]) prove, construction of the SMML estimator is NP-hard. They were, for example, unable to find a polynomial-time algorithm for the trinomial distribution, although they were able to achieve quite good results with an heuristic argument.

Fortunately, there are approximations to the SMML method that are

computationally tractable. There are two basic ideas. The most computationally intractable parts of the SMML procedure are the calculation of the marginal probabilities of observing any data and the construction of the partition of all possible observations. These can both be avoided if we replace reliance on the marginal probability of the data with an approximation based on the prior probability distribution. Secondly, we use a quadratic approximation to the log-likelihood function $\log f(x|h)$ in the neighbourhood of parameter vector h.[14]

Estimators derived using these approximations for the most part retain the desirable properties of the SMML estimator described above, namely, language invariance, independence from cost functions, generality (i.e., applicability to cases where the likelihood function has no useful maximum), consistency and efficiency.[15]The estimators for the problems discussed in section 6 are all derived from these approximations.

6 Criticisms of AIC

We turn now to direct comparisons of AIC and MML. AIC tells us to optimise a penalised maximum-likelihood function. Our criticism of AIC takes three forms. Firstly, there are cases where the use of any kind of maximum likelihood function leads to problems. Secondly, there are cases where the specific form of the penalty function chosen used by AIC to measure the complexity of a model can be shown to be too crude. Thirdly, since AIC is applicable only to problems containing real-valued parameters, it cannot be be applied to the many kinds of problems where the parameters are discrete.[16]

Forster and Sober say they are impressed by the generality of AIC. We are not. Overall, we think AIC is insufficiently general. There is nothing very startling about this. Statistics is replete with techniques that are applicable only within a limited range of cases. If our criticisms were purely negative, they would not be particularly interesting. What is more significant is the existence of an alternative method, based on the Minimum Message Length principle, that can be successfully applied to all of the different cases that we are about to exhibit.

6.1 Problems with Maximum Likelihood

Let's look again at AIC. As we've said, it's a technique for doing model selection. The criterion gives us a measure of the goodness or badness of fit of a model or family (F) to the data.

The first part of AIC says that the likelihood of the family is given by the likelihood of the best fitting member of that family, where that is determined by the method of Maximum Likelihood. Then you correct this by adding a penalty for the complexity of the model.

The first thing we want to point out is: not every problem involves model selection. Many problems are just problems of parameter estimation. For those problems, using AIC is equivalent to using the method of Maximum Likelihood.

But there are many problems of parameter estimation where ML gives the wrong answers. Fundamentally, the reason for this is ML's incaution, its tendency to find patterns in the data that aren't really there. This tendency

is very pronounced when the amount of data per parameter is small. This can happen either when the absolute amount of data itself is small, or when the number of parameters to be estimated is large, e.g., growing with the amount of data. We illustrate with examples of both kinds. Note that for all of the problems discussed in this section, the MML estimators behave well on finite samples and converge on the true parameter values in the limit.

6.1.1 Small sample bias in a Gaussian distribution

This is a well known result, and so on its own not especially impressive. But it's a good warm up exercise.

Consider the simple univariate Gaussian distribution with mean μ and standard deviation σ. The probability density function for this distribution is given by:

$$f(\underline{x}) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x_i-\mu}{\sigma}\right]^2}$$

whose negative logarithm can be written as

$$L = -\sum_{i=1}^{N} \left(-\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}\left[\frac{(x_i-\mu)}{\sigma}\right]^2 \right)$$

Taking partial derivatives of L with respect to μ and σ gives

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

and the sample variance $\hat{\sigma}^2_{\mathrm{ML}}$ is given by:

$$\hat{\sigma}^2_{\mathrm{ML}} = \sum_{i=1}^{N} \frac{(x_i-\bar{x})^2}{N}$$

It is well known that this is a biased estimator of the true variance. For large sample sizes, the bias is quite small; for small sample sizes the bias is

considerable. Replacing the divisor N by $N-1$ gives an unbiased estimator, but this move away from the Maximum Likelihood estimator seems ad hoc. Sure, it works, but *why* does it work? Some classical statisticians have suggested that changing the divisor to $N-1$ is justified because $\sum_{i=1}^{N}(x_i - \bar{x})^2$ is distributed as a constant of proportionality (namely, the unbiased estimate of σ^2) times χ^2_{N-1}, but this special case of a fix hints more at problems than it does at solutions in the general case.

The MML treatment of this problem is instructive. Because MML is a Bayesian technique, the derivation of the estimator relies on the choice of a prior. Critics of Bayesianism view this as a weakness, seeing the choice of prior as arbitrary and difficult to justify. A typical objection in a problem like this one might be that we have no particular reason to expect one value of σ rather than another. Here, however, we can choose a prior that represents an important belief that we do have about the data: that the spread of the data is independent of the magnitude of the data, i.e., that σ is *scale-invariant*. That is, we do not know, and it does not matter, whether the data are measured in nanometres or light-years, molehills or mountains. Hence we choose a prior which is uniform in $\log \sigma$, namely, $1/\sigma$. With this choice of prior, $\hat{\mu}_{\text{MML}} = \hat{\mu}_{\text{ML}} = \hat{\mu}$, and the MML estimate of σ^2,

$$\hat{\sigma}^2_{\text{MML}} = \sum_{i=1}^{N} \frac{(x_i - \hat{\mu})^2}{N-1},$$

which as we noted above is unbiased for any sample size.[17]Hence we see that what looks to be a kludge from the perspective of classical statistics is actually justified by broader principles from the perspective of MML.

6.1.2 The von Mises circular and von Mises-Fisher spherical distributions

The von Mises circular and von Mises-Fisher spherical distributions are angular analogues of the Gaussian distribution. These distributions with mean direction, μ, and concentration parameter, κ, can be thought of as the long-term distribution of the direction of a compass needle in the plane (for the circular distribution), or in *Euclidean 3-space* (for the spherical distribution), subjected to something like magnetic field strength, κ.

$\kappa = 0$ corresponds to a uniform distribution around the circle (through the sphere), and for large κ the distribution approximates a Gaussian distribution with mean μ and variance, $\sigma^2 = 1/\kappa$. The von Mises circular distribution has been used to model protein dihedral angles (see (Wallace and Dowe [2000]) and references therein) and hospital arrival times around a 24-hour clock.

It is generally agreed that (as with the Gaussian distribution) the best way to estimate μ is to average the sample data—this corresponds to the direction of a head-to-tail vector addition. The difficult issue is how to estimate κ.

Wallace and Dowe ([1993]) considered a variety of estimation criteria for the circular distribution, including Maximum Likelihood and MML.[18] They found that for *all* measurements of error (bias, absolute error, squared error and Kullback-Leibler distance), for *all* (true) values of κ, for *all* sample sizes, N, Maximum Likelihood, and therefore AIC, was the worst (or equal worst)-performing method on *all* occasions. The differences are especially

apparent for small sample sizes.[19] By contrast, MML was in most cases the best, and otherwise nearly the best of the estimators.

Dowe et al. ([1996]) similarly showed that ML is the worst-performing estimation criterion, and in general MML the best, for the von Mises-Fisher spherical distribution.

6.1.3 The Neyman-Scott problem

The small sample bias of Maximum Likelihood in the Gaussian distribution led Neyman and Scott ([1948]) to wonder how ML would perform on a problem where the number of data per parameter to be estimated is necessarily small, because the number of parameters to be estimated grows with the data.

The Neyman-Scott problem is one where we have $2N$ measurements arising as 2 measurements each from N things. Given measurements $\{x_{i1}, x_{i2} : i = 1, ..., N\}$, assuming x_{i1} and x_{i2} to come from a population of mean μ_i and standard deviation σ (independent of i), the problem is to estimate σ and the μ_i.

MML is consistent (Dowe and Wallace [1997]) but the uncorrected small sample bias of Maximum Likelihood for the Gaussian distribution prevails here, and we see the Maximum Likelihood estimate of σ^2 inconsistently converging to $\frac{1}{2}\sigma^2$.[20] Similar inconsistencies in Maximum Likelihood can be seen on other problems where there is a finite amount of data per parameter to be estimated, even as the number of data increases: for example, single and multiple factor analysis[21](Wallace and Freeman [1992]; Wallace [1995])

and also for fully-parametrised mixture modelling (Wallace ([2005], Ch. 6.8); Wallace and Dowe ([2000], sec. 4.3)).

6.1.4 Neyman-Scott, predictive accuracy and Minimum Expected Kullback-Leibler Distance

Strict MML and its approximations are not the only statistically invariant Bayesian methods of point estimation. Another is the Minimum Expected Kullback-Leibler Distance (MEKLD) estimator (Kullback and Leibler [1951]; Dowe et al. [1998]; Wallace [2005], pp.205–209). The Kullback-Leibler (KL) distance is a measure of the difference between two probability distributions for the same variable. Let $a(x)$ and $b(x)$ be two probability distributions of a discrete random variable x. Then the KL-distance of $b()$ from $a()$ is defined as

$$KLD(a,b) = \sum_x a(x) \log(a(x)/b(x))$$

The KL-distance is a measure of how surprising observed values of x will appear if we think they are being generated according to $b()$ when in fact they come from $a()$. Because it is invariant under re-parametrisation, the KL-distance is useful as a general purpose cost-function, giving the cost of mistaking $a()$ for $b()$. Of course, we do not in general know what the true distribution $a()$ is, and so we cannot calculate directly how far away from it is some other distribution, $b()$, that we have estimated from the data. But we can minimise the *expected* cost with respect to the posterior distribution of the true value. This MEKLD estimate is also called a *predictive distribution*; it is the maximum expected log-likelihood estimate based, not on the data we have, but on what data we might expect to get from the same source.

According to the predictive accuracy framework advocated by Forster, finding this estimate is the goal of science, and AIC is the means to achieve this goal.

The Neyman-Scott problem raises difficulties for both of these claims. If AIC provides the means to achieve predictive accuracy, then the estimates recommended by AIC should converge to MEKLD estimates. As with all the problems discussed in this section, the estimates recommended by AIC are just those derived using the method of Maximum Likelihood. For the Neyman-Scott problem, ML inconsistently *over*-fits the data, with $\hat{\sigma}^2_{\text{ML}} \rightarrow \frac{1}{2}\sigma^2$, MML consistently converges on σ^2, while MEKLD inconsistently *under*-fits the data, with $\hat{\sigma}^2_{\text{MEKLD}} \rightarrow \frac{3}{2}\sigma^2$ (Wallace [2005], p.207). So even if we accept the goal of predictive accuracy, AIC is not the means to achieve that goal.

The Neyman-Scott problem also casts doubt on the adequacy of predictive accuracy as the single goal of statistics, because it shows an interesting distinction between inference and prediction. We just noted above that MEKLD under-fits the data, that is, it overestimates the value of σ. This is not to say that MEKLD is failing in its task. While MEKLD is overly conservative in terms of inference, in terms of expected predictive error it makes perfectly good sense to let all theories have their say and consequently, to overestimate the noise term. Our position is not that inference is better than prediction, only that they can be different things. If you have some data for the Neyman-Scott problem and you want to minimise the surprise you get when you see new data, you should use the predictive distribution. But if you are interested in inferring the source from which the data actually came, then MML is the method you need.

An example may help clarify the distinction. Say you want to predict tomorrow's maximum temperature. Today it's hot, but there's a strong cold front on its way. However, it's hard to know exactly when the front will arrive. According to your weather model, if the front arrives early tomorrow, temperatures will be cool, with a maximum of 20°C, whereas if the front arrives late, the maximum temperature will be 30°C. According to your model, there's a 40% chance that the front arrives early. What should your prediction be? If you want to minimise your surprise on learning what the temperature was, you should predict a maximum temperature of 26°C, even though you are almost certain that the maximum temperature will not have this value. If the cost of being wrong is proportional to your distance from the right answer, this prediction makes sense. However, if it is more important to get the right answer, you should predict a temperature of 30°C.[22]

6.2 Other problems with AIC

The preceding criticisms of AIC may strike some readers as unfair, since the difficulties encountered are actually difficulties with the Maximum Likelihood method of estimating parameters, whereas AIC is being advanced as a method of doing model selection. However, it is not so easy to get AIC off the hook. AIC maximises a penalised maximum likelihood function. Any problems with maximum likelihood are therefore going to be conceptual problems for AIC. In addition, the problems with ML identified above will become problems for AIC as soon as AIC is asked to discriminate between different models, some or all of which take these forms.

In any case, not all of the problems with AIC can be attributed to difficulties with the method of Maximum Likelihood. In this section we demonstrate several problems in which AIC performs very badly in selecting a model, even where Maximum Likelihood or similar related methods can be used to estimate the parameters of the model. These cases show that the penalty function used in AIC to balance model complexity against goodness-of-fit with the data is too crude to work effectively in many commonly occurring statistical problems.

6.2.1 Univariate polynomial regression

The key model selection problem discussed in (Forster and Sober [1994]) is that of univariate polynomial regression, which they call the 'curve-fitting problem'. In this problem, the task is to choose the degree of a polynomial approximation to an unknown function. It is specifically to this problem that Forster claims Akaike has provided the solution.

Wallace ([1997]) compares the performance of five methods for selecting the order of a polynomial approximation to a function. The methods compared include AIC and MML.[23] The tests were conducted as follows: we have some unknown target function, $t(x)$.[24] We are given some training data comprising N pairs $\{x_n, y_n : n = 1, \ldots, N\}$, where the x-values are chosen randomly from the Uniform distribution on [-1,1], and the y-values are given by $y_n = t(x_n) + \varepsilon_n$, where each of the noise values $\{\varepsilon_n : n = 1, \ldots, N\}$ are selected from a Gaussian distribution of zero mean and unknown variance, v.

The task is to construct some polynomial function $f(d, x)$ of degree d

which may be used to predict the value of $t(x)$ in the interval [-1,1]. Only polynomials of degree up to 20 are considered, and for any degree, only the polynomial which minimises the squared error on the training data (i.e., the maximum likelihood polynomial) is considered. The model selection problem is thereby reduced to the choice of degree, d.

The success of the chosen approximation is measured by its Expected Squared Prediction Error, i.e. the average value of $[f(d,x)-t(x)]^2$, estimated using a Monte Carlo estimate as

$$ESPE[f(d,x)] = \frac{1}{M}\sum_{m=1}^{M}[f(d,x_m)-t(x_m)]^2$$

with the test points $\{x_m : m = 1,\ldots,M\}$ chosen randomly and uniformly in [-1,1]. A test consists of 1000 cases of this kind, and the ESPE is averaged over the 1000 cases.

[Table 2 about here.]

The results clearly show that AIC is not competitive with MML, except under the very favourable conditions of low noise and large sample, where the methods give similar results. The difference between the two methods is especially stark when the amount of data is small. The results of one such experiment are shown in Table 2, adapted from (Wallace [1997]).[25] This experiment shows the average results over 1000 cases for approximation of a trigonometric function with only 10 data points,[26] under conditions of low noise. The first row of the table shows the average ESPE (AV), followed by its standard deviation (SD), and percentile points and maximum of the ESPE distribution over the cases. Finally, the average ESPE is shown for those cases where a method has selected a polynomial of particular degree

(avERR), together with the number of times it chose that degree (CNT).

AIC's pronounced tendency to over-fit the data is clear from the table. While both AIC and MML chose a 6th-degree polynomial more often than any other, AIC's other choices were heavily weighted towards the higher degrees. In 1000 cases, AIC never selected a polynomial of degree zero or one, whereas MML selected one of these on 255 occasions. AIC chose the highest possible 8th-degree polynomial 154 times, as against 10 times for MML.

The tendency to overfit is reflected in the average errors. We can see from the table that the average ESPE for AIC is nearly a hundred times greater than that for MML, and that AIC's worst case is more than two hundred times worse than MML's worst case.

Nor is this case particularly unusual among the many different trials conducted in (Wallace [1997]). In general, the average squared prediction error for AIC is orders of magnitude greater than that for MML. Moreover, the great superiority of MML over AIC in these tests cannot be attributed to the use of helpful priors: the only prior beliefs assumed are that all degrees are equally likely, and that the signal and noise are likely to be of similar size. These assumptions are at best uninformative and at worst misleading in this context.

6.2.2 Autoregressive econometric time series

Econometric autoregression has quite a bit in common with univariate polynomial regression. Univariate polynomial regression consists of regressing

a variable on polynomial powers of another variable and a noise term. Econometric autoregression consists of regressing a variable on lags of itself (which can be thought of as powers of the Backshift operator: $B(x_t) = x_{t-1}, B^2(x_t) = x_{t-2}, \ldots$) and a noise term.

The study in (Fitzgibbon et al. [2004]) compares several criteria, including AIC, a corrected version of AIC (AIC_c), and an MML estimator.[27]The comparison was performed by using Maximum Likelihood to estimate the parameters for each of the models, with the different criteria then being used to select the model order. In all cases, MML was the best at choosing the model order and the one most prone to under-estimating the model order while AIC was the worst at choosing the model order and the one most prone to over-estimating the model order. With T being the training sample size and the Mean Squared Prediction Error (MSPE) for model order p with parameters $(\phi_1, \ldots, \phi_p)$ being given by

$$\text{MSPE}(p) = \frac{1}{T} \sum_{i=T+1}^{2T} (y_i - (\hat{\phi}_1 y_{i-1} + \ldots + \hat{\phi}_p y_{i-p}))^2$$

in all cases, MML gave the least squared prediction error and AIC gave the largest squared prediction error.

6.2.3 Multivariate second-order polynomial model selection

A variation of the polynomial regression problem arises when the order of the model is fixed and known, while the models vary in selecting from among a large set of variables. The task is both to select the correct variables and to determine how the chosen variables influence the model directly and in collaboration with other variables.

Rumantir and Wallace ([2003]) discuss this problem as it arises in the context of predicting the intensity of tropical cyclones. The order of the model is presumed to be a second-order polynomial, while the models vary in selecting from among 36 variables, representing a wide variety of meteorological, geographical and hydrological data. Rumantir and Wallace compare twelve different model selection techniques, including MML and AIC, using both artificially generated and real data. MML was found reliably to converge to the true model for the artificially generated data, and to a reasonably parsimonious model when tested on real data. Indeed, MML was shown to outperform existing cyclone prediction techniques, and to discover new and useful regularities in the data. The models selected by AIC on the artificial data were found to have over-fitted the data. AIC was not sufficiently competitive with the other techniques to warrant its being tested on the real data.

6.2.4 Gap or no gap: A clustering-like problem for AIC

The last statistical problem to be discussed in this section is a new problem which has not (to our knowledge) previously been discussed in the literature. Suppose that we have to select between the following two models: either a quantity is uniformly distributed over $[0, 1]$, or else it is uniformly distributed over the intervals $[0, a] \cup [b, 1]$, where $0 < a < b < 1$. (In this second case, there is a gap—the interval (a, b) —over which the quantity is absent.) Although this problem is different from, and simpler than, typical problems studied in statistical mixture modelling, clustering, numerical taxonomy, and intrinsic classification, we choose it because it shows a weakness of AIC in an instructive way.

Suppose that we have N data points. The expected average gap size is $1/N$ and the expected size of the largest gap is $(\log N)/N$ (see Appendix B.1 for a proof).

Consider the case in which $b = a + (\log N)/N$. As we show in Appendix B.2, in this case AIC inconsistently prefers the gappy model to the model without the gap, even though a gap of this size is just what we would expect if the data actually came from a distribution with no gap in it. By contrast, as we show in Appendix B.3, in this case MML correctly prefers the no-gap model. We also show that MML will not prefer the gappy model until the size of the largest gap in the data is approximately $(2 \log N)/N$.[28]

In defence of AIC, it might be pointed out that AIC's estimate of the gap size tends to zero as the number of data approaches infinity. But any tendency to overestimate the gap size—and especially inconsistently so—augurs particularly badly for AIC, since one of the stated aims of AIC is to maximise the predictive accuracy (i.e., minimise the KL-distance). Yet overestimation of the gap size gives an infinite KL-distance.

6.3 Conclusions from the comparison of MML and AIC

In this section we have compared the performance of MML and AIC on a wide variety of statistical problems. We conclude the section with a summary of the three most significant implications arising from the comparison.

First, the very great superiority of MML over AIC in the inference problems examined should be apparent. MML can be used to do either parameter estimation or model selection. Where the problem is parameter estimation

(the Gaussian distribution, the von Mises circular and von Mises-Fisher spherical distributions, the Neyman-Scott problem), AIC is equivalent to Maximum Likelihood estimation, which either performs badly or is inconsistent on these problems. MML provides consistent estimates of the parameters in each case.

Where the problem is model selection (univariate polynomial regression, econometric time series autoregression, multivariate second-order polynomial model selection, the gap/no gap problem), AIC is either inconsistent or very bad at choosing the model order, while MML performs well.

Second, Forster ([2002], pp.S129–31) has in the past defended AIC against claims of inconsistency. But we are not merely repeating claims that Forster has already rebutted. Forster's defence of AIC concerns the alleged inconsistency of AIC on only one of the problems we have discussed here, that of univariate polynomial regression. We were careful not to repeat the charge; on that problem we claimed only that the performance of MML is orders of magnitude better than that of AIC, not that AIC is inconsistent. In fact, we accept Forster's defence of the consistency of AIC on this problem, since it is similar to that which we give for Strict MML below in section 7.1.[29]

However, Forster's defence of AIC on the problem of univariate polynomial regression is specific to that problem and will not generalise to other problems. For example, it cannot be used to defend AIC against the charge that it inconsistently prefers the gappy model in the gap/no gap problem. Nor to our knowledge, has Forster ever considered the important class of cases, represented in this paper by the Neyman-Scott problem but also including single and multiple factor analysis and fully-parametrised mixture modelling,

in which the amount of data per parameter is bounded above. In these problems, Maximum Likelihood, and hence AIC, is inconsistent, MML consistent.

Third, the Neyman-Scott problem and the other finite data per parameter problems have highly significant implications for Forster's defence of the goal of predictive accuracy as the goal of statistics. These problems show an important and under-appreciated distinction between inference and prediction: minimising the error in one's predictions of future data is not the same as inferring the true model. It's not that Forster is unaware of this distinction (see, for example Forster [2001], p.95). But since he believes that Bayesian inference is impossible, he has concluded—too quickly, we think—that the field has been left clear for predictive accuracy to claim as its own. What we take ourselves to have shown here is that Bayesian inference is a viable goal for statistics, and that predictive accuracy cannot be a substitute for it. We discuss this point in more detail below in section 7.2.

Two final points: while either prediction or inference can be legitimate goals for statistics, it is important to suit one's means to one's goals: maximising predictive accuracy in Neyman-Scott leads to inconsistent estimates (and, therefore, inferences). If you want to maximise predictive accuracy, you should minimise the expected Kullback-Leibler distance (MEKLD); if you want the best inference, you should use MML. Finally, it is important to note that on the Neyman-Scott problem, AIC does not converge on the MEKLD estimate, and hence does not deliver predictive accuracy: while MEKLD overestimates the noise in the data, and MML correctly estimates the noise, ML (and hence AIC) underestimates it.

7 Meeting Forster's objections to Bayesianism

We have already seen above how the method of minimising message length is immune to the problem of language variance. In this section, we show how MML enables Bayesians to meet two other objections frequently raised by Forster against Bayesian statistics.

7.1 The sub-family problem

The sub-family problem is a generalised version of the curve-fitting problem. Consider any model selection problem in which one family of models, A, is a subset of another family, B. Then for any consistent assignment of priors and for any possible data, $p(B|\,\text{data}) \geq p(A|\,\text{data})$. How, ask Forster and Sober (Forster and Sober [1994]; Forster [2000], p.214), can Bayesians explain the fact we sometimes prefer model family A to model family B?

On the face of it, the SMML estimator described above is vulnerable to this objection. The reason is that in its purest form, the first part of a message constructed according to the strict MML procedure makes no reference to families of models. The SMML estimator is not interested in families of models per se. It partitions the data into subsets each of which is represented by an estimate, which is to say, by a fully specified model. So consider a case of the curve-fitting problem where the true curve is a second degree polynomial (with Gaussian noise added). If asked to select the best curve to account for some data from among all polynomials of, say, degree 8 or less, the particular estimate chosen according to the SMML method is likely to fall very *near* those axes of the parameter space where

the higher order co-efficients have values of zero, but it is unlikely to fall precisely on those axes. So the SMML estimator will typically select a polynomial of degree 8, without noticing that the co-efficients of all the terms in x^3 and higher are very close to zero. Thus, while the polynomial chosen as the SMML estimate will, across the range of the data, be virtually indistinguishable from a quadratic, it will not actually be a quadratic.

How then can the Bayesian solve the sub-family problem? There is a clue to the answer in the description of the problem. Why in the case described above does the SMML estimator choose a high-degree polynomial? Because it is not interested in families of models, only in fully specified models. If we are interested in the question of which is the simplest family of models containing the true model, then we can modify the SMML estimator to direct its attention to that question. As it happens, this can be achieved by a simple modification: we modify the first part of the message so that a model family is asserted. The first part of the message now consists of a statement of the model family, followed by a statement fully specifying the parameters of the chosen family. This modified form of the estimator results in a negligible increase in expected message length and is essentially equivalent[30]to that used in the comparison reported on in section 6.2.1, where the superior performance of MML shows that the estimator successfully solves the curve-fitting problem.

The treatment of this specific problem illustrates the general form of the answer to the sub-family problem within the Bayesian/MML framework. Since models from family B are more complex than those from family A, asserting a model from family B will lengthen the first part of the message. If this lengthening is not compensated for by a greater shortening in the

second part of the message—as will be the case if the extra complexity in model family B brings no extra explanatory power—then MML will prefer the simpler model from family A as its explanation of the data. And this is true notwithstanding the fact that model family B has overall the higher posterior probability. In general, the MML estimate is not equal to the mode of the posterior.

7.2 The problem of approximation, or, which framework for statistics?

At bottom, the competition between MML and AIC is not just a disagreement about which of two statistical methods is superior. There is, as Forster has noted, a fundamental disagreement about the right framework for doing statistics, a disagreement about what statisticians should be trying to do.

Forster defends the *predictive accuracy* framework. In this framework, the basic object of interest is the *model family*, and the basic aim is *prediction*. In the Bayesian/MML framework, the basic object of interest is the *fully specified model*, and the basic aim is *inference*.

This is not to say that in the Bayesian/MML view, prediction is not a legitimate goal of statistics. It is to say that we are sometimes legitimately interested in which fully specified model is most probably true, and not nearly so much in inferring from which class of models that theory comes, or in predicting what future data sampled from a posterior distribution over that class of models would look like.[31]

Why does Forster think we should be interested in model families rather

than in fully specified models? Forster gives two related arguments for preferring the predictive accuracy framework. Firstly, he objects to the Bayesian framework because there is no sensible way to define the probability that a fully specified model (with at least one real-valued parameter) is exactly true. As a result, he thinks that the predictive accuracy framework is truer to the way real scientists behave.

Here is Forster's first argument:

> 'We work with families of curves because they deliver the most reliable estimates of the predictive accuracy of a few curves; namely their best fitting cases. There is no reason to suspect that such an enterprise can be construed as maximising the probability that these best fitting cases are true. Why should we be interested in the *probability* of these curves' being true, when it is intuitively clear that no curve fitting procedure will ever deliver curves that are *exactly* true? If we have to live with false hypotheses, then it may be wise to lower our sights and aim at hypotheses that have the highest possible *predictive accuracy*.' (Forster and Sober [1994], p.26, emphasis in the original)

As we've already noted, we agree with Forster that no method of statistical induction delivers inferences that are exactly true of fully-specified models with real-valued parameters.[32] However, what if there were a method capable of delivering inferences which are *approximately* true?

Forster ([2000], p.214) considers a crude version of this suggestion, where a theory is considered to be approximately true if the true values of parameters differ only infinitesimally from the asserted values. Forster rightly dismisses this attempt;[33] we agree with him that the idea of infinitesimal closeness is

not the right way to define what it means for a theory to be approximately true.

However, Forster concludes from this that no better definition of what it means for a theory to be approximately true is available to the Bayesian (or to anyone). His pessimistic response to this ('If we have to live with false hypotheses, then it may be wise to lower our sights...') is to give up entirely on the idea of inferring the best fully-specified model, and move to a framework within which the focus of interest is on model families instead of fully specified models.

We think this pessimism is hasty and unfounded. The Minimum Message Length Principle provides a rigorous, consistent and useful way of defining the approximate truth of theories. This is possible because there is a trade-off involved in choosing the precision of an estimate. On the one hand, if the estimates are stated too coarsely, then although the first part of the message, asserting some particular estimate, will be short, the second part of the message, encoding the data values, will have to be longer, because more information is required to locate the values within the large region represented by the coarsely stated estimate. On the other hand, if the estimates are stated too precisely, the length of the first part of the message will be unnecessarily lengthened by redundant information.

The result of the trade-off is that MML methods deliver point estimates stated to a precision *consistent with the expected error in their estimation.* It is in this sense that the assertion of a theory chosen by an MML estimator is approximately true: the data give us no reason for being any more (or less) precise in our estimates.

The availability of this Bayesian method of point estimation undermines Forster's second argument:

> 'The goals of probable truth and predictive accuracy are clearly different, and it seems that predictive accuracy is the one that scientists care about most. Wherever parameter values are replaced by point estimates, there is zero chance of that specific value being the true one, yet scientists are not perturbed by this. Economists don't care if their predictions of tomorrow's stock prices are *exactly* right; being close would still produce huge profits. Physicists don't care whether their estimate of the speed of light is *exactly* right, so long as it has a high degree of accuracy. Biologists are not concerned if they fail to predict the exact corn yield of a new strain, so long as they are approximately right. If the probability of truth were something that they cared about, then point estimation would be a puzzling practice. But if predictive accuracy is what scientists value, then their methodology makes sense.' (Forster [2001], p.95)

Since the point estimate asserted by an MML estimator is stated only to a precision warranted by the data, we can see that this argument relies on a false contrast. Scientists can care about probability of truth and still use theories that they know to be only approximately true. The MML framework thus preserves the strong intuition that scientists care about *getting to the truth*, that is, inferring the model from which the data actually came.

8 Conclusion

The mathematical formalisation of the notion of theoretical simplicity is an exciting development for statistics, with profound implications for both statistical theory and practice, and for philosophy of science.

In this paper we have argued for the theoretical and empirical superiority of the Bayesian method of Minimum Message Length over the non-Bayesian Akaike Information Criterion, defended by Malcolm Forster and Elliott Sober. The theoretical arguments show that MML is a consistent and invariant Bayesian procedure. It is thus immune to the problem of language variance raised by Forster and Sober as a general objection to Bayesianism. It is also a general method, one which has been applied successfully to problems of a wide range of statistical forms, and originating in many different fields. It can be used for model selection and for parameter estimation, regardless of whether the parameters are continuous or discrete. AIC, by contrast, can only be used for model selection problems with continuous parameters, as far as we know.

Our empirical arguments show that even in those cases where AIC can be applied, in every case where AIC and MML have been directly compared, MML outperforms AIC. For example, in the key case of univariate polynomial regression (the "curve fitting problem"), AIC is not competitive with MML except under favourable conditions of low noise and large sample. For autoregressive time-series, and for the von-Mises circular and von-Mises-Fisher spherical distributions, the performance of AIC is even worse, being unable to compete with MML at all.

Of special significance is the Neyman-Scott problem, an example of an important class of cases where the amount of data per parameter to be estimated is bounded above. These cases show the inadequacy of the predictive accuracy framework advocated by Forster and Sober: if we aim to maximise the expected log-likelihood of re-sampled data, we end up with inconsistent estimates. By contrast, MML gives consistent estimates in the Neyman-Scott problem and similar cases.

Having laid out the many advantages of the Bayesian MML approach over the non-Bayesian AIC approach, we leave the reader with Dowe's question ((Dowe et al. [1998], p.93); (Wallace and Dowe [1999a], p.282); (Wallace and Dowe [2000], sec. 5); (Comley and Dowe [2005], sec. 11.3.1)) as to whether only MML and closely related Bayesian methods can, in general, infer fully specified models with both statistical consistency and invariance. Even if non-Bayesian methods can achieve this, we doubt that they will be as efficient as MML.

Appendices

A Details of the derivation of the Strict MML estimator

We begin by observing that the set of possible observations, X, is countable. This must be so because any apparatus that we can construct with which to take measurements can do so only with finite accuracy.

We also observe that the number of actual observations must be finite. Since a countable quantity raised to the power of a finite quantity is countable, it follows that the set of distinct hypotheses that the data could justify us in

making, $H^* = \{m(x) : x \in X\}$, is also countable, i.e. $H^* = \{h_j : j \in N\}$.

For example, if want to estimate the probability that a coin lands heads from the number of heads in a sequence of 100 tosses, it would not make sense to consider more than 101 different estimates of the bias of the coin, because there are only 101 different possible observations we could make.

Since the number of assertable estimates is countable, we can define $c_j = \{i : m(x_i) = h_j\}$ for each $h_j \in H^*$, and $C = \{c_j : h_j \in H^*\}$. That is, each c_j is a group of possible data points, with all the points in the group being mapped by the estimator m to a single estimate, h_j. C is the set of these data groups, which together form a partition of the data space. Given some set of assertable estimates H^* as defined above, we assign non-zero prior probabilities to the members of H^*—constrained by the prior density function—and then, for each $x \in X$, choose $h \in H^*$ to maximise $p(h|x)$. This defines $m^* : X \to H^*$, and so shows the existence in theory of a solution to our point estimation problem, $\{H, X, f, p\}$, provided that we have selected the right H^*. We now consider how best to choose H^*.

For all j, let $q_j = \sum_{i \in c_j} m(x_i)$ be the prior probability assigned to h_j, which in sec. 5.1 we called the coding prior. We must have $\sum_j q_j = 1$ and $q_j > 0$ for all j. Moreover, if $i \in c_j$ then $p(x_i|h_j)q_j \geq p(x_i|h_{j'})q_{j'}$ for all i and $j \neq j'$. That is, given a set of data points all of which are mapped to a single estimate h_j, no other choice of estimate in H^* gives a greater joint probability of estimate and data.

To get a good approximation, perhaps the best approach would be to compare $r(x_i) = \int_H p(x_i|h)f(h)dh$, the marginal probability of making some observation, with $r^*(x_i) = \sum_j p(x_i|h_j)q_j = \sum_j b(h_j, x_i)$. Each $b(h_j, x_i)$ is the

joint probability of the estimate h_j and a datum x_i. $r^*(x_i)$ is therefore the marginal probability of datum x_i given the coding prior. However, we shall instead maximise an average of $b(x, h)$ over all x.

Suppose that we conduct N trials, with observations $Y = \{y_1, \ldots, y_n, \ldots, y_N\}$, parameter values $G = \{g_1, \ldots, g_n, \ldots, g_N\}$, and estimated values $K = \{k_1, \ldots, k_n, \ldots, k_N\}$. The joint probability of Y and G, $J(G, Y) = \prod_{n=1}^{N} b(g_n, y_n)$.

Our aim will be to choose the discrete model of the problem and the estimate sequence, K, so as to give the highest possible joint probability, i.e. to choose the model and K to maximise $J(K, Y) = \prod_{n=1}^{N} b(k_n, y_n)$.[34]

Let D_i be the number of times x_i occurs in Y, for all i. Then $J(K, Y) = \prod_i \{b(m(x_i), x_i)\}^{D_i}$, where m is some yet-to-be determined function.

Now, if N is large enough, then $D_i \approx Nr(x_i)$. So, we aim to choose the model H^* and m to maximise:

$$\begin{aligned} B &= \frac{1}{N} \sum_i Nr(x_i) \ \log b(m(x_i), x_i) \\ &= \sum_j (\sum_{i \in c_j} r(x_i) \ \log q_j) + \sum_j (\sum_{i \in c_j} r(x_i) \ \log p(x_i|h_j)) \end{aligned}$$

The first term gives the negative expected length of the assertion of the model, and the second term gives the negative expected length of the data, encoded under the assumption that the asserted model is the true model.

B MML, AIC and the Gap vs. No Gap Problem

We have data $\{0 \leq x_i \leq 1 : i = 1...N\}$. Each of the x_i are stated to some accuracy $\varepsilon > 0$.[35]We are to decide whether the data come from a uniform distribution over the interval [0,1] with no gap, or from a uniform distribution over $[0, a] \cup [b, 1], 0 < a < b < 1$, where there is an interval (a, b) over which the quantity is absent.

B.1 Expected size of the largest gap

We begin by showing that the expected size of the largest gap between two of N data points from the non-gappy distribution is approximately $(\log N)/N$.

To get things started, we assume that the N data points come approximately from a negative exponential distribution, with rate $r = N + O(\sqrt{N}) \approx N$; whereupon $r/N = 1 + O(1/\sqrt{N}) = N/r$ and $\lim_{N\to\infty} r/N = 1 = \lim_{N\to\infty} N/r$.

The points are therefore uniformly distributed in $[0, 1]$ and the average gap size is $1/N$. Then,

$$\begin{aligned}\Pr(\text{largest gap} \leq y) &= \Pr(\text{all gaps} \leq y) = [\Pr(\text{a random gap} \leq y)]^N \\ &= \left[\int_0^y N e^{-tN}\, dt\right]^N \\ &= \left[\int_0^{yN} e^{-u}\, du\right]^N \\ &= (1 - e^{-yN})^N\end{aligned}$$

We now introduce a few results to be used in the rest of this appendix. As is

well known,

$$\lim_{M\to\infty}\left(1+\frac{k}{M}\right)^M = e^k, \text{ and} \tag{B.1}$$

$$\lim_{M\to\infty}\frac{(M+2)^2}{M^2} = 1 \tag{B.2}$$

We also make use of a result suggested to us by Gérald R. Petit, generalising B.1 to the case of $f(M) = o(\sqrt{M})$, i.e., where $\lim_{M\to\infty}\frac{f(M)}{\sqrt{M}} = 0$:

$$\lim_{M\to\infty}\frac{\left(1+\frac{f(M)}{M}\right)^M}{e^{f(M)}} = 1 \tag{B.3}$$

Letting $\gamma(y,N) = e^{Ny}/N$ and so $y = (\log\gamma N)/N,\ 0<\gamma<\infty$,

in the special case that $y = 0$ then $\gamma = 1/N$ and

$\Pr(\text{largest gap} \le y) = \Pr(\text{largest gap} \le 0) = 0 = \lim_{N\to\infty} e^{-N} = \lim_{N\to\infty} e^{-1/\gamma}$.

Otherwise, for $y > 0$, noting that $-1/\gamma = -N/e^{Ny} = o(\sqrt{N})$ for each y in turn, by equation (B.3) we have that for each $y > 0$,

$$\begin{aligned}\Pr(\text{largest gap} \le y) &= (1-e^{-(N\log\gamma N)/N})^N \\ &= (1-e^{-(\log\gamma+\log N)})^N \\ &= (1-1/\gamma N)^N \\ &\to e^{-1/\gamma} \text{ in the asymptotic limit of large } N. \end{aligned} \tag{B.4}$$

Eqn. B.4 is a cumulative distribution for the size of the largest gap. Differentiating Eqn. B.4 can similarly be shown to give the distribution for the largest gap itself; we can then find the expectation of this distribution.

$$\begin{aligned} f(\gamma) = \frac{\partial}{\partial\gamma}\left[\Pr\left(\text{largest gap} \le y = \frac{\log N+\log\gamma}{N}\right)\right] &\to \frac{\partial}{\partial\gamma}\, e^{-1/\gamma} \\ &= \frac{1}{\gamma^2}\, e^{-1/\gamma} \end{aligned} \tag{B.5}$$

Since $0 \leq y \leq 1$ and $y(\gamma)$ is a continuous bounded function of γ for $0 < \gamma \leq \frac{e^N}{N}$, the expected size of the largest gap, $E[f(\gamma)]$, is

$$\begin{aligned}
E[f(\gamma)] &= \int_0^\infty f(\gamma)\, y(\gamma)\, d\gamma \\
&= \int_0^\infty f(\gamma)\ \frac{\log\gamma + \log N}{N}\, d\gamma \\
&= \frac{\log N}{N} \int_0^\infty f(\gamma)\, d\gamma + \frac{1}{N} \int_0^\infty f(\gamma) \log\gamma\, d\gamma \\
&= \frac{\log N}{N} + \frac{1}{N} \int_0^\infty O(1) \frac{e^{-1/\gamma}}{\gamma^2} \times \log\gamma\ d\gamma \\
&= \frac{\log N}{N} + O(\frac{1}{N}) \\
&\approx \frac{\log N}{N}
\end{aligned}$$

$$\text{I.e., } \lim_{N\to\infty} \frac{N}{\log N} E[f(\gamma)] = 1 \tag{B.6}$$

B.2 Performance of AIC on the Gap vs. No Gap Problem

Recall that AIC minimises $2\log(L(F)) + 2k$, where k is the number of estimated parameters, and $L(F)$ the member of the family of curves that best fits the data. (By contrast, maximum likelihood minimises $\log(L(F))$. AIC differs from maximum likelihood by the inclusion of the penalty function $+2k$.)

We consider a case in which $b = a + (\log N)/N$. The likelihood function is $\left[1 - \frac{\log N}{N}\right]^N$ for the gappy model (for which $k = 2$), and 1^N for the non-gappy model (for which $k = 0$).

So, on the one hand, $\text{AIC}_{NG} = -2\log(1^N) + 2 \times 0 = 0$.

On the other hand,

$$
\begin{aligned}
\mathrm{AIC}_G &= 2N\log(1-(b-a)) + 2\times 2 \\
&\approx 2N\log(1-\frac{\log N}{N}) + 4 \\
&= 2N(-\frac{\log N}{N} + \frac{1}{2}\left(\frac{\log N}{N}\right)^2 + \cdots) + 4 \\
&= -2\log N + \frac{(\log N)^2}{N} + \cdots + 4 \\
&< 0, \text{ for moderately sized N.}
\end{aligned}
$$

So AIC selects the model with the gap.[36]

B.3 Performance of MML in the Gap vs. No Gap Problem

To encode the models, we employ the following coding scheme: we use 1 bit ($= \log_e 2$ nits[37]) to indicate which of the two models is being used, then we encode the parameters of the model (none for the no-gap model, the two parameters a and b for the gappy model), then we encode the data.

For sufficiently small ε, the message length for the no-gap model is given straightforwardly by:

$$
L_{\mathrm{NG}} = \overbrace{\log 2}^{\text{model}} + \overbrace{0}^{\text{parameters}} \overbrace{-N\log\varepsilon}^{\text{data}}. \tag{B.7}
$$

For the gappy model (see Figure 1), we must decide to what precision to state the parameters of the model a and b. Initially, let us say that we state a to precision δ_1 and b to δ_2.

Then the message length for the gappy model is given by:

$$
L_{\mathrm{G}} = \overbrace{\log 2}^{\text{model}} \overbrace{-\log\delta_1 - \log\delta_2 - \log 2}^{\text{parameters}[38]} \overbrace{-N\log\varepsilon + N\log(1-[(b-\delta_2)-(a+\delta_1)])}^{\text{data}}
$$

Assume (from symmetry) that $\delta_1 = \delta_2 = \frac{\delta}{2}$. So[39]

$$L_{\mathrm{G}} = -2\log\frac{\delta}{2} - N\log\varepsilon + N\log(1 - [(b-a) - \delta]). \tag{B.8}$$

[Figure 1 about here.]

We wish to choose δ to minimise L_{G}. Solving

$$0 = \frac{\partial L_{\mathrm{G}}}{\partial \delta} = -\frac{2}{\delta} + \frac{N}{1 - [(b-a) - \delta]},$$

we find that L_{G} is minimised when

$$\delta = \frac{2[1 - (b-a)]}{N-2}, \text{for } N \geq 3. \tag{B.9}$$

Given N, we wish to know for what size gap MML will prefer the gappy model to the no-gap model. This is equivalent to asking at what point the message lengths L_{G} and L_{NG} are the same. Subtracting B.7 from B.8 we get the expression

$$\begin{aligned} L_{\mathrm{G}} - L_{\mathrm{NG}} &= -2\log\frac{\delta}{2} - N\log\varepsilon + N\log(1 - [(b-a) - \delta]) \\ &\quad - \log 2 + N\log\varepsilon \\ &= -2\log\frac{\delta}{2} + N\log(1 - [(b-a) - \delta]) - \log 2 \end{aligned}$$

Substituting $\delta = \frac{2[1-(b-a)]}{N-2}$ from equation B.9 above, and letting the expression equal zero, we get:

$$
\begin{aligned}
0 = L_{\text{G}} - L_{\text{NG}} &= -\log 2 - 2\log\frac{1-(b-a)}{N-2} \\
&\quad + N\log\left[(1-(b-a))\left(1+\frac{2}{N-2}\right)\right] \\
&= -2\log(1-(b-a)) + 2\log(N-2) + N\log(1-(b-a)) \\
&\quad + N\log N - N\log(N-2) - \log 2 \\
&= (N-2)\log(1-(b-a)) - (N-2)\log(N-2) \\
&\quad + N\log N - \log 2 \\
&= (N-2)\log\frac{(1-(b-a))}{N-2} + N\log N - \log 2 \\
&= 2\log N + (N-2)\log\frac{N(1-(b-a))}{N-2} - \log 2
\end{aligned}
$$

Exponentiating, and given N, we seek $(b-a)$ satisfying

$$
\begin{aligned}
1 = e^0 &= \tfrac{1}{2}N^2 \cdot \left[\frac{N(1-(b-a))}{N-2}\right]^{N-2} \\
1 &= \tfrac{1}{2}N^2 \cdot (1-(b-a))^{N-2} \cdot \left[\frac{N}{N-2}\right]^{N-2} \\
1 &= \tfrac{1}{2}N^2 \cdot (1-(b-a))^{N-2} \cdot \left[1+\frac{2}{N-2}\right]^{N-2}
\end{aligned}
$$

For $N \geq 3$, we can let $M = N-2$, and say that

$$
1 = \frac{M^2(M+2)^2}{2M^2} \cdot (1-(b-a))^M \cdot \left[1+\frac{2}{M}\right]^M \tag{B.10}
$$

Using equations B.1 and B.2 we can rewrite equation B.10 as follows:

$$
\lim_{M\to\infty} \tfrac{1}{2}M^2e^2(1-(b-a))^M = 1, \tag{B.11}
$$

and using equation B.3 it now follows that

$$
\lim_{M\to\infty} \frac{\left[(1+\frac{2}{M})(1+\frac{2\log M}{M})\right]^M}{e^2M^2} = 1.
$$

Therefore, equation B.11 can be rewritten as

$$
\lim_{M\to\infty} \tfrac{1}{2}\left[(1+\frac{2}{M})(1+\frac{2\log M}{M})(1-\frac{M(b-a)}{M})\right]^M = 1. \tag{B.12}
$$

Expanding,

$$\lim_{M\to\infty} \tfrac{1}{2}\Big[1+\frac{1}{M}(-M(b-a)+2+2\log M) \\ +\frac{1}{M^2}(-2M(b-a)-2M(b-a)\log M+4\log M) \\ +\frac{1}{M^3}(4M(b-a)\log M)\Big]^M = 1.$$

Under the very reasonable assumption that $M(b-a) = o(\sqrt{M})$, the terms in $\frac{1}{M^2}$ and $\frac{1}{M^3}$ are small enough to be ignored, and we can re-apply Petit's result in equation B.3, letting $f(M) = -M(b-a)+2+2\log M$. We see that

$$\lim_{M\to\infty} M^2 e^2 e^{-M(b-a)} = 2$$
$$\lim_{M\to\infty} 2(\log M+1) - M(b-a) = \log 2$$
$$\lim_{M\to\infty} \frac{M(b-a)}{2(\log M+1)-\log 2} = 1 \tag{B.13}$$

Recalling that $M = N-2$ for $N \geq 3$, we can say that

$$\lim_{N\to\infty} \frac{N(b-a)}{2\log N} = 1 \tag{B.14}$$

What is the meaning of this result? We have shown that MML will prefer the gappy model if there is a gap, $b-a$, in the data of approximately $(2\log N)/N$. This is commendable. For if the true distribution were non-gappy, then the expected size of the largest gap would still be $(\log N)/N$. Therefore, to prefer the gappy model when the largest gap in the data is $(\log N)/N$ would be to leap to unwarranted conclusions. In fact, as we showed, this is just what AIC does. Waiting until the the size of the largest gap in the data is $(2\log N)/N$ before switching over to the gappy model is therefore very sensible behaviour.

Notes

[1]We assume here that no two points are vertically aligned. If the data contains two points that are vertically aligned, then polynomials of degree $N - 1$ can get arbitrarily close to the data.

[2]The case discussed in 6.1.4, the Neyman-Scott problem, shows the problems for AIC in a particularly stark way. But even in friendlier cases, where the amount of data per parameter is not strictly bounded above (for example, the cases of univariate polynomial regression (section 6.2.1), or econometric time series regression (section 6.2.2)), AIC is generally the worst or nearly the worst estimator of those studied for predictive accuracy.

[3]Minimum Message Length has important similarities and differences to Rissanen's Minimum Description Length (MDL) (Rissanen [1978], [1989], [1999]). They are similar in that they share the aim of making an inference about the source of the data (and not just making predictions about future data); they also share the insight that inference is fundamentally connected to achieving a brief encoding of the data. However, they differ in two important respects: firstly, Rissanen's work is non-Bayesian, and MDL tries to avoid any use of priors. Secondly, MDL (like AIC) typically aims to infer only the model class from which the true (fully specified) model comes, while MML aims to infer a single, fully specified model. For detailed discussions of the differences between MML and MDL, see (Wallace and Dowe [1999a]) and the other articles in that special issue of The Computer Journal, as well as (Wallace [2005], Ch. 10.2) and (Comley and Dowe [2005], sec. 11.4.3). It is also worth mentioning that the 1978 version of MDL, although not the later versions, is equivalent to Schwartz's Bayes Information Criterion (BIC) (Schwartz [1978]).

[4]Forster and Sober ([1994]) also discuss the consequences of their arguments

for Bayesianism. However, since their main question is whether Bayesianism can somehow replicate Akaike's work, the significance of their discussion is rather diminished when the shortcomings of AIC, demonstrated below, are fully appreciated!

[5]Unlike Maximum A Posteriori (MAP), which maximises a posterior *density* and is generally not statistically invariant. See footnote 14 for more details.

[6]The equality is approximate rather than strict because code lengths must be integers, while probabilities need not be. But the difference is negligible; in fact $-\log p_i \approx l_i < -\log p_i + 1$. Shannon ([1948]) proved that it is always possible to construct such a codebook, and demonstrated a method for doing so. See (Wallace [2005], Ch. 2) for a clear exposition.

[7]Unless the parameter estimates render x_i impossible, such as if x_i were 50 Heads followed by 50 Tails but the group parameter estimate was that the probability of Heads was 1.

[8]This problem was originally treated in (Wallace and Boulton [1975]). A more detailed exposition than we offer here can be found in (Wallace [2005], Ch. 3.2.3).

[9]Wallace ([2005], p.160) notes, 'If the set of possible models were discrete...the correspondence would be exact. [But see (Wallace [2005], Ch. 3.2.1) and (Comley and Dowe [2005], sec. 11.3.1) for caveats.] However, when h is a continuum...$q(h)$ is not the prior probability that "h is true": indeed no non-zero probability could be attached to such a statement. Nonetheless, the difference can play the role of a negative log posterior probability, and its expectation is a good measure of the "believability" of the estimates.' An example of the need for some caution in the interpretation of Eqn. (2) can be seen in (Wallace and Boulton [1975], Table 3, p.29), where the ratio exceeds unity for some possible data, typically those

having relatively high likelihood given their associated parameter estimate.

[10]In case this use of improper priors should alarm friends or critics of Bayesianism, Wallace includes some mathematical techniques for renormalising improper priors. For example, a uniform prior distribution for a location parameter is improper, but we can renormalise this prior by observing that the prior belief that the uniform distribution is meant to capture is that the location is equally likely to be anywhere within a large but finite range. As long as there is only negligible probability that the data will reflect a location from outside this range, we need not even specify what the range is. He also observes that improper priors, when combined with real data, usually lead to proper posterior densities.

[11]We say this because Akaike's argument in support of penalising the number of parameters applies to continuous-valued parameters rather than discrete-valued parameters. An example of an inference problem with discrete-valued parameters is the inference of decision trees. If AIC can be defined for decision trees, then it can presumably only take the form of penalising the log-likelihood with twice the number of nodes. In the case of binary decision trees, this is equivalent to a penalty of the number of nodes, which is the penalty function adopted in the binary tree study by Murphy and Pazzani ([1994]). Even if we are to permit this interpretation of AIC, MML has empirically been shown to work decidedly better on this problem in (Needham and Dowe [2001]).

[12]Wallace and Freeman ([1987], p.241) give a laconic hint to this effect, mentioning the second result but leaving the reader to draw out for themselves the conclusions which follow from it. A much more detailed argument can be found in (Wallace [2005], ch.3.4.5, pp.190-91), and an independent derivation of an almost identical result can be found in (Barron and Cover [1991]).

[13]Strictly speaking, as Wallace ([2005], p.191) notes, because optimal encoding ideally requires a non-integral number of binary digits (as in fn. 6), the above arguments imply only that the assertion chosen by the SMML estimator lacks at most one binary digit of information and contains at most one binary digit of noise.

[14]The estimator derived using these approximations is often referred to as MML87; for continuous-valued attributes, it can be thought of as maximising the posterior density divided by the square root of the the Fisher information, which shows clearly its difference from MAP estimation. More detailed expositions can be found in (Wallace and Freeman [1987], sec. 5), (Wallace and Dowe [2000], sec. 2), and (Wallace [2005], Ch. 5). For non-quadratic approximations to SMML, based on different ideas, see (Wallace [2005], Ch. 4).

[15]For the precise conditions under which the approximations of (Wallace and Freeman [1987]) can be applied, see (Wallace [2005], Ch. 5.1.1).

[16]In addition to the inference of decision trees mentioned in fn. 11, other such problems include the hierarchical mixture modelling of (Boulton and Wallace [1973]), as well as the generalised Bayesian networks of (Comley and Dowe [2003]) and (Comley and Dowe [2005]), which treat model families with mixtures of continuous and discrete variables. These are but a few of many inference problems well-studied in the machine learning and artificial intelligence literature for which AIC appears to be undefined.

[17]Despite the fact that the choice of prior is independently motivated, some sceptics about Bayesianism might still be troubled that the MML estimator has been 'cooked up' by careful choice of prior. So it is worth noting that in fact the choice of prior on its own makes the situation worse, changing the divisor in the

estimator from N to $N+1$. What saves MML, and brings the divisor back from $N+1$ to $N-1$, is the use it makes of the Fisher information, the expectation of the determinant of the matrix of second partial derivatives of the negative log-likelihood, which can be thought of as relating to the uncertainty of our estimates. See (Wallace and Dowe [2000], sec. 2.1) for more details.

[18]The other criteria considered were N. I. Fisher's modification to Maximum Likelihood (Fisher [1993]), and G. Schou's marginalised maximum likelihood (Schou [1978]).

[19]Indeed, a theorem of Dowe's shows that for the von Mises circular distribution, when $N = 2$, regardless of the true value of κ (even with $\kappa = 0$), the expected value of the ML estimate of κ is infinity.

[20]See also (Wallace [2005], Ch. 4.5) for further discussion.

[21]In section 5 ("Control of Improper Solutions by a Bayesian Modeling") of his paper on AIC and factor analysis (Akaike [1987], p.325), the inconsistency of AIC on this problem not only led Akaike to adopt a Bayesian prior, but moreover, a "prior" whose logarithm is proportional to the sample size.

[22]See (Wallace and Dowe [1999a], sec. 8) for other examples and further discussion.

[23]To be more precise, a version of AIC called *final prediction error* (FPE) is used (Akaike [1970]). FPE is derived from AIC by estimating the variance of the noise in the data independently for each model family (see Cherkassky and Ma [2003], p.1694). The other methods compared are Vapnik-Chervonenkis dimension (Vapnik [1995]), Schwartz's Bayes Information Criterion (Schwartz [1978]), and Craven and Wahba's Generalised Cross-Validation technique (Craven and Wahba [1979]).

[24]The functions tested are a trigonometric function, a logarithmic function, a function with a discontinuous derivative, and a discontinuous function. It is important to note that none of the 'true curves' here are actually polynomial functions. As Cherkassky and Ma ([2003]) point out, this violates an assumption of AIC that the true model is among the possible models. However, we do not therefore consider the comparison between MML and AIC to be unfair to AIC, for three reasons: (1) the theory of minimum message length is in part motivated by the same assumption, so violating it seems equally unfair to both methods; (2) as Cherkassky and Ma ([2003]) also note, AIC is often used in contexts where this assumption does not hold; and (3) the violation of the assumption is realistic, in the sense that many real-world settings require us to pick the best polynomial approximation to a function where we cannot be sure that the true model is a polynomial function. The problem of predicting the intensity of tropical cyclones by modelling them as second-order polynomials, tackled in (Rumantir and Wallace [2003]) and discussed below in section 6.2.3, is an example.

[25]The results for the other model selection criteria mentioned in fn. 23 have been removed from this table. AIC was clearly the worst of all the methods studied.

[26]This is why the maximum degree considered is only 8, rather than 20.

[27]The other criteria tested were Schwartz's Bayesian Information Criterion (Schwartz [1978]) and a criterion due to Hannan and Quinn (Hannan and Quinn [1979]). MML also outperformed both these criteria in the test.

[28]Related, but more complex, illustrations of this point may be found in (Wallace and Boulton [1975]) and (Wallace and Dowe [1999b]).

[29]Like AIC, Strict MML overshoots the simplest model family containing the

true model in the sense that it typically selects a model from a more complex family. (Although SMML doesn't overshoot in the same way—for SMML, unlike AIC, the terms in the higher co-efficients are insignificant.) Like AIC, Strict MML converges on the true model as the number of data goes to infinity. And we agree with Forster that the definition of the simplest model family containing the true model is dependent on how the parameter space is defined, and so is not something we should be interested in *per se*. Rather, we should be interested in inferring the true model itself. However, as we show below in section 7.1, a simple modification to Strict MML solves the problem of inferring the simplest model family containing the true model if we are interested in solving it.

[30]The form of the estimator described here and that used in (Wallace [1997]) and reported on in section 6.2.1 are not exactly the same, but the differences are inessential. The purpose of Wallace's study was to compare different model selection criteria. He therefore estimated the parameters of each model family using the method of Maximum Likelihood. However, since univariate polynomial regression is one of those problems in which the likelihood function is well-behaved, using MML (rather than Maximum Likelihood) to estimate the parameters of each model family would not have given significantly different answers.

[31]A nice example of this from a real world application of MML comes from spam detection (Oliver [2005]). Spammers often use templates that can be filled with random text to make each spam unique. Spam-detectors are therefore interested in the problem of inferring the template from which a given spam was generated. As Oliver explicitly notes, this is a problem where we are interested very much in inference to the best fully specified model, and not at all in predicting the precise form that future spam generated from that model will take.

[32]Strictly speaking, while Forster's claim is very nearly right, we don't completely agree with it. But there isn't space here to explore the reasons for our slight dissent.

[33]Here's his argument: consider an example in which there are two models A and B, where A asserts that $\theta = 0$ and B asserts that $\theta \neq 0$. If what it means to say that A is approximately true is that the true value of θ is infinitesimally close to 0, then there is a member of B that is also approximately true in virtue of being infinitesimally close to 0.

[34]Wallace and Boulton ([1975], sec. 3.3) give a justification of *why* we should choose this particular discrete model of the problem, appealing to a frequentist interpretation of probability. For problems whose context makes a frequentist interpretation inappropriate, Wallace elsewhere (Wallace and Dowe [1999a]; Wallace [2005], Ch. 6.7.2, p.275) appeals to an interpretation of probability in terms of Turing machines and algorithmic complexity.

[35]Classical statisticians do not normally discuss the accuracy to which the data is stated, despite the fact that all data must be recorded with some finite accuracy in order that it be recorded at all. ε can, however, be arbitrarily small. This is an important technical point: if ε were bounded below for all data points, AIC could escape the charge that it always prefers the gappy model: for then at some point, the uncertainty regions around the data points would overlap with each other, leaving no room for any gaps. However, this escape route out of statistical inconsistency for AIC only appears because of the introduction of an MML measurement accuracy, and the inconsistency returns if we simply let the measurement accuracy ε_i of data point x_i (previously ε) depend upon i and tend to 0 sufficiently quickly.

[36]Since there are regularity conditions that must be satisfied in order to apply AIC, it might be that one of the regularity conditions is not satisfied. But, in

that case, AIC selects no model at all!

[37]In everything that follows all logarithms are natural logarithms ($\log_e$), and hence all message lengths are in 'natural bits' or 'nits' (Boulton and Wallace [1970]; Comley and Dowe [2005], sec. 11.4.1). 1 nit = $\log_2 e$ bits; 1 bit = $\log_e 2$ nits.

[38]Because it does not matter in what order we state the parameters of the model, a and b, it possible to devise a code that saves $\log_e 2$ nits in the length of the assertion of that part of the gappy model.

[39]There is a potential concern with the expression $L_{\rm G}$ in the case that $b-a < \delta$, because in that case, the uncertainty regions around the gap-limits cross over (see Figure 1). However, as we show below at expression B.14, we only use $L_{\rm G}$ (in preference to $L_{\rm NG}$) in the case that $b - a > \approx (2 \log N)/N$, which for large N is substantially greater than $\delta = 2[1 - (b - a)]/(N - 2) \approx \frac{2}{N}$.

Dedication and Acknowledgements

This paper is dedicated to the memory of Chris Wallace (1933–2004). Our thanks to Gérald R. Petit and Vladimir Cherkassky. Our research was made possible by Australian Research Council Discovery Grant DP0343319.

Addresses

David Dowe

Clayton School of Information Technology

Monash University,

Clayton, VIC.

Australia 3800

URL: http://www.csse.monash.edu.au/∼dld

Steve Gardner

School of Philosophy and Bioethics

Monash University,

Clayton, VIC.

Australia 3800

email: Steven.Gardner@arts.monash.edu.au

Graham Oppy

School of Philosophy and Bioethics

Monash University,

Clayton, VIC.

Australia 3800

email: Graham.Oppy@arts.monash.edu.au

URL: http://www.arts.monash.edu.au/phil/department/Oppy/index.htm

References

Akaike, H. [1970]:'Statistical prediction information.', *Ann. Inst. Statist. Math*, **22**, pp. 203–217.

Akaike, H. [1973]:'Information theory and an extension of the maximum likelihood principle.', in B. N. Petrov and F. Csaki (eds.) *2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267–281.

Akaike, H. [1987]:'Factor analysis and AIC', *Psychometrika*, **52(3)**, pp. 317–332.

Barron, A. and Cover, T. [1991]:'Minimum complexity density estimation', *IEEE Transactions on Information Theory*, **37**, pp. 1034–1054.

Boulton, D. M. and Wallace, C. S. [1970]:'A program for numerical classification', *Computer Journal*, **13**, pp. 63–69.

Boulton, D. M. and Wallace, C. S. [1973]:'An information measure for hierarchic classification', *The Computer Journal*, **16**, pp. 254–261.

Cherkassky, V. and Ma, Y. [2003]:'Comparison of model selection for regression', *Neural Computation*, **15**, pp. 1691–1714.

Comley, J. W. and Dowe, D. L. [2003]:'General Bayesian networks and asymmetric languages', in *Proc. Hawaii International Conference on Statistics and Related Fields, 5-8 June, 2003*.

Comley, J. W. and Dowe, D. L. [2005]:'Minimum Message Length, MDL and generalised Bayesian networks with asymmetric languages', in P. Grünwald, M. A. Pitt and I. J. Myung (eds.) *Advances in Minimum Description Length: Theory and Applications*, M.I.T. Press, chap. 11, pp. 265–294.

Craven, P. and Wahba, G. [1979]:'Smoothing noisy data with spline

functions', *Numerische Mathematik*, **31**, pp. 377–403.

Dowe, D. L., Baxter, R. A., Oliver, J. J. and Wallace, C. S. [1998]:'Point Estimation using the Kullback-Leibler Loss Function and MML', in *Proc. 2nd Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, Melbourne, Australia: Springer Verlag, pp. 87–95.

Dowe, D. L., Oliver, J. J. and Wallace, C. S. [1996]:'MML estimation of the parameters of the spherical Fisher distribution', in S. Arikawa and A. K. Sharma (eds.) *Proc. 7th Conf. Algorithmic Learning Theory (ALT'96), LNAI 1160*, Sydney, Australia: Springer, pp. 213–227.

Dowe, D. L. and Wallace, C. S. [1997]:'Resolving the Neyman-Scott problem by Minimum Message Length', in *Proc. Computing Science and Statistics - 28th Symposium on the Interface*, vol. 28, pp. 614–618.

Earman, J. [1992]:*Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, MIT Press.

Farr, G. E. and Wallace, C. S. [2002]:'The complexity of Strict Minimum Message Length inference', *The Computer Journal*, **45(3)**, pp. 285–292.

Fisher, N. I. [1993]:*Statistical Analysis of Circular Data*, Cambridge Univeristy Press.

Fitzgibbon, L. J., Dowe, D. L. and Vahid, F. [2004]:'Minimum Message Length autoregressive model order selection.', in M. Palanaswami, C. C. Sekhar, G. K. Venayagamoorthy, S. Mohan and M. K. Ghantasala (eds.) *International Conference on Intelligent Sensors and Information Processing (ICISIP), (special session on 'Coding and Compression for Model Selection')*, pp. 439–444, URL `www.csse.monash.edu.au/~dld/Publications/2004/Fitzgibbon+Dowe%+Vahid2004.ref`.

Forster, M. R. [1995]:'Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation.', *British Journal for the*

Philosophy of Science, **46**, pp. 399–424.

Forster, M. R. [1999]:'Model selection in science: The problem of language variance.', *British Journal for the Philosophy of Science*, **50**, pp. 83–102.

Forster, M. R. [2000]:'Key concepts in model selection: Performance and generalizability.', *Journal of Mathematical Psychology*, **44**, pp. 205–231.

Forster, M. R. [2001]:'The new science of simplicity.', in A. Zellner, H. A. Keuzenkamp and M. McAleer (eds.) *Simplicity, Inference and Modelling*, University of Cambridge Press, pp. 83–119.

Forster, M. R. [2002]:'Predictive accuracy as an achievable goal in science.', *Philosophy of Science*, **69**, pp. S124–S134, URL `http://philosophy.wisc.edu/forster/PSA2000.htm`.

Forster, M. R. and Sober, E. [1994]:'How to tell when simpler, more unified or less ad hoc theories will provide more accurate predictions', *British Journal for the Philosophy of Science*, **45**, pp. 1–35.

Hannan, E. J. and Quinn, B. G. [1979]:'The determination of the order of an autoregression', *Journal of the Royal Statistical Society, Series B (Methodological)*, **41(2)**, pp. 190–195.

Kullback, S. and Leibler, R. A. [1951]:'On information and sufficiency.', *Annals of Mathematical Statistics*, **22**, pp. 79–86.

Murphy, P. M. and Pazzani, M. J. [1994]:'Exploring the decision forest: An empirical investigation of Occam's razor in decision tree induction', *Journal of Artifical Intelligence Research*, **1**, pp. 257–275.

Needham, S. L. and Dowe, D. L. [2001]:'Message length as an effective Ockham's Razor in decision tree induction', in *Proc. 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001)*, pp. 253–260, URL `http://www.csse.monash.edu.au/~dld/Publications/2001/Needham+%Dowe2001.ref`.

Neyman, J. and Scott, E. [1948]:'Consistent estimates based on partially consistent observations', *Econometrika*, **16**, pp. 1–32.

Oliver, J. J. [2005]:'Using lexigraphical distancing to block spam', in *Proceedings of the MIT Spam Conference*, URL `http://www.mailfrontier.com/docs/mit_jan_2005.pdf`.

Rissanen, J. J. [1978]:'Modeling by shortest data description', *Automatica*, **14**, pp. 465–471.

Rissanen, J. J. [1989]:*Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific.

Rissanen, J. J. [1999]:'Hypothesis selection and testing by the MDL principle', *The Computer Journal*, **42**, pp. 223–239.

Rumantir, G. W. and Wallace, C. S. [2003]:'Minimum Message Length criterion for second-order polynomial model selection applied to tropical cyclone intensity forecasting', in M. R. Berthold, H.-J. Lenz, E. Bradley, M. Kruse and C. Borgelt (eds.) *Advances in Intelligent Data Analysis V: Fifth International Symposium on Intelligent Data Analysis, IDA (2003)*, Springer-Verlag, pp. 486–496.

Schou, G. [1978]:'Estimation of the concentration parameter in von Mises-Fisher distributions', *Biometrika*, **65(1)**, pp. 369–77.

Schwartz, G. [1978]:'Estimating the dimension of a model', *Annals of Statistics*, **6**, pp. 461–464.

Shannon, C. E. [1948]:'A mathematical theory of communication', *Bell System Technical Journal*, **27**, pp. 379–423, 623–656.

Vapnik, V. [1995]:*The Nature of Statistical Learning Theory*, Springer.

Wallace, C. S. [1995]:'Multiple factor analysis by MML estimation', *Technical Report 95/218*, Dept. of Computer Science, Monash University, Clayton, Victoria 3800, Australia.

Wallace, C. S. [1996]:'False oracles and SMML estimators', in D. Dowe, K. Korb and J. Oliver (eds.) *Proc. Information, Statistics and Induction in Science conference (ISIS'96)*, Singapore: World Scientific, pp. 304–316.

Wallace, C. S. [1997]:'On the selection of the order of a polynomial model', *Tech. rep.*, Royal Holloway College.

Wallace, C. S. [2005]:*Statistical and Inductive Inference by Minimum Message Length*, Berlin, Germany: Springer.

Wallace, C. S. and Boulton, D. M. [1968]:'An information measure for classification', *Computer Journal*, **11**, pp. 185–194.

Wallace, C. S. and Boulton, D. M. [1975]:'An invariant Bayes method for point estimation', *Classification Society Bulletin*, **3(3)**, pp. 11–34.

Wallace, C. S. and Dowe, D. L. [1993]:'MML estimation of the von Mises concentration parameter', *Tech. Report TR 93/193*, Dept. of Comp. Sci., Monash Univ., Clayton 3800, Australia.

Wallace, C. S. and Dowe, D. L. [1999a]:'Minimum Message Length and Kolmogorov complexity', *Computer Journal*, **42(4)**, pp. 270–283, URL `http://comjnl.oxfordjournals.org/cgi/reprint/42/4/270`. Special issue on Kolmogorov complexity.

Wallace, C. S. and Dowe, D. L. [1999b]:'Refinements of MDL and MML coding', *Computer Journal*, **42(4)**, pp. 330–337. Special issue on Kolmogorov Complexity.

Wallace, C. S. and Dowe, D. L. [2000]:'MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions', *Journal of Statistics and Computing*, **10(1)**, pp. 73–83.

Wallace, C. S. and Freeman, P. R. [1987]:'Estimation and inference by compact coding', *Journal of the Royal Statistical Society (Series B)*, **49**, pp. 240–252.

Wallace, C. S. and Freeman, P. R. [1992]:'Single factor analysis by MML estimation', *Journal of the Royal Statistical Society (Series B)*, **54**, pp. 195–209.

Figure 1. The gappy model, showing parameters a and b stated to precision $\frac{\delta}{2}$, and one data point x_i stated to accuracy ε.

Table 1
Groups, success count ranges and estimates of the Strict MML estimator for the Binomial distribution, 100 trials, uniform prior.

j	c_j	h_j
1	0	0
2	1–6	0.035
3	7–17	0.012
4	18–32	0.25
5	33–49	0.41
6	50–66	0.58
7	67–81	0.74
8	82–93	0.875
9	94–99	0.965
10	100	1.0

Table 2
Comparison of MML and AIC on the task of selecting a polynomial approximation to a non-polynomial function. Adapted from (Wallace [1997]).

Target function: $y = \sin^2(\pi(x + 1.0))$ 1000 Cases, N = 10, NoiseSD = 0.61 signal/noise = 10.0				
KEY	MML		AIC	
AV	0.1857		15.8055	
SD	0.2633		63.8077	
5pc	0.0078		0.0091	
25pc	0.0385		0.0863	
50pc	0.1236		0.7974	
75pc	0.1880		5.4448	
95pc	0.6075		60.7315	
99pc	1.3700		306.5231	
Max	3.0411		771.4965	
DEG	avERR	CNT	avERR	CNT
0	0.141	222		0
1	0.281	33		0
2	0.406	27	7.587	2
3	0.698	23	13.099	17
4	0.303	177	39.709	78
5	0.421	30	18.996	214
6	0.106	426	10.882	340
7	0.112	52	18.547	195
8	0.095	10	7.068	154