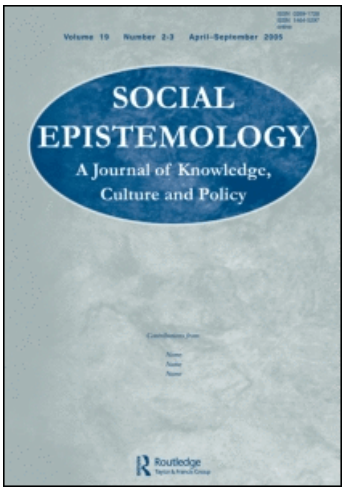

Social Epistemology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713765921

Minimum Message Length and Statistically Consistent Invariant (Objective?) Bayesian Probabilistic Inference—From (Medical) "Evidence"

David L. Dowe

Online Publication Date: 01 October 2008

To cite this Article Dowe, David L.(2008)'Minimum Message Length and Statistically Consistent Invariant (Objective?) Bayesian Probabilistic Inference—From (Medical) "Evidence"',Social Epistemology,22:4,433 — 460
To link to this Article: DOI: 10.1080/02691720802576291
URL: http://dx.doi.org/10.1080/02691720802576291

Social Epistemology
Vol. 22, No. 4, October–December 2008, pp. 433–460

Minimum Message Length and Statistically Consistent Invariant (Objective?) Bayesian Probabilistic Inference—From (Medical) "Evidence"

David L. Dowe

"Evidence" in the form of data collected and analysis thereof is fundamental to medicine, health and science. In this paper, we discuss the "evidence-based" aspect of evidence-based medicine in terms of statistical inference, acknowledging that this latter field of statistical inference often also goes by various near-synonymous names—such as inductive inference (amongst philosophers), econometrics (amongst economists), machine learning (amongst computer scientists) and, in more recent times, data mining (in some circles).

Three central issues to this discussion of "evidence-based" are (i) whether or not the statistical analysis can and/or should be objective and/or whether or not (subjective) prior knowledge can and/or should be incorporated, (ii) whether or not the analysis should be invariant to the framing of the problem (e.g. does it matter whether we analyse the ratio of proportions of morbidity to non-morbidity rather than simply the proportion of morbidity?), and (iii) whether or not, as we get more and more data, our analysis should be able to converge arbitrarily closely to the process which is generating our observed data.

For many problems of data analysis, it would appear that desiderata (ii) and (iii) above require us to invoke at least some form of subjective (Bayesian) prior knowledge. This sits uncomfortably with the understandable but perhaps impossible desire of many medical publications that at least all the statistical hypothesis testing has to be classical non-Bayesian—i.e. it is not permitted to use any (subjective) prior knowledge.

Keywords: Minimum Message Length; MML; Inference; Bayesianism; Statistical Invariance; Statistical Consistency; Evidence; Evidence-Based Medicine

David L. Dowe is Associate Professor of Computer Science at Monash University in suburban Melbourne. Correspondence to: Clayton School of Information Technology, Monash University, Clayton, Vic. 3800, Australia. Email: david.dowe@infotech.monash.edu.au.

ISSN 0269–1728 (print)/ISSN 1464–5297 (online)
DOI: 10.1080/02691720802576291

Introduction

Data is collected in medical and other scientific studies to provide "evidence" in support of or against a variety of hypotheses. Ultimately, we hope that collection and analysis of such data evidence in turn both enables us to accurately infer any underlying process from which the data is generated and also to accurately predict as yet unmeasured outcomes.

We will examine here several desirable properties—or desiderata—for a statistical inference technique in analysing medical and other data. We will address the issue of whether or not all of these desiderata can be simultaneously satisfied and when some sort of trade-off might be necessary.

One property that we want from our statistical inference technique is that of **statistical consistency**. Informally, this says that as the amount of data collected increases, we converge closer and closer and arbitrarily close to whatever underlying process can be said to be generating the data. Single and multiple latent factor analysis are but a few examples of problems for which frequently-used modelling tools are statistically inconsistent.

Another property which we want from our statistical inference tool is the ability to make probabilistic models and accurately quantify noise. Diagnoses such as "yes"/"no" or "presence"/"absence" of some condition are less useful than models which give a probability of a diagnosis. Rather than respond with "no", a system returning a probability of (say) 10% of some condition enables the medical experts to decide upon possible treatment and further tests; and certainly there is much more difference between (say) 10% and 45% than there is between two (less informative) responses of "no".

Another property which we presumably also want from our statistical inference tool is that of **statistical invariance**—namely, that the inferred value is independent of the framing of the problem. By "*framing*", I don't particularly mean linguistic framing but rather a statistical or (statistically) parametric framing. (For example, variations of Bertrand's paradox say that in a cube of side-length between 1 and 2, the side-length has probability 1/2 of being less than 1.5 but the volume has probability 1/2 of being less than 4.5, which—paradoxically?—is not 1.5^3.) To elaborate, if we know a skin lesion or tumour to be circular, then statistical invariance would require that the estimated area is equal to π times the square of the estimated radius. Whether we parameterise in terms of radius or area, we get the same answer.

Perhaps the single main other issue to mention in the use of "evidence" is the difference between *inference* and *prediction*. Inference is the use of one—ideally the "best"—theory to model the observed data and find a pattern within it. Prediction is concerned with forecasting as yet unseen data. Unless the currently observed data has one outstanding single best theory, prediction is often best done by combining more than one theory.

Desiderata in (Probabilistic) Inference and (Probabilistic) Prediction

Data can be both time-consuming and expensive to collect and obtain. It is often useful to know the accuracy to which the data was measured (Wallace and Dowe 1993, 1–3,

1994, 38, secs 2 and 2.1, 2000, sec. 2, 74, col. 2; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651; Comley and Dowe 2003, sec. 9, 2005, sec. 11.3.3, 270; Fitzgibbon, Dowe and Vahid 2004, eqn (19); Wallace 2005, secs 3.1.1 and 3.3; Dowe, Gardner and Oppy 2007; Dowe 2008, sec. 0.2.4)—as, after all, no-one knows their height or weight to infinitely many decimal places (if such a notion were even to make sense). It is also important to make good use of the data—whether we are making some sort of (probabilistic) inference, doing some sort of (probabilistic) prediction or perhaps (Dowe 2008, sec. 0.2.5) doing some kind of hypothesis test.

When doing (probabilistic) inference to some hypothesis, H, from (observed) data, D, we look at some possible desiderata, or properties that we might desire in our inference technique(s).

Statistical Invariance

Many problems can be phrased in several equivalent ways. Informally, *statistical invariance* says that we infer the same answer no matter how we phrase (or parameterise) the problem. Let us give several examples to clarify this point:

1. if p is the proportion of the population with a certain condition (or illness, diagnosis, prognosis, etc.) and q is the relative "odds ratio" proportion of those thus affected divided by those unaffected, then $q = p/(1-p)$ and $p = q/(1+q)$;
2. if r and A are the radius and area of a circle respectively through which an epidemic has spread (or, alternatively, of a surface lesion), then $A = \pi r^2$ and $r = \sqrt{A/\pi}$;
3. if a cube (maybe call it C) has side length l, face area A and volume V, then $l = A^{1/2} = V^{1/3}$, $A = l^2 = V^{2/3}$ and $V = l^3 = A^{3/2}$;
4. if a vector in the plane (such as direction and strength of a magnetic field) has direction θ and distance (or strength), κ (in polar co-ordinates) and can also be thought of (in Cartesian co-ordinates) as (x, y), then $(x, y) = (\kappa \cos\theta, \kappa \sin\theta)$ and[1] $(\kappa, \theta) = (\sqrt{x^2 + y^2}, \tan^{-1}(y/x))$.

In the language of statistical inference, $\hat{\theta}$ denotes the estimated value of θ. The hat (or circumflex), $\hat{}$, denotes an estimated value. Recall that, informally, statistical invariance says that we get the same answer no matter how we phrase (or parameterise) the problem. So, for example, with item 1 above, statistical invariance of an estimator would give us that $\hat{q} = \hat{p}/(1-\hat{p})$ and equivalently $\hat{p} = \hat{q}/(1+\hat{q})$. Not all problems have a "natural" parameterisation (or framing), so if we don't have statistical invariance then we have to get a different estimate for each re-parameterisation (or re-framing), potentially leading to awkward situations where for some cube (as in item 3) we might perhaps rather curiously estimate poorly matching values such as (e.g.) $\hat{l} = 0.98$, $\hat{A} = 1.03$ and $\hat{V} = 0.97$.

Notice also that many notions of "error", like bias and squared error, are not invariant to re-parameterisation. However, the notion of Kullback–Leibler divergence (or Kullback–Leibler distance) from the next section is one measure which is invariant under re-parameterisation.

Kullback–Leibler Divergence (or Kullback–Leibler Distance)

The Kullback–Leibler divergence is a measure of the difference between two probability distributions. It has the property of being invariant to re-parameterisation. The divergence is typically not symmetrical, though, meaning that the distance from distribution *f* to distribution *g* is not necessarily the same as the distance from distribution *g* to distribution *f*. This lack of symmetry is why some prefer the term "*divergence*" to "*distance*".

If *f* and *g* are both discrete distributions, with probabilities $f_1, \ldots, f_N$ and $g_1, \ldots, g_N$ for an *N*-state multinomial distribution (such as a two-sided coin with $N = 2$, or a six-sided dice with $N = 6$), then the Kullback–Leibler distance from *f* to *g* is defined as $KL(f,g) = \Delta(g\|f) = \sum_{i=1}^{N} f_i \log(f_i / g_i)$.

If *f* and *g* are both continuous-valued distributions, then we replace the summation by an integral and the Kullback–Leibler distance from *f* to *g* is defined as $KL(f, g) = \Delta(g||f) = \int f \log(f/g)$.

The Kullback–Leibler distance between two Bayesian networks (or graphical models) *f* and *g* can be defined as in Tan and Dowe (2006, sec. 4.2) or Dowe (2008, sec 0.2.5) and the Kullback–Leibler distance between two mixture models can be defined similarly. And there is no problem having a hybrid of both discrete- and continuous-valued variables.

Statistical Consistency

Informally, statistical consistency says that, as we get more and more data, we converge more and more closely—and, ultimately, arbitrarily closely—to the true underlying model. More formally, if θ is a parameter value, *N* is a sample size and $\hat{\theta}$ is a parameter estimate from a sample of size *N*, then *statistical consistency* says that

$$\forall \theta \, \forall \in > 0 \, \exists N_0 \, \forall N \geq N_0 \; Pr(|\theta - \hat{\theta}| < \in) > 1 - \in.$$

In other words, as we get more and more data, then with arbitrarily large probability we can converge arbitrarily closely to any true underlying model. Given our intuition that more and more data should enable us to infer more and more accurately, and given how expensive and time-consuming it can be to collect data, statistical consistency—that more data will ultimately take us to the correct answer—seems like one of the very least things we should seek in an inference method.

The notion of statistical consistency raises at least four other issues. First, it raises the issue of *efficiency* (Wallace 2005, sec. 3.4.5; Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5, especially footnote 162), the idea of not just converging on the true model (consistency) but of converging on the true value as quickly—or as efficiently—as possible. A second issue, perhaps subtly different to (asymptotic) efficiency, is that of—loosely speaking—performing well on small sample sizes. Statistical consistency guarantees asymptotic convergence, and (asymptotic) efficiency guarantees performing as well as possible on large (asymptotic) sample sizes. Efficiency and excellent small-sample performance are surely related, but surely also not identical. Third, it raises the issue of consistency when the model is misspecified (or, equivalently, when the true

model is not in the class of models being considered by the estimators) (Grünwald and Langford 2007; Dowe 2008, sec. 0.2.5). Given that many, if not perhaps most, inference problems have to contend with misspecification (e.g. Normal distributions are often used to model heights and other variables that can't take negative values), misspecification and inference methods which might or might not be susceptible to its vagaries (Grünwald and Langford 2007; Dowe 2008, sec. 0.2.5) should be paid greater attention. Fourth (and last), there is the issue of methods which are statistically consistent for (easier) problems where the number of parameters remains fixed but which do or don't remain statistically consistent for (harder) problems where the number of parameters increases as the amount of data increases (to the point where the amount of data per parameter is always bounded above, as in section "Amount of Data per Parameter Bounded Above"). It is known that some inference methods (such as Maximum Likelihood and AIC from sections "Maximum Likelihood" and "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" often become statistically inconsistent in such cases (Neyman and Scott 1948), while at least one other method (MML from section "Minimum Message Length") appears to remain statistically consistent (Dowe and Wallace 1997; Wallace 2005, secs 4.2–4.5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5).

Probabilistic Inference—vs. Mere Non-probabilistic Classification

Many inference problems are in ([supervised or] extrinsic) classification and—especially within the machine learning community—these are often regarded as problems of right vs. wrong. Probabilities are often neglected, whereas for certain ([moderately] serious) medical conditions the threshold for further investigation or treatment might well be other than 50%. If a patient presenting with chest pains were deemed to "only" have a 40% probability of heart attack or even deemed to "only" have a probability of 15% of heart attack, it would be somewhere along the lines of irresponsible, negligent and legally challenging to classify this as a "no" and not give treatment. This remains true whether the patient was presenting in person or telephoning a service such as *Nurse On Call* for a provisional symptom-based assessment over the phone. Even in DNA microarray classification, it is more prudent and probably also safer to report probabilities. While "right"/"wrong" is a fairly easy and seemingly natural scoring system to use, it is not invariant to re-framing of questions. As an example, consider a four-class problem which can be divided in three reasonable different ways into two two-class problems. The "right"/"wrong" score will depend upon the relevant division. However, probabilistic inference with log-loss scoring (Dowe and Krusel 1993, 4, Table 3; Dowe et al. 1998, sec. 3; Needham and Dowe 2001, Figs 3–5; Tan and Dowe 2002, sec. 4, 2004, sec. 3.1, 2006, secs 4.2–4.3; Kornienko, Dowe and Albrecht 2002, Table 2; Comley and Dowe 2003, sec. 9, 2005, sec. 11.4.2; Tan and Dowe 2003, sec. 5.1; Kornienko, Albrecht and Dowe 2005a, Tables 2–3, 2005b; Tan, Dowe and Dix 2007, sec. 4.3; Dowe 2008, sec. 0.2.5, especially footnote 175 [and 176]) not only scores probabilities, but it also has the desirable feature that the optimal long-term strategy is to give the true probabilities (if known).

If you assign probabilities $\{p_i : i = 1, \ldots, N\}$ for events $\{e_i : i = 1, \ldots, N\}$ such that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$, then for some constant c (however chosen), if event e_j is the event that actually happened, then log-loss (or "*probabilistic bit-cost*") scoring awards a score of $c + \log p_j$. This scoring system has been used for Australian Football League (AFL) matches since early 1995 (Dowe, Farr et al. 1996; Dowe et al. 1998, sec. 3; Dowe 2008, sec. 0.2.5). With a probability of p on one team (and $1 - p$ on the other—in a match between two teams), using the constant $c = 1$, this competition at www.csse.monash.edu.au/~footy gives scores of $1 + \log_2 p$ if you're right, and $1 + \log_2 (1 - p)$ if you're wrong.

Log-loss scoring is invariant under re-framing of the problem (Dowe 2008, sec. 0.2.5, especially footnote 175 [and 176]), and appears—rather importantly—to enjoy the property of being unique in this respect.

And just as log-loss scoring appears to be unique in its invariance under re-framing of the problem, so, too, in some sense the (analogous) Kullback–Leibler divergence from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)" seems to also be unique in retaining invariance under re-framing of a problem. Both $KL(f, g) = \Delta(g||f)$ and $KL(g, f) = \Delta(f||g)$ are invariant to the level of detail of re-framing of the problem and appear to be unique in having this property—although, clearly, any linear combination $\alpha KL(f, g) + (1 - \alpha)KL(g, f)$ (with $0 \leq \alpha \leq 1$) will also share this invariance. (Interestingly, the difference between the approaches in Dowe (2008, sec. 0.2.2, footnotes 64 and 65) mentioned in section "Properties of MML (and Approximations)" largely comes down to the difference between $KL(f, g)$ and $KL(g, f)$. This said, for those interested in the finer detail of current state-of-the-art MML approximations, it seems opportune here to re-visit an issue from (Dowe, 2008, sec. 0.2.2, footnote 65). Upon reflection, (Dowe, 2008, sec. 0.2.2, footnote 64, eq (3)) should be *further* from Maximum Likelihood than the method from (Dowe, 2008, sec. 0.2.2, footnote 65). I wrap up this note by idly speculating about the merits of returning to (Dowe, 2008, sec. 0.2.2, footnotes 64 and 65) with a hybrid method involving $\alpha KL(\theta^*, \theta) + (1 - \alpha)KL(\theta, \theta^*)$ with $\alpha = 1/2$.)

And just as log-loss scoring retains the above uniqueness in its invariance under re-framing of the problem when we add (or subtract) the entropy of the prior (or a multiple thereof) (Dowe, 2008, footnote 176), again, so, too, the Kullback–Leibler divergence—or even any linear combination $\alpha KL(f, g) + (1 - \alpha)KL(g, f)$ (with $0 \leq \alpha \leq 1$)—retains its invariance when we add (or subtract) the entropy of the prior (or a multiple thereof).

Bayesianism vs. Non-Bayesianism

Much metaphorical "blood" has been spilt on the issue of whether or not prior beliefs should be incorporated into analysing data. Bayesians are those who contend that any prior knowledge should be used. While this seems fairly clear (to me), it opens some cans of worms. One issue is exactly how should we quantify our prior beliefs? Another issue is to ask what two different people with different prior beliefs—or, more

extremely, (as expert witnesses) in opposing sides of a class action, malpractice suit or other legal battle—should do to reconcile the fact that their different prior beliefs will give rise to different answers.

Some classical (non-Bayesian) statisticians have suggested quite wrongly that Bayesian methods are not statistically invariant. While it is true that some Bayesian methods are not statistically invariant, some most certainly are (Wallace and Boulton 1975).

Some Bayesian statisticians are almost self-conscious about the presence of a prior probability distribution representing prior beliefs, and try in a variety of ways to make such a term as objective as possible. Taking the log-likelihood from section "Maximum Likelihood" the Fisher information is the determinant of the matrix of the expected second partial derivatives of the log-likelihood (Wallace 2005, sec. 5.1). One of many attempts to be as objective as possible while still being Bayesian is to use the observation of Jeffreys that the Fisher information has the same mathematical form as a prior (Jeffreys 1946), and to use it as a prior. Jeffreys himself never advocated this (Wallace 2005, sec. 1.15.3), it seems rather odd that our *prior* beliefs should depend upon the observed data, and this (so-called) "Jeffreys prior" frequently either has an infinite integral or other failings (Wallace and Dowe 1999a, sec. 5, 277, col. 2, 1999b, sec. 2.3; Comley and Dowe 2005, sec. 11.4.3, 273; Wallace 2005, sec. 10.2.1; Dowe 2008, footnote 75).

It would be fair to say that the community is still a long way from being unified in the best way to analyse data. But it must be pointed out that not only can Bayesian methods be statistically invariant (Wallace and Boulton 1975; Wallace 2005) but that, furthermore, it has been conjectured (Dowe et al. 1998, 93; Edwards and Dowe 1998, sec. 5.3; Wallace and Dowe 1999a, 282, 2000, sec. 5; Comley and Dowe 2005, sec. 11.3.1, 269; Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5) that only Bayesian methods can give both statistical invariance and statistical consistency on the harder problems (with the amount of data per parameter bounded above) in sections "Statistical Consistency" and "Amount of Data per Parameter Bounded Above".

We now look at (probabilistic) prediction in the next section and then, in section "Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc." we look at some classical (non-Bayesian) and some Bayesian approaches to inference, including (in section "Bayes's Theorem and Bayesianism") a discussion of Bayes's theorem.

(Probabilistic) Prediction

The distinction between *inference* and *prediction* is that inference is concerned with finding the single best theory while prediction is concerned with finding the most probable future data—see also Wallace and Dowe (1999a, sec. 8) and Wallace (2005, sec. 10.1.2). Classical non-Bayesians seem either to conflate these two notions or to regard the single best inference as necessarily being the best predictor. In the Bayesian approach, theories have a prior probability (distribution) before the data is seen and then a posterior probability (distribution) after the data is seen.

The optimal Bayesian predictor combines all theories available, weighting them according to their respective posterior probabilities. Classical non-Bayesian inference tends to over-fit and err on the side of under-estimating any spread in the data. The single best (Bayesian) inference tends to give the best estimate of spread in the data. The best (Bayesian) predictor makes a weighted Bayesian combination of theories, as in section "Prediction". This results in a slightly conservative over-estimate of the spread in the data, due to the combination of diverse theories (Wallace 2005, sec. 4.9).

And, of course, the quality of any probabilistic predictions can be measured using the log-loss ("*probabilistic bit cost*") scoring method from section "Probabilistic Inference—vs. Mere Non-probabilistic Classification".

Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc.

Given data, D, how do we (best) choose which hypothesis, H, to infer? Recalling the discussion of Bayesianism from section "Bayesianism vs. Non-Bayesianism", we look at several approaches below. We consider classical (non-Bayesian) approaches in sections "Maximum Likelihood", "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" and "Other: Other Classical, Other Bayesian, etc.", and we consider Bayesian approaches in sections "Bayes's Theorem and Bayesianism", "Maximum A Posteriori (MAP)", "Other: Other Classical, Other Bayesian, etc." and "Minimum Message Length (MML)".

Maximum Likelihood

Maximum Likelihood says that, given data D, we should choose the hypothesis, H, for which the likelihood $Pr(D|H)$ is maximised. Given the monotonicity of the likelihood function, Maximum Likelihood is equivalent to minimising $-\log Pr(D|H)$.

This classical approach to inference is statistically invariant—and a hand-waving argument for this is that stretching the likelihood function in and out sideways will not affect the maximum height or any height. But Maximum Likelihood tends to over-fit (especially on small sample sizes) (Wallace and Dowe 1993), "finding" non-existent patterns in random noise. One simple case in point is, where even in the case of the Gaussian distribution, the Maximum Likelihood estimator of the variance has to be corrected and multiplied by $\frac{N}{N-1}$ for sample size, N (Dowe, Gardner and Oppy 2007, sec. 6.1.1). Another simple case in point is the *bus number problem* (Dowe 2008, footnote 116, 535–536), where we arrive in a new town with θ buses numbered consecutively from 1 to θ. If we see only one bus and observe its number, x_{obs}, then Maximum Likelihood tells us to estimate θ as x_{obs}. This will typically be a silly under-estimate.

At least two or three more issues arise with Maximum Likelihood.

One issue is how do we choose between models of increasing complexity and increasingly good fit—e.g. constant, linear, quadratic, cubic, …? Maximum Likelihood advocates an unambiguous approach when all is parameterised (e.g. we know that the function is linear with Gaussian noise), but when models are nested it doesn't give a way of avoiding the most complicated model.

A second issue is that Maximum Likelihood chooses the hypothesis to make the already observed data as likely as possible. But the data has already been observed—so, philosophically, choosing the hypothesis to make the already observed data as (retrospectively) probable as possible seems to be stating the problem back to front. Shouldn't we instead find some way of choosing H so as to maximise $Pr(H|D)$? Plenty of Bayesians might consider this to be self-evident or at worst close to conclusive (Berger and Wolpert 1988; Bernardo and Smith 1994) (but classical likelihood-based reasoning and its advocates do live on (Glymour 1981; Forster and Sober 1994), as per section "Other: Other Classical, Other Bayesian, etc.").

A third issue, which is mentioned in section "Statistical Consistency", is that Maximum Likelihood is known to be statistically inconsistent for a wide range of problems where the amount of data per parameter is bounded above (Neyman and Scott 1948; Wallace and Freeman 1992; Wallace 1995; Wallace and Dowe 2000, sec. 5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5).

Akaike's Information Criterion (AIC)—see next section—is one attempt to address the first issue. The second issue is the contentious "Bayesianism vs. non-Bayesianism" issue of section "Bayesianism vs. non-Bayesianism". If we think that maximising $Pr(H|D)$ makes more sense than maximising $Pr(D|H)$, then it makes sense to explore Bayesian approaches—such as Maximum A Posteriori (MAP) and Minimum Message Length (MML) from sections "Maximum A Posteriori (MAP)" and "Minimum Message Length (MML)" respectively, both of which use Bayes's theorem (from section "Bayes's Theorem and Bayesianism").

Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood

Where Maximum Likelihood advocates minimising $-\log Pr(D|H)$, the Akaike Information Criterion (AIC) advocates minimising $2.(-\log Pr(D|H) + k)$, or equivalently $(-\log Pr(D|H) + k)$, where k is the number of free parameters (Akaike 1970, 1973). (It is worth mentioning that there is a substantial literature on AIC where the penalty, k, has been changed to, e.g. $\frac{3}{2}k$ and a variety of other constants multiplied by k. This will typically not change things overly much.) So, in the special case when the model class is known, the relevant variables have already been selected and we only need to do parameter estimation (e.g. we are fitting a univariate cubic polynomial with Gaussian noise, $y = (\sum_{i=0}^{3} a_i x^i) + N(0,\sigma^2)$ as per section "Problems with Increasing Numbers of Parameters" for some (a_0, a_1, a_2, a_3, σ^2) to be inferred), AIC reduces to Maximum Likelihood. The fact that AIC is the likelihood function with a penalty term (namely, k) means that AIC can be regarded as a form of *penalised likelihood*.

For the wide range of problems where the amount of data per parameter is bounded above (Neyman and Scott 1948; Wallace and Freeman 1992; Wallace 1995; Wallace and Dowe 2000, sec. 5; Dowe, Gardner and Oppy 2007, secs 6.1 and 8; Dowe 2008, sec. 0.2.5) and there is no variable selection, AIC reduces to Maximum Likelihood and suffers the same problems of statistical inconsistency. For a comparison of AIC and the

Bayesian MML approach (from section "Minimum Message Length (MML)"), see Wallace and Dowe (1999a, sec. 9) and Dowe, Gardner and Oppy (2007).

Bayes's Theorem and Bayesianism

Following discussions such as that in section "Bayesianism vs. non-Bayesianism" let us explore Bayesianism—the notion that we should look at $Pr(H|D)$ rather than at $Pr(D|H)$.

The Bayesian approach takes into account our prior beliefs over the space of possible hypotheses. We will write the prior probability of H as Pr(H). This is the probability distribution over the space of hypotheses prior to—or *before*—seeing any data. We can combine the prior, $Pr(H)$ and the (statistical) likelihood function, $Pr(D|H)$, to calculate the posterior distribution—which is the probability of hypotheses *after* seeing the data. The relationship between the prior ($Pr(H)$), the likelihood ($Pr(D|H)$) and the posterior ($Pr(H|D)$) can be shown using Bayes's theorem, which can also be thought of in terms of a Venn diagram. Repeated application of Bayes's theorem thus gives

$$Pr(H) \cdot Pr(D \mid H) = Pr(H \,\&\, D) = Pr(D \,\&\, H) = Pr(D) \cdot Pr(H \mid D). \tag{1}$$

So, this now gives the posterior probability of H given D as

$$\begin{aligned} \text{posterior}\,(H \mid D) &= \Pr(H \mid D) = \frac{\Pr(H) \cdot \Pr(D \mid H)}{\Pr(D)} = \frac{1}{\Pr(D)}(\Pr(H) \cdot \Pr(D \mid H)) \\ &= \frac{\text{prior}(H) \cdot \text{likelihood}(D \mid H)}{\text{marginal}(D)}, \end{aligned} \tag{2}$$

where the marginal probability of D, $Pr(D)$ or marginal(D), is the prior probability that D is the data-set generated. Informally, depending upon whether H is a discrete space over which we sum or a continuous space over which we integrate, we can write $Pr(D) = \sum_H Pr(H)Pr(D \mid H)$ or $Pr(D) = \int_H Pr(H)Pr(D \mid H)\mathrm{d}H$. Discrete spaces include cases such as all attributes are categorical (e.g. drinker or non-drinker, smoker or non-smoker, male or female, etc.) and continuous spaces include attributes such as height or weight. Of course, hybrid spaces with both discrete (categorical) and continuous attributes exist, and the above formula for Pr(D) is modified to sum over the discrete attributes and integrate over the continuous attributes. The term *attribute* sometimes goes by alternative names including (e.g.) *dimension*, *feature*, *field* and *variable*. Note that all the hypotheses are used to calculate $Pr(D)$ but that they are all summed or integrated out—and, as such, $Pr(D)$ is independent of any individual hypothesis or rival hypotheses being considered for inference. So, from Equation 2, maximising $Pr(H|D)$ is equivalent to maximising $Pr(H) \cdot Pr(D|H)$.

The Bayesian interested in doing inference is quite probably going to be interested in choosing H to maximise $Pr(H|D)$—or, equivalently, to maximise $Pr(H) \cdot Pr(D|H)$. But issues arise here and, if we are not careful and principled, we might be

left open to a criticism from a classical statistician along the lines of "Classical approaches based on likelihood and penalised likelihood are invariant under re-parameterisation, but maximising the Bayesian posterior usually isn't". One issue here will be when we are dealing exclusively with discrete (categorical) attributes—and so are dealing with probabilities—and when instead at least one of our attributes is continuous and so we are dealing not with probabilities per se but with *probability densities*.

Maximum A Posteriori (MAP)

As its name suggests, the Bayesian method of Maximum A Posteriori (or MAP) maximises the posterior probability (or density), $Pr(H|D)$, or equivalently, the prior multiplied by likelihood. When all attributes are discrete (categorical), this is statistically invariant under re-parameterisation. However, when at least one of the attributes is continuous, then both $Pr(H)$ and $Pr(H|D)$ are *densities*. As an example, if the hypothesis, H, concerns a height, then $Pr(H)$ and $Pr(H|D)$ must be measured in units of 1/length, or length^{-1}, in order that the integral of the prior along the height axis gives a probability of 1. In other words, if we're multiplying something by a height (in cm) and the answer is 1, then that something must be in cm^{-1}. This gives us some insight into why MAP is generally not statistically invariant. The Bayesian prior on a length will look quite different to the prior on its square, an area—and, indeed, their maxima and minima, etc. will generally be different. The statistical likelihood ($Pr(D|H)$) is invariant but the prior ($Pr(H)$) isn't, so $Pr(H) \cdot Pr(D|H)$ and the posterior also won't be invariant in general, and therefore the maximum of the posterior—namely, the MAP estimate—also won't be invariant (Dowe, Oliver and Wallace 1996; Wallace and Dowe 1999b, secs 1.2–1.3, 1999c, sec 2, col. 1, 2000, secs 2 and 6.1; Comley and Dowe 2005, sec. 11.3.1; Dowe 2008, sec. 0.2.3). The similarity between MAP and Maximum Likelihood means that MAP inherits the statistical inconsistency results of Maximum Likelihood described in section "Maximum Likelihood" for problems where the amount of data per parameter is bounded above. Even when all attributes are discrete (and so issues of density do not arise), even then MAP can inherit the statistical inconsistency tendencies of Maximum Likelihood for problems where the amount of data per parameter is bounded above (Dowe 2008, footnote 158).

The good news is that if we re-visit MAP very carefully and make sure that our posterior is a *probability* and not a *density*, then we arrive at something which is statistically invariant. If we take some more care (when required), then we also get statistical consistency for the hard problems where the amount of data per parameter is bounded above. This approach is Minimum Message Length (MML) (Wallace and Boulton 1968, 1975; Wallace 2005), which we will discuss in section "Minimum Message Length (MML)". But, first, we all too briefly gloss over some of the many other approaches to inference in the next section and then—in the remainder of section "Some Methods of Inference: Maximum Likelihood, AIC, (Bayesian) MAP, etc."—touch on other issues pertaining to inference.

Other: Other Classical, Other Bayesian, etc.

In this section, we attempt to mention some of the myriad of alternative estimation techniques used in the literature and not yet discussed above. One can only do one's best with such an impossible task, but it is worth re-emphasising the point from section "Bayesianism vs. Non-Bayesianism" that the community remains far from unified in how best to do inference. The classical (non-Bayesian) community is far from unified. And, whether or not MML is "*the*" way to do inference, the Bayesian community currently remains a long way from unified.

Schwarz's (1978) Bayesian Information Criterion (BIC) is independent from and coincidentally equivalent to the 1978 version of Minimum Description Length (MDL) (Rissanen 1978) (which, in turn, shares much in common with MML, as per Wallace and Dowe (1999a, secs 6.2 and 7, 1999b), Wallace (2005, sec. 10.2), Comley and Dowe (2005, sec. 11.4.3) and Dowe (2008, sec. 0.2.2), although MML pre-dates MDL by a decade (Wallace and Dowe 1999a, sec. 1, 271, col. 1; Comley and Dowe 2005, sec. 11.1; Dowe 2008, secs 0.2.2–0.2.4)). Recall from section "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" that AIC was a penalised likelihood of the form $-\log Pr(D|H) + k$ where k is the number of free parameters. BIC advocates minimising $-\log Pr(D|H) + \frac{k}{2} \log N$, where N is the sample size of the data. For sufficiently large N (indeed, once $N \geq 8 > e^2$ and $\log N > 2$), we see that the BIC penalty of $\frac{k}{2} \log N$ becomes greater than the AIC penalty of k. So, for larger sample sizes, BIC tends to give a larger (and, we contend, more appropriate) penalty than AIC.

The Vapnik–Chervonenkis dimension, Structural Risk Minimisation (SRM) and Support Vector Machine (SVM) approach (Vapnik 1995) is a (classical or) non-Bayesian approach which came from the machine learning community and is only slowly working its way through statistics and econometrics. That said, there have been efforts to do this in a Bayesian way and also in a (Bayesian) MML way (Vapnik 1995, sec. 4.6; Tan and Dowe 2004; Dowe 2007, 2008, sec. 0.2.2), including explicitly modelling (Dowe 2008, footnote 53, fourth way, 527–528) the distribution of *all* the variables, including the input variables.

The minimum expected Kullback–Leibler distance (MEKLD, or MEKL, or minEKL) estimator (Dowe et al. 1998; Wallace 2005, secs 4.7–4.9; Dowe, Gardner and Oppy 2007, sec. 6.1.4) is a Bayesian estimator which uses the notion of Kullback–Leibler distance (from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)") to attempt to optimise the (average) log-loss probabilistic score from section "Probabilistic Inference—vs. Mere Non-probabilistic Classification" on future, as yet unseen, data. It does this by taking the Bayesian posterior distribution $Pr(H|D)$ over hypotheses H to then get a distribution $f(y|D)$ on "expected" future data, y. Having such a probability distribution on the "expected" future data, it then seeks a hypothesis H which - in average expectation - optimises the "expected" log-loss penalty. The purpose of minEKL is to be the best possible (Bayesian) predictor within the parameter space. It is perhaps curious that this was the original motivation behind AIC (Akaike 1970, 1973), although Akaike tried to do this without use of a Bayesian prior. A comparison between MEKLD and AIC is given in Dowe, Gardner and Oppy (2007, sec. 6.1.4).

We now say something about hypothesis testing (as a form of inference), experimental design and prediction in the next three sections respectively, and then talk about Bayesian Minimum Message Length (MML) inference in section "Minimum Message Length (MML)".

Hypothesis Testing

Recall from section "Maximum Likelihood" that Maximum Likelihood tries to choose a hypothesis, H, to make the already observed data, D, as retrospectively likely as possible. Classical hypothesis tests do the same curious thing, trying to say how probable the observed data would be if the actual hypothesis were true—rather than how probable the hypothesis is given the data. As such, classical hypothesis tests—like maximum likelihood—often neglect how complicated or even tightly-peaked the hypothesis is (Dowe 2008, sec. 1 and footnotes 57 and 58). In fairness, the classical hypothesis test tries to objectively side-step any use of Bayesian priors, although they often (inadvertently?) include a prior which can be slightly curious (Dowe 2008, sec. 0.2.5).

Experimental Design, Data Collection Protocol and Likelihood Principle

This (brief) section is partly to mention that any experiment should be designed "randomly" to collect as much information as possible (Dowe 2008, sec. 0.2.7, 544). Whatever the data collection protocol, the statistical *likelihood principle* says, roughly, that the likelihood function $Pr(D|H)$ is all that we need to know about the data (Berger and Wolpert 1988; Grossman forthcoming). As such, Maximum Likelihood will always honour the likelihood principle. In changing from a Binomial protocol (when we sample a fixed number of times) to a Negative Binomial protocol (when we sample until a fixed number of successes), MML (from section "Minimum Message Length (MML)") gives at worst a minor violation of the likelihood principle (Wallace 2005, sec. 5.8) (although I am not convinced that this constitutes a valid criticism of MML). But, as discussed in Wallace and Dowe (1999b, sec. 2.3.5) and Wallace (2005, sec. 10.2.2), some inference methods—even those doing all they can to avoid using a Bayesian prior—can be in substantial violation of the likelihood principle.

Having said above something one could interpret as meaning that we wish to design our experiment to have the maximum expected information gain, over the next 16 lines or so I'd now like to change tack and put forward something of a paradox—or the making thereof. Consider experiments (or tests) T_1, T_2, …, T_s, …, T_i, … such that, for some $n > 0$, experiment T_i has probability $2^{-2^{-(n+i)}}$ of yielding 0 information. The experiments can be independent of one another, and perhaps are such that with probability $1-2^{-2^{-(n+i)}}$, experiment T_i yields $2^i/(1-2^{-2^{-(n+i)}})$ bits, thus yielding an expected information gain from experiment T_i of $(1-2^{-2^{-(n+i)}}) \times (2^i/(1-2^{-2^{-(n+i)}})) = 2^i$ bits. Letting $s \geq 1$, the probability that, starting

with experiment T_s, all the experiments T_s, T_{s+1}, … yield 0 information is $2^{-2^{-(n+s)}} \times 2^{-2^{-(n+s+1)}} \times \ldots = 2^{-(2^{-(n+s)}+2^{-(n+s+1)}+\ldots)} = 2^{-(2^{-(n+s-1)})}$. On the other hand, the expected information gain from starting with experiment T_s is $2^s + 2^{s+1} + \ldots = \infty$. Paradoxically, for a finite number (say j) of experiments, T_s, …, T_{s+j-1}, the larger the value of s that we start with, the larger our expected information gain but the greater the probability that all the experiments from T_s to T_{s+j-1} and forever will all yield 0 information. With $n > 0$, this probability $2^{-(2^{-(n+s-1)})}$ of getting no information rapidly approaches 1 for increasing s. We can extend the paradox by averaging the information gains of T_s, T_{s+1}, …, T_u and then letting u tend to infinity. As u gets larger, the expected average information gain (divided by (u-s+1)) tends to infinity (on the one hand) but—(on the other hand) curiously—the probability that the average information gain (divided by (u–s+1)) is arbitrarily close to 0 becomes arbitrarily close to 1.

And, last, while not about design *per se* but rather about protocol, an issue about the reporting and collection of results (rather than directly about the collection of data) is the unfortunate trend of not reporting negative results (Dowe 2008, sec. 0.2.5) and of only reporting positive results. There seems to be some sort of widespread—but fortunately not universal—implicit result-reporting protocol in some communities whereby negative results only get to be published when following on in response to a reported positive result. This can easily become data censoring of a primitive kind and can give rise to all sorts of bias. (Although it is not in a medical area, this view is presumably shared by the recently formed *Journal of Interesting Negative Results in Natural Language Processing*.)

Prediction

Recall the distinction(s) between inference and prediction in section "(Probabilistic) Prediction" (Wallace and Dowe 1999a, sec. 8; Wallace 2005, sec. 10.1.2). As in the AIC approach of Akaike (1970, 1973), the classical approach to prediction seems to be to find the single best theory—and use that for both inference and prediction. One could argue that it makes more sense intuitively—even in the classical (non-Bayesian) approach, such as Akaike's—to combine several theories which perform similarly (Wallace and Dowe 1999c, sec. 4). Empirical results would certainly suggest (Dowe, Gardner and Oppy 2007) that this is true in the case of AIC.

The Bayesian approach to prediction consists of taking every available theory in the parameter space and weighting it according to its posterior probability, and then using this to get a predictive distribution over expected future data. The resultant distribution will not necessarily be in the original parameter space—e.g. if our distribution is known to be $N(0, \sigma^2)$ and we consider all such possible distributions weighted by the posterior density of σ^2 (or, equivalently, of σ), the result will be an infinite mixture of Normal distributions.

Note that where the predictive distribution is not in the parameter space, the best fit to the predictive distribution from within the parameter space turns out to be the

MEKLD estimator (Dowe et al. 1998; Wallace 2005, secs 4.7–4.9; Dowe, Gardner and Oppy 2007, sec. 6.1.4) from section "Other: Other Classical, Other Bayesian, etc.". This makes sense, because MEKLD gives the best expected log-loss score (amongst hypotheses within the parameter space) and also because (recalling section "Probabilistic Inference—vs. Mere Non-probabilistic Classification") log-loss scoring rewards the true probabilities (if known) and appears to be unique in doing so (Dowe 2008, sec. 0.2.5).

Where there is one outstandingly good theory, then prediction and inference come very much to the same thing (Dowe 2008, sec. 0.3.1). However, they can vary when there is no outstandingly good theory, whereupon it is a good idea for predictive purposes to make a weighted combination of good inferences (Wallace and Dowe 1999a, sec. 8, 1999c, sec. 4), ideally—where feasible—weighting over the entire posterior.

Minimum Message Length (MML)

From Equation 2 and section "Bayes's Theorem and Bayesianism", we have that choosing H to maximise $Pr(H|D)$ is equivalent to choosing H to maximise $Pr(H) \cdot Pr(D|H)$. By the monotonicity of the logarithm function, this is equivalent to minimising $-\log(Pr(H) \cdot Pr(D|H)) = -\log Pr(H) - \log Pr(D|H)$. Simply changing notation, we can equivalently write

$$\begin{aligned} \arg\max_H Pr(H \mid D) &= \arg\max_H Pr(H) \cdot Pr(D \mid H) \\ &= \arg\min_H -\log Pr(H) - \log Pr(D \mid H). \end{aligned} \tag{3}$$

All data-sets—or at least all the ones I've used and/or heard of—are finite. This is partly so because, as mentioned at the start of section "Desiderata in (Probabilistic) Inference and (Probabilistic) Prediction", all heights and weights, etc. are measured to finite accuracy and finitely many decimal places (Wallace and Dowe 1993, 1–3, 1994, 38, secs 2 and 2.1, 2000, sec. 2, 74, col. 2; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651; Comley and Dowe 2003, sec. 9; Fitzgibbon, Dowe and Vahid 2004, eqn (19); Comley and Dowe 2005, sec. 11.3.3, 270; Wallace 2005, secs 3.1.1 and 3.3; Dowe, Gardner and Oppy 2007; Dowe 2008, sec. 0.2.4).

Given that heights, weights and other measurements are measured and recorded to finite accuracy, a fact often neglected by statisticians is that the likelihood, $Pr(D|H)$, can be viewed as a *probability* rather than as a *density* (of zero point mass). Let us clarify with an example. The probability of measuring a height of (say) 1.84 m is the probability that the height is between (say) 1.835 and 1.845 m. So, $Pr(D|H)$ is a probability (as measured), even if (in some sort of theory) perhaps a density. In the case that our parameter space is a continuum—such as the mean and variance, μ and σ^2, of heights—if we suitably quantise this down into at most countably many permissible (or usable) estimates, then $Pr(H)$ will also correspond to a probability and not a density (and we can even argue that the standard deviation, σ should be bounded below by a multiple of the measurement accuracy (Wallace and Dowe 1994, sec. 2.1; Dowe, Allison et al. 1996, sec. 2; Kissane, Bloch, Dowe et al. 1996, 651;

Comley and Dowe 2003, sec. 9; Comley and Dowe 2005, sec. 11.3.3; Dowe 2008, sec. 0.2.4)). As alluded to at the end of section "Maximum A Posteriori (MAP)", this will give us $Pr(H)$, $Pr(D|H)$ and (consequently) $Pr(H|D)$ all as probabilities—and *not* as densities.

Information Theory, Compression and MML

It should be clear to the reader from section "Maximum A Posteriori (MAP)" and/or from Dowe (2008, footnote 158 and sec. 0.2.3) that MML is, in general, different from MAP.

Let us now mention a result from information theory Shannon (1948) that an event e_i of probability p_i can be encoded in a prefix code by a code-word of length l_i where $-\log p_i \approx l_i < (-\log p_i) + 1$. This can be achieved with a Huffman code, which successively joins the two least probable events together and iterates. For details of code construction, see Wallace (2005, chap. 2, especially sec. 2.1), including the example in Wallace (2005, sec. 2.1.4 and fig. 2.5). As a simple example, if we have $\{p_1, p_2, p_3, p_4, p_5\} = \{1/2, 1/4, 1/16, 1/16, 1/8\}$ and we are considering binary codes, then a Huffman code would first join e_3 and e_4 (call this $e_{3,4}$) to give a probability of $1/16 + 1/16 = 1/8$, then next join $e_{3,4}$ and e_5 (call this $e_{3,4,5}$) to give a probability of $1/8 + 1/8 = 1/4$, then join e_2 to $e_{3,4,5}$ (call this $e_{2,3,4,5}$) to give a probability of $1/4 + 1/4 = 1/2$, and then finally join e_1 and $e_{2,3,4,5}$, resulting in a probability of $1/2 + 1/2 = 1$, whereupon it would stop. With "up" branches being given the bit 0 and "down" branches being given the bit 1, this would result in a binary code tree as in Figure 1 where e_1, e_2, e_3, e_4 and e_5 respectively have the code-words 0, 10, 1100, 1101 and 111. In this friendly example, we note that $l_i = -\log_2 p_i$ in each case.

Given that an event of probability p_i can be represented by a code-word of length l_i, looking at Equation 3, we see that $-\log Pr(H)$ can be regarded as the length of a message for encoding a hypothesis and $-\log Pr(D|H)$ can be regarded as the length of a message for encoding the data given this hypothesis. So, maximising the posterior *probability*, $Pr(H|D)$, is equivalent to minimising the length of a two-part

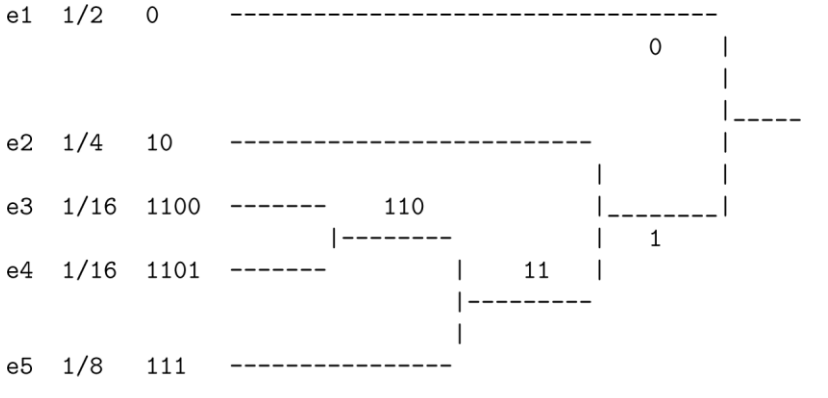

Figure 1 A simple Huffman code tree.

message, $-\log Pr(H)$ $-\log Pr(D|H)$, for jointly encoding the hypothesis and the observed data given this hypothesis. Hence the name minimum message length (MML). Given that MML is maximising a *probability* and *not* a density, and given the benefits of this (as per section "Properties of MML (and Approximations)"), MML can be thought of as MAP done properly (Wallace and Dowe 1999b, secs 1.2–1.3, 1999c, sec. 2, col. 1, 2000, secs 2 and 6.1; Comley and Dowe 2005, sec. 11.3.1; Dowe, Gardner and Oppy 2007, sec. 5.1, coding prior; Dowe 2008, footnote 158).

Philosophers wanting to know more about MML might wish to read Wallace (2005), Dowe, Gardner and Oppy (2007) and Dowe and Oppy (2001). Some of the many other articles of interest include Wallace and Boulton (1968), Comley and Dowe (2003, 2005), Wallace and Dowe (1999a) and Dowe (2008).

Ockham's Razor and MML

We recall the idea from Ockham's razor—or a common interpretation thereof—that if two theories fit the data equally well then one should prefer the simpler. Given MML's desire to quantitatively find (relatively) simple theories that fit the data (relatively) well, one can regard MML as being not only a quantitative version of Ockham's razor, but perhaps also a generalisation. Where Ockham's razor only seems to tell us which theory to prefer when both fit the data equally well, MML gives us a quantitative trade-off between simplicity and goodness of fit.

Figure 2 gives an example of two rival hypotheses, H_1 and H_2, for the data. We see that the encoding of Data given H_2 is shorter than that of Data given H_1, meaning that H_2 fits the data with a better log-likelihood than did H_1. However, we also see that the code length for H_1 is far shorter than that of H_2, meaning that H_1 is far more probable a priori (or simpler) than H_2. In this example, the shorter two-part message length is the explanation involving H_1, and so it would be the preferred MML inference. For further comments on MML and Ockham's razor, see, e.g. Needham and Dowe (2001), Comley and Dowe (2005, sec. 11.4.3) and Dowe (2008, footnotes 18 and 182).

Turing Machines, Algorithmic Information Theory and MML

A *Turing machine* (TM) (Wallace 2005, sec. 2.2.1) is an abstract mathematical model of a computer program. It can be written in a language from a certain alphabet of symbols (such as 1 and (blank) " "). We assume that Turing machines have a read/write

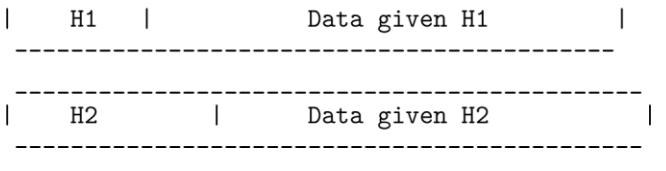

Figure 2 Two-part message lengths for two rival hypotheses for some Data.

head on an infinitely long tape. A Turing machine in a given state (with the read/write head) reading a certain symbol either moves to the left (L) or to the right (R) or stays where it is and writes a specified symbol. The instruction set for a Turing machine can be written as:

$$f: States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols)$$

Without loss of generality we can assume that the alphabet is the binary alphabet {0,1}, whereupon the instruction set for a Turing machine can be written as:

$$f: States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols).$$

Any known computer program can be represented by a Turing Machine. *Universal* Turing Machines (UTMs) are like compilers and can be made to emulate *any* Turing Machine (TM).

An example of a Turing machine would be a program which, for some a_0 and a_1, when given any input x, calculates (or outputs) $a_0 + a_1x$. In this case, x would input in binary (base 2), and the output would be the binary representation of $a_0 + a_1x$.

A *Universal Turing machine* (UTM) (Wallace, 2005, sec. 2.2.5) is a Turing machine which can simulate any other Turing machine. So, if U is a UTM and M is a TM, then there is some input c_M such that for any string s, $U(c_M s) = M(s)$ and the output from U when given the input $c_M s$ is identical to the output from M when given input s. In any other words, given any TM M, there is an emulation program (or code) c_M so that once U is input c_M it forever after behaves as though it were M.

The notion of *algorithmic information theory* (or *Kolmogorov complexity*) (Solomonoff 1964; Kolmogorov 1965; Chaitin 1966) of a string x is the length of the shortest input l_x to a UTM U such that $U(l_x) = x$—i.e. U will output x if given input l_x. Informally, if the length of l_x is the same as (or larger than) the length of x, then we can say that x is random in some sense. Similarly, if the length of l_x is much less than the length of x, then we can say that x is non-random.

Of these works from the mid-1960s, Kolmogorov (1965) and Chaitin (1966) study this important new concept while Solomonoff (1964) is also interested in using it for prediction. This work from the mid-1960s was very shortly before the first appearance of MML (Wallace and Boulton 1968). And just as MML can be regarded as a quantitative version of Ockham's razor (or as a generalisation thereof) as per section "Ockham's Razor and MML" and Figure 2, MML can also be regarded as (two-part) Kolmogorov complexity (Wallace and Dowe 1999a; Comley and Dowe 2005; sec. 11.4.3; Wallace 2005, chap. 2; Dowe 2008, secs 0.2.2, 0.2.7 and 0.3.1). Here, the first ("hypothesis") part of the message tells the Turing machine what hypothesis program is to be emulated but no output is written yet. The bit string in the second ("data given hypothesis") part of the message then causes the emulation program to output the original data string.

Note the analogy between MML as information theory in section "Information Theory, Compression and MML" and MML as *algorithmic* information theory in this section. And, very relatedly, note that the Kolmogorov complexity is dependent on the choice of UTM, and that this is a *Bayesian* choice (Wallace and Dowe 1999a, secs 2.4 and 7; Comley and Dowe 2005, sec 11.3.2; Dowe 2008, footnote 133).

Properties of MML (and Approximations)

The approach obtained from strictly minimising the message length as above is called Strict Minimum Message Length—or Strict MML, or SMML (Wallace and Boulton 1975; Wallace and Freeman 1987; Wallace and Dowe 1999a, sec. 6.1; Wallace 2005, chap. 3; Dowe, Gardner and Oppy 2007, sec. 5; Dowe 2008, footnotes 12, 153, 158 and 196 and sec. 0.2.2). Despite the many desirable properties of SMML (Wallace 2005, sec. 3.4), it can be computationally intractable even for relatively simple problems (Wallace 2005, sec. 3.2.9). (Historically, Strict MML (Wallace and Boulton 1975) came 7 years after MML (Wallace and Boulton 1968).) In practice, we often use a variety of applications, and these are also known to share many of the desirable properties of SMML.

Given data, D, the MMLD (or I_{1D}) approximation (Wallace 2005, secs 4.10 and 4.12.2; Dowe 2008, sec. 0.2.2) seeks a region R which minimises

$$-\log\left(\int_R h(\vec{\theta})\mathrm{d}\theta\right) \quad -\frac{\int_R h(\vec{\theta})\cdot \log f(\vec{D}\,|\,\vec{\theta})\mathrm{d}\theta}{\int_R h(\vec{\theta})\mathrm{d}\theta}. \tag{4}$$

The length of the first part is the negative log of the probability mass inside the region, R. The length of the second part is the (prior-weighted) average over the region R of the log-likelihood of the data, D.

An earlier approximation similar in motivation which actually inspired MMLD is the Wallace-Freeman approximation (Wallace and Dowe 1999a, sec. 6.1.2; Wallace 2005, chap. 5),

$$-log\left(h(\vec{\theta})\cdot\frac{1}{\sqrt{\kappa_d^d\ \mathrm{Fisher}\,(\vec{\theta})}}\right) \quad -log\, f(\vec{x}\,|\,\vec{\theta})+\frac{d}{2}, \tag{5}$$

which was first published in the statistics literature (Wallace and Freeman 1987).

The term $1/\sqrt{\kappa_d^d\ \mathrm{Fisher}\,(\vec{\theta})}$ gives a measure of uncertainty or quantisation in hypothesis space, where d is the number of continuous-valued parameters, κ_d is a constant (Fitzgibbon, Dowe and Vahid 2004, 441; Wallace 2005, table 3.4) between $1/12$ and $1/(2\pi e)$ and the Fisher information, $\mathrm{Fisher}\,(\vec{\theta})$, is as described in section "Bayesianism vs. Non-Bayesianism". (More specifically, in maths-speak, κ_d corresponds to the geometry of the optimally tesselating—or tiling—Voronoi region in d dimensions. In plainspeak, circles are compact but don't tile because they leave gaps, squares tile the plane, but hexagons tile optimally. $\kappa_2 = 5/(36\sqrt{3})$ corresponds to the geometry of a hexagon.) The term $d/2$ is the round-off in the second part of the message due to the uncertainty in the parameter estimate.

Perhaps the first thing to mention about Strict MML is its generality (Wallace 2005, sec. 3.4.3), that it is always defined—as likewise is Kolmogorov complexity. Strict MML, Wallace–Freeman and MMLD are all statistically invariant (Wallace 2005), as also are the estimators from Dowe (2008, sec. 0.2.2, footnotes 64 and 65) alluded to near the end of the section "Probabilistic Inference—vs. Mere Non-probabilistic Classification". Various theoretical results exist about the statistical consistency and efficiency of Strict MML (Wallace and Freeman 1987, 241; Barron and Cover 1991; Wallace 1996, 2005, sec. 3.4.5; Dowe, Gardner and Oppy 2007, sec. 5.3.4), and specific examples demonstrate the statistical consistency of Wallace–Freeman (Dowe and Wallace 1997) and similar approximations (Wallace and Freeman 1987; Wallace 1995). Many papers (e.g. Wallace and Freeman 1992; Wallace and Dowe 1993, 1999a, sec. 9, 1999b; Wallace 1995; Dowe, Oliver and Wallace 1996; Fitzgibbon, Dowe and Vahid 2004; Tan and Dowe 2002, 2003, 2004; Dowe, Gardner and Oppy 2007) attest to excellent small-sample performances of the Wallace–Freeman (or similar) approximation.

Another, possibly prophetic, thing to mention is that Strict MML first appeared in 1975 (Wallace and Boulton 1975) and the approximation from Equation 5 with the lattice constants (κ_d) first appeared in 1987 (Wallace and Freeman 1987), where κ_2 corresponds to the hexagon. When the trinomial (or 3-state multinomial) distribution—which has $d = 2$, as the parameters are p_1 and p_2 (because $p_3 = 1 - p_1 - p_2$)—was first done using Strict MML well over a decade later, with a uniform prior and $N = 60$ data-points, the partition (of the triangle) for the trinomial distribution turned out to contain an absolute abundance of hexagons (Wallace 2005, fig. 3.1, 166).

Problems with Increasing Numbers of Parameters

Consider the univariate polynomial regression problem of Dowe, Gardner and Oppy (2007, sec. 6.2) and Dowe (2008, sec. 0.2.3). Given data $(x, y)_{j=1,\ldots,N}$, we seek d, a_0, …, a_d, σ^2 such that $y = (\sum_{i=0}^{d} a_i x^i) + N(0,\sigma^2)$. This is a problem of nested models (or sub-families) (Dowe, Gardner and Oppy, 2007, sec. 7.1), in that (e.g.) every quadratic is also a cubic.

Studies (Wallace 1997; Dowe 2008, ref. 281; Dowe, Gardner and Oppy 2007, sec. 6.2.1; Dowe 2008, ref. 281) show that the classical Maximum Likelihood and AIC methods from sections "Maximum Likelihood" and "Akaike's Information Criterion (AIC) and Penalised (Maximum) Likelihood" over-fit, over-estimating the model order and (as in the section "Maximum Likelihood") under-estimating the variance, σ^2. MML gets the model order correct more often, sometimes under-estimating it (Dowe, 2008, footnote 153) and certainly getting a smaller squared error in prediction.

A different problem with nested models is that of econometric autoregressive time series. Models with terms from only the recent past are a special case of models including all of these terms and terms from the more distant past. Studies (Fitzgibbon, Dowe and Vahid 2004; Dowe, Gardner and Oppy 2007, sec. 6.2.2) similarly show the classical Maximum Likelihood and AIC methods over-fitting, and MML managing to give better predictions using a lower model order.

Amount of Data per Parameter Bounded Above

In the classic Neyman–Scott problem (Neyman and Scott 1948; Dowe and Wallace 1997; Wallace 2005, secs 4.2–4.5; Dowe, Gardner and Oppy 2007, sec. 6.1; Dowe 2008, secs 0.2.5 and 0.2.3), we measure N people's heights J times each (say $J = 2$) and then infer

1. the heights $\mu_1, \ldots, \mu_N$ of each of the N people,
2. the accuracy (σ) of the measuring instrument.

We have JN measurements from which we need to estimate $N + 1$ parameters. $JN/(N + 1) \leq J$, so the amount of data per parameter is bounded above (by J), the notion of which we flagged in section "Statistical Consistency"

$$\hat{\sigma}^2_{\text{Maximum Likelihood}} \rightarrow \frac{J-1}{J}\sigma^2,$$

and so for fixed J as $N \rightarrow \infty$ we have that Maximum Likelihood is statistically inconsistent—under-estimating σ and "finding" patterns that aren't there. As alluded to in the section "Properties of MML (and Approximations)", MML remains statistically consistent for the Neyman–Scott problem (Dowe and Wallace 1997).

What makes the Neyman–Scott problem difficult is that, even though the amount of data is increasing unboundedly, the amount of data *per parameter* is bounded above. This is sufficient to preserve the small sample bias from section "Maximum Likelihood". This is somewhat awful for Maximum Likelihood and Akaike's Information Criterion (AIC).

Other examples of the amount of data being bounded above include

- latent factor analysis—single (Wallace and Freeman 1992; Edwards and Dowe 1998) and multiple (Wallace 1995, 2005, sec. 6.9; Dowe, Gardner and Oppy 2007, sec. 6.1.3; Dowe 2008, sec. 0.2.3), and
- fully-parameterised mixture modelling (Wallace and Dowe 2000, sec. 4.3; Wallace 2005, sec. 6.8; Dowe, Gardner and Oppy 2007, sec. 6.1.3; Dowe 2008, sec. 0.2.5).

These problems are more commonplace than one might at first realise. The factors from latent factor analysis correspond to notions like I.Q. or octane rating. More specifically, if we get N people to sit J aptitude tests or if we test N petrols on J engines, then what we wish to infer are statistical factors—such as I.Q. and octane rating. These I.Q.s (for each of the N people in turn) and the octane ratings (for each of the N petrols in turn) are known as the factor scores. But we also need to estimate the factor loads, or the load vector. This basically tells us how important, relevant or otherwise—and, if relevant, how difficult/easy—each aptitude test or engine test is. Both Maximum Likelihood and AIC again struggle in such cases, with Akaike (1987) himself adopting a Bayesian prior—actually, a "prior" which changes as the sample size changes (Akaike 1987, sec. 5, 325; Dowe, Gardner and Oppy 2007, sec. 6.1.3 and footnote 22)—for latent factor analysis. Empirical studies (Wallace and Freeman 1992; Wallace 1995) again show MML outperforming these methods—even when they have been helped

out with Bayesian priors—and doing so with simpler models. For these types of problems (with data per parameter bounded above), classical methods often appeal to Bayesianism for help (Wallace 2005, sec. 4.5).

By acknowledging *uncertainty* (or quantising) when doing parameter estimation, MML is statistically consistent on all of these problems. MML is about *inference*, seeking the *truth* (Dowe 2008, secs 0.2.4 and 0.2.6). (Indeed, Steven L. Gardner would like to relate MML to the notion in philosophy of *approximate truth*.) It gives a statistically invariant—and statistically consistent—Bayesian method of point estimation. It gives general consistency results where classical non-Bayesian approaches are known to break down. It is also efficient, working well on all real inference problems currently known to the author.

The above evidence and experience has led to the following two conjectures.

Conjecture 1 (Dowe et al. 1998, 93; Edwards and Dowe 1998, sec. 5.3; Wallace and Dowe 1999a, 282, 2000, sec. 5; Comley and Dowe 2005, sec. 11.3.1, 269) Only MML and very closely-related Bayesian methods are in general both statistically consistent and invariant.

Conjecture 2 (*Back-up Conjecture*) (Dowe, Gardner and Oppy 2007, sec. 8; Dowe 2008, sec. 0.2.5) If there are any such non-Bayesian methods, they will be far less efficient than MML.

Before proceeding to a final discussion and conclusion, it seems appropriate to first mention some medical and humanities applications of MML.

Medical, Biological and Other Applications of MML

Some Medical-related Applications of MML

The second application ever of MML to real-world data was in a classification of depression (Pilowsky, Levine and Boulton 1969). Studies classifying and clustering grieving families include (Kissane, Bloch, Dowe et al. 1996; Kissane, Bloch, Onghena et al. 1996), with a classification of sub-groups within autism given in Prior et al. (1998) and a classification of distress syndromes in Clarke et al. (2003). Another study of a diagnostic nature was McKenzie et al. (1993).

A fairly routine application of some MML clustering software (Edgoose, Allison and Dowe 1998, sec. 6; Dowe, Allison et al. 1996, sec. 5, 253; Wallace 1998, sec. 4.2; Dowe 2008, footnote 85) gave that proteins apparently fold with the Helices (and Extendeds) forming first and then the "Other" turn classes forming subsequently to accommodate these structures. Some further applications of MML clustering are cited in Wallace and Dowe (1994) (and Dowe 2008).

DNA microarray data (Tan, Dowe and Dix 2007) can be studied by MML (Dowe 2008, sec. 0.2.7, footnote 196), and MML image analysis (Wallace 1998; Visser and Dowe 2007) is also ripe for medical applications.

And, although it doesn't just apply to medical data, if we take a noise-free unstructured, unnormalised database of sufficient size and then apply MML Bayesian nets

(Comley and Dowe 2003, 2005; Dowe 2008, sec. 0.2.5) (from section "Kullback–Leibler Divergence (or Kullback–Leibler Distance)") to this, we get the elegant result that the MML Bayesian net inference will result in a normalised database (Dowe 2008, sec. 0.2.6, footnote 187). If there is sufficient data, this will be a full normalisation.

Some Applications of MML in the Humanities

Many applications of MML to real-world data-sets and a variety of subject areas exist—see, e.g. Wallace (2005), Dowe (2008) and elsewhere. For the curious reader, I'd like to give some admittedly all too brief pointers to reading on MML in philosophy and humanities. These include

- MML and an argument that—contrary to widely-held views in physics, philosophy and many fields—entropy is *not* time's arrow (Wallace 2005, chap. 8; Dowe 2008, sec. 0.2.5),
- MML, existence of "miracles" (Dowe 2008, sec. 0.2.7), cosmological arguments and "Intelligent Design" (I.D.),
- MML and linguistics—inferring "dead" languages (Ooi and Dowe 2005; Dowe 2008, sec. 0.2.4),
- MML, Kolmogorov complexity (Wallace and Dowe 1999a), measures of "intelligence" (Dowe and Hajek 1998; Hernandez-Orallo 2000; Legg and Hutter 2007; Dowe 2008, sec. 0.2.5) and a possible variation on the (so-called) Lucas–Penrose argument in the philosophy of mind that humans are (supposedly) more intelligent than machines can be (Dowe 2008, footnotes 70–71 and sec. 0.2.3),
- MML and the Efficient Markets Hypothesis, in which appeals to the relationship between MML and Kolmogorov complexity (as per section "Turing Machines, Algorithmic Information Theory and MML") tell us that markets are *not* provably efficient (Dowe and Korb 1996; Wallace 2005, sec. 9.1, 387; Dowe 2008, sec. 0.2.5), and
- varying the *elusive model paradox* (Dowe 2008, footnote 211) so that each bit (0 or 1) in a sequence of bits is to be the bit which was *not* predicted to be the (most probable) next bit in the sequence. (Recalling sections "(Probabilistic) Prediction" and "Prediction", we can do this by inferring a model from the past bits—as per the original elusive model paradox—or by combining several models and predicting.) For mathematicians and computer scientists, this gives us a new non-computable number. In terms of game theory, psychology, sociology and making of "thriller" movies with lots of plot "twists and turns", this relates to leaving a trail so that those trying to follow you will (almost) always—or at least as often as possible—be surprised by your next step.

Discussion and Conclusion

When seeking to draw conclusions from medical (or other) data evidence, sometimes the problem is particularly friendly—there are few parameters to be estimated, there is

an abundance of data and there is relatively little noise in the data. Inference here should be sound, the best predictor should be close to the derived inference, and the reported power of hypothesis tests should not be unreasonable.

But in the not uncommon case that the amount of data per parameter is limited, great care should be taken. Not only do the classical approaches to inference start to differ, but the ones in common usage at the time of writing tend to over-estimate the relevance of the explanatory variables and under-estimate the noise. Answers obtained by classical methods will typically improve when replaced by the Bayesian MML approach, and this seems to hold regardless of sample size. Here, care should be taken in the choice of a Bayesian prior, lest one be accused of fudging one's results.

The seeming objectivity of the classical approach versus the seeming more reliable results from the Bayesian MML approach leaves us with something of a quandary. As a *first recommendation*, at the very least, the data analyst should be aware of these issues and should ideally at least mention something along these general lines when publishing. As a *second recommendation*, if one is using a classical approach, then—where possible—the data should also be analysed in a Bayesian (MML) way alongside whatever classical analysis is chosen. If one is understandably concerned about (e.g.) how one's peers might regard a Bayesian analysis, one can repeat the study with a different set of Bayesian priors. One can then report discrepancies and—one hopes—similarities amongst different approaches. As a *third recommendation*, probabilistic predictions should be made and—given the apparent uniqueness result(s) (from the section "Probabilistic Inference—vs. Mere Non-probabilistic Classification" and Dowe (2008, footnote 175))—should probably be scored with log-loss.

Acknowledgements

I thank Chris Wallace (1933–2004) (Dowe 2008) for giving me a proper appreciation of Bayesianism and for introducing me to—and training me in—Minimum Message Length (Wallace and Boulton 1968; Wallace 2005). I thank Dr Jason Grossman for alerting me to the (subtle) distinction in the section "Statistical Consistency" between statistical consistency (or even efficiency) and small-sample performance, and also for discussion motivating the section "Experimental Design, Data Collection Protocol and Likelihood Principle". As in Dowe (2008, footnote 116), I am grateful to Claire Leslie for telling me about the *bus number problem* from the section "Maximum Likelihood". And I thank those who anonymously reviewed the paper for useful guidance in helping me make the paper clearer and better.

Notes

[1] If $\tan^{-1}$ ranges from $-\pi/2$ to $\pi/2$, then we take the negative square root for $x < 0$. We can write this more properly as $(\kappa, \theta) = (\text{sign}(x) \cdot \sqrt{x^2 + y^2},\ \tan^{-1}(y/x))$.

References

Akaike, H. 1970. Statistical prediction information. *Annals of the Institute of Statistical Mathematics* 22: 203–17.

———. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd international symposium on information theory,* edited by B.N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

———. 1987. Factor Analysis and AIC. *Psychometrika* 52 (3): 317–332.

Barron, A. R., and T. M. Cover. 1991. Minimum complexity density estimation. *IEEE Transactions on Information Theory* 37: 1034–54.

Berger, J. O., and R. L. Wolpert. 1988. *The likelihood principle,* 2nd edition, Institute of Mathematical Statistics monograph series. California, USA: Hayward.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory.* New York: Wiley.

Chaitin, G. J. 1966. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery* 13: 547–69.

Clarke, D. M., G. C. Smith, D. L. Dowe, and D. P. McKenzie. 2003. An empirically-derived taxonomy of common distress syndromes in the medically ill. *Journal of Psychosomatic Research* 54: 323–30.

Comley, Joshua W., and David L. Dowe. 2003. General Bayesian networks and asymmetric languages. Paper presented at Proceedings of the Hawaii International Conference on Statistics and Related Fields, 5–8 June.

———. 2005. Minimum message length and generalized Bayesian nets with asymmetric languages. Chap. 11 in *Advances in minimum description length: Theory and applications (MDL handbook),* edited by P. Grünwald, M. A. Pitt, and I. J. Myung, pp. 265–94. Cambridge, MA: MIT Press.

Dowe, D. L. 2007. Discussion following "Hedging predictions in machine learning, A. Gammerman and V. Vovk". *Computer Journal* 2 (50): 167–8.

———. 2008. Foreword re C. S. Wallace. *Computer Journal* 51 (5): 523–560.

Dowe, D. L., L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. 1996. Circular clustering of protein dihedral angles by minimum message length. In *Pacific symposium on biocomputing '96,* edited by L. Hunter and T. Klein, pp. 242–55. Singapore: World Scientific.

Dowe, D. L., R. A. Baxter, J. J. Oliver, and C. S. Wallace. 1998. Point estimation using the Kullback–Leibler loss function and MML. In *Proceedings of the 2nd Pacific-Asia conference on research and development in knowledge discovery and data mining (PAKDD-98),* Volume 1394 of *LNAI,* edited by X. Wu, Ramamohanarao Kotagiri, and K. Korb, pp. 87–95. Berlin: Springer.

Dowe, D. L., G. E. Farr, A. J. Hurst, and K. L. Lentin. 1996. Information-theoretic football tipping. Paper presented at the 3rd Conference on Maths and Computers in Sport, pp. 233–41. [See also Technical Report TR 96/297, Dept. Computer Science, Monash University, Australia 3168, Dec 1996.]

Dowe, D. L., S. Gardner, and G. R. Oppy. 2007. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science* 58 (4): 709–54.

Dowe, D. L., and A. R. Hajek. 1998. A non-behavioural, computational extension to the Turing test. Paper presented at the International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98), Gippsland, Australia, February, pp. 101–6.

Dowe, D. L., and K. B. Korb. 1996. Conceptual difficulties with the efficient market hypothesis: Towards a naturalized economics. Paper presented at the Proceedings on Information, Statistics and Induction in Science (ISIS), pp. 212–23. [See also Technical Report TR 94/215, Dept. Computer Science, Monash University, Australia 3168, 1994.]

Dowe, D. L., and N. Krusel. 1993. *A decision tree model of bushfire activity.* Technical report TR 93/190, Dept. of Computer Science, Monash University, Clayton, Vic. 3800, Australia, September.

Dowe, D. L., J. J. Oliver, and C. S. Wallace. 1996. MML estimation of the parameters of the spherical Fisher distribution. In *Algorithmic learning theory, 7th international workshop, ALT '96,*

Sydney, Australia, October 1996, proceedings, Volume 1160 of *Lecture notes in artificial intelligence,* edited by S. Arikawa and A. Sharma, pp. 213–227. Berlin: Springer.

Dowe, D. L., and G. R. Oppy. 2001. Universal Bayesian inference? *Behavioral and Brain Sciences (BBS)* 24 (4): 662–3.

Dowe, D. L., and C. S. Wallace. 1997. Resolving the Neyman–Scott problem by Minimum Message Length. In *Proceedings of computing science and statistics – 28th symposium on the interface,* Volume 28, edited by L. Billard and N. I. Fisher, pp. 614–18. Interface Foundation of North America.

Edgoose, T., L. Allison, and D. L. Dowe. 1998. An MML classification of protein structure that knows about angles and sequence. In *Pacific symposium on biocomputing '98,* edited by R. B. Altman, A. K. Dunker, L. Hunter, and T. Klein, pp. 585–96. Singapore: World Scientific.

Edwards, R. T., and D. L. Dowe. 1998. Single factor analysis in MML mixture modelling. In *Proceedings of the 2nd Pacific-Asia conference on research and development in knowledge discovery and data mining (PAKDD-98),* Volume 1394 of *Lecture notes in artificial intelligence (LNAI),* edited by Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, pp. 96–109. Berlin: Springer.

Fitzgibbon, L. J., D. L. Dowe, and F. Vahid. 2004. Minimum message length autoregressive model order selection. Paper presented at the Proceedings of the International Conference on Intelligent Sensors and Information Processing, Chennai, India, January, pp. 439–44. IEEE (IEEE Press).

Forster, M., and E. Sober. 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: 1–35.

Glymour, C. 1981. Why I am not a Bayesian. *Theory and Evidence,* edited by C. Glymour and D. Stalker, pp. 63–93. Princeton: Princeton University Press.

Grossman, J. Forthcoming. The likelihood principle. In *Handbook for philosophy of science,* Volume 7, *Philosophy of statistics.* New York: Elsevier.

Grünwald, Peter D., and John Langford. 2007. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning* 66 (31): 119–149.

Hernández-Orallo, José. 2000. Beyond the Turing test. *Journal of Logic, Language and Information* 9 (4): 447–66.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A* 186: 453–4.

Kissane, D. W., S. Bloch, D. L. Dowe, R. D. Snyder, P. Onghena, D. P. McKenzie, and C. S. Wallace. 1996. The Melbourne family grief study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry* 153: 650–8.

Kissane, D. W., S. Bloch, P. Onghena, D. P. McKenzie, R. D. Snyder, and D. L. Dowe. 1996. The Melbourne family grief study, II: Psychosocial morbidity and grief in bereaved families. *American Journal of Psychiatry* 153: 659–66.

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

Kornienko, L., D. W. Albrecht, and D. L. Dowe. 2005a. A preliminary MML linear classifier using principal components for multiple classes. In *Proceedings of the 18th Australian joint conference on artificial intelligence (AI'2005),* Volume 3809 of *Lecture notes in artificial intelligence (LNAI),* Sydney, Australia, edited by S. Zhang, and Ray Jarvis, pp. 922–6. Berlin: Springer.

———. 2005b. *A preliminary MML linear classifier using principal components for multiple classes.* Technical report CS 2005/179, School of Computer Sci. & Softw. Eng., Monash Univ., Melb., Australia.

Kornienko, Lara, David L. Dowe, and David W. Albrecht. 2002. Message length formulation of support vector machines for binary classification – A preliminary scheme. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence,* Volume 2557 of *Lecture notes in artificial intelligence (LNAI),* edited by B. McKay, and J. K. Slaney, pp. 119–130. Berlin: Springer-Verlag.

Legg, S., and M. Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17 (4): 391–444.

McKenzie, D. P., P. D. McGorry, C. S. Wallace, L. H. Low, D. L. Copolov, and B. S. Singh. 1993. Constructing a minimal diagnostic decision tree. *Methods in Information in Medicine* 32: 161–6.

Needham, S. L., and D. L. Dowe. 2001. Message length as an effective Ockham's razor in decision tree induction. Paper presented at the 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001), pp. 253–60.

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

Ooi, J. N., and D. L. Dowe. 2005. Inferring phylogenetic graphs of natural languages using minimum message length. Paper presented at CAEPIA 2005 (11th Conference of the Spanish Association for Artificial Intelligence), Volume 1, pp. 143–52.

Pilowsky, I., S. Levine, and D.M. Boulton. 1969. The classification of depression by numerical taxonomy. *British Journal of Psychiatry* 115: 937–45.

Prior, M., R. Eisenmajer, S. Leekam, L. Wing, J. Gould, B. Ong, and D. L. Dowe. 1998. Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *Journal of Child Psycholology and Psychiatry* 39 (6): 893–902.

Rissanen, J. J. 1978. Modeling by shortest data description. *Automatica* 14: 465–71.

Schwarz, G. 1978. Estimating dimension of a model. *Annals of Statistics* 6: 461–4.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423 and 623–56.

Solomonoff, R. J. 1964. A formal theory of inductive inference. *Information and Control* 7: 1–22, 224–54.

Tan, P. J., and D. L. Dowe. 2002. MML inference of decision graphs with multi-way joins. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence,* Volume 2557 of *Lecture notes in artificial intelligence (LNAI),* edited by R. McKay and J. Slaney, pp. 131–42. Berlin: Springer Verlag.

———. 2003. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence,* Volume 2903 of Lecture Notes in Artificial Intelligence (*LNAI*), edited by T. D. Gedeon, and L. Chun Che Fung, pp. 269–81. Berlin: Springer.

———. 2004. MML inference of oblique decision trees. In *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence,* Volume 3339 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by G. I. Webb, and Xinghuo Yu, pp. 1082–8. Berlin: Springer.

———. 2006. Decision forests with oblique decision trees. In *Proceedings of the 5th Mexican international conference on artificial intelligence,* Volume 4293 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by A. F. Gelbukh, and C. A. Reyes García, pp. 593–603. Berlin: Springer.

Tan, P. J., D. L. Dowe, and T. I. Dix. 2007. Building classification models from microarray data with tree-based classification algorithms. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence,* Volume 4830 of *Lecture Notes in Artificial Intelligence (LNAI),* edited by M. A. Orgun, and J. Thornton, pp. 589–98. Berlin: Springer.

Vapnik, V. N. 1995. *The nature of statistical learning theory.* Berlin: Springer.

Visser, Gerhard, and D. L. Dowe. 2007. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In *Proceedings of the 6th IEEE international conference on computer and information science (ICIS) 2007,* pp. 17–22. Piscataway, NJ: IEEE Press.

Wallace, C. S. 1995. *Multiple factor analysis by MML estimation.* Technical report CS TR 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Clayton, Melbourne, Australia.

———. 1996. False oracles and SMML estimators. In *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference,* edited by D. L. Dowe, K. B. Korb, and J. J Oliver,

pp. 304–316. Singapore: World Scientific. [Was previously Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia, June 1989.]

———. 1997. *On the selection of the order of a polynomial model.* Technical report, Royal Holloway College, England, UK. Chris released this in 1997 (from Royal Holloway) in the belief that it would become a Royal Holloway Tech Rept dated 1997, but it is not clear that it was ever released there. Soft copy certainly does exist, though. Perhaps see www.csse.monash.edu.au/~dld/CSWallacePublications; INTERNET.

———. 1998. Intrinsic classification of spatially correlated data. *Computer Journal* 41 (8): 602–611.

———. 2005. *Statistical and inductive inference by minimum message length.* Information Science and Statistics series. Berlin: Springer Verlag.

Wallace, C. S., and D. M. Boulton. 1968. An information measure for classification. *Computer Journal* 11 (2): 185–94.

———. 1975. An invariant Bayes method for point estimation. *Classification Society Bulletin* 3 (3): 11–34.

Wallace, C. S., and D. L. Dowe. 1993. *MML estimation of the von Mises concentration parameter.* Technical Report 93/193, Dept. of Computer Science, Monash University, Clayton 3168, Australia, December.

———. 1994. Intrinsic classification by MML – the Snob program. In *Proceedings of the 7th Australian Joint Conference on Artificial Intelligence,* edited by C. Zhang, J. Debenham and D. Lukose pp. 37–44. Singapore: World Scientific.

———. 1999a. Minimum message length and Kolmogorov complexity. *Computer Journal* 42 (4): 270–283.

———. 1999b. Refinements of MDL and MML coding. *Computer Journal* 42 (4): 330–7.

———. 1999c. Rejoinder. *Computer Journal* 42 (4): 345–7.

———. 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing* 10 (1): 73–83.

Wallace, C. S., and P. R. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B* 49 (3): 240–52. See also Discussion on pp. 252–65.

———. 1992. Single-factor analysis by minimum message length estimation. *Journal of the Royal Statistical Society B* 54 (1): 195–209.