MML, HYBRID BAYESIAN NETWORK GRAPHICAL MODELS, STATISTICAL CONSISTENCY, INVARIANCE AND UNIQUENESS

David L. Dowe

1 INTRODUCTION

The problem of statistical — or inductive — inference pervades a large number of human activities and a large number of (human and non-human) actions requiring 'intelligence'. Human and other 'intelligent' activity often entails making inductive inferences, remembering and recording observations from which one can make inductive inferences, learning (or being taught) the inductive inferences of others, and acting upon these inductive inferences.

The Minimum Message Length (MML) approach to machine learning (within artificial intelligence) and statistical (or inductive) inference gives us a trade-off between simplicity of hypothesis (H) and goodness of fit to the data (D) [Wallace and Boulton, 1968, p. 185, sec 2; Boulton and Wallace, 1969; 1970, p. 64, col 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec 1 col 1; Wallace and Boulton, 1975, sec. 3; Boulton, 1975; Wallace and Georgeff, 1983; Wallace and Freeman, 1987; Wallace and Dowe, 1999a; Wallace, 2005; Comley and Dowe, 2005, secs. 11.1 and 11.4.1; Dowe, 2008a, sec 0.2.4, p. 535, col. 1 and elsewhere]. There are several different and intuitively appealing ways of thinking of MML. One such way is to note that files with structure compress (if our file compression program is able to find said structure) and that files without structure don't compress. The more structure (that the compression program can find), the more the file will compress.

Another, second, way to think of MML is in terms of Bayesian probability, where $Pr(H)$ is the prior probability of a hypothesis, $Pr(D|H)$ is the (statistical) likelihood of the data D given hypothesis H, $-\log Pr(D|H)$ is the (negative) log-likelihood, $Pr(H|D)$ is the posterior probability of H given D, and $Pr(D)$ is the marginal probability of D — i.e., the probability that D will be generated (regardless of whatever the hypothesis might have been). Applying Bayes's theorem twice, with or without the help of a Venn diagram, we have $Pr(H|D) = Pr(H\&D)/Pr(D) = (1/Pr(D))\ Pr(H)Pr(D|H)$.

Choosing the most probable hypothesis (a posteriori) is choosing H so as to maximise $Pr(H|D)$. Given that $Pr(D)$ and $1/Pr(D)$ are independent of the choice of

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

hypothesis H, this is equivalent to choosing H to maximise $Pr(H)$. $Pr(D|H)$. By the monotonicity of the logarithm function, this is in turn equivalent to choosing H so as to minimise $-\log Pr(H) - \log Pr(D|H)$. From Shannon's information theory (see sec. 2.1), this is the amount of information required to encode H (in the first part of a two-part message) and then encode D given H (in the second part of the message). And this is, in turn, similar to our first way above of thinking about MML, where we seek H so as to give the optimal two-part file compression.

We have shown that, given data D, we can variously think of the MML hypothesis H in at least two different ways: (a) as the hypothesis of highest posterior probability and also (b) as the hypothesis giving the two-part message of minimum length for encoding H followed by D given H; and hence the name Minimum Message Length (MML).

Historically, the seminal Wallace and Boulton paper [1968] came into being from Wallace's and Boulton's finding that the Bayesian position that Wallace advocated and the information-theoretic (conciseness) position that Boulton advocated turned out to be equivalent [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546 and footnote 213]. After several more MML writings [Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1] (and an application paper [Pilowsky *et al.*, 1969], and at about the same time as David Boulton's PhD thesis [Boulton, 1975]), their paper [Wallace and Boulton, 1975, sec. 3] again emphasises the equivalence of the probabilistic and information-theoretic approaches. (And all of this work on Minimum Message Length (MML) occurred prior to the later Minimum Description Length (MDL) principle discussed in sec. 6.7 and first published in 1978 [Rissanen, 1978].)

A third way to think about MML is in terms of *algorithmic* information theory (or Kolmogorov complexity), the shortest input to a (Universal) Turing Machine [(U)TM] or computer program which will yield the original data string, D. This relationship between MML and Kolmogorov complexity is formally described — alongside the other two ways above of thinking of MML (probability on the one hand and information theory or concise representation on the other) — in [Wallace and Dowe, 1999a]. In short, the first part of the message encodes H and causes the TM or computer program to read (without yet writing) and prepare to output data, emulating as though it were generated from this hypothesis. The second part of the input then causes the (resultant emulation) program to write the data, D.

So, in sum, there are (at least) three equivalent ways of regarding the MML hypothesis. It variously gives us: (i) the best two-part compression (thus best capturing the structure), (ii) the most probable hypothesis (a posteriori, after we've seen the data), and (iii) an optimal trade-off between structural complexity and noise — with the first-part of the message capturing all of the structure (no more, no less) and the second part of the message then encoding the noise.

Theorems from [Barron and Cover, 1991] and arguments from [Wallace and Freeman, 1987, p241] and [Wallace, 2005, chap. 3.4.5, pp. 190-191] attest to the

general optimality of this two-part MML inference — converging to the correct answer as efficiently as possible. This result appears to generalise to the case of model misspecification, where the model generating the data (if there is one) is not in the family of models that we are considering [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5]. In practice, we find that MML is quite conservative in variable selection, typically choosing less complex models than rival methods [Wallace, 1997; Fitzgibbon *et al.*, 2004; Dowe, 2008a, footnote 153, footnote 55 and near end of footnote 135] while also appearing to typically be better predictively.

Having introduced Minimum Message Length (MML), throughout the rest of this chapter, we proceed initially as follows. First, we introduce information theory, Turing machines and algorithmic information theory — and we relate all of those to MML. We then move on to Ockham's razor and the distinction between inference (or induction, or explanation) and prediction. We then continue on to relate MML and its relevance to a myriad of other issues.

2 INFORMATION THEORY — AND VARIETIES THEREOF

2.1 Elementary information theory and Huffman codes

Tossing a fair unbiased coin n times has 2^n equiprobable outcomes of probability 2^{-n} each. So, intuitively, it requires n bits (or binary digits) of information to encode an event of probability 2^{-n}, so (letting $p = 2^{-n}$) an event of probability p contains $-\log_2 p$ bits of information. This results holds more generally for bases $k = 3, 4, ...$ other than 2.

The Huffman code construction (for base k), described in (e.g.) [Wallace, 2005, chap. 2, especially sec. 2.1; Dowe, 2008b, p. 448] and below ensures that the code length l_i for an event e_i of probability p_i satisfies $-\log_k p_i \approx l_i < -\log_k p_i + 1$.

Huffman code construction proceeds by taking m events $e_1, \ldots, e_m$ of probability $p_1, \ldots, p_m$ respectively and building a code tree by successively (iteratively) joining together the events of least probability. So, with $k = 2$, the binary code construction proceeds by joining together the two events of least probability (say e_i and e_j) and making a new event $e_{i,j}$ of probability $p_{i,j} = p_i + p_j$. (For a k-ary code construction of arity k, we join the k least probable events together — see, e.g., fig. 3, with arity $k = 3$. We address this point a little more later.) Having joined two events into one event, there is now 1 less event left. This iterates one step at a time until the tree is reduced to its root.

An example with $k = 2$ from [Dowe, 2008b, p. 448, Fig. 1] is given in Figure 1.

Of course, we can not always expect all probabilities to be of the form k^{-n}, as they are in the friendly introductory example of fig. 1.

One example with $k = 2$ (binary) and where the probabilities are not all some k raised to the power of a negative (or zero) integer is 1/21, 2/21, 3/21, 4/21, 5/21, 6/21, as per fig. 2, which we now examine.

Immediately below, we step through the stages of the binary Huffman code construction in fig. 2. The two events of smallest probability are e_1 and e_2 of

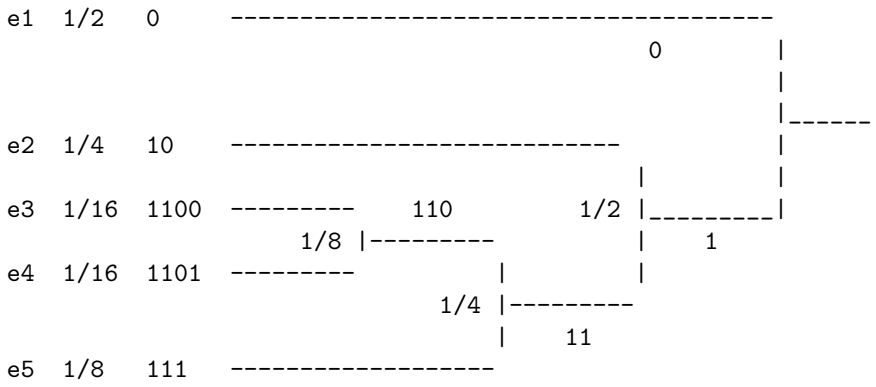

Figure 1. A simple (binary) Huffman code tree with k = 2

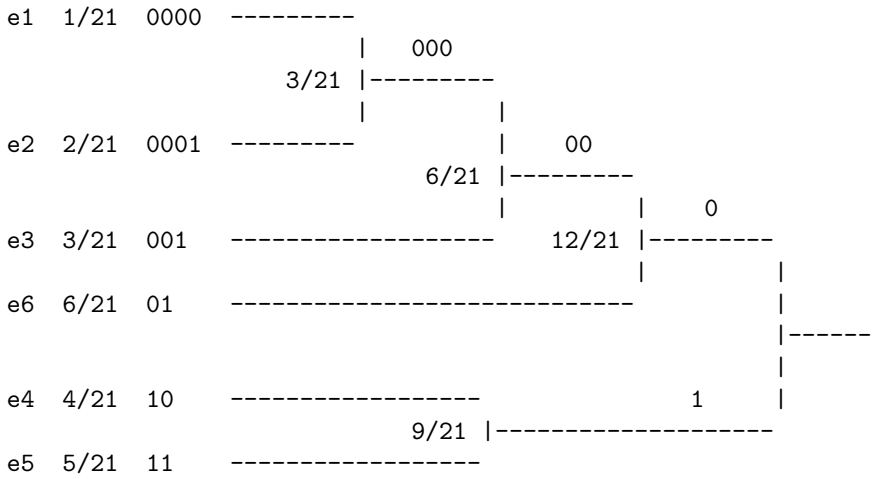

Figure 2. A not so simple (binary) Huffman code tree with k = 2

probabilities 1/21 and 2/21 respectively, so we join them together to form $e_{1,2}$ of probability 3/21. The two remaining events of least probability are now $e_{1,2}$ and e_3, so we join them together to give $e_{1,2,3}$ of probability 6/21. The two remaining events of least probability are now e_4 and e_5, so we join them together to give $e_{4,5}$ of probability 9/21. Three events now remain: $e_{1,2,3}$, $e_{4,5}$ and e_6. The two smallest probabilities are $p_{1,2,3} = 6/21$ and $p_6 = 6/21$, so they are joined to give $e_{1,2,3,6}$ with probability $p_{1,2,3,6} = 12/21$. For the final step, we then join $e_{4,5}$ and $e_{1,2,3,6}$. The code-words for the individual events are obtained by tracing a path from the root of the tree (at the right of the code-tree) left across to the relevant event at the leaf of the tree. For a binary tree ($k = 2$), every up branch is a 0 and every down branch is a 1. The final code-words are e_1: 0000, e_2: 0001, e_3: 001, etc. (For the reader curious as to why we re-ordered the events e_i, putting e_6 in the middle and not at an end, if we had not done this then some of the arcs of the Huffman code tree would cross — probably resulting in a less elegant and less clear figure.)

For another such example with $k = 2$ (binary) and where the probabilities are not all some k raised to the power of a negative (or zero) integer, see the example (with probabilities 1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 7/36, 8/36) from [Wallace, 2005, sec. 2.1.4, Figs. 2.5–2.6].

An example with $k = 3$ is given in Fig. 3. The Huffman code construction for $k = 3$ is very similar to that for $k = 2$, but it also sometimes has something of a small pre-processing step. Each step of the k-ary Huffman construction involves joining k events together, thus reducing the number of events by $(k - 1)$, which is equal to 2 for $k = 3$. So, if the number of events is even, our initial pre-processing step is to join the two least probable events together. That done, we now have an odd number of events and our code tree construction is, at each step, to join the three least probable remaining events together. We continue this until just one event is left, when we are left with just the root node and we have finished. The assignment of code-words is similar to the binary case, although the ternary construction (with $k = 3$) has 3-way branches. The top branch is 0, the middle branch is 1 and the bottom branch is 2. The reader is invited to construct and verify this code construction example (in fig. 3) and the earlier examples referred to above.

For higher values of k, the code construction joins k events (or nodes) into 1 node each time, reducing the number of nodes by $(k-1)$ at each step. If the number of nodes is $q(k-1)+1$ for some $q \geq 0$, then the Huffman code construction does not need a pre-processing step. Otherwise, if the number of nodes is $q(k-1)+1+r$ for some $q \geq 0$ and some r such that $1 \leq r \leq k-2$, then we require a pre-processing step of first joining together the $(r + 1)$ least probable events into one, reducing the number of nodes by r to $q(k - 1) + 1$.

The result mentioned earlier that

$$-\log_k p_i \approx l_i < -\log_k p_i + 1 \tag{1}$$

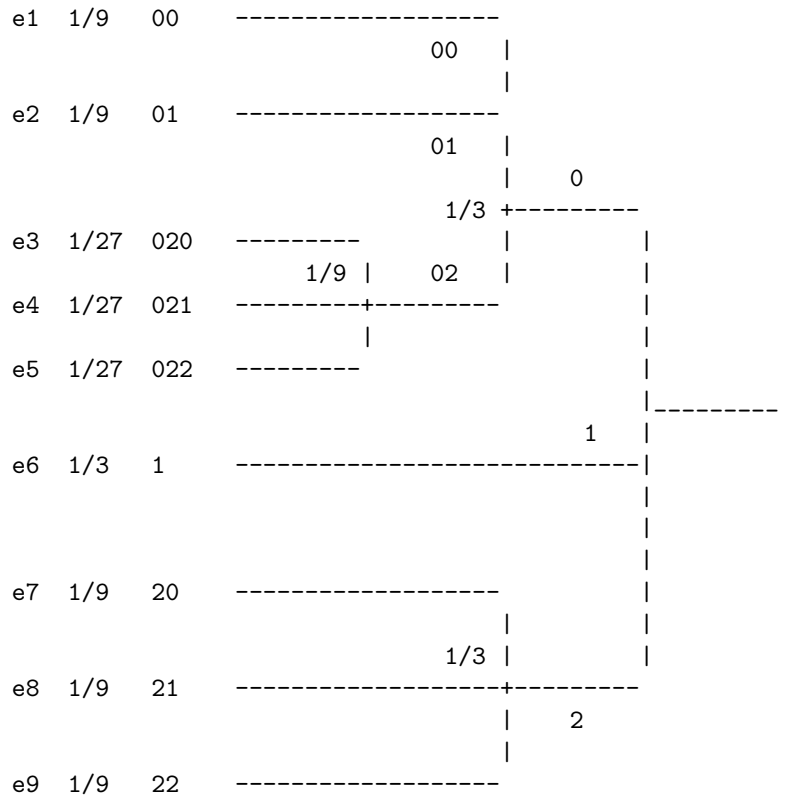

Figure 3. A simple (ternary) Huffman code tree with k = 3

follows from the Huffman construction. It is customary to make the approximation that $l_i = -\log_k p_i$.

Because of the relationship between different bases of logarithm a and b that $\forall x > 0 \;\; \log_b x = (\log_a x)/(\log_a b) = (\log_b a) \log_a x$, changing base of logarithms has the effect of scaling the logarithms by a multiplicative constant, $\log_b a$. As such, the choice of base of logarithm is somewhat arbitrary. The two most common bases are 2 and e. When the base is 2, the information content is said to be in bits. When the base is e, the information content is said to be in nits [Boulton and Wallace, 1970, p. 63; Wallace, 2005, sec. 2.1.8; Comley and Dowe, 2005, sec. 11.4.1; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1], a term which I understand to have had its early origins in (thermal) physics. Alternative names for the nit include the *natural ban* (used by Alan Turing (1912-1954) [Hodges, 1983, pp. 196-197]) and (much much later) the *nat*.

2.2 Prefix codes and Kraft's inequality

Furthermore, defining a prefix code to be a set of (k-ary) strings (of arity k, i.e., where the available alphabet from which each symbol in the string can be selected is of size k) such that no string is the prefix of any other, then we note that the 2^n binary strings of length n form a prefix code. Elaborating, neither of the $2^1 = 2$ binary strings 0 and 1 is a prefix of the other, so the set of code words $\{0, 1\}$ forms a prefix code. Again, none of the $2^2 = 4$ binary strings 00, 01, 10 and 11 is a prefix of any of the others, so the set of code words $\{00, 01, 10, 11\}$ forms a prefix code. Similarly, for $k \geq 2$ and $n \geq 1$, the k^n k-ary strings of length n likewise form a prefix code. We also observe that the fact that the Huffman code construction leads to a (Huffman) *tree* means that the result of any Huffman code construction is always a prefix code. (Recall the examples from sec. 2.1.)

Prefix codes are also known as (or, perhaps more technically, are equivalent to) *instantaneous codes*. In a prefix code, as soon as we see a code-word, we instantaneously recognise it as an intended part of the message — because, by the nature of prefix codes, this code-word can not be the prefix of anything else. Non-prefix (and therefore non-instantaneous) codes do exist, such as (e.g.) $\{0, 01, 11\}$. For a string of the form 01^n, we need to wait to the end of the string to find out what n is and whether n is odd or even before we can decode this in terms of our non-prefix code. (E.g., 011 is 0 followed by 11, 0111 is 01 followed by 11, etc.) For the purposes of the remainder of our writings here, though, we can safely and do restrict ourselves to (instantaneous) prefix codes.

A result often attributed to Kraft [1949] but which is believed by many to have been known to at least several others before Kraft is Kraft's inequality — namely, that in a k-ary alphabet, a prefix code of code-lengths $l_1, \ldots, l_m$ exists if and only if $\sum_{i=1}^{m} k^{-l_i} \leq 1$. The Huffman code construction algorithm, as carried out in our earlier examples (perhaps especially those of figs. 1 and 3), gives an informal intuitive argument as to why Kraft's inequality must be true.

2.3 Entropy

Let us re-visit our result from equation (1) and the standard accompanying approximation that $l_i = -\log p_i$.

Let us begin with the 2-state case. Suppose we have probabilities p_1 and $p_2 = 1 - p_1$ which we wish to encode with code-words of length $l_1 = -\log q_1$ and $l_2 = -\log q_2 = -\log(1 - q_1)$ respectively. As per the Huffman code construction (and Kraft's inequality), choosing such code lengths gives us a prefix code (when these code lengths are non-negative integers).

The negative of the expected code length would then be

$$p_1 \log q_1 + (1 - p_1) \log(1 - q_1),$$

and we wish to choose q_1 and $q_2 = 1 - q_1$ to make this code as short as possible on average — and so we differentiate the negative of the expected code length with respect to q_1.

$$\begin{aligned} 0 &= \frac{d}{dq_1}(p_1 \log q_1 + (1 - p_1) \log(1 - q_1)) = (p_1/q_1) - ((1 - p_1)/(1 - q_1)) \\ &= (p_1(1 - q_1) - q_1(1 - p_1))/(q_1(1 - q_1)) = (p_1 - q_1)/(q_1(1 - q_1)) \end{aligned}$$

and so $(p_1 - q_1) = 0$, and so $q_1 = p_1$ and $q_2 = p_2$.

This result also holds for p_1, p_2, p_3, q_1, q_2 and q_3 in the 3-state case, as we now show. Let $P_2 = p_2/(p_2 + p_3)$, $P_3 = p_3/(p_2 + p_3) = 1 - P_2$, $Q_2 = q_2/(q_2 + q_3)$ and $Q_3 = q_3/(q_2 + q_3) = 1 - Q_2$, so $p_2 = (1 - p_1)P_2$, $p_3 = (1 - p_1)P_3 = (1 - p_1)(1 - P_2)$, $q_2 = (1 - q_1)Q_2$ and $q_3 = (1 - q_1)Q_3 = (1 - q_1)(1 - Q_2)$.

Encoding the events of probability p_1, p_2 and p_3 with code lengths $-\log q_1$, $-\log q_2$ and $-\log q_3$ respectively, the negative of the expected code length is then $p_1 \log q_1 + (1 - p_1)P_2 \log((1 - q_1)Q_2) + (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))$. To minimise, we differentiate with respect to both q_1 and Q_2 in turn, and set both of these to 0.

$$\begin{aligned} 0 &= \frac{\partial}{\partial q_1}(p_1 \log q_1 + (1 - p_1)P_2 \log((1-_1)Q_2) + \\ &\qquad (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))) \\ &= (p_1/q_1) - ((1 - p_1)P_2)/(1 - q_1) - ((1 - p_1)(1 - P_2))/(1 - q_1) \\ &= (p_1/q_1) - (1 - p_1)/(1 - q_1) \\ &= (p_1 - q_1)/(q_1(1 - q_1)) \end{aligned}$$

exactly as in the 2-state case above, where again $q_1 = p_1$.

$$\begin{aligned} 0 &= \frac{\partial}{\partial Q_2}(p_1 \log q_1 + (1 - p_1)P_2 \log((1 - q_1)Q_2) + \\ &\qquad (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))) \end{aligned}$$

$$\begin{aligned} &= (((1-p_1)P_2)/Q_2) - ((1-p_1)(1-P_2)/(1-Q_2)) \\ &= (1-p_1) \times ((P_2/Q_2) - (1-P_2)/(1-Q_2)) \\ &= (1-p_1) \times (P_2-Q_2)/(Q_2(1-Q_2)) \end{aligned}$$

In the event that $p_1 = 1$, the result is trivial. With $p_1 \neq 1$, we have, of very similar mathematical form to the two cases just examined, $0 = (P_2/Q_2) - (1-P_2)/(1-Q_2)$, and so $Q_2 = P_2$, in turn giving that $q_2 = p_2$ and $q_3 = p_3$.

One can proceed by the principle of mathematical induction to show that, for probabilities $(p_1, ..., p_i, ..., p_{m-1}, p_m = 1 - \sum_{i=1}^{m-1} p_i)$ and code-words of respective lengths $(-\log q_1, ..., -\log q_i, ..., -\log q_{m-1}, -\log q_m = -\log(1 - \sum_{i=1}^{m-1} q_i))$, the expected code length $-(p_1 \log q_1 + ... + p_i \log q_i + ... + p_{m-1} \log q_{m-1} + p_m \log q_m)$ is minimised when $\forall i \; q_i = p_i$.

This expected (or average) code length,

$$\sum_{i=1}^{m} p_i \times (-\log p_i) = -\sum_{i=1}^{m} p_i \log p_i \tag{2}$$

is called the *entropy* of the m-state probability distribution $(p_1, ..., p_i, ..., p_m)$.

Note that if we sample randomly from the distribution p with code-words of length $-\log p$, then the (expected) average long-term cost is the entropy.

Where the distribution is continuous rather than (as above) discrete, the sum is replaced by an integral and (letting x be a variable being integrated over) the entropy is then defined as

$$\begin{aligned} \int f \times (-\log f) \; dx &= -\int f \log f \; dx = -\int f(x) \log f(x) \; dx \\ &= \int f(x) \times (-\log f(x)) \; dx \end{aligned} \tag{3}$$

And, of course, entropy can be defined for hybrid structures of both discrete and continuous, such as Bayesian network graphical models (of sec. 7.6) — see sec. 3.6, where it is pointed out that for the hybrid continuous and discrete Bayesian net graphical models in [Comley and Dowe, 2003; 2005] (emanating from the current author's ideas in [Dowe and Wallace, 1998]), the log-loss scoring approximation to Kullback-Leibler distance has been used [Comley and Dowe, 2003, sec. 9].

The next section, sec. 2.4, introduces Turing machines as an abstract model of computation and then discusses the formal relationship between MML and minimising the length of some (constrained) input to a Turing machine. The section can be skipped on first reading.

2.4 Turing machines and algorithmic information theory

The area of algorithmic information theory was developed independently in the 1960s by Solomonoff [1960; 1964], Kolmogorov [1965] and Chaitin [1966], independently of and slightly before the seminal Wallace & Boulton paper on MML [1968].

Despite the near-simultaneous independent work of the young Chaitin [1966] and the independent earlier work of Solomonoff [1960; 1964] pre-dating Kolmogorov, the area of algorithmic information theory is often referred to by many as Kolmogorov complexity (e.g., [Wallace and Dowe, 1999a; Li and Vitányi, 1997]). Before introducing the notion of the algorithmic complexity (or Kolmogorov complexity) of a string s, we must first introduce the notion of a Turing machine [Turing, 1936; Wallace, 2005, sec. 2.2.1; Dowe 2008b, pp. 449-450]. Following [Dowe, 2008b, pp. 449-450], a *Turing machine* (TM) [Wallace, 2005, sec. 2.2.1; Dowe, 2008b, pp. 449-450] 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) " ", also denoted by "␣"). We assume that Turing machines have a read/write head on an infinitely long tape, finitely bounded to the left and infinitely long to the right. Turing machines have a set of instructions — or an instruction set — as follows. 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: $f : States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols)$.

So, the definition that we are using is that a Turing Machine M is a set of quadruples $\{Q_n\} = \{\langle q_i, q_j, s_k, \{s_l, H\}\rangle\}$ where

- $q_i, q_j \in \{1, \ldots, m\}$ (the machine states)
- $s_k, s_l \in \{s_0, \ldots, s_r\}$ (the symbols)
- $H \in \{R, L\}$ (tape head direction)

(such that no two quadruples have the same first and third elements). The Turing machine in state q_i given input s_k goes into state q_j and either stays where it is and writes a symbol (s_l) or moves to the left or right (depending upon the value of H) without writing a symbol.

An alternative equivalent definition of a Turing Machine M which we could equally well use instead is a set of quintuples $\{Q_n\} = \{\langle q_i, q_j, s_k, s_l, H\rangle\}$ where

- (the machine states) $q_i, q_j \in \{1, \ldots, m\}$
- (the symbols) $s_k, s_l \in \{s_0, \ldots, s_r\}$
- (tape head direction) $H \in \{R, L\}$

and the Turing machine in state q_i given input s_k then goes into state q_j, writes symbol s_l and moves the head in direction H (and, again, we require that no two quintuples have the same first and third elements).

Note that the Turing Machine (TM) in the first definition *either* writes a (new) symbol *or* moves the head at each step whereas the TM in the second of these two equivalent definitions *both* writes a (new) symbol *and* moves the head.

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: $f : States \times \{0, 1\} \rightarrow States \times (\{L, R\} \cup \{0, 1\})$.

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

An example of a Turing machine would be the program from fig. 4, which, given two inputs, x_0 and x_1, adds them together, writing $x_0 + x_1$ and then stopping[1]. This machine adds two unary numbers (both at least 1), terminated by blanks (and separated by a single blank). In unary, e.g., 4 is represented by "1111␣". In general in unary, n is represented by n 1s followed by a blank.

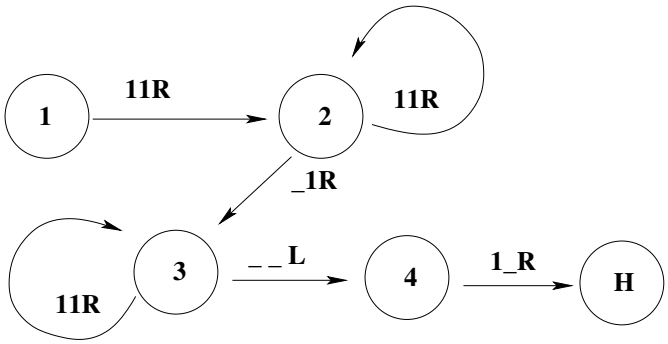

Figure 4. A Turing machine program for adding two numbers

Alternatively, recalling our notation of quintuples, $< q_i, q_j, s_k, s_l, H >$, this Turing machine adding program from fig. 4 can be represented as:

$$\{\langle 1, 2, 1, 1, R\rangle, \langle 2, 2, 1, 1, R\rangle, \langle 2, 3, \sqcup, 1, R\rangle, \langle 3, 3, 1, 1, R\rangle, \langle 4, 5, 1, \sqcup, R\rangle\}$$

(where state 5 is the Halting — or stop — state, also referred to as H).
(This Turing machine program over-writes the blank (␣) in the middle with a 1 and removes a 1 from the right end of the second number — and, in so doing, leaves behind the unary representation of the sum.)

Another 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) $a_0 + a_1 x$. In this case, x would input in binary (base 2), and the output would be the binary representation of $a_0 + a_1 x$.

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

[1]Wherever he might or might not have inherited it from, I acknowledge obtaining the figure in fig. 4 from Kevin B. Korb.

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 translation program] (or code) c_M so that once U is input c_M it forever after behaves as though it were M.

The algorithmic complexity (or Kolmogorov complexity), $K_U(x)$, of a string x is the length of the shortest input (l_x) to a Universal Turing Machine such that, given input l_x, U outputs x and then stops. (This is the approach of Kolmogorov [1965] and Chaitin [1966], referred to as stream one in [Wallace and Dowe, 1999a].)

Algorithmic information theory can be used to give the algorithmic probability [Solomonoff, 1960; 1964; 1999; 2008] of a string (x) or alternatively also to insist upon the two-part MML form [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.2–2.3].

Let us elaborate, initially by recalling the notion of a prefix code (from sec. 2.2) and then by considering possible inputs to a UTM. Let us consider the two binary strings 0 and 1 of length 1, the four binary strings 00, 01, 10 and 11 of length 2, and (in general) the 2^n binary strings of length n. Clearly, if a Turing Machine stops on some particular input (of length n), then it will stop on that input with any suffix appended.

The (unnormalized) probability that a UTM, U, will generate x from random input is $P_U(x) = \sum_{s:U(s)=x} 2^{-\text{length}(s)}$, summing over the strings s such that U taking input s will output x and then stop. In Solomonoff's original predictive specification [Solomonoff, 1960; 1964] (stream two from [Wallace and Dowe, 1999a]), the (unnormalized) summation actually includes more strings (and leads to a greater sum), including [Wallace, 2005, sec. 10.1.3] those strings s such that U on input s produces x and possibly a suffix — i.e., outputs a string for which x is a prefix. For this sum to be finite, we must add the stipulation that the strings s (over which we sum) must form a prefix code. In choosing the strings s to form a prefix code, the sum is not affected by insisting that the strings s are all chosen so that (e.g.) for no prefix s' of s does $U(s') = x$ and then halt. And, for the sum $\sum_x P_U(x)$ to be useful, we must again make sure that the strings x are prefix-free — i.e., that the strings x all together form a prefix code - so as to avoid double-counting.

Clearly, $2^{-K_U(x)} < P_U(x)$, since $K_U(x)$ takes only the shortest (and biggest) [input] term outputting x, whereas $P_U(x)$ takes all the terms which output x (whether or not we wish to also include terms which append a suffix to x). The earlier mention above of "(unnormalized)" is because, for many inputs, the UTM will not halt [Turing, 1936; Chaitin, 2005; Dowe, 2008a, footnote 70]. For the purposes of prediction, these considerations just discussed are sufficient. But, for the purposes of *inference* (or, equivalently, *explanation* or *induction*), we need a *two-part* construction — as per theorems from [Barron and Cover, 1991] and arguments from [Wallace and Freeman, 1987, p. 241; Wallace, 2005, sec. 3.4.5, pp. 190–191] (and some examples of what can go wrong when we don't have a two-part construction [Wallace and Dowe, 1999a, sec. 6.2; 1999b, secs. 1.2, 2.3 and 3; 1999c,

sec. 2]). Our two-part input to the (Universal) Turing Machine will be such that [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.3.6–2.3.9] the first part results in no output being written but rather the Turing machine is programmed with the hypothesis, H. Now programmed with the hypothesis, H, the Turing machine now looks at the second part of the message (which is possibly the output of a Huffman code) and uses H to write out the data, D. The MML inference will be the hypothesis, H, which is represented by the first part of the shortest two-part input giving rise to the data, D.

The other thing to mention here is the Bayesianism inherent in all these approaches. The Bayesian and (two-part [file compression]) information-theoretic interpretations to MML are both clearly Bayesian. And, although some authors have been known to neglect this (by sweeping Order 1, $O(1)$, terms under the carpet or otherwise neglecting them), the choice of (Universal) Turing Machine in algorithmic information theory is (obviously?) also a Bayesian choice [Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450].

2.5 Digression on Wallace non-universality probability

This section is a digression and can be safely skipped without any loss of continuity or context, but it does follow on from sec. 2.4 — which is why it is placed here.

The Wallace non-universality probability [Dowe, 2008a, sec. 0.2.2, p. 530, col. 1 and footnote 70] of a UTM, U, is the probability that, given a particular infinitely long random bit string as input, U will become non-universal at some point. Quite clearly, the Wallace non-universality probability (WNUP) equals 1 for all non-universal TMs. Similarly, WNUP(U) is greater than 0 for all TMs, U; and WNUP equals 1 for some UTM if and only if it equals 1 for all UTMs. Wallace, others and I believed it to equal 1. In unpublished private communication, George Barmpalias argues that it isn't equal to 1, appealing to a result of Kucera. George is correct (and Chris and I mistaken) if and only if $\inf_{\{U:U \text{ a UTM}\}} \text{WNUP(U)} = 0$.

This section was a digression and could be safely skipped without any loss of continuity or context.

3 PROBABILISTIC INFERENCE, LOG-LOSS SCORING AND KULLBACK-LEIBLER DISTANCE — AND UNIQUENESS

There are many measures of predictive accuracy. The simplest of these, such as on a quiz show, is the number of correct answers (or "right"/"wrong" scoring). There are likewise many measures of how close some estimated function is to the true function from which the data is really coming.

We shall present the notions of probabilistic scoring and of measuring a distance between two functions.

From the notion of probabilistic scoring, we shall present our new apparent uniqueness property of log-loss scoring [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438]. From the notion of measuring a distance between two functions, we shall present a related result showing uniqueness (or two versions of uniqueness) of Kullback-Leibler distance [Dowe, 2008a, p. 438]. For those interested in causal decision theory and scoring rules and to those simply interested in scoring rules and scoring probabilities, I highly recommend log-loss scoring and Kullback-Leibler distance — partly for their invariance and partly for their apparent uniqueness in having this invariance.

3.1 "Right"/"wrong" scoring and re-framing

Imagine two different quizzes which are identical apart from their similar but not quite identical beginnings. Quiz 1 begins with a multiple-choice question with 4 possible answers: $0, 1, 2, 3$ or (equivalently, in binary) $00, 01, 10, 11$.

Quiz 2 begins with 2 multiple-choice questions:

- Q2.1: is the 2^{nd} last bit a 0 or a 1?, and
- Q2.2: is the last bit a 0 or a 1?

Getting a score of 1 correct at the start of quiz 1 corresponds to getting a score of 2 at the start of quiz 2. Getting a score of 0 at the start of quiz 1 corresponds to a score of either 0 or 1 at the start of quiz 2. This seems unfair — so we might try to attribute the problem to the fact that quiz 1 began with 1 4-valued question where quiz 2 began with 2 2-valued questions and explore whether all is fair when (e.g.) all quizzes have 2 2-valued questions.

But consider now quiz 3 which, like quiz 2, begins with 2 2-valued questions, as follows:

- Q3.1: is the 2^{nd} last bit a 0 or a 1?, and
- Q3.2: are the last two bits equal or not equal?

Getting Q3.2 correct means that on quiz 2 we either get 0 (if we get Q2.1 wrong) or 2 (if we get Q2.2 correct and therefore all questions correct).

We see that no matter how we re-frame the question — whether as one big question or as lots of little questions — we get all answers correct on one quiz if and only if we get all answers correct on all quizzes. But, however, as the following example (of Quiz 4 and Quiz 5) demonstrates, we also see that even when all questions are binary (yes/no), it is possible to have two different framings of n questions such that in one such quiz (here, Quiz 4) we have $(n-1)$ questions answered correctly (and only 1 incorrectly) and in the re-framing to another quiz (here, Quiz 5) all n questions are answered incorrectly.

Quiz 4 (of n questions):

- Q4.1: What is the 1^{st} of the n bits?
- Q4.i ($i = 2, ..., n$): Is the 1^{st} bit equal to the i^{th} bit?

Quiz 5 (of n questions):

- Q5.1: What is the 1^{st} of the n bits?
- Q5.i ($i = 2, ..., n$): What is the i^{th} bit?

If the correct bit string is $0^n = 0...0$ and our guess is $1^n = 1...1$, then on Quiz 4 we will get $(n-1)$ correct (and 1 wrong) whereas on quiz 5 we will get 0 correct (and all n wrong).

This said by way of motivation, we now look at forms of prediction that remain invariant to re-framing — namely, probabilistic prediction with log(arithm)-loss — and we present some recent uniqueness results [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437-438] here.

3.2 Scoring predictions, probabilistic predictions and log-loss

The most common form of prediction seems to be a prediction without a probability or anything else to quantify it. Nonetheless, in some forms of football, the television broadcaster sometimes gives an estimated probability of the kicker scoring a goal — based on factors such as distance, angle and past performance. And, of course, if it is possible to wager a bet on the outcome, then accurately estimating the probability (and comparing this with the potential pay-out if successful) will be of greater interest.

Sometimes we don't care overly about a probability estimate. On some days, we might merely wish to know whether or not it is more probable that it will rain or that it won't. On such occasions, whether it's 52% probable or 97% probable that it will rain, we don't particularly care beyond noting that both these numbers are greater than 50% and we'll take our umbrella with us in either case.

And sometimes we most certainly want a good and reliable probability estimate. For example, a patient reporting with chest pains doesn't want to be told that there's only a 40% chance that you're in serious danger (with a heart attack), so you can go now. And nor does an inhabitant of an area with the impending approach of a raging bush-fire want to be told that there's only a 45% chance of your dying or having serious debilitation if you stay during the fire, so you might as well stay. The notion of "reasonable doubt" in law is pertinent here — and, without wanting to seem frivolous, so, too, is the notion of when a cricket umpire should or shouldn't give the "benefit of the doubt" to the person batting (in l.b.w. or other contentious decisions).

Now, it is well-known that with logarithm-loss function ($\log p$) for scoring probabilistic predictions, the optimal strategy is to give the true probability, if known.

This property also holds true for quadratic loss ($(1-p)^2$) and has also been shown to be able to hold for certain other functions of probability [Deakin, 2001]. What we will show here is our new result that the logarithm-loss (or log-loss) function has an apparent uniqueness property on re-framing of questions [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438].

Let us now consider an example involving (correct) diagnosis of a patient. (With apologies to any and all medically-informed human earthlings of the approximate time of writing, the probabilities in the discussion(s) below might be from non-earthlings, non-humans and/or from a different time.) We'll give four possibilities:

1. no diabetes, no hypothyroidism
2. diabetes, but no hypothyroidism
3. no diabetes, but hypothyroidism
4. both diabetes and hypothyroidism.

Of course, rather than present this as one four-valued diagnosis question, we could have presented it in a variety of different ways.

As a second possibility, we could have also asked, e.g., the following two two-valued questions:

1. no diabetes
2. diabetes

and

1. hypothyroidism
2. no hypothyroidism.

As another (third) alternative line, we could have begun with

1. no condition present
2. at least one condition present,

and then finished if there we no condition present but, if there were at least one condition present, instead then continued with the following 3-valued question:

1. diabetes, but no hypothyroidism
2. no diabetes, but hypothyroidism
3. both diabetes and hypothyroidism.

To give a correct diagnosis, in the original setting, this requires answering one question correctly. In the second setting, it requires answering exactly two questions correctly — and in the third setting, it might require only answering one question correctly but it might require answering two questions correctly.

Clearly, then, the number of questions answered correctly is not invariant to the re-framing of the question. However, the sum of logarithms of probabilities *is* invariant, and — apart from (quite trivially, multiplying it by a constant, or) adding a constant multiple of the entropy of the prior distribution - would appear to be unique in having this property.

Let us give two examples of this. In the first example, our conditions will be independent of one another. In the second example, our conditions will be dependent upon one another.

So, in the first case, with the conditions independent of one another, suppose the four estimated probabilities are

1. no diabetes, no hypothyroidism; probability 1/12
2. diabetes, but no hypothyroidism; probability $2/12 = 1/6$
3. no diabetes, but hypothyroidism; probability $3/12 = 1/4$
4. both diabetes and hypothyroidism; probability $6/12 = 1/2$.

Then, in the second possible way we had of looking at it (with the two given two-valued questions), in the first case we have

1. no diabetes; probability 1/3
2. diabetes; probability 2/3

and — because the diseases are supposedly independent of one another in the example — we have

1. no hypothyroidism; probability 1/4
2. hypothyroidism; probability 3/4.

The only possible way we can have an additive score for both the diabetes question and the hypothyroid question separately is to use some multiple of the logarithms. This is because the probabilities are multiplying together and we want some score that adds across questions, so it must be (a multiple of) the logarithm of the probabilities.

In the third alternative way that we had of looking at it, Pr(no condition present) = 1/12. If there is no condition present, then we do need need to ask the remaining question. But, in the event (of probability 11/12) that at least one condition is present, then we have

1. diabetes, but no hypothyroidism; probability 2/11

2. no diabetes, but hypothyroidism; probability 3/11

3. both diabetes and hypothyroidism; probability 6/11.

And the logarithm of probability score again works.

So, our logarithm of probability score worked when the conditions were assumed to be independent of one another.

We now consider an example in which they are not independent of one another.

Suppose the four estimated probabilities are

1. no diabetes, no hypothyroidism; probability 0.1

2. diabetes, but no hypothyroidism; probability 0.2

3. no diabetes, but hypothyroidism; probability 0.3

4. both diabetes and hypothyroidism; probability 0.4.

Then, in the second possible way we had of looking at it (with the two given two-valued questions), for the first of the two two-valued questions we have

1. no diabetes; probability 0.4

2. diabetes; probability 0.6

and then for the second of which we have either

1. no hypothyroidism; prob(no hypothyroidism | no diabetes) = $0.1/(0.1 + 0.3) = 0.1/0.4 = 0.25$

2. hypothyroidism; prob(hypothyroidism | no diabetes) = $0.3/(0.1 + 0.3) = 0.3/0.4 = 0.75$

or

1. no hypothyroidism; prob(no hypothyroidism | diabetes) = $0.2/(0.2 + 0.4) = 0.2/0.6 = 1/3$

2. hypothyroidism; prob(hypothyroidism | diabetes) = $0.4/(0.2+0.4) = 0.4/0.6 = 2/3$.

And in the third alternative way of looking at this, prob(at least one condition present) = 0.9. If there is no condition present, then we do need need to ask the remaining question. But, in the event (of probability 9/10) that at least one condition is present, then we have the following three-way question:

1. diabetes, but no hypothyroidism; probability 2/9

2. no diabetes, but hypothyroidism; probability 3/9

3. both diabetes and hypothyroidism; probability 4/9.

We leave it to the reader to verify that the logarithm of probability score again works, again remaining invariant to the phrasing of the question. (Those who would rather see the above examples worked through with general probabilities rather than specific numbers are referred to a similar calculation in sec. 3.6.)

Having presented in sec. 3.1 the problems with "right"/"wrong" scoring and having elaborated on the uniqueness under re-framing of log(arithm)-loss scoring [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437-438] above, we next mention at least four other matters.

First, borrowing from the spirit of an example from [Wallace and Patrick, 1993], imagine we have a problem of inferring a binary (2-class) output and we have a binary choice (or a binary split in a decision tree) with the following output distributions. For the "no"/"left" branch we get 95 in class 1 and 5 in class 2 (i.e., 95:5), and for the "yes"/"right" branch we get 55 in class 1 and 45 in class 2 (i.e., 55:45).

Because both the "no"/"left" branch and the "yes"/"right" branch give a majority in class 1, someone only interested in "right"/"wrong" score would fail to pick up on the importance and significance of making this split, simply saying that one should always predict class 1. Whether class 1 pertains to heart attack, dying in a bush-fire or something far more innocuous (such as getting wet in light rain), by reporting the probabilities we don't run the risk of giving the wrong weights to (so-called) type I and type II errors (also known as false positives and false negatives). (Digressing, readers who might incidentally be interested in applying Minimum Message Length (MML) to hypothesis testing are referred to [Dowe, 2008a, sec. 0.2.5, p. 539 and sec. 0.2.2, p. 528, col. 1; 2008b, p. 433 (Abstract), p. 435, p. 445 and pp. 455–456; Musgrave and Dowe, 2010].) If we report a probability estimate of (e.g.) 45%, 10%, 5%, 1% or 0.1%, we leave it to someone else to determine the appropriate level of risk associated with a false positive in diagnosing heart attack, severe bush-fire danger or getting caught in light drizzle rain.

And we now mention three further matters.

First, some other uniqueness results of log(arithm)-loss scoring are given in [Milne, 1996; Huber, 2008]. Second, in binary (yes/no) multiple-choice questions, it is possible (and not improbable for a small number of questions) to serendipitously fluke a good "right"/"wrong" score, even if the probabilities are near 50%-50%, and with little or no risk of downside. But with log(arithm)-loss scoring, if the probabilities are near 50%-50% (or come from random noise and are 50%-50%), then predictions with more extreme probabilities are fraught with risk. And, third, given all these claims about the uniqueness of log(arithm)-loss scoring in being invariant to re-framing (as above) [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438] and having other desirable properties [Milne, 1996; Huber, 2008], we discuss in sec. 3.3 how a quiz show contestant asked to name (e.g.) a city or a person and expecting to be scored on "right"/"wrong" could instead give a probability distribution over cities or people's names and be scored by

log(arithm)-loss.

3.3 Probabilistic scoring on quiz shows

Very roughly, one could have probabilistic scoring on quiz shows as follows. For a multiple-choice answer, things are fairly straightforward as above. But what if, e.g., the question asks for a name or a date (such as a year)?

One could give a probability distribution over the length of the name. For example, the probability that the name has length l might be $1/2^l = 2^{-l}$ for $l = 1, 2, ...$ Then, for each of the l places, there could be 28 possibilities (the 26 possibilities a, ..., z, and the 2 possibilities space " " and hyphen "-") of probability 1/28 each. So, for example, as a default, "Wallace" would have a probability of $2^{-7} \times (1/28)^7$. Call this distribution Default. (Of course, we could refine Default by, e.g., noticing that the first character will neither be a space " " nor a hyphen "-" and/or by (also) noticing that both the space " " and the hyphen "-" are never followed by a space " " or a hyphen "-".) For the user who has some idea rather than no proverbial idea about the answer, it is possible to construct hybrid distributions. So, if we wished to allocate probability 1/2 to "Gauss" and probability 1/4 to "Fisher" and otherwise we had no idea, then we could give a probability of 1/2 to "Gauss", 1/4 to "Fisher" and for all other answers our probability we would roughly be 1/4 × Default. A similar comment applies to someone who thinks that (e.g.) there is a 0.5 probability that the letter "a" occurs at the start and a 0.8 probability that the letter "c" appears at least once or instead that (e.g.) a letter "q" not at the end can only be followed by an "a", a "u", a space " " or a hyphen "-".

If a question were to be asked about the date or year of such and such an event, one could give a ([truncated] Gaussian) distribution on the year in the same way that entrants having been giving (Gaussian) distributions on the margin of Australian football games since early 1996 as per sec. 3.5.

It would be nice to see a (television) quiz (show) one day with this sort of scoring system. One could augment the above as follows. Before questions were asked to single contestants — and certainly before questions were thrown upon to multiple contestants and grabbed by the first one pressing the buzzer — it could be announced what sort of question it was (e.g., multiple-choice [with n options] or open-ended) and also what the bonus in bits was to be. The bonus in bits (as in the constant to be added to the logarithm of the contestant's allocated probability to the correct answer) could relate to the ([deemed] prior) difficulty of the question, (perhaps) as per sec. 3.4 — where there is a discussion of adding a term corresponding to (a multiple of) the log(arithm) of the entropy of some Bayesian prior over the distribution of possible answers. And where quiz shows have double points in the final round, that same tradition could be continued.

One comment perhaps worth adding here is that in typical quiz shows, as also in the probabilistic and Gaussian competitions running on the Australian Football League (AFL) football competitions from sec. 3.5, entrants are regularly updated

on the scores of all their opponents. In the AFL football competitions from sec. 3.5, this happens at the end of every round. In quiz shows, this often also happens quite regularly, sometimes after every question. The comment perhaps worth adding is that, toward the end of a probabilistic competition, contestants trying to win who are not winning and who also know that they are not winning might well nominate "aggressive" choices of probabilities that do not represent their true beliefs and which they probably would not have chosen if they had not been aware that they weren't winning.

Of course, if (all) this seems improbable, then please note from the immediately following text and from sec. 3.5 that each year since the mid-1990s there have been up to hundreds of people (including up to dozens or more school students, some in primary school) giving not only weekly probabilistic predictions on results of football games but also weekly predictions of probability distributions on the margins of these games. And these have all been scored with log(arithm)-loss.

3.4 Entropy of prior and other comments on log-loss scoring

Before finishing with some references to papers in which we have used log(arithm)-loss probabilistic ("bit cost") scoring and mentioning a log-loss probabilistic competition we have been running on the Australian Football League (AFL) since 1995 in sec. 3.5, two other comments are worth making in this section.

Our first comment is to return to the issue of quadratic loss ($(1-p)^2$, which is a favourite of many people) and also the loss functions suggested more recently by Deakin [2001]. While it certainly appears that only log(arithm)-loss (and multiples thereof) retains invariance under re-framing, we note that

$$-\log(p_1^n\ p_2^n\ \ldots\ p_m^n) = -n\sum_{i=1}^{m}\log p_i \tag{4}$$

So, although quadratic loss ($(1-p)^2$) does not retain invariance when adding scores between questions, if we were for some reason to want to *multiply* scores between questions (rather than add, as per usual), then the above relationship in equation (4) between a power score (quadratic score with $n=2$) and log(arithmic) score — namely, that the log of a product is the sum of the logs — might possibly enable some power loss to be unique in being invariant under re-framing (upon *multiplication*).

The other comment to make about the uniqueness of log(arithm)-loss (upon addition) under re-framing is that we can also add a term corresponding to (a multiple of) the entropy of the Bayesian prior [Tan and Dowe, 2006, sec. 4.2; Dowe, 2008a, footnote 176; 2008b, p. 438]. (As is hinted at in [Tan and Dowe, 2006, sec. 4.2, p. 600] and explained in [Dowe, 2008a, footnote 176], the idea for this arose in December 2002 from my correcting a serious mathematical flaw in [Hope and Korb, 2002]. Rather than more usefully use the fact that a logarithm of probability ratios is the difference of logarithms of probabilities, [Hope and Korb,

2002] instead suggests using a ratio of logarithms — and this results in a system where the optimal score will rarely be obtained by using the true probability.)

Some of many papers using log(arithm)-loss scoring include [Good, 1952] (where it was introduced for the binomial distribution) and [Dowe and Krusel, 1993, p. 4, Table 3; Dowe *et al.*, 1996d; Dowe *et al.*, 1996a; Dowe *et al.*, 1996b; Dowe *et al.*, 1996c; Dowe *et al.*, 1998, sec. 3; Needham and Dowe, 2001, Figs. 3–5; Tan and Dowe, 2002, sec. 4; Kornienko *et al.*, 2002, Table 2; Comley and Dowe, 2003, sec. 9; Tan and Dowe, 2003, sec. 5.1; Comley and Dowe, 2005, sec. 11.4.2; Tan and Dowe, 2004, sec. 3.1; Kornienko *et al.*, 2005a, Tables 2–3; Kornienko *et al.*, 2005b; Tan and Dowe, 2006, secs. 4.2–4.3] (and possibly also [Tan *et al.*, 2007, sec. 4.3]), [Dowe, 2007; 2008a, sec. 0.2.5, footnotes 170–176 and accompanying text; 2008b, pp. 437–438]. The 8-class multinomial distribution in [Dowe and Krusel, 1993, p. 4, Table 3] (from 1993) is the first case we are aware of in which log-loss scoring is used for a distribution which is not binomial, and [Dowe *et al.*, 1996d; 1996a; 1996b; 1996c] (from 1996) are the first cases we are aware of in which log-loss scoring was used for the Normal (or Gaussian) distribution.

3.5 Probabilistic prediction competition(s) on Australian football

A log(arithm)-loss probabilistic prediction competition was begun on the outcome of Australian Football League (AFL) matches in 1995 [Dowe and Lentin, 1995; Dowe, 2008b, p. 48], just before Round 3 of the AFL season. In 1996, this was extended by the author to a log(arithm)-loss Gaussian competition on the margin of the game, in which competition entrants enter a μ and a σ — in order to give a predictive distribution $N(\mu, \sigma)$ on the margin — for each game [Dowe *et al.*, 1996d; 1996a; 1996b; 1996c; 1998, sec. 3; Dowe, 2008a, sec. 0.2.5]. These log-loss compression-based competitions, with scores in bits (of information), have been running non-stop ever since their inceptions, having been put on the WWW in 1997 and at their current location of `www.csse.monash.edu.au/~footy` since 1998 [Dowe, 2008a, footnote 173]. (And thanks to many, especially Torsten Seemann for all the unsung behind the scenes support in keeping these competitions going [Dowe, 2008a, footnote 217].) The optimal long-term strategy in the log(arithm)-loss probabilistic AFL prediction competition would be to use the true probability if one knew it. Looking ahead to sec. 3.6, the optimal long-term strategy in the Gaussian competition would be to choose μ and σ so as to minimise the Kullback-Leibler distance from the (true) distribution on the margin to $N(\mu, \sigma^2)$. (In the log-loss probabilistic competition, the "true" probability is also the probability for which the Kullback-Leibler distance from the (true) distribution is minimised.) Competitions concerned with minimising sum of squared errors can still be regarded as compression competitions motivated by (expected) Kullback-Leibler distance minimisation, as they are equivalent to Gaussian competitions with σ fixed (where σ can be presumed known or unknown, as long as it's fixed).

3.6 Kullback-Leibler "distance" and measuring "distances" between two functions

Recall from sec. 2.3 that the optimal way of encoding some probability distribution, f, is with code-words of length $-\log f$, and that the average (or expected) cost is $\sum_{i=1}^{m} f_i \times (-\log f_i)$, also known as the entropy. The log(arithm)-loss score is obtained by sampling from some real-world data. If we are sampling from some true (known) distribution, then the optimal long-run average that we can expect will be the entropy.

One way of thinking about the Kullback-Leibler divergence (or "distance") between two distributions, f and g, is as the inefficiency (or sub-optimality) of encoding f using g rather than (the optimal) f. Equivalently, one can think about the Kullback-Leibler distance between two distributions, f and g, as the average (or expected) cost of sampling from distribution f and coding with the corresponding cost of $-\log g$ minus the entropy of f — and, of course, the entropy of f (namely, $-\sum f \log f$) is independent of g.

Recalling equations (2) and (3) for the entropy of discrete and continuous models respectively, the Kullback-Leibler distance from f to g is

$$(\sum_{i=1}^{m} f \times (-\log g_i)) \; - \; (\sum_{i=1}^{m} f \times (-\log f_i)) \quad = \quad \sum_{i=1}^{m} f \times \log(f_i/g_i) \tag{5}$$

for the discrete distribution. As one might expect, for a continuous distribution, the Kullback-Leibler distance from f to g is

$$\begin{aligned} & (\int f(x) \times (-\log g(x)) \; dx) \quad - \quad (\int f(x) \times (-\log f(x)) \; dx) \\ = \quad & \int f \times \log(f(x)/g(x)) \; dx \quad = \quad \int dx \; f \times \log(f/g) \end{aligned} \tag{6}$$

We should mention here that many refer to the Kullback-Leibler "distance" as Kullback-Leibler *divergence* because it is not — in general — symmetrical. In other words, there are plenty of examples when $KL(f, g)$, or, equivalently, $\Delta(g||f)$, is not equal to $KL(g, f) = \Delta(f||g)$.

In sec. 3.2, we showed a new result about uniqueness of log(arithm)-loss scoring in terms of being invariant under re-framing of the problem [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438]. It turns out that there is a similar uniqueness about Kullback-Leibler distance in terms of being invariant to re-framing of the problem [Dowe, 2008b, p. 438]. Despite some mathematics to follow which some readers might find slightly challenging in places, this follows intuitively because (the entropy of the true distribution, f, is independent of g and) the $-\log g$ term in the Kullback-Leibler distance is essentially the same as the log(arithm)-loss term in sec. 3.2.

Before proceeding to this example, we first note that the Kullback-Leibler distance is quite clearly invariant under re-parameterisations such as (e.g.) transforming from polar co-ordinates (x, y) to Cartesian co-ordinates ($r = \text{sign}(x)$.

$\sqrt{x^2+y^2}$, $\theta = \tan^{-1}(y/x)$) and back from Cartesian co-ordinates (r, θ) to polar co-ordinates ($x = r\cos\theta$, $y = r\sin\theta$).

We give an example of our new result (about the uniqueness of Kullback-Leibler distance in terms of being invariant to re-framing) below, letting f and g both have 4 states, with probabilities $f_{1,1}$, $f_{1,2}$, $f_{2,1}$, $f_{2,2}$, $g_{1,1}$, $g_{1,2}$, $g_{2,1}$ and $g_{2,2}$ respectively. The reader is invited to compare the example below with those from sec. 3.2. The reader who would prefer to see specific probabilities rather than these general probabilities is suggested more strongly to compare with sec. 3.2.

$$KL(f,g) = \Delta(g||f) = \sum_{i=1}^{2}\sum_{j=1}^{2} f_{i,j}(\log f_{i,j} - \log g_{i,j})$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{2} f_{i,j}\log(f_{i,j}/g_{i,j}) \qquad (7)$$

Another way of looking at the Kullback-Leibler "distance" (or Kullback-Leibler divergence, or KL-distance) is to say that a proportion $f_{1,\cdot} = f_{1,1}+f_{1,2}$ of the time, we have the Kullback-Leibler distance to the corresponding cross-section $(g_{1,\cdot})$ of g, and then the remaining proportion $1 - f_{1,\cdot} = f_{2,\cdot} = f_{2,1} + f_{2,2}$ of the time, we have the Kullback-Leibler distance to the corresponding (other) cross-section $(g_{2,\cdot})$ of g.

To proceed down this path, we have to do the calculations at two levels. (Analogously with sec. 3.2, we could do the calculation in one step involving four terms or break it up into two levels of parts each involving two terms.) At the top level, we have to calculate the KL-divergence from the binomial distribution $(f_{1,\cdot}, f_{2,\cdot})$ to the binomial distribution $(g_{1,\cdot}, g_{2,\cdot})$. This top-level KL-divergence is

$$f_{1,\cdot}\log(f_{1,\cdot}/g_{1,\cdot}) \;+\; f_{2,\cdot}\log(f_{2,\cdot}/g_{2,\cdot}) \qquad (8)$$

It then remains to go to the next level (or step) and first look at the binomial distribution $f_{1,1}/(f_{1,1}+f_{1,2})$ and $f_{1,2}/(f_{1,1}+f_{1,2})$ versus $g_{1,1}/(g_{1,1}+g_{1,2})$ and $g_{1,2}/(g_{1,1}+g_{1,2})$ (on the first or left branch), and then to look at the binomial distribution $f_{2,1}/(f_{2,1}+f_{2,2})$ and $f_{2,2}/(f_{2,1}+f_{2,2})$ versus $g_{2,1}/(g_{2,1}+g_{2,2})$ and $g_{2,2}/(g_{2,1}+g_{2,2})$ (on the second or right branch). Note that the first of these (KL-divergences or) coding inefficiencies will only occur a proportion $f_{1,1}+f_{1,2} = f_{1,\cdot}$ of the time, and the second of these (KL-divergences or) coding inefficiencies will only occur a proportion $f_{2,1}+f_{2,2} = f_{2,\cdot} = 1 - f_{1,\cdot}$ of the time.

The first of these KL-divergences is the (expected or) average coding inefficiency when encoding $(f_{1,1}/(f_{1,1}+f_{1,2}), f_{1,2}/(f_{1,1}+f_{1,2}))$ not using itself, but rather instead (sub-optimally) using $(g_{1,1}/(g_{1,1}+g_{1,2}), g_{1,2}/(g_{1,1}+g_{1,2}))$. This first KL-divergence is

$$f_{1,1}/(f_{1,1}+f_{1,2})\log((f_{1,1}/(f_{1,1}+f_{1,2}))/(g_{1,1}/(g_{1,1}+g_{1,2})))$$
$$+\quad f_{1,2}/(f_{1,1}+f_{1,2})\log((f_{1,2}/(f_{1,1}+f_{1,2}))/(g_{1,2}/(g_{1,1}+g_{1,2})))$$

$$\begin{aligned}
&= (f_{1,1}/(f_{1,1}+f_{1,2})) \times ((\log f_{1,1}/g_{1,1}) - (\log((f_{1,1}+f_{1,2})/(g_{1,1}+g_{1,2}))))\\
&+ (f_{1,2}/(f_{1,1}+f_{1,2})) \times ((\log f_{1,2}/g_{1,2}) - (\log((f_{1,1}+f_{1,2})/(g_{1,1}+g_{1,2}))))\\
&= (f_{1,1}/f_{1,\cdot}) \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot}))\\
&+ (f_{1,2}/f_{1,\cdot}) \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot})) \qquad (9)
\end{aligned}$$

Changing the very first subscript from a 1 to a 2, we then get that the second of these KL-divergences, namely the KL-divergence from the binomial distribution $(f_{2,1}/(f_{2,1}+f_{2,2})), f_{2,2}/(f_{2,1}+f_{2,2}))$ to the binomial distribution $(g_{2,1}/(g_{2,1}+g_{2,2})), g_{2,2}/(g_{2,1}+g_{2,2}))$, is

$$\begin{aligned}
&f_{2,1}/(f_{2,1}+f_{2,2}) \log((f_{2,1}/(f_{2,1}+f_{2,2}))/(g_{2,1}/(g_{2,1}+g_{2,2})))\\
&+ f_{2,2}/(f_{2,1}+f_{2,2}) \log((f_{2,2}/(f_{2,1}+f_{2,2}))/(g_{2,2}/(g_{2,1}+g_{2,2})))\\
&= (f_{2,1}/(f_{2,1}+f_{2,2})) \times ((\log f_{2,1}/g_{2,1}) - (\log((f_{2,1}+f_{2,2})/(g_{2,1}+g_{2,2}))))\\
&+ (f_{2,2}/(f_{2,1}+f_{2,2})) \times ((\log f_{2,2}/g_{2,2}) - (\log((f_{2,1}+f_{2,2})/(g_{2,1}+g_{2,2}))))\\
&= (f_{2,1}/f_{2,\cdot}) \times (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot}))\\
&+ (f_{2,2}/f_{2,\cdot}) \times (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \qquad (10)
\end{aligned}$$

The first coding inefficiency, or KL-divergence, given in equation (9), occurs a proportion $f_{1,\cdot} = (f_{1,1}+f_{1,2})$ of the time. The second coding inefficiency, or KL-divergence, given in equation (10), occurs a proportion $f_{2,\cdot} = (f_{2,1}+f_{2,2}) = (1-(f_{1,1}+f_{1,2})) = 1 - f_{1,\cdot}$ of the time.

So, the total expected (or average) coding inefficiency of using g when we should be using f, or equivalently the KL-divergence from f to g, is the following sum: the inefficiency from equation (8) $+$ ($f_{1,\cdot} \times$ (the inefficiency from equation (9))) $+$ ($f_{2,\cdot} \times$ (the inefficiency from equation (10))).

Before writing out this sum, we note that

$$\begin{aligned}
&(f_{1,\cdot} \times \text{(the inefficiency from equation (9))})\\
&= f_{1,1} \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot}))\\
&+ f_{1,2} \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot})) \qquad (11)
\end{aligned}$$

and similarly that

$$\begin{aligned}
&(f_{2,\cdot} \times \text{(the inefficiency from equation (10))})\\
&= f_{2,1} \times (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot}))\\
&+ f_{2,2} \times (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \qquad (12)
\end{aligned}$$

Now, writing out this sum, summing equations (8), (11) and (12), it is

$$\begin{aligned}
&f_{1,\cdot} \log(f_{1,\cdot}/g_{1,\cdot}) + f_{2,\cdot} \log(f_{2,\cdot}/g_{2,\cdot})\\
&+ f_{1,1} \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot}))\\
&+ f_{1,2} \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot}))
\end{aligned}$$

$$
\begin{aligned}
&+ \quad f_{2,1} \ \times \ (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
&+ \quad f_{2,2} \ \times \ (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
&= \quad f_{1,1} \log(f_{1,1}/g_{1,1}) + f_{1,2} \log(f_{1,2}/g_{1,2}) \\
&\quad +f_{2,1} \log(f_{2,1}/g_{2,1}) + f_{2,2} \log(f_{2,2}/g_{2,2}) \\
&= \quad \sum_{i=1}^{2} \sum_{j=1}^{2} f_{i,j} \log(f_{i,j}/g_{i,j}) \\
&= \quad \sum_{i=1}^{2} \sum_{j=1}^{2} f_{i,j} (\log f_{i,j} - \log g_{i,j}) \ = \ KL(f,g) \ = \ \Delta(g||f) \qquad (13)
\end{aligned}
$$

thus reducing to our very initial expression (7).

Two special cases are worth noting. The first special case is when events are independent — and so $f_{i,j} = f_{i,\cdot} \times \phi_j = f_i \times \phi_j$ and $g_{i,j} = g_{i,\cdot} \times \gamma_j = g_i \times \gamma_j$ for some ϕ_1, ϕ_2, γ_1 and γ_2 (and $f_1 + f_2 = 1$, $g_1 + g_2 = 1$, $\phi_1 + \phi_2 = 1$ and $\gamma_1 + \gamma_2 = 1$). In this case, following from equation (7), we get

$$
\begin{aligned}
KL(f,g) \quad &= \quad \Delta(g||f) = \sum_{i=1}^{2} \sum_{j=1}^{2} f_{i,j} \log(f_{i,j}/g_{i,j}) = \sum_{i=1}^{2} \sum_{j=1}^{2} f_i \phi_j \log(f_i \phi_j / (g_i \gamma_j)) \\
&= \quad \sum_{i=1}^{2} \sum_{j=1}^{2} (f_i \phi_j \log(f_i/g_i) + f_i \phi_j \log(\phi_j/\gamma_j)) \\
&= \quad (\sum_{i=1}^{2} f_i \log(f_i/g_i)) \ + \ (\sum_{j=1}^{2} \phi_j \log(\phi_j/\gamma_j)) \qquad (14) \\
&= \quad (\sum_{i=1}^{2} f_i \log(1/g_i)) \ + \ (\sum_{j=1}^{2} \phi_j \log(1/\gamma_j)) \\
&\quad - (\sum_{i=1}^{2} \sum_{j=1}^{2} -f_i \phi_j \log(f_i \phi_j)) \qquad (15) \\
&= \quad (\sum_{i=1}^{2} f_i \log(1/g_i)) \ + \ (\sum_{j=1}^{2} \phi_j \log(1/\gamma_j)) \ - \ \text{(Entropy of f)} \qquad (16)
\end{aligned}
$$

We observe in this case of the distributions being independent that the Kullback-Leibler scores de-couple in exactly the same uniquely invariant way as they do for the probabilistic predictions in sec. 3.2.

The second special case of particular note is when probabilities are correct in both branching paths — i.e., $f_{1,1}/f_{1,\cdot} = g_{1,1}/g_{1,\cdot}$, $f_{1,2}/f_{1,\cdot} = g_{1,2}/g_{1,\cdot}$, $f_{2,1}/f_{2,\cdot} = g_{2,1}/g_{2,\cdot}$ and $f_{2,2}/f_{2,\cdot} = g_{2,2}/g_{2,\cdot}$. In this case, starting from equation (7), we get

$$
KL(f,g) \quad = \quad \Delta(g||f) = \sum_{i=1}^{2} \sum_{j=1}^{2} f_{i,j} \log(f_{i,j}/g_{i,j})
$$

$$
\begin{aligned}
&= \sum_{i=1}^{2}\sum_{j=1}^{2} f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log((f_{i,\cdot}(f_{i,j}/f_{i,\cdot}))/(g_{i,\cdot}(g_{i,j}/g_{i,\cdot})))\\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log((f_{i,\cdot}/g_{i,\cdot}) \times [(f_{i,j}/f_{i,\cdot})/(g_{i,j}/g_{i,\cdot})])\\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log(f_{i,\cdot}/g_{i,\cdot})\\
&= \sum_{i=1}^{2} f_{i,\cdot} \log(f_{i,\cdot}/g_{i,\cdot}) = \sum_{i=1}^{2} f_{i,\cdot} \log(1/g_{i,\cdot}) - (\text{Entropy of f}) \quad (17)
\end{aligned}
$$

We observe in this second special case of probabilities being correct in both branching paths that the divergence between the two distributions is the same as that between the binomial distributions ($f_{1,\cdot}$, $f_{2,\cdot} = 1 - f_{1,\cdot}$) and ($g_{1,\cdot}$, $g_{2,\cdot} = 1 - g_{1,\cdot}$), exactly as it should and exactly as it would be for the uniquely invariant log(arithm)-loss ("bit cost") scoring of the probabilistic predictions in sec. 3.2.

Like the invariance of the log(arithm)-loss scoring of probabilistic predictions under re-framing (whose uniqueness is introduced in [Dowe, 2008a, footnote 175 (and 176)] and discussed in [Dowe, 2008b, pp. 437–438] and sec. 3.2), this invariance of the Kullback-Leibler divergence to the re-framing of the problem is [Dowe, 2008b, p. 438] due to the fact(s) that (e.g.) $\log(f/g) = \log f - \log g$ and $-\log(f_{1,1}/(f_{1,1} + f_{1,2})) + \log f_{1,1} = \log(f_{1,1} + f_{1,2})$.

A few further comments are warranted by way of alternative measures of "distance" or divergence between probability distributions. First, where one can define the notion of the distance remaining invariant under re-framing for these, the Bhattacharyya distance, Hellinger distance and Mahalanobis distance are all not invariant under re-framing. Versions of distance or divergence based on the Rényi entropy give invariance in the trivial case that $\alpha = 0$ (where the distance will always be 0) and the case that $\alpha = 1$ (where we get the Shannon entropy and, in turn, the Kullback-Leibler distance that we are currently advocating).

A second further comment is that, just as in [Tan and Dowe, 2006, sec. 4.2], we can also add a term corresponding to (a multiple of) the entropy of the Bayesian prior. Just like the Kullback-Leibler divergence (and any multiple of it), this (and any multiple of it) will also remain invariant under re-parameterisation or other re-framing [Dowe, 2008a, footnote 176; 2008b, p. 438].

A third — and important — further comment is that it is not just the Kullback-Leibler distance from (say, the true distribution) f to (say, the inferred distribution) g, $KL(f, g) = \Delta(g||f)$, that is invariant and appears to be uniquely invariant under re-framing, but clearly also $KL(g, f) = \Delta(f||g)$ is invariant, as will also be a sum of any linear combination of these, such as (e.g.) $\alpha KL(f, g) + (1-\alpha)KL(g, f)$ (with $0 \leq \alpha \leq 1$, although this restriction is not required for invariance) [Dowe, 2008b, p. 438]. The case of $\alpha = 1/2$ gives the symmetric Kullback-Leibler distance.

The notion of Kullback-Leibler distance can be extended quite trivially — as

above — to hybrid continuous and discrete Bayesian net graphical models [Tan and Dowe, 2006, sec. 4.2; Dowe 2008a, sec. 0.2.5; 2008b, p. 436] (also see sec. 7.6) or mixture models [Dowe, 2008b, p. 436], etc. For the hybrid continuous and discrete Bayesian net graphical models in [Comley and Dowe, 2003; 2005] (which resulted at least partly from theory advocated in [Dowe and Wallace, 1998]), the log-loss scoring approximation to Kullback-Leibler distance has been used [Comley and Dowe, 2003, sec. 9].

4 OCKHAM'S RAZOR (AND MISUNDERSTANDINGS) AND MML

Let us recall Minimum Message Length (MML) from secs. 1 and 2.4, largely so that we can now compare and contrast MML with Ockham's razor (also written as Occam's razor). Ockham's razor, as it is commonly interpreted, says that if two theories fit the data equally well then prefer the simplest (e.g., [Wallace, 1996b, sec. 3.2.2, p. 48, point b]). Re-phrasing this in statistical speak, if $Pr(D|H_1) = Pr(D|H_2)$ and $Pr(H_1) > Pr(H_2)$, then Ockham's razor advocates that we prefer H_1 over H_2 — as would also MML, since $Pr(H_1)Pr(D|H_1) > Pr(H_2)Pr(D|H_2)$. It is not clear what — if anything — Ockham's razor says in the case that $Pr(H_1) > Pr(H_2)$ but $Pr(D|H_1) < Pr(D|H_2)$, although MML remains applicable in this case by comparing $Pr(H_1) \times Pr(D|H_1)$ with $Pr(H_2) \times Pr(D|H_2)$. In this sense, I would at least contend that MML can be thought of as a generalisation of Ockham's razor — for MML tells us which inference to prefer regardless of the relationships of $Pr(H_1)$ with $Pr(H_2)$ and $Pr(D|H_1)$ with $Pr(D|H_2)$, but it is not completely clear what Ockham's razor *per se* advocates unless we have that $Pr(D|H_1) = Pr(D|H_2)$.

Our earlier arguments (e.g., from sec. 1) tell us why the MML theory (or hypothesis) can be thought of as the most probable hypothesis. Informal arguments of Chris Wallace's from [Dowe, 2008a, footnote 182] (in response to questions [Dowe and Hajek, 1997, sec. 5.1; 1998, sec. 5]) suggest that, if $Pr(D|H_1) = Pr(D|H_2)$ and $Pr(H_1) > Pr(H_2)$, then we expect H_1 to be a better predictor than H_2. But, in addition, there is also an alternative, more general and somewhat informal argument for, in general, preferring the predictive power of one hypothesis over another if the former hypothesis leads to a shorter two-part message length. This argument is simply that the theory of shorter two-part message length contributes more greatly (i.e., has a greater Bayesian weighting) in the optimal Bayesian predictor. In particular, the MML model will essentially have the largest weight in the predictor. In those cases where the optimal Bayesian predictor is statistically consistent (i.e., converges to any underlying data when given sufficient data), the optimal Bayesian predictor and the MML hypothesis appear always to converge.

Several papers have been written with dubious claims about the supposed ineffectiveness of MML and/or of Ockham's razor. Papers using inefficient (Minimum Description Length [MDL] or MML) coding schemes lead quite understandably to sub-optimal results — but a crucial point about minimum *message length* is to make sure that one has a reliable message length (coding scheme) before one

sets about seeking the minimum of this "message length". For corrections to such dubious claims in such papers by using better coding schemes to give better results (and sometimes vastly better coding schemes to get vastly better results), see (e.g.) examples [Wallace and Dowe, 1999a, secs. 5.1 and 7; 1999c, sec. 2; Wallace, 2005, sec. 7.3; Viswanathan *et al.*, 1999; Wallace and Patrick, 1993; Comley and Dowe, 2005, secs. 11.3 and 11.4.3; Needham and Dowe, 2001; Wallace, 2005, sec. 5.1.2; Grünwald, 2007, sec. 17.4, An Apologetic Remark; Dowe, 2008a, p. 536] such as those in [Dowe, 2008a, footnote 18]. Not unrelatedly, a variety of misconceptions have led a variety of authors to make ill-founded criticisms of Ockham's razor. One (such) interpretation (I think I should say, *mis*interpretation) of Ockham's razor seems to go along the lines that Ockham's razor supposedly advocates the simplest hypothesis, regardless of any data — and so (e.g.) DNA should supposedly be shaped in a single-helix rather than a double-helix. [And it seems a pity to misinterpret Ockham's razor so — especially in a biological framework — because interpreting Ockham's razor more properly using MML enables us to make a strong case that proteins fold with the Helices (and Extendeds) forming first and then the "Other" turn classes forming subsequently to accommodate these structures [Edgoose *et al.*, 1998, sec. 6; Dowe *et al.*, 1996, sec. 5, p. 253] (see also [Dowe *et al.*, 1995]) [Wallace, 1998a, sec. 4.2; Dowe, 2008a, footnote 85; 2008b, p. 454].

What seems like a variation of this misconception is an argument in one paper that if we fit data from within some model family (such as fitting the data with a decision tree) and then subsequently find that a more complicated model predicts better, then this is somehow supposedly empirical evidence against Ockham's razor. (See also a comment here from [Jorgensen and Gentleman, 1998, Some Criticisms].)

Using MML as our (more general) form of Ockham's razor, these supposed criticisms based on using overly simple models and paying insufficient attention to the data seem somewhat silly. For a discussion of the adverse consequences of not giving equals weights to the lengths of the two parts of an MML message, see, e.g., [Dowe, 2008a, footnote 130].

For those who would like every function — both the simple functions and the more complicated functions — to have the same prior probability, not only does this seem counter-intuitive, but — furthermore — it is not possible when there are infinitely many theories. When there are infinitely many theories, it necessarily follows that, as we look at progressively more and more complicated theories, it must necessarily follow that the prior probability must tend asymptotically to zero so that the countable sum over the prior probabilities of the theories can equal 1 (or unity).

Another criticism of the Bayesian approach — and therefore of our Bayesian MML interpretation of Ockham's razor — is that this approach can be undone by a pathological (sabotaging) form of prior. If we look at Bayes's theorem (from sec. 1) and its consequences in terms of MML, we see what our intuition tells us — that we get a reasonable posterior distribution over the hypotheses if we start off with a

reasonable prior distribution. Our Bayesian priors (as we use them in problems of inference) should be somewhere between what we genuinely suspect *a priori* and (partly politically, so that we are less open to being accused of fudging our results, and perhaps partly to protect ourselves from ourselves) something innocuous (and seemingly "objective").

For problems where the number of parameters is bounded above — or grows sufficiently slowly when the amount of data increases — Bayesian inference will converge to the underlying model given sufficient data. To criticise Bayesian MML and/or Ockham's razor after being sabotaged by a counter-intuitive and misrepresentative pathological prior is somewhat akin to criticising any inference method when the bulk of the relevant explanatory variables are not made available (or at least not made available until after much data has been seen) but in their stead is a plethora of essentially irrelevant variables.

4.1 Inference (or explanation) and prediction

Inference — also variously known as *explanation* [Wallace, 2005, sec. 1.1, first sentence and sec. 1.5], *induction* and/or *inductive inference* — pertains to finding the single best explanation for a body of data. *Prediction* pertains to the activity of anticipating the future, whether this is done using a single inference or a combination of more than one inference. To give an example, someone doing inference would be interested in a model of stock market prices which gives a theory of how the stock market works. An investor would certainly find that useful, but an investor would perhaps be more interested in whether prices are expected to be going up or down (and a probability distribution over these events and the magnitude of movement). To give a second example [Dowe *et al.*, 2007, sec. 6.1.4], when two models of slightly different (Bayesian posterior) probabilities give substantially different answers, inference would advocate going with the more probable theory where prediction would advocate doing some sort of averaging of the theories.

In the classical (non-Bayesian) approach, inference and prediction are perhaps the same thing. Certainly, an inference can be used to predict — and, to the classically (non-Bayesian) minded, prediction seems to be done by applying the single best inference. But, to the Bayesian, the best predictor will often result from combining more than one theory [Wallace, 1996b, sec. 3.6, p. 55; Oliver and Hand, 1996; Wallace and Dowe, 1999a, sec. 8; Tan and Dowe, 2006].

Herein lies a difference between the predictive approach of Solomonoff and the MML inductive inference approach of Wallace from sec. 2.4. By taking the single best theory, MML is doing induction. Despite the potentially confusing use of the term "Solomonoff induction" by some others, Solomonoff (is not doing induction [and not really inductive inference *per se*, either] but rather) is doing prediction [Solomonoff, 1996; Wallace, 1996b, sec. 3.6, p. 55; 2005, sec 10.1].

On the relative merits of induction (or [inductive] inference) vs prediction, there can be no doubting that humans acknowledge and reward the intelligence behind

inductive inferences. When we ask for a list of great human intellects, whoever else is on the list, there will be people who have made prominent inductive inferences. Examples of such people and theories include Isaac Newton for the theory of gravity, Charles Darwin and Alfred Russel Wallace for the theory of natural selection and evolution, Albert Einstein for the theories of special and general relativity, Alfred Wegener for the theory of "continental drift", and countless Nobel laureates and/or others in a variety of areas for their theories [Sanghi and Dowe, 2003, sec. 5.2]. (And when a human is paid the compliment of being called "perceptive", my understanding of this term is that one thing that is being asserted is that this "perceptive" person is good at making inductive inferences about human behaviour.) Of course, those such theories as are accepted and whose developers are rewarded usually are not just good pieces of induction but typically also lead to good predictions. And whether or not predictions are done with the single best available theory or with a combination of theories, people are certainly interested in having good predictors.

In trying to re-construct or restore a damaged image, the argument in support of inference is that we clearly want the single best inference rather than a probability distribution over all possible re-constructions. On the other hand, if there are a few inferences almost as good as the best (MML) inference, we would also like to see these alternative models [Dowe, 2008a, sec. 0.3.1]. Let me now essentially repeat this example but modify its context. If you think that the sentence you are currently reading is written in grammatically correct unambiguous English (and that you correctly understand the author's intended meaning), then you are using several little innocuous inferences — such as (e.g.) the author and you (the reader) have at least sufficiently similar notions of English-language word meaning, English-language grammar and English-language spelling. However, if the writing were smudged, the spelling was questionable and the grammar and punctuation were poor (several of which can happen with the abbreviated form of a telegram, some e-mails or a mobile text message), inference would advocate going with the single best interpretation. A related case in point is automatic (machine) translation. Whether for the smudged poorly written (ambiguous) sentence or for automatic translation, prediction (in its pure platonic form) would advocate having a probability distribution over all interpretations. In reality, if there is one outstandingly clear interpretation to the sentence, then someone doing prediction would most probably be satisfied with this interpretation, (as it were) "beyond a reasonable doubt". But, as with the damaged image, if there are a few inferences almost as good as the best (MML) inference, we would again also like to see these alternative models.

The distinction between explanation (or inference, or inductive inference, or induction) and prediction is something which at least some other authors are aware of [Wallace and Dowe, 1999a, sec. 8; Wallace 2005, sec. 10.1.2; Shmueli and Koppius, 2007, Dowe *et al.*, 2007, secs. 6.1.4, 6.3 and 7.2; Dowe, 2008b, pp. 439–440], and we believe that both have their place [Dowe *et al.*, 2007, sec. 6.3]. Whether or not because of our newly discussed uniqueness in invariance properties of Kullback-

Leibler distance (from [Dowe, 2008a, p. 438] and sec. 3.6), some authors regard prediction as being about minimising the expected log-likelihood error — or equivalently, minimising the expected Kullback-Leibler distance between the true model (if there is one) and the inferred model. While the reasons (of many) for doing this might be (more) about minimising the expected log-likelihood error, the uniqueness in invariance properties of Kullback-Leibler distance suggest it is certainly a worthy interpretation of the term "*prediction*" and that doing prediction this way is worthy of further investigation.

Recalling the invariance of the Kullback-Leibler distance (from, e.g., sec. 3.6), taking the Bayesian approach to minimising the expected Kullback-Leibler distance will be invariant under re-parameterisation (e.g., from polar to Cartesian co-ordinates) [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2]. Recalling α at the very end of sec. 3.6 [from the expression $\alpha KL(f,g) + (1-\alpha)KL(g,f) = \alpha\Delta(g||f) + (1-\alpha)\Delta(f||g)$], the extreme of $\alpha = 1$ sees us choose a function (g) so that the *expected* coding inefficiency of using our function (g) rather than the (ideal) truth (true function, f) is minimised, weighting over our posterior distribution on f; and the other extreme of $\alpha = 0$ sees us choose a function (g) so that (under the hypothetical assumption that the data were being sampled from distribution, g) the expected inefficiency of using a function (f) sampled from the (actual) Bayesian posterior rather than using our function (g) is minimised. Although both of these are statistically invariant, convention is that we are more interested in choosing a function of minimal expected coding inefficiency relative to the (ideal) truth (true function) — equivalently minimising the expected log-likelihood error (and hence choosing $\alpha = 1$).

As a general rule of thumb, the MML estimator lies between the Maximum Likelihood estimator (which is given to over-fitting) on the one hand and the Bayesian minEKL estimator (which is, curiously, given to under-fitting) on the other hand [Wallace, 2005, secs. 4.7–4.9]. (Wallace makes an excellent intuitive case for this in [Wallace, 2005, sec. 4.9].) Four examples of this are the multinomial distribution, the Neyman-Scott problem (see sec. 6.4) [Wallace, 2005, sec. 4.2–4.8], the "gap or no gap" ("gappy") problem [Dowe *et al.*, 2007, sec. 6.2.4 and Appendix B] and the bus number problem [Dowe, 2008a, footnote 116; 2008b, p. 440]. We outline these below, and then mention at the end not yet totally explored possible fifth and sixth (which would probably begin from [Schmidt and Makalic, 2009b]) examples.

For the multinomial distribution, with counts $s_1, ..., s_m, ..., s_M$ in classes $1, ..., m, ..., M$ respectively and $S = s_1 + ... + s_m + ... + s_M$, Maximum Likelihood gives $\hat{p}_m = s_m/S$. With a uniform prior, the minEKL estimator (also known as the Laplace estimate or the posterior mean) is $(s_m+1)/(S+M)$, whereas the Wallace-Freeman MML approximation [Wallace and Freeman, 1987; Wallace, 2005, sec. 5.4] with this same prior is $(\hat{p}_m)_{MML} = (s_m + 1/2)/(S + M/2)$.

For the particular case of the "gap or no gap" ("gappy") problem [Dowe *et al.*, 2007, sec. 6.2.4 and Appendix B], data ($\{x_i : 0 \leq x_i \leq 1, i = 1, ..., N\}$ for

increasing N) are being generated uniformly either from the closed interval $[0, 1]$ or from a sub-region $[0, a] \cup [b, 1]$ for some a and b such that $a < b$. We see Maximum Likelihood and Akaike's Information Criterion (AIC) over-fitting here, surmising a gap even whether there isn't one. At the other extreme, we see the minEKL estimator *under*-fitting, stating no gap even in extreme cases such as (e.g.) $[0, 0.001] \cup [0.999, 1]$ with $a = 0.001$ and $b = 0.999$. The curious behaviour of minEKL is due to the fact that the posterior probability that the region is $[0, 1]$ will get arbitrarily small for large N but never down to 0, and there is an infinite penalty in Kullback-Leibler distance for ascribing a probability of 0 to something which can actually happen. Unlike the over-fitting Maximum Likelihood and AIC, and unlike the under-fitting minEKL, MML behaves fine in both cases [Dowe *et al.*, 2007].

For the Neyman-Scott problem (of sec. 6.4), see [Wallace, 2005, sec. 4.2–4.8]. For the bus number problem [Dowe, 2008a, footnote 116; 2008b, p. 440] (where we arrive in a new town with θ buses numbered consecutively from 1 to θ, and we see only one bus and observe its number, x_{obs}, and are then asked to estimate the number of buses in the town), the Maximum Likelihood estimate is the number of the observed bus, x_{obs}, which is an absolute lower bound and seems like a silly under-estimate. At the other extreme, minEKL will behave in similar manner to how it did with the abovementioned "gappy" problem. It will choose the largest positive integer (no less than x_{obs}) for which the prior (and, in turn, the posterior) is non-zero. In the event that the prior never goes to 0, it will return infinity. It seems fairly trivial that the MML estimate must fall between the Maximum Likelihood estimate (the lowest possible value) and the minEKL estimate (from a Bayesian perspective, the highest possible estimate). For further discussion of the behaviour of MML here, see [Dowe, 2008a, footnote 116]. In addition to the four examples we have just given of the MML estimate lying between (over-fitting) Maximum Likelihood and (under-fitting) minEKL, of possible interest along these lines as potential fifth and sixth examples worthy of further exploration, see sec. 6.5 on panel data (as a probable fifth example) and the treatment of MML shrinkage estimation in [Schmidt and Makalic, 2009b].

5 DESIDERATA: STATISTICAL INVARIANCE, STATISTICAL CONSISTENCY, EFFICIENCY, SMALL-SAMPLE PERFORMANCE, ETC.

In this section, we look at several *desiderata* — or properties that we might desire — from statistical estimators.

5.1 Statistical invariance

Statistical invariance [Wallace, 2005, sec. 5.2; Dowe *et al.*, 2007, sec. 5.3.2; Dowe, 2008b, p. 435] says, informally, that we get the same answer no matter how we phrase the problem.

So, if we know that the relationship between the area A of a circle and its radius r is given by $A = \pi r^2$ (and, equivalently, $r = \sqrt{A/\pi}$), then statistical invariance requires that our estimate of the area is π times our the square of our estimate of the radius. The estimator function is often denoted by a hat (or circumflex), $\hat{}$, above. So, for a circle, statistical invariance in the estimator would require that $\hat{A} = \pi \hat{r}^2$.

If we replace r by κ in the Cartesian and polar co-ordinates example from sec. 3.6, ($\kappa = \text{sign}(x) \, . \, \sqrt{x^2 + y^2}$, $\theta = \tan^{-1}(y/x)$) and ($x = \kappa \cos\theta$, $y = \kappa \sin\theta$). If we are estimating the strength and direction (κ, θ) of a magnetic field or equivalently the x and y co-ordinates (x, y) [Wallace and Dowe, 1993], then statistical invariance requires that $\hat{x} = \hat{\kappa} \cos\theta$, $\hat{y} = \hat{\kappa} \sin\theta$.

Statistical invariance is surely an aesthetic property of an estimate. In many problems, we are not committed to only one parameterisation - and, in those cases, statistical invariance is more useful than a simple aesthetic nicety.

Maximum Likelihood, Akaike's Information Criterion, Strict Minimum Message Length (SMML) [Wallace and Boulton, 1975; Wallace, 2005, chap. 3] and many MML approximations [Wallace and Freeman, 1987; Wallace, 2005, chaps. 4–5; Dowe 2008a, sec. 0.2.2 and footnote 159; Schmidt, 2008; Dowe, 2008b, p. 438 and p. 451] are statistically invariant, but there do exist approaches — such as the Bayesian Maximum A Posteriori (MAP) approach [Oliver and Baxter, 1994; Dowe *et al.*, 1996e; 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 *et al.*, 2007, sec. 5.1, coding prior; Dowe 2008a, footnote 158; 2008b, p. 443 and pp. 448–449] — which are not statistically invariant.

5.2 Statistical consistency

For those who like collecting larger and larger data sets in the hope and belief that this will bring us closer and closer to whatever model or process underlies the data, statistical consistency — the notion of getting arbitrary close to any true underlying model given sufficient data [Dowe *et al.*, 2007, secs. 5.3.4, 6.1.3 and later; Dowe 2008b, pp. 436–437] — is of paramount importance.

More formally, we might write it as [Dowe, 2008b, p. 436]

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

and we could even venture to write it (in a parameterisation-invariant way) as (e.g.)

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

Of course, as highlighted by Grünwald and Langford [2004; 2007], cases of model misspecification do occur. In other words, it might be that the true model θ (if there is one) is not contained in the family (or class) of models over which we conduct our search for $\hat{\theta}$. In such cases, we can modify (or generalise) the notion

of statistical consistency to be that (as implicitly described in [Dowe, 2008a, sec. 0.2.5, p. 540, col. 1]), as we get more and more data, we get the Kullback-Leibler distance arbitrarily close to that of the closest available member in our model space. Or, more formally,

$$\forall\theta \;\; \forall\epsilon > 0 \;\; \exists N_0 \;\; \forall N \geq N_0 \;\; Pr(\Delta(\hat{\theta}||\theta) = KL(\theta,\hat{\theta}) < KL(\theta,\hat{\theta}_{\rm best}) + \epsilon) > 1 - \epsilon,$$

where $\hat{\theta}_{\rm best}$ is as close as one can get in Kullback-Leibler distance to θ from within the space of models being considered.

I should (and do now) qualify this slightly. It is possible that $\hat{\theta}_{\rm best}$ might not exist in the same sense that there is no number in the list $1, 1/2, 1/3, 1/4, ...$ which is the list's smallest element. One of a few ways of dealing with this is simply to replace the start of the above with "$\forall\theta'$ in our model space" and to replace the finish of the above with "$Pr(KL(\theta,\hat{\theta}) < KL(\theta,\theta') + \epsilon) > 1 - \epsilon$", thus now making it

$$\begin{array}{l} \forall\theta' \text{ in our model space} \\ \forall\epsilon > 0 \;\; \exists N_0 \;\; \forall N \geq N_0 \;\; Pr(KL(\theta,\hat{\theta}) < KL(\theta,\theta') + \epsilon) > 1 - \epsilon. \end{array}$$

(As a second point, replacing the first ϵ by $\epsilon/2$ does not change the semantics of the definition. Similarly, replacing one of the ϵ terms by a δ and adding a quantifier $\forall\delta > 0$ out front also does not change the semantics, as we can (e.g.) re-set $\epsilon' = \min\{\delta,\ \epsilon\}$.)

With an eye to secs. 6.4 and 6.5 and this issue of statistical consistency under misspecification, it is worth bearing in mind that — even though there is misspecification — the class (or family) of models over which we conduct our search might be dense in the space of possible models. In other words, if you have a non-negative valued function (or probability distribution) on the real line which integrates to 1 (and can't be written as a finite mixture model), it can still be possible to find a sequence of finite (Gaussian) mixture models which fit it arbitrarily closely. (For reading on mixture models, see, e.g., [Jorgensen and McLachlan, 2008].)

5.3 Efficiency, small-sample performance, other considerations, etc.

The notion of *efficiency* is perhaps ambiguous in that it has been used in the literature with at least two different meanings. On the one hand, *efficiency* has been taken to mean that the message length calculations and approximations are both optimal or near-optimal (with $l_i \approx -\log p_i$) [Wallace, 2005, sec. 5.2.4]. On the other hand, *efficiency* of an estimator has been taken to mean the speed with which that estimator converges to the true model generating the data as the amount of data increases [Dowe *et al.*, 2007, secs. 5.3.4 and 8].

While these two notions are different, it should be pointed out that, insofar as reliable MML coding schemes lead to good inferences and less reliable coding schemes lead to less reliable inferences [Quinlan and Rivest, 1989; Wallace and

Patrick, 1993, Kearns *et al.*, 1997; Viswanathan *et al.*, 1999; Wallace, 2005, sec. 7.3; Murphy and Pazzani, 1994; Needham and Dowe, 2001; Wallace and Dowe, 1999a, secs. 5.1 and 7; 1999c, sec. 2; Comley and Dowe, 2005, secs. 11.3 and 11.4.3; Wallace, 2005, sec. 5.1.2; Dowe, 2008a, footnote 18], the two notions are very related.

As well as the notions of statistical invariance (from sec. 5.1), statistical consistency (from sec. 5.2) and efficiency, there are also issues of performing well on small-sample sizes [Dowe, 2008b, p. 436 and p. 456]. The issue of which likelihood function(s), sample size(s), parameterisation(s), Bayesian prior(s) and protocol(s) (or which parts of LNPPP-space) are important when comparing the efficacy of two estimators is discussed in [Dowe, 2008a, sec. 0.2.7, pp. 543–544].

6 MINIMUM MESSAGE LENGTH (MML) AND STRICT MML

As in sec. 1, historically, the seminal Wallace and Boulton paper [1968] came into being from Wallace's and Boulton's finding that the Bayesian position that Wallace advocated and the information-theoretic (conciseness) position that Boulton advocated turned out to be equivalent [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546 and footnote 213]. After several more MML writings [Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1] (and an application paper [Pilowsky *et al.*, 1969], and at about the same time as David Boulton's PhD thesis [Boulton, 1975]), their paper [Wallace and Boulton, 1975, sec. 3] again emphasises the equivalence of the probabilistic and information-theoretic approaches. (Different but not unrelated histories are given by Solomonoff [1997a] and a review of much later work by Kontoyiannis [2008]. For those interested in the formative thinking of Wallace (and Boulton) leading up to the seminal Wallace and Boulton MML paper [1968], see evidence of the young Bayesian (but pre-MML) Wallace in his mid-20s in the 1950s [Brennan, 2008, sec. 4; Brennan *et al.*, 1958, Appendix] and see Wallace's accounts of his early discussions with David Boulton [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546, col. 2 and footnote 213 (and sec. 1)] which resulted in [Wallace and Boulton, 1968]. If you can obtain it, then I also commend [Wallace, 1992] for background.)

As in sec. 1 and following the principles of information theory from sec. 2.1, given data D, we wish to choose a hypothesis H so as to minimise the length of a two-part message conveying H (in part 1) followed (in part 2) by D given H. The length of this message is

$$-\log Pr(H) \; - \; \log Pr(D|H).$$

A one-part form of the message was examined in [Boulton and Wallace, 1969], but various pieces of theory and practice (e.g., [Barron and Cover, 1991]) point to the merits of the two-part form of the message.

We now point to the Strict Minimum Message Length (SMML) formulation from Wallace and Boulton [1975] in sec. 6.1, and then go on to talk about some

"MML" approximations to SMML, some conjectures about the possible uniqueness of Strict MML in being both statistically invariant and statistically consistent for certain classes of problems, and some applications of MML to a variety of problems in inference and other areas in science and philosophy.

6.1 Strict MML (SMML)

The Strict Minimum Message Length (SMML) formulation from Wallace and Boulton [Wallace and Boulton, 1975; Wallace and Freeman, 1987; Wallace, 1996c; Dowe *et al.*, 1998; Wallace and Dowe, 1999a; 1999b; Farr and Wallace, 2002; Fitzgibbon *et al.*, 2002b, Fitzgibbon, 2004; Agusta, 2005; Wallace, 2005, chap. 3; Comley and Dowe, 2005, sec. 11.2; Dowe *et al.*, 2007; Dowe, 2008a, footnotes 12, 153, 158 and 196, and sec. 0.2.2] shows how to generate a code-book whose expected two-part message length is minimised, but this turns out to be computationally intractable except in the simplest of cases — such as the binomial distribution [Farr and Wallace, 2002; Wallace, 2005, chap. 3].

Of historical interest is the fact [Dowe, 2008a, sec. 0.1, p. 524, col. 1] that, even though MML had been in print many times over since 1968 [Wallace and Boulton, 1968, p. 185, sec. 2; Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1; Boulton, 1975], referees delayed the publication of Strict MML until Wallace and Boulton [1975].

Strict MML (SMML) partitions in *data*-space and optimises a formula of the form

$$(-\sum_j (q_j \log q_j)) \qquad + \qquad (-\sum_j \sum_{i \in c_j} (q_j . \frac{r(x_i)}{q_j} . \log f(x_i|\theta_j))) \tag{18}$$

Note here first that i indexes over the data. This set must be countable, as all recorded measurements are truncated and recorded to some finite accuracy. (See [Dowe, 2008a, footnote 63] for a discussion of consequences of attempting to sidestep such an insistence.) This point established, we now assign data to groups indexed by j. The number of groups will certainly be countable (and to date I am not aware of any cases where there are infinitely many groups). Letting $h(\cdot)$ be the prior and $f(\cdot|\cdot)$ denote the statistical likelihood, $r(x_i) = \int h(\theta) f(x_i|\theta)\ d\theta$ is the marginal probability of datum x_i. Note that $r(x_i)$ is a probability and *not* a density, and also that $\sum_i r(x_i) = 1$. The term $q_j = \sum_{i \in c_j} r(x_i)$ is the amount of prior probability associated with the data group c_j.

The groups c_j form a partition of the data, with each datum being assigned to exactly one group — from which it follows that $\sum_j q_j = 1$.

For each data group c_j, we choose the estimate θ_j which maximises the weighted log-likelihood $\sum_{i \in c_j} r(x_i) \log f(x_i|\theta_j)$.

As we have written equation (18), the first term is the expected length of encoding the hypothesis (see, e.g., sec. 2.3) and the second term is the expected length

of encoding the data given this hypothesis — namely the hypothesis that datum x_i lies in group c_j with (prior) probability q_j and estimate θ_j.

The computational intractability of Strict MML (except in the simplest of cases — such as the binomial distribution [Farr and Wallace, 2002; Wallace, 2005, chap. 3]) is largely due to its discrete nature — or its being "gritty" (as Chris Wallace once put it) — requiring shuffling of data between data groups, then re-estimating the q_j and θ_j for each data group c_j, and then re-calculating the message length. The code-book with the shortest expected message length as per equation (18) is the SMML code-book, and the SMML estimator for each datum x_i is the θ_j corresponding to the group c_j to which x_i is assigned.

6.2 Strict Strict MML (SSMML)

Recall the notion of (algorithmic information theory or) Kolmogorov complexity from sec. 2.4 and sec. 1. It could be said that the relationship between Strict MML and Kolmogorov complexity [Wallace and Dowe, 1999a; Wallace, 2005, chaps. 2–3] might be slightly enhanced if we turn the negative logarithms of the probabilities from equation (18) into integer code lengths — such as would seem to be required for constructing a Huffman code from sec. 2.1 (or other fully-fledged kosher code). From [Wallace, 2005, sec. 3.4, p. 191] and earlier writings (e.g., [Wallace and Freeman, 1987]), it is clear that Wallace was aware of this issue but chose to neglect and not be distracted by it.

Although it is hard to imagine it having anything other than the most minor effect on results, we take the liberty here of introducing here what I shall call Strict Strict MML (SSMML), where the constituent parts of both the first part (currently, for each j, of length $-\log q_j$) and the second part (currently, for each j, for each i, of length $-\log f(x_i|\theta_j)$) of the message have non-negative integer lengths.

One reason for preferring Strict MML to Strict Strict MML is that, as can be seen from inspecting equation (18), the Strict MML data groups, estimates and code-book will all be independent of the base of logarithm — be it 2, 10, e or whatever — and the (expected) message length will transform in the obvious invariant way with a change of base of logarithms. However, Strict Strict MML will require an integer greater than or equal to 2 to be the base of logarithms, and will not be independent of this choice of base. The simplest response to this objection is to insist that the base of logarithms is always 2 for Strict Strict MML.

My guess is that Strict Strict MML (with base of logarithms set to 2, although any larger positive integer base should work both fine and similarly as the amount of data increases) will typically be very similar to Strict MML. By construction, Strict Strict MML (with fixed base of logarithms, such as 2) will necessarily be statistically invariant, and Strict Strict MML will presumably share statistical consistency and other desirable properties of Strict MML.

There is another issue which arises when relating Strict MML (from sec. 6.1) and Strict Strict MML to Kolmogorov complexity (or algorithmic information

theory). As per equation (18) and associated discussion(s), both Strict MML and Strict Strict MML require the calculation of the marginal probability, $r(x_i) = \int h(\theta) f(x_i|\theta)\ d\theta$, of each datum, x_i. As in sec. 6.1, these marginal probabilities are then used to calculate what we call the "*coding prior*" [Dowe *et al.*, 2007, sec. 5.1], namely the discrete set of possible estimates $\{\theta_j\}$ and their associated prior probabilities, $\{q_j\}$, with $\sum_j q_j = 1$. (Strict MML is then equivalent to using the coding prior in combination with the given statistical likelihood function and doing conventional Bayesian Maximum A Posteriori (MAP).) As per [Wallace and Dowe, 1999a] and [Wallace, 2005, chaps. 2–3], the input to the Turing machine will be of a two-part form such that the first part of this input message (which conveys the hypothesis, theory or model) programs the Turing machine (without any output being written). The second part is then input to the program resulting from the first part of the message, and this input causes the desired Data to be output. (In the example with noise in sec. 7.2, the first part of the message would encode the program together with an estimate of the noise, and the second part would encode the data with code lengths depending upon the probabilities as per sec. 2.1.)

The particular additional issue which arises when relating Strict MML and Strict Strict MML to Kolmogorov complexity (or algorithmic information theory) occurs when dealing with *universal* Turing machines (UTMs) and the Halting problem (Entscheidungsproblem) — namely, we can get lower bounds on the marginal probability ($r(x_i)$) of the various data (x_i) but, due to the Halting problem, typically for at least many values of x_i we will not be able to calculate $r(x_i)$ exactly but rather only give a lower bound. If the Turing machine (TM) representing our prior is not universal (e.g., if we restrict ourselves to the family of multivariate polynomials with one of the Bayesian priors typically used in such a case), then we can calculate $r(x_i)$ to arbitrary precision for each x_i. But if the TM representing our prior is a UTM, then we might have to live with only having ever-improving lower bounds on each of the $r(x_i)$. If we stop this process after some finite amount of time, then we should note that the coding prior corresponding to the grouping arising from Strict MML (and ditto from Strict Strict MML) would appear to have the potential to be different from the prior emanating from our original UTM. That said, if we don't go to the trouble of summing different terms contributed from different programs in the calculation of $r(x_i)$ but rather simply take the largest available such term, then we quite possibly get something very similar or identical to our intuitive notion of a two-part Kolmogorov complexity.

Finally, it is worth changing tack slightly and adding here that Strict MML is a function of the sufficient statistics in the data [Wallace, 2005, sec. 3.2.6], as also should be Strict Strict MML. When some authors talk of the Kolmogorov sufficient statistics, it is as though they sometimes forget or are unaware that sometimes — such as for the Student t distribution or the restricted cut-point segmentation problem from [Fitzgibbon *et al.*, 2002b] — the minimal sufficient statistic can be the entire data set [Comley and Dowe, 2005, sec. 11.3.3, p. 270].

6.3 Some MML approximations and some properties

Given the typical computational intractability of Strict MML from sec. 6.1 (which would only be worse for Strict Strict MML from sec. 6.2), it is customary to use approximations.

Given data, D, the MMLD (or I_{1D}) approximation [Dowe, 2008a, sec. 0.2.2; Fitzgibbon *et al.*, 2002a; Wallace, 2005, secs. 4.10 and 4.12.2 and chap. 8, p. 360; Dowe, 2008b, p. 451, eqn (4)] seeks a region R which minimises

$$-\log(\int_R h(\vec{\theta})\ d\theta) \quad - \frac{\int_R h(\vec{\theta}).\log f(\vec{D}|\vec{\theta})\ d\theta}{\int_R h(\vec{\theta})\ d\theta} \tag{19}$$

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 Dowe's MMLD approximation from eqn (19) above is the Wallace-Freeman approximation [Wallace and Dowe, 1999a, sec. 6.1.2; Wallace, 2005, chap. 5; Dowe, 2008b, p. 451, eqn (5)]

$$\begin{aligned} & -\log(h(\vec{\theta}).\frac{1}{\sqrt{\kappa_d^d\ \mathrm{Fisher}(\vec{\theta})}}) \qquad - \log f(\vec{x}|\vec{\theta}) + \frac{d}{2} \\ = \quad & -\log(h(\vec{\theta})) \ + \ L \ + \ (1/2)\log(\mathrm{Fisher}(\vec{\theta})) \ + \ (d/2)(1+\log(\kappa_d)) \end{aligned} \tag{20}$$

which was first published in the statistics literature [Wallace and Freeman, 1987].

(Digressing, note that if one approximates $-\log(1/\mathrm{Fisher}(\vec{\theta}))$ in equation (20) very crudely as $k \log N$, then equation (20) reduces to something essentially equivalent to Schwarz's Bayesian Information Criterion (BIC) [Schwarz, 1978] and Rissanen's original 1978 version of Minimum Description Length (MDL) [Rissanen, 1978], although it can be strongly argued [Wallace and Dowe, 1999a, sec. 7, p. 280, col. 2] that the $-\log(1/\mathrm{Fisher}(\vec{\theta}))$ term from equation (20) is best not idly approximated away.)

A very recent approximation certainly showing promise is due to Schmidt [2008] and in [Dowe, 2008a, footnotes 64–65]. This MMLFS estimator [Schmidt, 2008], upon close examination, would appear to be based on an idea in [Fitzgibbon *et al.*, 2002a, sec. 2, especially equation (7)] (which in turn uses Wallace's FSMML Boundary rule as from [Wallace, 1998e]) and [Fitzgibbon *et al.*, 2002b, sec. 4] (again using Wallace's FSMML Boundary rule as from [Wallace, 1998e] and [Fitzgibbon *et al.*, 2002b, sec. 3.2], but see also [Wallace, 2005, sec. 4.11]).

My MMLD estimator from equation (19) gives the message length (MsgLen) for a region. The MMLFS estimator just mentioned gives MsgLen for a point (as does the Wallace-Freeman [1987] estimator from equation (20)). Both MMLD and MMLFS are calculated using Markov Chain Monte Carlo (MCMC) methods.

These approximations above, together with the Dowe-Wallace Ideal Group (or IG) estimator [Wallace, 2005, secs. 4.1, 4.3 and 4.9; Agusta, 2005, sec. 3.3.3,

pp. 60–62; Fitzgibbon, 2004, sec. 5.2, p. 70, footnote 1; Dowe, 2008a, p. 529, col. 1 and footnote 62] and other estimators (e.g., TAIG) discussed in [Dowe, 2008a, footnotes 62–65] are all statistically invariant. Recalling the notion of statistical consistency from sec. 5.2, we now show (in secs. 6.4 and 6.5) that MML is statistically consistent where a variety of other estimators fail — either under-estimating (which is typical of most alternatives to MML) or over-estimating (which is typical of the Bayesian minEKL estimator) the degree of noise.

6.4 Neyman-Scott problem and statistical consistency

In the Neyman-Scott problem [Neyman and Scott, 1948; Dowe, 2008b, p. 453], we measure N people's heights J times each (say $J = 2$) and then infer

1. the heights $\mu_1, ..., \mu_N$ 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).

It turns out that $\hat{\sigma}^2_{\text{MaximumLikelihood}} \rightarrow \frac{J-1}{J}\sigma^2$, and so for fixed J as $N \rightarrow \infty$ we have that Maximum Likelihood is statistically inconsistent — under-estimating σ [Neyman and Scott, 1948] and "finding" patterns that aren't there. The Bayesian Maximum A Posteriori (MAP) approach (from sec. 5.1) is likewise not statistically consistent here [Dowe, 2008a, footnote 158].

Curiously, the Bayesian minimum expected Kullback-Leibler distance (minEKL) estimator [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2; 2008b, p. 444] from sec. 4.1 is also statistically inconsistent for the Neyman-Scott problem, conservatively over-estimating σ [Wallace, 2005, sec. 4.8]. Recall a discussion of this (the Neyman-Scott problem) and of the "gappy" ("gap or no gap") problem in sec. 4.1.

However, the Wallace-Freeman MML estimator from equation (20) and the Dowe-Wallace Ideal Group (IG) estimator have both been shown to be statistically consistent for the Neyman-Scott problem [Dowe and Wallace, 1996; 1997a; 1997b; Wallace, 2005, secs. 4.2–4.5 and 4.8; Dowe *et al.*, 2007, secs. 6.1.3–6.1.4; Dowe, 2008a, secs. 0.2.3 and 0.2.5; 2008b, p. 453]. An interesting discussion of the intuition behind these results is given in [Wallace, 2005, sec. 4.9].

We now use MML to re-visit a Neyman-Scott(-like) panel data problem from [Lancaster, 2002], as hinted at in [Dowe, 2008a, sec. 0.2.3, footnote 88].

6.5 Neyman-Scott panel data problem (from Lancaster)

Following the concise discussion in [Dowe, 2008a, sec. 0.2.3, footnote 88], we use MML here to re-visit the panel data problem from [Lancaster, 2002, 2.2 Example 2, pp. 651–652].

$$y_{i,t} \;=\; f_i + x_{i,t}\beta + u_{i,t} \qquad (i = 1, ..., N;\ t = 1, ..., T) \tag{21}$$

where the $u_{i,t}$ are independently Normal$(0, \sigma^2)$ conditional on the regressor sequence, f_i, and $\theta = (\beta, \sigma^2)$.

We can write the (negative) log-likelihood as

$$L \;=\; \frac{NT}{2}\log 2\pi \;+\; \frac{NT}{2}\log\sigma^2 \;+\; \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t} - f_i - x_{i,t}\beta)^2 \tag{22}$$

Using the Wallace-Freeman approximation [Wallace and Dowe, 1999a, sec. 6.1.2; Wallace, 2005, chap. 5; Dowe, 2008b, p. 451, eqn (5)] from equation (20), we require a Bayesian prior (which does not have a great effect but which does, among other things, keep the estimator statistically invariant) and the determinant of the expected Fisher information matrix of expected second-order partial derivatives (with respect to the f_i $[i = 1, ..., N]$, β and σ^2).

Before taking the expectations, let us first take the second-order partial derivatives — starting with the diagonal terms.

$$\frac{\partial L}{\partial f_i} \;=\; -\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{i,t} - f_i - x_{i,t}\beta)\ , \qquad \text{and} \qquad \frac{\partial^2 L}{\partial {f_i}^2} \;=\; T/(\sigma^2) \tag{23}$$

Not dissimilarly,

$$\frac{\partial L}{\partial \beta} = -\frac{1}{\sigma^2}\sum_{t=1}^{T} x_{i,t}(y_{i,t} - f_i - x_{i,t}\beta), \qquad \text{and} \qquad \frac{\partial^2 L}{\partial \beta^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T} {x_{i,t}}^2 \tag{24}$$

$$\frac{\partial L}{\partial(\sigma^2)} \;=\; \frac{NT}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t} - f_i - x_{i,t}\beta)^2\ , \qquad \text{and} \tag{25}$$

$$\frac{\partial^2 L}{\partial(\sigma^2)^2} \;=\; -\frac{NT}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t} - f_i - x_{i,t}\beta)^2 \tag{26}$$

Still looking at the second derivatives, let us now look at the off-diagonal terms and then return to take expectations.

$$\frac{\partial^2 L}{\partial f_i \partial f_j} \;=\; \frac{\partial^2 L}{\partial f_j \partial f_i} \;=\; \frac{\partial}{\partial f_j}\left(-\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{i,t} - f_i - x_{i,t}\beta)\right) \;=\; 0 \quad (\text{for } i \neq j) \tag{27}$$

$$\frac{\partial^2 L}{\partial f_i \partial \beta} \;=\; \frac{\partial^2 L}{\partial \beta \partial f_i} \;=\; \frac{1}{\sigma^2} \sum_{t=1}^{T} x_{i,t} \tag{28}$$

$$\frac{\partial^2 L}{\partial f_i \partial (\sigma^2)} \;=\; \frac{\partial^2 L}{\partial (\sigma^2) \partial f_i} \;=\; \frac{1}{2(\sigma^2)^2} \sum_{t=1}^{T} (y_{i,t} - f_i - x_{i,t}\beta) \tag{29}$$

$$\frac{\partial^2 L}{\partial \beta \partial (\sigma^2)} \;=\; \frac{\partial^2 L}{\partial (\sigma^2) \partial \beta} \;=\; \frac{1}{(\sigma^2)^2} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t}(y_{i,t} - f_i - x_{i,t}\beta) \tag{30}$$

Now taking expectations to get the terms contributing to the determinant of the expected Fisher information matrix, namely the expected Fisher information, let us first use equation (27) (dealing with the off-diagonal cases $i \neq j$) and equation (23) (dealing with the diagonal cases $i = j$) to give

$$E(\frac{\partial^2 L}{\partial f^2}) \;=\; \prod_{i=1}^{N} E(\frac{\partial^2 L}{\partial {f_i}^2}) \;=\; \prod_{i=1}^{N} \frac{\partial^2 L}{\partial {f_i}^2} \;=\; \prod_{i=1}^{N} T/(\sigma^2) \;=\; T/((\sigma^2)^N) \tag{31}$$

Re-visiting equation (29), we have that

$$E(\frac{\partial^2 L}{\partial f_i \partial (\sigma^2)}) \;=\; 0 \qquad (\text{for } i = 1, ..., N) \tag{32}$$

Equation (28) gives us a term proportional to $1/(\sigma^2)$, namely

$$E(\frac{\partial^2 L}{\partial f_i \partial \beta}) \;=\; \frac{1}{\sigma^2} \sum_{t=1}^{T} E(x_{i,t}) \;, \tag{33}$$

and equation (26) gives us

$$\begin{aligned} E(\frac{\partial^2 L}{\partial (\sigma^2)^2}) \quad &= \quad -\frac{NT}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} NT \; E((y_{i,t} - f_i - x_{i,t}\beta)^2) \\ &= \quad -\frac{NT}{2(\sigma^2)^2} + \frac{NT\sigma^2}{(\sigma^2)^3} \;=\; \frac{NT}{2(\sigma^2)^2} \end{aligned} \tag{34}$$

From [Lancaster, 2002, p. 651, 2.2 Example 2, equation (2.8)] and our equation (30), we have

$$E(\frac{\partial^2 L}{\partial \beta \partial (\sigma^2)}) \;=\; 0 \tag{35}$$

Looking at the $(N+2) \times (N+2)$ expected Fisher information matrix, we first note that the only non-zero entry in the σ^2 column is also the only non-zero entry in σ^2 row, namely that from equation (34).

Looking at the rest of the matrix, namely the $(N+1) \times (N+1)$ sub-matrix in the top left, we see that the only non-zero off-diagonal terms are the $E(\partial L/(\partial f_i \; \partial \beta))$ terms from equation (33) in the row and column corresponding to β. Looking at equations (31) and (33), we see that these few off-diagonal terms from equation (33) and all the diagonal terms are of the form $\text{Const}_1/(\sigma^2)^2$.

Combining this with equation (34), we see that the Fisher information is given by $\text{Const}_2 \times (1/(\sigma^2)^{N+1}) \times (NT)/(2(\sigma^2)^2) \; = \; \text{Const}/((\sigma^2)^{N+3})$.

Anticipating what we need for the Wallace-Freeman (1987) MML approximation in equation (20), this expression for $\text{Fisher}(\vec{\theta})$ and equation (22) then give that

$$
\begin{aligned}
& L + (\log \text{Fisher}(\vec{\theta}))/2 \\
= \quad & \frac{NT}{2} \log 2\pi \; + \; \frac{NT}{2} \log \sigma^2 \; + \; \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - f_i - x_{i,t}\beta)^2 \\
& + (\frac{1}{2} \log(\text{Const}) \; - \; \frac{N-3}{2} \log \sigma^2) \\
= \quad & \frac{1}{2} \log((2\pi)^{NT} \; \text{Const}) \; + \; \frac{(N-1)T-2}{2} \log(\sigma^2) \\
& + \; \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - f_i - x_{i,t}\beta)^2 \qquad (36)
\end{aligned}
$$

Re-capping, leaving aside the Bayesian priors (which will give statistical invariance to the MML estimator) and some constant terms, we see that the Wallace-Freeman MML approximation gives us what [Lancaster, 2002, p. 652] calls "the "correct" degrees of freedom, $N(T-1)$, apart from a negligible term".

As per [Dowe, 2008a, sec. 0.2.3, footnote 88], this can be extended to also deal with the subsequent panel data problem from [Lancaster, 2002, 2.3 Example 2, pp. 652-653].

Having made these points about statistical invariance and statistical consistency of MML, we perhaps digress slightly and note that Grünwald and Langford [2004; 2007] have shown statistical inconsistency under model misspecification for various Bayesian estimators and various forms of the Minimum Description Length (MDL) principle, but we are not aware of any current evidence for a statistical inconsistency in MML [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5].

The above — and other — evidence and experience has led to the following conjectures. The first two conjectures deal with the case that there is a true model in the space of models being examined, and the subsequent conjectures deal with the case of model misspecification.

Conjecture 1 [Dowe *et al.*, 1998, p. 93; Edwards and Dowe, 1998, sec. 5.3; Wallace and Dowe, 1999a, p. 282; 2000, sec. 5; Comley and Dowe, 2005, sec. 11.3.1, p. 269; Dowe, 2008a, sec. 0.2.5, pp. 539–540; 2008b, p. 454]: Only MML and very closely-related Bayesian methods are in general both statistically consistent and invariant.

(This first conjecture was once the subject of a panel discussion at a statistics conference [Dowe *et al.*, 1998a].)

Conjecture 2 (*Back-up Conjecture*) [Dowe *et al.*, 2007, sec. 8; Dowe, 2008a, sec. 0.2.5; 2008b, p. 454]: If there are (hypothetically) any such non-Bayesian methods, they will be far less efficient than MML.

Re the issue of statistical consistency under model misspecification as per sec. 5.2, first, suppose that the space where we conduct our search is dense in the space from which the true model comes (e.g., suppose the true model space is that of friendly non-Gaussian t distributions and our search space is the space of finite mixtures of Gaussian distributions). Then, in this case, if our inference method is statistically consistent when the true model comes from the search space (i.e., in this example, if our inference method is statistically consistent within the space of finite Gaussian mixture models) then we would expect our inference method to still be statistically consistent for the misspecified true model from the larger class (i.e., in this example, we would expect our inference method to remain statistically consistent when the true model is a friendly non-Gaussian t distribution from ["just"] outside our search space). Paraphrasing, if it's consistent in the search space, it can get arbitrarily close within the search space — and, if the search space is dense in the true space, then it would appear that we can get arbitrarily close to something arbitrarily close, seemingly implying statistical consistency.

Still on this issue of statistical consistency under model misspecification from sec. 5.2, we know that MML will be statistically invariant and we further conjecture [Dowe, 2008a, sec. 0.2.5, especially p. 540] that MML will still — in this more challenging setting — be statistically consistent. If there are (hypothetically) any non-Bayesian methods which are statistically consistent in this setting, then we further conjecture that they will be less efficient than MML.

Note throughout that the statistical consistency is coming from the information-theoretic properties of MML and the statistical invariance is coming from the Bayesian priors.

If there is any truth to these conjectures — and I am yet to see anything I could legitimately call a counter-example — then it would seem to suggest that inference done properly must inherently be Bayesian. I make this claim because

- recalling sec. 5.1, *statistical invariance* says that we get the same answer whether we model in (e.g.) polar or Cartesian co-ordinates, or in (e.g.) side length, face area or volume of a cube, etc.

- recalling sec. 5.2, *statistical consistency* — whether for a properly specified model or a misspecified model — merely says that collecting extra data (as people seem very inclined to do) is a worthwhile activity.

For other related arguments in support of Bayesianism and the Bayesian MML approach, recall Wallace's notions of Bayesian bias [Wallace, 1996c, sec. 4.1] and false oracles [Wallace, 1996c, sec. 3; Dowe *et al.*, 2007], Wallace's intuitive but nonetheless important proof that sampling from the Bayesian posterior is a false oracle [Wallace, 1996c, sec. 3.4] and Wallace's arguments that the Strict MML estimator (which is deterministic) approximates a false oracle [Wallace, 1996c, secs. 5–7]. For different arguments in support of Bayesianism and the Bayesian MML approach for those who like the notions of Turing machines and (Kolmogorov complexity or) algorithmic information theory from sec. 2.4, recall that (as per the end of sec. 2.4) the choice of (Universal) Turing Machine in algorithmic information theory is (obviously?) also a Bayesian choice [Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450].

6.6 Further MML work, such as MML Support Vector Machines

The relationship between MML and Kolmogorov complexity (or algorithmic information theory) [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.2–2.3] from sec. 2.4 means that MML can be applied universally across all inference problems and even compare and contrast two models from very different families.

Let us saying something about the statistical learning theory of Vapnik and Chervonenkis, VC dimension, Support Vector Machines (SVMs) and Structural Risk Minimisation (SRM) [Vapnik, 1995] before discussing how this might be put in an MML framework.

The statistical learning theory of Vapnik and Chervonenkis uses the notion of the Vapnik-Chervonenkis (VC) dimension of a set to give a classical non-Bayesian way of doing regression (typically using a technique called Structural Risk Minimisation [SRM]) and classification (typically using Support Vector Machines [SVMs]). Recalling the distinction between inference and prediction from sec. 4.1, both statistical learning theory and Akaike's Information Criterion (AIC) seem to be motivated by prediction, whereas MML is motivated by inference — a point noted in [Wallace, 1997] (ref. [281] in [Dowe, 2008a]).

It is not clear to this writer what statistical learning theory advocates for (comparing) models from different families (e.g., polynomial vs exponential), or for decision trees (classification trees) (e.g., [Wallace and Patrick, 1993; Wallace, 2005, sec. 7.2]), where each split corresponds to a conjunctive "AND", discretely partitioning the data. It is less clear what statistical learning theory will advocate for the generalisation of decision trees (classification trees) called decision graphs (classification graphs) [Oliver and Wallace, 1991; 1992; Oliver, 1993; Tan and Dowe, 2002; 2003], in which a disjunctive "OR" in the formula allows selected branches of the tree to join — making the model space more general, as now we have two

discrete operators (both split and join) in addition to various continuous-valued parameters (such as the multinomial class probabilities in the leaves).

Problems where this writer is either not sure what statistical learning theory will advocate and/or where I suspect that it might advocate (and have to bear the statistical inconsistency of) Maximum Likelihood include those mentioned above and also the Neyman-Scott and Neyman-Scott panel data problems from secs. 6.4 — 6.5 and the "gappy" problem and the bus number problem of sec. 4.1.

Whether or not my ignorance of how statistical learning theory will behave in certain situations is a shortcoming in statistical learning theory or in me is something for the reader to decide. Meanwhile, MML is universal — due to its relationship with Kolmogorov complexity (as per secs. 1 and 6.2) and likewise because it always has a message length as an objective function.

The original treatments (e.g., [Vapnik, 1995]) of the Vapnik-Chervonenkis (VC) notion of statistical learning theory, support vector machines (SVMs) and structural risk minimisation (SRM) are not Bayesian. Efforts have been made to put the notions of Vapnik-Chervonenkis statistical learning theory into a Bayesian MML (or similar) framework (starting with [Vapnik, 1995, sec. 4.6]). At least one motivation for doing this is to be able to apply statistical learning theory to problems where it might not have otherwise been possible to do so.

Fleshing out ideas hinted at in [Vapnik, 1995, sec. 4.6], MML has been applied to Support Vector Machines (SVMs) in [Tan and Dowe, 2004] (where we do not just have SVMs, but we also have decision trees — and, in fact, we have a hybrid model with SVMs in the leaves of decision trees), with discussions on alternative and refined coding schemes given in [Dowe, 2007; 2008a, footnote 53; 2008b, p. 444] including [Dowe, 2008a, footnote 53, fourth way, pp. 527–528; 2008b, p. 444] explicitly modelling the distribution of *all* the variables, including the input variables. It is MML's abovementioned relationship with Kolmogorov complexity (or algorithmic information theory) that enables us to consider alternative coding schemes. Explicitly modelling the distribution of all the variables (including the input variables) would amount to making generalized hybrid Bayesian network graphical models (as per sec. 7.6), some of whose properties are discussed in secs. 2.3 and and 3.6. (Perhaps digressing, as per [Dowe, 2008a, footnote 56], [Rubinstein *et al.*, 2007] might also be of some use here.)

Staying with the Vapnik-Chervonenkis VC dimension but moving from SVMs to SRM, MML was compared with SRM for univariate polynomial regression in [Wallace, 1997]. See the discussion in [Dowe, 2008a, sec. 0.2.2., p. 528, col. 1, including also footnotes 57 and 58].

MML has also been applied (e.g.) to hierarchical classification in [Boulton and Wallace, 1973b; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1 and sec. 0.2.4, p. 537, col. 2] (and elsewhere), with an application of hierarchical MML mixture modelling in [Wallace and Dale, 2005], to image recognition in [Torsello and Dowe, 2008b; 2008a], with other work on MML mixture modelling in sec. 7.6, and MML applied to James-Stein estimation in [Schmidt and Makalic, 2009b]. For some of many more examples, see also (e.g.) sec. 7.6 (in particular) and (in general) all of sec. 7.

6.7 A note on Minimum Description Length (MDL)

I forget how many times and how regularly I have been asked to summarise and/or highlight the similarities and differences between MML and the much later Minimum Description Length (MDL) principle. Because of this and also because of a request from at least one and possibly all of my referees, I include the current section. For want of somewhere to put it, I have placed it here, but the reader can probably safely skip from sec. 6.6 to sec. 6.8 with probably greater continuity and perhaps no great loss.

Historically, the Minimum Description Length (MDL) principle [Rissanen, 1978] (following formative ideas in [Rissanen, 1976]) was first published 10 years, 6 journal papers [Wallace and Boulton, 1968, p. 185, sec. 2; Boulton and Wallace, 1969; 1970, p. 64, col. 1; 1973b, sec. 1, col. 1; 1975, sec. 1, col. 1; Wallace and Boulton, 1975, sec. 3] (it would be 7 journal papers if we were hypothetically to count [Boulton and Wallace, 1973a]), 1 Master's thesis [Boulton, 1970], at least one conference abstract [Boulton and Wallace, 1973c] and 1 PhD thesis [Boulton, 1975] after the seminal Wallace and Boulton MML paper [1968], including 3 years after the Wallace and Boulton [1975] paper introducing Strict MML (whose original publication was delayed as per sec. 6.1 and [Dowe, 2008a, sec. 0.1, p. 524, col. 1]).

The ideas in MML of being Bayesian and of having a two-part message have been unwaveringly constant throughout since the original 1968 inception [Wallace and Boulton, 1968]. A variety of theoretical justifications for Bayesianism are given in (e.g.) sec. 6.5 and in [Wallace, 1996c, secs. 3 (especially 3.4), 4.1 and 5–7; Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe *et al.*, 2007; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450]. A variety of theoretical justifications for the two-part form of the MML message are given in (e.g.) [Wallace and Freeman, 1987, p. 241; Barron and Cover, 1991; Wallace, 2005, sec. 3.4.5, p. 190, note use of "agrees"; Dowe *et al.*, 2007, sec. 5.3.4].

The objectives — or at the least the way(s) of attempting to achieve the objectives — of the Minimum Description Length (MDL) principle would appear to have changed over the years since the first MDL paper in 1978 [Rissanen, 1978], where part of the motivation appears [Rissanen, 1978, p. 465] to be (algorithmic information theory or) Kolmogorov complexity, a term repeated in [Rissanen, 1999a, sec. 2, p. 261]. It is the prerogative of any scientist or any researcher to change and/or refine their ideas, and I attempt to survey various developments and changes in the presentations I have seen of MDL.

Rissanen appears throughout his MDL works to want to avoid being Bayesian. This seems slightly curious to me for a few reasons. First, there are countably infinitely many Universal Turing Machines (UTMs) and, as per secs. 2.4 and 6.5, the choice of a UTM is a Bayesian choice. As such, in relating MDL to Kolmogorov complexity, it seems difficult not to relate MDL to Bayesianism. Second, although Rissanen does not seem to want to use a Bayesian prior, his Normalised Maximum Likelihood (NML) uses the Jeffreys "prior" [Rissanen, 1996; 1999a], an approach

one might facetiously call "Bayesian". The Jeffreys "prior" is not without issue — it is based on the data (thus apparently weakening the claimed relationship with Kolmogorov complexity), it doesn't always normalise [Wallace and Dowe, 1999b, sec. 2], and it will typically depend upon things which we would not expect to be overly relevant to our prior beliefs — namely, the strength and location of our measuring instruments [Dowe *et al.*, 1996e, p. 217; Wallace and Dowe, 1999a, sec. 2.3.1; Comley and Dowe, 2005, sec. 11.4.3, p. 273]. There are also other concerns [Wallace and Freeman, 1987, sec. 1, p. 241; Wallace and Dowe, 1999a, sec. 5, p. 277, col. 2]. Efforts to normalise the Jeffreys prior by restricting its domain are said by other authors to be "unsatisfying" [Dawid, 1999, sec. 5, p. 325, col. 2] and would certainly appear to be reverting back to Bayesianism (rather than "Bayesianism"). Another opinion about the Bayesianism or otherwise in MDL is "... we see that Rissanen's approach is not incompatible with a Bayesian approach" [Clarke, 1999, sec. 2, p. 338, col. 2]. And while discussing the Jeffreys "prior" and Normalised Maximum Likelihood (NML) approach(es) [Rissanen, 1996; 1999a], it is worth inviting the reader to compare with the approximately contemporaneous PIC (Phillips Information Criterion, Posterior Information Criterion) [Phillips and Ploberger, 1996] and the much earlier and very similar Wallace-Freeman approximation [Wallace, 1984a; Wallace and Freeman, 1987] from equation (20) of no later than 1987 — see also the discussion in [Wallace, 2005, sec. 10.2.1].

The MDL notion of 'completing the code' (or complete coding) [Rissanen, 1996; Grünwald *et al.*, 1998, sec. 4] seems to break down for a variety of relatively simple cases [Wallace and Dowe, 1999b, secs. 1.2 and 2.3] and would appear to be in violation of contravening the convergence conditions of the two-part message form from which the results in [Barron and Cover, 1991] emanate, a variety of theoretical justifications for which are cited above. The latest versions of MDL seem to advocate using Normalised Maximum Likelihood (NML) to select the "model class" (see [Wallace and Dowe, 1999b, sec. 2.1] re issues of ambiguity here) and the order of the model but not to do the point estimation of the parameters. Given the issues with Maximum Likelihood of over-fitting and statistical inconsistency raised in secs. 4.1 and 5.2, we endorse the avoidance of Maximum Likelihood. But then Normalised Maximum Likelihood (NML) starts to look quite similar to the earlier Wallace-Freeman [1987] approximation for model order selection but without necessarily easily being able to advocate a point estimate.

And, as in sec. 6.5, Grünwald and Langford [2004; 2007] have shown statistical inconsistency for various Bayesian estimators and various forms of the Minimum Description Length (MDL) principle (under model misspecification), but none of us are aware of any current evidence for a statistical inconsistency in MML [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5]. (It is probably worth mentioning here an attack on MML [Grünwald *et al.*, 1998] which was later retracted [Grünwald, 2007, sec. 17.4, An Apologetic Remark; Dowe, 2008a, sec. 0.2.4, p. 536].)

As we near the conclusion of this sub-section, it is worth pointing out that many authors use MDL as a generic term for any MDL-like or MML-like coding

scheme based on some similar method. People should take as much care as they are able to here — [Comley and Dowe, 2005, sec. 11.4.3] gives plenty of examples of poor MDL-like coding schemes whose performances vastly improved when they were re-visited using MML [Wallace and Patrick, 1991; 1993, Viswanathan *et al.*, 1999; Needham and Dowe, 2001]. (See also sec. 4.) One can make bad wine from good grapes, and a poor coding scheme will not do justice to the likes of MDL and MML.

Despite the above challenges to and/or criticisms of much MDL work to date as it compares with earlier MML work, considering the issues which Rissanen raises — as he attempts to maintain all the niceties (of MML) while also attempting to avoid being Bayesian — and contemplating responses can certainly yield insights at the very least, and of course possibly much more. One such purported insight is described in part of sec. 7.1, a section in which objective Bayesianism is discussed. (Also worth mentioning here in passing is a way in which MDL could be re-visited stating parameter estimates "using whatever 'code' or representation was used in the presentation of the raw data" [Wallace and Dowe, 1999b, sec. 3, p. 336, col. 2].) And, finally, re comparing MDL and MML, as per the final sentence of this section, I refer the reader to [Wallace and Dowe, 1999c, abstract].

Given that when we take logarithms base 2 ($\log_2$) we typically refer to the unit as a bit, for some historical context on what to call the units when we take natural logarithms (base e, $\log_e$), see (e.g.) [Hodges, 1983, pp. 196–197] for early names (e.g., '*ban*'), see (e.g.) [Boulton and Wallace, 1970, p. 63; Comley and Dowe, 2005, p. 271, sec. 11.4.1] re '*nit*', and see (e.g.) much later MDL writings for the term '*nat*'.

Other treatments of this topic of contrasting MDL and MML are given in (e.g.) [Wallace, 1999; Wallace and Dowe, 1999b, sec. 3; Wallace, 2005, sec. 10.2; Comley and Dowe, 2005, sec. 11.4.3, pp. 272–273; Baxter and Oliver, 1995] and — perhaps most especially in summary — [Wallace and Dowe, 1999c, abstract].

6.8 Comparing "Right"/"Wrong" and Probabilistic scores

The original idea behind the notion of boosting was to more heavily (penalise or) weight incorrect answers in a decision tree (or classification tree) so as to grow the tree and ultimately have less errors — that is, right/wrong errors. Sec. 3.1 showed us that "right"/"wrong" scoring is not invariant to re-framing of questions, and sec. 3.2 re-iterated some recent results on the uniqueness of log(arithm)-loss scoring in being invariant to the re-framing of questions. This said, before we examine boosting more closely in sec. 6.9, we might ask what a good "right"/"wrong" score tells us about the log(arithm)-loss score and vice versa.

By rights, in a multiple-choice question of c choices, even if $c >> 2$, giving the correct answer a probability of just less than 0.5 can still result in a higher probability of at least 0.5 being given to an incorrect answer and so a "right"/"wrong" score of "wrong" (or 0). So, regardless of the number of choices, c, an incorrect answer guarantees a score of at least 1 bit. Correspondingly, a score of less than 1 bit

guarantees a "right" answer. At the other extreme, it is possible that a probability of just over $1/c$ allocated to the correct answer will give us a 'right"/"wrong" score of "right" (or 1). Correspondingly, a score of more than $\log(c)$ will surely give us a "right"/"wrong" score of "wrong" (or 0). So, if we have n multiple-choice questions of $c_1, ..., c_i, ..., c_n$ options each, then a "right"/"wrong" score of 0 corresponds to a log(arithm)-loss cost of at least $\log(2^n) = n$ bits, and a "right"/"wrong" score of (all correct) n correct corresponds to a log(arithm)-loss cost of at most $\sum_{i=1}^{n} \log(c_i) = \log(\prod_{i=1}^{n} c_i)$.

So, slightly paradoxically, on a quiz of 1 ternary (3-valued) question, someone (with probabilities $\{0.498, 0.501, 0.001\}$) might get a wrong answer for a minimum total of 0 "right" with a log-loss penalty score of just over $\log(2)$ whereas someone else (with probabilities $\{0.334, 0.333, 0.333\}$) might get a correct answer for a maximum total of 1 "right" but with a worse log-loss penalty score of just under $\log(3)$. I put it to the reader that the person with the better log-loss score actually has a better claim to having been correct on this question than the person given a score of 1 "right".

And, of course, more emphatically, on a quiz of n questions, someone with a "right"/"wrong" score of 0 "right" might have a log(arithm)-loss penalty score of little over $\log(2^n)$ whereas someone who got $(n-1)$ out of n correct might have an arbitrarily large (or infinite) log(arithm)-loss penalty score by assigning an arbitrarily small (or zero) probability to the correct answer in the one question that this person got "wrong". (And, similarly, as at least implicitly pointed out in sec. 3.2, one can use boosting to make all sorts of guesses and predictions in data which is just random noise, and although much damage could be done to the log(arithm)-loss penalty score, no damage will be done to the "right"/wrong" score.) We address this issue of getting good "right"/wrong" scores without unduly damaging the log(arithm)-loss penalty score again in sec. 6.9.

6.9 Boosting

One suggestion of Chris Wallace's (in private communication) was that the right/wrong predictive accuracy of MML decision trees could be improved by going to each leaf in turn and doing one additional split beyond the MML split. I understand that this was the motivation behind the subsequent [Oliver and Hand, 1994]. Of course, as per sec. 4.1, optimal prediction is given not just by using the MML tree, but by combining several trees — ideally as many as possible — together [Oliver and Hand, 1996; Tan and Dowe, 2006].

However, given our new apparent uniqueness results for probabilistic log(arithm)-loss scoring from [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438] and sec. 3.2, it perhaps makes more sense to carefully focus on improving the probabilistic log(arithm)-loss score.

One option is to do MML inference but with the "boosting priors" from [Dowe, 2008a, sec. 0.2.6]. The idea behind these "boosting priors" is that, rather than fix our Beta/Dirichlet prior [Wallace, 2005, p. 47 and sec. 5.4] to have $\alpha = 1$, we

can try "boosting priors", whose rough form [Tan and Dowe, 2006, sec. 3.4, p. 598; Dowe, 2008a, sec. 0.2.6] on α could be, e.g., $3/(2\sqrt{\alpha}(1+\sqrt{\alpha})^4)$ or $(e^{-\alpha/\pi})/(\pi\sqrt{\alpha})$. The idea is simply to retain a mean of (approximately) 1 but to have a large spike near $\alpha = 0$, which in turn increases our propensity to have pure classes.

Another option is to re-visit boosting and to think of it not as in its original form of minimising the number of right/wrong errors but rather instead in the similar form of trying to optimise the expected predictive score. This modification to boosting is related to the original form of boosting in that each individual (right/wrong) mistake will typically correspond to a poor right/wrong (yes/no) score. The predictive score should not be done using the log-likelihood, but rather should be done using the minimum expected Kullback-Leibler (minEKL) probability estimate [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2, 2008b, p. 444] from sec. 4.1. In other words, if there are m classes and a given leaf has counts $c_1, ..., c_i, ..., c_m$ for the m classes and $C = \sum_{i=1}^{m} c_i$, Maximum Likelihood would advocate a probability estimate of $\hat{p}_{i\ \mathrm{MaxLhood}} = c_i/C$ for each class. However, if we think of α_i from a Dirichlet prior as denoting a "pre-count" in class i (before any data is actually counted), then the probability in each class can be regarded as $\hat{p}_i = (c_i + \alpha_i)/(C + (\sum_{j=1}^{m} \alpha_i))$. Of course, we can set $\alpha_i = \alpha$ for each class and then use the so-called "boosting priors" on α as per [Tan and Dowe, 2006, sec. 3.4, p. 598; Dowe, 2008a, sec. 0.2.6].

Let us finish with three further comments. First, by way of digression, an attempt to give "objective" ways of choosing α is given in sec. 7.1. Second, for those who wish to boost to improve the right/wrong score or simply wish to get a good to excellent right/wrong score, given the (apparent uniqueness of) invariance of the log-loss score, we make the simple recommendation that predictors that give good right/wrong scores be checked so that they also give a good log-loss score — this might involve moving extreme probabilities away from the extremities of 0 and 1 (such as can arise from using Maximum Likelihood). (Some possible estimators for doing this are given in, e.g., [Wallace, 2005, secs. 4.8 and 5.4]. A prediction method which is good enough to genuinely get a good "right"/"wrong" score can surely be gently modified or toned down to give a good log-loss score.) As a third comment, for those not wishing to sacrifice statistical consistency in their efforts to improve predictive accuracy, it might be worth considering comments [Dowe, 2008a, footnote 130] about potential dangers of placing too much weight on the likelihood of the data.

7 MML AND SOME APPLICATIONS IN PHILOSOPHY AND ELSEWHERE

MML has been applied to a variety of problems in philosophy — including (e.g.) the philosophy of science [Dowe and Oppy, 2001], the philosophy of statistics and inference [Dowe *et al.*, 2007], the philosophy of mind (see, e.g., sec. 7.3), the philosophy of language and the philosophy of religion. We mention these and some other — mainly philosophical — issues here.

7.1 Objective Bayesianism (and Bertrand's paradox) and some new invariant "objective priors"

Bertrand's paradox is essentially concerned with the issue that we can not choose a uniform Bayesian prior in all parameterisations. Certainly, many authors would like an objective form of Bayesianism — or, equivalently, a parameterisation in which our Bayesian prior can be uniform.

Recalling the notion of Universal Turing Machine (UTM) from sec. 2.4, one can claim that a simplest UTM is one with the smallest product of its number of states and its number of symbols (as this is the number of rows in the instruction table) [Dowe *et al.*, to appear (b)].

Simplest UTMs have been used in inference [Wallace, 2005, sec. 2.3.12; Gammerman and Vovk, 2007a; 2007b; Martin-Löf, 1966; Dowe, 2007], and they are one way of attempting to be objective (while still being Bayesian) and — as a consequence — side-stepping Bertrand's paradox. Of course, such objectivity — where possible — will potentially be useful in places such as legal battles [Dowe, 2008a, pp. 438–439].

Much good and interesting work has been done in the area of objective Bayesianism by (e.g.) J. M. Bernardo [Bernardo and Smith, 1994] and others. Above, we follow Wallace in offering simplest UTMs as objective priors. Below, as per the end of sec. 6.7, we now change tack and re-visit the Jeffreys "prior" (whose use, incidentally, was not advocated by Jeffreys [1946]) as an objective invariant prior.

Some things that we know to be invariant upon re-parameterisation include the likelihood function, Maximum Likelihood, the marginal probability $r(x_i) = \int h(\theta)f(x_i|\theta)\ d\theta$ of datum x_i (from sec. 6.1), the message length and many of its variants (from secs. 6 — 6.3), Minimum Message Length (when using an invariant form), the Fisher information and (recalling secs. 4.1 and 6.9) the minimum Expected Kullback-Leibler divergence (minEKL) estimator.

Given these invariant building blocks, we now take the Jeffreys "prior" (which we recall from sec. 6.7 and pointers therein does not always normalise), and construct a family of other invariant priors. To kick off with an example, let us take a (m-state, $(M-1)$-dimensional, $M \geq 2$) multinomial distribution — as per [Wallace, 2005, sec. 5.4] — with prior of the form Const ${p_1}^{\alpha_1}...{p_M}^{\alpha_M}$ (where we use α_i where Wallace [2005, sec. 5.4] writes $\alpha_i - 1$). We will have counts s_i $(i = 1, ..., M)$ and we let $N = \sum_{i=1}^{M} s_i$ and $A = \sum_{i=1}^{M} \alpha_i$. The Wallace-Freeman (1987) MML estimate from equation (20) is $(\hat{p}_i)_{MMLWF1987} = (s_i+\alpha_i+1/2)/(N+A+M/2)$. And, recalling sec. 6.7, the minEKL (or Laplace) estimate (equivalently here, the posterior mean [Boulton and Wallace, 1969]) is $(\hat{p}_i)_{MinEKL} = (s_i+\alpha_i+1)/(N+A+M)$. As such, we observe that the Wallace-Freeman [1987] MML estimate from equation (20) with the Jeffreys "prior" ($\alpha_i = -1/2$) gives the Maximum Likelihood estimate. We similarly observe that the Wallace-Freeman estimate with the uniform prior ($\alpha_i = 0$) is equivalent to getting the minEKL estimate with the Jeffreys "prior" ($\alpha_i = -1/2$).

In this particular case of the multinomial distribution, we note that we can trans-

form from the (invariant) Wallace-Freeman [1987] MML estimate to (invariant) minEKL by adding 1/2 to the α_i. As such, if the Jeffreys prior $h_0 = h_{Jeffreys} = h_{FisherInfo}$ (with $\alpha_i = -1/2$) is to be called objective, then a case can be made that so, too, is the uniform prior h_1 (with $\alpha_i = 0$). We can iterate again to get further invariant priors: h_2 (with $\alpha_i = 1/2$), h_3 (with $\alpha_i = 1$), etc. One could also iterate in the opposite direction: h_{-1} (with $\alpha_i = -1$), h_{-2} (with $\alpha_i = -3/2$), etc. All such priors — or at least those which normalise — are invariant (by construction) and can be regarded in some sense as "objective". One could then choose the prior h_j for the smallest value of j for which h_j normalises.

This method of using the Jeffreys "prior" to generate further invariant objective priors (via invariant transformations) and then taking the "first" to normalise certainly generalises — well beyond the above example of the multinomial distribution — to other distributions. In general, for some given distribution, start again with $h_0 = h_{Jeffreys} = h_{FisherInfo}$ and then, given prior h_i, let h_{i+1} be the prior such that (some invariant form of) the MML estimate with prior h_{i+1} is as close as possible in Kullback-Leibler distance (and ideally equal) to the minEKL estimate with prior h_i. With however much ease or difficulty, we can then generate this sequence of invariant priors $h_0 = h_{Jeffreys} = h_{FisherInfo}$, h_1, h_2, ... and perhaps also h_{-1}, h_{-2}, etc. (As a general rule, because of MML's tendency to fit just right and minEKL's tendency to under-fit as per sec. 4.1, we expect to see a corresponding progression in this sequence of priors — as is perhaps best seen from the above example with the multinomial distribution. In that case, h_i has $\alpha_i = (i-1)/2$, meaning that, as i increases, it takes increasingly much data to move the estimate away from the centroid where for all i, $\hat{p}_i = 1/M$.) If there does exist some smallest j for which h_j normalises, a case could be made that this is an objective invariant prior which might be more suitable than the Jeffreys "prior", h_0.

Penultimately, cases could be made for investigating combining two such priors, as in considering (e.g.) $h_{hybrid} = \sqrt{h_{j_1} h_{j_2}}$. Cases could also be made for attempting to allow j not to be an integer but rather somehow to be fractional. We will not investigate these here.

And, finally, returning to issues from sec. 6.7, it would perhaps be nice if Normalised Maximum Likelihood (NML) could be re-visited but with use of one of these alternative invariant priors to the Jeffreys "prior", h_0. This should retain the statistical invariance but might reduce some of the vulnerabilities (such as over-fitting and statistical inconsistency, both mentioned earlier) associated with Maximum Likelihood.

7.2 Goodman's "Grue" paradox (and choice of language)

Nelson Goodman's "grue" paradox raises the issue of why notions like "green" and "blue" should be more natural than notions of "grue" (green before time t_0 [say year 3000] and blue thereafter) and "bleen" (blue before time t_0 [say year 3000] and green thereafter). This has been discussed in the Solomonoff predictive

and Wallace MML inductive frameworks, with relevant writings being [Solomonoff, 1996; 1997b, sec. 5; Comley and Dowe, 2005, sec. 11.4.4; Dowe, 2008a, footnotes 128, 184 and 227]. Among other things, an adequate solution of when to arrive at a notion like "grue" and when to arrive at a notion like "green" (which is, after all, grue before time t_0 and bleen thereafter) is presumably necessary when trying to evolve language (for those beings not yet with language) or when trying to communicate with non-human terrestrials or extra-terrestrials [Dowe, 2008a, footnote 184]. Wallace's approach from [Dowe, 2008a, footnote 128], elaborating upon [Comley and Dowe, 2005, sec. 11.4.4], was summarised as follows: "Suppose someone is growing and harvesting crops, commencing (much) before t_0 and finishing (much) after t_0. We expect the grass and certain moulds to be green, and we expect the sky and certain weeds to be blue. The notions of grue and bleen here offer at most little in return other than sometimes to require (time-based) qualification and to make the language sometimes unnecessarily cumbersome." This said, there are times of event changes which can be of interest. If t_0 were the time of the next expected reversal of the earth's magnetic field, then in talking on such a time-scale we have reason to disambiguate between magnetic north and geographic north in our language — as these notions are approximately equal before t_0 and approximately antipodal (for at least some time) after t_0 [Dowe, 2008a, footnote 128]. But the terms 'grue' and 'bleen' cost us but seem to gain us nothing. By and large, languages will develop, import, qualify and/or abbreviate terms when these terms warrant (sufficient) use.

And, while on that very issue of abbreviation, the reader will note at least one place in this article where we have written "Minimum Message Length (MML)". This convention of putting an acronym or other abbreviation in brackets immediately after the term it abbreviates enables us to use the abbreviation (rather than the full term) elsewhere — thus enabling us to shorten the length of our message.

And, digressing, while on the earlier issue of languages, MML has been used to model evolution of languages [Ooi and Dowe, 2005; Dowe, 2008a, sec. 0.2.4; 2008b, p. 455] (not to mention finite state automata [Wallace and Georgeff, 1983] and DNA string alignment [Allison *et al.*, 1990a; 1990b; 1990; 1991; 1992a; 1992b; Allison and Wallace, 1993, 1994a, 1994b]).

An able philosopher colleague, Toby Handfield, has told me in private communication — while discussing Lewis [1976] and the "laws of nature" — that if MML were able to recognise a number constructed as (say) the sum without carries of e (the base of natural logarithms) expanded in hexadecimal (base 16) and π expanded in decimal (base 10), then this would go a long way towards convincing him that MML can solve Goodman's grue paradox. Using the relationship between MML and (algorithmic information theory or) Kolmogorov complexity [Wallace and Dowe, 1999a; Wallace, 2005, chaps. 2–3] discussed in sec. 6, we outline the argument below. In short, MML will have no difficulty with doing this (in principle) — the caveat being that the search might take quite some time.

We can specify e as $\sum_{i=0}^{\infty} 1/i!$ in a (Turing machine) program of length P_e, and we can specify the h^{th} hex(adecimal) digit of e in a program of length $P_e + C_1 + l(h)$

for some constant C_1, where $l(h)$ is the length of some prefix code (recall sec. 2.2) over the positive integers (e.g., the unary code from sec. 2.4). We could use a code of length 1 for $h = 1$ and of length $\leq 1 + \lceil 1 + \log_2(h) + 2\log_2(\log_2(h)) \rceil < 2 + 1 + \log_2(h) + 2\log_2(\log_2(h))$ for $h \geq 2$. Similarly, we can specify π as (e.g.) $\sum_{i=0}^{\infty}(4(-1)^i)/(2i+1)$ in a (Turing machine) program of length P_π, and we can specify the h^{th} hex(adecimal) digit of π in a program of length $P_\pi + C_2 + l(h)$ for some constant C_2, where $l(h)$ is as above.

The program for addition without carry/ies simply entails addition without carry (or modulo addition) in each place, h, for $h = 1, 2, 3, 4, ...$ So, for the h^{th} hex(adecimal) digit, we can say that the h^{th} hex digit, composite_h, of our composite number is given as follows:

```
if        (e_{h, 16} + pi_{h, 10} < 15)
then  composite_h  =  e_{h, 16} + pi_{h, 10}
else   composite_h  =  e_{h, 16} + pi_{h, 10} - 16;
```

Given that this is how the composite number is being generated, given sufficiently many hex digits of this number, the Minimum Message Length (MML) inference will be the algorithm for generating this composite number. Again, the search might be slow, but this will be found.

We can actually take this further by randomly adding noise. Let us suppose that, with probability p, hex digit h comes from some probability distribution $(q_1, q_2, ..., q_{14}, q_{15}, q_{16} = 1 - \sum_{i=1}^{15} q_i)$ and with probability $1 - p$ this h^{th} hex digit will be composite_h. So, $Pr(\text{CompositeWithNoise}_h = \text{composite}_h) = pq_{\text{composite}_h} + (1 - p)$. For each $i \neq \text{composite}_h$, $Pr(\text{CompositeWithNoise}_h = i) = pq_i$. In the case that $p = 0$, this reduces to the noiseless case. Here, the search will be even slower, but with sufficiently many digits and with sufficient search time, we will converge upon the noiseless program above generating the digits in addition to having an increasingly good quantification of the noise.

7.3 MML, inductive inference, explanation and intelligence

As intimated in sec. 1, MML gives us the inductive inference (or induction, or inference, or explanation) part of intelligence [Dowe and Hajek, 1997; 1998, especially sec. 2 (and its title) and sec. 4; Sanghi and Dowe, 2003, sec. 5.2]. And Ockham's razor tells us that we should expect to improve on Searle's "Chinese room" look-up table [Searle, 1980] by having a compressed representation — as per our commonsense intuition and arguments in [Dowe and Hajek, 1997, sec. 5.1 and elsewhere; 1997, p. 105, sec. 5 and elsewhere; Dowe, 2008a, footnote 182 and surrounding text] and sec. 4.

Let us consider an assertion by Hutter [Legg and Hutter, 2007] that compression is equivalent to (artificial) intelligence (although subsequent work by Hutter now seems instead to equate intelligence with a weighted sum of reward scores across different environments). This assertion is later than a similar idea of Hernandez-Orallo [Hernández-Orallo and Minaya-Collado, 1998; Hernández-Orallo, 2000]. It

is also stronger than an earlier idea [Dowe and Hajek, 1997; 1998, especially sec. 2 (and its title) and sec. 4; Sanghi and Dowe, 2003, sec. 5.2 and elsewhere; Dowe, 2008a, sec. 0.2.5, p. 542, col. 2 and sec 0.2.7, p. 545, col. 1] that (the part of intelligence which is) inductive inference (or inductive inference) is equivalent to (two-part) compression. Let us look at the two issues separately of

- (i) first, whether all of (artificial) intelligence or perhaps just inductive inference is equivalent to (two-part) compression, and
- (ii) second, whether it is satisfactory simply to talk about (one-part) compression or whether we should insist upon two-part compression.

First, the components of intelligence would appear to include (at least) memory, deductive inference, inductive inference and ability to receive direct instruction. (By deductive inference, we mean and include mathematical calculations and logical reasoning, such as modus ponens — Socrates is a man, all men are mortal, therefore Socrates is mortal. To illustrate the distinction with an example, inductive inference is more along the lines of all men are mortal, Socrates is mortal, therefore we assert some probability that Socrates is a man.) We need memory to store observations for making inductive inferences, for remembering inductive inferences, for remembering our progress through mathematical calculations or other (logical) deductions and for remembering those direct instructions (perhaps the deductions or inferences of others) that we receive. For example, a good human player of a game where the search space is too vast to be exhaustively searched (like chess or Go) needs inductive inference and direct instruction to help with an evaluation function (such as, in chess, the advantages of passed pawns, the weaknesses of isolated and backward pawns, and the approximate equivalence between a queen and three minor pieces), memory to remember these, memory to remember the rules, and deduction (and memory again) to do the lookahead calculations in the search tree. It is clear that all these aspects of intelligence are useful to a human player of such a game. However, to the mathematician, the logician, and especially a mathematician or a logician checking the validity of a proof (or someone double-checking that a Sudoku solution is correct), the main forms of intelligence required would surely appear to be deduction and memory. It is fair to say that the harder aspects of inductive learning and (two-part) compression also require memory and deductive inference. And we have argued elsewhere that we are more likely to attribute intelligence to someone performing an act of great memory if they have done this using a compressed representation [Dowe and Hajek, 1997; 1998]. But we ask the reader whether we should not attribute intelligence to the chess player or the mathematician (or the person checking a Sudoku solution) when performing (difficult) activities involving at most little inductive inference.

Second, in many cases, doing straight (one-part) compression rather than two-part compression can lead to an incorrect model (as in the statistical inconsistency of the minEKL estimator from sec. 4.1 for the "gappy" problem mentioned in sec. 4.1 and for the Neyman-Scott problem in sec. 6.4) — and this remains true asymptotically regardless of how much data we have.

As per [Dowe, 2008a, sec. 0.2.7, p. 545, col. 1], I have discussed with J. Hernandez Orallo the notion of quantifying the intelligence of a system of agents and endeavouring to quantify how much of this comes from the individual agents (in isolation) and how much comes from their communication. Let us try to take this further in a couple of different (related) ways. First, it would be good to (artificially) evolve such a communal intelligence, including (perhaps inevitably) evolving a language. (As a tiny step, one of my 4^{th} year Honours project students in 2009, Jeffrey R. Parsons, has made slight progress in evolving Mealy and/or Moore machines with the message length as a guiding fitness function. I do not wish to overplay his current progress, but it is in a useful direction.) And, second, re the topics of swarm intelligence and ant colony optimisation, perhaps only a very small range of parameter values (where the parameters describe the individual agents and/or their communication) permit the different parts to interact as an "intelligent" whole. This raises a couple of further issues: the issue of using MML to analyse data (as per sec. 7.6 and [Dowe, 2008a, sec. 0.2.7, p. 545]) and infer the parameter values (or setting) giving the greatest communal intelligence, and the additional issue(s) of whether or not greater prior weight should be given to those systems giving the interesting outcome of intelligence, and (similarly) — in the *fine tuning* argument of sec. 7.7 and [Dowe *et al.*, to appear (b)] — whether greater prior probability should be given to parameter settings in which interesting universes (like our own) result.

Having mentioned here the issues of intelligence, non-human intelligence and communication, it is worth mentioning some of Chris Wallace's comments about trying to communicate with an alien intelligence [Dowe, 2008a, sec. 0.2.5, p. 542, col. 2, and also footnote 184 and perhaps text around footnote 200] (and possibly also worth recalling Goodman's notion of "grue" from sec. 7.2 and [Dowe, 2008a, footnote 128]).

We conclude here by saying that further discussion on some of the topics in this sub-section will appear in [Hernández-Orallo and Dowe, 2010].

7.4 (So-called) Causality

Chris Wallace did much work on "causal nets" using MML, including doing the (MML) mathematics and writing the software behind several papers on this topic [Wallace and Korb, 1994; Wallace, 1996b; Wallace *et al.*, 1996a; 1996b; Dai *et al.*, 1996a; 1996b; 1997a; 1997b; Wallace and Korb, 1997; 1999; Korb and Wallace, 1997; 1999] (with the possible exception of [Neil *et al.*, 1999a; 1999b]). I have no objection to the quality of Wallace's MML statistical inference — from the available data — in this work. Indeed, I have (at most little or) nothing but the highest praise for it. However, there are at least two or so matters about which one should express caution when interpreting the results from such inference.

One issue is that getting the wrong statistical model (because we didn't have enough data, we hadn't searched thoroughly enough and/or our statistical inference method was sub-optimal) can lead to having arrows pointing the wrong way.

And even if we did have enough data, our statistical inference method was ideal and we searched thoroughly, it could still be the case that the true (underlying) model (from which the data has been generated) is outside the family of models that we are considering — e.g., our model family might be restricted to linear regressions (on the parent "explanatory" variables to the "target" child variable) with Gaussian noise while the real data-generating process might be more complicated. In such cases, slightly modifying the family of models being considered might change the directions of arrows in the inference, suggesting that the directions of these arrows should not all be regarded as directly "*causal*" [Dowe, 2008a, footnote 169].

As a related case in point, we might have data of (at least) two variables, including (i) height/co-ordinates of (weather) (monitoring) station and (ii) (air) pressure reading. Our best statistical model might have arrows from pressure reading to height of monitoring station, but we surely shouldn't interpret this arrow as being in any way causal.

Of course, temporal knowledge (of the order in which things occur) is also important for attributing causality.

Whether or not this is well-known and well-documented (and please pardon my medical ignorance, as per sec. 3.2), there would appear to be a substantial overlap between cancer patients and stroke patients. Let's suppose that in many cases the patient has a stroke and then cancer is detected some months later. It could appear that the stroke caused the cancer, but it is perhaps more probable that cancer-induced changes in the tissue and/or the bloodstream caused the stroke — even if the primary cancer was not in the brain and the metastatic cancer did not present in the brain until after the stroke. If this is all true, then it would suggest that the actual mechanism is that the Cancer is causing the Stroke — despite the possibility that an analysis of the data might easily lead one to conclude that the Stroke is causing the (Brain) Cancer.

We also have to be careful about issues such as (e.g.) hidden (or unknown) variables. As an example, a hidden latent variable might cause both A (which takes place slightly before B) and B. B might do the same exam paper (of mathematical calculations) as A but starting and finishing slightly later. Or perhaps B is a newspaper which goes to print after newspaper A goes to print but before newspaper A appears on the stands. We expect B's answers and stories to be very similar to A's, but this is because A and B have common (hidden) causes; and it seems loose to say that A causes B.

As another example, standing on a podium of a Grand Prix tends to greatly increase one's chances of winning at a subsequent Grand Prix event. But this wouldn't be true of one of the scaffolding constructors who tested the podium before the race ceremony, and nor would it be true of some overly exuberant spectator who managed to somehow get access to the podium. Rather, there is a (not very) hidden cause of ability causing someone to do well in two races, and the result of doing well in the first of these races caused that racer to stand on the podium at the end of the first of the two races.

Lecturers (at university, college, or wherever), tutors, teachers and instructors who are able to give lectures without notes (of whom Chris Wallace is but one notable example) often give excellent lectures. Let us assume that this is the norm for such people. The cause of the good lecturing is surely the excellent memory and strong command of the subject of the lecturer rather than any supposed benefit that the average person might supposedly gain by trying to lecture without notes.

As another example, if A causes B and B causes C and we only know about A and C but have not yet even conceived of B (and, of course, we might be open to the existence of B but simply don't know), then I think we can say A "causations" C but we have to be careful about saying that (e.g., supposedly) A *causes* C. A specific case might be the old example of A being living alone, C being having (few or) no rodents at home. B is the owning of a pet cat — the single person keeps the pet for company (and has no flat-mate or house-mate to complain), and the cat keeps the rodents at bay.

Possibly see also [Dowe, 2008a, sec. 0.2.7, pp. 543–544] re LNPPP and causality.

7.5 Elusive model paradox (and encryption)

Gödel's incompleteness theorem consists of constructing a mathematical statement which can be interpreted as saying that "This statement is not provable" [Gödel, 1931]. Clearly, this statement can't be false, or it would be provable and hence true, leading to a logical contradiction. Hence, the statement must be both true (of the natural numbers) and not provable.

The original version of the elusive model paradox gives us a sequence where the next number is one (or unity) more than what we would expect it to be [Dowe, 2008a, footnote 211]. The subsequent version of the paradox essentially takes modulo 2 (so that even numbers are transformed to 0 and odd numbers are transformed to 1) and then gives us a binary sequence (or bit string) (of 0s and 1s) in which we can (paradoxically) be sure that the next bit is not the bit that we expect (or would have predicted) based on what we have seen so far (before it). This leads to a contradiction from which the only escape would appear to be the undecidability of the Halting problem (or *Entscheidungsproblem*), the notion that there are many calculations which will never terminate but for which we can never know that they will not terminate [Turing, 1936].

Whether one takes the elusive model paradox as being over a sequence of (increasing) positive integers (as per the original version [Dowe, 2008a, footnote 211]) or over a binary bit string sequence of 0s and 1s (as per the later version [Dowe, 2008b, p. 455]), each of these versions in turn can be thought of in two (essentially) equivalent ways. One of these ways is to play this as a game, where we have one agent (which can be represented by a Turing machine) generating the sequence and a group of one or more agents (which can also be represented by a Turing machine) trying to guess the next bit — while the (Turing machine) agent generating the sequence is attempting to generate the opposite bit to what (the Turing machine representing) those guessing will guess. (It might possibly help to

think of the generating agent as a soccer player taking a penalty kick, trying to kick the ball where the goalie won't be — or as a tennis player trying to serve the ball to where the receiver won't be; and, in turn, the guessing agent as the goalie trying to guess the location of the kick or the tennis receiver trying to anticipate the serve.) To give both the generating Turing machine agent and the guessing Turing machine agent the best chances to do their respective jobs properly, we will assume that — recalling sec. 2.4 — these Turing machines are universal. As such, among other things, there will be finite emulation programs (or translation programs) causing one machine to emulate the other, and vice versa. As the generating program and the guessing program start out on small sequences being the early short initial segments of the generated bits and the guess(ed) bits, the programs will quite possibly have different models of the data. But, as the sequences get longer and longer, after they become at least kilobits, megabits, gigabits, terabits, etc. long and vastly longer than the abovementioned translation programs, the models that these two UTMs have of the available data will start to converge. The guessing UTM will have had a *very* good look at the generating UTM and — given that the generating UTM is a finite deterministic machine — the guessing UTM would appear to be able at some stage to lock in on the behaviour of the generating UTM, thereafter guessing all subsequent bits correctly. Similarly, at some stage, the generating UTM would appear to be able at some stage to lock in on the behaviour of the guessing UTM, thereafter anticipating all its guesses and then flipping the bit before generating it. After both these stages have occurred, we have the contradiction that the guessing UTM always guesses correctly and the generating UTM anticipates the guess, flips the bit that it knows will be guessed and ensures that all subsequent guesses are incorrect. The Halting problem gets us out of this paradox (and seems like the only way out), as both the generating UTM and the guessing UTM can and very often want more time before they are content that they have modelled the other correctly. The second (essentially) equivalent way of thinking of the elusive model paradox is simply that the generating UTM agent and the guessing UTM agent are the same — as at the end of the previous paragraph. After starting off the sequence, we guess which bit should most probably come next, and then generate the bit which is least probable to come next — and then continue this indefinitely. We get (essentially) the same paradox, and again the Halting problem seems like the only way out of the paradox.

The above all said by way of introduction, we now present some variations on the elusive model paradox [Dowe, 2008a, footnote 211; 2008b, p. 455], including — recalling sec. 4.1 — one using inference and one using prediction. (Recall that inference uses the single best model whereas prediction weights over all available models.) One variation is that we can restrict ourselves to multinomial Markov models where the n^{th} order Markov model has (a maximum of) 2^n binomial distributions.

Let $j = j_m = j_m(i) \leq i$ be some unbounded non-decreasing computable function of i. At step i, having bits $b_1, b_2, ..., b_i$, we choose b_{i+1} as follows, from the following two similar but (slightly) different methods — noting that both these constructions

are computable.

Method 1 (inference — using restricted "memory"): We infer the best (MML) Markov model of order $\leq j_m$ based on $b_1, b_2, ..., b_i$. We then use the predictive distribution from this MML inference to give a probability distribution for b_{i+1}. We then choose b_{i+1} to be the bit with the least predicted probability.

Method 2 (prediction — using restricted "memory"): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1}. Again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

With both of these methods — method 1 (inference) and method 2 (prediction) — the resultant sequence is "random" in the sense that no Markov model of finite order is going to be able to compress it And this is so because the construction of the sequence is to destroy any such structure at the first viable opportunity upon its detection.

Recall that both these constructions immediately above based on restricting "memory" are computable. Two (or more) alternative computable constructions — based on restricting computation *time* rather than "memory" — are given below. Let $j = j_t = j_t(i) > i$ be some strictly increasing computable function of i.

Method 3 (inference — with restricted computation time): We infer the best (Minimum Message Length [MML]) inference from all computable functions (that we search over) within a search time of $\leq j_t$ based on $b_1, b_2, ..., b_i$. As in method 1, we then use the predictive distribution from this MML inference to give a probability distribution for b_{i+1}. We then choose b_{i+1} to be the bit with the least predicted probability.

Method(s) 4 (prediction — with restricted computation time): We use Bayesian model averaging. There are two ways of proceeding further in restricted finite computation time — method 4(a) and (with tighter restriction) method 4(b).

Method 4(a): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1}. Here, time restriction of $\leq j_t$ is applied to each of the individual Markov models in turn. They are then averaged as per Method 2. Again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

Method 4(b): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1}. But, here, the time restriction of $\leq j_t$ is tighter in that it is applied to the entire calculation, including the final Bayesian model averaging. And, yet again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

We might refer to these various sequences emerging from variations of our elusive model paradox as "*red herring*" sequences. Among other things, these (red herring sequences) have the potential to be used in encryption. If various people or agents studying the sequence have varying computational resources (e.g., different lag lengths in the Markov models they can consider), a variant of the sequence can be constructed in such a way as to guide some sub-population (perhaps those from whom we wish to conceal some data) to believe in the presence or absence

of a particular pattern while guiding a different sub-population (perhaps those to whom we wish to divulge the data) to be aware of the presence (or absence) of some (particular) pattern.

Finally, let us return to the notes at the start of this subsection about how one (apparently) needs the Halting problem (or *Entscheidungsproblem*) [Turing, 1936] to resolve the elusive model paradox [Dowe, 2008a, footnote 211; 2008b, p. 455]. The Halting problem is something which people do not normally encounter before their undergraduate university years. I put it to the reader that the elusive model paradox is something from which we can deduce the halting problem yet which should be accessible to school students.

7.6 Some of many other issues which MML can address

- In numerical integration (or numerical quadrature), we see a variety of approaches such as the rectangular rule, the trapezoidal rule and Simpson's rule, etc. Of course, where the function we are trying to integrate is not generated from a polynomial and especially when it is generated from a noisy process, then it will typically be better to use MML or a related method to guide the fitting process rather than use arbitrarily complicated polynomials and suffer from the over-fitting problems that come with Maximum Likelihood and similar methods;

- generalized hybrid Bayesian network graphical models [Dowe and Wallace, 1998; Comley and Dowe, 2003; Tan and Dowe, 2004, sec. 5; Comley and Dowe, 2005] deal with the issue of "discriminative vs generative" studied by Jebara [2003] and others (e.g., [Long and Servedio, 2006]). Many authors have claimed that discriminative learning can often outperform generative learning. However, if one follows the ideas in [Dowe and Wallace, 1998; Comley and Dowe, 2003; Tan and Dowe, 2004, sec. 5; Comley and Dowe, 2005] and carefully uses MML — recalling the discussion of poor coding schemes in sec. 6.7, taking care with the coding scheme — to construct one's generalized hybrid Bayesian network graphical model (of which inference of a logic program via Inductive Logic Programming [ILP] [Dowe *et al.*, to appear (a)] is one possible outcome, as can be SVMs from sec. 6.6 or also, e.g., the sort of model from [Oliver and Dowe, 1995]), then the statistical consistency results of MML from [Dowe, 2008a, sec. 0.2.5] and discussed in sec. 6 should guarantee that "generative" learning (when done like this) works fine.

  Some properties of these generalized hybrid Bayesian network graphical models (which can include both continuous and discrete variables [Comley and Dowe, 2003; 2005]) are discussed in secs. 2.3 (where it is mentioned in passing that entropy can be defined on such hybrid structures) and 3.6 (where we mention that there is no difficulty in defining Kullback-Leibler divergence over such structures);

- following this point, where an unnormalised database is sufficiently large,

then MML inference will lead to database normalisation [Dowe, 2008a, sec. 0.2.6, footnote 187; 2008b, pp. 454–455; Dowe and Zaidi, 2010], often resulting in several tables, conveying a generalised Bayesian net. We can adjust our priors to require this process to be noise-free;

- experimental design [Dowe, 2008a, sec. 0.2.7, p. 544; 2008b, pp. 445–446];
- MML can be applied to statistical hypothesis testing [Dowe, 2008a, sec. 0.2.5, p. 539 and sec. 0.2.2, p. 528, col. 1; 2008b, p. 433 (Abstract), p. 435, p. 445 and pp. 455–456; Musgrave and Dowe, 2010], as can also MDL [Rissanen, 1999a, sec. 3]. (Perhaps see also [Dowe, 2008a, sec. 1].) Daniel F. Schmidt and Enes Makalic have recently presented work in front of an audience including me showing their desire to take this further. As per sec. 7.1, I harbour some concerns about associating the Maximum Likelihood estimate with Normalised Maximum Likelihood;
- association rules (from "data mining" and machine learning) can be incorporated within a generalised Bayesian network structure;
- re-visiting A. Elo's Elo system and M. Glickman's Glicko system for chess ratings. Whether for chess players with the advantage of the white pieces and the first move or whether for sports teams with a home ground advantage, MML can both select the relevant model and do the parameter estimation. Of interest would be a Neyman-Scott-like situation in which many games are being played but there are relatively few games per player. If similar interest would be a situation with several groups of players where there are many games played within each of the groups but very games played between members of different groups;
- directional angular data, such as the von Mises circular distribution [Wallace and Dowe, 1993; 1994a; Dowe *et al.*, 1995a; 1995b] and the von Mises-Fisher spherical distribution [Dowe *et al.*, 1996e; 1996f];
- inference of megalithic stone circle (or non-circle) geometries [Patrick and Wallace, 1977; Patrick, 1978; Patrick and Wallace, 1982];
- polynomial regression [Wallace, 1997; Wallace, 1998c; Viswanathan and Wallace, 1999; Rumantir and Wallace, 2001; Fitzgibbon *et al.*, 2002a; Rumantir and Wallace, 2003] (and perhaps also [Schmidt and Makalic, 2009c]);
- inference of MML neural nets [Makalic *et al.*, 2003];
- inference of MML decision trees (or classification trees) and decision graphs (or classification graphs) [Oliver and Wallace, 1991; Oliver *et al.*, 1992; Oliver and Wallace, 1992; Oliver, 1993; Uther and Veloso, 2000; Tan and Dowe, 2002; Tan and Dowe, 2003; Tan and Dowe, 2004], including (as per sec. 6.6) decision trees with support vector machines (SVMs) in their leaves [Tan and

Dowe, 2004] — with applications of MML decision trees and graphs in a variety of areas including (e.g.) protein folding [Dowe *et al.*, 1992; 1992a; 1993] and medical diagnosis [McKenzie *et al.*, 1993];

- MML clustering, mixture modelling and intrinsic classification via the *Snob* program [Wallace and Boulton, 1968; Wallace, 1984b; 1986; 1990b; 1990c; Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000] was originally for the multinomial and Gaussian distributions, but this was extended to also include the Poisson and von Mises circular distributions [Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000] — with applications in a variety of areas including (e.g.) spectral modelling [Papp *et al.*, 1993], protein folding [Zakis *et al.*, 1994], psychology and psychiatry [Kissane *et al.*, 1994; 1996; 1996a; Prior *et al.*, 1998]. Also of interest is determining whether our data appears to contain one line segment or a mixture of more than one line segment [Georgeff and Wallace, 1984a; 1984b; 1985] (and much later work on engineering bridge deterioration using a mixture of a Poisson distribution and a uniform distribution with total assignment [Maheswaran *et al.*, 2006]). (After the MML mixture modelling of multinomial, Gaussian, Poisson and von Mises circular distributions from 1994 [Wallace and Dowe, 1994b; 1996] came a slightly different paper doing only MML Gaussian mixture modelling [Oliver *et al.*, 1996] but emphasising the success of MML in empirical comparisons.) The MML single linear factor analysis from [Wallace and Freeman, 1992] was incorporate into [Edwards and Dowe, 1998] — although [Edwards and Dowe, 1998] did total (rather than partial) assignment and only did single rather than multiple [Wallace, 1995a; 1998b] factor analysis. This has also been extended to a variety of forms of sequential clustering [Edgoose and Allison, 1999; Molloy *et al.*, 2006], with an extension of [Edgoose and Allison, 1999] being (as in the next item) to MML *spatial* clustering. See also [Boulton and Wallace, 1973b; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1 and sec. 0.2.4, p. 537, col. 2] for a discussion of MML hierarchical clustering. As well as the abovementioned multinomial, Gaussian, Poisson and von Mises circular distributions [Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000], this work — without sequential and spatial clustering (following in the next item) has been extended to other distributions [Agusta and Dowe, 2002a, 2002b; 2003a; 2003b; Bouguila and Ziou, 2007];

- extensions of MML spatial clustering [Wallace, 1998a; Visser and Dowe, 2007] to tomographic [Visser *et al.*, 2009a] and climate [Visser *et al.*, 2009b] models. Variations on this work (and possibly other MML image analysis work [Torsello and Dowe, 2008a; 2008b]) should lend themselves both to training a robot to learn a model for and then recognise a particular class of object (such as a coloured shape, like a particular type of fruit) for robotic hand-eye co-ordination and also to analysing data in constructing a bionic eye;

- inference of systems of one or more probabilistic/stochastic (partial or) or-

dinary (difference or) differential equations (plus at least one noise term) from (presumably noisy) data (as per wishes from [Dowe, 2008a, sec. 0.2.7, p. 545]). Uses for this should include the likes of (e.g.) inferring parameter settings for ant colonies and other swarms so as to model them and/or suggesting settings giving better "intelligence" (as per sec. 7.3) and medical applications such as (e.g.) cardiac modelling or modelling stem cells;

- MML, particle physics and the analysis of the data in the search for the Higgs boson [Dowe, 2008a, sec. 0.2.7, p. 544, col. 2];

- etc.

Also of interest here might be

- the relationship between MML and the likelihood principle of statistics [Wallace, 2005, sec 5.8; Wallace and Dowe, 1999b, sec. 2.3.5], for which MML's violation is "innocent enough — a misdemeanour rather than a crime" [Wallace, 2005, sec. 5.8, p. 254; Dowe, 2008a, sec. 0.2.4, p. 535, col. 2];

- the relationship between MML and Ed Jaynes's notion of maximum entropy (or MaxEnt) priors [Jaynes, 2003; Wallace, 2005, secs. 1.15.5 and 2.1.11; Dowe, 2008a, sec. 0.2.4, p. 535, col. 1]. While MML and MaxEnt are different, while still on the topic of entropy and MML, it turns out that — within the MML mixture modelling literature — the term used to shorten the message length when going from (the inefficient coding scheme of) total assignment to (the efficient coding scheme of) partial assignment equates to the entropy of the posterior probability distribution of the class assignment probabilities [Visser *et al.*, 2009b; Wallace, 1998a; Visser and Dowe, 2007];

- etc.

7.7 MML and other philosophical issues

When time next permits, here are some of the many other philosophical issues to which MML pertains

- entropy is *not* time's arrow [Wallace, 2005, chap. 8 (and p. vii); Dowe, 2008a, sec. 0.2.5; 2008b, p. 455], and note the ability of MML to detect thermal fluctuations and not over-fit them where some other statistical methods might be tempted to regard the standard noisy fluctuations as being something more [Wallace, 2005, chap. 8]. (One wonders whether the formation of these ideas might be evident in [Wallace, 1973a].) Recalling sec. 4.1 on inference (or explanation) and prediction, one interesting thing here is Wallace's take on why it is that we wish to predict the future but (only) infer (or explain) the past [Wallace, 2005, chap. 8];

- being able to accord something the title of a "*miracle*" [Wallace, 2005, sec. 1.2, p. 7; Dowe, 2008a, sec. 0.2.7, p. 545, col. 1; 2008b, p. 455], the *fine tuning* argument in *intelligent design* [Dowe *et al.*, to appear (b)] (possibly see also sec. 7.3) and evidence that there is an intelligent supervisor/shepherd listening to our prayers and overseeing — and sometimes intervening in — our lives. Just as we can use MML to decide whether or not to accord something the title of a miracle, so, too, we can set about being objective and using MML to quantify the probability of certain coincidences and whether or not there could be an intelligent supervisor/shepherd intervening in our lives. Of course, such a shepherd might be able to make discrete minor changes effectively impossible to notice in one place which lead to substantial changes in other places. (I am not offering my opinion one way or another here, but rather merely raising how this issue might be addressed in an MML framework);

- Efficient Markets [Dowe and Korb, 1996; Dowe, 2008a, sec. 0.2.5; 2008b, p. 455] — due to the Halting problem, MML and Kolmogorov complexity essentially say (in short) that financial markets are very unlikely to be efficient and next to impossible to be proved efficient. Attempts to make this point more accessible by showing the effects of having a variety of trading approaches equal in all ways but one where one trader is better in terms of (e.g.) inference method, speed or memory, are given in [Collie *et al.*, 2005; 2005a];

- redundant Turing Machines (unlike those of sec. 2.4), for which pre- and post-processing can be used to effectively emulate a (Universal) Turing Machine by non-conventional means [Dowe, 2008a, sec. 0.2.7, p. 544];

- undecidability in (optimal) engineering tuning and design (ultimately due to the Halting problem);

- probabilities of conditionals and conditional probabilities [Dowe, 2008a, sec. 0.2.7, p. 546];

- information and MML re originality of an idea, degree of creativity of an act or design — or humour [Dowe, 2008a, sec. 0.2.7, p. 545] (the reader is welcome to inspect not unrelated ideas in [Solomonoff, 1995; Schmidhuber, 2007] in order to determine the originality of this idea). Puns typically entail finding commonality between at least two different subject matters. The finding of such commonality is crucial to the creation of the pun, and the recognising of such commonality is crucial to the understanding of the pun. A similar comment applies to the creation and solving of (cryptic) clues from a (cryptic) crossword. This said, in my experience, creating puns seems to be more difficult than understanding them — whereas solving cryptic crossword clues seems to be more difficult than creating them;

- mnemonics (or memory aids), whether or not this should be regarded as a philosophical issue. Certain mnemonics are compressions or contractions from which we can re-construct that "data" that we ultimately wish to recall. However, there is a seemingly slightly curious phenomenon here. People might recall (e.g.) the periodic table of elements or (e.g.) the base 10 decimal expansion of π by recalling a mnemonic sequence of words which tells a story. In the case of the periodic table, these words (can) start with the first one or so letters of the chemical elements in sequence. In the case of π, these words in sequence (can) have lengths corresponding to the relevant digit: so, the length of the i^{th} word is the i^{th} digit of π — e.g., "How I want a drink, alcoholic of course, after the heavy chapters involving quantum mechanics" (for 3.14159265358979). These little stories are fairly easily remembered — one might say that they are compressible, so one can re-construct them fairly easily, from where one can go on to re-construct what one was really trying to recall. However, the slight curiousity is that for all the easy compressible niceties of the mnemonic sequence, it is actually longer than the original. Perhaps the resolution is that whatever slight redundancies there are in the mnemonics serve as error corrections. So, perhaps such mnemonics are quite compressible in their own right so that they can easily be re-constructed but have sufficiently much redundancy to reduce errors. I think there is room for further discussion on this topic;

- fictionalism is an area of philosophy which (according to my understanding of it) is concerned about the sense in which we can talk about some fictional character (e.g., Elizabeth Bennett from "*Pride and Prejudice*") as though they were real — and then go on to discuss how said character(s) might behave in some scenario not presented in the story in which said character(s) appear(s). This seems very much to be an MML matter. We form a model of the character(s) based on what we know about the(se) character(s). We have a model of how various types of real-world character behave in certain scenarios, and we go from there. In similar vein, MML has much to say about the philosophical notion of *counterfactuals* and *possible worlds*, although here there is a minor issue of (recall sec. 4.1) whether we are interested in inference as to how things would most probably be in the nearest possible world or instead a weighted prediction of how things might be — obtained by doing a weighted combination of predictions over a variety of possible worlds;

- etc.

Having made the above list, I now mention some issues to which I would like to be able to apply MML.

- virtue — Confucius (the Chinese philosopher) [Confucius, 1938] and (about 500 years later) Jesus (from whom we have the Christian religion) are well-known for their comments on consideration, and Confucius further for his

other comments on virtue. In game theory (e.g., prisoners' dilemma), perhaps virtue is being/doing as all would need to do to give the best all-round solution (in similar vein to the merits of co-operation described in, e.g., [Wallace, 1998d]), perhaps it is doing the optimum by the others on the presumption that the others will all act out of self-interest. Perhaps MML can offer some insight here; and

- the Peter principle — I hear of and sometimes see far too many shocking appointments of the undeserving, of a candidate who "peaked in the interview". Perhaps in terms of some sort of experimental design or more properly knowing how to analyse the available data on candidates, it would be nice to put MML to use here.

8 ACKNOWLEDGEMENTS

I first thank Prasanta Bandyopadhyay, Malcolm Forster and John Woods for permitting me to contribute this piece. I thank the anonymous referee — whose helpful feedback clearly indicated that said referee had read my submission closely. I especially thank Prasanta for his regular patient editorial monitoring of my progress (or lack thereof), at least comparable in magnitude to that of an additional referee.

I next thank my mother and my family [Comley and Dowe, 2005, p. 287, sec. 11.6]. I now quote Confucius [1938, Book I, saying 15]: "Tzu-kung said, 'Poor without cadging, rich without swagger.' What of that? The Master said, Not bad. But better still, 'Poor, yet delighting in the Way; rich, yet a student of ritual.' "

I thank Chris Wallace [Gupta *et al.*, 2004; Dowe, 2008a, footnote 218] (for a variety of matters — such as, e.g., the letter referred to in [Dowe, 2008a, footnote 218] and all the training he gave me over the years), for whose Christopher Stewart WALLACE (1933-2004) memorial special issue of the *Computer Journal* [Dowe, 2008a; Brennan, 2008; Solomonoff, 2008; Jorgensen and McLachlan, 2008; Brent, 2008; Colon-Bonet and Winterrowd, 2008; Castro *et al.*, 2008] it was an honour to be guest editor. (For other mention and examples of the range and importance of his work, see, e.g., [Parry, 2005] and [Clark and Wallace, 1970].) I talk there of his deserving at least one Turing Award [Dowe, 2008a, sec. 0.2.2, p. 526 and sec. 0.2.4, p. 533, col. 2] and of how his work on entropy not being time's arrow might have stood him in contention for the Nobel prize in Physics if it could be experimentally tested [Dowe, 2008a, sec. 0.2.5, footnote 144]. (And, possibly, as per [Dowe, 2008a, sec. 0.2.7, p. 544, col. 2] and sec. 7.6, MML will have a role to play in the discovery of the Higgs boson.) I forgot to point out that, if he'd lived a few decades longer, the increasing inclusion of his work in econometrics (see, e.g., [Fitzgibbon *et al.*, 2004; Dowe, 2008a, sec. 0.2.3, footnote 88] and sec. 6.5 — and possibly also these papers on segmentation and cut-points [Oliver *et al.*, 1998; Viswanathan *et al.*, 1999; Fitzgibbon *et al.*, 2002b] — for teasers) might have one day earned him the Nobel prize in Economics. (Meanwhile, it would be good to re-visit both ARCH [Autoregressive Conditional Heteroskedasticity] and

GARCH [Generalised ARCH] using MML.) For a list of his publications (MML and other), see either the reference list to [Dowe, 2008a] (which also lists the theses he supervised) or `www.csse.monash.edu.au/~dld/CSWallacePublications`.

And, finally, I thank Fran Boyce [Dowe, 2008a, footnote 217; Obituaries, 2009], one of the nicest, gentlest, most thoughtful and most gracious people one could meet or ever even hope to meet.

BIBLIOGRAPHY

[Obituaries, 2009] Obituaries: Drive to learn despite struggle with lupus. *The Herald Sun*, page 79, Thu. 6 Aug., 2009. Melbourne, Australia; Author: M. Sonogan.

[Agusta, 2005] Y. Agusta. *Minimum Message Length Mixture Modelling for Uncorrelated and Correlated Continuous Data Applied to Mutual Funds Classification*. PhD thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia, 2005.

[Agusta and Dowe, 2002b] Y. Agusta and D. L. Dowe. Clustering of Gaussian and t distributions using minimum message length. In *Proc. International Conference on Knowledge Based Computer Systems (KBCS 2002)*, pages 289–299. Vikas Publishing House, 2002. http://www.ncst.ernet.in/kbcs2002.

[Agusta and Dowe, 2002a] Y. Agusta and D. L. Dowe. MML clustering of continuous-valued data using Gaussian and t distributions. In B. McKay and J. Slaney, editors, *Lecture Notes in Artificial Intelligence (LNAI), Proc. Australian Joint Conference on Artificial Intelligence*, volume 2557, pages 143–154, Berlin, Germany, December 2002. Springer-Verlag.

[Agusta and Dowe, 2003b] Y. Agusta and D. L. Dowe. Unsupervised learning of correlated multivariate Gaussian mixture models using MML. In *Lecture Notes in Artificial Intelligence (LNAI) 2903 (Springer), Proc. 16th Australian Joint Conf. on Artificial Intelligence*, pages 477–489, 2003.

[Agusta and Dowe, 2003a] Y. Agusta and D. L. Dowe. Unsupervised learning of Gamma mixture models using minimum message length. In M. H. Hamza, editor, *Proceedings of the 3rd IASTED conference on Artificial Intelligence and Applications*, pages 457–462, Benalmadena, Spain, September 2003. ACTA Press.

[Allison and Wallace, 1993] L. Allison and C. S. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimisation of multiple alignments. Technical report CS 93/188, Dept Computer Science, Monash University, Melbourne, Australia, 1993.

[Allison and Wallace, 1994a] L. Allison and C. S. Wallace. An information measure for the string to string correction problem with applications. *17th Australian Comp. Sci. Conf.*, pages 659–668, January 1994. Australian Comp. Sci. Comm. Vol 16 No 1(C) 1994.

[Allison and Wallace, 1994b] L. Allison and C. S. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *J. Mol. Evol.*, 39(4):418–430, October 1994. an early version: TR 93/188, Dept. Computer Science, Monash University, July 1993.

[Allison *et al.*, 1990] L. Allison, C. S. Wallace, and C. N. Yee. Induction inference over macro–molecules. Technical Report 90/148, Monash University, Clayton, Victoria, Australia, 3168, 1990.

[Allison *et al.*, 1990b] L Allison, C S Wallace, and C N Yee. Inductive inference over macromolecules. In *Working Notes AAAI Spring Symposium Series*, pages 50–54. Stanford Uni., Calif., U.S.A., 1990.

[Allison *et al.*, 1990a] L. Allison, C. S. Wallace, and C. N. Yee. When is a string like a string? In *International Symposium on Artificial Intelligence and Mathematics*, January 1990.

[Allison *et al.*, 1991] L Allison, C S Wallace, and C N Yee. Minimum message length encoding, evolutionary trees and multiple-alignment. Technical report CS 91/155, Dept Computer Science, Monash University, Melbourne, Australia, 1991.

[Allison *et al.*, 1992b] L. Allison, C. S. Wallace, and C. N. Yee. Finite-state models in the alignment of macro-molecules. *J. Mol. Evol.*, 35(1):77–89, July 1992. extended abstract titled: Inductive inference over macro-molecules in joint sessions at AAAI Symposium, Stanford, Mar 1990 on (i) Artificial Intelligence and Molecular Biology, pp5-9 & (ii) Theory and Application of Minimal-Length Encoding, pp50-54.

[Allison *et al.*, 1992a] L. Allison, C. S. Wallace, and C. N. Yee. Minimum message length encoding, evolutionary trees and multiple alignment. *25th Hawaii Int. Conf. on Sys. Sci.*, 1:663–674, January 1992. Another version is given in TR 91/155, Dept. Computer Science, Monash University, Clayton, Vic, Australia, 1991.

[Barron and Cover, 1991] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[Baxter and Oliver, 1995] R. A. Baxter and J. J. Oliver. MDL and MML: Similarities and differences. Technical report TR 94/207, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995.

[Bernardo and Smith, 1994] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley, New York, 1994.

[Bouguila and Ziou, 2007] N. Bouguila and D. Ziou. High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1716–1731, October 2007.

[Boulton, 1970] D. M. Boulton. Numerical classification based on an information measure. Master's thesis, M.Sc. thesis, Basser Computing Dept., University of Sydney, Sydney, Australia, 1970.

[Boulton, 1975] D. M. Boulton. *The Information Measure Criterion for Intrinsic Classification.* PhD thesis, Dept. Computer Science, Monash University, Clayton, Australia, August 1975.

[Boulton and Wallace, 1969] D. M. Boulton and C. S. Wallace. The information content of a multistate distribution. *J. Theor. Biol.*, 23:269–278, 1969.

[Boulton and Wallace, 1970] D. M. Boulton and C. S. Wallace. A program for numerical classification. *Computer Journal*, 13(1):63–69, February 1970.

[Boulton and Wallace, 1973c] D. M. Boulton and C. S. Wallace. A comparison between information measure classification. In *Proc. of the Australian & New Zealand Association for the Advancement of Science (ANZAAS) Congress*, August 1973. abstract.

[Boulton and Wallace, 1973b] D. M. Boulton and C. S. Wallace. An information measure for hierarchic classification. *Computer Journal*, 16(3):254–261, 1973.

[Boulton and Wallace, 1973a] D. M. Boulton and C. S. Wallace. Occupancy of a rectangular array. *Computer Journal*, 16(1):57–63, 1973.

[Boulton and Wallace, 1975] D. M. Boulton and C. S. Wallace. An information measure for single link classification. *Computer Journal*, 18(3):236–238, 1975.

[Brennan, 2008] M. H. Brennan. Data processing in the early cosmic ray experiments in Sydney. *Computer Journal*, 51(5):561–565, September 2008.

[Brennan *et al.*, 1958] M. H. Brennan, D. D. Millar, and C. S. Wallace. Air showers of size greater than 10^5 particles - (1) core location and shower size determination. *Nature*, 182:905–911, Oct. 4 1958.

[Brent, 2008] R. P. Brent. Some comments on C. S. Wallace's random number generators. *Computer Journal*, 51(5):579–584, September 2008.

[Castro *et al.*, 2008] M. D. Castro, R. D. Pose, and C. Kopp. Password-capabilities and the Walnut kernel. *Computer Journal*, 51(5):595–607, September 2008.

[Chaitin, 1966] G. J. Chaitin. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, 13:547–569, 1966.

[Chaitin, 2005] G. J. Chaitin. *Meta Math! The Quest for Omega.* Pantheon, 2005. ISBN 0-375-42313-3 (978-0-375-42313-0).

[Clark and Wallace, 1970] G. M. Clark and C. S. Wallace. Analysis of nasal support. *Archives of Otolaryngology*, 92:118–129, August 1970.

[Clarke, 1999] B. Clarke. Discussion of the papers by Rissanen, and by Wallace and Dowe. *Computer J.*, 42(4):338–339, 1999.

[Collie *et al.*, 2005] M. J. Collie, D. L. Dowe, and L. J. Fitzgibbon. Stock market simulation and inference technique. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, Rio de Janeiro, Brazil, Nov 2005.

[Collie *et al.*, 2005a] M. J. Collie, D. L. Dowe, and L. J. Fitzgibbon. Trading rule search with autoregressive inference agents. Technical report CS 2005/174, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia, 2005.

[Colon-Bonet and Winterrowd, 2008] G. Colon-Bonet and P. Winterrowd. Multiplier evolution — a family of multiplier VLSI implementations. *Computer Journal*, 51(5):585–594, September 2008.

[Comley and Dowe, 2003] Joshua W. Comley and David L. Dowe. General Bayesian networks and asymmetric languages. In *Proc. Hawaii International Conference on Statistics and Related Fields*, 5-8 June 2003.

[Comley and Dowe, 2005] Joshua W. Comley and David L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, M. A. Pitt, and I. J. Myung, editors, *Advances in Minimum Description Length: Theory and Applications (MDL Handbook)*, pages 265–294. M.I.T. Press, April 2005. Chapter 11, ISBN 0-262-07262-9. Final camera-ready copy submitted in October 2003. [Originally submitted with title: "Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages".].

[Confucius, 1938] Confucius. *The Analects of Confucius (Lun Yü)*. Vintage Books, 1989. (Published earlier with Macmillan in 1938.) Translated by Arthur Waley. Online at http://myweb.cableone.net/subru/Confucianism.html.

[Dai *et al.*, 1996a] H Dai, K B Korb, and C S Wallace. The discovery of causal models with small samples. In *1996 Australian New Zealand Conference on Intelligent Information Systems Proceedings ANZIIS96*, pages 27–30. IEEE, Piscataway, NJ, USA, 1996.

[Dai *et al.*, 1996b] H Dai, K B Korb, and C. S. Wallace. A study of causal discovery with weak links and small samples. Technical report CS 96/298, Dept Computer Science, Monash University, Melbourne, Australia, 1996.

[Dai *et al.*, 1997b] H Dai, K B Korb, C S Wallace, and X Wu. A study of causal discovery with weak links and small samples. Technical report SD TR97-5, Dept Computer Science, Monash University, Melbourne, Australia, 1997.

[Dai *et al.*, 1997a] Honghua Dai, Kevin B. Korb, C. S. Wallace, and Xindong Wu. A study of causal discovery with weak links and small samples. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97*, pages 1304–1309, 1997.

[Dawid, 1999] A. P. Dawid. Discussion of the papers by Rissanen and by Wallace and Dowe. *Computer J.*, 42(4):323–326, 1999.

[Deakin, 2001] M. A. B. Deakin. The characterisation of scoring functions. *J. Australian Mathematical Society*, 71:135–147, 2001.

[Dowe, 2007] D. L. Dowe. Discussion following "Hedging predictions in machine learning, A. Gammerman and V. Vovk". *Computer Journal*, 2(50):167–168, 2007.

[Dowe, 2008a] D. L. Dowe. Foreword re C. S. Wallace. *Computer Journal*, Christopher Stewart WALLACE (1933–2004) memorial special issue, 51(5):523–560, September 2008.

[Dowe, 2008b] D. L. Dowe. Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference — from (medical) "evidence". *Social Epistemology*, 22(4):433–460, October–December 2008.

[Dowe *et al.*, 1995] D. L. Dowe, L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. Circular clustering by minimum message length of protein dihedral angles. Technical report CS 95/237, Dept Computer Science, Monash University, Melbourne, Australia, 1995.

[Dowe *et al.*, 1996] D. L. Dowe, L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. Circular clustering of protein dihedral angles by minimum message length. In *Pacific Symposium on Biocomputing '96*, pages 242–255. World Scientific, January 1996.

[Dowe *et al.*, 1998] D. L. Dowe, R. A. Baxter, J. J. Oliver, and C. S. Wallace. Point estimation using the Kullback-Leibler loss function and MML. In X. Wu, Ramamohanarao Kotagiri, and K. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 87–95, Berlin, April 15–17 1998. Springer.

[Dowe *et al.*, 1996a] D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. *3rd Conf. on Maths and Computers in Sport*, pages 233–241, 1996. See also Technical Report TR 96/297, Dept. Computer Science, Monash University, Australia 3168, Dec 1996.

[Dowe *et al.*, 1996b] D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. Technical report TR 96/297, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1996.

[Dowe *et al.*, to appear (a)] D. L. Dowe, C. Ferri, J. Hernandez-Orallo, and M. Jose Ramirez-Quintana. An MML Scheme for ILP. Submitted.

[Dowe *et al.*, 2007] D. L. Dowe, S. Gardner, and G. R. Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, 58(4):709–754, December 2007.

[Dowe *et al.*, to appear (b)] D. L. Dowe, S. Gardner, and G. R. Oppy. MML and the fine tuning argument. To appear, 2011.

[Dowe and Hajek, 1997] D. L. Dowe and A. R. Hajek. A computational extension to the Turing test. Technical Report 97/322, Dept. Computer Science, Monash University, Australia 3168, October 1997.

[Dowe and Hajek, 1998] D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing test. In *Proceedings of the International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98)*, pages 101–106, Gippsland, Australia, February 1998.

[Dowe *et al.*, 1996c] D. L. Dowe, A.J. Hurst, K.L. Lentin, G. Farr, and J.J. Oliver. Probabilistic and Gaussian football prediction competitions - Monash. *Artificial Intelligence in Australia Research Report*, June 1996.

[Dowe *et al.*, 1998a] D. L. Dowe, M Jorgensen, G McLachlan, and C S Wallace. Information-theoretic estimation. In W Robb, editor, *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 125, Queensland, Australia, July 1998.

[Dowe and Korb, 1996] D. L. Dowe and K. B. Korb. Conceptual difficulties with the efficient market hypothesis: Towards a naturalized economics. In *Proc. Information, Statistics and Induction in Science (ISIS)*, pages 212–223, 1996. See also Technical Report TR 94/215, Dept. Computer Science, Monash University, Australia 3168, 1994.

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

[Dowe and Lentin, 1995] D. L. Dowe and K. L. Lentin. Information-theoretic footy-tipping competition — Monash. *Computer Science Association Newsletter (Australia)*, pages 55–57, December 1995.

[Dowe *et al.*, 1996d] D. L. Dowe, K.L. Lentin, J.J. Oliver, and A.J. Hurst. An information-theoretic and a Gaussian footy-tipping competition. *FCIT Faculty Newsletter, Monash University, Australia*, pages 2–6, June 1996.

[Dowe *et al.*, 1992] D. L. Dowe, J. J. Oliver, L. Allison, T. I. Dix, and C. S. Wallace. Learning rules for protein secondary structure prediction. In C. McDonald, J. Rohl, and R. Owens, editors, *Proc. 1992 Department Research Conference*. Dept. Computer Science, University of Western Australia, July 1992.

[Dowe *et al.*, 1992a] D. L. Dowe, J. J. Oliver, L. Allison, C. S. Wallace, and T. I. Dix. A decision graph explanation of protein secondary structure prediction. Technical report CS 92/163, Dept Computer Science, Monash University, Melbourne, Australia, 1992.

[Dowe *et al.*, 1995a] D. L. Dowe, J. J. Oliver, R. A. Baxter, and C. S. Wallace. Bayesian estimation of the von Mises concentration parameter. In *Proc. 15th Int. Workshop on Maximum Entropy and Bayesian Methods, Santa Fe*, July 1995.

[Dowe *et al.*, 1995b] D. L. Dowe, J. J. Oliver, R. A. Baxter, and C. S. Wallace. Bayesian estimation of the von Mises concentration parameter. Technical report CS 95/236, Dept Computer Science, Monash University, Melbourne, Australia, 1995.

[Dowe *et al.*, 1993] D. L. Dowe, J. J. Oliver, T. I. Dix, L. Allison, and C. S. Wallace. A decision graph explanation of protein secondary structure prediction. *26th Hawaii Int. Conf. Sys. Sci.*, 1:669–678, January 1993.

[Dowe *et al.*, 1996e] D. L. Dowe, J. J. Oliver, and C. S. Wallace. 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*, pages 213–227. Springer, October 1996.

[Dowe *et al.*, 1996f] D. L. Dowe, J. J. Oliver, and C. S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. Technical report CS 96/272, Dept Computer Science, Monash University, Melbourne, Australia, 1996.

[Dowe and Oppy, 2001] D. L. Dowe and G. R. Oppy. Universal Bayesian inference? *Behavioral and Brain Sciences (BBS)*, 24(4):662–663, Aug 2001.

[Dowe and Wallace, 1996] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by minimum message length (abstract). In *Proc. Sydney Int. Stat. Congress*, pages 197–198, 1996.

[Dowe and Wallace, 1997a] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics — 28th Symposium on the interface*, volume 28, pages 614–618, 1997.

[Dowe and Wallace, 1997b] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. Technical report TR no. 97/307, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, February 1997. Also in Proc. Sydney Int. Stat. Congr. (SISC-96), Sydney, pages 197-198; and in IMS Bulletin (1996), 25 (4), pp410-411.

[Dowe and Wallace, 1998] D. L. Dowe and C. S. Wallace. Kolmogorov complexity, minimum message length and inverse learning. In W. Robb, editor, *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 144, Queensland, Australia, July 1998.

[Dowe and Zaidi, 2010] D. L. Dowe and N. A. Zaidi. Database normalization as a by-product of minimum message length inference. Im *Proc. 23rd Australian Joint Conference on Artificial Intelligence (AI'2010)*, Adelaide, Australia, 7-10 December 2010, pp. 82–91. Springer Lecture Notes in Artificial Intelligence (LNAI), vol. 6464, Springer, 2010.

[Edgoose and Allison, 1999] T. Edgoose and L. Allison. MML Markov classification of sequential data. *Stats. and Comp.*, 9(4):269–278, September 1999.

[Edgoose *et al.*, 1998] T. Edgoose, L. Allison, and D. L. Dowe. An MML classification of protein structure that knows about angles and sequence. In *Pacific Symposium on Biocomputing '98*, pages 585–596. World Scientific, January 1998.

[Edwards and Dowe, 1998] R. T. Edwards and D. L. Dowe. Single factor analysis in MML mixture modelling. In Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *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)*, pages 96–109, Berlin, April 15–17 1998. Springer.

[Farr and Wallace, 2002] G. E. Farr and C. S. Wallace. The complexity of strict minimum message length inference. *Computer Journal*, 45(3):285–292, 2002.

[Fitzgibbon, 2004] L. J. Fitzgibbon. *Message from Monte Carlo: A Framework for Minimum Message Length Inference using Markov Chain Monte Carlo Methods.* PhD thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia, 2004.

[Fitzgibbon *et al.*, 2002b] L. J. Fitzgibbon, D. L. Dowe, and Lloyd Allison. Change-point estimation using new minimum message length approximations. In *Lecture Notes in Artificial Intelligence (LNAI) 2417, 7th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 244–254. Springer, 2002.

[Fitzgibbon *et al.*, 2002a] L. J. Fitzgibbon, D. L. Dowe, and Lloyd Allison. Univariate polynomial inference by Monte Carlo message length approximation. In *19th International Conference on Machine Learning (ICML)*, pages 147–154, 2002.

[Fitzgibbon *et al.*, 2004] L. J. Fitzgibbon, D. L. Dowe, and F. Vahid. Minimum message length autoregressive model order selection. In *Proc. Int. Conf. on Intelligent Sensors and Information Processing*, pages 439–444, Chennai, India, January 2004.

[Gammerman and Vovk, 2007a] Alex Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *Computer Journal*, 2(50):151–163, 2007.

[Gammerman and Vovk, 2007b] Alex Gammerman and Vladimir Vovk. Rejoinder: Hedging predictions in machine learning. *Computer Journal*, 2(50):173–177, 2007.

[Georgeff and Wallace, 1984a] M. P. Georgeff and C. S. Wallace. A general selection criterion for inductive inference. In T. O'Shea, editor, *Advances in Artificial Intelligence: Proc. Sixth European Conference on Artificial Intelligence (ECAI-84)*, pages 473–482, Amsterdam, September 1984. Elsevier Science Publishers B.V. (North Holland).

[Georgeff and Wallace, 1984b] M. P. Georgeff and C. S. Wallace. A general selection criterion for inductive inference. Technical report TR 44, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, June 1984.

[Georgeff and Wallace, 1985] M. P. Georgeff and C. S. Wallace. Minimum information estimation of structure. In T. O'Shea, editor, *Advances in Artificial Intelligence*, pages 219–228. Elsevier, 1985.

[Gödel, 1931] K. Gödel. On formally undecidable propositions of *Principia mathematica* and related systems I. *Monatshefte fur Mathematik und Physik*, 38:173–198, 1931.

[Good, 1952] I. J. Good. Rational decisions. *J. Roy. Statist. Soc. B*, B 14:107–114, 1952.

[Grünwald, 2007] P. D. Grünwald. *The Minimum Description Length principle (Adaptive Computation and Machine Learning)*. M.I.T. Press, 2007.

[Grünwald *et al.*, 1998] P. D. Grünwald, P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, 1998.

[Grünwald and Langford, 2004] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. In *COLT*, pages 331–347, 2004.

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

[Gupta *et al.*, 2004] G. K. Gupta, I. Zukerman, and D. W. Albrecht. Obituaries: Australia's inspiring leader in the computing revolution — Christopher Wallace computer scientist. *The Age*, page 9, 2004. Melbourne, Australia; near and probably on Fri. 1 Oct. 2004.

[Hernández-Orallo, 2000] José Hernández-Orallo. Beyond the Turing test. *Journal of Logic, Language and Information*, 9(4):447–466, 2000.

[Hernández-Orallo and Dowe, 2010] José Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence Journal*, 174(18):1508–1539, 2010.

[Hernández-Orallo and Minaya-Collado, 1998] José Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems, ICSC Press*, pages 146–163, 1998.

[Hodges, 1983] Andrew Hodges. *Alan Turing: The Enigma*. Simon and Schuster, 1983.

[Hope and Korb, 2002] L. R. Hope and K. Korb. Bayesian information reward. In R. McKay and J. Slaney, editors, *Proc. 15th Australian Joint Conference on Artificial Intelligence — Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Germany, ISSN: 0302-9743, Vol. 2557*, number 2557 in Lecture Notes in Artificial Intelligence (LNAI), pages 272–283. Springer Verlag, 2002.

[Huber, 2008] F. Huber. Milne's argument for the log-ratio measure. *Philosophy of Science*, pages 413–420, 2008.

[Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[Jebara, 2003] Tony Jebara. *Machine Learning: Discriminative and Generative*. Kluwer Academic Publishers, Norwell, MA, U.S.A., 2003.

[Jeffreys, 1946] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. of the Royal Soc. of London A*, 186:453–454, 1946.

[Jorgensen and Gentleman, 1998] M. A. Jorgensen and R. Gentleman. Data mining. *Chance*, 11:34–39, 42, 1998.

[Jorgensen and McLachlan, 2008] M. A. Jorgensen and G. J. McLachlan. Wallace's approach to unsupervised learning: the Snob program. *Computer Journal*, 51(5):571–578, September 2008.

[Kearns *et al.*, 1997] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

[Kissane *et al.*, 1994] D. W. Kissane, S. Bloch, W. I. Burns, J. D. Patrick, C. S. Wallace, and D. P. McKenzie. Perceptions of family functioning and cancer. *Psycho-oncology*, 3:259–269, 1994.

[Kissane *et al.*, 1996] D. W. Kissane, S. Bloch, D. L. Dowe, R. D. Snyder, P. Onghena, D. P. McKenzie, and C. S. Wallace. The Melbourne family grief study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry*, 153:650–658, May 1996.

[Kissane *et al.*, 1996a] D. W. Kissane, S. Bloch, P. Onghena, D. P. McKenzie, R. D. Snyder, and D. L. Dowe. The Melbourne family grief study, II: Psychosocial morbidity and grief in bereaved families. *American Journal of Psychiatry*, 153:659–666, May 1996.

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

[Kontoyiannis, 2008] I. Kontoyiannis. Review of "Information and Complexity in Statistical Modeling by Jorma Rissanen". *American Mathematical Monthly — Reviews*, 115:956–960, 2008.

[Korb and Wallace, 1997] K B Korb and C. S. Wallace. In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. Technical report CS 97/315, Dept Computer Science, Monash University, Melbourne, Australia, 1997.

[Korb and Wallace, 1999] K. B. Korb and C. S. Wallace. In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. *British Jrnl. for the Philosophy of Science*, pages 543–553, 1999. TR 97/315, Mar 1997, Dept. Computer Science, Monash University, Australia 3168.

[Kornienko *et al.*, 2005a] L. Kornienko, D. W. Albrecht, and D. L. Dowe. A preliminary MML linear classifier using principal components for multiple classes. In *Proc. 18th Australian Joint Conference on Artificial Intelligence (AI'2005)*, volume 3809 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 922–926, Sydney, Australia, Dec 2005. Springer.

[Kornienko *et al.*, 2005b] L. Kornienko, D. W. Albrecht, and D. L. Dowe. 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, 2005.

[Kornienko *et al.*, 2002] Lara Kornienko, David L. Dowe, and David W. Albrecht. Message length formulation of support vector machines for binary classification — A preliminary scheme. In *Lecture Notes in Artificial Intelligence (LNAI), Proc. 15th Australian Joint Conf. on Artificial Intelligence*, volume 2557, pages 119–130. Springer-Verlag, 2002.

[Kraft, 1949] L. G. Kraft, 1949. Master's thesis, Dept. of Elec. Eng., M.I.T., U.S.A.

[Lancaster, 2002] A. Lancaster. Orthogonal parameters and panel data. *Review of Economic Studies*, 69:647–666, 2002.

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

[Lewis, 1976] David K. Lewis. Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85(3):297–315, July 1976.

[Li and Vitányi, 1997] Ming Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its applications*. Springer Verlag, New York, 1997.

[Long and Servedio, 2006] Phil Long and Rocco Servedio. Discriminative learning can succeed where generative learning fails. In *The 19th Annual Conference on Learning Theory*, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A., 2006.

[Maheswaran *et al.*, 2006] T. Maheswaran, J. G. Sanjayan, D. L. Dowe, and P. J. Tan. MML mixture models of heterogeneous Poisson processes with uniform outliers for bridge deterioration. In *Lecture Notes in Artificial Intelligence (LNAI) (Springer), Proc. 19th Australian Joint Conf. on Artificial Intelligence*, pages 322 – 331, Hobart, Australia, Dec. 2006.

[Makalic *et al.*, 2003] E. Makalic, L. Allison, and D. L. Dowe. MML inference of single-layer neural networks. In *Proc. of the 3rd IASTED Int. Conf. on Artificial Intelligence and Applications*, pages 636–642, September 2003. See also Technical Report TR 2003/142, CSSE, Monash University, Australia Oct. 2003.

[Martin-Löf, 1966] P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[McKenzie *et al.*, 1993] D. P. McKenzie, P. D. McGorry, C. S. Wallace, L. H. Low, D. L. Copolov, and B. S. Singh. Constructing a minimal diagnostic decision tree. *Methods in Information in Medicine*, 32:161–166, 1993.

[Milne, 1996] P. Milne. $\log[Pr(H|E \cap B)/Pr(H|B)]$ is the one true measure of confirmation. *Philosophy of Science*, 63:21–26, 1996.

[Molloy *et al.*, 2006] S. B. Molloy, D. W. Albrecht, D. L. Dowe, and K. M. Ting. Model-Based Clustering of Sequential Data. In *Proceedings of the 5th Annual Hawaii International Conference on Statistics, Mathematics and Related Fields*, January 2006.

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

[Musgrave and Dowe, 2010] S. Musgrave and D. L. Dowe. Kinship, optimality and typology, *Behavioral and Brain Sciences (BBS)*, 33(5), 2010.

[Needham and Dowe, 2001] S. L. Needham and D. L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In *Proc. 8th Int. Workshop on Artif. Intelligence and Statistics (AI+STATS 2001)*, pages 253–260, Jan. 2001.

[Neil *et al.*, 1999a] J. R. Neil, C. S. Wallace, and K. B. Korb. Learning Bayesian networks with restricted causal interactions. In Kathryn B. Laskey and Henri Prade, editors, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 486–493, S.F., Cal., July 30–August 1 1999. Morgan Kaufmann Publishers.

[Neil *et al.*, 1999b] Julian R. Neil, C. S. Wallace, and K. B. Korb. Bayesian networks with non-interacting causes. Technical Report 1999/28, School of Computer Science & Software Engineering, Monash University, Australia 3168, September 1999.

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

[Oliver, 1993] J. J. Oliver. Decision graphs — an extension of decision trees. In *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, pages 343–350, 1993. Extended version available as TR 173, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia.

[Oliver and Baxter, 1994] J. J. Oliver and R. A. Baxter. MML and Bayesianism: similarities and differences. Technical report TR 94/206, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994.

[Oliver *et al.*, 1998] J. J. Oliver, R. A. Baxter, and C. S. Wallace. Minimum message length segmentation. In Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 222–233, Berlin, April 15–17 1998. Springer.

[Oliver *et al.*, 1996] J. J. Oliver, Rohan A. Baxter, and Chris S. Wallace. Unsupervised learning using MML. In *Proc. 13th International Conference on Machine Learning*, pages 364–372. Morgan Kaufmann, 1996.

[Oliver and Dowe, 1995] J. J. Oliver and D. L. Dowe. Using unsupervised learning to assist supervised learning. In *Proc. 8th Australian Joint Conf. on Artificial Intelligence*, pages 275–282, November 1995. See also TR 95/235, Dept. Comp. Sci., Monash University, Australia 3168, Sep 1995.

[Oliver *et al.*, 1992] J. J. Oliver, D. L. Dowe, and C. S. Wallace. Inferring decision graphs using the minimum message length principle. In *Proc. of the 1992 Aust. Joint Conf. on Artificial Intelligence*, pages 361–367, September 1992.

[Oliver and Hand, 1996] J. J. Oliver and D. J. Hand. Averaging on decision trees. *Journal of Classification*, 1996. An extended version is available as Technical Report TR 5-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.

[Oliver and Hand, 1994] J. J. Oliver and D.J. Hand. Fanned decision trees. Technical report TR 5-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK, 1994.

[Oliver and Wallace, 1991] J. J. Oliver and C. S. Wallace. Inferring decision graphs. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91), workshop 8*, January 1991.

[Oliver and Wallace, 1992] J. J. Oliver and C. S. Wallace. Inferring decision graphs. Technical report CS 92/170, Dept Computer Science, Monash University, Melbourne, Australia, 1992.

[Ooi and Dowe, 2005] J. N. Ooi and D. L. Dowe. Inferring phylogenetic graphs of natural languages using minimum message length. In *CAEPIA 2005 (11th Conference of the Spanish Association for Artificial Intelligence)*, volume 1, pages 143–152, Nov. 2005.

[Papp *et al.*, 1993] E. Papp, D. L. Dowe, and S. J. D. Cox. Spectral classification of radiometric data using an information theory approach. In *Proc. Advanced Remote Sensing Conf.*, pages 223–232, UNSW, Sydney, Australia, July 1993.

[Parry, 2005] Leigh Parry. Midas touch. *The Age newspaper, Melbourne, Australia (Education section)*, page 6 (in Education section), Mon. 20 June 2005. www.TheAge.com.au , www.monash.edu.au/policy/midas.htm.

[Patrick, 1978] J. D. Patrick. *An Information Measure Comparative Analysis of Megalithic Geometries*. PhD thesis, Department of Computer Science, Monash University, Australia, 1978.

[Patrick and Wallace, 1977] J. D. Patrick and C. S. Wallace. Stone circles: A comparative analysis of megalithic geometry. In *Proc. 48th Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conference*. 1977. abstract.

[Patrick and Wallace, 1982] J. D. Patrick and C. S. Wallace. Stone circle geometries: an information theory approach. In D. Heggie, editor, *Archaeoastronomy in the New World*, pages 231–264. Cambridge University Press, 1982.

[Phillips and Ploberger, 1996] P. C. B. Phillips and W. Ploberger. An asymptotic theory of Bayesian inference for time series. *Econometrica*, 64(2):240–252, 1996.

[Pilowsky *et al.*, 1969] I. Pilowsky, S. Levine, and D.M. Boulton. The classification of depression by numerical taxonomy. *British Journal of Psychiatry*, 115:937–945, 1969.

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

[Quinlan and Rivest, 1989] J.R. Quinlan and R.L. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, 1989.

[Rissanen, 1976] J. J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Develop.*, 20(3):198–203, May 1976.

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

[Rissanen, 1996] J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, 42(1):40–47, January 1996.

[Rissanen, 1999a] J. J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Journal*, 42(4):260–269, 1999.

[Rubinstein *et al.*, 2007] B. Rubinstein, P. Bartlett, and J. H. Rubinstein. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. MIT Press, Cambridge, MA, U.S.A., 2007.

[Rumantir and Wallace, 2001] G. W. Rumantir and C. S. Wallace. Sampling of highly correlated data for polynomial regression and model discovery. In *The 4th International Symposium on Intelligent Data Analysis (IDA)*, pages 370–377, 2001.

[Rumantir and Wallace, 2003] G. W. Rumantir and C. S. Wallace. Minimum message length criterion for second-order polynomial model selection applied to tropical cyclone intensity forecasting. In *The 5th International Symposium on Intelligent Data Analysis (IDA)*, pages 486–496, 2003.

[Sanghi and Dowe, 2003] P.S̃anghi and D. L. Dowe. A computer program capable of passing I.Q. tests. In *4th International Conference on Cognitive Science (and 7th Australasian Society for Cognitive Science Conference)*, volume 2, pages 570–575, Univ. of NSW, Sydney, Australia, Jul 2003.

[Schmidhuber, 2007] J. Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Lecture Notes in Computer Science (LNCS) 4755*, pages 26–38. Springer, 2007.

[Schmidt, 2008] D. F. Schmidt. *Minimum Message Length Inference of Autoregressive Moving Average Models*. PhD thesis, Faculty of Information Technology, Monash University, 2008.

[Schmidt and Makalic, 2009b] D. F. Schmidt and E. Makalic. Minimum message length shrinkage estimation. *Statistics & Probability Letters*, 79(9):1155–1161, 2009.

[Schmidt and Makalic, 2009c] D. F. Schmidt and E. Makalic. MML invariant linear regression. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*. Springer, December 2009.

[Schwarz, 1978] G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[Searle, 1980] J. R. Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.

[Shmueli and Koppius, 2007] G. Shmueli and O. Koppius. Predictive vs. Explanatory Modeling in IS Research. In *Proc. Conference on Information Systems & Technology*, 2007. Seattle, Wa, U.S.A. (URL www.citi.uconn.edu/cist07/5c.pdf).

[Solomonoff, 1960] R. J. Solomonoff. A preliminary report on a general theory of inductive inference. Report V-131, Zator Co., Cambridge, Mass., U.S.A., 4 Feb. 1960.

[Solomonoff, 1964] R. J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22,224–254, 1964.

[Solomonoff, 1995] R. J. Solomonoff. The discovery of algorithmic probability: A guide for the programming of true creativity. In P.V̄itanyi, editor, *Computational Learning Theory: EuroCOLT'95*, pages 1–22. Springer-Verlag, 1995.

[Solomonoff, 1996] R. J. Solomonoff. Does algorithmic probability solve the problem of induction? In D. L. Dowe, K. B. Korb, and J. J Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 7–8, Melbourne, Australia, August 1996. World Scientific. ISBN 981-02-2824-4.

[Solomonoff, 1997a] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

[Solomonoff, 1997b] R. J. Solomonoff. Does algorithmic probability solve the problem of induction? Report, Oxbridge Research, P.O.B. 400404, Cambridge, Mass. 02140, U.S.A., 1997. See http://world.std.com/∼rjs/isis96.pdf.

[Solomonoff, 1999] R. J. Solomonoff. Two kinds of probabilistic induction. *Computer Journal*, 42(4):256–259, 1999. Special issue on Kolmogorov Complexity.

[Solomonoff, 2008] R. J. Solomonoff. Three kinds of probabilistic induction: Universal and convergence theorems. *Computer Journal*, 51(5):566–570, September 2008.

[Tan and Dowe, 2002] P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins. In R. McKay and J. Slaney, editors, *Proc. 15th Australian Joint Conference on Artificial Intelligence — Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Germany, ISSN: 0302-9743, Vol. 2557*, number 2557 in Lecture Notes in Artificial Intelligence (LNAI), pages 131–142. Springer Verlag, 2002.

[Tan and Dowe, 2003] P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Lecture Notes in Artificial Intelligence (LNAI) 2903 (Springer), Proc. 16th Australian Joint Conf. on Artificial Intelligence*, pages 269–281, Perth, Australia, Dec. 2003.

[Tan and Dowe, 2004] P. J. Tan and D. L. Dowe. MML inference of oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 3339 (Springer), Proc. 17th Australian Joint Conf. on Artificial Intelligence*, volume 3339, pages 1082–1088, Cairns, Australia, Dec. 2004.

[Tan and Dowe, 2006] P. J. Tan and D. L. Dowe. Decision forests with oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 5th Mexican International Conf. Artificial Intelligence*, pages 593–603, Apizaco, Mexico, Nov. 2006.

[Tan *et al.*, 2007] P. J. Tan, D. L. Dowe, and T. I. Dix. Building classification models from microarray data with tree-based classification algorithms. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 20th Australian Joint Conf. on Artificial Intelligence*, Dec. 2007.

[Torsello and Dowe, 2008a] A. Torsello and D. L. Dowe. Learning a generative model for structural representations. In *Lecture Notes in Artificial Intelligence (LNAI)*, volume 5360, pages 573–583, 2008.

[Torsello and Dowe, 2008b] A. Torsello and D. L. Dowe. Supervised learning of a generative model for edge-weighted graphs. In *Proc. 19th International Conference on Pattern Recognition (ICPR2008)*. IEEE, 2008. 4pp. IEEE Catalog Number: CFP08182 , ISBN: 978-1-4244-2175-6 , ISSN: 1051-4651.

[Turing, 1936] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc. 2*, 42:230–265, 1936.

[Uther and Veloso, 2000] W. T. B. Uther and M. M. Veloso. The Lumberjack Algorithm for Learning Linked Decision Forests. In *Proc. 6th Pacific Rim International Conf. on Artificial Intelligence (PRICAI'2000), Lecture Notes in Artificial Intelligence (LNAI) 1886 (Springer)*, pages 156–166, 2000.

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

[Visser and Dowe, 2007] Gerhard Visser and D. L. Dowe. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In *Proc. 6th IEEE International Conf. on Computer and Information Science (ICIS) 2007*, pages 17–22, July 2007.

[Visser *et al.*, 2009a] Gerhard Visser, D. L. Dowe, and I. D. Svalbe. Information-theoretic image reconstruction and segmentation from noisy projections. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*, pp. 170–179. Springer, December 2009.

[Visser *et al.*, 2009b] Gerhard Visser, D. L. Dowe, and J. Petteri Uotila. Enhancing MML clustering using context data with climate applications. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*, pp. 350–359. Springer, December 2009.

[Viswanathan and Wallace, 1999] M. Viswanathan and C. S. Wallace. A note on the comparison of polynomial selection methods. In D Heckerman and J Whittaker, editors, *Proceedings of Uncertainty 99: The Seventh International Workshop on Artificial Intelligence and Statistics*, pages 169–177, Fort Lauderdale, Florida, USA, January 1999. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA.

[Viswanathan *et al.*, 1999] Murlikrishna Viswanathan, C. S. Wallace, David L. Dowe, and Kevin B. Korb. Finding cutpoints in noisy binary sequences — a revised empirical evaluation. In *Proc. 12th Australian Joint Conference on Artificial Intelligence*, volume 1747 of *Lecture Notes in Artificial Intelligence*, pages 405–416. Springer Verlag, 1999.

[Wallace, 1973a] C. S. Wallace. Simulation of a two-dimensional gas. In *Proc. of the Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conf.*, page 19, August 1973. abstract.

[Wallace, 1984b] C. S. Wallace. An improved program for classification. Technical Report 47, Department of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Melbourne, 1984.

[Wallace, 1984a] C. S. Wallace. Inference and estimation by compact coding. Technical Report 84/46, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, August 1984.

[Wallace, 1986] C. S. Wallace. An improved program for classification. In *Proc. of the 9th Australian Computer Science Conference (ACSC-9)*, pages 357–366, February 1986. Published as Proc. of ACSC-9, volume 8, number 1.

[Wallace, 1990c] C. S. Wallace. Classification by minimum-message-length encoding. In S. G. Akl et al, editor, *Advances in Computing and Information — ICCI '90*, volume 468 of *Lecture Notes in Computer Science (LNCS)*, pages 72–81. Springer-Verlag, May 1990.

[Wallace, 1990b] C S Wallace. Classification by minimum-message-length inference. In *Working Notes AAAI Spring Symposium Series*, pages 65–69. Stanford Uni., Calif., U.S.A., 1990.

[Wallace, 1992] C. S. Wallace. A Model of Inductive Inference. Seminar, November 1992. Also on video, Dept. of Computer Science, Monash University, Clayton 3168, Australia, Wed. 25 Nov. 1992.

[Wallace, 1995a] C. S. Wallace. 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, 1995.

[Wallace, 1996c] C. S. Wallace. False oracles and SMML estimators. In D. L. Dowe, K. B. Korb, and J. J Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 304–316, Melbourne, Australia, August 1996. World Scientific. ISBN 981-02-2824-4. Was previously Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia, June 1989.

[Wallace, 1996b] C. S. Wallace. MML inference of predictive trees, graphs and nets. In A. Gammerman, editor, *Computational Learning and Probabilistic Reasoning*, chapter 3, pages 43–66. Wiley, 1996.

[Wallace, 1997] C. S. Wallace. On the selection of the order of a polynomial model. Technical report, Royal Holloway College, England, U.K., 1997. 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.

[Wallace, 1998d] C. S. Wallace. Competition isn't the only way to go, a Monash FIT graduation address, April 1998. (Perhaps see www.csse.monash.edu.au/∼dld/CSWallacePublications).

[Wallace, 1998a] C. S. Wallace. Intrinsic classification of spatially correlated data. *Computer Journal*, 41(8):602–611, 1998.

[Wallace, 1998b] C. S. Wallace. Multiple factor analysis by MML estimation. In *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 144, Queensland, Australia, July 1998.

[Wallace, 1998c] C. S. Wallace. On the selection of the order of a polynomial model. In W. Robb, editor, *Proc. of the 14th Biennial Australian Statistical Conf.*, page 145, Queensland, Australia, July 1998.

[Wallace, 1998e] C. S. Wallace. PAKDD-98 Tutorial: Data Mining, 15-17 April 1998. Tutorial entitled "*Data Mining*" at the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98), held in Melbourne, Australia. This

partly constituted an early draft of Chris Wallace's 2005 book "*Statistical and Inductive Inference by Minimum Message Length*".

[Wallace, 1999] C. S. Wallace. *The MIT Encyclopedia of the Cognitive Sciences (MITECS)*, chapter Minimum description length, (major review), pages 550–551. The MIT Press, London, England, ISBN: 0-262-73124-X, 1999.

[Wallace, 2005] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer Verlag, May 2005. ISBN 0-387-23795X.

[Wallace and Boulton, 1968] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

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

[Wallace and Dale, 2005] C. S. Wallace and M. B. Dale. Hierarchical clusters of vegetation types. *Community Ecology*, 6(1):57–74, 2005. ISSN: 1585-8553.

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

[Wallace and Dowe, 1994a] C. S. Wallace and D. L. Dowe. Estimation of the von Mises concentration parameter using minimum message length. In *Proc. 12th Aust. Stat. Soc. Conf.*, 1994. 1 page abstract.

[Wallace and Dowe, 1994b] C. S. Wallace and D. L. Dowe. Intrinsic classification by MML — the Snob program. In *Proc. 7th Australian Joint Conf. on Artificial Intelligence*, pages 37–44. World Scientific, November 1994.

[Wallace and Dowe, 1996] C. S. Wallace and D. L. Dowe. MML mixture modelling of Multi-state, Poisson, von Mises circular and Gaussian distributions. In *Proc. Sydney International Statistical Congress (SISC-96)*, page 197, Sydney, Australia, 1996.

[Wallace and Dowe, 1997a] C. S. Wallace and D. L. Dowe. MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions. *Proc 28th Symp. on the Interface*, pages 608–613, 1997.

[Wallace and Dowe, 1997b] C. S. Wallace and D. L. Dowe. MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions. In *Sixth International Workshop on Artificial Intelligence and Statistics, Society for AI and Statistics*, pages 529–536, San Francisco, USA, 1997.

[Wallace and Dowe, 1999a] C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.

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

[Wallace and Dowe, 1999c] C. S. Wallace and D. L. Dowe. Rejoinder. *Computer Journal*, 42(4):345–347, 1999.

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

[Wallace and Freeman, 1987] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B*, 49(3):240–252, 1987. See also Discussion on pp252-265.

[Wallace and Freeman, 1992] C. S. Wallace and P. R. Freeman. Single-factor analysis by minimum message length estimation. *J. Royal Stat. Soc. B*, 54(1):195–209, 1992.

[Wallace and Georgeff, 1983] C. S. Wallace and M. P. Georgeff. A general objective for inductive inference. Technical Report #83/32, Department of Computer Science, Monash University, Clayton, Australia, March 1983. Reissued in June 1984 as TR No. 44.

[Wallace and Korb, 1994] C. S. Wallace and K. B. Korb. A Bayesian learning agent. In C. S. Wallace, editor, *Research conference: Faculty of Computing and Information Technology*, page 19. Monash University Melbourne, 1994.

[Wallace and Korb, 1997] C. S. Wallace and K B Korb. Learning linear causal models by MML sampling. Technical report CS 97/310, Dept Computer Science, Monash University, Melbourne, Australia, 1997.

[Wallace and Korb, 1999] C. S. Wallace and K. B. Korb. Learning linear causal models by MML sampling. In A. Gammerman, editor, *Causal Models and Intelligent Data Management*, pages 89–111. Springer Verlag, 1999. see TR 97/310, Dept. Comp. Sci., Monash Univ., Australia, June 1997.

[Wallace *et al.*, 1996b] C. S. Wallace, K B Korb, and H Dai. Causal discovery via MML. Technical report CS 96/254, Dept Computer Science, Monash University, Melbourne, Australia, 1996.

[Wallace *et al.*, 1996a] C. S. Wallace, Kevin B. Korb, and Honghua Dai. Causal discovery via MML. In *13th International Conf. on Machine Learning (ICML-96)*, pages 516–524, 1996.

[Wallace and Patrick, 1991] C. S. Wallace and J D Patrick. Coding decision trees. Technical report CS 91/153, Dept Computer Science, Monash University, Melbourne, Australia, 1991.

[Wallace and Patrick, 1993] C. S. Wallace and J. D. Patrick. Coding decision trees. *Machine Learning*, 11:7–22, 1993.

[Zakis *et al.*, 1994] J. D. Zakis, I. Cosic, and D. L. Dowe. Classification of protein spectra derived for the resonant recognition model using the minimum message length principle. *Australian Comp. Sci. Conf. (ACSC-17)*, pages 209–216, January 1994.