MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions

Chris S. Wallace and David L. Dowe

Department of Computer Science, Monash University, Clayton, Victoria 3168, Australia
e-mail: {csw, dld}@cs.monash.edu.au

Abstract

Minimum Message Length (MML) is an invariant Bayesian point estimation technique which is also consistent and efficient. We provide a brief overview of MML inductive inference (Wallace and Boulton (1968), Wallace and Freeman (1987)), and how it has both an information-theoretic and a Bayesian interpretation. We then outline how MML is used for statistical parameter estimation, and how the MML mixture modelling program, Snob (Wallace and Boulton (1968), Wallace (1986), Wallace and Dowe(1994)) uses the message lengths from various parameter estimates to enable it to combine parameter estimation with selection of the number of components. The message length is (to within a constant) the logarithm of the posterior probability of the theory. So, the MML theory can also be regarded as the theory with the highest posterior probability. Snob currently assumes that variables are uncorrelated, and permits multi-variate data from Gaussian, discrete multi-state, Poisson and von Mises circular distributions.

1 Introduction - About Minimum Message Length (MML)

The Minimum Message Length (MML)[37, p185][43] (and, e.g., [5, pp63-64][38]) principle of inductive inference is based on information theory, and hence lies on the interface on computer science and statistics. A Bayesian interpretation of the MML principle is that it variously states that the best conclusion to draw from data is the theory with the highest posterior probability or, equivalently, that theory which maximises the product of the prior probability of the theory with the probability of the data occuring in light of that theory. We quantify this immediately below.

Letting D be the data and H be an hypothesis (or theory) with prior probability $\Pr(H)$, we can write the posterior probability $\Pr(H|D) = \Pr(H\&D)/\Pr(D) = \Pr(H).\Pr(D|H)/\Pr(D)$, by repeated application of Bayes's Theorem. Since D and $\Pr(D)$ are given and we wish to infer H, we can regard the problem of maximising the posterior probability, $\Pr(H|D)$, as one of choosing H so as to maximise $\Pr(H).\Pr(D|H)$.

An information-theoretic interpretation of MML is that elementary coding theory tells us that an event of probability p can be coded (e.g. by a Huffman code) by a message of length $l = -\log_2 p$ bits. (Negligible or no harm is done by ignoring effects of rounding up to the next positive integer.)

So, since $-\log_2(\Pr(H).\Pr(D|H)) = -\log_2(\Pr(H)) - \log_2(\Pr(D|H))$, maximising the posterior probability, $\Pr(H|D)$, is equivalent to minimising

$$MessLen = -\log_2(\Pr(H)) - \log_2(\Pr(D|H)) \qquad (1)$$

the length of a two-part message conveying the theory, H, and the data, D, in light of the theory, H. Hence the name "minimum message length" (principle) for thus choosing a theory, H, to fit observed data, D. The principle seems to have first been stated by Solomonoff[31, p20] and was re-stated and apparently first applied in a series of papers by Wallace and Boulton[37, p185][4][5, pp63-64][6, 8, 7, 38] dealing with model selection and parameter estimation (for Normal and multi-state variables) for problems of mixture modelling (also known as clustering, numerical taxonomy or, e.g. [3], "intrinsic classification").

An important special case of the Minimum Message Length principle is an observation of Chaitin[9] that data can be regarded as "random" if there is no theory, H, describing the data which results in a shorter total message length than the null theory results in. For a comparison with the related Minimum Description Length (MDL) work of Rissanen[28, 29], see, e.g., [32]. We discuss later some (other) applications of MML.

2 Parameter Estimation by MML

Given data x and parameters $\vec{\theta}$, let $h(\vec{\theta})$ be the prior probability distribution on $\vec{\theta}$, let $p(x|\vec{\theta})$ be the likelihood, let $L = -\log p(x|\vec{\theta})$ be the negative log-likelihood and let

$$F = E(\sum \partial^2 L/\partial\vec{\theta}\partial\vec{\theta}') \qquad (2)$$

be the Fisher information, the determinant of the (Fisher information) matrix of expected second partial derivatives of the negative log-likelihood. Then the MML estimate of $\vec{\theta}$ [43, p245] is that value of $\vec{\theta}$ minimising the message length,

$$-\log(h(\vec{\theta})p(x|\vec{\theta})/\sqrt{F(\vec{\theta})}) + 1/2\log(e/12) \qquad (3)$$

(If ϵ is the measurement accuracy of the data and N is the number of data things, then we add the constant term $N\log(1/\epsilon)$ to the length of the message. This is elaborated upon elsewhere[43, p245][39, pp1–3].)

The two-part message describing the data thus comprises first, a theory, which is the MML parameter estimate(s), and, second, the data given this theory.

It is reasonably clear to see that a finite coding can be given when the data is discrete or multi-state. For continuous data, we also acknowledge that it must only have been stated to finite precision by virtue of the fact that it was able to be (finitely) recorded. (In practice[41], as below equation (3), we assume that, for a given continuous or circular attribute, all measurements are made to some accuracy, ϵ.) Just as all recorded data is finitely recorded and can be finitely represented, by acknowledging an uncertainty region in the MML estimate of approximately[43, 36, 39] $\sqrt{12/F(\theta)}$, the MML estimate is stated to a (non-zero) finite precision. The MML estimate has a genuine, non-zero, prior *probability* (*not* a density) and can be encoded by a genuine finite code. (Indeed, the object of MML is to choose a finitely stated estimate or hypothesis, H, to make the two-part message of length $-\log_2(\Pr(H)) - \log_2(\Pr(D|H))$ stating H followed by D given H as short as possible.) The MML theory is thus different, in general, from the standard Bayesian maximum a posteriori (MAP) theory.

In the remainder of this section, we give several examples of the result of using the MML formula to obtain parameter estimates from "innocuous" priors. For the Gaussian, multi-state and Poisson distributions, the MML estimate can be written in a simple analytic form and closely approximates the Maximum Likelihood (ML) estimate. For the von Mises distribution, the estimators take a messier form[30, 18, 39, 14].

The following two sections are on extending MML parameter estimation to MML mixture modelling, and on the invariance[43, 38] of MML and the consistency and efficiency[43, 36, 1] of MML. Further sections mention alternative mixture modelling programs, and applications of and extensions to the Snob program.

2.1 Gaussian Variables

For a Normal distribution (with sample size, N), assuming a uniform prior on μ and a scale-invariant, $1/\sigma$ prior on σ, we get that the Maximum Likelihood (ML) and MML estimates of the mean concur, i.e., that $\hat{\mu}_{MML} = \hat{\mu}_{ML} = \bar{x}$. Letting $s^2 = \sum_i (x_i - \bar{x})^2$, we get that $\sigma^2{}_{ML} = s^2/N$ and[37, p190] that

$$\sigma^2{}_{MML} = s^2/(N-1) \qquad (4)$$

corrects this minor but well-known bias in the Maximum Likelihood estimate.

2.2 Discrete, Multi-State Variables

Since multi-state attributes are discrete, the above issues of measurement accuracy do not arise.

For a multi-state distribution with M states, a ("colourless") uniform prior, $h(\vec{p}) = (M-1)!$, is assumed over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$ given by $p_1 + p_2 + ... + p_M = 1;\ \ p_i \geq 0$.

Letting n_m be the number of things in state m and $N = n_1 + ... + n_M$, minimising the message length formula gives that the MML estimate of p_m is given[37, p187(4), pp191-194][41] by

$$\hat{p}_m = (n_m + 1/2)/(N + M/2) \qquad (5)$$

This nominally gives rise to a (minimum) message length[37, p187(4),p194(28)] of

$$(M-1)\log(N/12+1)/2 - \log(M-1)! - \sum_m (n_m + 1/2)\log \hat{p}_m \qquad (6)$$

for both stating the parameter estimates and then encoding the things in light of these parameter estimates.

2.3 Poisson Variables

Earlier versions of Snob originally[37, 33, 34] permitted models of classes whose variables were assumed to come from a combination of either (discrete) multi-state or (continuous) Normal distributions. Snob has since been augmented by permitting Poisson distributions and von Mises circular distributions[39, 40, 14].

With α the population rate, c the total count and t the total time, with a prior on the rate, r, of $h(r) = (1/\alpha).e^{-r/\alpha}$, we get an MML estimate of

$$\hat{r}_{MML} = (c + 1/2)/(t + 1/\alpha) \qquad (7)$$

2.4 von Mises Circular Variables

The von Mises distribution, $M_2(\mu, \kappa)$, with mean direction μ, and concentration parameter, κ, is a circular analogue of the Normal distribution[18, 21, 39], - both being maximum entropy distributions. Letting $I_0(\kappa)$ be the relevant normalisation constant, it has probability density function (p.d.f.)

$$f(x|\mu, \kappa) = (1/2\pi I_0(\kappa))e^{\kappa \cdot \cos(x-\mu)} \tag{8}$$

and corresponds to the distribution of the angle, x, of a circular pendulum in a uniform field (at angle μ) subjected to thermal fluctuations, with κ representing the ratio of field strength to temperature. For small κ, it tends to a uniform distribution and for large κ, it tends to a Normal distribution with variance $1/\kappa$. Circular data arises commonly in many fields[18, 12].

MML estimation of the von Mises concentration parameter, κ, is obtained by minimising the earlier formula for the message length, using[39] a uniform prior on μ in $[0, 2\pi)$ and the prior $h_3(\kappa) = \kappa/(1+\kappa^2)^{3/2}$ on κ. The contrast between MML and ML estimation is sharper for the von Mises distribution than it is for the Normal, multi-state and Poisson distributions, with Monte Carlo simulations[39, pp12-18] showing a very impressive performance by the MML estimator against ML and other classical rivals (e.g. marginalised Maximum Likelihood)[30, 18].

Being able to associate a message length both with the number of components and, in turn, with each component enables us to use (the minimisation of) the message length as a natural metric for model selection.

2.5 Corrections (and missing data)

Additionally, in calculating the length of the part 2 of the message, D given H, appropriate corrections are made (e.g. Shepherd's approximation for the Normal distribution, or when $M > N$ for the multi-nomial distribution) to account for expected effects on this length of rounding-off parameter values to limited precision.

We further note that, in principle, a separate codeword of some length can be set aside for missing data. The transmission of the missing data will thus be of constant length regardless of the hypothesised classification, and as such will affect neither the minimisation of the message length nor the (statistical) inference.

2.6 A note on higher dimensions

A slight saving can be made in the length of the statement of a message of two or more parameters by generalising the 1-dimensional case at the start of this section to permit (e.g.) in 2 dimensions, the uncertainty region to be a hexagon rather than a rectangle since (in short) both hexagons and rectangles tile the Euclidean plane but a hexagon has a smaller (average or) expected squared distance from its centre than a rectangle or any other tiling shape. This is quantified elsewhere[43, 39] in terms of lattice constants[11] for optimally[1] tesselating Voronoi regions.

3 Applying MML to Mixture Modelling - the Snob Program

Snob uses MML for both the model selection (number of components and assignment of data things to components) and parameter estimation (estimating means and standard deviations, etc.). Snob will prefer to hypothesise the existence of an additional component in the data precisely when the information cost of stating the parameter estimates for this additional component is more than offset by the information gain in stating the things assigned to this new component in terms of the newer, more appropriate, parameter estimates. Recall throughout the equivalence[38] between the probability paradigm and the message length paradigm, with an event of probability p corresponding to a message of length $l = -\log_2 p$ bits, and a message of length l bits corresponding to a probability of $p = 2^{-l}$. That stated and understood, it seems conceptually simpler to continue below in the message length paradigm.

3.1 Stating the message – a first draft

Following earlier work[37, 33, 34, 41], we suppose the data (for mixture modelling) to be given as a matrix of D attribute values for each of N "things", with some attribute values possibly missing. We assume the variables to be independent of one another.

The first part of the message, stating the hypothesis, H, comprises several concatenated message fragments, stating in turn:

1a. The number of components. (All numbers are considered equally likely a priori, although this could easily be modified.)

1b. The relative abundance of each component. (Creating names or labels for each component of length $-\log_2$ of the relative abundance, via a Huffman code, gives us a way of referring to components later when, e.g., we wish to say which component a particular data thing belongs to.)

[1] in terms of minimum average squared distance from the centre for a region of unit hyper-volume

1c. For each component, the distribution parameters of the component (as discussed in Section 2). Each parameter is considered to be specified to a precision of the order of its expected estimation error or uncertainty (see Section 2 or, e.g., [39, pp3–4]). For a larger component, the parameters will be encoded to greater precision and hence by longer fragments than for a less abundant component.

1d. For each thing, the component to which it is estimated to belong. (This can be done using the Huffman code referred to in 1(b) above.)

Having stated in part 1 of the message above, our hypothesis, H, about how many components there are and what the distribution parameters (μ, σ, etc.) are for each attribute for each component, in part 2 of the message we need to state the data, D, in light of this hypothesised model, H.

The details of the encoding and of the calculation of the length of part 1 of the message may be found in Section 2 and elsewhere[37, 39]. It is perhaps worth noting here that since our objective is to minimise the message length (and maximise the posterior probability), we never need construct a message - we only need be able to calculate its length.

Given that part 1(d) of the message told us which component each thing was estimated to belong to and that, for each component, part 1(c) gives us the (MML) estimates of the distribution parameters for each attribute, part 2 of the message now encodes each attribute value of each thing in turn in terms of the distribution parameters (for each attribute) for the thing's component.

3.2 Stating the message more concisely using partial assignment

Part 1(d) of the message described in the previous section (§3.1) implicitly restricts us to hypotheses, H, which assert with 100% definiteness which component each thing belongs to. Given that the population that we might encounter could consist of two different but highly over-lapping distributions, forcing us to state definitely which component each thing belongs to is bound to cause us to mis-classify outliers from one distribution as belonging to another. In the case of two overlapping (but distinguishable) 1-dimensional Normal distributions, this would cause us to over-estimate the difference in the component means and under-estimate the component standard deviations.

Since what we seek is a message which enables us to encode the attribute values of each thing as concisely as possible, we note that a shorter message than that of Section 3.1 can be obtained by a probabilistic (or partial) assignment of things to components. The reason for this is that[33, §3][34, p77] if $p(j,x), j = 1, ..., J$, is the probability of component j generating datum x, then the total assignment of x to its best component results in a message length of $-\log(\max_j p(j,x))$ to encode x whereas, letting $P(x) = \sum_j p(j,x)$, a partial assignment of x having probability $p(j,x)/P(x)$ of being in component j results in a shorter message length of $-\log(P(x))$ to encode x. As shown in [33, §3][34, p77][41], this shorter length is achievable by a message which asserts definite membership of each thing by use of a special coding trick.

4 Consistency, invariance and efficiency of MML estimates

If the outcomes of any random process are encoded using a code that is optimal for that process, the resulting binary string forms a completely random process[43, p241]. This fact and the fact that general MML codes are (by definition) optimal implicitly suggest that, given sufficient data, MML will converge as closely as possible to any underlying model. Indeed, MML can be thought of as extending Chaitin's idea of randomness[9] to always trying to fit given data with the shortest possible computer program (plus noise) for generating it. This general convergence result for MML has been explicitly re-stated elsewhere[36, 1]. Similar arguments show that MML estimates are not only consistent, but that they are also efficient, i.e., that they converge to any true underlying parameter value as quickly as possible.

The fact that $\sqrt{F}$ transforms like a prior is a basis used by some to choose $\sqrt{F}$ as a Jeffrey's prior. Although we do not wish to advocate the use of a Jeffrey's prior, we do note that $h/\sqrt{F}$ is invariant under parameter transformation. Since the likelihood function is also invariant under parameter transformation, we see from equation (3) that MML is also invariant under parameter transformation[43, 38].

The problem of model selection and parameter estimation in mixture modelling can, at its worst, be thought of as a problem for which the number of parameters to be estimated grows with the data. It is well known that Maximum Likelihood can become inconsistent (or very inefficient) with such problems, e.g. multiple factor analysis[35] and the Neyman-Scott problem[24, 16].

5 Alternative Bayesian methods

In doing inductive inference of mixture models from data, there are several levels of inference that we might conceivably wish to make. We might wish simply to infer the most likely number of components. Or, alternatively, we might wish to infer the number of components, their relative abundances and the parameter values associated with each component. Or, we might further wish to infer the above and a probabilistic assignment of things to components. It is these last two variations which are variously understood by the term "mixture modelling". Finally, one might wish to infer the number of components and the identities of their members without regard to parameter estimation. This form is often termed "clustering".

MAP (maximum a posteriori) operates on a density and must marginalise over (or integrate out) parameters to estimate memberships, and must likewise marginalise over memberships to estimate parameters. MAP (like penalised likelihood methods) is unable consistently to estimate both parameter values and class memberships. Let us see why this is: consider some estimate of the number of components followed by parameter estimates for each of these components. (We could, for example, have two equally abundant and substantially overlapping 1-dimensional Normal distributions with the same standard deviation, σ.) If we assign each thing to its most probable class, there will be a neat division of things to classes, a division which will not be consistent with the original estimates of means and σ.

Rather than obtain probabilities from densities of real-valued parameters by integrating (as MAP does), MML obtains such probabilities by rounding-off (or quantising)[2] the possible parameter estimates into coding blocks (or uncertainty regions) as discussed in Section 2. By shortening the length of the message to a minimum, MML arrives at the (quantised) theory of the highest *probability* (see Section 1) whose resulting binary string forms[43, p 241][41, p 41] a completely random process. The fact that the first part of the message string[3] and the second part of the message are completely random (and "noise") means that the coding trick[4] causes the assignment of data things to components to be done (pseudo-)randomly in a way which is consistent with the parameter estimates. If we do not minimise the message length (by taking advantage of the coding trick), as with MAP estimation, inconsistencies will arise.

Results of Barron and Cover[1] show MML to be consistent for any i.i.d. problem, and other results[36][43, p 241] show MML (and Strict MML[38, 43]) to be consistent and efficient for problems of arbitrary generality.

Furthermore, whereas MML is known to be invariant[38, 43] under 1-to-1 transformations, the MAP (posterior mode) estimate is known generally not to be invariant under 1-to-1 transformations - e.g., von Mises circular parameter estimation[14] in polar and Cartesian co-ordinates.

While the authors do not advocate MAP, another Bayesian method which the authors do advocate is estimation by minimising the Expected Kullback-Leibler distance (min EKL). Like the MML estimator, min EKL is invariant under re-parameterisation. Work to appear[5] follows Wallace[36] and shows strong similarities between Strict MML[38, 43] and min EKL (as is easily seen in the case of M-state Bernoulli sampling).

6 Alternative mixture modelling programs

The first Snob program (since out-dated)[37] was possibly the first program for Gaussian mixture modelling, although many statistical and machine learning approaches to this problem have been developed since (e.g., McLachlan et al.[23, 22], D. Fisher's CobWeb[17]). Discussions of early alternative algorithms for Gaussian mixture modelling have been given by Boulton[3].

6.1 Comparison with AutoClass II

Like Snob, AutoClass II [10] assumes[6] a prior distribution over the number of classes and independent prior densities over the distribution parameters of the sample class densities. However[34], AutoClass II is not based on a message length criterion, but instead makes a more direct inference of the number of classes, J.

Let V be the vector of abundance and distribution parameters needed to specify a model with J components. Let $P(J)$ be the prior probability of having J components, and let $h(V)$ be the prior probability of the parameters, V. Let X denote the data, i.e. the set of attribute values for all things, and let $P(X|V)$ be the probability of obtaining data X given the J-component model specified by V. The joint probability $P(J,X)$ of J and X is then

$$P(J,X) \;=\; \int h(V)P(X|V)dV \qquad (9)$$

and the posterior probability, $P(J|X)$, of J given the data, X, is

$$P(J|X) \;=\; P(J,X)/(\sum_j P(j,X)) \qquad (10)$$

[2] hence, Peter Cheeseman (private communication) refers to MML as "quantised Bayes"

[3] and part 1d in particular see Section 3.1

[4] see Section 3.2

[5] by the current authors, and R. Baxter and J. Oliver

[6] This sub-section is very much a re-writing of [34, pp78–80].

The calculation of the posterior, $P(J|X)$, requires the calculation of an integral for each possible number of classes, J, in order to obtain the joint probability, $P(J, X)$. The integrand is proportional to the posterior density of the parameters of a J-class model, $h(V) \times P(X|V)$.

AutoClass II approximates the integral by making the assumption that most of the contribution to the integral will come from the neighbourhood of the highest peak value of the integrand. It effectively fits a Gaussian function to the integrand at this peak and uses the integral of the Gaussian as its estimate of the true integral. Letting F be the Fisher information (from Section 2), the estimate is very similar, both analytically and numerically, to the quantity $h(V) \times P(X|V)/\sqrt{F}$, which is what MML (in general) and Snob (in particular) endeavour to maximise. Thus, although AutoClass II is differently motivated from Snob, in practice it gives almost identical results.

6.2 Comparison with other methods

Oliver et al.[25] re-wrote the Gaussian mixture modelling part of Snob[41, 42] by modifying the Bayesian priors and introducing lattice constants[43, 39] (see Section 2.5) and then empirically showed a successful performance of (this slightly modified) Snob against AIC (Akaike's Information Criterion), BIC [28] and other methods.

The literature does not yet seem to contain any alternative algorithms for mixture modelling of von Mises circular and Poisson distributions.

In general, with problems such as mixture modelling or multiple factor analysis where the number of parameters to be estimated increases with (and is potentially proportional to) the amount of data, one must beware Maximum Likelihood and MAP methods, which are both liable[24, 16] to give inconsistent results.

7 Snob (and MML) Applications

Earlier applications of Snob include several to medical, psychological, biological and exploratory geological data, with a survey in [41]. The Poisson module seems to be accurately able to discriminate between pseudo-randomly generated classes from different Poisson distributions. It has also been used to analyse word-counts from a data-set of 17th Century texts. On this data-set, a shorter message length was obtained by using a Normal model than a Poisson model, and hence MML advocated the Normal model. The von Mises module has found clusters in data of several thousand sets of protein dihedral angles[12]. The Poisson module is currently being used to model run lengths of helices and other protein conformations as being a mixture of Poisson distributions. This work should indirectly lead to a better way of predicting protein conformations.

Extensive surveys of Snob applications are given in Patrick[27] and Wallace and Dowe[41], with a recent application of Gaussian mixture modelling to data on members of grieving families is given in Kissane et al.[20].

In applying Snob, a difference of more than 5 to 6 bits[43, p251] or of more than 10 bits[33] might be deemed to be statistically significant under certain modelling conditions.

As well as having been applied to mixture models (discussed here), MML has also been successfully applied to a variety of problems of parameter estimation[37, 38, 43, 36, 39, 40, 14, 16], hypothesis testing[43, 39], Hidden Markov Models[19] and other multi-variate models[43, 36, 44, 35, 16]. Further references are given in [13].

8 Notes on further work and Snob program extensions

The Snob program currently implicitly assumes that variables are independent and uncorrelated. This could be modified to permit single linear (Gaussian) factor analysis[44] or multiple linear (Gaussian) factor analysis[35], or to model correlations via an inverse Wishart or some other such prior.

It would not be too difficult[41] to permit the user to modify the colourless priors (see Section 2) used by Snob to better represent the user's prior beliefs (or knowledge, or bias).

MML estimators have been obtained for the spherical Fisher distribution[15] and work is currently underway[26] to deal with the mixture modelling of these.

When there are two or more overlapping components, a slight inefficiency will arise in the message length calculations since parameters will be stated to a slightly higher than necessary degree of precision. The correction for this can be computationally very slow and has been inspected in the Gaussian case by Baxter[2].

9 Availability of the Snob program

The current version of the Snob program (written in Fortran 77) is freely available for not-for-profit, academic research, and not for re-distribution, from ftp://ftp.cs.monash.edu.au/pub/snob/Snob.README (or from C. S. Wallace). Published or otherwise recorded work using Snob should cite the current paper. User guidelines are given in[41] and in the documentation file, snob.doc .

10 Acknowledgments Section

This work was supported by Australian Research Council (ARC) Grant A49330656 and ARC 1992 small grant 9103169 . The authors wish to thank Graham Farr for valuable feedback.

References

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

[2] R.A. Baxter. Finding Overlapping Distributions with MML. Technical Report 95/244, Dept. of Computer Science, Monash University, Clayton 3168, Australia, November 1995.

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

[4] D.M. Boulton and C.S. Wallace. The information content of a multistate distribution. *Journal of Theoretical Biology*, 23:269–278, 1969.

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

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

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

[8] D.M. Boulton and C.S. Wallace. A comparison between information measure classification. In *Proceedings of ANZAAS Congress*, Perth, August 1973.

[9] G.J. Chaitin. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, 13:547–549, 1966.

[10] P. Cheeseman, M. Self, J. Kelly, W. Taylor, D. Freeman, and J. Stutz. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, pages 607–611, Saint Paul, Minnesota, 1988.

[11] J.H. Conway and N.J.A Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag, London, 1988.

[12] 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 *Proceedings of the 1st Pacific Symposium on Biocomputing (PSB-1)*, pages 242–255, Hawaii, U.S.A., January 1996.

[13] D.L. Dowe and K.B. Korb. Conceptual difficulties with the Efficient Market Hypothesis: Towards a naturalized economics. In D.L. Dowe, K.B. Korb, and J.J. Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 212–223, Melbourne, Australia, August 1996. World Scientific.

[14] D.L. Dowe, J.J. Oliver, R.A. Baxter, and C.S. Wallace. Bayesian estimation of the von Mises concentration parameter. In *Proc. 15th Maximum Entropy Conference, Santa Fe, New Mexico*, August 1995.

[15] D.L. Dowe, J.J. Oliver, and C.S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. In A. et al. Sharma, editor, *Proc. 7th Conf. Algorithmic Learning Theory (ALT'96), LNAI 1160*, pages 213–227, Sydney, Australia, October 1996.

[16] D.L. Dowe and C.S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Sydney International Statistical Congess (SISC-96)*, pages 197–198, Sydney, Australia, 1996.

[17] D.H. Fisher. Conceptual clustering, learning from examples, and inference. In *Machine Learning: Proceedings of the Fourth International Workshop*, pages 38–49. Morgan Kaufmann, 1987.

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

[19] M.P. Georgeff and C.S. Wallace. A general criterion for inductive inference. In T. O'Shea, editor, *Advances in Artificial Intelligence : Proc. Sixth European Conference on Artificial Intelligence*, pages 473–482, Amsterdam, 1984. North Holland.

[20] 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, 1996.

[21] K.V. Mardia. *Statistics of Directional Data*. Academic Press, 1972.

[22] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[23] G.J. McLachlan and K.E. Basford. *Mixture Models*. Marcel Dekker, New York, 1988.

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

[25] J. Oliver, Baxter R., and Wallace C. Unsupervised learning using MML. In *Proc. 13th International Conf. Machine Learning (ICML 96)*, pages 364–372. Morgan Kaufmann, San Francisco, CA, 1996.

[26] J.J. Oliver and D.L. Dowe. Minimum Message Length Mixture Modelling of Spherical von Mises-Fisher distributions. In *Proc. Sydney International Statistical Congess (SISC-96)*, page 198, Sydney, Australia, 1996.

[27] J.D. Patrick. Snob: A program for discriminating between classes. Technical Report TR 91/151, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1991.

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

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

[30] G. Schou. Estimation of the concentration parameter in von Mises-Fisher distributions. *Biometrika*, 65:369–377, 1978.

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

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

[33] C.S. Wallace. An improved program for classification. In *Proceedings of the Nineteenth Australian Computer Science Conference (ACSC-9)*, volume 8, pages 357–366, Monash University, Australia, 1986.

[34] C.S. Wallace. Classification by Minimum-Message-Length inference. In G. Goos and J. Hartmanis, editors, *Advances in Computing and Information – ICCI '90*, pages 72–81. Springer-Verlag, Berlin, 1990.

[35] C.S. Wallace. Multiple Factor Analysis by MML Estimation. Technical Report 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995. submitted to J. Multiv. Analysis.

[36] 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. Was Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia. June 1989.

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

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

[39] C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Technical report TR 93/193, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1993. prov. accepted, Aust. J. Stat.

[40] C.S. Wallace and D.L. Dowe. Estimation of the von Mises concentration parameter using Minimum Message Length. In *Proc. 12th Australian Statistical Soc. Conf.*, Monash University, Australia, 1994.

[41] C.S. Wallace and D.L. Dowe. Intrinsic classification by MML – the Snob program. In C. Zhang, J. Debenham, and D Lukose, editors, *Proc. 7th Australian Joint Conf. on Artificial Intelligence*, pages 37–44. World Scientific, Singapore, 1994. See ftp://ftp.cs.monash.edu.au/pub/snob/Snob.README.

[42] 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 Congess (SISC-96)*, page 197, Sydney, Australia, 1996.

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

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