Statistics and Computing (2000) **10,** 73–83

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

CHRIS S. WALLACE and DAVID L. DOWE

Computer Science and Software Engineering, Monash University, Clayton, Vic. 3168, Australia (csw@cs.monash.edu.au), (dld@cs.monash.edu.au)

Submitted November 1998 and accepted April 1999

Minimum Message Length (MML) is an invariant Bayesian point estimation technique which is also statistically consistent and efficient. We provide a brief overview of MML inductive inference (Wallace C.S. and Boulton D.M. 1968. Computer Journal, 11: 185–194; Wallace C.S. and Freeman P.R. 1987. J. Royal Statistical Society (Series B), 49: 240–252; Wallace C.S. and Dowe D.L. (1999). Computer Journal), 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 C.S. and Boulton D.M. 1968. Computer Journal, 11: 185–194; Wallace C.S. 1986. In: Proceedings of the Nineteenth Australian Computer Science Conference (ACSC-9), Vol. 8, Monash University, Australia, pp. 357–366; Wallace C.S. and Dowe D.L. 1994b. In: Zhang C. *et al.* (Eds.), Proc. 7th Australian Joint Conf. on Artif. Intelligence. World Scientific, Singapore, pp. 37–44. See http://www.csse.monash.edu.au/-dld/Snob.html) uses the message lengths from various parameter estimates to enable it to combine parameter estimation with selection of the number of components and estimation of the relative abundances of the components. The message length is (to within a constant) the logarithm of the posterior probability (*not* a posterior density) 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 within each component, and permits multi-variate data from Gaussian, discrete multi-category (or multi-state or multinomial), Poisson and von Mises circular distributions, as well as missing data. Additionally, Snob can do fully-parameterised mixture modelling, estimating the latent class assignments in addition to estimating the number of components, the relative abundances of the parameters and the component parameters. We also report on extensions of Snob for data which has sequential or spatial correlations between observations, or correlations between attributes.

Keywords: clustering, mixture modelling, minimum message length, MML, Snob, induction, coding, information theory, statistical inference, machine learning, classification, intrinsic classification, unsupervised learning, numerical taxonomy

1. Introduction – about minimum message length (MML)

The Minimum Message Length (MML) (Wallace and Boulton 1968, p. 185, Wallace and Freeman 1987) (and, e.g., Boulton and Wallace (1970, pp. 63, 64), Wallace and Boulton (1975) and Wallace and Dowe (1999)) principle of inductive inference, machine learning and "data mining" is based on information theory, and hence lies on the interface of 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 occurring 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 $\Pr(H \mid D) = \Pr(H\&D)/\Pr(D) = \Pr(H) \cdot \Pr(D \mid H)/\Pr(D)$, 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, $\Pr(H \mid D)$, as one of choosing H so as to maximise $\Pr(H) \cdot \Pr(D \mid H)$.

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 $-\log_2(\Pr(H) \cdot \Pr(D \mid H)) = -\log_2(\Pr(H)) - \log_2(\Pr(D \mid H))$, maximising the posterior probability, $\Pr(H \mid D)$, is equivalent to minimising

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

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. A related principle seems to have first been stated by Solomonoff (1964, p. 20) and was independently stated and apparently first applied in a series of papers by Wallace and Boulton (1968, p. 185) (Boulton and Wallace 1969, 1970 (pp. 63, 64), 1973a,b, 1975, Wallace and Boulton 1975, Boulton 1975) 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. Boulton (1975), "intrinsic classification").

An important special case of the Minimum Message Length principle is an observation of Chaitin (1966) 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 an elaboration on the relation between MML, Chaitin's work (Chaitin 1966) and Kolmogorov complexity, see Wallace and Dowe (1999). For a general justification by the authors of the Bayesian paradigm, see Wallace (1996), Dowe *et al.* (1998), and Wallace and Dowe (1999). For a comparison with the related Minimum Description Length (MDL) work of Rissanen (1978), (1989), (1994), see, e.g., Solomonoff (1995) and Wallace and Dowe (1999).

Beginning with MML parameter estimation, this paper both describes and updates the status of the Snob program (Wallace and Boulton 1968, Wallace 1986, Wallace and Dowe 1994b, 1997) for MML mixture modelling, largely updating and expanding upon Wallace and Dowe (1997). We discuss later, in Section 8, some applications of MML.

2. Parameter estimation by MML

Before we move on to the problem of mixture modelling, we deal with the special case of parameter estimation, which can be thought of as corresponding to mixture modelling with one component.

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

$$F = \det\left\{E\left(\frac{\partial^2 L}{\partial\vec{\theta}\,\partial\vec{\theta}'}\right)\right\} \qquad (2)$$

be the Fisher information, the determinant of the (Fisher information) matrix of expected second partial derivatives of the negative log-likelihood. We now follow equation (1), where the hypothesis, H, is to be a quantised statement of parameter estimates. A Taylor expansion as far as the second-order term of the log-likelihood function, L, gives that (Wallace and Dowe 1993, pp. 1–3, Wallace and Freeman 1987, p. 245, Wallace and Dowe 1999 (Sec. 6.1.2)) the MML estimate of $\vec{\theta}$ (Wallace and Freeman 1987, p. 245) is that value of $\vec{\theta}$ minimising the message length,

$$\begin{aligned}
&-\log\left(\left\{h(\vec{\theta}) \times \sqrt{\frac{12^k}{F(\vec{\theta})}}\right\} \times p(x \mid \vec{\theta})\right) + \frac{k}{2}\\
&= -\log\left(\left\{h(\vec{\theta}) \times \sqrt{\frac{12^k}{F(\vec{\theta})}}\right\}\right) + \left(-\log p(x \mid \vec{\theta}) + \frac{k}{2}\right)\\
&= -\log\left\{\frac{h(\vec{\theta})p(x \mid \vec{\theta})}{\sqrt{F(\vec{\theta})}}\right\} + k\left(-\frac{1}{2}\log 12 + \frac{1}{2}\right) \qquad (3)
\end{aligned}$$

where k is the number of parameters to be estimated.[1]

(If ϵ is the measurement accuracy of the data and N is the number of data things,[2] then we add the constant term $N \log(1/\epsilon)$ to the length of the message. This is elaborated upon elsewhere (Wallace and Freeman 1987, p. 245, Wallace and Dowe 1993, pp. 1–3).

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 (Wallace and Dowe 1994b), as earlier in this section, 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 (Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1993) $\sqrt{12^k/F(\theta)}$, the MML estimate is stated to a (non-zero) finite precision.

The MML estimate thus 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 $-\log_2(\Pr(H)) - \log_2(\Pr(D \mid H))$ 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. This last point seems to be misunderstood by many. To re-iterate (Wallace and Dowe 1999), MML optimises a posterior probability and is invariant under 1-to-1 re-parameterisation whereas MAP optimises a posterior density and (see, e.g., Dowe, Oliver and Wallace (1996) and Dowe *et al.* (1995)) is typically not invariant under 1-to-1 re-parameterisation. To put it another way, the presence of the $\sqrt{F}$ term in equation (3) should more than highlight the difference between MML and MAP.

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 estimator can be written in a simple analytic form and closely approximates the Maximum Likelihood (ML) estimator. For the von Mises distribution, the estimators take a messier form (Schou 1978, Fisher 1993, Wallace and Dowe 1993, Dowe *et al.* 1995) and the MML estimator is less similar to the ML estimator (Wallace and Dowe 1993).

From here through Section 4, we define the mixture modelling (or clustering) problem and then extend MML parameter estimation to MML mixture modelling. Section 5 mentions desirable statistical properties of MML. Sections 6 and 7 mention alternative approaches and mixture modelling programs, and the final sections mention applications of, extensions to and availability of 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 (which corresponds to a uniform prior on $\log\sigma$ and, equivalently, to a $1/\sigma^2$ prior on σ^2), we get that the Maximum Likelihood (ML) and MML estimates of the mean concur, i.e., that $\hat{\mu}_{\mathrm{MML}} = \hat{\mu}_{\mathrm{ML}} = \bar{x}$. Letting $s^2 = \sum_i (x_i - \bar{x})^2$, we get that $\hat{\sigma}^2_{\mathrm{ML}} = s^2/N$. Also, either from Wallace and Boulton (1968, p. 190) or instead by noting (c.f. Dowe and Wallace (1997, Section 4.2.1)) that

$$F(\mu, \sigma^2) = \frac{N}{\sigma^2} \times \frac{N}{2(\sigma^2)^2} = \frac{N^2}{2(\sigma^2)^3}$$

and then minimising expression (3), it follows that

$$\hat{\sigma}^2_{\mathrm{MML}} = \frac{s^2}{N-1} \tag{4}$$

This corrects a minor but well-known small sample bias in the Maximum Likelihood estimate, $\hat{\sigma}^2_{\mathrm{ML}}$.

2.2. *Discrete, multi-state variables*

Since multi-state (or multi-category, or multinomial) 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 $p_1 + p_2 + \cdots + p_M = 1$; $p_i \geq 0$.

Letting n_m be the number of things in state m and $N = n_1 + \cdots + n_M$, minimising the message length formula (3) gives that the MML estimate $\hat{p}_m$ of p_m is given by Wallace and Boulton (1968, p. 187(4), pp. 191–194) and Wallace and Dowe (1994b).

$$\hat{p}_m = \frac{n_m + 1/2}{N + M/2} \tag{5}$$

The slight difference between the MML and the Maximum Likelihood estimates is due to the Fisher information term, a term given for $M = 2$ by $F = N/(p_1(1-p_1))$.

Substituting equation (5) immediately above into the message length equation (3) nominally gives rise to a (minimum) message length (Wallace and Boulton 1968, p. 187(4), p. 194(28)) of

$$\frac{M-1}{2}\log\left(\frac{N}{12}+1\right) - \log(M-1)!$$
$$-\sum_m \left(n_m + \frac{1}{2}\right)\log\hat{p}_m \tag{6}$$

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 (Wallace and Boulton 1968, Wallace 1986, 1990) 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 (Wallace and Dowe 1994b, 1997) by permitting Poisson distributions and von Mises circular distributions (Wallace and Dowe 1993, 1994a, Dowe *et al.* 1995).

With r (or λ) the Poisson rate parameter to be inferred, c the total count and t the total time, we have that

$$F(r) = \frac{t}{r}$$

With α the population rate and a prior on the rate, r, of $h(r) = (1/\alpha)\cdot e^{-r/\alpha}$, minimising the message length expression (3) gives us (Wallace and Dowe 1994b, 1996, 1997) an MML estimate of

$$\hat{r}_{\mathrm{MML}} = \frac{c + 1/2}{t + 1/\alpha} \tag{7}$$

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 (Fisher 1993, Mardia 1972, Wallace and Dowe 1993) – both being maximum entropy distributions. Letting $I_0(\kappa)$ be the relevant normalisation constant, it has probability density function (p.d.f.)

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

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 (Fisher 1993, Dowe *et al.* 1996, Edgoose, Allison and Dowe 1998).

MML estimation of the von Mises concentration parameter, κ, is obtained by minimising the earlier formula (3) for the message length, using (Wallace and Dowe 1993) a uniform prior on μ in $[0, 2\pi)$, the prior $h_3(\kappa) = \kappa/(1+\kappa^2)^{3/2}$ on κ and the Fisher information calculated in equation (9) below. Letting $I_1(\kappa) = dI_0(\kappa)/d\kappa$ and letting $A(\kappa) = d \log(I_0(\kappa))/d\kappa = I_1(\kappa)/I_0(\kappa)$, the Fisher information is given (Wallace and Dowe 1993, 1994b, 1997) by

$$\begin{aligned} F(\mu, \kappa) &= N\kappa A(\kappa) \times N\left(1 - \frac{A(\kappa)}{\kappa} - (A(\kappa))^2\right) \\ &= N^2 \kappa A(\kappa)\left(1 - \frac{A(\kappa)}{\kappa} - (A(\kappa))^2\right) \end{aligned} \tag{9}$$

The contrast between MML and ML estimation is sharper for the von Mises distribution than it is for the Normal, multistate and Poisson distributions, with Monte Carlo simulations (Wallace and Dowe 1993, pp. 12–18) showing a very impressive performance by the MML estimator against ML and other classical rivals (e.g. marginalised Maximum Likelihood) (Schou 1978, Fisher 1993). We have also obtained encouraging results against rival Bayesian methods (Dowe *et al.* 1995).

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. *Missing data (and corrections)*

Additionally, in calculating the length of the second part 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 multinomial 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 code-word 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 (Wallace and Freeman 1987, Wallace and Dowe 1993) in terms of lattice constants (Conway and Sloane 1988) for optimally[3] tesselating Voronoi regions.

3. Mixture modelling

Mixture modelling (Everitt and Hand 1981, Titterington, Smith and Makov 1985, McLachlan and Basford 1988), otherwise known as clustering, intrinsic classification (Boulton 1975, Wallace and Dowe 1994b, Wallace 1998) or numerical taxonomy, involves modelling data as coming from several classes (or components, or clusters).

In mixture modelling problems, we want to estimate the number of components, the relative abundances of the components, and the distributional parameters for each component. The problem changes in a slight but subtly important way when we consider the distinction regarding whether or not we also wish to estimate the *latent class assignments*, i.e., the assignment of all data things to components.

We deal below with this problem of fully parameterised mixture modelling, part of which includes the latent class assignment of data things to components. The section immediately following will address the problem of mixture modelling by using MML. Mention of and comparison with some alternative approaches (Everitt and Hand 1981, Titterington, Smith and Makov 1985, Fisher 1987, McLachlan and Basford 1988, McLachlan 1992, McLachlan *et al.* 1999, Roeder 1994, Cheeseman *et al.* 1988, Stutz and Cheeseman 1994, Dellaportas, Karlis and Xekalaki 1997, Neal 1998, Fraley and Raftery 1998, Jorgensen and Hunt 1996, Hunt and Jorgensen 1999) will be given in Sections 7 and 6.

4. Applying MML to mixture modelling – the Snob program

The MML mixture modelling 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 (Wallace and Boulton 1975, Wallace and Dowe 1999 (Section 4)) 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.

4.1. *Stating the message – a first draft*

Following earlier work (Wallace and Boulton 1968, Wallace 1986, 1990, Wallace and Dowe 1994b), 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 (see Section 2.5). 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 up to some constant (Wallace and Boulton 1968) such as 100, although both the choice of constant and the general choice of prior 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.) With the number of components, M, stated in part 1(a) of the message, the relative abundances are encoded as coming from a multinomial distribution, as in Section 2.2.

1c. For each component, the distribution parameters of the component (as discussed for the various distributions 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., Wallace and Dowe (1993, pp. 3, 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.[4] (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 (Wallace and Boulton 1968, Wallace and Dowe 1993). It is perhaps worth noting here that since our objective is to minimise the message length (and maximise the posterior probability), we never need to construct a message – we only need to 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.

The encoding scheme above is the coding scheme from the original MML mixture modelling paper (Wallace and Boulton 1968), and it certainly does encode both an hypothesis H, and the data, D, given H, or $D \mid H$. However, this coding scheme above is inefficient (Wallace 1986, 1990, Wallace and Dowe 1994b, 1997), as we demonstrate below.

4.2. *Stating the message more concisely using partial assignment*

Part 1(d) of the message described in the previous section (Section 4.1) 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 by choosing the most probable component is bound to cause us to mis-classify outliers from one distribution as belonging to another. In the case of two over-lapping (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 4.1 can be obtained by a probabilistic (or partial) assignment of things to components. The reason for this is that (Wallace 1986 (Section 3), Wallace 1990, p. 77) if $p(j, x)$, $j = 1, \ldots, J$, 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, a saving on datum x of $\log_2(P(x)/\max_j p(j, x))$ bits. As shown by Wallace (1986, Section 3), Wallace (1990, p. 77) and Wallace and Dowe (1994b), this shorter length is achievable by a message which asserts definite membership of each thing by use of a special coding trick.

The essence of this special coding trick which enables us to encode more cheaply using partial assignment than using (as in Section 4.1) total assignment to the most probable component is given in the simple case when we have two equally abundant components with datum x sitting in the middle so that $p(1, x) = p(2, x) = P(x)/2$. Recall that at the relevant point of the message, parts 1(a), 1(b) and 1(c) have already been transmitted, so our job now is to encode parts 1(d) and 2, the choice of component followed by the data given the choice of component, as concisely as possible. In this case, with $p(1, x) = p(2, x) = P(x)/2$, nothing is to be gained by choosing component 1 over component 2 for x, or vice versa. We can save $\log_2(P(x)/\max_j p(j, x)) = 1$ bit of information from our message by assigning x to either component 1 or component 2 at random, since either choice would give rise to the same cost, $-\log p(1, x) = -\log p(2, x)$, in part 2 of the message.

4.3. *Fully-parameterised mixture modelling (using strict MML)*

For fully-parameterised mixture modelling, part 1(d) of the message in Section 4, in which data things are assigned to classes, is essential. This simplifies parameter estimation and message length calculations, since the off-diagonal elements of the Fisher information matrix in equation (2) corresponding to distributional parameters from different components can be assumed to be 0.

The message length expression (3) in MML arises from a quadratic Taylor series approximation to what is known as Strict MML (Wallace and Boulton 1975, Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1999 (Sec. 6.1.1)). The partial assignments of data things to classes in Section 4.2 is a good approximation to the total assignments that Strict MML will do. However, unlike Section 4.1 where the total assignments were always to the more likely class, Strict MML (Wallace 1996) will do the total assignments in such a way that for all intents and purposes, data things will appear to have been randomly assigned to classes with a probability given by $p(j, x)/P(x)$ in Section 4.2.

The approximation to this given in Section 4.2 is what is currently used by the Snob program.

5. 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 (Wallace and Freeman 1987, p. 241, Wallace 1996). This fact and the fact that general (Strict) MML codes are (by definition) optimal implicitly suggest that, given sufficient data, (Strict) MML will converge as closely as possible to any underlying model. (Recall that MML arises from a quadratic Taylor series approximation to Strict MML (Wallace and Boulton 1975, Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1999 (Sec. 6.1.1.)). Indeed, MML can be thought of as extending Chaitin's idea of randomness (Chaitin 1966) to always trying to fit given data with the shortest possible computer program (plus noise) for generating it (Wallace and Dowe 1999). This general convergence result for MML has been explicitly re-stated elsewhere (Barron and Cover 1991, Wallace 1996, Wallace and Freeman 1987). 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 Jeffreys "prior". Although we do not wish to advocate the use of a Jeffreys prior, we do note that $h/\sqrt{F}$ is invariant under 1-to-1 parameter transformations. Since the likelihood function is also invariant under parameter transformation, we see from expression (3) that MML is also invariant under 1-to-1 parameter transformation (Wallace and Freeman 1987, Wallace and Boulton 1975).

The problem of model selection and parameter estimation in mixture modelling can, at its worst,[5] 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. single and multiple factor analysis (Wallace and Freeman 1992, Wallace 1995) and the Neyman-Scott problem (Neyman and Scott 1948, Dowe and Wallace 1997); and the second author has conjectured (Dowe *et al.* 1998, p. 93, Wallace and Dowe 1999) that Maximum Likelihood and many other non-Bayesian techniques are doomed to inconsistency on many problems of this nature.

For this and other reasons, we now consider in Section 6 some alternative approaches (both Bayesian and non-Bayesian) to inference and, in Section 7, some alternative approaches to mixture modelling.

6. Alternative 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, as discussed in Sections 3 and 4.3, 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". Elsewhere throughout this paper, we have tended to use the terms "mixture modelling" and "clustering" interchangeably.

6.1. *Alternative Bayesian methods*

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, as in the discussion in Section 4.2.) 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)[6] 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 Sections 1 and 2) whose resulting binary string forms (Wallace and Freeman 1987, p. 241, Wallace 1996, Wallace and Dowe 1994b, p. 41) a completely random process. The fact that the first part of the message string[7] and the second part of the message are completely random (and "noise") means that the coding trick[8] 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 (as discussed in Section 5) will arise.

Results of Barron and Cover (1991) show MML to be consistent for any i.i.d. problem, and other results (Wallace 1989, Wallace and Freeman 1987, p. 241) show (MML and) Strict MML (Wallace and Boulton 1975, Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1999 (Sec. 6.1.1)) to be consistent and efficient for problems of arbitrary generality.

Furthermore, whereas MML is known to be invariant (Wallace and Boulton 1975, Wallace and Freeman 1987) 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 (Dowe *et al.* 1995) in polar and Cartesian co-ordinates. See Section 2 for further stark contrast between MML and MAP.

While the authors do not advocate MAP, another Bayesian method which the authors do advocate as a point estimation technique is estimation by minimising the Expected Kullback-Leibler distance (min EKL). Like the MML and Strict MML estimators, min EKL is invariant under 1-to-1 re-parameterisation. It also has much in common with MML and Strict MML (Dowe *et al.* 1998). An alternative Bayesian approach is not to do point estimation at all, but rather to sample from the posterior (Neal 1998). It is perhaps interesting to note that results from Wallace (1996) suggest that, as the amount of data increases, a method closely related to Strict MML (such as described in Section 4.3) will resemble more and more closely a single Gibbs sampling from the posterior.

6.2. *Other alternative methods*

With regard to the general problem of inductive inference rather than just the specific inductive problem of mixture modelling, the method of Generalised Cross-Validation (GCV) (Wahba 1990) and the Vapnik-Chervonenkis method of Support Vector Machines (SVMs) (Vapnik 1995) are non-Bayesian, whereas MML is Bayesian. Regarding specific comparisons between GCV, SVM and MML, the authors are aware of a comparison between MML and Vapnik-Chervonenkis (V-C) dimension for the problem of polynomial regression (Viswanathan and Wallace 1999), a problem for which V-C was typically more error-prone than MML and for which V-C was likewise more inclined to over-estimate the degree of the polynomial than MML. In earlier work (see references in Viswanathan and Wallace (1999)) on this same problem of polynomial regression, MML and V-C both considerably outperformed GCV.

Some preliminary investigations (M. Viswanathan *et al.*, 1999) have been carried out comparing MML, GCV and other methods for the problem (due to Kearns *et al.* (1997)) of inferring the model underlying a switching binomial process. Our work to date suggests that MML substantially outperforms GCV and the other methods proposed by Kearns *et al.* (1997) for this problem. These are the only comparisons between MML, SVM and GCV of which the authors are currently aware at the time of writing.

For a general comparison of MML with the related Minimum Description Length (MDL) work of Rissanen (1978), (1989) and Rissanen and Ristad (1994), work which some would argue is Bayesian, see, e.g., Solomonoff (1995) and Wallace and Dowe (1999).

7. Alternative mixture modelling programs

The first Snob program (since out-dated) (Wallace and Boulton 1968) was possibly one of the first programs for Gaussian mixture modelling, although many statistical and machine learning approaches to this problem have been developed since (e.g., McLachlan *et al.* (McLachlan and Basford 1988, McLachlan 1992, McLachlan *et al.* 1999), D. Fisher's CobWeb (Fisher 1987), Everitt and Hand 1981, Titterington, Smith and Makov 1985, AutoClass (Cheeseman *et al.* 1988, Stutz and Cheeseman 1994), MULTIMIX (Jorgensen and Hunt 1996, Hunt and Jorgensen 1999), MCLUST (Fraley and Raftery 1998) and others (Roeder 1994, Neal 1998, Dellaportas, Karlis and Xekalaki 1997)). Discussions of early alternative algorithms for Gaussian mixture modelling have been given by Boulton (1975). An excellent survey of AutoClass (Stutz and Cheeseman 1994), Snob (Wallace and Dowe 1997) and MCLUST (Fraley and Raftery 1998) is given in an article on MULTIMIX by Hunt and Jorgensen (1999).

7.1. *Comparisons with AutoClass*

7.1.1. Comparison with AutoClass II

Like Snob, AutoClass II (Cheeseman *et al.* 1988) assumes[9] a prior distribution over the number of classes and independent prior densities over the distribution parameters of the sample class densities. However (Wallace 1990), AutoClass II is not based on a message length criterion, but instead makes a more direct inference of the number of classes, J (cf. Section 6.1).

Let $\boldsymbol{V}$ 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 $h(\boldsymbol{V})$ be the prior probability of the parameters, $\boldsymbol{V}$. Let X denote the data, i.e. the set of attribute values for all things, and let $P(X \mid \boldsymbol{V})$ be the probability of obtaining data X given the J-component model specified by $\boldsymbol{V}$. The joint probability $P(J, X)$ of J and X is then

$$P(J, X) = \int h(\boldsymbol{V}) P(X \mid \boldsymbol{V})\, d\boldsymbol{V} \tag{10}$$

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

$$P(J \mid X) = \frac{P(J, X)}{\sum_j P(j, X)} \tag{11}$$

The calculation of the posterior, $P(J \mid X)$, 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, $h(\boldsymbol{V}) \times P(X \mid \boldsymbol{V})$.

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 $h(\boldsymbol{V}) \times P(X \mid \boldsymbol{V})/\sqrt{F}$, which is what MML (in general, recall expression (3)) and Snob (in particular) endeavour to maximise. Thus, although AutoClass II is differently motivated from Snob, in practice it gives almost identical results.

7.1.2. More Recent Comparisons with AutoClass

As well as referring the reader to the discussion of AutoClass (Stutz and Cheeseman 1994) in the survey in Hunt and Jorgensen (1999), the authors are also led to believe that more recent versions of AutoClass are likely to use MML, and thus presumably bear a greater resemblance to Snob.

7.2. *Comparison with other methods*

Oliver, Baxter and Wallace (1996) re-wrote the Gaussian mixture modelling part of Snob (Wallace and Dowe 1994b, 1996) by modifying the Bayesian priors and introducing lattice constants (Wallace and Freeman 1987, Wallace and Dowe 1993) (see Section 2.6) and then empirically showed a successful performance of (this slightly modified) Snob against AIC (Akaike's Information Criterion), BIC (Rissanen 1978) and other methods.

The literature (Everitt and Hand 1981, Titterington, Smith and Makov 1985, McLachlan and Basford 1988, Jorgensen and Hunt 1996, Hunt and Jorgensen 1999)[10] does contain at least one alternative algorithm (Dellaportas, Karlis and Xekalaki 1997) for mixture modelling of Poisson (Wallace and Dowe 1994b, 1997) distributions. For mixture modelling of von Mises circular (Wallace and Dowe 1994, 1997, Edgoose and Allison 1998) distributions, the idea is at least discussed for AutoClass by Stutz and Cheeseman (1994).

It is not clear to the authors whether the non-parametric method of Roeder (1994) will be able to distinguish two or more multi-dimensional components all having identical means but not having identical standard deviations. Given sufficient data, Snob can identify such a mixture.

As surveyed by Hunt and Jorgensen (1999) and touched upon here in Section 4.2, for substantially overlapping distributions, MCLUST (Fraley and Raftery 1998) would appear likely from our understanding to under-estimate the class variances and to over-estimate the difference in the class means.

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, recalling Sections 4.2, 5 and 6.1, one must beware Maximum Likelihood and MAP methods, which are both liable (Neyman and Scott 1948, Dowe and Wallace 1997) to give inconsistent results.

8. Snob (and MML) applications

Earlier applications of Snob include several to medical, psychological, biological and exploratory geological data, with a survey by Wallace and Dowe (1994b). 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 (Dowe *et al.* 1996, Edgoose, Allison and Dowe 1998). The Poisson module is currently being used to model run lengths of helices and other protein conformations as being a mixture of Poisson distributions. Recently developed theory (Dowe and Wallace 1998) suggests that this work should indirectly lead to a better way of predicting protein conformations.

Extensive surveys of Snob applications are given by Patrick (1991) and Wallace and Dowe (1994b), with an application of Gaussian mixture modelling to data on members of grieving families being given by Kissane *et al.* (1996) and an application to data on autistic children being given by Prior *et al.* (1998).

In applying Snob, a difference of more than 5 to 6 bits (Wallace and Freeman 1987, p. 251) or of more than 10 bits (Wallace

1986) 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 (Wallace and Boulton 1968, 1975, Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1993, 1994a, Dowe *et al.* 1995, Dowe and Wallace 1997, Viswanathan and Wallace 1999), hypothesis testing (Wallace and Freeman 1987, Wallace and Dowe 1993), Hidden Markov Models (Georgeff and Wallace 1984, Edgoose and Allison 1998) and other multi-variate models (Wallace and Freeman 1987, Wallace 1996, Wallace and Freeman 1992, Wallace 1995, Dowe and Wallace 1997). Further references are given in Dowe and Korb (1996) and Wallace and Dowe (1999).

9. Algorithmic issues and further work

9.1. *Algorithmic issues in Snob*

The default of Snob is the "adjust" command (see Wallace and Dowe (1994b) and snob.doc in Section 10 for further and related details) to iteratively adjust the parameter estimates and the assignment of things to classes using an EM algorithm (Wallace and Boulton 1968, McLachlan and Krishnan 1996). Classes can be forcibly combined and can be split either randomly or with some structure. Batch jobs can be run with a "control" file. With the message length as the objective function, Snob proceeds greedily, although, as in Section 9.2, the current Snob search heuristic could be modified to include simulated annealing.

9.2. *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 (Wallace and Freeman 1992) or multiple linear (Gaussian) factor analysis (Wallace 1995), or to model correlations via an inverse Wishart or some other such prior. Work has been done to deal with sequential correlations (Edgoose and Allison 1998) and spatial correlations (Wallace 1998) in mixture modelling data, and preliminary investigations have been carried out (Edwards and Dowe 1998) in incorporating single factor analysis into MML mixture modelling.

It would be desirable to re-introduce hierarchical clustering (Boulton and Wallace 1973) to Snob, and the current Snob search heuristic could be modified to include simulated annealing.

It would not be too difficult 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 (Dowe, Oliver and Wallace 1996) and work has been done (Oliver and Dowe 1996) 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 and Oliver (1997).

Finally, recently developed theory (Dowe and Wallace 1998, Wallace and Dowe 1999) shows how, in addition to the unsupervised clustering for which it was originally developed, Snob can also be developed and used for supervised learning.

10. Availability and use 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 http://www.csse.monash.edu.au/∼dld/Snob.html (or from C.S. Wallace). Published or otherwise recorded work using Snob should cite the current paper. Detailed user guidelines are given by Wallace and Dowe (1994b) and in the documentation file, snob.doc .

Acknowledgments

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

Notes

1. For refinements to the quadratic Taylor expansion leading to equation (3), see the discussion of lattice constants (Conway and Sloane 1988) in Section 2.6, and similarly see Wallace and Dowe (1999, Sec. 6.1.2). It is perhaps important to note that the quadratic Taylor expansion (Wallace and Dowe 1993, pp. 1–3, Wallace 1987, pp. 245) leading to all these similar expressions essentially assumes that the value of the prior density remains reasonably constant over the uncertainty region of size (Wallace and Freeman 1987, Wallace 1996, Wallace and Dowe 1993) approximately $\sqrt{12^k/F(\theta)}$. Although such an assumption will not always be valid, it is certainly more than reasonable for the Gaussian, multinomial, Poisson and von Mises circular distributions being considered here.
2. The terminology "data thing" dates back to (Wallace and Boulton 1968) where "thing" was preferred to other words such as, e.g., 'item", on the grounds that, in that particular case, "item" is merely the Latin word for "thing".
3. In terms of minimum average squared distance from the centre for a region of unit hyper-volume.
4. Since we are doing fully-parameterised mixture modelling, which includes these latent class assignments.
5. For the fully-parameterised mixture modelling problem (with latent class assignments) in Section 4.3.
6. Hence, Peter Cheeseman (private communication) refers to MML as "quantised Bayes".
7. And part 1d in particular.
8. See Section 4.2.

9. This sub-section is very much a re-writing of Wallace (1990, pp. 78–80).
10. See also http://www.csse.monash.edu.au/∼dld/cluster.html.

References

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

Baxter R.A. and Oliver J.J. 1997. Finding overlapping distributions with MML. Statistics and Computing 10(1): 5–16.

Boulton D.M. 1975. The information criterion for intrinsic classification. Ph. D. Thesis, Dept. Computer Science, Monash University, Australia.

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

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

Boulton D.M. and Wallace C.S. 1973a. An information measure for hierarchic classification. The Computer Journal 16: 254–261.

Boulton D.M. and Wallace C.S. 1973b. A comparison between information measure classification. In: Proceedings of ANZAAS Congress, Perth.

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

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

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

Conway J.H. and Sloane N.J.A. 1988. Sphere Packings, Lattices and Groups. London, Springer Verlag.

Dellaportas P., Karlis D., and Xekalaki E. 1997. Bayesian Analysis of Finite Poisson Mixtures. Technical Report No. 32. Department of Statistics, Athens University of Economics and Business, Greece.

Dowe D.L., Allison L., Dix T.I., Hunter L., Wallace C.S., and Edgoose T. 1996. Circular clustering of protein dihedral angles by minimum message length. In: Proc. 1st Pacific Symp. Biocomp., HI, U.S.A., pp. 242–255.

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

Dowe D.L. and Korb K.B. 1996. Conceptual difficulties with the efficient market hypothesis: towards a naturalized economics. In: Dowe D.L., Korb K.B., and Oliver J.J. (Eds.), Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference, Melbourne, Australia.World Scientific, pp. 212–223.

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

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

Dowe D.L. and Wallace. C.S 1997. Resolving the Neyman-Scott problem by minimum message length. In: Proc. Computing Science and Statistics – 28th Symposium on the Interface, Vol. 28, pp. 614–618.

Dowe D.L. and Wallace C.S. 1998. Kolmogorov complexity, minimum message lenth and inverse learning. In: Proc. 14th Australian Statistical Conference (ASC-14), Gold Coast, Qld., Australia, pp. 144.

Edgoose T.C. and Allison L. 1998. Unsupervised markov classification of sequenced data using MML. In: McDonald C. (Ed.), Proc. 21st Australasian Computer Science Conference (ACSC'98), Singapore. Springer-Verlag, ISBN: 981-3083-90-5, pp. 81–94.

Edgoose T.C., Allison L., and Dowe D.L. 1998. An MML classification of protein structure that knows about angles and sequences. In: Proc. 3rd Pacific Symp. Biocomp. (PSB-98) HI, U.S.A., pp. 585–596.

Edwards R.T. and Dowe D.L. 1998. Single factor analysis in MML mixture modelling. In: Proc. 2nd Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98), Melbourne, Australia. springer Verlag, pp. 96–109.

Everitt B.S. and Hand D.J. 1981. Finite Mixture Distributions. London, Chapman and Hall.

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

Fisher N.I. 1993. Statistical Analysis of Circular Data. Cambridge University Press.

Fraley C. and Raftery A.E. 1998. Mclust: software for model-based clustering and discriminant analysis. Technical Report TR 342, Department of Statistics, Univeristy of Washington, U.S.A. Journal of Classification, to appear.

Georgeff M.P. and Wallace C.S. 1984. A general criterion for inductive inference. In: O'Shea T. (Ed.), Advances in Artificial Intelligence: Proc. Sixth European Conference on Artificial Intelligence, Amsterdam. North Holland, pp. 473–482.

Hunt L.A. and Jorgensen M.A. 1999. Mixture model clustering using the multimix program. Australian and New Zealand Journal of statistics 41(2): 153–171.

Jorgensen M.A. and Hunt L.A. 1996. Mixture modelling clustering of data sets with categorical and continous variables. In: Dowe D.L., Korb K.B., and Oliver J.J. (Eds.), Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference, Melbourne, Australia. World Scientific, pp. 375–384.

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

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

Mardia K.V. 1972. Statistics of Directional Data. Academic Press.

McLachlan G.J. 1992. Discriminant Analysis and Statistical Pattern Recognition. New York, Wiley.

McLachlan G.J. and Basford. K.E. 1998. Mixture Models. New York, Marcel Dekker.

McLachlan G.J. and Krishnan T. 1996. The EM Algorithm and Extensions. New York, Wiley.

McLachlan G.J., Peel D., Basford K.E., and Adams P. 1999. The EMMIX software for the fitting of mixtures of Normal and t-components. Journal of Statistical Software 4, 1999.

Neal R.M. 1998. Markov chain sampling methods for dirichlet process mixture models. Technical Report 9815, Dept. of Statistics and Dept. of Computer Science, University of Toronto, Canada, pp. 17.

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

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

Oliver J.J. and Dowe D.L. 1996. Minimum message length mixture modelling of spherical von Mises-Fisher distributions. In: Proc. Sydney International Statistical Congress (SISC-96), Sydney, Australia, p. 198.

Patrick J.D. 1991. Snob: A program for discriminating between classes. Technical report TR 151, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia.

Prior M., Eisenmajer R., Leekam S., Wing L., Gould J., Ong B., and Dowe D.L. 1998. 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.

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

Rissanen. J.J. 1989. Stochastic Complexity in Statistical Inquiry. Singapore, World Scientific.

Rissanen J.J. and Ristad E.S. 1994. Unsupervised Classfication with stochastic complexity. In: Bozdogan H. *et al.* (Ed.), Proc. of the First US/Japan Conf. on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, pp. 171–182.

Roeder K. 1994. A graphical technique for determining the number of components in a mixture of normals. Journal of the American Statistical Association 89(426): 487–495.

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

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

Solomonoff R.J. 1995. The discovery of algorithmic probability: A guide for the programming of true creativity. In: Vitanyi P. (Ed.), Computational Learning Theory: EuroCOLT'95. Springer-Verlag, pp. 1–22.

Stutz J. and Cheeseman P. 1994. Autoclass: A Bayesian approach to classfication. In: Skilling J. and Subuiso S. (Eds.), Maximum Entropy and Bayesian Methods. Dordrecht, Kluwer academic.

Titterington D.M., Smith A.F.M., and Makov U.E. 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, Inc.

Vapnik V.N. 1995. The Nature of Statistical Learning Theory. Springer.

Viswanathan M. and Wallace C.S. 1999. A note on the comparison of polynomial selection methods. In: Proc. 7th Int. Workshop on Artif. Intelligence and Statistics. Morgan Kaufmann, pp. 169–177.

Viswanathan M., Wallace C.S., Dowe D.L., and Korb K.B. 1999. Finding cutpoints in Noisy Binary Sequences. In: Proc. 12th Australian Joint Conf. on Artif. Intelligence.

Wahba G. 1990. Spline Models for Observational Data. SIAM.

Wallace C.S. 1986. An improved program for classfication. In: Proceedings of the Nineteenth Australian Computer Science Conference (ACSC-9), Vol. 8, Monash University, Australia, pp. 357–366.

Wallace C.S. 1990. Classfication by Minimum Message Length inference. In: Goos G. and Hartmanis J. (Eds.), Advances in Computing and Information – ICCI'90. Berlin, Springer-Verlag, pp. 72–81.

Wallace C.S. 1995. Multiple factor analysis by MML estimation. Technical Report 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia. J. Multiv. Analysis, (to appear).

Wallace C.S. 1989. False Oracles and SMML Estimators. In: Dowe D.L., Korb K.B., and Oliver J.J. (Eds.), Proceedings of the Information, Statistics and Induction in science (ISIS) Conference, Melbourne, Australia. World Scientific, pp. 304–316, Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia.

Wallace C.S. 1998. Intrinsic Classification of Spatially-Correlated Data. Computer Journal 41(8): 602–611.

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

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

Wallace C.S. and Dowe D.L. 1993. MML estimation of the von Mises concentration parameter. Technical Report TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton 3168, Australia. Aust. and N.Z. J. Stat, prov. accepted.

Wallace C.S. and Dowe D.L. 1994. Estimation of the von Mises concentration parameter using minimum message length. In: Proc. 12th Australian Statistical Soc. Conf., Monash University, Australia.

Wallace C.S. and Dowe D.L. 1994. Intrinsic classification by MML – the Snob program. In: Zhang C. *et al.* (Eds.), Proc. 7th Australia Joint Conf. on Artif. Intelligence. World Scientific, Singapore, pp. 37–44. See http://www.csse.monash.edu.au/-dld/Snob.html.

Wallace C.S. and Dowe D.L. 1996. MML mixture modelling of Multistate, Poisson, von Mises circular and Gaussian distributions. In: Proc. Sydney International Statistical Congress (SISC-96), Sydney, Australia, p. 197.

Wallace C.S. and Dowe D.L. 1997. MML mixture modelling of Multistate, Poisson, von Mises circular and Gaussian distributions. In: Proc. 6th Int. Workshop on Artif. Intelligence and Statistics, pp. 529–536.

Wallace C.S. and Dowe D.L. 1999. Minimum Message Length and Kolmogorov Complexity. Computer Journal (Special issue on Kolmogorov Complexity) 42(4): 270–283.

Wallace C.S. and Freeman P.R. 1987. Estimation and inference by compact coding. J. Royal Statistical Society (Series B), 49: 240–252.

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