UNSUPERVISED LEARNING OF GAMMA MIXTURE MODELS USING MINIMUM MESSAGE LENGTH

YUDI AGUSTA
Computer Science and Software Engineering
Monash University
Clayton, VIC 3800 Australia
email: yagusta@bruce.csse.monash.edu.au

DAVID L. DOWE
Computer Science and Software Engineering
Monash University
Clayton, VIC 3800 Australia

ABSTRACT
Mixture modelling or unsupervised classification is a problem of identifying and modelling components in a body of data. Earlier work in mixture modelling using Minimum Message Length (MML) includes the multinomial and Gaussian distributions (Wallace and Boulton, 1968), the von Mises circular and Poisson distributions (Wallace and Dowe, 1994, 2000) and the t distribution (Agusta and Dowe, 2002a, 2002b). In this paper, we extend this research by considering MML mixture modelling using the Gamma distribution. The point estimation of the distribution was performed using the MML approximation proposed by Wallace and Freeman (1987) and gives impressive results compared to Maximum Likelihood (ML). We then considered mixture modelling on artificially generated datasets and compared the results with two other criteria, AIC and BIC. In terms of the resulting number of components, the results were again impressive. Application to the Heming Pike dataset was then examined and the results were compared in terms of the probability bitcostings, showing that the proposed MML method performs better than AIC and BIC. A further application also shows that our method works well with datasets containing left-skewed components such as the Palm Valley (Australia) image dataset.

KEY WORDS
Unsupervised Classification, Mixture Modelling, MML, Gamma

1 Introduction

Mixture modelling [1, 2, 3] - generally known as clustering or unsupervised classification - models, as well as partitions, an unknown number of components (or classes or clusters) of a dataset into a finite number of components. In this paper, we discuss, in particular, a classification problem which models a statistical distribution by a mixture (a weighted sum) of other distributions. This type of classification results in a model described by four elements: (1a) the number of components, (1b) the relative abundances (or mixing proportions) of each component, (1c) their distribution parameters and (1d) the members (or things) that belong to the components.

In selecting the most appropriate number of components in a dataset, the problem we often face is keeping the balance between model complexity and goodness of fit. In other words, the best model for a dataset must be sufficiently complex in order to cover all information in the dataset, but not so complex as to over-fit. A series of papers by Wallace and co-authors [1, 4, 5], dealt with model selection and parameter estimation problems using the Minimum Message Length (MML) principle, which provides a fair comparison between models by stating each of them as a two-part message which encodes both model and the data in light of the model stated. Various related principles have also been stated independently by Solomonoff [6], Kolmogorov [7], Chaitin [8], and subsequently by Rissanen [9]. For an overview, see [5].

The MML mixture modelling proposed in [1] dealt with classification problems of discrete multinomial and continuous Gaussian distributions. It was extended by Wallace and Dowe [10, 11, 12] to accommodate two other distributions - Poisson and von Mises circular. The method was further broadened by Agusta and Dowe [13, 14] to accommodate the t distributions with known and unknown degrees of freedom.

Beginning with parameter estimation, this paper extends the application of the MML principle to the problem of mixture modelling by considering the Gamma distribution. We compare the parameter estimation results with the Maximum Likelihood method in terms of their Kullback-Leibler distances from the true model to the inferred model. We also analyse mixture modelling results and compare them with the results of two other commonly used criteria, AIC and BIC, in terms of the resulting number of components. Applications to two real-world datasets: Heming Pike dataset and Palm Valley (Australia) image dataset are also provided.

2 MML Parameter Estimation

The Minimum Message Length (MML) principle is an invariant Bayesian point estimation and model selection technique based on information theory. The basic idea of MML is to find a model that minimises the total length of a two-part message encoding the model, and the data in light of that model [1, 4, 5].

Letting D be the data and H be a model with prior probability $P(H)$, using Bayes's theorem, the point estimation and model selection problems can be regarded simultaneously as a problem of maximising the posterior probability $P(H) \cdot P(D|H)/P(D)$. From the information-theoretic point of view, where an event with probability p is encoded by a message of length $l = -\log_2 p$ bits, the problem is then equivalent to minimising

$$\mathrm{MessLen} = -\log_2(P(H)) - \log_2(P(D|H)) \qquad (1)$$

where the first term is the message length of the model and the second term is the message length of the data in light of the model.

In applying the MML principle to the mixture problem of Gamma distributions, we need firstly to perform parameter estimation of the Gamma distribution. The parameter estimation used here utilises the MML approximation proposed by Wallace and Freeman [4].

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

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

the Fisher information - that is the determinant of the matrix of expected second derivatives of the negative log-likelihood. Based on (1), and by expanding the negative log-likelihood, L, as far as the second term of the Taylor series about the parameter $\vec{\theta}$, the message length is then calculated by [4, 12]:

$$\begin{aligned}\mathrm{MessLen} &= -\log\Big(\frac{h(\vec{\theta})}{\sqrt{\kappa_D^D\, F(\vec{\theta})}}\Big) + L + \frac{D}{2}\\ &= -\log\Big(\frac{h(\vec{\theta})\; f(x|\vec{\theta})}{\sqrt{F(\vec{\theta})}}\Big) + \frac{D}{2}(1+\log\kappa_D) \qquad (3)\end{aligned}$$

where D is the dimension of the parameter space and κ_D is a D-dimensional lattice constant with $\kappa_1 = 1/12$ and $\kappa_D \leq 1/12$ [4]. The MML estimate of $\vec{\theta}$ can be obtained by minimising (3).

Considering that the Gamma distribution is continuous, a finite coding for the message can be obtained by acknowledging that all recorded continuous data and measurements must be stated to a finite precision, which is, in practice, made only to some precision, ϵ. In this way, a constant of $N\log(1/\epsilon)$ is added to the message length expression above, where N is the number of data [12, p74] [13, Sec. 2] [10, p38].

2.1 Multi-state Distribution

For a multi-state distribution with M states (and sample size, N), the likelihood of the distribution is given by:

$$f(n_1, n_2, \cdots, n_M|p_1, p_2, \cdots, p_M) = p_1^{n_1} p_2^{n_2} \cdots p_M^{n_M}$$

where $p_1 + p_2 + \cdots + p_M = 1$, for all m: $p_m \geq 0$ and $n_1 + n_2 + \cdots + n_M = N$.

Using (2), it follows that:

$$F(p_1, p_2, \cdots, p_M) = N^{(M-1)}/p_1 p_2 \cdots p_M.$$

The derivation of this equation is also shown elsewhere for $M = 2$ [12, p75].

Assuming a uniform prior of $h(\vec{p}) = (M-1)!$ over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$, and minimising (3), the MML estimate $\hat{p}_m$ is [10, 11, 12, 13, 14, 15]:

$$\hat{p}_m = (n_m + 1/2)/(N + M/2) \qquad (4)$$

Substituting (4) into (3) provides the total two-part message length [1, p187 (4)] [1, p194 (28)] [12, p75]:

$$\begin{aligned}&-\log(M-1)! + ((M-1)/2)(\log(N\kappa_{M-1}) + 1)\\ &-\sum_{m=1}^{M}(n_m + 1/2)\log\hat{p}_m \qquad (5)\end{aligned}$$

2.2 Gamma Distribution

The Gamma distribution has a likelihood function:

$$f(x|\beta,\gamma) = \frac{(\frac{x}{\beta})^{\gamma-1} e^{(-\frac{x}{\beta})}}{\beta\Gamma(\gamma)} \qquad x \geq 0; \gamma, \beta > 0$$

where β is the scale parameter, γ is the shape parameter and $\Gamma(x)$ is the Gamma function, given by:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

where we let here that $\psi(x) = \mathrm{d}\Gamma(x)/\mathrm{d}x$ and $\psi^{(1)}(x) = \mathrm{d}^2\Gamma(x)/\mathrm{d}x^2$. For any positive integer, x, $\Gamma(x) = (x-1)!$. For large x, the direct calculation from the Gamma function definition above results in a very large value which can not be calculated precisely. Instead, Stirling's asymptotic representation of the Gamma function can be used to approximate the function:

$$\Gamma(x) \approx e^{-x} x^{x-\frac{1}{2}} \sqrt{2\pi}(1 + \frac{1}{12x} + \frac{1}{288x^2} + O(|x|^{-3}))$$

In estimating the second parameter, γ, we consider two separate cases: firstly, γ as a known parameter and secondly, γ as an unknown continuous parameter. Using (2), the Fisher information for the first case, $F(\beta)$, is:

$$F(\beta) = \frac{N\gamma}{\beta^2}$$

and for the second case, $F(\beta,\gamma)$, is:

$$F(\beta,\gamma) = \frac{N^2}{\beta^2}\big(\gamma\psi^{(1)}(\gamma) - 1\big)$$

Here, we assume a $1/\beta$ prior on β over the range [e^{-8}, e^{8}] for both cases and a $2/\pi(1+\gamma^2)$ prior on γ over

β=1.0 & γ=		1.0	3.0	10.0	50.0	100.0	200.0
N=10	ML	0.153±0.23	0.220±0.36	0.227±0.59	0.182±0.22	0.179±0.27	0.113±0.11
	MML	0.112±0.16	0.137±0.21	0.140±0.36	0.125±0.11	0.118±0.14	0.099±0.09
N=100	ML	0.011±0.01	0.012±0.01	0.010±0.01	0.011±0.01	0.011±0.01	0.013±0.02
	MML	0.011±0.01	0.011±0.01	0.010±0.01	0.011±0.01	0.010±0.01	0.012±0.02
β=5.0 & γ=		1.0	3.0	10.0	25.0	100.0	200.0
N=10	ML	0.196±0.30	0.138±0.19	0.148±0.17	0.193±0.36	0.149±0.21	0.115±0.12
	MML	0.141±0.19	0.097±0.11	0.100±0.10	0.138±0.25	0.103±0.11	0.096±0.09
N=100	ML	0.010±0.01	0.011±0.01	0.010±0.01	0.013±0.01	0.011±0.01	0.010±0.01
	MML	0.010±0.01	0.010±0.01	0.010±0.01	0.012±0.01	0.010±0.01	0.010±0.01

Table 1. Kullback-Leibler distances of the ML and MML estimations of 100 datasets for Gamma distributions with $\beta = \{1.0, 5.0\}$ and $\gamma = \{1.0, 3.0, 10.0, 50.0, 100.0, 200.0\}$(with ± standard errors).

the range (0, ∞] for the second case. Minimising (3), the first MML estimate, $\hat{\beta}_{\text{MML}}$, gives:

$$\hat{\beta}_{\text{MML}} = \sum_{i=1}^{N} \frac{x_i}{N\gamma} \tag{6}$$

for both cases. Since there is no sufficient statistic, the second MML estimate, $\hat{\gamma}_{\text{MML}}$, is estimated by setting $\partial \text{MessLen}/\partial\gamma = 0$ and performing a binary search over the parameter space. The search process is terminated when a certain precision of estimation is obtained.

2.3 Point Estimation of One Univariate Gamma Model

In this subsection, we compare the MML parameter estimations of one-component univariate Gamma models to the results obtained using the Maximum Likelihood (ML) method in terms of their Kullback-Leibler (KL) distances.

The Kullback-Leibler distance between two continuous models $P(X)$ and $Q(X)$ is calculated by:

$$D(P||Q) = \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} dx$$

where, in this calculation, $P(X)$ is regarded as the true model and $Q(X)$ is the inferred model.

The experiment datasets for Table 1 were generated repeatedly (100 times) from Gamma distributions with $\beta = \{1.0, 5.0\}$, $\gamma = \{1.0, 3.0, 10.0, 50.0, 100.0, 200.0\}$ and the number of data, $N = \{10, 100\}$. Both methods infer the two estimates as unknown continuous parameters. As shown in Table 1, the MML estimations *always* resulted in estimates closer to the true model, with smaller Kullback-Leibler distances for *all* values of β, γ and N. The MML also performed a more robust estimation, with smaller or equal standard errors of the resulting Kullback-Leibler distances. This case of MML outclassing ML is reminiscent of the von Mises circular distribution [16] as well as the t distribution [13, 14].

3 MML Mixture Modelling

In order to apply MML to a mixture modelling problem, a two-part message conveying the mixture model needs to be constructed. Recall that from Section 2, the model for the mixture comprises several concatenated message fragments, stating in turn:

1a. The number of components: Assuming that all numbers are considered as equally likely up to some constant, (say, 100), this part can be encoded using a uniform distribution over the range.

1b. The relative abundances (or mixing proportions) of each component: Considering the relative abundances of an M-component mixture, this is the same as the condition for an M-state multinomial distribution. The parameter estimation and the message length calculation of the multi-state distribution have been elaborated upon in subsection 2.1.

1c. For each component, the distribution parameters of its attributes: In this case, the component attribute is inferred as a Gamma distribution as elaborated in subsection 2.2.

1d. For each thing, the component to which the thing is estimated to belong.

For the item (1d), instead of the total assignment as proposed in [1], the partial assignment is utilised [17, 10, 11, 12, 13, 14]. In the partial assignment, the data are assigned partially to each component with a certain probability which costs $-\log(P(x))$, where $P(x)$ is the total probability of any component generating datum x. For further discussion, see [17, Sec. 3] [12, pp77-78][13, Sec. 5].

Once the first part of the message is stated, the second part of the message will encode the data in light of the model stated in the first part of the message. Since the objective of the MML principle is to find the model that minimises the message length, we do not need to actually encode the message. In other words, we only need to calculate the length of the message and find the model that gives the shortest/minimum message length.

4 Alternative Model Selection Criteria - AIC and BIC

For comparison, two criteria are considered here. These are the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC).

AIC was first developed by Akaike [18] in order to identify the model of a dataset and is given by:

$$\mathrm{AIC} = -2L + 2N_p$$

where L is the logarithm of the likelihood at the maximum likelihood solution for the investigated mixture model and N_p is the number of parameters estimated. Drawing on the work by Sclove [19], for the Gamma mixture model, N_p is set equal to $k - 1 + kn$ when γ is known and $k - 1 + 2kn$ when γ is unknown, where k is the number of components and n is the number of variables in the dataset. Despite some drawbacks of this criterion in selecting the number of components for a mixture modelling problem, it is still often used to assess the order of a mixture model [3]. The model which results in the smallest AIC is the model selected for the dataset.

The second criterion, BIC, was first introduced by Schwarz [20] and is given by:

$$\mathrm{BIC} = -2L + N_p \log N$$

where L is the logarithm of the likelihood at the maximum likelihood solution for the investigated mixture model, N_p is the number of parameters estimated and N is the number of data. Based on the discussion by Fraley and Raftery [21], for the Gamma mixture model, N_p is set equal to nk when γ is known and $2kn$ when γ is unknown, where k is the number of components and n is the number of variables in the dataset. (This model selection criterion is formally, not conceptually, the same as the Minimum Description Length (MDL) criterion proposed by Rissanen [9].) The model which results in the smallest BIC is selected as the best model for the dataset.

5 Experiments

We applied the proposed method to some artificial datasets and compared the results with two other criteria: AIC and BIC. Applications to the Heming Pike dataset and the Palm Valley (Australia) image dataset are also considered.

5.1 Artificial Datasets

In this experiment, we generated two artificial univariate three-component mixture datasets with $\beta_1 = 10$, $\{\beta_2, \beta_3\} = \{\{25, 40\}, \{27, 44\}, \{29, 48\}, \{31, 52\}, \{33, 56\}, \{35, 60\}\}$ and $\gamma_1 = \gamma_2 = \gamma_3 = 5$ for the first dataset, and $\gamma_1 = 5$, $\{\gamma_2, \gamma_3\} = \{\{16, 27\}, \{18, 31\}, \{20, 35\}, \{22, 39\}, \{24, 43\}, \{26, 47\}\}$ and $\beta_1 = \beta_2 = \beta_3 = 5$ for the second. The mixing proportions of the three components in both datasets were 1:1:1, out of 300 samples. The two datasets were modelled with the parameter γ, set as a known parameter and unknown continuous parameter, respectively. The measurement accuracy, ϵ, is set equal to 0.01 for both cases. We repeated each modelling 20 times.

Comparison of the MML modelling results with those obtained using AIC and BIC can be seen in Figure 1 for the first case, and Figure 2 for the second case. As shown in both graphs, MML performed better, in most cases, than AIC and BIC. AIC performed badly as none of the datasets were correctly modelled.

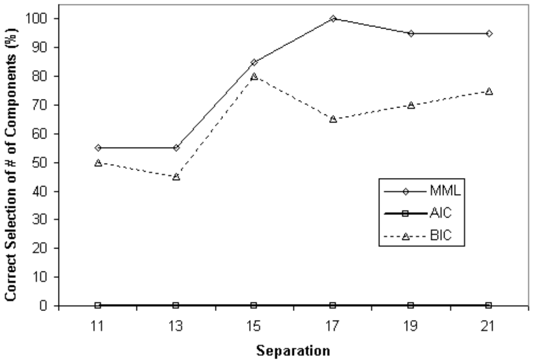

Figure 1. Comparison of MML, AIC and BIC, when modelling the first artificial dataset.

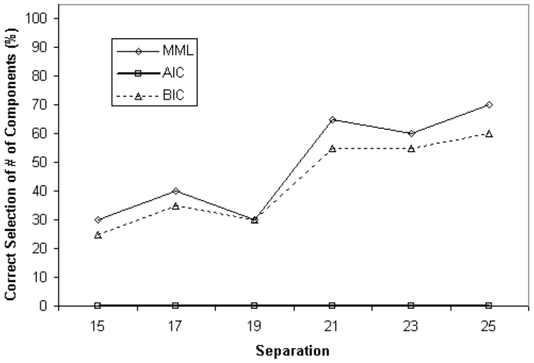

Figure 2. Comparison of MML, AIC and BIC, when modelling the second artificial dataset.

5.2 Heming Pike

The Heming Pike dataset consists of 523 data, measuring the length of northern pikes (*Esox lucius*). This was sampled from Heming Lake, Manitoba, Canada in 1965, and was reported earlier in [22, 23]. The data comprises five different age-groups of pikes. In [23], the dataset was modelled by setting the second parameter, γ, as a known parameter with $\gamma = 101.415$ (or $\gamma^{-\frac{1}{2}} = 0.0993$). In this modelling, the same setting was applied. The measurement accuracy, ϵ, for the modelling is set equal to 1.0.

Comparison between the modelling criteria is performed by dividing the original dataset into training and test datasets with proportions of 471:52. We first find the model for the training dataset and then fit the dataset to the selected model. The latter is performed by measuring the probability bit-costing, $-\log(P(x))$, of each datum x in the test dataset (see [15] and the references therein). This process is repeated 10 times. The resulting averages of the probability bit-costings for the three criteria, MML, AIC and BIC, are 184.27, 188.34, and 184.50 nits (1 nit = $\log_2 e$ bits), respectively. These results show that MML performs better than both AIC and BIC with a smaller average of probability bit-costings.

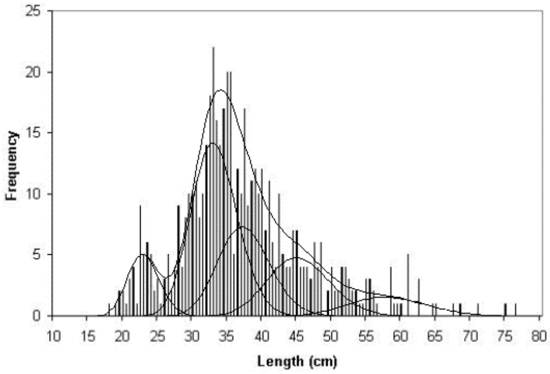

Figure 3. Mixture model of the Heming Pike dataset.

We further analysed the original dataset using MML, AIC and BIC. The modelling results using MML and the histogram of the dataset are presented in Figure 3. Using the proposed method, the Heming Pike dataset was grouped into five components with three components strongly overlapping in the middle. This result is the same as that reported in [23] and is also the same as the true groupings. We further compared the modelling results with those performed using AIC and BIC. BIC modelled the original dataset into five-component mixture model which was the same as our MML results with a slight different structure, whereas AIC grouped the data into three-component mixture model.

5.3 Palm Valley, Australia

The Palm Valley (Australia) image is a cropped and grey-scaled image from a colour image displayed on Visible Earth, NASA web page with VE record ID = 5687. The original was acquired by the Spaceborne Imaging Radar-C/X-band Synthetic Aperture Radar (SIR-C/X-SAR) onboard the space shuttle Endeavour on April 13, 1994. The image consists of 66,482 colour data. We modelled the dataset with the second parameter, γ, which was assumed to be unknown and continuous. The measurement accuracy, ϵ, is set equal to 1.0.

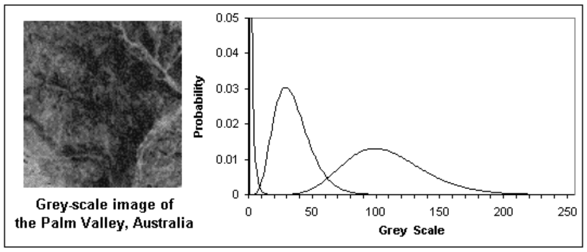

Figure 4. Mixture model of the Palm Valley (Australia) image dataset.

The modelling results as well as the Palm Valley image can be seen in Figure 4, with the dataset being grouped into three components. The first component has a small $\gamma < 1.0$, so that it is skewed strongly to the left. The second and the third components have $\gamma > 1.0$ and are almost normally distributed with a slight skew to the left. These results suggest that the proposed method can also be used to analyse image datasets with left-skewed components.

6 Conclusion

In conclusion, we draw the reader's attention to the following results from Subsection 2.3 and Section 5:

1. MML parameter estimation for one Gamma distribution shows a better performance than Maximum Likelihood (ML) for all values of β, γ and N. Smaller standard errors of the estimations further show that the method robustly performs parameter estimation (see Section 2.3).

2. In most cases, our MML scheme performed better than BIC when modelling the artificial datasets. AIC performed badly for all experiments (see Section 5.1).

3. The proposed MML also performed better than AIC and BIC, as shown in the analysis of the probability bit-costings for the Heming Pike dataset (see Section 5.2). The method can also be applied to the modelling of datasets with left-skewed components such as image dataset shown in Section 5.3.

References

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

[2] L. A. Hunt and M. A. Jorgensen, Mixture model clustering using the Multimix program, *Australian and*

New Zealand Journal of Statistics, 41(2), 1999, 153-171.

[3] G. J. McLachlan and D. Peel, Finite Mixture Models, (New York: John Wiley and Sons, 2000).

[4] C. S. Wallace and P. R. Freeman, Estimation and Inference by Compact Coding, *Journal of the Royal Statistical Society* (B), 49(3), 1987, 240-265.

[5] C. S. Wallace and D. L. Dowe, Minimum Message Length and Kolmogorov Complexity, *Computer Journal*, 42(4), 1999, 270-283. Special issue on Kolmogorov Complexity.

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

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

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

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

[10] C. S. Wallace and D. L. Dowe, Intrinsic classification by MML - the Snob program, *Proc. 7th Australian Joint Conference on Artificial Intelligence*, Armidale, Australia, 1994, 37-44.

[11] C. S. Wallace and D. L. Dowe, MML Mixture Modelling of multi-state, Poisson, von Mises Circular and Gaussian Distributions, *Proc. Sixth International Workshop on Artificial Intelligence and Statistics*, Florida, USA, 1997, 529-536.

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

[13] Y. Agusta and D. L. Dowe, MML Clustering of Continuous-Valued Data Using Gaussian and t Distributions, in B. McKay and J. Slaney (Ed.), *Lecture Notes in Artificial Intelligence*, 2557, (Berlin: Springer-Verlag, 2002) 143-154.

[14] Y. Agusta and D. L. Dowe, Clustering of Gaussian and t Distributions using Minimum Message Length, *Proc. International Conference Knowledge Based Computer Systems (KBCS-2002)*, Mumbai, India, 2002, 289-299.

[15] P. Tan and D. L. Dowe, MML Inference of Decision Graphs with Multi-way Joins, in B. McKay and J. Slaney (Ed.), *Lecture Notes in Artificial Intelligence*, 2557, (Berlin: Springer-Verlag, 2002) 131-142.

[16] C. S. Wallace and D. L. Dowe, MML estimation of the von Mises concentration parameter, Technical Report TR 93/19, Computer Science Department, Monash University, Clayton, 3800 Australia, 1993.

[17] C. S. Wallace, An improved program for classification, *Proc. Ninth Australian Computer Science Conference (ACSC-9)*, 8, Monash University, Australia, 1986, 357-366.

[18] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC-19(6), 1974, 716-723.

[19] S. L. Sclove, Application of model-selection criteria to some problems in multivariate analysis, *Psychometrika*, 52(3), 1987, 333-343.

[20] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics*, 6, 1978, 461-464.

[21] C. Fraley and A. E. Raftery, How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, *Computer Journal*, 41(8), 1998, 578-588.

[22] P. D. M. Macdonald and T. J. Pitcher, Age-groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures, *Journal of the Fisheries Research Board of Canada*, 36, 1979, 987-1001.

[23] P. D. M. Macdonald, Analysis of length-frequency distributions, in R. C. Summerfelt and G. E. Hall (Ed.), *Age and Growth of Fish*, (Iowa: The Iowa State University Press, 1987) 371-384.