Unsupervised Learning of Correlated Multivariate Gaussian Mixture Models Using MML

Yudi Agusta and David L. Dowe
{yagusta, dld}@bruce.csse.monash.edu.au
Computer Science & Software Eng., Monash University, Clayton, 3800 Australia

Abstract. Mixture modelling or unsupervised classification is the problem of identifying and modelling components (or clusters, or classes) in a body of data. We consider here the application of the Minimum Message Length (MML) principle to a mixture modelling problem of multivariate Gaussian distributions. Earlier work in MML mixture modelling includes the multinomial, Gaussian, Poisson, von Mises circular, and Student t distributions and in these applications all variables in a component are assumed to be uncorrelated with each other. In this paper, we propose a more general type of MML mixture modelling which allows the variables within a component to be correlated. Two MML approximations are used. These are the Wallace and Freeman (1987) approximation and Dowe's MMLD approximation (2002). The former is used for calculating the relative abundances (mixing proportions) of each component and the latter is used for estimating the distribution parameters involved in the components of the mixture model. The proposed method is applied to the analysis of two real-world datasets - the well-known (Fisher) Iris and diabetes datasets. The modelling results are then compared with those obtained using two other modelling criteria, AIC and BIC (which is identical to Rissanen's 1978 MDL), in terms of their probability bit-costings, and show that the proposed MML method performs better than both these criteria. Furthermore, the MML method also infers more closely the three underlying Iris species than both AIC and BIC.

Keywords. Unsupervised Classification, Mixture Modelling, Machine Learning, Knowledge Discovery and Data Mining, Minimum Message Length, MML, Classification, Clustering, Intrinsic Classification, Numerical Taxonomy, Information Theory, Statistical Inference.

1 Introduction

Mixture modelling [14, 17, 27] - generally known as unsupervised classification or clustering - models, as well as partitions, a dataset with an unknown number of components (or classes or clusters) into a finite number of components. The problem is also known as intrinsic classification, latent class analysis or numerical taxonomy. Mixture modelling is widely acknowledged as a useful and powerful method to perform pattern recognition - as well as being useful in other areas, such as image and signal analysis.

In this paper, we discuss, in particular, an unsupervised classification that models a statistical distribution by a mixture (a weighted sum) of other distributions. The likelihood function - or objective function (see also Sec. 4, part 1d) - of the mixture modelling problem takes the form of:

$$f(x|M,\pi_1,\cdots,\pi_M,\widetilde{\theta}_1,\cdots,\widetilde{\theta}_M) = \sum_{m=1}^{M} \pi_m \times f_m(x|\widetilde{\theta}_m),$$

where there are M components, π_m is the relative abundance or mixing proportion of the m^{th} component, and $f_m(x|\widetilde{\theta}_m)$ is the probability distribution of m^{th} component (given the component distributional parameters).

There are two processes involved in performing mixture modelling. These are model selection for the model that best describes the dataset and point estimation for the parameters required. The former includes the selection of the most appropriate number of components. The problem we often face in choosing the best model is keeping the balance between model complexity and goodness of fit. In other words, the best model selected for a dataset must be sufficiently complex in order to cover all information in the dataset, but not so complex as to over-fit. Here, we apply the Minimum Message Length (MML) principle simultaneously for both parameter estimation and model selection.

The MML principle [27, 33, 30, 31] was first proposed by Wallace and Boulton [27] in 1968. It provides a fair comparison between models by stating each of them into a two-part message which, in turn, encodes each model (H) and the data in light of that model (D given H). Various related principles have also been stated independently by Solomonoff [24], Kolmogorov [15], Chaitin [4], and subsequently by Rissanen [20]. For a more comprehensive overview, see [30, 31].

Previous applications of MML to the problem of mixture modelling [27, 26, 28, 29, 32, 1, 2] includes the multinomial, Gaussian, Poisson, von Mises circular, and Student t distributions. In these applications, all variables in a component are assumed to be uncorrelated with one another.

For the correlated multivariate problem, various methods have also been proposed including AutoClass by Cheeseman *et. al.* [5], EMMIX (using AIC, BIC, and one other approach) by McLachlan *et. al.* [18], MCLUST (using BIC [22] - which is also the 1978 MDL [20]) by Fraley and Raftery [13] and MULTIMIX by Jorgensen *et. al.* [14]. (Relatedly, see also [7].) Figueiredo and Jain [8] also proposed a mixture modelling method for the same problem using an MML-like criterion. In their method, non-informative Jeffreys priors were utilised for the parameters estimated. A discussion of the appropriateness or otherwise of the Jeffreys prior can be found in [31] and the references therein.

Beginning with an elaboration of the MML principle and its approximations in Section 2, this paper proposes an MML mixture modelling method of correlated multivariate Gaussian distributions, where the variables within a component are assumed to be correlated with one another. This involves elaborations on point estimations for multinomial and multivariate Gaussian distributions (Section 3) and the coding scheme for MML mixture modelling of multivariate Gaussian distributions (Section 4). The proposed method is then applied to the analysis of two real-world datasets, the well-known (Fisher) Iris dataset and a

diabetes dataset. The modelling results are compared with those obtained using two other commonly used criteria, BIC [22] (which is also the 1978 MDL [20]) and AIC (see Section 5), in terms of their probability bit-costings (see [25] and references therein). Comparisons in terms of the resulting number of components and the structure of the resulting components are also provided.

2 MML Principle and Its Approximations

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 [27, 33, 30, 31].

Letting D be the data and H be a model with a prior probability distribution $P(H)$, using Bayes's theorem, the point estimation and model selection problems can be regarded simultaneously as a problem of maximising the posterior probability $P(H) \cdot P(D|H)$. 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

$$\text{MessLen} = -\log_2(P(H)) - \log_2(P(D|H)) \tag{1}$$

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 dealing with the mixture modelling problem of multivariate Gaussian distributions, it is required to perform parameter estimations of the multi-state and the multivariate Gaussian distributions. The parameter estimation of the multi-state distribution can be performed using the MML approximation proposed by Wallace and Freeman (1987) [33]. However, for the correlated multivariate Gaussian distribution, some mathematical challenges arise when using the 1987 MML approximation [33]. As an alternative more tractable approach, Dowe's recent MMLD approximation [16, 11, 10] is applied. These two approximations differ in the way they determine the optimal coding region of the possible models for a given dataset.

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

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

the Fisher information - i.e., the determinant of the matrix of expected second derivatives of the negative log-likelihood. Using (1), and expanding the negative log-likelihood, L, as far as the second term of the Taylor series about $\boldsymbol{\theta}$, the message length for the 1987 MML approximation is then given by [33, 32, 30]:

$$\text{MessLen} = -\log\Big(\frac{h(\boldsymbol{\theta})}{\sqrt{\kappa_D^D F(\boldsymbol{\theta})}}\Big) + L + \frac{D}{2} = -\log\Big(\frac{h(\boldsymbol{\theta}) f(x|\boldsymbol{\theta})}{\sqrt{F(\boldsymbol{\theta})}}\Big) + \frac{D}{2}(1 + \log \kappa_D) \tag{3}$$

where D is the dimension of the dataset and κ_D is a D-dimensional lattice constant [33] with $\kappa_1 = 1/12$ and $\kappa_D \leq 1/12$. The MML estimate of $\boldsymbol{\theta}$ can be obtained by minimising (3).

In Dowe's MMLD approximation, on the other hand, the optimal coding region, R, is determined by specifying the total two-part message length as follows [16, 11, 10]:

$$\text{MessLen} = -\log \int_R h(\theta')d\theta' - \frac{1}{\int_R h(\theta')d\theta'} \int_R h(\theta') \log f(x|\theta')d\theta' \tag{4}$$

Minimising (4) with respect to θ results in the following expression:

$$-\log f(x|\theta')\Big|_{\partial R} = -\frac{1}{\int_R h(\theta')d\theta'} \int_R h(\theta') \log f(x|\theta')d\theta' + 1 \tag{5}$$

where the first term on the right hand side of the equation above represents the second part of the message length. The expression above is known as the MMLD boundary rule - which means that the negative log-likelihood at the boundary, ∂R, of the optimal coding region is equal to the expected negative log-likelihood (with respect to the prior) throughout the region, R, plus one. However, using this approximation, it can be difficult to find the exact optimal coding region, R. Therefore, it is necessary to apply the approximation numerically.

The MMLD message length has been numerically approximated in [11], where the use of the importance sampling distribution was proposed. The importance sampling distribution is useful when we know the most likely area from where the possible models will be derived. With the posterior probability as the importance sampling distribution and applying Monte Carlo integration, the MMLD message length expression (4) can be numerically approximated as follows [11]:

$$\text{MessLen} = -\log \Big(\frac{\sum_{\theta \in Q} f(x|\theta)^{-1}}{\sum_{\theta \in S} f(x|\theta)^{-1}} \Big) - \Big(\frac{\sum_{\theta \in Q} f(x|\theta)^{-1} \log f(x|\theta)}{\sum_{\theta \in Q} f(x|\theta)^{-1}} \Big) \tag{6}$$

where S is the parameter space for the sampling distribution and Q is a subset of the optimal coding region, R.

Considering that the Gaussian is a continuous distribution, a finite coding for the message can be obtained by acknowledging that all recorded continuous data and measurements must only be stated to a finite 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 [32, p74] [2, Sec. 2] [28, p38].

3 Parameter Estimation by Minimum Message Length

As mentioned earlier, for a mixture modelling problem involving multivariate Gaussian distributions, we need to perform parameter estimations of the multi-state distribution for the relative abundances of each component and the multivariate Gaussian distribution for the distribution parameters of each component. Unlike most previous MML mixture modelling works, it is assumed here that the data in each component are required to be normally distributed but that also the variables within each component can be correlated with one another.

3.1 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 $n_1 + n_2 + \cdots + n_M = N$.

The distribution parameters are estimated using the 1987 MML approximation [33]. It follows from (2) that $F(p_1, p_2, \cdots, p_M) = N^{(M-1)}/p_1 p_2 \cdots p_M$. The derivation is also shown elsewhere for $M = 2$ [32, p75].

Assuming a uniform prior of $h(p) = (M-1)!$ over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$, and minimising (3), the MML estimate $\hat{p}_m$ is obtained by [25, sec. 4.2]:

$$\hat{p}_m = (n_m + 1/2)/(N + M/2) \tag{7}$$

Substituting (7) into (3) provides the following two-part message length [27, p187 (4)] [27, p194 (28)] [32, p75 (6)] [1, p291 (5)]:

$$-\log(M-1)! + ((M-1)/2)(\log(N\kappa_{M-1}) + 1) - \sum_{m=1}^{M} (n_m + 1/2) \log \hat{p}_m \tag{8}$$

3.2 Multivariate Gaussian Distribution

The multivariate Gaussian distribution has a likelihood function:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)}$$

where μ is the vector of means, Σ is the covariance matrix of the distribution (allowing correlations), and n is the number of variables in the dataset.

As explained in Section 2, the parameter estimation for this distribution is to be performed using the MMLD approximation. The parameter estimation in this approximation is conducted numerically using the following algorithm:

1. Sample a number of models from the importance sampling distribution.
2. Sort the models according to their likelihood values in decreasing order.
3. Apply the MMLD boundary rule (5) and select models that lie in the region.
4. Find the estimates using the Minimum Expected Kullback-Leibler (minEKL) distance method (weighted by the posterior [6], instead [11, 10] of the prior).

In the first step, the posterior probability is chosen as the importance sampling distribution. For this purpose, two prior probabilities on both parameters, μ and Σ, are required. Here, an improper uniform prior on μ over the $[-\infty, \infty]$ n-dimensional real space ($\Re^n$) and an improper conjugate prior, $|\Sigma|^{(-\frac{n+1}{2})}$, on Σ are considered [21, Sec. 5.2.3]. Both priors are the limiting form of the conjugate normal-inverted Wishart prior [21, Chapter 5]. We notice here that it is impossible to normalise both priors. However, as shown in (6), our numerical message

length calculation only involves the priors via the posterior - which is proper. Therefore, it at least appears that we do not need to explicitly normalise the (possibly improper) priors. With these priors, the posterior probability becomes:

$$\mu|\Sigma, X \sim N(\bar{x}, N^{-1}\Sigma) \tag{9}$$

$$\Sigma|X \sim W^{-1}(N-1, (NS)^{-1}) \tag{10}$$

where N is the number of data and S is the data covariance matrix. Utilising the Gibbs sampling method, the possible models are sampled from (9) and (10).

Once the models are sampled, they are sorted according to their likelihood values in decreasing order, starting with the model at the Maximum Likelihood solution. We then apply the MMLD boundary rule (5) and select the models that lie inside the optimal coding region. Referring to equation (6), the following algorithm is utilised in simultaneously selecting models lying in the coding region and calculating the first and second parts of the resulting message length. This algorithm is a variation of that proposed in [11].

```
BEGIN ALGORITHM
 //Setting the first model (ML model) into the selected models
 Allocate first model into selected models;
 //Setting each expression involves in the message length
 Set FIRSTPARTNUMERATOR = 1.0;
 Set SECONDPARTNUMERATOR = minusLogLikelihood of first model;
 Set SECONDPARTDENOMINATOR = 1.0;
 //Calculating the second part of the message length
 Set SECONDPART = SECONDPARTNUMERATOR/SECONDPARTDENOMINATOR;
 Move to next model;
 While(not reaching the end of the sorted models) {
  //Applying the MMLD boundary rule for the rest of the models
  If(minusLogLikelihood of current model<=SECONDPART+1.0) {
   //Setting the model into the selected models, if it is inside
   Allocate current model into selected models;
   Set likelihood = exp(minusLogLikelihood of current model -
    minusLogLikelihood of first model);
   //Updating each expression involves in the message length
   FIRSTPARTNUMERATOR += likelihood;
   SECONDPARTNUMERATOR += minusLogLikelihood of current model *
    likelihood;
   SECONDPARTDENOMINATOR += likelihood;
   //Calculating the second part of the message length
   SECONDPART = SECONDPARTNUMERATOR/SECONDPARTDENOMINATOR;
  }
  Else -> Exit loop;
  Move to next model;
 }
```

```
 //Calculating the denominator of the first part until
 //the last model of the sorted models
 Set FIRSTPARTDENOMINATOR = FIRSTPARTNUMERATOR;
 While(not reaching the end of the sorted models) {
  Set likelihood = exp(minusLogLikelihood of current model-
   minusLogLikelihood of first model);
  FIRSTPARTDENOMINATOR += likelihood;
  Move to next model;
 }
 //Calculating the first part of the message length
 FIRSTPART = -log(FIRSTPARTNUMERATOR/FIRSTPARTDENOMINATOR);
END ALGORITHM
```

From the selected models, estimates are derived using the Minimum Expected Kullback-Leibler (minEKL) distance point estimation method, which is numerically calculated by taking the maximum likelihood of the (posterior-weighted) future samples, randomly sampled from the selected models [6]. This point estimation method is statistically invariant under 1-1 re-parameterisation.

4 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 (in principle). Recall that from Section 1, the encoding of the mixture model hypothesis comprises several concatenated message fragments [27, 26, 28, 29, 32], 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 3.1.

1c For each component, the distribution parameters of the component attribute. In this case, each component is inferred as a multivariate correlated Gaussian distribution as in subsection 3.2.

1d For each thing, the component to which the thing is estimated to belong. (Part 1d is typically included and discussed in MML mixture modelling but is often omitted in other mixture modelling literature.)

For part (1d), instead of the total assignment as originally proposed in [27], partial assignment is used to approximate the improved cost. Further discussion of this can be found in [26, Sec. 3] [28, Sec. 3.2] [29, Sec. 3.2] [32, pp. 77-78][2, 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.

5 Alternative Model Selection Criteria - AIC and BIC

In order to justify the proposed MML method, two criteria are considered for comparison. These are the *Akaike Information Criterion* (AIC) and Schwarz's *Bayesian Information Criterion* (BIC).

AIC, first developed by Akaike [3], is given by:

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

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 to be estimated in the model. For the multivariate Gaussian mixture, N_p is set equal to $k-1+k[n+n(n+1)/2]$ as explained by Sclove [23] ($k-1$ mixing proportions, and $n+n(n+1)/2$ parameters for the n means and the matrix per component), where k is the number of components and n is the number of variables in the dataset. The model which results in the smallest AIC is the model selected.

The second criterion, BIC, first introduced by Schwarz [22] is given by:

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

where L is the logarithm of the likelihood at the maximum likelihood solution for the investigated mixture model, N_p is the number of independent parameters to be estimated, and N is the number of data. For the multivariate Gaussian mixture, N_p is again equal to $k-1+k[n+n(n+1)/2]$ ($k-1$ mixing proportions, and $n+n(n+1)/2$ parameters per component), where k is the number of components and n is the number of variables in the dataset. (The number of components is not considered an independent parameter for the purposes of calculating the BIC as explained by Fraley and Raftery [12].) (The BIC model selection criterion is formally, not conceptually, the same as the 1978 Minimum Description Length (MDL) criterion proposed by Rissanen [20].) The model which results in the smallest BIC is selected as the best model.

6 Experiments

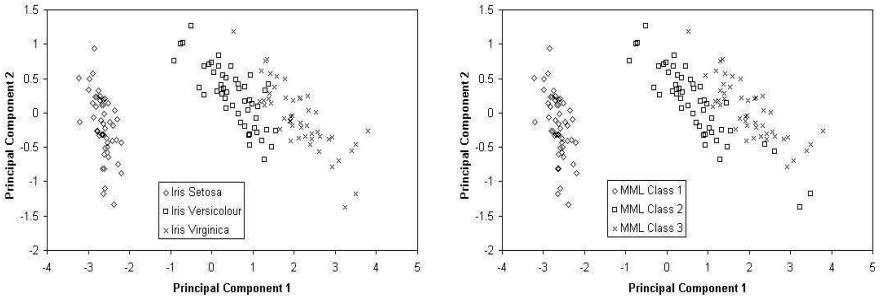
6.1 Iris Dataset

The Iris dataset was first analysed in 1936 by Fisher [9]. It comprises 150 iris plants belonging to three species, namely Iris Setosa (S), Iris Versicolour (Ve), and Iris Virginica (Vi). Four variables measuring sepal and petal length and width of the species are involved. Each group is represented by 50 plants. The measurement accuracies, ϵ, in this dataset (see Sec. 2) were set to 1.0 for all variables.

The analysis here is performed by dividing the original dataset into training and test datasets with proportions of 135:15. We first find the model for the training dataset and then fit the test 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 [25] and the references therein). This process was repeated 20 times. The resulting averages ($\pm$ the standard deviations) of the probability bit-costings for the three criteria, MML, AIC and BIC, are 21.37 ($\pm$ 5.8), 23.14 ($\pm$ 10.1), and 21.83 ($\pm$ 6.2) nits (1 nit = $\log_2 e$ bits), respectively. These results suggest that MML performs better than both AIC and BIC.

In a further analysis using the proposed MML method, MML grouped the entire dataset into three components. Fig. 1 shows the original (Fisher) Iris dataset and the resulting three-component MML mixture of the dataset, plotted with the first two principal components as axes. Here, the relative abundances of the resulting MML components were 0.333:0.339:0.328, which were almost the same as those of the true model (0.333:0.333:0.333), and the MML fit appeared pleasing. The entire dataset was also analysed using AIC and BIC. The modelling using AIC resulted in a four-component mixture, whereas the modelling using BIC resulted in a two-component mixture in which the highly overlapping Versicolour and Virginica iris groups were modelled into one component. The second best model for BIC was a three-component mixture. However, the relative abundances in this model were 0.333:0.436:0.231, which were substantially different from those of the true model.



(a) Original Iris Data (b) MML Model

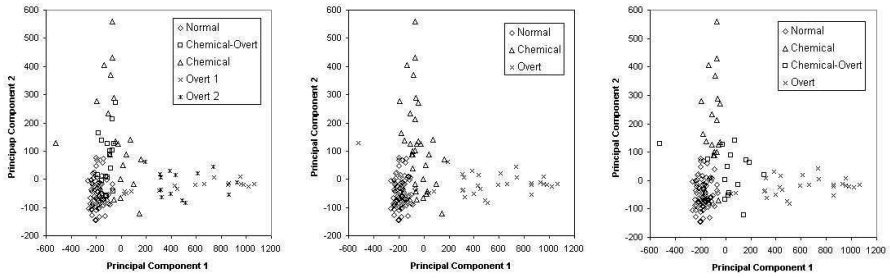
Fig. 1. The original (Fisher) Iris dataset and the resulting three-component MML mixture, plotted with the first two principal components as axes.

6.2 Diabetes Dataset

This diabetes dataset was first reported in 1979 by Reaven and Miller [19] and comprises 145 samples with three variables measuring glucose area, insulin area and the steady state plasma glucose response (SSPG). The modelling reported in [19] was performed based on the groupings established using conventional clinical criteria. In this conventional classification, diabetes was grouped into three categories: Normal, Chemical and Overt. A subsequent analysis which also resulted in a three-component mixture has been reported by Fraley and Raftery [12]. In the latter analysis [12], BIC was used to select the number of components. In the present application, we aimed to compare these earlier results [19, 12] with the analysis obtained using the proposed MML method.

The measurement accuracies, ϵ, in this modelling were set equal to 1.0 for all three variables.

We applied the same analysis as in Sec. 6.1, where the proportions of the training and test datasets are set equal to 130:15. The experiment was repeated 20 times. The averages ($\pm$ the standard deviations) of the probability bit-costings on this diabetes dataset for MML, AIC and BIC were 235.94 ($\pm$ 8.9), 237.49 ($\pm$ 12.4), and 236.54 ($\pm$ 10.4) nits (1 nit $= \log_2 e$ bits), respectively. Again, MML performed better than AIC and BIC.



(a) AIC Model (b) BIC Model (c) MML Model

Fig. 2. Modelling using AIC, BIC and MML (plotted on 2 principal component axes).

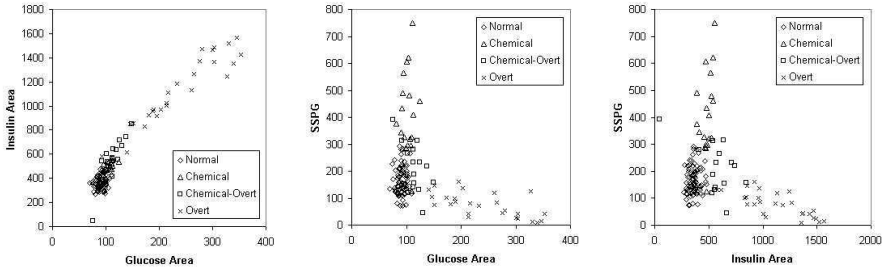


Fig. 3. Modelling using the proposed MML method (the results are pair-plotted).

We further analysed the original diabetes dataset using AIC, BIC and MML, with the results (plotted with the first two principal components as axes) being shown in Fig. 2. The modelling using AIC (see Fig. 2(a)) resulted in five components with two components dividing the Overt group, and one component overlapping with the Chemical and Overt groups. The modelling using BIC resulted in three components, which are the same as those reported in [12] and shown in Fig. 2(b).

In the modelling using the proposed MML method, a four-component mixture resulted: this is plotted against the first two principal components in Fig. 2(c) and its ${}^3C_2 = 3$ cross-sectional pair plots are shown in Fig. 3. The additional component to the original classification appears to highly overlap with the Chemical and Overt groups, and consists of members that originally belonged to both groups. Although the results are different from the original classification,

this does not imply that the proposed method has modelled the dataset incorrectly. As mentioned earlier, the original groupings used to justify the analysis in [19] (and possibly also in [12]) were performed based on conventional clinical criteria. Thus, no true model exists which can be used to justify which classification is correct. Conversely, the results obtained here and the performance of the proposed MML method compared to both AIC and BIC in terms of the probability bit-costings might suggest an alternative diabetes classification by the addition of a Chemical-Overt group.

7 Conclusion

In conclusion, we draw the attention of the reader to the following results:

1. The proposed method broadens the scope of problems handled by MML mixture modelling, by now modelling correlated multivariate data. This provides flexibility since most real-world datasets contain variables that are correlated within each component in their mixture models.
2. The proposed MML method performs better than two other modelling criteria, AIC and BIC (or 1978 MDL), as shown in the analysis of the probability bit-costings for both the (Fisher) Iris and diabetes datasets.

References

1. Y. Agusta and D. L. Dowe. Clustering of Gaussian and t Distributions using Minimum Message Length. In *Proc. Int'l. Conf. Knowledge Based Computer Systems - KBCS-2002*, pp. 289-299, Mumbai, India, 2002. Vikas Publishing House Pvt. Ltd.
2. Y. Agusta and D. L. Dowe. MML Clustering of Continuous-Valued Data Using Gaussian and t Distributions. In *Lecture Notes in Artificial Intelligence*, vol. 2557, pp. 143-154, 2002. Springer-Verlag, Berlin.
3. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.
4. G. J. Chaitin. On the length of programs for computing finite sequences. *J. the Association for Computing Machinery*, 13:547–569, 1966.
5. P. Cheeseman and J. Stutz. Bayesian Classification (AutoClass): Theory and Results. In *Advances in Knowledge Discovery and Data Mining*, pp. 153-180, 1996. AAAI Press/MIT Press.
6. D. L. Dowe, R. A. Baxter, J. J. Oliver, and C. S. Wallace. Point Estimation using the Kullback-Leibler Loss Function and MML. In *Lecture Notes in Artificial Intelligence*, vol. 1394, pp. 87-95, 1998. Springer-Verlag, Berlin.
7. R. T. Edwards and D. L. Dowe. Single factor analysis in MML mixture modelling. In *Lecture Notes in Artificial Intelligence*, vol. 1394, pp. 96-109, 1998. Springer-Verlag, Berlin.
8. M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.
9. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

10. L. J. Fitzgibbon, D. L. Dowe, and L. Allison. Change-Point Estimation Using New Minimum Message Length Approximations. In *Lecture Notes in Artificial Intelligence*, vol. 2417, pp. 244-254, 2002. Springer-Verlag, Berlin.
11. L. J. Fitzgibbon, D. L. Dowe, and L. Allison. Univariate Polynomial Inference by Monte Carlo Message Length Approximation. In *Proc. 19th International Conf. of Machine Learning (ICML-2002)*, pp. 147-154, Sydney, 2002. Morgan Kaufmann.
12. C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Computer J.*, 41(8):578–588, 1998.
13. C. Fraley and A. E. Raftery. MCLUST: Software for Model-Based Cluster and Discriminant Analysis. Technical Report 342, Statistics Dept., Washington Uni., Seattle, USA, 1998.
14. L. A. Hunt and M. A. Jorgensen. Mixture model clustering using the Multimix program. *Australian and New Zealand Journal of Statistics*, 41(2):153–171, 1999.
15. A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:4–7, 1965.
16. E. Lam. Improved approximations in MML. Honours Thesis, School of Computer Science and Software Engineering, Monash Uni., Clayton 3800 Australia, 2000.
17. G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley, NY, 2000.
18. G. J. McLachlan, D. Peel, K. E. Basford, and P. Adams. The EMMIX software for the fitting of mixtures of Normal and t-components. *J. Stat. Software*, 4, 1999.
19. G. M. Reaven and R. G. Miller. An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis. *Diabetologia*, 16:17–24, 1979.
20. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
21. J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.
22. G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
23. S. L. Sclove. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343, 1987.
24. R. J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22, 224–254, 1964.
25. P. J. Tan and D. L. Dowe. MML Inference of Decision Graphs with Multi-way Joins. In *Lecture Notes in Artificial Intelligence*, vol. 2557, pp. 131-142, 2002. Springer-Verlag, Berlin.
26. C. S. Wallace. An improved program for classification. In *Proc. 9th Aust. Computer Science Conference (ACSC-9)*, vol. 8, pp. 357-366, Monash Uni., Australia, 1986.
27. C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer J.*, 11(2):185–194, 1968.
28. C. S. Wallace and D. L. Dowe. Intrinsic classification by MML - the Snob program. In *Proc. 7th Aust. Joint Conf. on AI*, pp. 37-44, 1994. World Scientific, Singapore.
29. C. S. Wallace and D. L. Dowe. MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions. In *Proc. 6th International Workshop on Artificial Intelligence and Statistics*, pp. 529-536, Florida, 1997.
30. C. S. Wallace and D. L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Comp. J.*, 42(4):270–283, 1999. Special issue on Kolmogorov Complexity.
31. C. S. Wallace and D. L. Dowe. Refinements of MDL and MML Coding. *Computer J.*, 42(4):330–337, 1999. Special issue on Kolmogorov Complexity.
32. C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, Jan. 2000.
33. C. S. Wallace and P. R. Freeman. Estimation and Inference by Compact Coding. *J. Royal Statistical Society (B)*, 49(3):240–265, 1987.