MML Clustering of Continuous-valued Data using Gaussian and t Distributions

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

Abstract. Clustering, also known as mixture modelling or intrinsic 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 clustering problem of Gaussian and t distributions. Earlier work in the MML clustering was conducted in regards to the multinomial and Gaussian distributions (Wallace and Boulton, 1968) and in addition, the von Mises circular and Poisson distributions (Wallace and Dowe, 1994, 2000). Our current work extends this by applying the Gaussian distribution to the more general t distribution. Point estimation of the t distribution is performed using the MML approximation proposed by Wallace and Freeman (1987). A comparison of the MML estimations of the t distribution to those of the Maximum Likelihood (ML) method in terms of their Kullback-Leibler (KL) distances is also provided. Within each component, our application also performs a model selection on whether a particular group of data is best modelled as a Gaussian or a t distribution. The proposed modelling method is then applied to several artificially generated datasets. The modelling results are compared to the results obtained when using the MML clustering of Gaussian distributions. Our modelling method compares quite well to an alternative clustering program (EMMIX) which uses various modelling criteria such as the Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC).

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

1 Introduction

The Minimum Message Length (MML) principle [18][23][21] is an invariant Bayesian point estimation and model selection technique based on information theory. The basic idea of MML is to find an hypothesis (or theory) of a distribution or a model that minimises the total length of a two-part message encoding the hypothesis, and the data in light of that hypothesis.

Letting D be the data and H be the hypothesis, with a prior probability distribution $P(H)$, based on Bayes's theorem, the point estimation and model selection problem can be regarded 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 states the message length of the hypothesis and the second term states the message length of the data in light of the hypothesis.

This principle was first stated and then applied in a series of papers by Wallace and Boulton dealing with model selections and parameter estimations of multi-state and Gaussian distributions for a clustering problem [18]. A related principle has also been stated independently by Solomonoff [15]. An important special case of the MML principle observed by Chaitin [4] is that data can be regarded as random if there is no hypothesis, H, that can encode the data in a shorter message length than the null hypothesis.

Beginning with parameter estimation, this paper proposes an MML clustering which extends the clustering problem of Gaussian distributions [18][17][22] by considering the t distribution as the distribution of the continuous data investigated. Since the Gaussian distribution is a special case of the t distribution, the application also performs a more general model selection on whether the data in a particular group fits a Gaussian or a t distribution.

2 Parameter Estimation by MML

In order to apply MML to the clustering problem of Gaussian and t distributions, we need parameter estimations of the multi-state, Gaussian and t distributions.

Given the 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 that is the determinant of the matrix of expected second derivatives of the negative log-likelihood. Based on equation (1), and by expanding the negative log-likelihood, L, as far as the second term of the Taylor series about the parameter $\boldsymbol{\theta}$, the message length is then calculated by [23]:

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

where D is the number of parameters to be estimated and κ_D is a D-dimensional lattice constant [5] , with $\kappa_D \leq 1/12$. The MML estimate of $\boldsymbol{\theta}$ can be obtained by minimising equation (3).

Considering that both distributions used here are continuous, 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, which is, in practice, made 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 [22].

2.1 Multi-state Variables

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 equation (2), it follows that the Fisher information is given by:

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

The derivation is also shown elsewhere for $M = 2$ [22].

Assuming a uniform prior $(M-1)!$ over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$, and minimising equation (3), the MML estimate $\hat{p}_m$ is obtained by [23][22, p75][18, p187 (4), p194 (28), p186 (2)]:

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

Substituting equation (4) into the message length expression (3) provides the following total two-part message length:

$$-\log(M-1)! + ((M-1)/2)\big(\log(N\kappa_{M-1}) + 1\big) - \sum_{m=1}^{M} \big(n_m + 1/2\big) \log \hat{p}_m \tag{5}$$

2.2 Gaussian Variables

For the Gaussian distribution, with a likelihood function

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

the Fisher information is given by:

$$F(\mu, \sigma) = 2N^2/\sigma^4 \quad \text{or by [22]:} \quad F(\mu, \sigma^2) = N^2/2(\sigma^2)^3$$

Assuming a uniform prior of $1/R$ on μ over a finite range of width, R, where $R = \max\{10$, the difference between the maximum and the minimum value of the data$\}$ and a $1/\sigma$ prior on σ (which corresponds to a uniform prior on $\log\sigma$ and equivalently to a $1/\sigma^2$ prior on σ^2) over the range $[e^{-4}, e^4]$, letting $s^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2$, and minimising equation (3), the MML estimates $\hat{\mu}_{\mathrm{MML}}$ and $\hat{\sigma}_{\mathrm{MML}}$ are then given by [22, p75]:

$$\hat{\mu}_{\mathrm{MML}} = \bar{x} = (\sum_{i=1}^{N} x_i)/N, \quad \hat{\sigma}^2_{\mathrm{MML}} = s^2/(N-1) \tag{6}$$

2.3 *t* Variables

The t distribution with mean, μ, standard deviation, σ, and degree of freedom, ν, is a continuous distribution which generalises some other distributions, such as the Gaussian ($\nu = \infty$) and Cauchy ($\nu = 1$) distributions. For large $\nu(> 100)$, the t distribution is closely approximated by a Gaussian distribution. The smaller

the value of ν, the longer the tail in the t distribution. Using this property, the t distribution is often used to model data with atypical observations, such as outliers. The distribution has a likelihood function:

$$f(x|\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})}\frac{1}{\sigma}\Big[1+\frac{(x-\mu)^2}{\nu\sigma^2}\Big]^{-\frac{(\nu+1)}{2}} \tag{7}$$

where $\Gamma(x)$ is the Gamma function, given by (for $x > 0$):

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \quad \text{with} \quad \psi^{(1)}(x) = \frac{\mathrm{d}^2\Gamma(x)}{\mathrm{d}x^2}$$

For any positive integer x, $\Gamma(x) = (x-1)!$. For large x, the value of the Gamma function can also be approximated using the following Stirling's approximation [9]:

$$\Gamma(x) \approx e^{-x}x^{x-\frac{1}{2}}\sqrt{2\pi}(1+\frac{1}{12x}+\frac{1}{288x^2}+O|x|^{-3}) \tag{8}$$

Using equation (2), the Fisher information, F, is given by:

$$F(\mu,\sigma,\nu) = \frac{N^3}{\sigma^4}\Big\{\frac{\nu(\nu+1)}{2(\nu+3)^2}\Big\{\psi^{(1)}\Big(\frac{\nu}{2}\Big)-\psi^{(1)}\Big(\frac{\nu+1}{2}\Big)\Big\}-\frac{1}{(\nu+1)(\nu+3)}\Big\} \tag{9}$$

Assuming the same priors on μ and σ as those used for Gaussian distribution, and $2/\pi(1+\nu^2)$ prior on ν with ν being an unknown continuous parameter in $(0,\ \infty]$, the MML estimation of the t distribution parameters are obtained by minimising equation (3) with respect to each parameter. Since there are no sufficient statistics to estimate the parameters and each parameter is dependent on each other, the inference is performed using a binary search by setting $\partial\mathrm{MessLen}/\partial\boldsymbol{\theta} = 0$ and iterating the search process until a certain precision of estimation is obtained.

3 One-component Univariate Model

In this section, we consider the inference of only one univariate component. We return in later sections to consider mixtures of several multivariate components.

The difference between the Maximum Likelihood (ML) and MML principles in estimating multi-state variables and Gaussian variables has been elaborated upon by Wallace and Dowe [22], and is quite pronounced for the von Mises [19] and some other distributions [21, p282]. Here, we compare the MML estimation of t variables to that of the ML method in terms of the resulting Kullback-Leibler (KL) distances. The ML estimation of t variables has been proposed in a paper by Liu and Rubin [10] in which the estimation is performed using the EM algorithm and its extensions, the ECM and ECME algorithms.

The KL 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 datasets for the experiment are repeatedly generated from artificial models with $N = \{10, 100\}$, $\mu = 0.0$, $\sigma = \{1.0, 5.0\}$ and $\nu = \{1.0, 3.0, 10.0, 25.0, 50.0, 100.0\}$. Both estimators infer the datasets with ν either taking the highest value of ν, which is, in this experiment, set to 100.0 or a certain value (< 100.0). The calculation results are shown in Table 1.

σ=1.0 & ν=		1.0	3.0	10.0	25.0	50.0	100.0
N=10	ML	0.655±0.89	0.419±0.70	0.170±0.21	0.169±0.25	0.172±0.19	0.173±0.26
	MML	0.292±0.48	0.205±0.45	0.118±0.17	0.140±0.20	0.131±0.12	0.134±0.17
N=100	ML	0.078±0.04	0.020±0.03	0.018±0.02	0.011±0.01	0.011±0.01	0.012±0.02
	MML	0.076±0.04	0.018±0.02	0.017±0.02	0.011±0.01	0.011±0.01	0.012±0.01

σ=5.0 & ν=		1.0	3.0	10.0	25.0	50.0	100.0
N=10	ML	0.684±1.00	0.427±0.61	0.190±0.22	0.183±0.23	0.199±0.28	0.191±0.22
	MML	0.283±0.48	0.222±0.35	0.141±0.14	0.133±0.14	0.157±0.17	0.145±0.15
N=100	ML	0.064±0.03	0.020±0.03	0.014±0.01	0.012±0.01	0.011±0.01	0.011±0.01
	MML	0.063±0.03	0.019±0.03	0.013±0.01	0.012±0.01	0.011±0.01	0.011±0.01

Table 1. Average of KL distances of the ML and MML estimations of 100 datasets with μ=0.0, σ=1.0 or 5.0 and N=10 or 100 (with ± standard errors).

As shown in Table 1, the MML estimators resulted in estimates which are closer than ML to the true model, with smaller KL distances for all cases except when $N = 100$ and ν is large (≥ 25.0). It is possible that a different choice of priors in subsection 2.3 will lead to MML outperforming ML in all cases. It is also possible that Strict MML [23][21], Dowe's MMLD or another refinement of equation (3) [23] will do likewise. The MML method performed a robust estimation with smaller standard errors of the resulting KL distances for all but one estimation (namely, $\sigma = 5.0, \nu = 50.0, N = 100$).

4 Clustering

Clustering, which is also known as mixture modelling [6][16][11][8][12], intrinsic classification [3][20], and numerical taxonomy, models a statistical distribution by a mixture (a weighted sum) of other distributions, as well as partitioning an unknown number of components (or classes or clusters) of a dataset into a finite number of components.

Such a cluster analysis will result in a description of the number of components, the relative abundances (or mixing proportions) of each component, their distribution parameters and the members that belong to them. In the latter, an issue arises as to whether each datum is assigned totally to the component

or not. This issue affects the application of the MML principle to the clustering problem, and is explained further in the next section (see part 1d of the message).

In the case of the clustering problem of the t distributions, McLachlan, Peel, Basford and Adams [13] have introduced the EMMIX software for the fitting of a mixture of the Gaussian and t components. This allows datasets to be fitted as Gaussian distributions as well as t distributions. A comparison of the proposed clustering method to the application of the t distribution in the EMMIX software is provided in subsection 6.4.

5 MML Clustering

The application of MML to the problem of clustering was first introduced by Wallace and Boulton [18]. This application involved discrete multinomial and continuous Gaussian distributions. Wallace and Dowe [20][22] extended this work by adding two other distributions - Poisson and von Mises circular. An alternative MML-based approach to mixture modelling was also given very recently by Figueiredo and Jain [7].

In order to apply the MML principle to a clustering problem, a two-part message conveying the mixture model needs to be constructed. The first part of the message encodes the hypothesis, H, and the second part encodes the data in light of the hypothesis. The hypothesis 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, an attribute of a component is inferred both as a Gaussian and a t distribution as in subsections 2.2 and 2.3, respectively. The model which results in a shorter message length is chosen.

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

The method of assignment of things to components has changed since the MML clustering was first introduced by Wallace and Boulton [18]. The original coding scheme [18] utilised a total assignment of things to components. That scheme was inefficient because of the possible savings that can be made when two components overlap substantially [17]. The original - total assignment - scheme can also lead to inconsistent estimates, where the difference between the means of components is over-estimated and the standard deviation of components are under-estimated, as shown in Fig. 1.

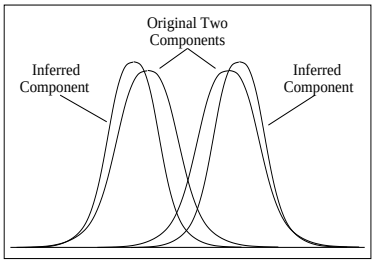

Fig. 1. Inferring two substantially overlapping components using a total assignment.

Instead of assigning things totally to a component, a partial assignment was proposed [17][22]. In a partial assignment, data things are partially assigned to each component with a certain probability. Below, we compare the total assignment with the partial assignment in terms of their message length.

In a total assignment, let $p(j,x), j = 1, \ldots, M$, be the probability of component j generating datum x. The message length to encode x which is assigned to its best component is equal to $-\log(\max_j\ p(j,x))$. On the other hand, for a partial assignment, let $P(x) = \sum_{j=1}^{M} p(j,x)$, be the total probability of any component generating datum x. The datum x will then be assigned to a component j with probability $p(j,x)/P(x)$. The message length of this assignment is equal to $-\log(P(x))$, which is shorter than that of a total assignment by $\log_2(P(x)/\max_j\ p(j,x))$ on each datum x.

In the case where a datum x has an equal probability of being assigned to more than one component, e.g. $p(1,x) = p(2,x) = P(x)/2$, a saving of 1 bit of information can be gained by assigning x to either component 1 or component 2 at random.

Once the first part of the message is stated, the second part of the message will encode the data in light of the hypothesis stated in the first part of the message. Since the objective of the MML principle is to find the hypothesis 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 hypothesis that gives the shortest/minimum message length.

6 Experimental Evaluations

In testing the MML clustering of Gaussian and t distributions, three examples of bivariate mixture datasets were generated artificially. The data points were generated from Gaussian as well as t distributions. The procedure in generating these artificial mixture datasets is similar to that used by Baxter and Oliver [2].

Below, we compare the modelling results of our method to those obtained using the MML clustering of Gaussian distributions only. The comparison of the latter clustering method to other criteria such as AIC, BIC, PC and ICOMP can be found in [2].

At the end of this section, we also compare the modelling results of 20 datasets generated using the same parameters as the datasets mentioned above. The comparison is performed in terms of the resulting number of components and includes two other criteria such as the Akaike Information Criterion (AIC) [1] and Schwarz's Bayesian Information Criterion (BIC) [14].

6.1 One-Component Bivariate Mixture

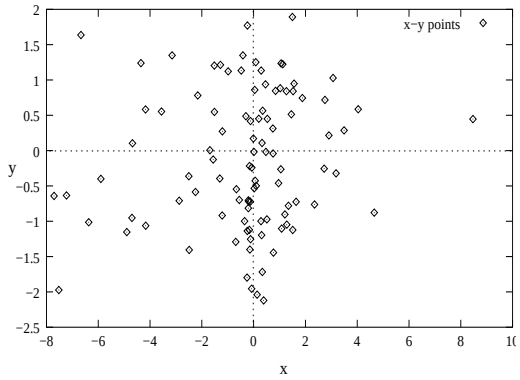

Fig. 2. One-component mixture of 100 bivariate data points generated from a combination of t and Gaussian distributions: $t_{\nu_x=1}(\mu_x = 0, \sigma_x^2 = 1) \times N(\mu_y = 0, \sigma_y^2 = 1)$.

	MML Gaussian or t	MML All Gaussian
MessLen	754.310 nits	771.976 nits
Attribute 1	(assume measurement accuracy $\epsilon = 1.0$)	
Mean(μ)	0.101	-0.372
SD(σ)	1.199	2.609
DegOfF(ν)	**1.612**	∞
Attribute 2	(assume measurement accuracy $\epsilon = 1.0$)	
Mean(μ)	-0.134	-0.134
SD(σ)	0.968	0.968
DegOfF(ν)	∞	∞

Table 2. Comparison of two different MML modelling methods: (i) using Gaussian and t distributions, and (ii) Gaussian distributions only. (Recall that a Gaussian distribution is a special case of the t distribution with $\nu = \infty$.) 1 nit = $\log_2 e$ bits.

Here, we extend the inference results from Section 3. The dataset used in this example (see Fig. 2) consists of 100 bivariate data points which are generated from a t distribution with three parameters: $\mu_x = 0.0$, $\sigma_x = 1.0$ and $\nu_x = 1.0$, and a Gaussian distribution with two parameters: $\mu_y = 0.0$ and $\sigma_y = 1.0$.

The modelling result (see Table 2) shows that the message length in modelling the dataset using the MML clustering of Gaussian and t distributions was shorter by roughly 18 nits than that when using Gaussian distributions only. This is an effect of the inference of the first attribute of the dataset, whereby using a combination of Gaussian and t distributions, the attribute was inferred as a t

distribution with degrees of freedom, $\nu = 1.612$. The result also shows that data from a t distribution can be inferred as coming from a Gaussian distribution with a larger standard deviation. However, since the latter inference resulted in a longer message length, the method automatically chose the t distribution as the inferred distribution.

6.2 Two-Component Bivariate Mixture

The dataset in this example, as shown in Fig. 3, is generated from a bivariate mixture with two components. The first component is a combination of a t distribution with three parameters: $\mu_{x1} = 0.0$, $\sigma_{x1} = 1.0$ and $\nu_{x1} = 1.0$ and a Gaussian distribution with two parameters: $\mu_{y1} = 0.0$ and $\sigma_{y1} = 1.0$. The data points in the second component are generated from two Gaussian distributions with two parameters, $\mu_{x2} = 2.0$ and $\sigma_{x2} = 1.0$ for the first attribute and $\mu_{y2} = 3.5$ and $\sigma_{y2} = 1.0$ for the second attribute. Both components have 50 data points.

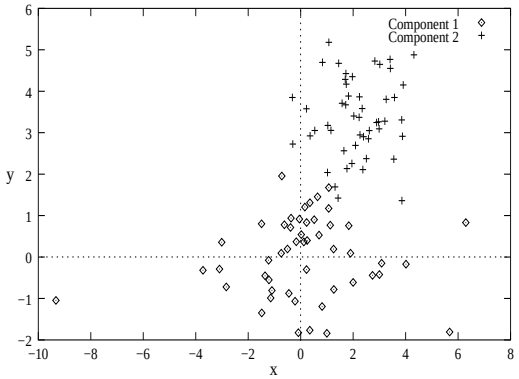

Fig. 3. Two-component mixture of 100 bivariate data points generated from a combination of t and Gaussian distributions: $0.5(t_{\nu_{x1}=1}(\mu_{x1} = 0, \sigma^2_{x1} = 1) \times N(\mu_{y1} = 0, \sigma^2_{y1} = 1)) + 0.5(N(\mu_{x2} = 2, \sigma^2_{x2} = 1) \times N(\mu_{y2} = 3.5, \sigma^2_{y2} = 1))$

Modelling the artificially generated dataset from Fig. 3, the first attribute of the first component was fitted as a t distribution with degrees of freedom, $\nu = 2.008$, instead of as a Gaussian distribution. Inferring the attribute as a t distribution resulted in a shorter message length by roughly 11 nits compared to when the attribute was inferred as Gaussian. In this result, the mixing proportions of the components for both clustering methods were almost the same with 0.527:0.473 for our MML Gaussian and t method and 0.524:0476 for the MML modelling using Gaussian distributions only.

6.3 Three-Component Bivariate Mixture

The example here (see Fig. 4) is generated from a bivariate mixture with three components. The first component is generated from two Gaussian distributions with two parameters each, $\mu_{x1} = -2.0$ and $\sigma_{x1} = 1.0$ and $\mu_{y1} = -3.5$ and $\sigma_{y1} = 1.0$, respectively. The second component is a combination of a t distribution with three parameters: $\mu_{x2} = 0.0, \sigma_{x2} = 1.0$ and $\nu_{x2} = 1.0$ and a Gaussian distribution

with two parameters: $\mu_{y2} = 0.0$ and $\sigma_{y2} = 1.0$. The third component is from two Gaussian distributions with two parameters each: $\mu_{x3} = 2.0$ and $\sigma_{x3} = 1.0$ and $\mu_{y3} = 3.5$ and $\sigma_{y3} = 1.0$, respectively. Each component has 50 data points.

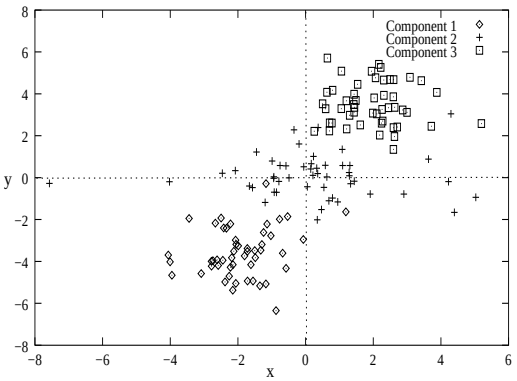

Fig. 4. Three-component mixture of 150 bivariate data points generated from a combination of t and Gaussians: $1/3(N(\mu_{x1} = -2, \sigma^2_{x1} = 1) \times N(\mu_{y1} = -3.5, \sigma^2_{y1} = 1)) +$
$1/3(t_{\nu_{x2}=1}(\mu_{x2} = 0, \sigma^2_{x2} = 1) \times N(\mu_{y2} = 0, \sigma^2_{y2} = 1)) +$
$1/3(N(\mu_{x3} = 2, \sigma^2_{x3} = 1) \times N(\mu_{y3} = 3.5, \sigma^2_{y3} = 1))$

Modelling the dataset illustrated in Fig. 4, the first attribute of the second component, which was generated from a t distribution, was inferred as a t distribution with degrees of freedom, $\nu = 2.225$. This inference resulted in a shorter message length by about 3 nits compared to that when the attribute was inferred as a Gaussian distribution. In this result, the mixing proportions of the components when using Gaussian and t distributions were mostly the same as those when using Gaussian distributions only.

6.4 Alternative Clustering of t Distributions: EMMIX

In 1999, McLachlan, Peel, Basford and Adams [13] introduced the EMMIX software, which allows a dataset to be modelled as either a mixture of only (correlated) Gaussian distributions or a mixture of only (correlated) t distributions. The software, which is mainly used to model datasets as a mixture of Gaussian distributions, is extended by providing an option to change the distribution of the components from Gaussian to t distributions. The parameter estimations are performed using the ML method by utilising the EM algorithm and its extensions, the ECM and ECME algorithms. The value of ν can be fixed in advance or estimated from the data for each component using the ECM algorithm.

Our MML clustering of uncorrelated Gaussian and uncorrelated t distributions allows all attributes in all components to be t or Gaussian. On the other hand, the EMMIX software permits attributes to be correlated within components (or clusters or classes) but it currently restricts either all attributes in all classes to be Gaussian or all to be t. Bearing this in mind, the empirical comparisons to follow - where all attributes are uncorrelated and some can be t and some can be Gaussian - are probably somewhat unfair in favour of our method.

6.5 Empirical Comparison: Number of Components

We consider here modelling the datasets, which are generated artificially from the same parameters as those used in subsections 6.1, 6.2 and 6.3, repeatedly 20 times. We also fed the datasets to the EMMIX software in order to see how the modelling criteria used in the software (AIC and BIC) behave toward the datasets.

As shown in Table 3, our proposed MML method showed good performance in determining the number of components in the datasets. AIC and BIC, on the other hand, rarely underfitted the datasets by inferring a smaller number of components. However, BIC showed better modelling than AIC, where for most modellings, AIC chose different numbers of components and tended to highly overfit the true number of these components. Compared to our method, BIC overfitted nearly half of the datasets investigated, especially when modelling one-component mixture datasets. See subsection 6.4 regarding the fairness of these comparisons.

Mixture	One-Component					Two-Component					Three-Component				
Component Number	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
MML Gaussian & t	**20**	0	0	0	0	0	**20**	0	0	0	0	2	**18**	0	0
AIC Gaussian Only*	0	0	5	7	8	0	0	0	7	13	0	0	0	6	14
BIC Gaussian Only*	0	18	2	0	0	0	13	6	1	0	0	2	12	6	0
AIC t Only*	0	0	2	3	15	0	0	2	2	16	0	0	0	3	17
BIC t Only*	7	12	1	0	0	0	15	5	0	0	0	3	15	2	0

*Modelled using the EMMIX software [13]

Table 3. Comparison of the modelling results in terms of the resulting number of components using MML and other criteria such as AIC and BIC, based on 20 trials.

7 Conclusion

In conclusion, we draw the attention of the reader to the following results from Sections 3 and 6:

1. The proposed method shows a better performance in estimating parameters of a single t distribution compared to the ML method for all settings but for large ν and large numbers of data. Smaller standard errors of the estimations proved that the proposed MML estimation is a robust method in performing parameter estimation (see Section 3).
2. The proposed method provides the flexibility for fitting an attribute of a component either as a Gaussian or a t distribution (see subsections 6.1, 6.2 and 6.3), although our attributes are currently uncorrelated.
3. The proposed method shows a good performance in determining the number of components in a dataset. MML rarely had more components than the true model (see subsection 6.5).

References

1. Akaike H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-19, 6 (1974) 716-723
2. Baxter R.A. and Oliver J.J.: Finding overlapping components with MML. Statistics and Computing, 10 (2000) 5-16
3. Boulton D.M.: The information criterion for intrinsic classification. Ph.D. Thesis, Dept. Computer Science, Monash University Clayton 3800 Australia (1975)
4. Chaitin G.J.: On the length of programs for computing finite sequences. Journal of the Association for Computing Machinery, 13 (1966) 547-569
5. Conway J.H. and Sloane N.J.A.: Sphere Packings Lattices and Groups. 3rd edn. Springer-Verlag, London (1998)
6. Everitt B.S. and Hand D.J.: Finite Mixture Distributions. Chapman and Hall, London (1981)
7. Figueiredo, M.A.T. and Jain A.K.: Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3) (2002) 381-396.
8. Hunt L.A. and Jorgensen M.A.: Mixture model clustering using the multimix program. Australian and New Zealand Journal of Statistics, 41(2) (1999) 153-171.
9. Lebedev N.N.: Special functions and their applications. Prentice-Hall, NJ (1965)
10. Liu C. and Rubin D.B.: ML Estimation of t distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 5 (1995) 19-39.
11. McLachlan G.J. and Basford K.E.: Mixture Models. Marcel Dekker, NY (1988)
12. McLachlan G.J. and Peel D.: Finite Mixture Models. John Wiley, NY USA (2000)
13. McLachlan G.J., Peel D., Basford K.E. and Adams P.: The EMMIX software for the fitting of mixtures of Normal and t-components. Journal of Statistical Software, 4 (1999)
14. Schwarz G.: Estimating the dimension of a model. Annals of Statistics, 6 (1978) 461-464
15. Solomonoff R.J.: A formal theory of inductive inference. Information and Control, 7 (1964) 1-22, 224-254
16. Titterington D.M., Smith A.F.M. and Makov U.E.: Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, Chichester (1985)
17. Wallace C.S. An improved program for classification. Proceedings of the Ninth Australian Computer Science Conference (ACSC-9), 8, Monash University Australia (1986) 357-366
18. Wallace C.S. and Boulton D.M.: An information measure for classification. Computer Journal, 11(2), (1968) 185-194
19. Wallace C.S. and Dowe D.L.: MML estimation of the von Mises concentration parameter. Technical Report TR 93/193, Dept. of Computer Science, Monash University Clayton 3800 Australia (1993)
20. Wallace C.S. and Dowe D.L.: Intrinsic classification by MML - the Snob program. In Zhang C. et al. (Eds.), Proc. 7th Australia Joint Conference on Artificial Intelligence. World Scientific, Singapore (1994) 37-44
21. Wallace C.S., and Dowe D.L.: Minimum Message Length and Kolmogorov Complexity. Computer Journal, 42(4) (1999) 270-283, Special issue on Kolmogorov Complexity.
22. Wallace C.S., and Dowe D.L.: MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing, 10(1) (2000) 73-83
23. Wallace C.S. and Freeman P.R.: Estimation and Inference by Compact Coding. Journal of the Royal Statistical Society Series B, Vol. 49(3) (1987) 240-265