Clustering of Gaussian and t Distributions using Minimum Message Length

Yudi Agusta **David L. Dowe**
Computer Sci. and Software Eng., Monash University, Clayton, VIC 3800 Australia
Email: {*yagusta,dld*} *at bruce.csse.monash.edu.au*

Abstract

Clustering or mixture modelling is a problem of identifying and modelling components 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 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). The current study extends this by generalising the Gaussian distribution to the more general, t, distribution. The inference problem of the univariate t distribution with an unknown degree of freedom has been elaborated upon in Agusta and Dowe (2002). In this paper, we modify this by considering the univariate t distribution with a known degree of freedom. Point estimation of the distribution is 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 and compared modelling results with various other criteria, such as AIC (Bozdogan) and BIC, on artificially generated datasets. In terms of the resulting number of components and the (numerically approximated) Kullback-Leibler distances, the results were again impressive. We also considered an application to the (real-world) Old Faithful geyser dataset.

1 Introduction

Clustering - also known as mixture modelling [Hunt and Jorgensen, 1999; McLachlan and Peel, 2000], intrinsic classification [Wallace and Dowe, 1994] and numerical taxonomy - 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, clustering which models a statistical distribution by a mixture (a weighted sum) of other distributions. This type of clustering results in a description of the number of components, the relative abundances (or mixing proportions) of each component, their distribution parameters and the members (or things) that belong to them.

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 Boulton, 1968] and subsequently by [Wallace and Freeman, 1987] dealt with this problem for clustering and parameter estimation problems. The model selection criterion proposed in these papers, i.e., Minimum Message Length (MML), provides a fair comparison between models by stating each of them into a two-part message which encodes each model and the data in light of the model stated. Various related principles have also been stated independently by [Solomonoff, 1964; Kolmogorov, 1965; Chaitin, 1966], and subsequently by [Rissanen, 1978]. For an overview, see [Wallace and Dowe, 1999].

The MML clustering method proposed in [Wallace and Boulton, 1968] dealt with a clustering problem of discrete multinomial and continuous Gaussian distributions. It was extended by [Wallace and Dowe, 1994; 2000] to accommodate two other distributions - Poisson and von Mises circular. The method was further broadened to accommodate a t distribution with an unknown degree of freedom in [Agusta and Dowe, 2002]. The generalisation of the Gaussian to t distribution was considered in order to provide a flexibility in modelling datasets which contain atypical observations, such as outliers. An alternative MML-based approach clustering was also proposed by [Figueiredo and Jain, 2002].

Beginning with parameter estimation, this paper proposes an MML clustering method of Gaussian and t distributions by considering a known, as well as an unknown, degree of freedom (ν) for the t distribution. We extend the results provided in [Agusta and Dowe, 2002] by comparing the proposed clustering method with two other commonly used criteria, AIC and BIC, in terms of the resulting number of components and the Kullback-Leibler distances from the true to the inferred model.

2 Parameter Estimation by MML

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 hypothesis [Wallace and Boulton, 1968; Wallace and Freeman, 1987; Wallace and Dowe, 1999].

Letting D be the data and H be an hypothesis 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

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

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

In applying the MML principle to the model selection of Gaussian and t distributions, we need parameter estimations of the Gaussian and t distributions and in addition, the multi-state distribution for the clustering problem. The parameter estimation used here utilises the MML approximation proposed by [Wallace and Freeman, 1987].

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

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

the Fisher information - that is the determinant of the matrix of expected second derivatives of the negative log-likelihood. Based on (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 [Wallace and Freeman, 1987; Wallace and Dowe, 2000]:

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

where D is the number of parameters to be estimated and κ_D is a D-dimensional lattice constant with $\kappa_1 = 1/12$ and $\kappa_D \leq 1/12$. The MML estimate of $\vec{\theta}$ can be obtained by minimising (3).

Considering that both Gaussian and t distributions 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 [Wallace and Dowe, 2000, p74] [Agusta and Dowe, 2002, Sec. 2] [Wallace and Dowe, 1994, p38]. In addition, for any continuous attributes, we assume the scale parameter σ is at least 0.4 ϵ.

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 (2), it follows 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 $M = 2$ [Wallace and Dowe, 2000, 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 obtained by:

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

Substituting (4) into (3) provides the following total two-part message length [Wallace and Boulton, 1968, p187 (4)] [Wallace and Boulton, 1968, p194 (28)] [Wallace and Dowe, 2000, p75]:

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

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:

$$F(\mu, \sigma) = 2N^2/\sigma^4 \text{ or by [Wallace and Dowe, 2000] : } \mathrm{F}(\mu, \sigma^2) = \mathrm{N}^2/2(\sigma^2)^3$$

Assuming a uniform prior on μ over a finite range R 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 (3), the MML estimates $\hat{\mu}$ and $\hat{\sigma}$ are then calculated by:

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

For the range of the prior on σ mentioned above, we need to normalise the prior by 0.125 and for the range $R = [-\frac{|R|}{2}, \frac{|R|}{2}]$ of the prior on μ, in practice we choose max$\{$10, the difference between the maximum and the minimum value of the sample data$\}$.

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 the 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:

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

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

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \text{ where we let } \psi(x) = \mathrm{d}\Gamma(x)/\mathrm{d}x \text{ and } \psi^{(1)}(x) = \mathrm{d}^2\Gamma(x)/\mathrm{d}x^2$$

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:

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

For estimation purposes, we consider two separate cases for the third parameter ν: firstly, ν as a known parameter (considered here) and secondly, ν as an unknown continuous parameter as elaborated in [Agusta and Dowe, 2002]. Using (2), the Fisher information, $F(\mu, \sigma)$, for the t distribution when the value of ν is known, is:

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

Assuming a uniform prior on μ over a finite range R and a $1/\sigma$ prior on σ over the range $[e^{-4}, e^4]$, the MML estimators of the t distribution (with known ν), $\hat{\mu}_{\mathrm{MML}}$ and $\hat{\sigma}_{\mathrm{MML}}$, are obtained by minimising (3) with respect to each parameter. Since there are no sufficient statistics to estimate the parameters and the parameter estimations are coupled, the inference is performed using a binary search by setting $\partial\mathrm{MessLen}/\partial\vec{\theta} = 0$ and iterating the search process until a certain precision of estimation is obtained.

2.4 Point Estimation of One Univariate t Model

In this subsection, we compare the MML parameter estimation of a one-component univariate t model with a known ν to the results obtained using the Maximum Likelihood (ML) method in terms of their Kullback-Leibler (KL) distances. The differences between the ML and MML methods in estimating multi-state and Gaussian variables have been elaborated upon by [Wallace and Dowe, 2000]. For the t distribution with unknown ν, refer to [Agusta and Dowe, 2002].

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.

σ=1.0 & ν=		1.0	3.0	10.0	25.0	50.0	100.0
N=10	ML	0.134±0.13	0.126±0.15	0.106±0.12	0.139±0.21	0.156±0.20	0.159±0.26
	MML	0.102±0.10	0.103±0.12	0.093±0.10	0.121±0.17	0.133±0.16	0.137±0.22
N=100	ML	0.012±0.01	0.010±0.01	0.012±0.01	0.009±0.01	0.010±0.01	0.012±0.02
	MML	0.011±0.01	0.010±0.01	0.012±0.01	0.009±0.01	0.010±0.01	0.011±0.02
σ=5.0 & ν=		1.0	3.0	10.0	25.0	50.0	100.0
N=10	ML	0.104±0.11	0.129±0.13	0.133±0.17	0.143±0.17	0.162±0.23	0.175±0.21
	MML	0.081±0.08	0.114±0.11	0.119±0.15	0.124±0.15	0.143±0.19	0.146±0.17
N=100	ML	0.015±0.02	0.010±0.01	0.010±0.01	0.010±0.01	0.009±0.01	0.010±0.01
	MML	0.014±0.01	0.010±0.01	0.009±0.01	0.010±0.01	0.009±0.01	0.009±0.01

Table 1: Kullback-Leibler distances of the ML and MML estimations of 100 datasets for t distributions with μ=0.0 when ν is known (with ± standard errors).

The datasets for Table 1 were generated repeatedly (100 times) from t distributions with $\nu = \{1.0, 3.0, 10.0, 25.0, 50.0, 100.0\}$, $\sigma = \{1.0, 5.0\}$, $\mu = \{0.0\}$ and the number of data, $N = \{10, 100\}$. Both estimators infer the datasets with the known value of ν set to the true value. 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 ν, σ and N. The MML also performed a more robust estimation, with smaller standard errors of the resulting Kullback-Leibler distances. This case of MML outclassing ML is reminiscent of the von Mises circular distribution [Wallace and Dowe, 1993].

3 MML Clustering

In order to apply MML to a clustering problem, a two-part message conveying the mixture model needs to be constructed. Recall from Section 2, the hypothesis for the mixture model 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 the component attribute: In this case, a component attribute is inferred as a Gaussian or a t distribution as in subsections 2.2 and 2.3, respectively. The MML model selection of Gaussian and t distributions is simultaneously applied - and, to specify which, for every attribute 1 bit ($= \log_e 2$ nits) is added to the message length.

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 MML clustering was first introduced. The original coding scheme [Wallace and Boulton, 1968] utilised a total assignment of things to components. In this paper, we utilise an assignment which assigns data 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. A comparison between total and partial assignments can be seen in [Wallace, 1986, Sec. 3] [Wallace and Dowe, 2000, pp77-78][Agusta and Dowe, 2002, 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 hypothesis that gives the shortest/minimum message length.

4 Alternative Model Selection Criteria - AIC (Bozdogan) and BIC

For comparison, two criteria are considered here. These are *Bozdogan's Akaike Information Criterion* ($\mathrm{AIC}_{\mathrm{Bozdogan}}$) and *Schwarz's Bayesian Information Criterion* (BIC).

AIC was first developed by Akaike [Akaike, 1974] in order to identify the model of a dataset. This was then extended by Bozdogan [Bozdogan, 1983] as a criterion for clustering problems. Bozdogan's AIC for model selection is given by: $\mathrm{AIC}_{\mathrm{Bozdogan}} = -2(N - 1 - d - K/2)L/N + 3N_p$, where N is the number of data, d is the number of parameters describing each component, K is the largest number of components considered, 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. This criterion was used elsewhere [Windham and Cutler, 1992, Sec.2] in comparisons between it and other criteria. The model which results in the smallest $\mathrm{AIC}_{\mathrm{Bozdogan}}$ is the model selected. Unable to find an explicit specification, we set $K = 100$. Because the number of parameters of a t distribution is different to that of a Gaussian, we set d to be the arithmetic mean of the number of parameters per component.

The second criterion, BIC, was first introduced by Schwarz [Schwarz, 1978] and is given by: $\mathrm{BIC} = -L + N_p \log N/2$, 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. The model which results in the smallest BIC is selected as the best model.

5 Experimental Evaluations of MML Clustering

5.1 Example 1

For this example, we use the same datasets as the datasets used in [Agusta and Dowe, 2002, Sec. 6.3], which were generated artificially from both Gaussian and t distributions according to the caption to Figure 1. The experiment was repeated 50 times (one example is in Figure 1) and each dataset was inferred using Gaussian and t distributions with a known, as well as an unknown, degree of freedom, ν. Table 2 shows the modelling results using our proposed MML method on the example dataset in Figure 1. As shown in Table 2, the first attribute of the second component was inferred as a t distribution both with $\nu = 2.594$ (when ν was regarded as unknown) and with (fixed/"known") $\nu = 1.0$. The other five attributes were inferred as Gaussian distributions, as they are in the true model.

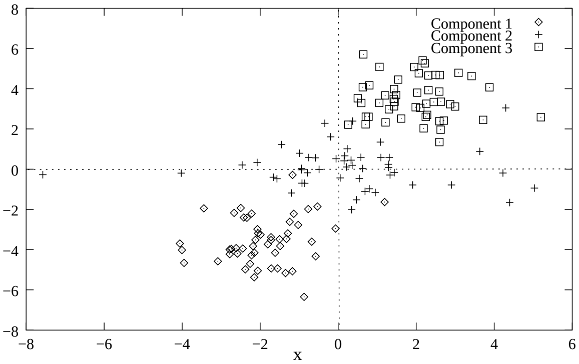

Figure 1: Three-component mixture of 150 bivariate data points generated from a combination of t and Gaussian distributions: $1/3(N(\mu_{x1} = -2, \sigma^2_{x1} = 1) \times N(\mu_{y1} = -3.5, \sigma^2_{y1} = 1)) +$
$1/3(t_{\nu_{x2}=1}(\mu_{x2} = 0, \sigma^2_{x2} = 1) \times N(\mu_{y2} = 0, \sigma^2_{y2}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

		MML: Fixed/Known $\nu(= 1.0)$			MML: Unknown ν		
Component Number		Comp. 1	Comp. 2	Comp. 3	Comp. 1	Comp. 2	Comp. 3
Mixing Proportion		0.316	0.323	0.360	0.314	0.326	0.360
Attribute 1	Mean(μ)	-1.999	0.166	1.926	-2.005	0.124	1.926
	SD(σ)	0.862	0.926	1.070	0.860	1.202	1.071
	DoF(ν)	∞	**1.000**	∞	∞	**2.594**	∞
Attribute 2	Mean(μ)	-3.704	-0.164	3.397	-3.716	-0.171	3.401
	SD(σ)	1.007	0.760	1.017	1.000	0.766	1.014
	DoF(ν)	∞	∞	∞	∞	∞	∞

Table 2: One example (see dataset in Figure 1) of MML modelling results with a known, as well as an unknown, degree of freedom, ν. (Recall that a Gaussian distribution is a t distribution with $\nu = \infty$.)

Inferred results giving both the the estimated number of components and the Kullback-Leibler distances from the true model were obtained for $\text{AIC}_{\text{Bozdogan}}$, BIC and our proposed MML method here. The data generation process is in the caption to Figure 1 and the results from 50 trials are presented in Table 3. Unlike our comparison with EMMIX in [Agusta and Dowe, 2002], in the comparison between $\text{AIC}_{\text{Bozdogan}}$, BIC and our proposed MML method in Table 3, all three criteria permit any attribute to be Gaussian or t in any component and all three criteria here assume that - within a component - all attributes are independent and uncorrelated. This provides a fair comparison.

In this comparison between modelling criteria (Table 3), when ν was regarded as a known parameter ($\nu = 1.0$), our proposed MML clustering method showed better performances in selecting the number of components - with 49 out of 50 datasets modelled the same as the true model. For unknown ν, our proposed MML clustering method selected the correct number of components for 48 out of the 50 datasets. In terms of the Kullback-Leibler distances, for both cases of known and unknown

ν, our proposed MML method inferred the dataset closest to the true model, with smaller average Kullback-Leibler distances compared with both $\text{AIC}_{\text{Bozdogan}}$ and BIC.

		Known $\nu(=1.0)$			Unknown ν		
		MML	AIC_B	BIC	MML	AIC_B	BIC
Avg. Kullback Leibler distances		0.889	1.542	1.564	0.365	1.546	1.565
Number of Components	1	0	0	0	0	0	0
	2	1	0	3	2	2	1
	3 true	**49**	36	38	**48**	32	36
	4	0	12	8	0	12	12
	5	0	2	1	0	4	1

Table 3: Comparison of modelling results (for the generating function in the caption to Figure 1) in terms of the average Kullback-Leibler distances and the resulting number of components between MML, $\text{AIC}_{\text{Bozdogan}}$ and BIC, based on 50 trials.

5.2 Example 2

In this subsection, we consider an application of our proposed MML clustering method to the Old Faithful geyser dataset. The dataset consists of 272 univariate measurements of eruption lengths of the Old Faithful geyser and was originally reported in [Silverman, 1986]. The dataset has been re-analysed recently in [McLachlan and Peel, 2000], with the data being lagged into a bivariate dataset and analysed for detecting the presence of outliers. In this paper, we modelled the dataset for the t distribution with both a known ν set to 1.0 and an unknown ν.

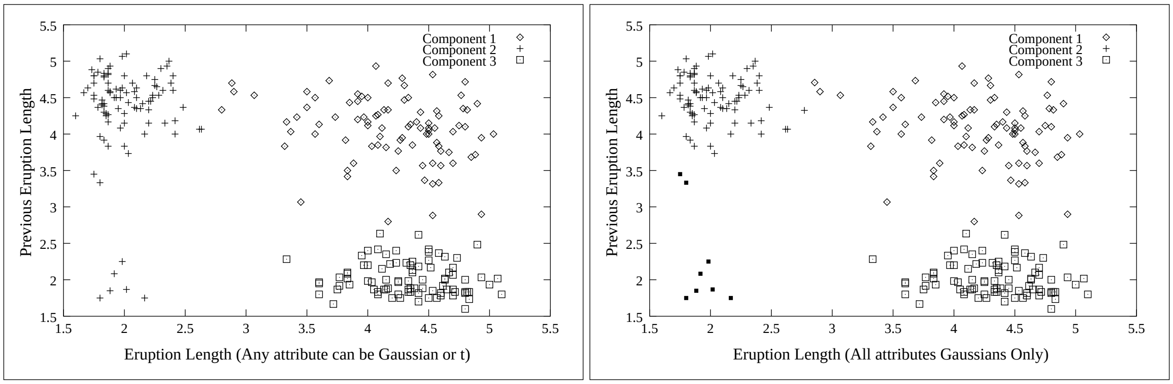

Figure 2: MML Modelling results for the Old Faithful geyser dataset using Gaussian-or-t with known and unknown ν (left) and Gaussian-only (right).

As shown in Figure 2 left, the dataset was modelled as a three-component mixture when using our proposed MML method for both cases of known and unknown ν. In these results, the second attribute of the second component was inferred as a t distribution - with $\nu = 1.0$ when ν was known (see Table 4 left), and with $\nu = 1.679$ when ν was unknown (see Table 4 middle). For the sake of comparison, we investigated further and found that the best MML model with all attributes required to be Gaussian had four components (see Figure 2 right and Table 4 right). In this model, eight observations in the bottom left and middle left of the graph (in dots) were grouped into a separate Gaussian component. Looking at the message lengths of the three models, it can be seen that modelling using Gaussian and t distributions with known, as well as unknown, ν resulted in slightly shorter message lengths compared to modelling with Gaussian distributions only. Our analysis contrasts with that of [McLachlan and Peel, 2000], who had a similar but different class of models to us. We permit any attribute to be Gaussian or t with no within-component correlations [Agusta and Dowe, 2002], whereas McLachlan and Peel permit correlations within components but insist that either all attributes be t with identical ν or all attributes be Gaussian ($\nu = \infty$).

		MML Gaussian or t w/ Known ν (= 1.0)			MML Gaussian or t with Unknown ν			MML Gaussians Only			
Component No.		1	2	3	1	2	3	1	2	3	4
Mix. Prop.		0.330	0.344	0.326	0.328	0.345	0.326	0.322	0.321	0.326	0.031
Att.1	Mean(μ)	4.174	2.007	4.342	4.179	2.009	4.342	4.213	2.035	4.350	1.818
	SD(σ)	0.507	0.213	0.371	0.500	0.215	0.371	0.465	0.236	0.372	0.166
	DoF(ν)	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Att.2	Mean(μ)	4.072	4.515	2.022	4.071	4.510	2.021	4.074	4.510	2.025	2.189
	SD(σ)	0.425	0.214	0.227	0.426	0.256	0.227	0.425	0.298	0.228	0.656
	DoF(ν)	∞	**1.000**	∞	∞	**1.679**	∞	∞	∞	∞	∞
Message Length		4319.281 nits			4317.763 nits			4320.312 nits			

Table 4: MML Modelling results for the (lagged, bivariate) Old Faithful geyser dataset.

6 Conclusion

In conclusion, we draw the reader's attention to the following results from Subsection 2.4 and Section 5:

1. The proposed MML inference method for a univariate t distribution with known degree of freedom shows a better performance in estimating the parameters compared to the Maximum Likelihood (ML) method for all values of ν, σ and N. Smaller standard errors of the estimations further show that the proposed MML method robustly performs parameter estimation (see Subsection 2.4). The strong performance of MML here is not unlike that with the von Mises circular distribution.
2. The proposed MML clustering method shows a better performance in determining the number of components compared to both $\mathrm{AIC}_{\mathrm{Bozdogan}}$ and BIC. Smaller Kullback-Leibler distances show that our method modelled the datasets closer to the true model compared to both $\mathrm{AIC}_{\mathrm{Bozdogan}}$ and BIC (see Subsection 5.1).

References

[Agusta and Dowe, 2002] Yudi Agusta and David L. Dowe. *MML Clustering of Continuous-valued Data using Gaussian and t Distributions.* To appear in Proc. 15th Australian Joint Conference on Artificial Intelligence, Canberra: Lecture Notes in Artificial Intelligence (LNAI), Springer-Verlag.

[Akaike, 1974] Hirotugu Akaike. *A new look at the statistical model identification.* IEEE Trans. on Automatic Control, AC-19, 6, 716-723.

[Bozdogan, 1983] Hamparsum Bozdogan. *Determining the number of component clusters in the standard multivariate normal mixture model using model-selection criteria.* TR UIC/DQM/A83-1, Quantitative Methods Dept., University of Illinois, Chicago, Illinois 60680.

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

[Figueiredo and Jain, 2002] Mario A.T. Figueiredo and Anil K. Jain. *Unsupervised Learning of Finite Mixture Models.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(3), 381-396.

[Hunt and Jorgensen, 1999] Lynette A. Hunt and Murray A. Jorgensen. *Mixture model clustering using the multimix program.* Australian and New Zealand Journal of Statistics, 41(2), 153-171.

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

[Liu and Rubin, 1995] Chuanhai Liu and Donald B. Rubin. *ML Estimation of t distribution using EM and its extensions, ECM and ECME.* Statistica Sinica, 5, 19-39.

[McLachlan and Peel, 2000] Geoff J. McLachlan and David Peel. *Finite Mixture Models.* Wiley, NY.

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

[Schwarz, 1978] Gideon Schwarz. *Estimating the dimension of a model.* Annals of Stat., 6, 461-464.

[Silverman, 1986] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

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

[Wallace, 1986] Chris S. Wallace. *An improved program for classification.* Proceedings of the Nineth Australian Computer Science Conference (ACSC-9), 8, Monash University, Australia, 357-366.

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

[Wallace and Dowe, 1993] Chris S. Wallace and David L. Dowe. *MML estimation of the von Mises concentration parameter.* Technical Report TR 93/193, Dept. of Computer Science, Monash University, Clayton 3168, Australia, 1993.

[Wallace and Dowe, 1994] Chris S. Wallace and David L. Dowe. *Intrinsic classification by MML - the Snob program.* In: Zhang C. et al. (Eds.), Proc. 7th Australia Joint Conference on Artificial Intelligence. World Scientific, Singapore, 37-44.

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

[Wallace and Dowe, 2000] Chris S. Wallace and David L. Dowe. *MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions.* Statistics and Computing, 10(1), 73-83.

[Wallace and Freeman, 1987] Chris S. Wallace and Peter R. Freeman. *Estimation and Inference by Compact Coding.* Journal of the Royal Statistical Society Series B, Vol. 49, No. 3, 240-265.

[Windham and Cutler, 1992] Michael P. Windham and Adele Cutler. *Information Ratios for Validating Mixture Analyses.* Journal of the American Statistical Association, 87, 1188-1192.