# Finding Overlapping Distributions with MML

ROHAN A. BAXTER (rohan@cs.monash.edu.au)
JONATHAN J. OLIVER (jono@cs.monash.edu.au)
*Computer Science Department*
*Monash University*
*Clayton, Victoria, 3168, AUSTRALIA*

**Abstract:** This paper considers an aspect of mixture modelling. Previous studies have shown minimum message length (MML) estimation to perform well in a wide variety of mixture modelling problems, including determining the number of components which best describes some data. In this paper, we focus on the difficult problem of overlapping components.

An advantage of the probabilistic mixture modelling approach is its ability to identify models where the components overlap and data items can belong probabilistically to more than one component. Significantly overlapping distributions require more data for their parameters to be accurately estimated than well separated distributions. For example, two Gaussian distributions are considered to significantly overlap when their means are within three standard deviations of each other. If insufficient data is available, only a single component distribution will be estimated, although the data originates from two component distributions.

In this paper, we quantify this difficulty in terms of the number of data items needed for the MML criterion to 'discover' two overlapping components. First, we perform experiments which compare the MML criterion's performance relative to other Bayesian criteria based on MCMC sampling. Second, we make two alterations to the existing MML estimates in order to improve its performance on overlapping distributions. Experiments are performed with the new estimates to confirm that they are effective.

## 1 Introduction

An advantage of the probabilistic mixture modelling approach is its ability to identify models where the components overlap and data items can belong probabilistically to more than one component. However this advantage is tempered by the additional difficulty in parameter estimation where the components overlap. In this paper, we quantify this difficulty in terms of the number of data items needed for the MML criterion to 'discover' two overlapping components. Two prominent probabilistic mixture model programs include Autoclass [CSK$^+$88, CS95] and Snob [WB68, Wal90, WD94]. For a comparison of Snob and Autoclass see [UN96, Upa95].

The MML criterion has previously been found to perform very well against non-Bayesian criteria [OBW96]. Recent Bayesian results have appeared recently based on MCMC sampling methods. It will be interesting to see whether these methods offer any advantages. In section 2.1 we conduct some preliminary experiments comparing MML with results from work in MCMC sampling.

The MML criterion is consistent, meaning that it will choose the correct model given enough data. How much data is enough? This has been answered theoretically by Barron and Cover [BC91]. The results in section 3 provide an empirical efficiency curve for the simple mixture models considered. We experimentally find how many data points are required for MML to find the correct number of components with high probability for different separation of components.

The MML estimates use approximations requiring the assumption of well separated components[OBW96]. We consider two alternative approximations which reduce the reliance on this assumption in section 3.1. The performance of the two approximations are then tested to verify their efficacy.

The results here will assist mixture modellers. They will help mixture modeller practitioners decide *a priori* whether they have enough data to successfully find two components for a particular separation. Intuition in this area can be misleading. We have observed unjustified optimism in identifying structure with little data and overlapping components. The results here have been used in our own application of mixture models in a number of domains.

## 2 Mixture Models

We consider the univariate case and concentrate on models where the $k$ component distributions are Gaussian, i.e., $f_j(x) \sim N(\mu_j,\ \sigma_j^2)$. Each component has a proportional parameter $p_j$.

Wallace and Freeman [WF87] provide a general message length expression applicable to mixture models:

$$MessLen(\tilde{x},\theta) \approx -\log h(\theta) + \frac{1}{2}\log det(F(\theta)) - \log f(\tilde{x}|\theta) + \frac{n_p}{2} + \frac{n_p}{2}\log \kappa_{n_p} \tag{1}$$

where $h(\theta)$ is a prior distribution over parameter values, $det(F(\theta))$ is the determinant of the *expected* Fisher Information matrix.

$f(\tilde{x})$ is the likelihood function of the mixture, $n_p$ is the number of parameters been estimated, where $\kappa_{n_p}$ is the $n_p$ dimensional optimal quantizing lattice constant, where $\kappa_1 = \frac{1}{12}$ and $\kappa_2 = \frac{5}{36\sqrt{3}}$. We used values for $\kappa_{n_p}$ from Table 2.3 of Conway and Sloane [CS88].

The prior distribution is described in [OBW96]. $\mu_j$ are considered to be uniformly distributed in the range, $[\mu_{pop}-\sigma_{pop}, \mu_{pop}+\sigma_{pop}]$, where $\sigma_{pop}$ is the standard deviation of all the data. $\sigma_j$ are considered to be uniformly distributed in the range, $(0, \sigma_{pop}]$. The prior for $p_j$ is uniform over the simplex.

For a single Gaussian distribution, $N(\mu_j, \sigma_j)$,

$$det(F_j(\mu_j,\ \sigma_j)) \ = \ \frac{2n_j^2}{\sigma_j^4} \tag{2}$$

The $p_j$ can be viewed as being the parameters of a multinomial distribution:

$$det(F(p)) \ \approx \ \frac{n}{\prod_{j=1}^{j=k} p_j} \tag{3}$$

We approximate the expected Fisher Information matrix of the mixture distribution in two senses. First, we treat the expected sufficient statistics for the incomplete data as if they were the actual sufficient statistics for the unavailable complete data [CH96]. Second, we only use the diagonal entries of the expected Fisher Information matrix. This approximation is accurate for well-separated components where the off-diagonal entries are then negligible. For overlapping components, this approximation is poor. In section 3.1, we consider two alternatives to improve the approximation for overlapping components.

Having instantiated the terms of Equation (1), the expression we wish to minimise is then [OBW96]:

$$\begin{aligned} MessLen(\tilde{x},\theta) \ \approx \ & -k\log\frac{1}{2\sigma_{pop}^2} \ - \ \log(k-1)! - \ \log k! \\ & + \sum_{j=1}^{k}\log\frac{\sqrt{2}n_j}{\sigma_j^2} \ + \ \frac{1}{2}\log n \ - \ \frac{1}{2}\sum_{j=1}^{k}\log p_j + \frac{n_p}{2}\log\kappa_{n_p} \\ & - \ \log f(\tilde{x}) + \frac{n_p}{2} \end{aligned}$$

## 2.1 MML's performanced compared to MCMC Methods

The MML criterion performs favourably on small samples relative to other criteria, such as various penalized likelihood criteria [OBW96, BOH96]. This is not surprising since the other criteria considered use approximations based on asymptotic behaviour (e.g. the Laplace approximation is used for the Bayesian Information Criterion(BIC)). The MML criterion does not use approximations based on asymptotics.

Bayesian MCMC methods for identifying the number of components are still under development [Rob96, RG96]. Mengersen and Robert [MR93] test for the presence of a mixture, while Richardson and Green [RG96] consider varying numbers of components.

We ran our program on the two mixtures of two normal distributions studied by Richardson and Green:

| | |
|---|---|
| model #6: | $0.5\mathcal{N}(-1,(\frac{2}{3})^2)+0.5\mathcal{N}(1,(\frac{2}{3})^2)$ |
| model #7: | $0.5\mathcal{N}(-1.5,(\frac{1}{2})^2)+0.5\mathcal{N}(1.5,(\frac{1}{2})^2)$ |

These are models #6 and #7 used by Marron and Wand [MW92], which represent bimodal distributions, moderately- and well-separated respectively. We generated $n = 50$ and $n = 250$ data points and report results from 50 replications in Table 1.

We cannot make any hard comparisons, because the priors and search (for estimates) used by the two methods differ. In particular, different priors may give one method 'more information' than the other method. However, at a high level we present the results alongside one another. It is reassuring to note that the results are somewhat similar. Such a small study is not very discriminatory. We note that other methods used by Richardson and Green, such as Bayes Factors and BIC, also give similar results.

| | Probabilities | | |
|---|---|---|---|
| | $\hat{k}=1$ | $\hat{k}=2$ | $\hat{k}=3$ |
| model #6 | | | |
| n = 50 | 0.56 | 0.44 | 0.00 |
| n = 250 | 0.00 | 0.94 | 0.06 |
| model #7 | | | |
| n = 50 | 0.00 | 0.94 | 0.06 |
| n = 250 | 0.00 | 0.98 | 0.02 |

| | Probabilities | | |
|---|---|---|---|
| | $\hat{k}=1$ | $\hat{k}=2$ | $\hat{k}=3$ |
| model #6 | | | |
| n = 50 | 0.58 | 0.40 | 0.02 |
| n = 250 | 0.00 | 1.00 | 0.00 |
| model #7 | | | |
| n = 50 | 0.00 | 0.98 | 0.02 |
| n = 250 | 0.00 | 0.98 | 0.02 |

Table 1: Richardson and Green's Results (left), MML's Results (right)

Mengersen and Robert use an interesting novel approach based on the Kullback-Leibler distance measure. Their method shares MML's advantage in being non-asymptotic. However, their method is restricted to comparing one and two component mixture models because of difficulties in approximating the Kullback-Leibler distance for more than two components. Their experiments comprised the following five mixture distributions:

| | |
|---|---|
| $0.5\mathcal{N}(0,1)+0.5\mathcal{N}(2,1)$ | (i) |
| $0.5\mathcal{N}(0,1)+0.5\mathcal{N}(2,0.5)$ | (ii) |
| $0.5\mathcal{N}(0,1)+0.5\mathcal{N}(0,0.25)$ | (iii) |
| $\mathcal{N}(0,1)$ | (iv) |
| $0.3\mathcal{N}(0,1)+0.7\mathcal{N}(2.5,0.8)$ | (v) |

These models reflect a range of separations, from strongly bimodal through to strongly mixed and truly homogeneous. We generated $n = 100$ data points and report results from 100 replications (following Mengersen and Robert) using the MML criterion. The results are shown in Table 2. The general trend of the results is similar to those of Mengersen and Robert.

| | Probabilities | | | True |
|---|---|---|---|---|
| | $\hat{k} = 1$ | $\hat{k} = 2$ | $\hat{k} = 3$ | $k$ |
| (i) | 0.94 | 0.06 | 0.00 | 2 |
| (ii) | 0.02 | 0.90 | 0.08 | 2 |
| (iii) | 0.06 | 0.86 | 0.08 | 2 |
| (iv) | 0.98 | 0.02 | 0.00 | 1 |
| (v) | 0.09 | 0.90 | 0.01 | 2 |

Table 2: MML's Results, (i) - (v)

# 3 Experiments

In this section, we examine how much data is required in order to discriminate between two distributions using the MML criterion for different separations.

We generated 100 datasets from $0.5N(0,1) + 0.5N(d,1)$ of size $n$ for $n$ varying from 5 to 500 and $d$ varying from 0.5 to 5. We chose 5 as the lower-bound on $n$ because that is the number of parameters in the two component model. We use the EM algorithm to estimate the parameters of the mixture models. We calculated the message lengths for a two-component model and for a one-component model using Equation (4).

When the true distribution contained two components, we examined how much data is required MML to discover the two components when the two components overlap significantly for more than half the datasets. We then chose the first $n$ where two components were chosen in 95 or more of the 100 cases (in 5 consecutive separate runs). The results are shown in Table 3 show two cases. In the first case, we used the EM-MML estimates for the parameters of the one- and two-component mixtures. In the second case, we 'cheated' and used the true generating parameters for the two-component mixture. The second case is useful, because it shows the MML performance independent of the parameter estimation accuracy. As the component separation decreases, the accuracy of parameter estimation decreases. A point of interest is separation 3. It is below this separation that the distribution changes from unimodal to bimodal. It is around this point that we observe the number of data items needed for the correct model to be identified increases dramatically.

## 3.1 Corrections to Message Length:The Observed Fisher Information

The Fisher Information approximation of Equation (2) implicitly assumes the component distributions are well separated. When this assumption is invalid, the resulting MML estimates are 'more certain' than they should be and so the model message length is inefficient.

We may avoid this assumption by using the observed Fisher Information as an estimate of the expected Fisher Information. The full derivation of the terms needed are in [Bax95]. We provide a brief outline here. Even without expectations, the observed Fisher Information of a mixture model is difficult to calculate analytically. Let the gradient vector be:

$$\Delta = (\frac{\partial}{\partial \theta_1}(-\log f(x|\theta)), \ldots, \frac{\partial}{\partial \theta_d}(-\log f(x|\theta)))$$

| Separation | $n$ for MML to select true model with | |
|---|---|---|
| of means | estimated parameters | true parameters |
| 8 | 18 | 22 |
| 7 | 21 | 27 |
| 6 | 20 | 34 |
| 5 | 24 | 55 |
| 4 | 67 | 102 |
| 3 | 193 | 302 |
| 2 | 1500 | 2380 |
| 1 | > 10000 | > 10000 |

Table 3: $n$ for which MML selects true model with greater than 0.95 probability

We can approximate the observed Fisher Information matrix by [MB88]

$$F_{obs}(\theta) = E_x(\Delta^T \Delta) \tag{4}$$

MML mixture modelling programs, such as Snob, will evaluate thousands of candidate models as part of the search for the best one. Unfortunately, the observed Fisher information calculation requires $2dn$ first derivative calculations, where $d$ is the number of dimensions and $n$ is the amount of data and also the inversion of a matrix with dimension $3 * d + (d - 1)$ (in order to compute the determinant). This makes it desirable to find a computationally cheaper approximation.

We call this approximation[1], the *Efficient Fisher* Information. This approximation is derived in [Bax95]. Consider computing the first derivative of the negative log-likelihood of the mixture model for parameter $\theta_j$, where $\theta_j$ is one of the parameters of component $j$ ($\mu_j, \sigma_j$,or $p_j$). For simplicity let us just describe the result for a two-component mixture so that $j = \{1, 2\}$:

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}(-\log f(x|\theta)) &= \frac{\frac{\partial p_j f(x|\mu_j,\sigma_j)}{\partial\theta_j}}{p_1 f(x|\theta_1) + p_2 f(x|\theta_2)} \\
&= \frac{\partial p_j f(x|\mu_j,\sigma_j)}{\partial\theta_j} \times \frac{1}{p_1 f(x|\theta_j)} \times \frac{p_1 f(x|\theta_j)}{p_1 f(x|\theta_1) + p_2 f(x|\theta_2)} \\
&= \frac{\partial}{\partial\theta_j}(-\log f(x|\theta_j)) \times w_j \qquad (5)
\end{aligned}$$

where $w_j = \frac{p_1 f(x|\theta_j)}{p_1 f(x|\theta_1)+p_2 f(x|\theta_2)}$. Consider only the diagonal entries of the observed Fisher Information matrix in Equation (4) and using the weight $w_j$, we can approximate the determinant of the observed Fisher Information by the following:

$$|F_{obs}(\theta)| \approx |\prod_{\theta_j}(\sum_{i=1}^{n}(\frac{\partial}{\partial\theta_j}(-\log f(x_i|\theta_j)) \times w_j)^2) \tag{6}$$

We can approximate the average of the square of the first derivative in Equation (6) by the expected Fisher Information for that entry:

$$|F_{eff}(\theta)| \approx F(\theta_j) \times \sum_{i=1}^{n} w_j^2 \tag{7}$$

We repeated the experiments of the last section. In Figure 1, we graph the log of determinant of the expected, observed and efficient Fisher information matrix,

[1] The suggestion for the approximation is due to Chris Wallace.

when the separation is 10, for differing $n$. We see that the three Fisher Informations roughly parallel each other. The efficient Fisher Information seemingly acts as a lower bound on the observed Fisher Information.

In Figure 3.1, we show the result when the separation is 1 (and the distributions overlap significantly). We now see that the observed Fisher Information is smaller than the expected Fisher Information. The efficient Fisher Information approximation is in between the other two.

Using the observed Fisher information results in a higher probability in selecting the correct probability for small $n$. For a separation of 3, this is shown in Figure 3.

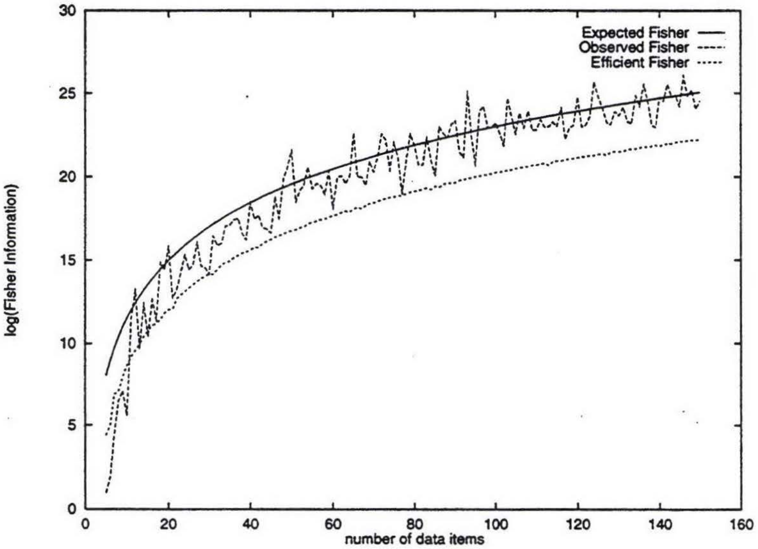

Figure 1: log of determinant of the Expected, Observed and Efficient Fisher Information estimates with separation = 10

# 4 Discussion and Conclusion

Although the MML criterion appears to perform as well as the Bayesian MCMC sampling methods, we have considered two alternative ways to improve its performance for overlapping components.

We have considered a one dimensional two-component problem. Applications using MML mixture models often have thirty or more dimensions. Each new dimension involves three new parameters for each component. Further results characterizing the increase in data items needed for overlapping multi-dimensional components will be of interest. The importance of the savings due to using the observed Fisher Information also increases with higher dimensions.

The usual MML estimate, using an approximation to the expected Fisher Information, is inefficient for overlapping distributions. The use of the observed Fisher information results in a slight improvement, but at an increased computational cost. The efficient observed Fisher Information approximation provides a compromise between computation cost and criterion performance.

## Acknowledgments

We would like to thank the anonymous referees for their helpful comments. This work was supported by a Research Fellowship funded by an Australian Research Council (ARC) Postdoctoral Research Fellowship F39340111.

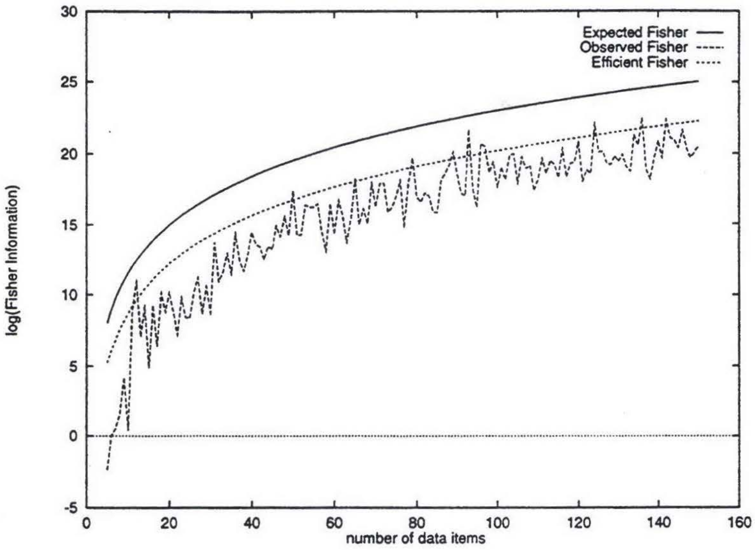

Figure 2: log of the determinant of the Expected, Observed and Efficient Fisher Information estimates with separation = 1

# References

[Bax95] R.A. Baxter. Finding overlapping distributions with MML. Technical Report 244, Dept. of Computer Science, Monash University, Clayton 3168, Australia, November 1995. Available on the WWW from `http://www.cs.monash.edu.au/~rohan`.

[BC91] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. on Info. Theory*, 37:1034–1054, 1991.

[BOH96] R.A. Baxter, J.J. Oliver, and D. Hand. Fitting finite Gaussian mixture models using minimum message length estimation. *The IMS Bulletin*, 25(4), Jul/Aug 1996.

[CH96] D. Chickering and D. Heckerman. Efficient Approximation for the Marginal Likelihood of Incomplete Data given a Bayesian Network. In *UAI'96*, pages 158–168. Morgan Kaufmann, 1996.

[CS88] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag,New York, 1988.

[CS95] P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* 1995.

[CSK+88] P. Cheeseman, M. Self, J. Kelly, W. Taylor, D. Freeman, and J. Stutz. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, pages 607–611, Saint Paul, Minnesota, 1988.

[MB88] G.I. McLachlan and K. Basford. *Mixture models: inference and applications to clustering.* Marcel Dekker. New York, 1988.

[MR93] K.L. Mengersen and C.P. Robert. Testing for mixtures via entropy distance and Gibbs sampling. Technical Report 9240, Document de travail, Crest, Insee. Paris, 1993.

[MW92] J.S. Marron and M.P. Wand. Exact Mean Integrated Squared Error. *Annals of Statistics*, 20:712–736, 1992.

[OBW96] J.J. Oliver, R.A. Baxter, and C.S. Wallace. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference*

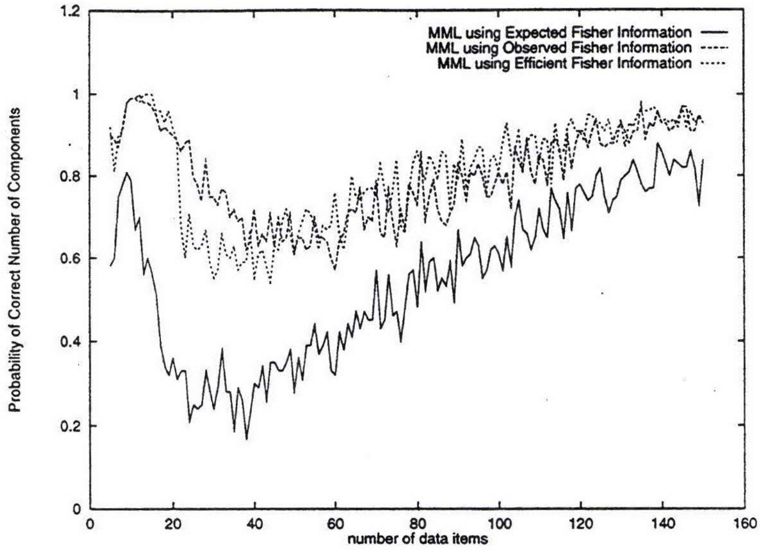

Figure 3: Use of Observed and Efficient Fisher Information approximations improves MML performance with separation = 3

*(ICML 96)*, pages 364–372. Morgan Kaufmann Publishers, 1996. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[RG96] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. Mathematics Research Report S-96-01, University of Bristol, 1996.

[Rob96] C. Robert. *Mixtures of distributions:inference and estimation*, chapter 24. Chapman and Hall, London, 1996.

[UN96] M.A. Upal and E.M. Neufeld. Comparison of Unsupervised Classifiers. In *Proceedings of the ISIS Information, Statistics and Induction in Science*, pages 342–353, Singapore, 20-23 August 1996. World Scientific.

[Upa95] M. A. Upal. Montel carlo comparison of non-hierarchical unsupervised classifiers, 1995.

[Wal90] C.S. Wallace. Classification by minimum-message-length inference. In S.G. Akl et al., editors, *Advances in Computing and Information- ICCI 1990*, pages 72–81, Niagara Falls, 1990.

[WB68] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11(2):195–209, 1968.

[WD94] C.S. Wallace and D.L. Dowe. Intrinsic classification by MML - the Snob program. In *Proceedings of the 7th Australian Joint Conference on Artificial Intelligence*, pages 37–44, Singapore, 1994. World Scientific.

[WF87] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *J. R. Statist. Soc B*, 49(3):240–265, 1987.