Draft 5.1, February 1998

Bayesian Approaches to Gaussian Mixture Modelling

Stephen J. Roberts, Dirk Husmeier, Iead Rezek & William Penny

Department of Electrical & Electronic Engineering
Imperial College of Science, Technology & Medicine
London, UK
{s.j.roberts,d.husmeier,i.rezek,w.penny}@ic.ac.uk

Abstract

A Bayesian-based methodology is presented which automatically penalises over-complex models being fitted to unknown data. We show that, with a Gaussian mixture model, the approach is able to select an 'optimal' number of components in the model and so partition data sets. The performance of the Bayesian method is compared to other methods of optimal model selection and found to give good results. The methods are tested on synthetic and real data sets.

Introduction

Scientific disciplines generate data. In the attempt to understand the patterns present in such data sets methods which perform some form of *unsupervised* partitioning or modelling are particularly useful. Such an approach is only of use, however, if it offers a *less complex* representation of the data than the data set itself. This introduces an apparent conflict, however, as any model *improves* its fit to the data monotonically with increases in its complexity (the number of model parameters) – a model as complex as the data will 'fit' perfectly, for example. This conflict may be solved in a particularly elegant manner if we adopt a Bayesian paradigm. The Bayesian approach may be crudely regarded as estimating the uncertainty of the model as a whole, given the data – the model 'fit' – and secondly the uncertainty in the estimated parameters themselves. Whilst the first measure decreases with the number of parameters, the second (often referred to as the 'Ockham factor' after the 13th century philosopher) *increases* as more parameters are estimated using a finite data set. Bayesian modelling, on a superficial level therefore, involves finding a 'balance' between these two measures. It is noted that this paper deals with the issue of estimating the probable number of component clusters, rather than concentrating on the popular issue of of finding an appropriate model for the data probability density function itself. Methods for the latter may be found in [9, 8, 25, 27, 24], for example.

Theory

Consider a finite data set $\mathcal{X} = \{\mathbf{x}^1, ..., \mathbf{x}^N\} \subset \Re^d$. We further consider fitting a K-component model, M_K whose free parameters are denoted by the vector $\boldsymbol{\theta}_K \in \Re^{N_p}$. The *evidence* of the data set may be obtained from the *marginal* integral:

$$P(\mathcal{X} \mid M_K) = \int_{\boldsymbol{\theta}_K} P(\mathcal{X} \mid M_K; \boldsymbol{\theta}_K) P(\boldsymbol{\theta}_K \mid M_K) d\boldsymbol{\theta}_K \tag{1}$$

For notational convenience, we will derive all results assuming a K-component Gaussian Mixture Model (GMM) and drop the explicit conditions on M_K in the above equation, thus Equation (1) is rewritten as

$$P(\mathcal{X}) = \int_{\boldsymbol{\theta}} P(\mathcal{X} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

We denote $L(\mathcal{X} \mid \boldsymbol{\theta})$ as the *log-likelihood* of the data given the parameters and the model i.e.

$$L(\mathcal{X} \mid \boldsymbol{\theta}) = \ln P(\mathcal{X} \mid \boldsymbol{\theta}) \tag{3}$$

and rewrite Equation (2) in the form

$$P(\mathcal{X}) = \int_{\boldsymbol{\theta}} \exp\{-E(\boldsymbol{\theta})\} \tag{4}$$

where the energy function, $E(\boldsymbol{\theta})$, is defined as

$$E(\boldsymbol{\theta}) = -L(\mathcal{X} \mid \boldsymbol{\theta}) - \ln P(\boldsymbol{\theta}) \tag{5}$$

If we are to avoid the computational expense of numerically integrating a high-dimensional parameter space we may make the assumption that $E(\boldsymbol{\theta})$ has a local quadratic form (the posterior parameter distribution is Gaussian and sharply peaked) around a *most-probable* state, represented via the most-probable parameter set, $\hat{\boldsymbol{\theta}}$. Expanding $E(\boldsymbol{\theta})$ as a Taylor series to second-order terms gives

$$E(\boldsymbol{\theta}) = E(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \left.\frac{\partial E(\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}}\right|_{\hat{\boldsymbol{\theta}}} + \frac{1}{2!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial^2 E(\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}^2}\right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \mathcal{R} \tag{6}$$

where $\mathcal{R}$ is the remainder of the series from third-order upwards which we will neglect in the following analysis. Setting

$$\mathbf{V}^{-1} = \left.\frac{\partial^2 E(\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}^2}\right|_{\hat{\boldsymbol{\theta}}}$$

(so that $\mathbf{V}$ is the inverse of the Hessian matrix of $E(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$) and noting that the first-order term of Equation (6) is equal to zero, we obtain the following from Equations (4,6)

$$P(\mathcal{X}) = \exp\{-E(\hat{\boldsymbol{\theta}})\} \int_{\boldsymbol{\theta}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{V}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \tag{7}$$

which is of standard form and equates to

$$P(\mathcal{X}) = (2\pi)^{N_p/2} |\mathbf{V}|^{1/2} \exp\{-E(\hat{\boldsymbol{\theta}})\} \tag{8}$$

where N_p is the number of parameters in the model. Taking logarithms of Equation (8) and combining it with Equation (5) gives

$$\ln P(\mathcal{X}) = L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}) + \ln P(\hat{\boldsymbol{\theta}}) + \frac{N_p}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathbf{V}| \tag{9}$$

Ripley [21] considers a normal distribution for the prior parameter expression, but we consider this to be inappropriate for a Gaussian mixture model in which the covariance elements are free parameters. Other authors have advocated the use of non-informative reference priors which we consider a more realistic reflection of our prior beliefs in parameter values. If we are to choose a reference prior such that it remains invariant to scalings of the parameters then a Jeffreys' [12] prior is favoured [13] which takes the form:

$$P(\hat{\boldsymbol{\theta}}) = \sqrt{|\mathbf{F}(\hat{\boldsymbol{\theta}})|} \tag{10}$$

where $\mathbf{F}$ is the Fisher information matrix. For a single, univariate, Gaussian in which both the mean and variance are unknown and are assumed to form a *dependent* joint prior distribution, the Jeffreys' prior is $P(\mu, \sigma) \propto \sigma^{-3/2}$ [13]. It is argued (also in [13]) that such dependence between parameters is not a viable reflection of prior parameter distributions and that a reference prior should be sought in which the mean and variance are assumed to be initially *independent*. In this case we would choose a prior distribution of the form $P(\mu, \sigma) = P(\mu)P(\sigma) \propto \sigma^{-2}$. Independent flat (improper) priors are advocated by Oliver *et al.* [18] and preferred over the Jeffreys' case by Lee [13]. Taking a set of such improper priors leads to a simplified analytic solution, as the prior becomes independent of the parameter values. We choose, for this reason, to use such flat priors and consider each component of the mean vector for any of the K Gaussians to have a flat distribution in the range $(-\alpha\sigma_{pop}, \alpha\sigma_{pop})$ where we define σ_{pop} to be the largest element along the diagonal of the covariance matrix of $\mathcal{X}$.

Similarly the diagonal covariance elements (σ_{ii}) of each Gaussian are taken to have a flat prior in the range $(0, \beta\sigma_{pop})$. This gives rise to a prior density for the internal parameters of K Gaussians in a d-dimensional space of

$$P(\hat{\boldsymbol{\theta}}_{internal}) = \frac{1}{(2\alpha\beta\sigma_{pop}^2)^{Kd}} \tag{11}$$

We note that this is the same prior distribution taken by Oliver *et al.* [18]. We furthermore follow the suggestion made in [18] to allow the prior distribution of the external model parameters (the set of model priors, or mixing fractions) to be of simple Dirichlet form such that:

$$P(\{P(k)\}) = (K-1)! \tag{12}$$

Combining Equations (11) and (12) and taking natural logarithms gives

$$\ln P(\hat{\boldsymbol{\theta}}) = -Kd \ln\left(2\alpha\beta\sigma_{pop}^2\right) + \ln(K-1)! \tag{13}$$

In the case discussed by Oliver *et al.* $\alpha = 1, \beta = 1$. Note, however, that these parameters will feature in the final equations governing the data evidence. As Fitzgerald [16] rightly points out, these scale parameters are essentially arbitrary (within reason, naturally - we would not expect Gaussian centre locations outside the convex hull of $\mathcal{X}$, for example) and thus our final equation for the evidence of $\mathcal{X}$ has an arbitrary component. By taking flat priors, however, such that any parameter set (within the bounds of the priors) is equally supported and the Hessian of the error function of Equation (5) is equal to the Hessian on the negative log-likelihood itself, i.e. $\mathbf{V}^{-1} = \mathbf{H}$ and the data evidence equation thus takes the form

$$\ln P(\mathcal{X}) = L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}) + f_{post}(\mathbf{H}) + f_{prior}(\hat{\boldsymbol{\theta}}, \mathcal{X}, \alpha, \beta) \tag{14}$$

where the two functions represent functions of the posterior curvature and prior parameter distributions respectively. In this form the parameters α, β, which govern the scale of the prior distribution, may be seen as *hyper-parameters* [14, 15] in a second level of Bayesian inference and as such they may, in principle, be inferred from the data itself, rather than being set arbitrarily. Such a scheme has been presented for the fitting of regularisation hyper-parameters in flexible models [14, 15] but adds considerably to the computational complexity of the modelling process. We would argue against such a course, despite its initial attraction of 'letting the data speak for itself' by noting that the dependence of the evidence on the prior hyper-parameters α and β is not strong and that a reasonable choice for these (such as unity) truly reflects our prior beliefs concerning the model. We choose to adopt this latter method, and allow $\alpha = \beta = 1$ in our analysis. We will thus write Equation (9) as

$$\ln P(\mathcal{X}) = L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}) - Kd \ln\left(2\alpha\beta\sigma_{pop}^2\right) + \ln(K-1)! + \frac{N_p}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{H}| \tag{15}$$

We note that this has the intuitive structure of a 'goodness-of-fit' term, $L(\mathcal{X} \mid \hat{\boldsymbol{\theta}})$, and a set of parameter uncertainty terms. The question remains, however, as to how the expressions of Equation (15) may be estimated, and this is dealt with in the next section.

Parameter estimation

All the quantities of Equation (15) are, in principle, uniquely defined given $(\mathcal{X}; \hat{\boldsymbol{\theta}})$. We seek, therefore, the parameter set, $\hat{\boldsymbol{\theta}}$, which is the most-probable (the maximum-likelihood set). We thus seek $\boldsymbol{\theta}$ which maximises $L(\mathcal{X} \mid \boldsymbol{\theta}) = \ln P(\mathcal{X} \mid \boldsymbol{\theta})$ or minimises an energy function $F = -L(\mathcal{X} \mid \boldsymbol{\theta})$. For a GMM with $k = 1 \ldots K$ components we may write

$$F = -\sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} P(\mathbf{x}^n \mid k)P(k)\right\} \tag{16}$$

The parameter vector $\boldsymbol{\theta}$ contains three quantities for each Gaussian; the centroid, $\boldsymbol{\mu}_k$, the covariance matrix $\boldsymbol{\Sigma}_k$ and prior (mixing fraction in the GMM), $P(k)$. Using standard techniques (see [4], for example) the following estimators are obtained from Equation (16)

$$\begin{aligned}
\hat{P}(k) &= \frac{1}{N}\sum_{n=1}^{N}\hat{P}(k \mid \mathbf{x}^n) \qquad (17)\\
\hat{\boldsymbol{\mu}}_k &= \frac{\sum_{n=1}^{N}\hat{P}(k \mid \mathbf{x}^n)\mathbf{x}^n}{\sum_{n=1}^{N}\hat{P}(k \mid \mathbf{x}^n)}\\
\hat{\boldsymbol{\Sigma}}_k &= \frac{\sum_{n=1}^{N}\hat{P}(k \mid \mathbf{x}^n)(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}^n - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{n=1}^{N}\hat{P}(k \mid \mathbf{x}^n)}
\end{aligned}$$

These represent a set of coupled, non-linear equations; they may, however, be used to form a successive re-estimation algorithm - most notably the expectation-maximisation (EM) algorithm [6] which gives estimates of the free parameters of the mixture model given $\mathcal{X}$. These estimates form the parameter vector $\hat{\boldsymbol{\theta}}$ and hence provide plug-in estimators for $L(\mathcal{X} \mid \hat{\boldsymbol{\theta}})$.

We must still, however, form an estimate for $\ln|\mathbf{H}|$. We recall that $\mathbf{H}$ is the Hessian matrix of $-L(\mathcal{X} \mid \hat{\boldsymbol{\theta}})$. We follow a similar approach to that taken in [18] in that a simplification of the Hessian evaluation is utilised. This simplification is argued for in [10] and assumes that the internal free parameters of one Gaussian component (the mean and covariance elements) are independent from those of all other components hence the determinant of the Hessian of the complete model will be approximated as the product of Hessians from each component. The coupling of the parameters between components is, from Equations (17), via the set of *posterior probabilities*. If we are to make the Hessian *analytically* tractable (so as to avoid numerical integration approaches such as applied in [20]) then adoption of this simplification is tantamount to treating this set of posterior probabilities as fixed quantities. This approximation may be partly justified by noting that as we assume that the posterior parameter distribution is sharply peaked around the maximum-likelihood solution (implicit in the local quadratic or Laplace approximation) so we consider only small changes in the parameters, to which the set of posterior probabilities are largely insensitive. Husmeier [11] considers an expansion of the Hessian for the GMM in terms of the approximate Hessian (obtained by following the approximation taken here) and a series of perturbation terms. The latter are found (empirically) to be negligable. He notes, furthermore, that neglecting these terms is equivalent to taking an inner-product approximation to the Hessian matrix (as discussed, for example, in [4, 21]).

We will introduce the theory first in the case for which each Gaussian has a diagonal covariance matrix with elements $\sigma_{k,i}^2$ ($k = 1...K$ and $i = 1...d$). From Equation (16)

$$F = -\sum_{n=1}^{N}\ln P(\mathbf{x}^n)$$

where

$$P(\mathbf{x}^n) = \sum_{k=1}^{K}P(\mathbf{x}^n \mid k)P(k)$$

Hence, considering the partial derivative of F with respect to some parameter π

$$\frac{\partial F}{\partial \pi} = -\sum_{n=1}^{N}\frac{1}{P(\mathbf{x}^n)}\frac{\partial}{\partial \pi}\left\{\sum_{k=1}^{K}P(\mathbf{x}^n \mid k)P(k)\right\} \qquad (18)$$

in which, for Gaussians with diagonal covariances

$$P(\mathbf{x}^n \mid k) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_{k,i}} \exp\left\{ -\frac{(x_i^n - \mu_{k,i})^2}{2\sigma_{k,i}^2} \right\} \tag{19}$$

and from Bayes' theorem we have

$$P(k \mid \mathbf{x}^n) = \frac{P(\mathbf{x}^n \mid k)P(k)}{P(\mathbf{x}^n)}$$

Under the assumption that the components decouple we may hence rewrite $|\mathbf{H}|$ as [18]

$$|\mathbf{H}| = |\mathbf{H}_p| \times |\mathbf{H}_g| \tag{20}$$

where $\mathbf{H}_p$ is the Hessian matrix with respect to the *external* parameters of the GMM (the set of mixture priors), and $\mathbf{H}_g$ the Hessian with respect to the *internal* parameters of the GMM (the set of centroids and covariances). We start our analysis by considering the Hessian components with respect to the set of mixture priors. From Equation (18) we obtain, noting that there are only $K-1$ free parameters, so the differentiation is performed subject to the constraint $\sum_k P(k) = 1$, i.e. $P(K) = 1 - \sum_{k=1}^{K-1} P(k)$,

$$\frac{\partial F}{\partial P(k)} = -\sum_{n=1}^{N} \frac{1}{P(\mathbf{x}^n)} \left(P(\mathbf{x}^n \mid k) - P(\mathbf{x}^n \mid K) \right) \tag{21}$$

where k indexes from $1..(K-1)$. Taking differentials again with respect to $P(k)$, as we will approximate the determinant of the Hessian from its diagonal components only (see [11]), gives

$$\left. \frac{\partial^2 F}{\partial P(k)^2} \right|_{\hat{\boldsymbol{\theta}}} = \sum_{n=1}^{N} \left(\frac{\hat{P}(k \mid \mathbf{x}^n)}{\hat{P}(k)} - \frac{\hat{P}(K \mid \mathbf{x}^n)}{\hat{P}(K)} \right)^2$$

hence our approximation to the determinant of $\mathbf{H}_p$ is

$$|\mathbf{H}_p| = \prod_{k=1}^{K-1} \sum_{n=1}^{N} \left(\frac{\hat{P}(k \mid \mathbf{x}^n)}{\hat{P}(k)} - \frac{\hat{P}(K \mid \mathbf{x}^n)}{\hat{P}(K)} \right)^2 \tag{22}$$

Note that $K = 1$ is dealt with as a special case as there is no degree of freedom in the mixing coefficient and so the corresponding Hessian does not exist.

Considering now the Hessian component with respect to $\mu_{k,l}$, the l–th component of the centroid vector of the k–th Gaussian in the mixture,

$$\frac{\partial F}{\partial \mu_{k,l}} = -\sum_{n=1}^{N} \frac{(x_l^n - \mu_{k,l})}{\sigma_{k,l}^2} \hat{P}(k \mid \mathbf{x}^n) \tag{23}$$

This expression, under the assumption of separable components, gives rise to cross terms which are zero upon differentiation with respect to all parameters except $\mu_{k,l}$ and $\sigma_{k,l}$. The latter term, evaluated at $\hat{\boldsymbol{\theta}}$, equates to

$$\left. \frac{\partial^2 F}{\partial \mu_{k,l} \partial \sigma_{k,l}} \right|_{\hat{\boldsymbol{\theta}}} = \frac{2}{\hat{\sigma}_{k,l}^3} \sum_{n=1}^{N} (x_l^n - \hat{\mu}_{k,l}) \hat{P}(\mathbf{x}^n \mid k)$$

and this term is zero upon substitution from the second of Equations (17). Differentiating Equation (23) once more with respect to $\mu_{k,l}$ gives

$$\left. \frac{\partial^2 F}{\partial \mu_{k,l}^2} \right|_{\hat{\boldsymbol{\theta}}} = \frac{1}{\hat{\sigma}_{k,l}^2} \sum_{n=1}^{N} \hat{P}(k \mid \mathbf{x}^n) = \frac{N\hat{P}(k)}{\hat{\sigma_{k,l}}} \tag{24}$$

In the evaluation of the Hessian component with respect to $\sigma_{k,l}$ we obtain

$$\frac{\partial F}{\partial \sigma_{k,l}} = -\sum_{n=1}^{N} \left[-\frac{1}{\sigma_{k,l}} + \frac{(x_l^n - \mu_{k,l})^2}{\sigma_{k,l}^3} \right] P(k \mid \mathbf{x}^n) \tag{25}$$

Note that all terms of the form

$$\frac{\partial^2 F}{\partial \sigma_{k,l} \partial \pi} = 0$$

for $\pi \neq \sigma_{k,l}$. This means that Equation (20) may be rewritten as

$$|\mathbf{H}| = |\mathbf{H}_p| \times \prod_{k=1}^{K} |\mathbf{H}_{g,k}| \tag{26}$$

where $\mathbf{H}_{g,k}$ is the Hessian of F with respect to the internal parameters of the k-th Gaussian in the mixture. Differentiating Equation (25) with respect to $\sigma_{k,l}$ gives

$$\left. \frac{\partial^2 F}{\partial \sigma_{k,l}^2} \right|_{\hat{\boldsymbol{\theta}}} = \sum_{n=1}^{N} \left[-\frac{1}{\hat{\sigma}_{k,l}^2} + \frac{3(x_l^n - \hat{\mu}_{k,l})^2}{\hat{\sigma}_{k,l}^4} \right] \hat{P}(k \mid \mathbf{x}^n)$$

Substitution in the above equation of the maximum-likelihood estimator for $\hat{\sigma}_{k,l}^2$ from Equations (17), namely,

$$\hat{\sigma}_{k,l}^2 = \frac{\sum_{n=1}^{N} \hat{P}(k \mid \mathbf{x}^n)(x_l^n - \hat{\mu}_{k,l})^2}{\sum_{n=1}^{N} \hat{P}(k \mid \mathbf{x}^n)}$$

and that for $\hat{P}(k)$ gives

$$\left. \frac{\partial^2 F}{\partial \sigma_{k,l}^2} \right|_{\hat{\boldsymbol{\theta}}} = \frac{2N\hat{P}(k)}{\hat{\sigma}_{k,l}^2} \tag{27}$$

From Equations (24,27) we write

$$|\mathbf{H}_{g,k}| = \left[\sqrt{2} N \hat{P}(k) \right]^{2d} \prod_{i=1}^{d} \frac{1}{\hat{\sigma}_{k,i}^4} \tag{28}$$

Combining Equations (22,26 & 28) gives

$$\ln |\mathbf{H}| = \sum_{k=1}^{K-1} \ln \sum_{n=1}^{N} \left(\frac{\hat{P}(k \mid \mathbf{x}^n)}{\hat{P}(k)} - \frac{\hat{P}(K \mid \mathbf{x}^n)}{\hat{P}(K)} \right)^2 + 2d \sum_{k=1}^{K} \ln \left[\sqrt{2} N \hat{P}(k) \right] - 2 \sum_{k=1}^{K} \sum_{i=1}^{d} \ln \hat{\sigma}_{k,i}^2 \tag{29}$$

This results in an estimate for the Hessian term in Equation (15).

Full covariance matrices

It is desirable, however, to deal with mixtures of Gaussians with *full* covariance matrices. As a covariance matrix, $\boldsymbol{\Sigma}$, is square-symmetric so we may write

$$\boldsymbol{\Sigma} = \mathbf{M} \boldsymbol{\Lambda} \mathbf{M}^T$$

where $\boldsymbol{\Lambda}$ is a *diagonal* matrix of the eigenvalues, λ_i, of $\boldsymbol{\Sigma}$ and $\mathbf{M}$ is a matrix of normalised eigenvectors. Under the transform

$$\mathbf{x} \mapsto \mathbf{M}\mathbf{x}$$

the previous theory for diagonal matrices may be re-applied with the sole adjustment that

$$\hat{\sigma}_{k,l}^2 \mapsto \lambda_{k,l}$$

the l–th eigenvalue of $\hat{\mathbf{\Sigma}}_k$. Equation (29) is hence re-written as

$$\ln|\mathbf{H}| = \sum_{k=1}^{K-1} \ln \sum_{n=1}^{N} \left(\frac{\hat{P}(k \mid \mathbf{x}^n)}{\hat{P}(k)} - \frac{\hat{P}(K \mid \mathbf{x}^n)}{\hat{P}(K)} \right)^2 + 2d \sum_{k=1}^{K} \ln \left[\sqrt{2} N \hat{P}(k) \right] - 2 \sum_{k=1}^{K} \sum_{i=1}^{d} \ln \lambda_{k,i} \tag{30}$$

The full evidence expression

Combining Equations (15) and (29/30) leads to the following estimate for the evidence of $\mathcal{X}$

$$\begin{aligned} \ln P(\mathcal{X}) \;=\;& L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}) - Kd \ln\left(2\alpha\beta\sigma_{pop}^2\right) + \ln(K-1)! + \frac{N_p}{2}\ln(2\pi) \qquad (31) \\ & - \frac{1}{2} \left(\sum_{k=1}^{K-1} \ln \sum_{n=1}^{N} \left(\frac{\hat{P}(k \mid \mathbf{x}^n)}{\hat{P}(k)} - \frac{\hat{P}(K \mid \mathbf{x}^n)}{\hat{P}(K)} \right)^2 + 2d \sum_{k=1}^{K} \ln \left[\sqrt{2} N \hat{P}(k) \right] - 2 \sum_{k=1}^{K} \sum_{i=1}^{d} \ln \lambda_{k,i} \right) \end{aligned}$$

We use this equation, with $\alpha = \beta = 1$, to obtain all the Bayesian method results presented in the paper. It is noted that, for computational simplicity, this same Equation may also be applied to mixtures with diagonal covariances, as $\lambda_{k,l} = \sigma_{k,l}^2$ in this case.

Results

Comparison with other methods

We compare the results from the Bayesian approach with those obtained using the same model parameters (from the EM algorithm) using other model-order selection criteria / techniques. The methods we compare are

1. Fuzzy hyper-volume (FHV) : this was detailed in [7] and looks at models with lowest total volume, defined via

$$V(K) = \sum_{k=1}^{K} \sqrt{|\mathbf{\Sigma}_k|} \tag{32}$$

2. Evidence density : this measure is argued for in [23] and uses the FHV measure of [7] to penalise the data likelihood measure. It is defined as

$$\varrho(K) = \frac{L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}_K)}{V(K)} \tag{33}$$

3. Minimum description length (MDL) : derived by Rissanen [22] from an information-theoretic perspective, MDL is defined via

$$MDL(K) = -L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}_K) + \frac{1}{2} N_p(K) \ln N \tag{34}$$

   where, as before, $N_p(K)$ is the number of parameters in the K Gaussian model.

4. Partition coefficient (PC) : this is defined as [3]

$$PC(K) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} P(k \mid \mathbf{x}^n)^2 \tag{35}$$

5. Minimum message length (MML) : derived by Wallace and Freeman [26] and tested and developed extensively by Oliver *et al.* [18, 17, 2]. The MML expression used is as given in [18]:

$$\begin{aligned} MML(K) &= K \ln(2\sigma_{pop}^2) - \ln(K-1)! + \frac{N_p}{2} \ln \kappa(N_p) - \ln K! \\ &+ \sum_{k=1}^{K} \ln \frac{\sqrt{2} N_k}{\sigma_k^2} + \frac{1}{2} \ln N - \frac{1}{2} \sum_{k=1}^{K} \ln P(k) \\ &- L(\mathcal{X} \mid \hat{\boldsymbol{\theta}}) + \frac{N_p}{2} \end{aligned} \tag{36}$$

Note that this expression is similar in parts to that developed under the Bayesian scheme. The constant, $\kappa(N_p)$, is the optimal lattice quantising constant in an N_p-dimensional space. These are set using the table of hypothesised lattice constants given in [5, Table 2.3]. It is noted that MML is developed for *diagonal* covariance matrices in [18]. We may, however, make the same transformation of variables as in the Bayesian case, and replace the set of σ_k^2 in the above Equation by the eigenvalues of a full covariance matrix. We find that this removes this obstacle. One further problem with this method is that optimal lattice constants are not known in some dimensions (e.g. $D = 9$) and are poorly specified in others. We utilise the expected *lower bounds* to the constants as given in [5] to provide some consistency. This lattice set is not well-defined for $D > 24$ are we use a linear extrapolant to provide values in higher dimensions. Similarly we use a linear interpolant to provide constants for the 'missing' values of lower dimension. It is not clear from the publications of Oliver *et al.* what scheme was used to overcome this problem and in the absence of any further information our simple linear regression scheme seems viable.

Synthetic data 1

We first investigate the properties of these model selection methods on a simple two-dimensional[1] toy problem, inspired in part by the tests of [18]. Data was generated from four Gaussians, each with a common (isotropic) width, σ_{gen}, and means given by

$$\begin{aligned} \boldsymbol{\mu}_1 &= (0,0)^T \\ \boldsymbol{\mu}_2 &= (2, \sqrt{12})^T \\ \boldsymbol{\mu}_3 &= (4,0)^T \\ \boldsymbol{\mu}_4 &= (-2, -\sqrt{12})^T \end{aligned} \tag{37}$$

a total of 30 samples per Gaussian were taken, making $N = 120$. The width was set to $\sigma_{gen} = \{0.66, 1.0, 1.2\}$. The resultant data sets are presented in Figure (1)[2].

Results are presented over ten runs of the EM algorithm, each with a different random number seed. Figures (2,3 & 4) show the mean and ± 1 SD for $K = 2 \ldots 7$ for:

- Bayesian method : we plot $-\ln P(\mathcal{X})$, hence model support is higher for lower values of this quantity.
- Fuzzy hyper-volume (FHV).
- Evidence density.
- Minimum description length (MDL).

[1] There is, of course, no requirement that the problem be of such low dimensionality and we choose $d = 2$ purely for ease of presentation (it does, however, become increasingly difficult to verify the results of any modelling when the dimensionality is high).

[2] Note that the data sets are normalised such that they have a zero mean and unity intra-component covariance - this is why the data appear re-scaled in the plots.

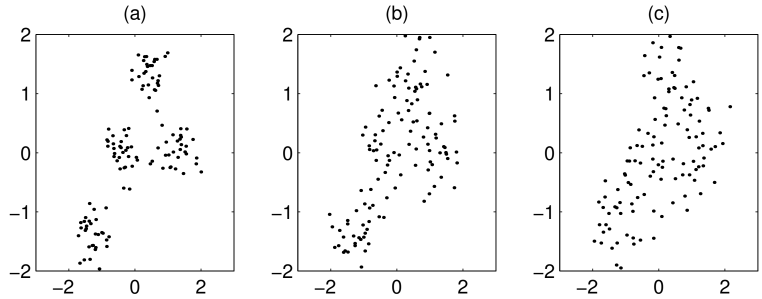

Figure 1: Synthetic data 1: (a) $\sigma_{gen} = 0.66$, (b) $\sigma_{gen} = 1.0$, (c) $\sigma_{gen} = 1.2$.

- Partition coefficient : we plot $1 - PC(K)$, hence model support is higher for lower values of this quantity. Note that $1 - PC(1) = 0$ always.
- Minimum message length (MML).

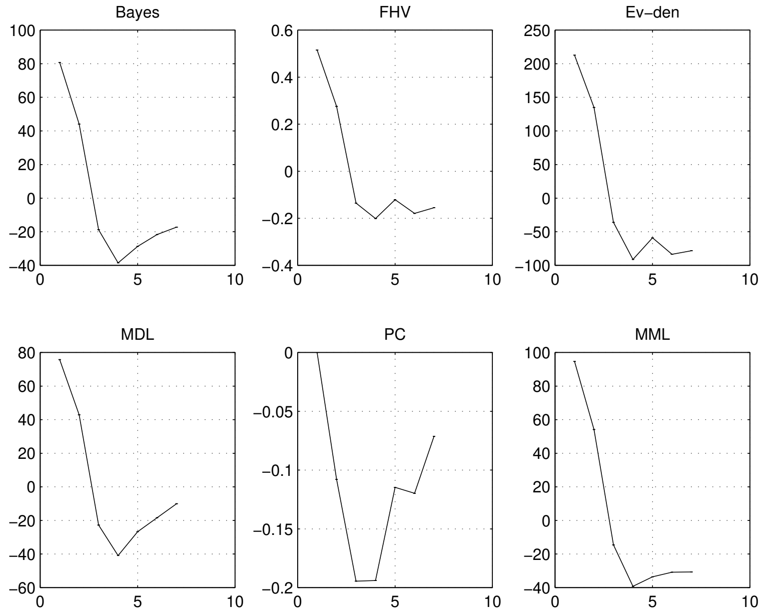

Figure 2: Model-selection functions for synthetic data set, $\sigma_{gen} = 0.66$.

We see clearly that, in the simple case of $\sigma_{gen} = 0.66$, all methods give greatest support to the 'true' model. As σ_{gen} increases, however, all methods save for the Bayesian, MDL and MML become erratic (higher variance) and make poor assessments. We see also that, as the data becomes more diffuse, both the Bayesian method and MDL give greater support to models of low complexity. MML, on the other hand, appears to support models with higher complexity under these circumstances. Note the very clear correlation between the Bayesian method and MDL (and, to a lesser extent MML). This is not a new observation - there are intimate links between the methods (see Oliver & Baxter's work [17, 2] for example).

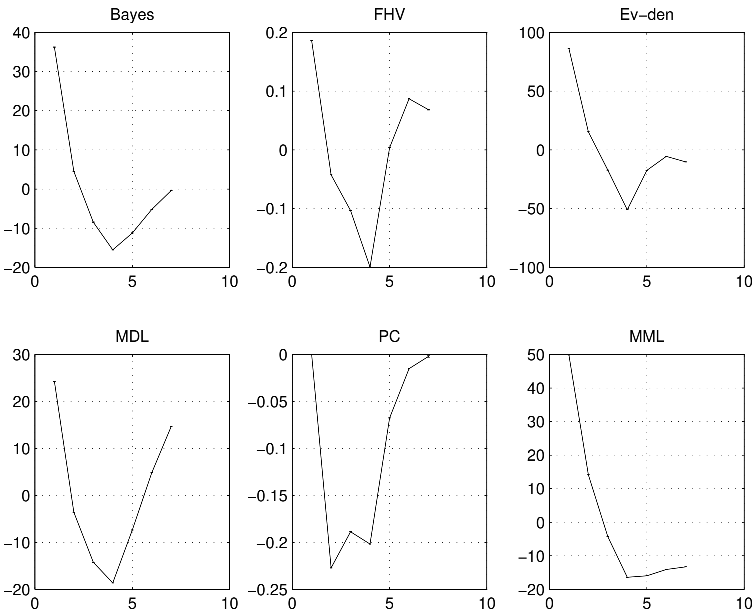

Figure 3: Model-selection functions for synthetic data set, $\sigma_{gen} = 1.0$.

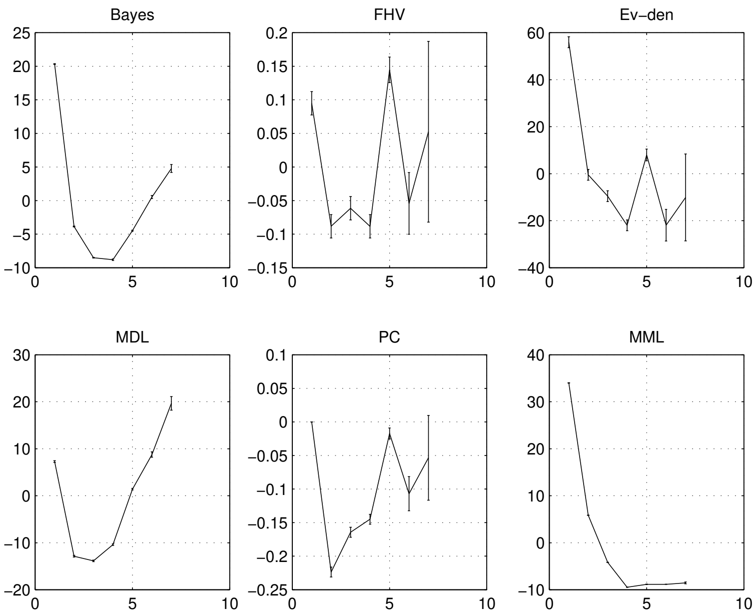

Figure 4: Model-selection functions for synthetic data set, $\sigma_{gen} = 1.2$. Note the graceful breakdown of both MDL and the Bayesian method which gradually give higher support to models of *lower* complexity as the data becomes more diffuse. It is interesting to note also that MML, under the same circumstances, gives much higher support for more complex models.

Synthetic data 2

The second toy problem we investigate takes the form, once more, of data generated from four Gaussians. In this case, however, they are paired such that each pair has a common mean, i.e. $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_4$. We set $\sigma_1 = \sigma_3 = 1$ and $\sigma_2 = \sigma_4 = 5$. We sample 1000 points from this distribution, taking 250 from each Gaussian. Once more, the entire data set is normalised to zero mean and unit variance and this is shown in Figure (5). We apply the same methodology as the previous section

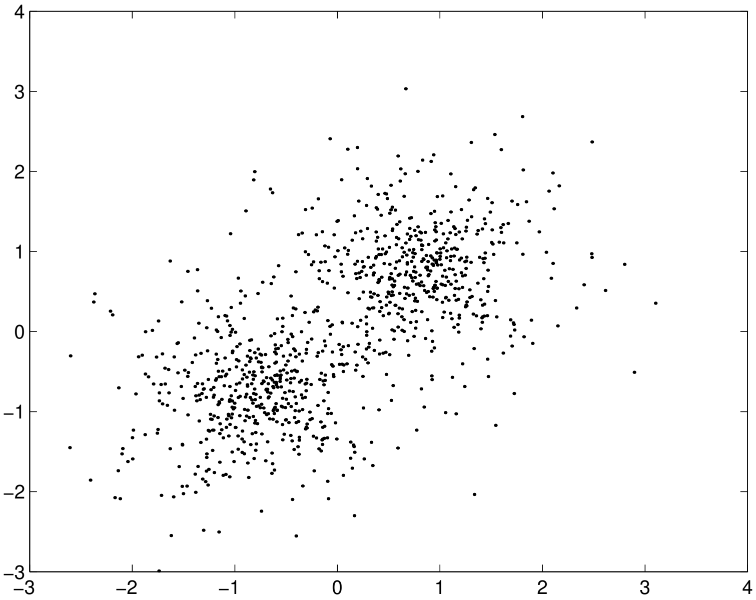

Figure 5: Synthetic data set 2: two pairs of joint-mean components (1000 data point).

and the results are shown in Figure (6). We note that MML, MDL and the Bayesian method clearly support two-cluster as well as the (true) four-cluster models, although in case of MML this support is only slightly greater than that for higher-complexity models. Other methods give the most support for two clusters, though there is a slight dip in the measures (more support) for $N_{clusters} = 4$. Note also that, due to large N, there is low variance in the plots - the ± 1 SD bars are barely visible.

Diagonal and full covariances

In the previous two examples, whilst the GMM was developed with full covariances, the data was generated from components with isotropic variance. It is clear that such simple data is not representative of many 'real-world' data sets. Figure (7) shows a set of four well-separated clusters whose covariances are full. The total number of data points is 120, (30 per cluster). Figures (8,9) respectively show model-selection estimates for models with diagonal and full covariance matrices. We see that in the diagonal case GMMs fail in general to give the true number of components (four) greatest support (although the Bayesian approach and MDL do support four clusters marginally more). With full covariances all methods clearly indicate highest support for the true component number.

Iris data

Anderson's 'iris' data set is well-known [1]. The data we analysed consisted of 50 samples for each of the three classes present in the data, *Iris Versicolor, Iris Virginica* and *Iris Setosa.* Each datum is four-dimensional and consists of measures of the plants morphology. Figure (10) shows the model support for the various methods. Note that both the Bayesian approach and MDL have most support for the true partitioning, as does the evidence density measure. MML, interestingly, has its first local minimum at this point, but favours more complex models marginally more. When we subsequently partition the data according to the $N_{clusters} = 3$ hypothesis we are in the position to compare the

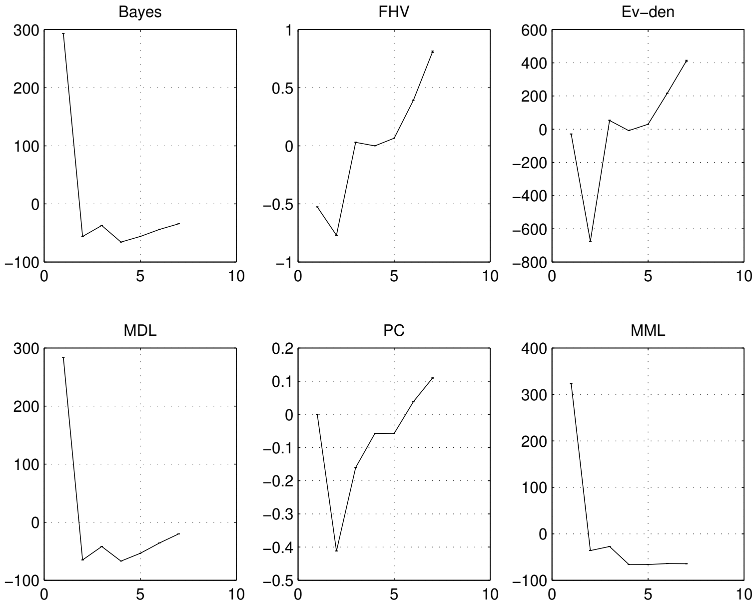

Figure 6: Synthetic data set 2: Model support results. Note that both MDL and the Bayesian method clearly support two-cluster as well as the (true) four-cluster model.

partitioning with the class labels of the data set. We find only three errors in 150 data samples, i.e. an accuracy of 98%. This is slightly better than the results presented in [7].

Physiological signals

The next data set we present results on comes from recordings of brain activity (EEG) in patients with a breathing disorder known as *obstructive sleep apnoea*. In this condition a cycling occurs between periods of breathing cessation and hyperventilation causing periodic arousals from sleep. The EEG signal was coded using features relating to its complexity over half-second time frames (details of the feature extraction methods may be found in [19]). This gave a three-dimensional data space. Figure (11) shows the model assessment measures. A partitioning of $N_{clusters} = 3$ is supported by both MDL and the Bayesian method and MML has a shallow local minimum at this point. Subsequent partitioning of the data gives rise to the traces of Figure (12). Plots (a), (b) & (c) show two-minute sections of the posterior partition probabilities and plot (d) the partition-label sequence. The periodic nature of the disorder is clear, as is the fact that label classes (a) and (c) represent system states at the two extremes of breathing and state (b) appears to represent the transition.

The next data set we present comes once more from recordings of brain activity (EEG) in a study to investigate changes of cortical activity during sleep. At the beginning of the recording the subject is awake and subsequently falls asleep ($t \approx 90$s into the recording). An external stimulus[3] is then given at roughly one minute intervals inducing arousals from sleep. The EEG signal was coded using three features relating to its complexity[4] over one-second time frames. Figure (13) shows the model support measure. A partitioning of $N_{clusters} = 3$ is supported Subsequent partitioning of the data gives rise to the traces of Figure (14). Plots (a), (b) & (c) show 4.5 minute sections of the posterior partition probabilities and plot (d) the partition-label sequence. The arousals are extremely clear as is the transition from wakefulness to sleep. It is interesting to note that the model support measures of Figure (13) are similar to those of the previous example.

[3] A small vibration under the pillow.

[4] The first two singular values of a phase-space embedding and the spectral entropy of the signal - full details of these feature extraction methods may be found in [19]

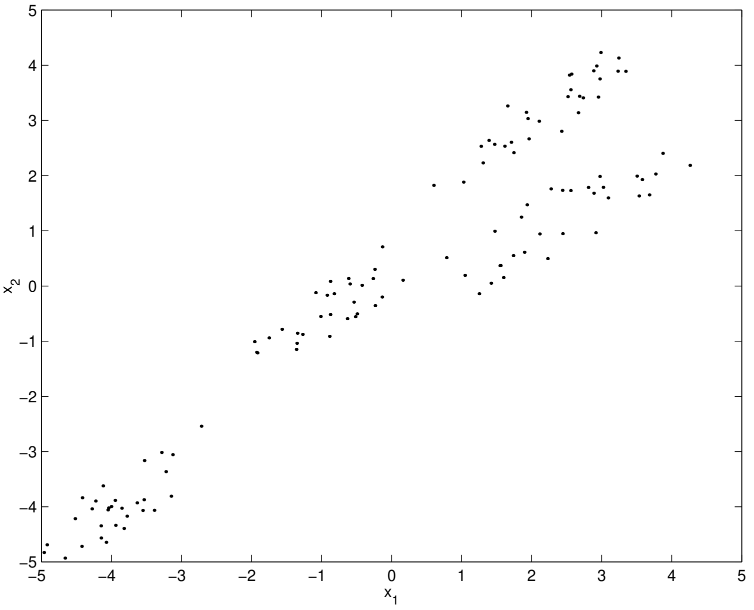

Figure 7: Data generated from normal distributions with full covariance matrices.

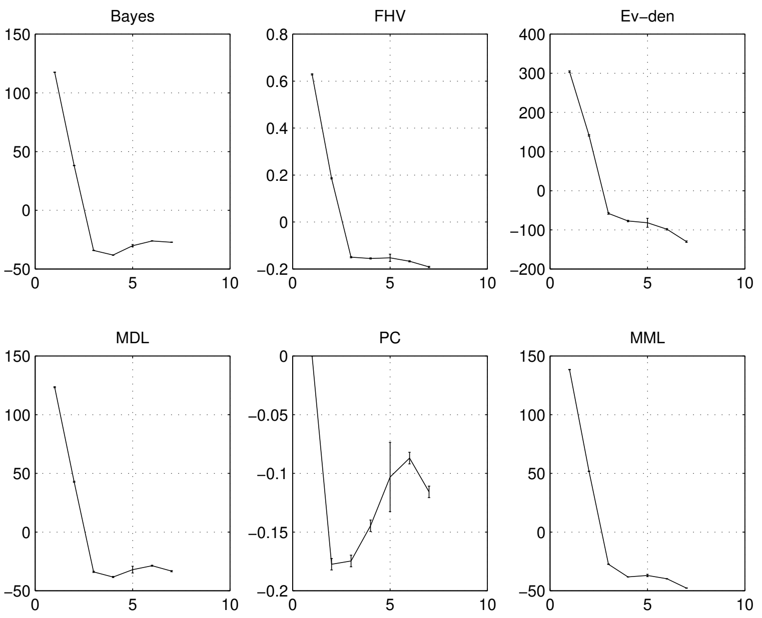

Figure 8: Model-selection functions for non-diagonal data set using diagonal-covariance models. Note the relatively inconclusive support for the true number of clusters (four).

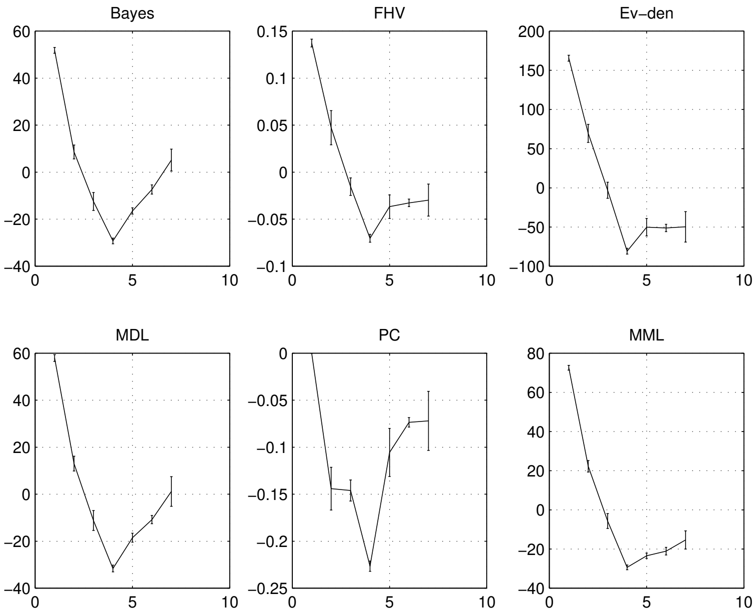

Figure 9: Model-selection functions for non-diagonal data set using full-covariance models. Model support is uniformly highest for the true cluster number (four).

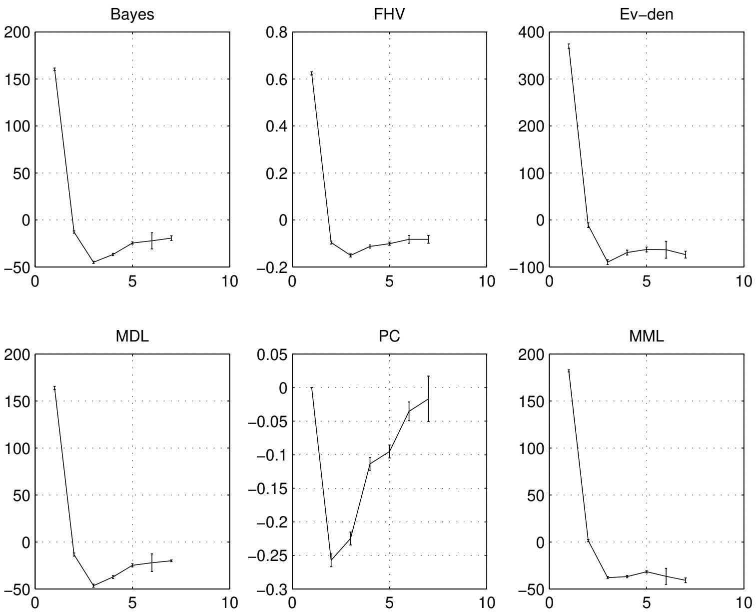

Figure 10: Anderson's *Iris* data: True partitioning is $N_{clusters} = 3$.

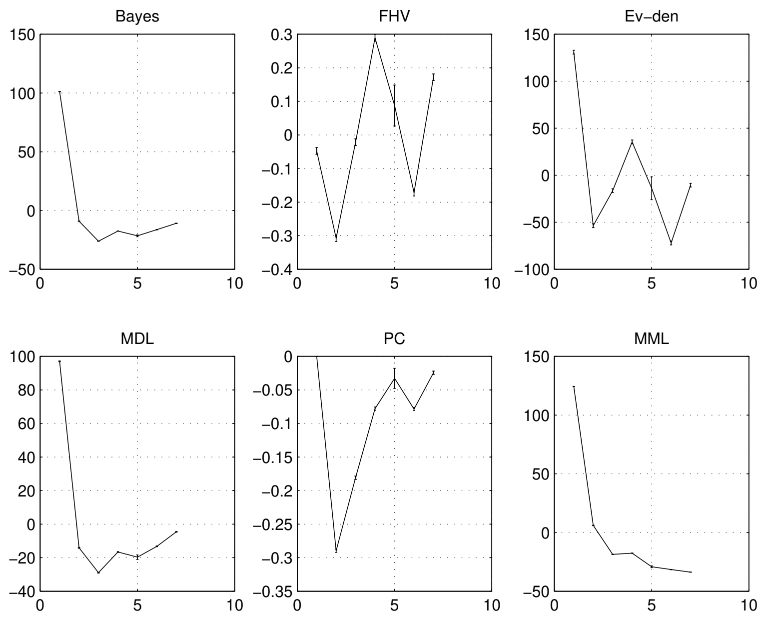

Figure 11: Obstructive sleep apnoea data: model support measures.

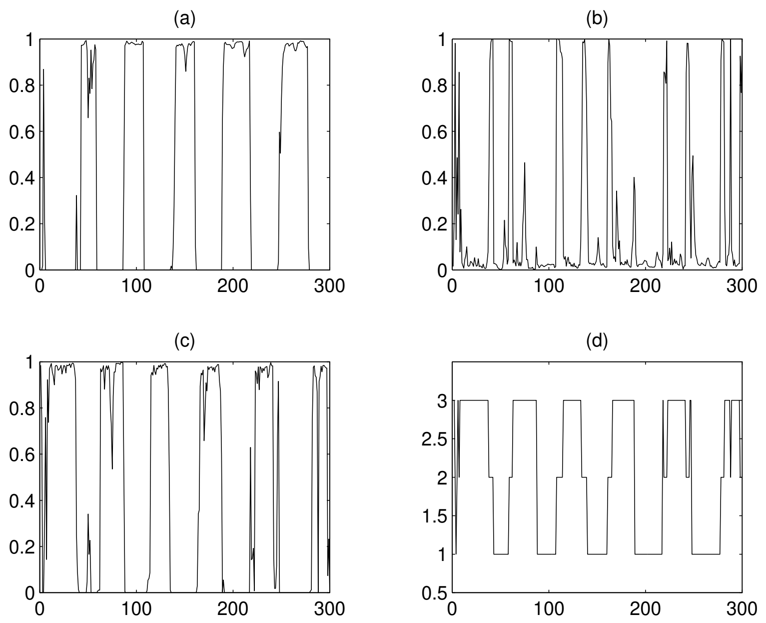

Figure 12: Obstructive sleep apnoea data: (a), (b) & (c) - posterior probabilities of $N_{cluster} = 3$ partitions. (d) - partition-label sequence.

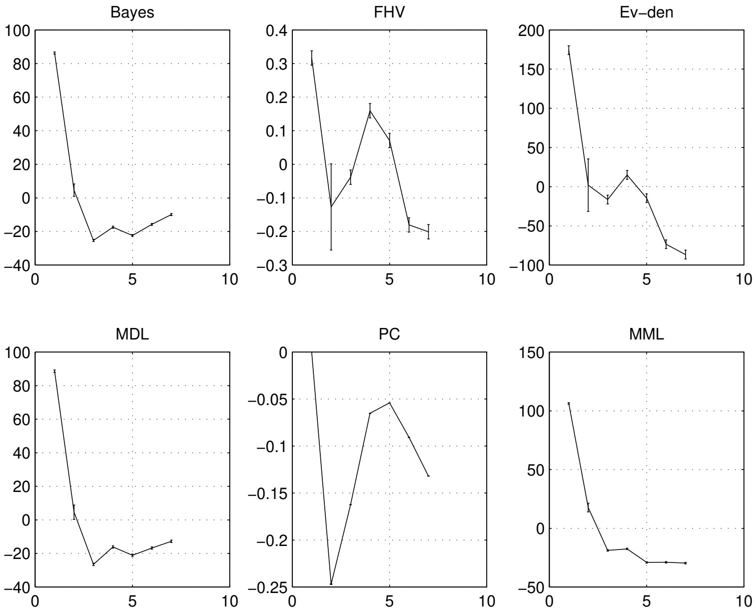

Figure 13: Model support measures for second EEG data set (arousals).

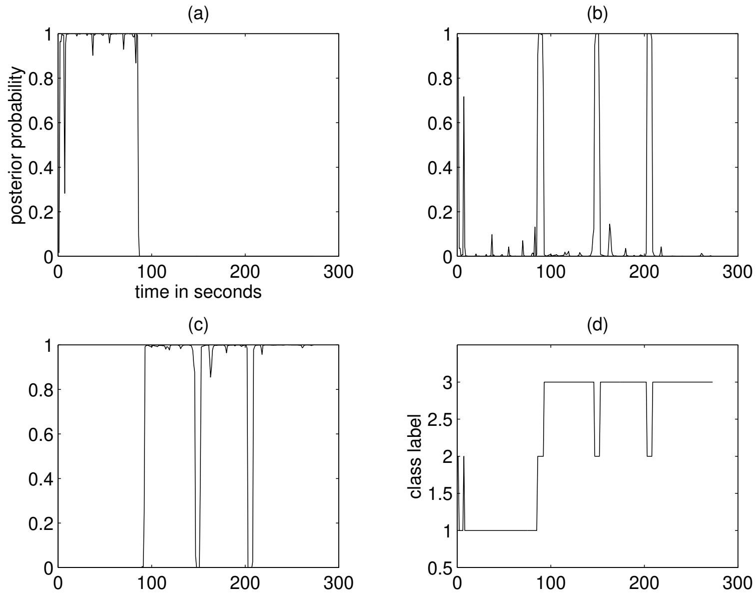

Figure 14: (a), (b) & (c) - posterior probabilities of $N_{cluster} = 3$ partitions. (d) - partition-label sequence.

Conclusions / discussion

We have presented results which clearly indicate that model selection methods based upon information theory, in the broadest sense, i.e. the Bayesian method, MDL and MML, outperform other, more heuristic methods. Not only do they give rise to reliable estimates in most cases, but their estimates remain sensible even when faced with difficult partitioning problems (i.e. few samples generated from high-width components). This is in contrast to methods such as the FHV of [7] which, although performing well on many problems, gives highly variable estimates with different local minima from the EM algorithm and does not 'die gracefully' in difficult situations. In any practical data-analysis problem, assuming diagonal covariance matrices may lead to erroneous results. All the methods presented in this paper have been adapted so as to assess model support for GMMs with full covariance matrices. It is noted that MML tends to prefer more complex models than either MDL or the Bayesian method. In some instances, of course, this may be advantageous, but we would argue that as data becomes more diffuse, for example, and cluster structure breaks down, the 'correct' response is to favour successively less complex models.

Is a GMM really a good model?

In many cases the simple answer is 'no'. As was shown in [23], the GMM itself will fail to discover 'true' structure in cases where the partitions are clearly non-Gaussian. The GMM remains, however, very appealing as it scales favourably with the dimensionality of the data, has good analytic properties and many data sets form clusters which are approximately Gaussian in nature. As was recommended in [23] a combination of methods may be more reliable and data visualisation is important.

Acknowledgements

The authors would like to thank the referees of the first version of this paper for their extremely helpful comments and criticism. Dirk Husmeier, Iead Rezek and Will Penny are, respectively, funded via the Jefferiss Research Trust, the Commission of the European Community (grant BMH4-CT97-2040) and the UK EPSRC (grant GR/K79062) whose support we gratefully acknowledge.

References

[1] E. Anderson. The Irises of the Gaspe peninsula. *Bull. Amer. Iris Soc.*, 59:2–5, 1935.

[2] R.A. Baxter and J.J. Oliver. MDL and MML: Similarities and differences. Technical report TR 207, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[3] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[4] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

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

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[7] I. Gath and B. Geva. Unsupervised Optimal Fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–781, 1989.

[8] P. Hall and D.M. Titterington. The use of uncategorized data to improve the performance of a nonparametric estimator of a mixture density. *Journal of the Royal Statistical Society (Series B)*, 47:155–163, 1985.

[9] D.J. Hand. *Kernel Discriminant Analysis.* Research Studies Press, 1994.

[10] A.C. Harvey. *The Econometric Analysis of Time Series.* Philip Allan, Oxford, 1981.

[11] D. Husmeier. *Modelling Conditional Probability Densities with Neural Networks.* PhD thesis, Department of Mathematics, King's College, University of London, 1997.

[12] H. Jeffreys. *Theory of Probability.* Oxford University Press, 1939.

[13] P.M. Lee. *Bayesian Statistics : An Introduction.* Edward Arnold, 1994.

[14] D.J.C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4:448–472, 1992.

[15] D.J.C. MacKay. The Evidence Framework applied to Classification Networks. *Neural Computation*, 4:720–736, 1992.

[16] J.J.K. O' Ruanaidth and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer, 1996.

[17] J.J. Oliver and R.A. Baxter. MML and Bayesianism: Similarities and differences. Technical report TR 206, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[18] J.J. Oliver, Baxter R.A., and Wallace C.S. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96)*, pages 364–372. Morgan Kaufmann Publishers, San Francisco, CA, 1996. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[19] I.A. Rezek and S.J. Roberts. Stochastic complexity measures for physiological signal analysis. Technical report, Imperial College, University of London, November 1996. Available from `http://www.ee.ic.ac.uk/hp/staff/sroberts.html`.

[20] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society (Series B)*, 59(4):731–758, 1997.

[21] B. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[22] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[23] S.J. Roberts. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272, 1997.

[24] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Number 26 in Monographs on statistics and applied probability. Chapman and Hall, London, 1986.

[25] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley, 1985.

[26] C.S. Wallace and P.R. Freeman. Estimation and Inference by Compact Coding. *Journal of the Royal Statistical Society (Series B).*, 49:240–252, 1987.

[27] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Number 60 in Monographs on statistics and applied probability. Chapman and Hall, London, 1995.