A Fast Distance Based Approach for Determining the Number of Components in Mixtures

Sujit K. Sahu and Russell C. H. Cheng
Faculty of Mathematical Studies
University of Southampton
Highfield, SO17 1BJ, UK.

August 23, 2002

A Fast Distance Based Approach for Determining the Number of Components in Mixtures

SUMMARY

The problem of determining the unknown number of components in mixtures is of considerable interest to researchers in many areas. This paper generalizes a Bayesian testing method based on the Kullback-Leibler distance proposed by Mengersen and Robert (1996). An alternative, weighted Kullback-Leibler distance is proposed as testing criterion. Explicit formulas for this distance are given for a number of mixture distributions. A step-wise testing procedure is proposed to select the minimum number of components adequate for the data. A fast, collapsing approach is proposed for reducing the number of components which does not require full re-fitting at each step. The method, using both distances, is compared to the Bayes factor approach. The method is easy to implement and is illustrated using `BUGS` (a general purpose software for Bayesian analysis).

Keywords: Bayes Factor, Kullback-Leibler distance, Gibbs sampler, Markov chain Monte Carlo, Mixture Model, Reversible Jump.

1 Introduction

Consider a family of probability density functions of the form

$$f^{(k)}(x) = \sum_{j=1}^{k} w_j f_j(x|\theta_j), \tag{1}$$

where the density $f_j(x|\theta_j)$ is given and parameterized by θ_j and the weights w_j, summing to 1. The density $f_j(x|\theta_j)$ may also depend on additional parameters and for notational convenience we will write f_j for $f_j(x|\theta_j)$ when the context is clear. Although we can work with different functional forms for the densities $f_j(x|\cdot)$, in this article we restrict ourselves to the case where the functional forms of the $f_j, j = 1, \dots, k$ are same. We observe $X_1, \dots, X_n$, a random sample of size n from (1). In this paper we allow the possibility of multi-dimensional observations as well.

We consider the problem of estimating k. In exploratory studies we will often be interested in a parsimonious explanation of the data and would be especially interested in identifying the smallest k that will nevertheless adequately explain the structure of the data. It is this particular problem that is the focus of this paper.

In classical non-Bayesian statistics this problem is non standard and is not straightforward. At first sight it might be thought that the problem is readily handled by Bayesian methods by calculating the posterior distribution for k. However, even here the problem is not straightforward because certain parameter combinations give rise to ambiguous k values making correct identification of k difficult. For example one component can be equally well, but unnecessarily, represented by two components:

$$w_k f_k(x|\theta_k) = w'_k f_k(x|\theta_k) + (w_k - w'_k) f_k(x|\theta_k).$$

In general there is no guarantee that the posterior distribution will not give probability mass to such two component representations of what is actually just one component. Thus the posterior distribution of k needs careful interpretation.

A commonly used approach by both Bayesians and non-Bayesians is the reductive stepwise-method. Here an initial $k = k_0$ (> 1) component mixture, $f^{(k_0)}$, is fitted to the data set where the value of k_0 is assumed to be more than adequate. The value of k is then progressively reduced until the fit is judged to be no longer acceptable. At any given stage, assuming $f^{(k)}$ is acceptable, the procedure is to

Step 1: Fit a $k-1$ component mixture, $f^{(k-1)}$.

Step 2: Compare $f^{(k-1)}$ with $f^{(k)}$ using a test of hypothesis to see if the reduced fit is still adequate.

These two steps are repeated until a reduction no longer gives an acceptable fit.

In the Bayesian approach, the Bayes factor is often used as the selection criterion for determining adequacy of the fit. See for example, Carlin and Chib (1995), Chib (1995), Raftery (1996) and Ishwaran *et al.* (2001). Other variants of the Bayes factor have also been considered for model selection in mixture settings. Dey *et al.* (1995) propose the use of pseudo-Bayes factor (Geisser and Eddy, 1979; Gelfand and Dey, 1994).

The calculation of the Bayes factor is usually difficult and requires rather elaborate numerical algorithms. The variable dimensional approach of Richardson and Green (1997) is not straightforward to implement in general mixture settings. The approaches based on the output of the MCMC algorithms to calculate the marginal likelihood (Chib, 1995) and hence the Bayes factor (Raftery, 1996), Roeder and Wasserman (1997) perform poorly, though improvements can be made by combining simulation and asymptotic approximations, see for example DiCiccio *et al.* (1997).

An interesting alternative is to use a distance measure to compare two distributions. For non-Bayesian approaches see e.g. Chen and Kalbfleisch (1996) and James *et al.* (2001). Mengersen and Robert (1996) have proposed the Kullback-Leibler (KL) distance. The KL distance is not a proper metric in the strict distance metric sense. However, it is generally recognized to be a convenient quantity for measuring the 'closeness' between distributions, so we shall in this paper use the term 'distance' in this informal sense. Although there are many possible distance measures between two densities available in the literature, the KL distance measure is attractive because of its simplicity and analytical tractability for mixtures of commonly used exponential family models, as we demonstrate here.

Our method is based on that of Mengersen and Robert, but in an unusual form. Our proposed version contains a modification to each of the previously described Steps 1 and 2.

Firstly, we note that a full re-fitting in Step 1 is not necessary for the following reason. Suppose that the data genuinely arises from a $k_1 (< k_0)$ component mixture. Then additional components will not represent true structure. Moreover, they will not significantly disturb or alter the fit of the k_1 component model. The only difficulty is the case mentioned previously where more than one component has been unnecessarily employed to explain what is really just one component.

The above consideration indicates that at an intermediate stage (where $k_1 < k \leq k_0$) then k_1 of the components will correspond to the true distribution and the remaining components correspond essentially to chance fluctuations. We do not therefore need to carry out a full re-fit in seeing if $(k-1)$ components are adequate. Instead we consider whether *merging* of two components in $f^{(k)}$ significantly worsens the fit. This merging operates by selecting two components from $f^{(k)}$ and replacing them by one component. The parameters of this component are recalculated, *but the other* $(k-2)$ *components and their weights are kept fixed.* We call such a reduced fit a *collapsed* version of $f^{(k)}$ and denote it by $f^{*(k-1)}_{(ij)}$, with the subscripts ij indexing one of the ${}^kC_2$ possible pairs of components that could be merged. Amongst all these collapsed versions we select the one whose distance (in a sense to be defined) from $f^{(k)}$ is smallest. We call this the *best collapsed version* and denote it by $f^{*(k-1)}$. Its distance from $f^{(k)}$ is denoted by $d(f^{(k)},\, f^{*(k-1)})$.

Once the best collapsed version is selected in this way we then have to decide if the fit is still adequate. A Bayesian test of hypothesis, based on the distance criterion, entails calculation of the posterior probability that $d(f^{(k)},\, f^{*(k-1)})$ is less than a pre-specified value. (A small value of this probability indicating that the two components in $f^{(k)}$ can be collapsed without significantly worsening the fit.). If the KL distance is used for $d(\cdot,\cdot)$ then this posterior probability can be expensive to calculate. Our second suggested modification is therefore to use an alternative distance criterion, still based on the KL distance, but that is much easier to calculate. The behavior of the test using this alternative distance measure is very similar to when the original KL distance is used, but it greatly reduces the computational effort to carry out the test of hypothesis.

In the overall method, we start with a reasonably large value of k, and reduce the number of components sequentially by collapsing until no further collapsing can be done without significant loss in the fit. A fresh fitting of the model with each reduction in the value of k is thus avoided. However, parameter estimates under a model with a smaller number of mixture components are more accurate in general. Hence a re-fit of the mixture model at the final step is advocated in order to ensure that the final posterior distribution has been unambiguously identified and that no further reduction is possible without loss of quality of fit.

In summary, the main contributions of the current article are as follows.

- We use accurate calculations (based on numerical quadratures) of the KL distance between $f^{(k)}$ and its collapsed version. This is in contrast to the Laplace approximations obtained by Mengersen and Robert (1996) which are only valid under suitable conditions.

- We also propose, as an alternative, an easy to implement weighted KL distance, with similar properties to the KL distance.

- We provide a step-wise method to determine k using a collapsing method to reduce the number of mixture components at each step. This method is generally applicable, fast and easy to implement in many mixture settings including multivariate normal mixtures and mixtures of gamma distributions.

It is well known that the Bayes factors choose the correct model under suitable assumptions asymptotically. The proposed method also enjoys the same property of consistency. In order to show this we need to consider consistency of the posterior distribution for the parameters of the mixture model $f^{(k)}$, since the distance is a function of the parameters under the mixture model. We prove the consistency of the proposed method in a separate paper (submitted elsewhere) since the proof is rather elaborate and technical.

The remainder of the paper is organized as follows. Section 2 develops the general testing procedure. Section 3 contains a number of mixture examples and provides expressions for the weighted KL distance in each case. We give numerical illustrations in Section 4. In Section 5 we conclude with a few summary remarks.

2 General Formulation

2.1 Test Procedure

Our proposed test procedure consists of the following five steps labelled T1 to T5.

T1 Select an initial k_0 and fit a k_0 component (Bayesian) mixture model, $f^{(k_0)}$, to the data. It is assumed that k_0 is definitely large enough to explain all the structure of the data.

Set $k = k_0$.

T2 Calculate the distances $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ between $f^{(k)}$ and each of its ${}^kC_2$ collapsed versions $f^{*(k-1)}_{(ij)}$.

T3 Select as the best collapsed version, that $f^{*(k-1)}_{(ij)}$ for which $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ is minimized. Label this best collapsed version by $f^{*(k-1)}$.

T4 Evaluate the posterior probability

$$P_{c_k}(k) = Pr\{d(f^{(k)}, f^{*(k-1)}) \le c_k \mid \text{ data }\}.$$

T5 If $P_{c_k}(k) > \alpha$ then replace $f^{(k)}$ by $f^{*(k-1)}$ and repeat from T2 with k reduced by 1.

A last checking step T6 which re-fits the mixture with the final value of k can also be implemented.

The test procedure avoids the non-identifiability problem where apparently different combinations of parameter values in a k component model actually identify the same mixture with $k-1$ components. However several aspects of the test procedure need discussion:

(i) The choice of $d(\cdot,\cdot)$, the distance measure.

(ii) The precise method of constructing the collapsed distribution $f^{*(k-1)}$.

(iii) The choice of the criteria parameters: c_k, α and k_0.

We discuss each of these in the following sections.

2.2 The Kullback-Leibler Distance d

As has been suggested by Mengersen and Robert (1996), the KL distance is one possible choice for the distance, d used in step T2 of the test procedure. The Kullback-Leibler distance between two densities f and g is defined as

$$d(f,g) = \int_{S(f)} f(x) \log \frac{f(x)}{g(x)} dx, \tag{2}$$

where $S(h)$ is the support of the density h, that is $S(h) = \{x : h(x) > 0\}$. The distance $d(f,g)$ is always non-negative, see for example, Rao (1973, p59). If the supports of f and g are the same, one simply writes

$$d(f,g) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

We use the distance $d(f^{(k)}, f^{*(k-1)})$ rather than $d(f^{*(k-1)}, f^{(k)})$, putting the more general model as the first argument. Although the supports of $f^{(k)}$ and $f^{*(k-1)}$ are same, $f^{(k)}$ is the more general model since the parameter space for $f^{*(k-1)}$ is a subset of the parameter space for $f^{(k)}$.

In general it is not possible to compute the KL distance $d(f^{(k)}, f^{*(k-1)})$ as defined by (2) *analytically.* Mengersen and Robert (1996) use a Laplace approximation for this in some special cases. The approximation they obtained was for the inverted distance $d(f^{*(1)}, f^{(2)})$ where $f^{*(1)}$ was chosen to be a univariate normal distribution and $f^{(2)}$ was the mixture of two univariate normal distributions.

We adopt the following approach in calculating $d(\cdot,\cdot)$. We have

$$d(f^{(k)}, f^{*(k-1)}) = \sum_{j=1}^{k} w_j \int f_j(x) \log \frac{f^{(k)}(x)}{f^{*(k-1)}(x)} dx.$$

Each integral in the summation can be carried out using numerical quadrature. In our version we used the routine given by Forsythe *et al.* (1977). The routine is an adaptive method, based on the Newton-Cotes 8-panel formula, that repeatedly checks for convergence of the integral in subintervals, refining the quadrature by adding additional points to subintervals where the estimate of error is larger than the selected allowed tolerance. The routine is thus well suited to evaluations involving mixture densities, with peaks in different places. For multidimensional integration the routine was used recursively. For example, a two dimensional integral was evaluated by calling the routine to evaluate the integral in the one dimensional form

$$\int\int f(x_1,x_2)dx_1dx_2 = \int g(x_2)dx_2.$$

and using a recursive call to the routine to evaluate the integrand $g(x_2) = \int f(x_1,x_2)dx_1$ within the first call.

Though evaluation of $d(f^{(k)}, f^{*(k-1)})$ by quadrature is straightforward, it is nevertheless an additional computational expense. In the next section we suggest an alternative weighted KL distance, $d^*(f, g)$, that can be used instead of the KL distance. In the present context, this alternative distance is much easier to calculate. Our numerical experience is that use of this distance is equally satisfactory, and so can be recommended as it is in general much easier to implement.

2.3 The Weighted Kullback-Leibler Distance d^*

Let w_j be a set of fixed weights. Then two k component mixture models with these same weights can be compared using the following distance measure. We let $f^{(k)}$ be as defined previously, and let $g^{(k)}$ be a k component mixture density with the same mixing weights as in $f^{(k)}$:

$$g^{(k)}(x) = \sum_{j=1}^{k} w_j g_j(x|\theta_j^*). \tag{3}$$

where the components of g_j may be different from the f_j and may possibly depend on different parameters θ_j^*. Then

$$d^*(f^{(k)}, g^{(k)}) = \sum_{j=1}^{k} w_j d(f_j, g_j), \tag{4}$$

defines a distance measure between $f^{(k)}$and $g^{(k)}$. For fixed weights, this measure clearly enjoys the same properties as the KL distance being simply a weighted sum of KL distances $d(f_j, g_j)$ between corresponding components. We shall refer to this as the *weighted KL distance.*

In our problem of determining k, the behavior of $d(f,g)$ and $d^*(f,g)$ is very similar. The following formula shows how they are related. We have,

$$
\begin{aligned}
d(f^{(k)}, g^{(k)}) &= \int f^{(k)}(x) \log \frac{f^{(k)}(x)}{g^{(k)}(x)} dx \\
&= \int \sum_{j=1}^{k} w_j f_j(x) \log \left[\frac{f_j(x)}{g_j(x)} \frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx \\
&= \sum_{j=1}^{k} w_j d(f_j, g_j) + \sum_{j=1}^{k} w_j \int f_j(x) \log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx \\
&= d^*(f^{(k)}, g^{(k)}) + \sum_{j=1}^{k} w_j \int f_j(x) \log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx. \qquad (5)
\end{aligned}
$$

The second term in (5) is non-positive. This follows immediately using $\ln z \le z - 1$ (for $z > 0$). We have

$$
\begin{aligned}
\textstyle\sum_{j=1}^{k} w_j \int f_j(x) \log \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} \right] dx &\le \textstyle\sum_{j=1}^{k} w_j \int f_j(x) \left[\frac{g_j(x)}{f_j(x)} \frac{f^{(k)}(x)}{g^{(k)}(x)} - 1 \right] dx \\
&= \textstyle\int \left[f^{(k)}(x) \right] dx - 1 = 0.
\end{aligned}
$$

Thus

$$
d(f^{(k)}, g^{(k)}) \le d^*(f^{(k)}, g^{(k)}).
$$

We propose that this weighted KL distance be used in our procedure for testing whether to merge components. Often it is easy to calculate $d^*(f_j, g_j)$. For example, if both the densities f_j and g_j belong to the exponential family then $d^*(f_j, g_j)$ has a simple form. Thus $d^*(f, g)$ often can be obtained explicitly when $d(f, g)$ cannot. In the context of our proposed collapsing method, the distance $d^*(f^{(k)}, g^{(k)})$ takes a particularly simple form. We give a number of examples in later sections.

2.4 Collapsing $f^{(k)}$

Let $f^{*(k-1)}_{(12)}$ be a collapsed version of $f^{(k)}$ which we assume is obtained by collapsing the first two components of $f^{(k)}$. To calculate $d^*(f^{(k)}, f^{*(k-1)}_{(12)})$, we view the collapsed version as being a special case of a k component mixture, i.e. $g^{(k)} = f^{*(k-1)}_{(12)}$. In the collapsing process the parameters of $f^{(k)}$ are regarded as fixed. The collapsed version is thus defined to be

$$
g^{(k)} = f^{*(k-1)}_{(12)} = \sum_{j=1}^{k} w_j f_j(x|\theta_j^*)
$$

with the condition that $\theta_1^* = \theta_2^* = \theta^*$, ($\theta^*$ is yet to be determined) and $\theta_j^* = \theta_j$ for $j = 3, ..., k$. Thus $g_j(x|\theta_j^*) = f_j(x|\theta_j)$ for $j = 3, ..., k$, and $d^*(f^{(k)},\ f^{*(k-1)}_{(12)})$ reduces to

$$
\begin{aligned}
d^*(f^{(k)},\ f^{*(k-1)}_{(12)}) &= \sum_{j=1}^{2} w_j d(f_j(.|\theta_j), f_j(.|\theta^*)) \\
&= \sum_{j=1}^{2} w_j E_j \ln \frac{f_j(X|\theta_j)}{f_j(X|\theta^*)} \qquad (6)
\end{aligned}
$$

where E_j denotes the expectation under the density $f(x|\theta_j)$ and

$$\begin{aligned}\theta^* &= \arg\left\{\min_\theta \sum_{j=1}^{2} w_j E_j \ln \frac{f_j(X|\theta_j)}{f_j(X|\theta)}\right\} \\ &= \arg\left\{\max_\theta \sum_{j=1}^{2} w_j E_j \ln f_j(X|\theta)\right\}.\end{aligned}$$

The above defines the collapsed version obtained from merging the first two components of $f^{(k)}$. The best collapsed version, denoted by $f^{*(k-1)}$, is simply the one that minimizes $d(f^{(k)}, f^{*(k-1)}_{(ij)})$ over all $i, j, i \neq j$ i.e. the best out of all ${}^kC_2$ possible collapsed versions of $f^{(k)}$.

Richardson and Green (1997) provide an intuitive moment matching method for collapsing two components. For their normal mixture examples our method produces similar results. However, they did not perform the last minimization over the ${}^kC_2$ possibilities.

2.5 Choice of c_k and k_0

We reduce k by 1 if $d(f^{(k)}, f^{*(k-1)})$ is small. A large distance emphasizes that the collapsing cannot be done without significantly deteriorating the fit. The Bayesian solution is to evaluate the posterior probability

$$P_{c_k}(k) = Pr\{d(f^{(k)}, f^{*(k-1)}) \leq c_k \mid \text{ data }\}.$$

A simple decision rule is to reduce k by 1 if the above posterior probability is high, (greater than $\alpha = 0.5$, for example). The formula (6) shows that the minimum distance $d(f^{(k)}, f^{*(k-1)})$ is a function of the parameters in $f^{(k)}$ only. Hence, to calculate $P_{c_k}(k)$ one needs to obtain the posterior density of the parameters in $f^{(k)}$ only.

The quantity c_k needs to be set to define the boundaries of the indifference zone for the k and $k-1$ component mixtures. We avoid being too prescriptive as a sensible choice is problem dependent. Obviously, the proposed method is sensitive to the choice of c_k. A large value of c_k will select a smaller number of components and a smaller value of c_k will select a higher number of components. In the practical examples presented in Section 4, we study the sensitivity. In effect, the freedom in choosing c_k is somewhat similar to the procedures using the Bayes factor. In the latter case one needs to compare the value of the Bayes factor using a calibration table and decide how large is the evidence.

An informed choice of c_k can be made before fitting the mixture distributions. We only *illustrate* with the following example. To fix ideas, suppose that it is desired to compare a mixture of two normal distributions with a standard normal distribution. (More details are given in Section 3.1). For the components of the mixture we assume that the standard deviations are the same (σ) and the mixing proportion is 50%. The two mean parameters are taken as $-\theta$ and θ. The parameters in the mixture

(θ and σ) are chosen so as to have mean zero and variance 1. This is to match the mean and variance of mixture with those of the standard normal distribution.

We are then interested to see how large the means should be in order to achieve a specific value of the weighted KL distance d^*. Figure 1 plots the densities of the standard normal distribution and the mixture distributions corresponding to different values of d^*. From this graph, an experimenter can choose the value of c_2 according to how far departure from the standard normal distribution s/he is willing to allow. For example, a value of c_2 between 0.1 and 0.3 does not let the mixture distribution assume bimodal shape and is therefore little bit small. Based on these considerations we recommend $c_2 = 0.5$ as a reasonable choice.

It is not desirable to choose one specific value of c_k for every problem. This is because neither distances (2) or (6) are invariant under different distributions. Consequently, the choice of c_k is problem specific and should be guided by the required size of the indifference zone between $f^{(k)}$ and its collapsed version. More discussion regarding the choice of c_k is provided in Section 5.

The maximum value of k denoted by k_0 should be chosen large enough so that $f^{(k_0)}$ is able to capture all the variations present in the data.

3 Examples: Theoretical Development

When the component distributions are from the exponential family we can often obtain expressions for the weighted KL distance in closed form. We illustrate with the class of multivariate normal and gamma distributions. In both cases we shall assume that the weight vector $\mathbf{w}$ is given a symmetric Dirichlet prior, $\mathcal{D}(\delta, \ldots, \delta)$ where δ is a known positive number.

3.1 Mixture of Multivariate Normals

3.1.1 Bayesian Formulation

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from the k-component mixture of p-variate normal distributions. Let $\boldsymbol{\theta}_j$ be the mean and Σ_j be the covariance matrix of the jth component ($j = 1, \ldots, k$). Let $Q_j = \Sigma_j^{-1}$. The density f_j of the jth mixture component is,

$$f_j(\mathbf{x}|\boldsymbol{\theta}, \Sigma_j) = N_p(\mathbf{x}; \boldsymbol{\theta}_j, \Sigma_j) = \frac{1}{(2\pi)^{p/2}} |\Sigma_j|^{-1/2} \exp\left\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\theta}_j)^T \Sigma_j^{-1} (\mathbf{x} - \boldsymbol{\theta}_j)\right\},$$

where $|A|$ is the determinant of a matrix A. The full mixture density is given by

$$f^{(k)}(\mathbf{x}) = \sum_{j=1}^{k} w_j f_j(x|\boldsymbol{\theta}_j, \Sigma_j). \tag{7}$$

Now we specify the prior distribution for the parameters in the mixture following Richardson and

Green (1997) and Stephens (2000). The mean vector $\boldsymbol{\theta}_j$ is assigned the independent normal prior distribution

$$\boldsymbol{\theta}_j \sim N_p(\boldsymbol{\xi}, \kappa^{-1})$$

where $\boldsymbol{\xi}$ is a p-vector and κ is a positive definite matrix. In our examples we choose each component of $\boldsymbol{\xi}$ to be the mid-point of the corresponding component of the p-dimensional data and κ is chosen to be a diagonal matrix with elements equal to 0.1 times the inverse squared range of the corresponding data component. As in Richardson and Green (1997) we keep κ fixed throughout.

The inverse covariance matrix Σ_j^{-1} is assigned the independent conjugate Wishart prior distribution as follows:

$$\Sigma_j^{-1} = Q_j \sim W_p(2\alpha, 2\beta)$$

where 2α is the assumed prior degrees of freedom ($\geq p$) and β is a positive definite matrix. The choice $2\alpha = p$ corresponds to the case that the prior is thought to be formed of p observations; and in p dimensional case this corresponds to weak prior information. In our example we have implemented this choice. The choice of β is discussed later.

We say that $\mathbf{Y}$ has the Wishart distribution $W_p(m, R)$ if its density is proportional to

$$|R|^{m/2}|y|^{\frac{1}{2}(m-p-1)}e^{-\frac{1}{2}\mathrm{tr}(Ry)}$$

if y is a $p \times p$ positive definite matrix. (Here $\mathrm{tr}(A)$ is the trace of a matrix A.) This is the parameterization used by for example, Anderson (1984, p249) and the `BUGS` (Spiegelhalter *et al.*, 1996) software. To choose the hyper-parameters, we note that $E(\mathbf{Y}) = (R/m)^{-1}$ and $E(\mathbf{Y}^{-1}) = R/(m-p-1)$ if the expectations exist.

We may assume the following hyper-prior for β,

$$\beta \sim W_p(2g, 2h)$$

where $g > 0$ and h is a positive definite matrix. Now the quantities g and h should be chosen sensibly. Note that $E(Q_j|\beta) = \alpha\,\beta^{-1}$. Using the iterated expectations $E(Q_j) = EE(Q_j|\beta)$, we now have $E(Q_j) = \alpha\,\frac{2h}{2g-p-1}$. In order to match the conditional and un-conditional expectations we may choose $2g = p+3$ and take $\beta^{-1} = h$ when β is kept fixed. In fact we distinguish between two cases: one treats β random and assigns the above hyper-prior distribution and the other keeps β fixed. Following Stephens (2000) we take h to be $100 \times g/\alpha I$ where I is the identity matrix.

3.1.2 Distance

Suppose that $g_j(x|\boldsymbol{\theta}_j^*, \Lambda_j) = N_p(\mathbf{x}; \boldsymbol{\theta}_j^*, \Lambda_j)$ and

$$g^{(k)}(\mathbf{x}) = \sum_{j=1}^{k} w_j g_j(x|\boldsymbol{\theta}_j^*, \Lambda_j).$$

Lemma 1

The distance (4) between $f^{(k)}$ and $g^{(k)}$ is given by,

$$2\ d^*(f^{(k)}, g^{(k)}) = \sum_{j=1}^{k} w_j \left\{ \log|\Lambda_j \Sigma_j^{-1}| + tr(\Lambda_j^{-1}\Sigma_j) - p + (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)^T \Lambda_j^{-1} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*) \right\}. \tag{8}$$

Let the first two components of $g^{(k)}$ have the common parameters $\boldsymbol{\theta}^*$ and Λ and all other components have the parameters equal to the corresponding parameters of $f^{(k)}$. That is $\boldsymbol{\theta}_1^* = \boldsymbol{\theta}_2^* = \boldsymbol{\theta}^*$ and $\Lambda_1 = \Lambda_2 = \Lambda$ and $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}_i, \Lambda_i = \Sigma_i, i = 3, \ldots, k$. Consequently, $g^{(k)}$ is the mixture of $k-1$ components. Note that the mixing weights $w_1, \ldots, w_k$ are unchanged.

Lemma 2

The choices of $\boldsymbol{\theta}^*$ and Λ which minimize the distance (8) between $f^{(k)}$ and the collapsed $g^{(k)}$ as described above are given by

$$\boldsymbol{\theta}^* = \frac{w_1\boldsymbol{\theta}_1 + w_2\boldsymbol{\theta}_2}{w_1 + w_2}$$

and

$$\begin{aligned}
\Lambda &= \tfrac{1}{w_1+w_2} \left\{ w_1\Sigma_1 + w_2\Sigma_2 + w_1(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)^T + w_2(\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*)^T \right\} \\
&= \tfrac{1}{w_1+w_2} \left\{ w_1\Sigma_1 + w_2\Sigma_2 + \tfrac{w_1 w_2}{w_1+w_2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \right\}.
\end{aligned}$$

The minimum distance is given by,

$$2\ d^*(f^{(k)}, g^{(k)}) = \sum_{j=1}^{2} w_j \left\{ \log|\Lambda\Sigma_j^{-1}| + tr(\Lambda^{-1}\Sigma_j) - p \right\} + \frac{w_1 w_2}{(w_1+w_2)^2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \Lambda^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2).$$

This matches exactly with Lemma 2.1 in Mengersen and Robert (1996). Their Lemma 2.2 is a special case of our Lemma 1.

3.2 Mixture of Gamma Distributions

3.2.1 Bayesian Formulation

Suppose that X is a mixture of gamma random variables and the jth component in the mixture has the density $f_j(x|\nu, \theta_j)$. This gamma distribution has mean ν/θ_j. Thus we have $f^{(k)}(x) = \sum_{j=1}^{k} w_j f_j(x|\nu, \theta_j)$. Following Dey *et al.* (1995) we assume a $\Gamma(\alpha, \beta)$ prior for ν, and independent $\Gamma(\xi, \kappa)$ prior for θ_j. Further, all the hyper-parameters $\alpha, \beta, \xi, \kappa$ are assumed to be known.

3.2.2 Distance

We first obtain the distance between $f^{(k)}$ and $g^{(k)}$ where the jth component of $g^{(k)}$ is $\Gamma(\nu, \theta_j^*)$.

Lemma 3

The distance (4) between $f^{(k)}$ and $g^{(k)}$ is given by,

$$d^*(f^{(k)}, g^{(k)}) = \nu \sum_{j=1}^{k} w_j \left\{ \log\left(\frac{\theta_j}{\theta_j^*}\right) + \frac{\theta_j^*}{\theta_j} - 1 \right\}. \tag{9}$$

Let the first two components of $g^{(k)}$ have the common parameter θ^* and all other components have the parameters equal to the corresponding parameters of $f^{(k)}$. That is $\theta_1^* = \theta_2^* = \theta^*$ and $\theta_i^* = \theta_i, i = 3, \ldots, k$. Consequently, $g^{(k)}$ is the mixture of $k-1$ components. Note that the mixing weights $w_1, \ldots, w_k$ are unchanged.

Lemma 4

The choice of θ^* which minimize the distance (9) between $f^{(k)}$ and the collapsed $g^{(k)}$ as described above is given by

$$\theta^* = \frac{w_1 + w_2}{w_1/\theta_1 + w_2/\theta_2}.$$

The minimum distance is given by,

$$d^*(f^{(k)}, g^{(k)}) = \nu \left[w_1 \log \theta_1 + w_2 \log \theta_2 - (w_1 + w_2) \log \theta^* \right].$$

4 Examples: Numerical Illustrations

4.1 Preliminaries

In mixture problems it is well known that the MCMC schemes experience slow convergence. We do not focus on convergence of the MCMC methods here. In all the examples given below we have used the Gibbs sampler. In each case we discarded first 100,000 iterates and used the next 100,000 iterates to make inference.

Bayes procedures including the Bayes factor are generally sensitive to the choice of the hyper-parameter values. For the examples based on the univariate normal distribution, we have implemented hyper-parameter values used by Richardson and Green (1997) so that we can compare their results with ours for the same prior specification. For the multivariate normal example we have used the hyper-parameter values similar to the ones adopted by Stephens (2000). Section 3.1 provides additional justifications for choosing the specific values.

A further problem in mixture models is in the non-identifiability of individual components. Here we have adopted the general idea of post-processing MCMC output by re-labeling, see e.g. Richardson and Green (1997), although other possibilities exist, see for example Celeux *et al.* (2000). In the univariate examples, we re-label the mixture components according to the order of the value of a parameter of interest. In the multi-dimensional case we re-label using the value of the first component mean.

4.2 Univariate Normal Mixtures

4.2.1 A Simulated Example

Our first example compares the behavior of our method with that using the Bayes factor method of identifying the posterior probability. We generate data from a two component normal mixture model with equal mixing proportion. The variances of the components are assumed to be 1 while the two means are set at 0 and 3. These choices ensure that the resulting model is indeed a two component mixture model. Starting with a sample of size 10, we run the simulation experiment with an increment of 5 further samples until 250 samples. We calculate the posterior probability of the two component model using the reversible jump MCMC method as described in Richardson and Green (1997) using k_0 to be 5. The same value of k_0 is used for the proposed method.

In each simulation we also calculate the posterior probabilities $P_{c_k}(k)$ for $k = 5, 4, 3$ and 2 using the weighted KL, d^*, distance. Since the true model is a two component one we expect $P_{c_k}(k)$ to be high for $k = 5, 4$ and 3 and low for $k = 2$ for reasonable choices of c_k. Guided by the example in Section 2.5, we take $c_k = 0.5$ for all k.

The posterior probabilities are plotted in Figure 2. The posterior probability of collapsing the fitted three component model rapidly rises to one as expected. The collapsing probabilities for the 4 and 5 component models were all calculated to be one. The graph also shows that using about 150 simulated data points from the two component model the proposed distance method reaches the correct decision of not to collapse the components of the fitted two component model. The posterior probability (corresponding to the Bayes factor) of the two component model increases slowly as sample size increases. The simulation experiment has been repeated and replicated many times to see the sensitivity. We report our findings below.

- **Exact (d) versus weighted distance (d^*):** The same model choice conclusions are arrived using the exact KL distance d. However, here the c_k values are to be set much lower than the choice of 0.5 used previously, as expected. Thus similar discrimination can be achieved with either distance. See the next example where we report both the exact and weighted distance.

- **Replication:** The simulation experiment has been replicated 10 times for fixed sample sizes of 100, 150 and 200. We observed insignificant changes in the posterior probabilities which are not worth reporting.

- k_0: The posterior model probability (corresponding to the Bayes factor) reduces if k_0 is increased. However, the posterior probabilities $P_{c_k}(k)$ do not change when k_0 is increased from 5 to 10.

Arbitrarily large values may cause problems when only a small number of data points are to be modeled, perhaps due to poor MCMC estimation.

- **Local minima:** A referee has expressed concern that the proposed method may reach a local minima and as a result would select the wrong model. In our experience this is not the case. However, to give some added reassurance on this point we extend the simulation experiment. We simulate $n = 200$ random samples from a two component normal mixture with components having unit standard deviation and means 0 and θ. We report the results for two values of $\theta = 2$ and 3. The mixing proportion is kept fixed at 50%. We compute the posterior distribution of the distance between a fully fitted three component mixture and its collapsed two component version. We compare the above posterior distribution with that of the posterior distribution of the distance between a collapsed three and its collapsed two component version. The collapsed three component mixture was first arrived at from fitting a full four component mixture to the data. These two posterior distributions are now compared since using either of these two distributions our method decides between a three and two component mixture model for the data.

  If the method were to reach a local minima then the two posterior distributions will be quite different and as a result different conclusions on the number of components will be arrived by the two distances. In Figure 3 we plot the kernel density estimates of the distances for $\theta = 2$ and 3. The corresponding true mixture density is plotted in the left panel. It is clear from the figure that there is no difference at all between the two distributions and the same conclusion is reached using either distributions. There would have been significant differences between the two distributions if the method were to reach a 'local minima' while collapsing from the fitted four component mixture to the true two component mixture.

  Figure 3 also reveals the possible effect of the choice of c_k. The same conclusion of collapsing to a two component mixture is reached for any value of c_k greater than 0.5. From the numerical experimentation in Section 2.5 we know that values of the weighted distance greater than 0.5 enable the mixture to become multimodal.

4.2.2 Enzyme Example

We implement our methodology for the enzyme data set discussed in Richardson and Green (1997). The posterior probability distribution of k from their method is given in Table 1. The posterior distribution of k does not provide strong evidence for any single value of k but generally suggests between 3–5 components.

In Figure 4 we plot the posterior densities of the distances $d(f^{(k)}, f^{*(k-1)})$ and $d^*(f^{(k)}, f^{*(k-1)})$ for $k = 2, 3$ and 4. The densities for the KL and weighted KL distances are plotted using the same plotting symbol. Obviously, the one on the right hand side is the density of the weighted KL distance in each case. As expected, the distribution of the distance $d(f^{(k)}, f^{*(k-1)})$ shifts towards 0 for increasing values

k	2	3	4	5	≥ 6	
$\pi(k\|y)$	0.047	0.343	0.307	0.200	0.103	

Table 1: Posterior distribution of k for the enzyme data example.

of k.

Consider the graphs from the $k = 4$ model. The values of $d(f^{(4)}, f^{*(3)})$ are very close to zero. In fact, the mean and median of $d(f^{(4)}, f^{*(3)})$ are 0.012 and 0.009, respectively. Hence it is concluded that the data does not support the $k = 4$ model. To compare the $k = 3$ versus the $k = 2$ model we note the densities plotted as dotted lines. The mean and median of $d(f^{(3)}, f^{*(2)})$ are 0.028 and 0.026, respectively. This provides some positive evidence for the $k = 3$ model.

The two densities on the far right (solid lines) are for the $k = 2$ case. Since the location of the distance here is very high, one clearly rejects the $k = 1$ model. Note the similarities between the densities of the KL and weighted KL distances.

The foregoing discussion suggests that a three component model is quite adequate for the data. However, a two component model is not too bad an approximation for the three component model. A more precise statement can be made by choosing a particular value of c_k. For example, if c_3 is chosen as 0.025 then $P_{0.025}(3) = 0.47$. Since this is not higher than 0.5, one selects the $k = 2$ model.

The effect of collapsing as opposed to fully fitting the model each time k is reduced, is illustrated in Figure 5. The plotted predictive densities are for the fully fitted two component model and the collapsed two component model obtained by collapsing the three component model. The plot does not show any significant difference between the two densities. This illustrates the discussion in the introduction that collapsed models will retain the essential character of fully fitted models.

4.3 Mixture of Multivariate Normals

Our next example illustrates the method applied to multivariate data. We consider the well known Old Faithful data. The data consist of 298 eruptions of the Old Faithful geyser in the Yellowstone National Park. Each observation has two components ($p = 2$): *duration* (in minutes) of the previous eruption and *waiting* time before the next eruption. A scatter plot of this bivariate data shows the possibility of two or more modes. Strictly speaking, there is time dependence between the observations, see for example Azzalini and Bowman (1990). However, we only use it to illustrate our technique as has been done by Stephens (2000).

The kernel density estimates of the distances are given in Figure 6. Observe that the distribution of the weighted KL distance gets closer to the distribution of the KL distance as k increases. By choosing $c_3 = 0.1$, we see that $P_{0.1}(3) = 0.37$. Since this probability is not high we are not able reject the $k = 3$ model for this choice of c_3. We do not attempt fitting mixtures with a larger number of components. Our experience suggests that in order to fit a larger number of components more precise prior distributions are required. However, our analysis here is based largely on vague prior distributions.

We also study the sensitivity of the distance measure to changes in prior distribution in this example. Figure 7 reports the kernel density estimates of the KL and weighted KL distances for the random and fixed β models. The plot shows that the proposed method is not very sensitive to changes resulting from prior distributions on the inverse covariance matrices.

4.4 Gamma Mixture Example

We consider a data set arising from geological measurements. The measurements are known to be mixtures of (non-negative) skewed distributions with component distributions arising from different geological components. An important problem thus is to estimate the number of components in the mixture.

The data set consist of 200 observations and there is a heavy tail in the observed data, see Figure 8. Here we attempt our mixture of gamma distributions and then determine the number of components. Various hyper-parameters are chosen as follows. The hyper-parameters α and β are set at unity. The parameter ξ is set to 2 and κ is set at the mean of the data. These choices lead to the prior predictive mean being equal to the mean of the data.

In this example we only use the weighted KL distance to determine the number of components since the distance itself is very small when comparing the $f^{(3)}$ versus $f^{*(2)}$. Also the weighted KL distance can be calculated using the general purpose software `BUGS` without much additional programming effort. The `BUGS` code and the data set can be obtained from the authors.

The posterior densities of the weighted KL distance (d^*) for $k = 2$ and 3 are plotted in Figure 9. We see that the two component gamma mixture model is better than the one component model. The estimate of $P_{0.1}(2)$ is 0.25. On the other hand the three component model does not provide a much better fit than the two component model. Here $P_{0.1}(3)$ is 0.997.

5 Discussion

This paper provides an easy to implement Bayesian method for determining the number of components in finite mixture densities. A weighted KL distance, an alternative to the KL distance, between a k component mixture and its collapsed $k-1$ component version has been developed in the paper. Often the weighted KL distance is easy to calculate and as a result univariate mixtures of standard exponential family distributions can be compared using the current version of `BUGS`. Accurate versions of the KL distance using numerical quadrature formula have been suggested and illustrated.

The proposed procedure can be applied even under non-informative prior distributions which are often improper. The Bayes factors are not interpretable for improper prior distributions. Moreover, the proposed method is consistent, see the simulation example. A theoretical proof of consistency has been submitted elsewhere for publication.

The collapsing scheme developed in the paper can also be used as a method when implementing the reversible jump MCMC. Since our scheme is based on minimizing the weighted KL distance d^* it may lead to a more efficient reversible jump MCMC, although we have not checked this empirically.

The methodology presented here does not necessarily converge to a local minima. In the current formulation it is of interest to see if two components in a mixture model can be collapsed without worsening the fit. In the next stage the procedure searches for the pair of components (among the ${}^kC_2$ possibilities) collapsing of which will produce the least discrepancy in fit. Finally, an appropriate Bayesian test of hypothesis decides whether to accept the proposed collapse. Thus the procedure does not end up finding a local minimum since the components to be collapsed are searched globally and a Bayesian test is used at the end.

Finally, some discussion regarding the constants c_k is appropriate. As suggested in the paper, the weighted KL distance should be calibrated between typical mixtures to be fitted to the data. A graphical diagnostics such as Figure 1 can be used to find guideline values of c_k. In addition, one can plot the posterior densities of the distance for comparing the mixture models as has been done in Figure 6. Relative locations of the posterior distributions will provide insights into the choice of c_k, and possible effect of choosing particular values. As a result such a plot provides an easy way to study the sensitivity of the method with respect to the choice of c_k.

REFERENCES

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis.* Chichester: John Wiley & Sons.

Azzalini, A. and Bowman, A. W. (1990) A Look at Some Data on the Old Faithful Geyser. *Applied Statistics*, **39**, 357–365.

Carlin, B. P. and Chib, S. (1995) Bayesian Model Choice via Markov Chain Monte Carlo. *Journal of the Royal Statistical Society*, B, **57**, 473–484.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

Chen J.H. and Kalbfleisch, J. D. (1996) Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, **24**, 167–175.

Chib, S. (1995) Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Dey, D. K., Kuo, L. and Sahu, S. K. (1995) A Bayesian predictive approach to determining the number of components in a mixture distribution. *Statistics and Computing*, **5**, 297–305.

DiCiccio, T. J., Kass, R. E., Raftery, A. Wasserman, L. (1997) Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, B, **56**, 363–375.

Forsythe, G.E., Malcolm, M.A. and Moler, C.B. (1977) *Computer Methods for Mathematical Computation.* Upper Saddle River, NJ: Prentice-Hall.

Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society*, B **56**, 501–514.

Ishwaran, H., James, L. F., and Sun, J. Y. (2001) Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, **96**, 1316–1332.

James, L. F., Preibe C. E. and Marchette, D. J. (2001) Consistent estimation of mixture complexity. *Annals of Statistics*, **29**, 1281–1296.

Mengersen, K. and Robert, C. P. (1996) Testing for mixtures: A Bayesian Entropic Approach (with discussion). In *Bayesian Statistics 5*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press, pp. 255–276.

Raftery, A. E. (1996) Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (Eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall, pp 163–187.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications.* New York: John Wiley and Sons.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian Density Estimation Using Mixture of Normals. *Journal of the American Statistical Association*, **92**, 894–902.

Richardson, S. and Green, P. J. (1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society*, B, **59**, 473–484.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1996) Computation on Bayesian graphical models. In *Bayesian Statistics 5*, (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press, pp. 407–426.

Stephens, M. (2000) Bayesian Analysis of Mixture Models with an Unknown Number of Components– an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

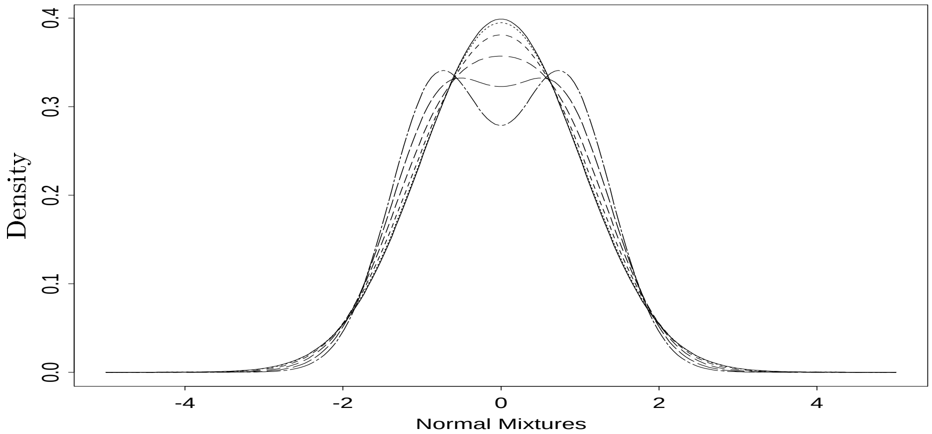

Figure 1: Examples of normal mixture densities corresponding to different values of the distance. The solid line is for the standard normal distribution. The weighted distance (d^*) increases as the height of the densities at zero decreases. There are five mixture densities corresponding to d^* equal to 0.1, 0.2, 0.3, 0.4 and 0.5.

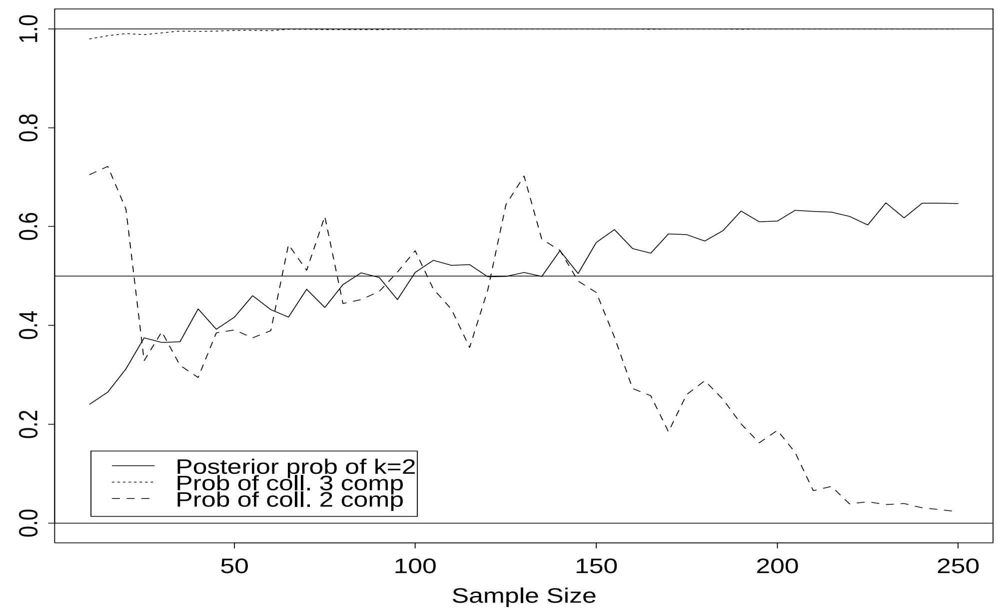

Figure 2: The probabilities of choosing the correct model (which has two components). Solid line is the posterior probability of the two component model calculated using the reversible jump method. The dotted line is $P_{0.5}(3)$ while the broken line is $P_{0.5}(2)$ using the weighted distance d^*.

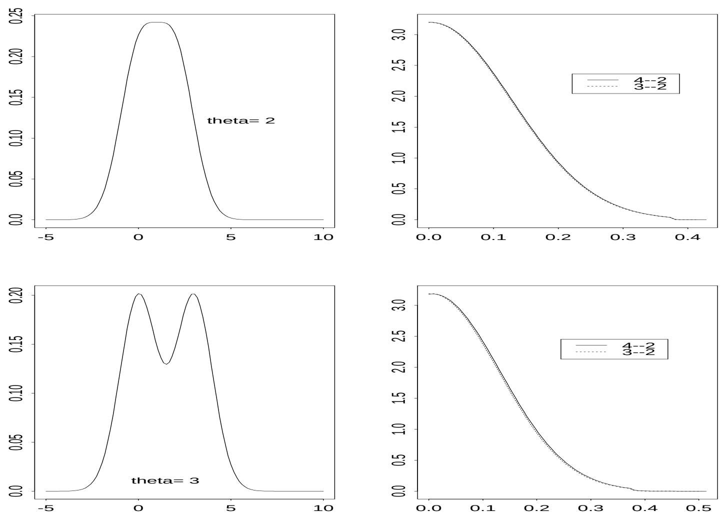

Figure 3: The true mixture density (left column) and the kernel density estimates (right column) of the posterior distributions of the distances.

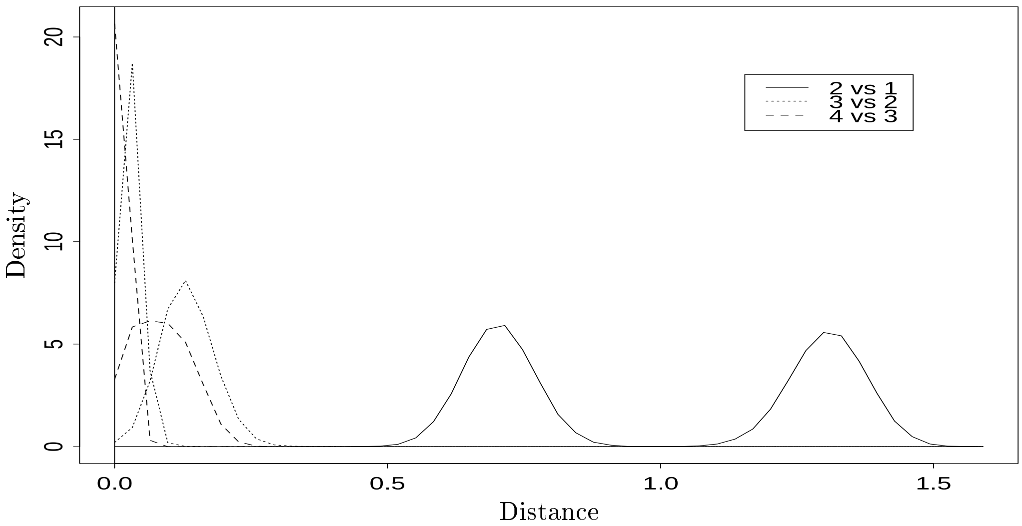

Figure 4: Kernel density estimates of the posterior distribution of the exact distance (d) and the weighted distance (d^*) for the enzyme data example. The density curve for d^* is on the right (with the same plotting symbol) in each case.

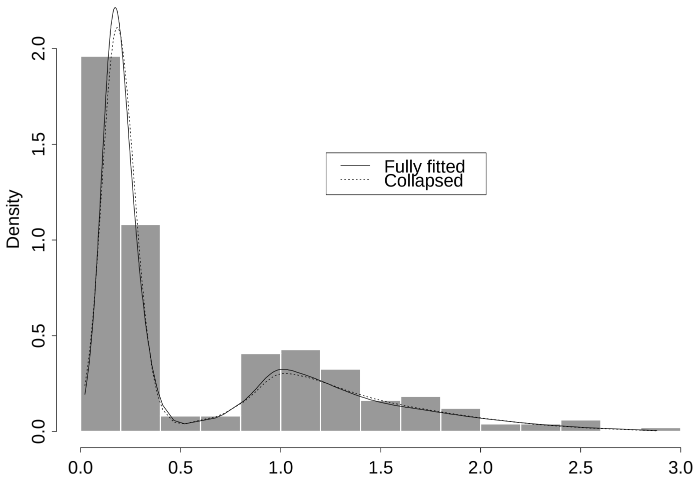

Figure 5: Predictive densities for the enzyme example.

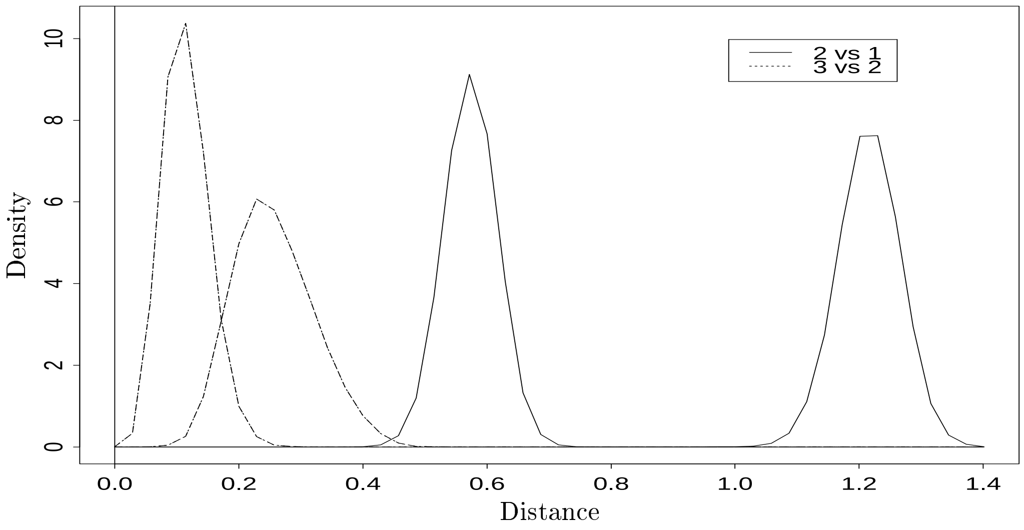

Figure 6: Kernel density estimates of the posterior distribution of the exact distance (d) and the weighted distance (d^*) for the Old Faithful data example. The density curve for d^* is on the right (with the same plotting symbol) in each case.

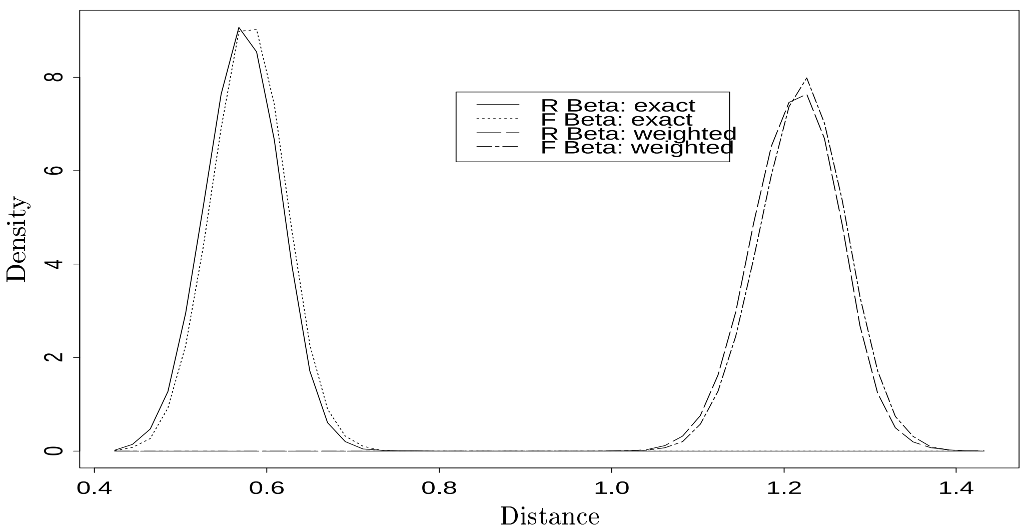

Figure 7: Kernel density estimates of the posterior distribution of the exact distance (d) and the weighted distance d^* for the Old Faithful data example. This plot shows the sensitivity of the distance measure for random and fixed β model.

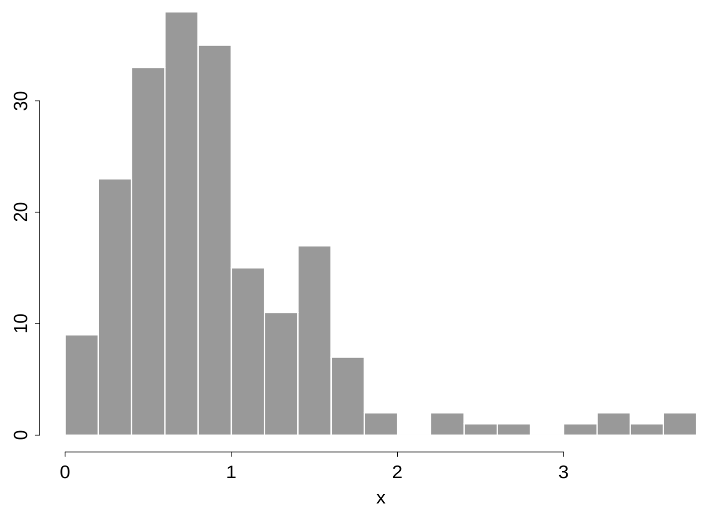

Figure 8: Histogram of the data for mixture of gamma example.

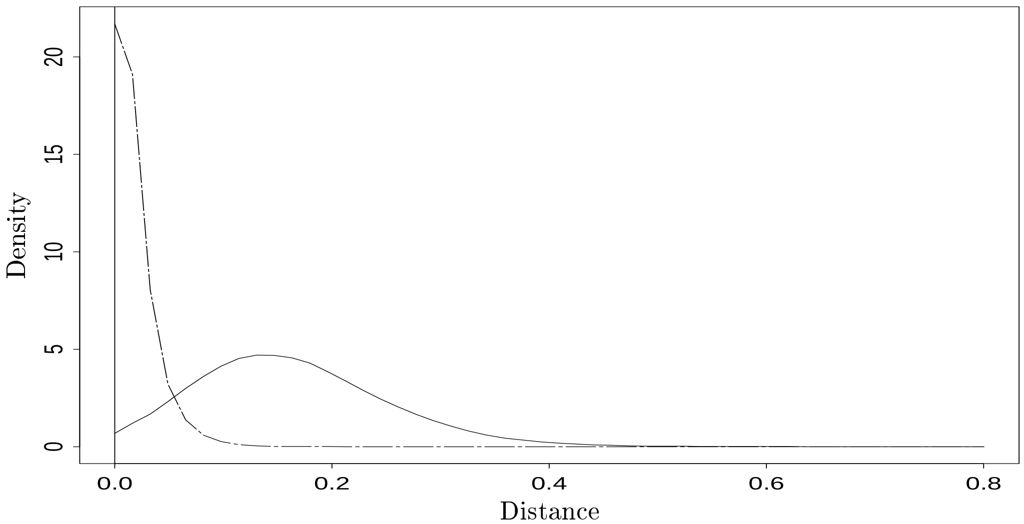

Figure 9: Kernel density estimates of the posterior distribution of the weighted distance d^* for the mixture of gamma example. Solid line is for the 2 vs. 1 case and the broken line is for the 3 vs. 2 case.