On Bayesian analysis of mixtures with an unknown number of components*

Sylvia Richardson[†] INSERM, France.

Peter J. Green[‡] University of Bristol, UK.

Revised version: 6 November 1996

Summary

New methodology for fully Bayesian mixture analysis is developed, making use of reversible jump Markov chain Monte Carlo methods, that are capable of jumping between the parameter subspaces corresponding to different numbers of components in the mixture. A sample from the full joint distribution of all unknown variables is thereby generated, and this can be used as a basis for a thorough presentation of many aspects of the posterior distribution. The methodology is applied here to the analysis of univariate normal mixtures, using a hierarchical prior model that offers an approach to dealing with weak prior information while avoiding the mathematical pitfalls of using improper priors in the mixture context.

Some key words: Birth and death process, Classification, Galaxy data, Heterogeneity, Lake acidity data, Markov chain Monte Carlo, Normal mixtures, Predictive distribution, Reversible jump algorithms, Sensitivity analysis.

1 Introduction

This article is a contribution to the methodology of fully Bayesian mixture modelling. We stress the word "fully" in two senses. First, we model the number of components and the mixture component parameters jointly and base inference about these quantities on their posterior probabilities. This is in contrast to most previous Bayesian treatments of mixture estimation, which consider models for different numbers of components separately, and use significance tests or other non-Bayesian criteria to infer the number of components. Secondly, we aim to present posterior distributions of our objects of inference (model parameters and predictive densities), and not just "best estimates".

There are three key ideas in our treatment.

First, we demonstrate that novel MCMC methods, the "reversible jump" samplers introduced by Green (1994, 1995), can be used to sample mixture representations with an unknown and hence varying number of components. We believe these methods are preferable on grounds of convenience,

*A version of this paper was presented at the workshop on Model Interpretation and Model Robustness in Highly Structured Stochastic Systems, Luminy, June 1995. First written version 12 February 1996.

[†]INSERM U.170, 16 avenue Paul Vaillant Couturier, 94807 Villejuif, France.
Email: richardson@vjf.inserm.fr

[‡]Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: P.J.Green@bristol.ac.uk.

accuracy and flexibility to the use of analytic approximations or other recently-proposed MCMC techniques.

Secondly, we show that a sample-based approach to computation in mixture models allows much more subtle extraction of information from posterior distributions. We give examples of presentation from posteriors, that tease out alternative explanations of the data, which would be difficult to discover by other approaches.

Finally, we base our experiments and discussion on a hierarchical model for mixtures that aims to provide a simple and generalisable way of being weakly informative about parameters of mixture models. We propose a specific model for univariate normal mixtures, used throughout to illustrate implementation and performance, but emphasise that the rest of our methodology is in no way restricted to this particular model.

Some of the issues considered in the paper, including the key ideas above, have relevance well beyond mixture problems. In particular, issues concerning presentation of posterior distributions arise in many problems of inference about functions, and the interesting questions raised about labelling of parameters occur whenever the statistical model has partial invariance to permutations of variables.

Since the beginning of the century, there has been strong and sustained interest in finite mixture distributions, attributable to the complementary aspects of mixture models: they provide, first, a natural framework for the modelling of heterogeneity, thereby establishing links with cluster analysis, and, secondly, an appealing semi-parametric structure in which to model unknown distributional shapes, whether the objective is density estimation or the flexible construction of Bayesian priors. The whole field is comprehensively discussed by Titterington, Smith and Makov (1985), and McLachlan and Basford (1988).

Statistical analysis of mixtures has not been straightforward, with non-standard problems posed by the geometry of the parameter space, and also computational difficulties. Headway has been made recently both in dealing with the challenging distributional problems in testing hypotheses about the number of mixture components (Dacunha-Castelle and Gassiat, 1996; Lindsay, 1995; Feng and McCulloch, 1996), and on the computational side, by implementation of variants of the EM algorithm (Celeux, Chauveau and Diebolt, 1996). However, we strongly believe that the Bayesian paradigm is particularly suited to mixture analysis especially with an unknown number of components.

Much previous work on finite mixture estimation, Bayesian or otherwise, has separated the issues of testing the number of components k from estimation with k fixed. For the fixed-k case, a comprehensive Bayesian treatment using MCMC methods was presented in Diebolt and Robert (1994). Early approaches to the general case where k is unknown typically adopted a different style of modelling, treating the problem as an example of "Bayesian nonparametrics", and basing prior distributions on the Dirichlet process; see Escobar and West (1995) for example. Other authors, for example Mengersen and Robert (1996), Raftery (1996) and Roeder and Wasserman (1995) have proposed to use, respectively, a Kullback-Leibler distance, a Laplace-Metropolis estimator or a Schwarz criterion to choose the number of components. The more direct line we adopt here, of modelling the unknown-k case by mixing over the fixed-k case, and making fully Bayesian inference, has been followed by only a few authors, including Nobile (1994), and Phillips and Smith (1996).

The paper is structured as follows. In Section 2, we present a Bayesian hierarchical model for mixtures. Markov chain Monte Carlo methods for variable-dimension problems are discussed in Section 3, and then adapted to the particular case of mixture analysis. In Section 4, performance of the methodology is assessed through application to three real data sets, and in Sections 5 and 6, sensitivity and MCMC performance issues are considered. Section 7 covers classification based on mixture modelling, and we conclude with a discussion of extensions and an outline of further work.

2 Bayesian models for mixtures

2.1 Basic formulation

We write the basic mixture model for independent scalar or vector observations y_i as

$$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \qquad \text{independently for } i = 1, 2, \dots, n, \tag{1}$$

where $f(\cdot|\theta)$ is a given parametric family of densities indexed by a scalar or vector parameter θ. The objective of the analysis is inference about the unknowns: the number k of components, the component parameters θ_j and the component weights w_j, summing to 1.

Such a model arises in two rather distinct contexts. In the first, we postulate a *heterogeneous population* consisting of groups $j = 1, 2, \dots, k$ of sizes proportional to w_j, from which our random sample is drawn. The identity or label of the group from which each observation is drawn is unknown. In this situation, it is natural to regard the group label z_i for the i^{th} observation as a latent *allocation variable.* The z_i are supposed independently drawn from the distributions

$$p(z_i = j) = w_j \qquad \text{for } j = 1, 2, \dots, k, \tag{2}$$

and, given the values of the z_i, the observations are drawn from their respective individual subpopulations:

$$y_i|z \sim f(\cdot|\theta_{z_i}) \qquad \text{independently for } i = 1, 2, \dots, n. \tag{3}$$

In the second context, not the prime focus of this paper, the mixture model (1) is thought of as a convenient parsimonious representation of a non-standard density, and the objective of inference is a kind of semi-parametric density estimation.

In either case, the formulation given by (2) and (3), is convenient for calculation and interpretation, and we will make continued use of it. Integrating z out from (2) and (3) brings us back to (1). Note that we have specified a population model; not all components are necessarily represented in a finite sample, so there may be "empty components".

2.2 Hierarchical model and priors

In a Bayesian framework, the unknowns k, w and θ are regarded as drawn from appropriate prior distributions. The joint distribution of all variables can be written in general as

$$p(k, w, z, \theta, y) = p(k)p(w|k)p(z|w, k)p(\theta|z, w, k)p(y|\theta, z, w, k), \tag{4}$$

where here and throughout the paper we are using $p(\cdot|\cdot)$ to denote generic conditional distributions consistent with this joint specification, and we use the notation $w = (w_j)_{j=1}^k$, $z = (z_i)_{i=1}^n$, $\theta = (\theta_j)_{j=1}^k$ and $y = (y_i)_{i=1}^n$. In (4), it is natural to impose further conditional independences, so that $p(\theta|z, w, k) = p(\theta|k)$ and $p(y|\theta, z, w, k) = p(y|\theta, z)$. Thus the joint distribution (4) simplifies to give the Bayesian hierarchical model

$$p(k, w, z, \theta, y) = p(k)p(w|k)p(z|w, k)p(\theta|k)p(y|\theta, z).$$

We only consider models in which $p(y|\theta, z)$ is given by (3), and $p(z|w, k)$ by (2).

For full flexibility, we now add an extra layer to the hierarchy, and allow the priors for k, w and θ to depend on hyperparameters λ, δ and η respectively. These will be drawn from independent hyperpriors. The joint distribution of all variables is then expressed by the factorisation

$$p(\lambda, \delta, \eta, k, w, z, \theta, y) = p(\lambda)p(\delta)p(\eta)p(k|\lambda)p(w|k, \delta)p(z|w, k)p(\theta|k, \eta)p(y|\theta, z). \tag{5}$$

2.3 Normal mixtures

From now on, our detailed exposition is limited to the case of univariate normal mixtures. However, the methodology is generic, and applies much more widely. Some specific generalisations, several of which are already implemented, are described in Section 8.4.

In the univariate normal case, the generic parameter θ is a vector of (expectation, variance) pairs $(\mu_j, \sigma_j^2), j = 1, 2, \ldots, k$, so that

$$f(y|\theta_j) = f(y|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{\frac{-(y-\mu_j)^2}{2\sigma_j^2}\right\}$$

Our prior distributions are that the μ_j and σ_j^{-2} are all drawn independently, with normal and gamma priors

$$\mu_j \sim N(\xi, \kappa^{-1}) \qquad \text{and} \qquad \sigma_j^{-2} \sim \Gamma(\alpha, \beta)$$

(in the latter, choosing the parameterisation in which the mean and variance are α/β and α/β^2 respectively), so that the generic η has become $(\xi, \kappa, \alpha, \beta)$. These are fairly natural choices of prior distributions, giving some of the advantages of conjugacy, advantages that are not actually needed when using MCMC computation. It is not the "natural conjugate" prior for (μ_j, σ_j^2), under which the parameters within each pair are *a priori* dependent.

We now come to the important issue of labelling the components. Note that our whole model is invariant to permutation of the labels $j = 1, 2, \ldots, k$. For identifiability, it is important to adopt a unique labelling. Unless stated otherwise, we use that in which the μ_j are in increasing numerical order; thus the joint prior distribution of the parameters is $k!$ times the product of the individual normal and gamma densities, restricted to the set $\mu_1 < \mu_2 < \cdots < \mu_k$.

The prior on w will always be taken as symmetric Dirichlet, $w \sim D(\delta, \delta, \ldots, \delta)$. It is necessary to adopt a proper prior distribution for k and a common choice is the Poisson with hyperparameter λ. For convenience of presentation and interpretation, we instead use a uniform distribution between 1 and a pre-specified integer $k_{\max}$, the choice of which is discussed when we come to our experiments.

2.4 Weak prior information for component parameters

In this paper, we only consider Bayesian mixture estimation in the set-up where one does not have (or want to use) strong prior information on the mixture parameters. There are cases where subjective priors are preferable, and our prior setting could be modified accordingly.

Being fully non-informative and obtaining proper posterior distributions is not possible in a mixture context. Since there is always the possibility that no observations are allocated to one or more components, and so the data are uninformative about them, standard choices of *independent* improper non-informative prior distributions for the component parameters cannot be used (Diebolt and Robert, 1994; Roeder and Wasserman, 1995). Some previous attempts to circumvent this problem, which involve dependent priors, are mentioned in Section 8.3.

It seems to us that for most purposes of mixture modelling, there is a case for keeping to the simple independence prior structure for the μ_j and σ_j^{-2} that we have outlined in Section 2.3 and defining *weakly informative priors*, which may or may not be data-dependent, a line also taken by Raftery (1996) and Nobile (1994). If, for example, one is interested in identifying subpopulations, there would be *a priori* information which can be translated into, say, a likely range for their spread. We thus introduce a hyperprior structure and default hyperparameter choices which correspond to making only "minimal" assumptions on the data.

It seems natural to take the $N(\xi, \kappa^{-1})$ prior for μ_j to be rather flat over an interval of variation of the data, either postulated *a priori*, or corresponding to the observed range. This can be achieved

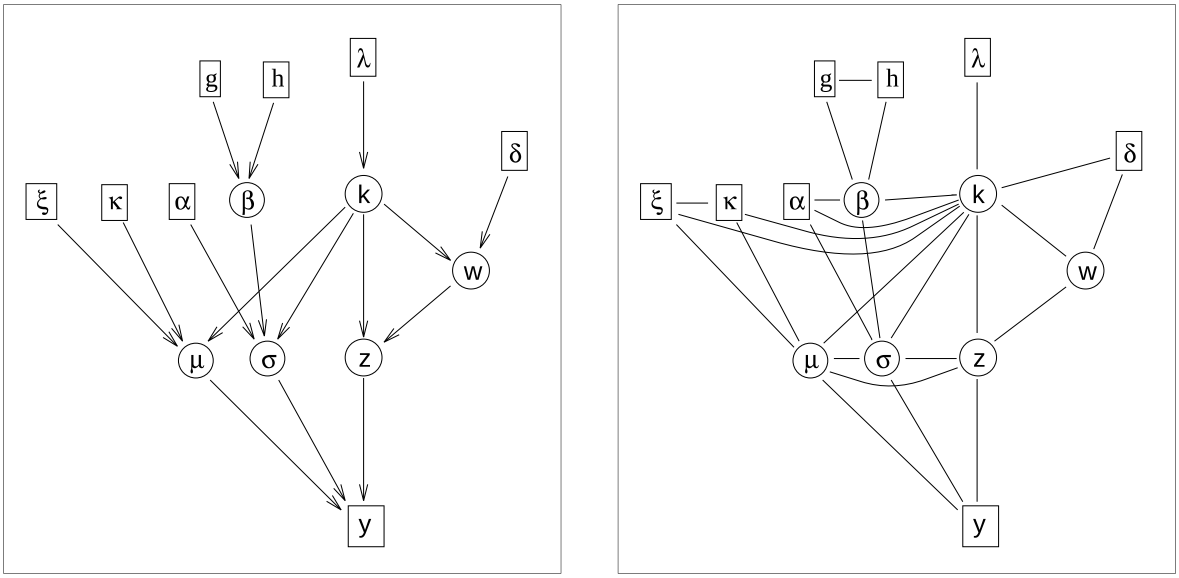

Figure 1: (a) the DAG specific to the normal mixture model implemented in this paper; (b) the corresponding conditional independence graph

in a simple way by letting ξ equal to the mid-point of this interval, and setting κ equal to a small multiple of $1/R^2$, where R is the length of the interval.

In contrast to the case of the means μ_j, it seems restrictive to suppose that knowledge of the range of the data implies much about the size of the σ_j^2; recall that $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$, independently. We therefore introduce an additional hierarchical level by allowing β to follow a Gamma distribution with parameters g and h. We will generally take $\alpha > 1 > g$ to express the belief that the σ_j^2 are similar, without being informative about their absolute size. The scale parameter h will be a small multiple of $1/R^2$.

Finally, λ and δ are held fixed in this paper.

The complete hierarchical model, which we call the random β model, is displayed as a directed acyclic graph (DAG) in Figure 1, with the usual convention of graphical models that square boxes represent fixed or observed quantities, and circles the unknowns.

3 A reversible jump MCMC algorithm for mixtures

3.1 MCMC algorithms for variable-dimension parameters

Markov chain Monte Carlo (MCMC) methods play a central role in modern Bayesian computation; see, for example, Tierney (1994), Besag, Green, Higdon and Mengersen (1995). Such methods were initially only available for problems where the posterior distribution has a density with respect to some fixed standard underlying measure, and so could not be used in cases, such as mixture estimation, where "the number of things you don't know is one of the things you don't know". Recently, MCMC methods for varying-dimension problems have been discussed (Grenander and Miller, 1994; Green, 1994; Phillips and Smith, 1996). One approach, termed reversible jump MCMC, is elaborated in Green (1995), including applications to change-point analysis in one and two dimensions, and to partition problems arising in Bayesian analysis of factorial experiments. In brief, reversible jump MCMC is a random-sweep Metropolis-Hastings method (Metropolis, *et al.*, 1953; Hastings, 1970) adapted for general state spaces. Letting x denote the state variable (in our application x is the complete set of unknowns $(\beta, \mu, \sigma, k, w, z)$), and $\pi(dx)$ the target probability measure (the posterior distribution), we consider a countable family of move types, indexed by $m = 1, 2, \ldots$. When the current state is x, a move type m and destination x' are proposed, with joint distribution given by an essentially arbitrary subprobability measure $q_m(x, dx')$. (With probability $\sum_m \int_{x'} q_m(x, dx')$, no move is attempted.) The move is accepted with probability

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')q_m(x', dx)}{\pi(dx)q_m(x, dx')}\right\}, \tag{6}$$

where the ratio of measures can be given a rigorous definition as a ratio of Radon-Nikodym derivatives with respect to a suitably-chosen common dominating measure. The existence of such a measure is ensured by a "dimension-balancing" condition on the $q_m(x, dx')$, that effectively matches the degrees of freedom of joint variation of the state and proposal as the dimension changes with k.

For a move type that does not change the dimension of the parameter, this rather abstract expression reduces to the familiar Metropolis-Hastings acceptance probability, using an ordinary ratio of densities (Hastings, 1970; Peskun, 1973); for dimension-changing moves, a little more care is needed. However, in a typical case, a more concrete form can be given. Suppose that a move of type m is proposed, from x to a point x' in a higher-dimensional space. This will very often be implemented by drawing a vector of continuous random variables u, independent of x, and setting x' using an invertible deterministic function $x'(x, u)$. The reverse of the move (from x' to x) can be accomplished by using the inverse transformation, so that the proposal is deterministic. Then the acceptance probability (6) reduces to

$$\min\left\{1, \frac{p(x'|y)r_m(x')}{p(x|y)r_m(x)q(u)}\left|\frac{\partial x'}{\partial(x, u)}\right|\right\}, \tag{7}$$

where $r_m(x)$ is the probability of choosing move type m when in state x, and $q(u)$ is the density function of u. Note that the final term in the ratio above is a Jacobian arising from the change of variable from (x, u) to x'.

3.2 Reversible jump moves for normal mixtures

Reversible jump MCMC is most simply derived mathematically in its random-scan form, but as usual with Metropolis-Hastings methods, the idea is equally valid when the available moves are scanned systematically, and that is the approach we have chosen to take here.

For our hierarchical normal mixture model, we will make use of 6 move types:

(a) updating the weights w;

(b) updating the parameters (μ, σ);

(c) updating the allocation z;

(d) updating the hyperparameter β;

(e) splitting one mixture component into two, or combining two into one;

(f) the birth or death of an empty component.

Moves (e) and (f) involve changing k by 1, and making necessary corresponding changes to (μ, σ, w, z).

The only randomness in the scanning is the random choice between splitting and combining in move (e), or birth and death in move (f). One complete pass over these six moves will be called a *sweep* and is the basic time step of the algorithm.

All MCMC algorithms make use of the full conditional distributions of some variables given all others, and in deriving these it is helpful to consult the conditional independence graph of the system, derived from the DAG by "moralising" and dropping the arrows (Lauritzen and Spiegelhalter, 1988). For our model this graph is displayed in Figure 1, from which we note that, given all other variables, w and (μ, σ) are conditionally independent, and similarly for z and β. Moves (a) and (b) can therefore be performed in parallel, and so can (c) and (d).

Move types (a), (b), (c) and (d) are conventional, largely following Diebolt and Robert (1994); they do not alter the dimension of the complete parameter vector $(\beta, \mu, \sigma, k, w, z)$, and we will not give much detail about them here. Through conjugacy, the full conditional distribution for the weights w remains Dirichlet in form:

$$w|\cdots \sim D(\delta + n_1, \ldots, \delta + n_k),$$

where $n_j = \#\{i : z_i = j\}$, and here and later we use "$|\cdots$" to denote conditioning on all other variables. Thus w can be updated by a Gibbs move, sampling from this full conditional by drawing independent Gamma random variables, and scaling them to sum to 1.

The full conditionals for $\{\mu_j\}$ are

$$\mu_j|\cdots \sim N\left(\frac{\sigma_j^{-2}\sum_{i:z_i=j} y_i + \kappa\xi}{\sigma_j^{-2} n_j + \kappa}, (\sigma_j^{-2} n_j + \kappa)^{-1}\right).$$

In order to preserve the ordering constraint on the $\{\mu_j\}$, the full conditional is used only to generate a proposal, and is accepted providing the ordering is unchanged.

The full conditionals for $\{\sigma_j^2\}$ are

$$\sigma_j^{-2}|\cdots \sim \Gamma(\alpha + \tfrac{1}{2} n_j, \beta + \tfrac{1}{2}\sum_{i:z_i=j}(y_i - \mu_j)^2),$$

for the allocation variables we have

$$p(z_i = j|\cdots) \propto \frac{w_j}{\sigma_j}\exp\left(-\frac{(y_i - \mu_j)^2}{2\sigma_j^2}\right), \tag{8}$$

while the only hyperparameter we are not treating as fixed, β, has a Gamma distribution

$$\beta|\cdots \sim \Gamma(g + k\alpha, h + \sum_j \sigma_j^{-2}).$$

For all of these variables, we use a Gibbs kernel.

For the split/combine move (e), the reversible jump mechanism is needed. Recall that we need to design these moves in tandem: they form a reversible pair. The strategy is to choose the proposal distributions according to informal considerations suggesting a reasonable probability of acceptance, but strictly subject to the requirement of dimension-matching. Having done so, conformation with the detailed balance condition is determined by the acceptance probability (7). This is the point at which the statistical model is used quantitatively.

In move (e), we make a random choice between attempting to split or combine, with probabilities b_k and $d_k = 1 - b_k$, respectively, depending on k. Of course, $d_1 = 0$ and $b_{k_{\max}} = 0$, where $k_{\max}$ is the maximum value allowed for k, and otherwise we choose $b_k = d_k = 0.5$, for $k = 2, 3, \ldots, k_{\max} - 1$. Our combine proposal begins by choosing a pair of components (j_1, j_2) at random, that are adjacent in terms of the current value of their means, that is

$$\mu_{j_1} < \mu_{j_2}, \quad \text{with no other } \mu_j \text{ in the interval } [\mu_{j_1}, \mu_{j_2}]. \tag{9}$$

These two components are merged, reducing k by 1. In doing so, forming a new component here labelled j^*, we have to reallocate all those observations y_i with $z_i = j_1$ or j_2, and create values for $(w_{j^*}, \mu_{j^*}, \sigma_{j^*})$. The reallocation is simply done by setting such $z_i = j^*$, while the other parameters are assigned by the expedient of matching the $0^{\text{th}}, 1^{\text{st}}$ and 2^{nd} moments of the new component to those of a combination of the two that it replaces:

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2} \\ w_{j^*}\mu_{j^*} &= w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2} \\ w_{j^*}(\mu_{j^*}^2 + \sigma_{j^*}^2) &= w_{j_1}(\mu_{j_1}^2 + \sigma_{j_1}^2) + w_{j_2}(\mu_{j_2}^2 + \sigma_{j_2}^2) \end{aligned} \tag{10}$$

This combine proposal is deterministic once the discrete choices of j_1 and j_2 have been made, so the expression (7) for the acceptance probability will be relevant.

The reverse split proposal is now largely determined. A component j^* is chosen at random, and split into two, labelled j_1 and j_2, with weights and parameters conforming to (10). There are three degrees of freedom in achieving this, so we need to generate a three-dimensional random vector u to specify the new parameters. We use Beta distributions

$$u_1 \sim Be(2,2), u_2 \sim Be(2,2), u_3 \sim Be(1,1)$$

for this, and set

$$\begin{aligned} w_{j_1} &= w_{j^*}u_1 \quad , \quad w_{j_2} = w_{j^*}(1-u_1) \\ \mu_{j_1} &= \mu_{j^*} - u_2\sigma_{j^*}\sqrt{\frac{w_{j_2}}{w_{j_1}}} \quad , \quad \mu_{j_2} = \mu_{j^*} + u_2\sigma_{j^*}\sqrt{\frac{w_{j_1}}{w_{j_2}}} \\ \sigma_{j_1}^2 &= u_3(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_1}} \quad , \quad \sigma_{j_2}^2 = (1-u_3)(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_2}}, \end{aligned}$$

which provide all 6 required weights and parameters, satisfying (10). It can be readily shown that these are indeed valid, with weights and variances positive. At this point, we check whether the adjacency condition (9) is satisfied. If not, the move is rejected forthwith, as the (split, combine) pair could not then be reversible. If the test is passed, it remains only to propose the reallocation of those y_i with $z_i = j^*$ between j_1 and j_2. This is done analogously to the standard Gibbs allocation move; see equation (8).

The acceptance probabilities for the split/combine moves, calculated from (7), have quite convoluted form. For the split move the probability is $\min(1, A)$, where

$$\begin{aligned} A &= (\text{likelihood ratio}) \times \frac{p(k+1)}{p(k)} \times (k+1) \times \frac{w_{j_1}^{\delta-1+l_1} w_{j_2}^{\delta-1+l_2}}{w_{j^*}^{\delta-1+l_1+l_2} B(\delta, k\delta)} \\ &\times \sqrt{\frac{\kappa}{2\pi}} \exp[-\tfrac{1}{2}\kappa\{(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 - (\mu_{j^*}-\xi)^2\}] \\ &\times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{\sigma_{j_1}^2\sigma_{j_2}^2}{\sigma_{j^*}^2}\right)^{-\alpha-1} \exp\left(-\beta(\sigma_{j_1}^{-2} + \sigma_{j_2}^{-2} - \sigma_{j^*}^{-2})\right) \\ &\times \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \{g_{2,2}(u_1)g_{2,2}(u_2)g_{1,1}(u_3)\}^{-1} \\ &\times \frac{w_{j^*}|\mu_{j_1}-\mu_{j_2}|\sigma_{j_1}^2\sigma_{j_2}^2}{u_2(1-u_2^2)u_3(1-u_3)\sigma_{j^*}^2}, \end{aligned} \tag{11}$$

where k is the number of components before the split, l_1 and l_2 are the numbers of observations proposed to be assigned to j_1 and j_2, $B(\cdot,\cdot)$ is the Beta function, P_{alloc} is the probability that this particular allocation is made, $g_{p,q}$ denotes the Beta(p,q) density, and (likelihood ratio) is the ratio of the product of the $f(y_i|\theta_{z_i})$ terms for the new parameter set to that for the old. For the corresponding combine move, the acceptance probability is $\min(1, A^{-1})$, using the same expression for A but with some obvious differences in the substitutions.

The correspondence between (7) and (11) is fairly straightforward; the first three lines of (11) form the ratio $p(x'|y)/p(x|y)$, the $(k+1)$ factor in the first line being the ratio $(k+1)!/k!$ from the order statistics densities for the parameters (μ, σ^2). The fourth line is the proposal ratio $r_m(x')/(r_m(x)q(u))$, and the final line is the Jacobian of the transformation from $(w_{j^*}, \mu_{j^*}, \sigma^2_{j^*}, u_1, u_2, u_3)$ to $(w_{j_1}, \mu_{j_1}, \sigma^2_{j_1}, w_{j_2}, \mu_{j_2}, \sigma^2_{j_2})$.

The birth and death move (f) is somewhat simpler. We first make a random choice between birth and death, using the same probabilities b_k and d_k as above. For a birth, a weight and parameters for the proposed new component are drawn using

$$w_{j^*} \sim Be(1,k), \qquad \mu_{j^*} \sim N(\xi, \kappa^{-1}) \qquad \text{and} \qquad \sigma_{j^*}^{-2} \sim \Gamma(\alpha, \beta).$$

To "make space" for the new component, the existing weights are re-scaled, so that all weights sum to 1, using $w'_j = w_j(1 - w_{j^*})$. For a death, a random choice is made between any existing empty components, the chosen component is deleted, and the remaining weights are re-scaled to sum to 1. No other changes are proposed to the variables: in particular, the allocations are unaltered.

Detailed balance holds for this move, provided we accept births and deaths according to (7), in which $(w_{j^*}, \mu_{j^*}, \sigma^2_{j^*})$ play the role of u. The use of the prior distributions in proposing values for μ_{j^*} and $\sigma^2_{j^*}$ leads to simplification of the resulting ratio. The acceptance probabilities for birth and death are $\min(1, A)$ and $\min(1, A^{-1})$ respectively, where

$$\begin{aligned} A &= \frac{p(k+1)}{p(k)} \frac{1}{B(k\delta, \delta)} w_{j^*}^{\delta-1} (1 - w_{j^*})^{n+k\delta-k} (k+1) \\ &\times \frac{d_{k+1}}{(k_0+1)b_k} \frac{1}{g_{1,k}(w_{j^*})} (1 - w_{j^*})^k. \end{aligned} \tag{12}$$

Here, k is the number of components and k_0 the number of empty components, before the birth. In (12), the first line is the prior ratio, and the second line contains the proposal ratio and Jacobian; the likelihood ratio is 1.

This completes the specification of the moves, which we do not claim is optimal; indeed, this is almost the first scheme that we tried. The validity of the algorithm is not compromised by the choice of proposals, since detailed balance is confirmed by using (7). In Metropolis-Hastings methods, it is rarely worth fine-tuning the proposal distribution, especially if doing so prevents simple and explicit random variate generation.

With detailed balance satisfied, it only remains to check that the Markov chain defined is irreducible and aperiodic. Aperiodicity is clear, since given any arbitrarily small neighbourhood of a current state $(\beta, \mu, \sigma, k, w, z)$ there is positive probability that after one sweep through moves (a) to (f) the chain lies in that neighbourhood. Irreducibility is also easily established, since the chain can move from any value of k to any other value in steps of one at a time, in move (c) all allocations have positive probability, and the parameters and hyperparameters are updated by drawing from continuous distributions whose supports are the natural parameter spaces.

4 Statistical performance of the proposed methodology

There are several inter-linked aspects of the proposed methodology to be demonstrated in illustrating its performance. Of course, we display examples of the results we obtain from real data sets. But presentation of substantive results is inevitably associated with questions of sensitivity to model specification, especially regarding the prior, and questions about the performance of the MCMC sampling method, that is how well it "mixes". These three aspects – results, sensitivity and mixing – interact closely. To avoid circularity in our presentation, we postpone discussion of sensitivity and mixing to Sections 5 and 6 respectively, and first present a description of the performance of the model itself with default settings for the hyperparameters. Of course, sensitivity issues informed our choices here.

4.1 Results for three data sets

Three real data sets are used throughout the paper, as a basis for our comparisons[1]. The first concerns the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of 245 unrelated individuals. The interest here is in identifying subgroups of slow or fast metabolisers as a marker of genetic polymorphism in the general population. This data set has been analysed by Bechtel *et al.* (1993), who identified a mixture of 2 skewed distributions using maximum likelihood techniques implemented in the program SKUMIX of Maclean *et al.* (1976). We shall refer to this data set as the 'Enzyme data'. The second data set, the 'Acidity data', concerns an acidity index measured in a sample of 155 lakes in the Northeastern United States and has been previously analysed as a mixture of gaussian distributions on the log scale by Crawford *et al.*(1992, 1994); we also use the log scale. The last data set, the 'Galaxy data', was first described in Roeder (1990), and subsequently analysed under different mixture models by several authors including Escobar and West (1995) and Phillips and Smith (1996). It consists of the velocities of 82 distant galaxies, diverging from our own galaxy. Histograms of the three data sets are shown in Figure 3.

The three data sets have been analysed with the hierarchical normal random β mixture model defined in Sections 2.3 and 2.4, with the following settings for previously unspecified constants: $\kappa = 1/R^2$, $\alpha = 2$, $g = 0.2$, $h = 10/R^2$ and $\delta = 1$. The prior on k is taken as uniform on the integers $1, 2, \ldots, k_{\max} = 30$, for which it is particularly easy to convert results to those corresponding to other priors on these values, using the identity

$$p^*(k, \theta^{(k)}|y) \propto p(k, \theta^{(k)}|y) \frac{p^*(k)}{p(k)},$$

where $p^*(\cdot|y)$ denotes the posterior for an alternative prior p^*.

For each of the three data sets, we report results corresponding to 100 000 sweeps, following a burn-in period also of 100 000 sweeps. We believe that these numbers exceed what is needed for reliable results. In all the runs, the number of components never exceeded 24, hence the chosen value of $k_{\max}$ was inconsequential.

Estimated posterior probabilities are given in Table 1. In each of the data sets, it is immediately apparent that there are a number of competing explanations of the data which are tenable. For the Enzyme data, the posterior for k favours 3 to 5 components. In this example, with the proviso that there is some prior evidence for enzyme level to be normally distributed, we could interpret the existence of three components in the mixture in terms of a simple underlying genetic model. For the Acidity data, there is again fairly equal support for 3 to 5 components; for the Galaxy data, the posterior for k is more widely spread and indicates a higher number of components, between 5 and 7. In each case, the high overall number of components can be related in part to the skewness of the data, two or three normals being sometimes needed to fit one skewed component, but also to our mixture model which imposes little structure *a priori*, in contrast to those considered by Roeder and Wasserman (1995) or Gruet *et al.*(1996).

[1] All three data sets can be obtained from the world wide web at `http://www.stats.bris.ac.uk/~peter/mixdata`

Table 1: Posterior distribution of k for the 3 data sets based on a mixture model with random β and default* parameter values.

Data set	n	$p(k\|y)$	Proportion *split/ combine* moves accepted	*birth/ death* moves accepted
Enzyme	245	$p(1)$=0.000 $p(2)$=0.024 $p(3)$=0.290 $p(4)$=0.317 $p(5)$=0.206 $p(6)$=0.095 $p(7)$=0.041 $p(8)$=0.017 $p(9)$=0.007 $p(10)$=0.002 $\sum_{k\geq 11} p(k)$=0.001	8%	4%
Acidity	155	$p(1)$=0.000 $p(2)$=0.082 $p(3)$=0.244 $p(4)$=0.236 $p(5)$=0.172 $p(6)$=0.118 $p(7)$=0.069 $p(8)$=0.037 $p(9)$=0.020 $p(10)$=0.011 $p(11)$=0.006 $p(12)$=0.003 $p(13)$=0.001 $\sum_{k\geq 14} p(k)$=0.001	14%	7%
Galaxy	82	$p(1)$=0.000 $p(2)$=0.000 $p(3)$=0.061 $p(4)$=0.128 $p(5)$=0.182 $p(6)$=0.199 $p(7)$=0.160 $p(8)$=0.109 $p(9)$=0.071 $p(10)$=0.040 $p(11)$=0.023 $p(12)$=0.013 $p(13)$=0.006 $p(14)$=0.003 $p(15)$=0.002 $\sum_{k\geq 16} p(k)$=0.003	11%	18%

* Range and default parameter values:
Enzyme data : $R = 2.86$, $\xi = 1.45$, $\kappa = 0.122$, $\alpha = 2$, $g = 0.2$, $h = 1.22$, $\delta = 1$
Acidity data : $R = 4.18$, $\xi = 5.02$, $\kappa = 0.057$, $\alpha = 2$, $g = 0.2$, $h = 0.573$, $\delta = 1$
Galaxy data : $R = 25.11$, $\xi = 21.73$, $\kappa = 0.0016$, $\alpha = 2$, $g = 0.2$, $h = 0.016$, $\delta = 1$

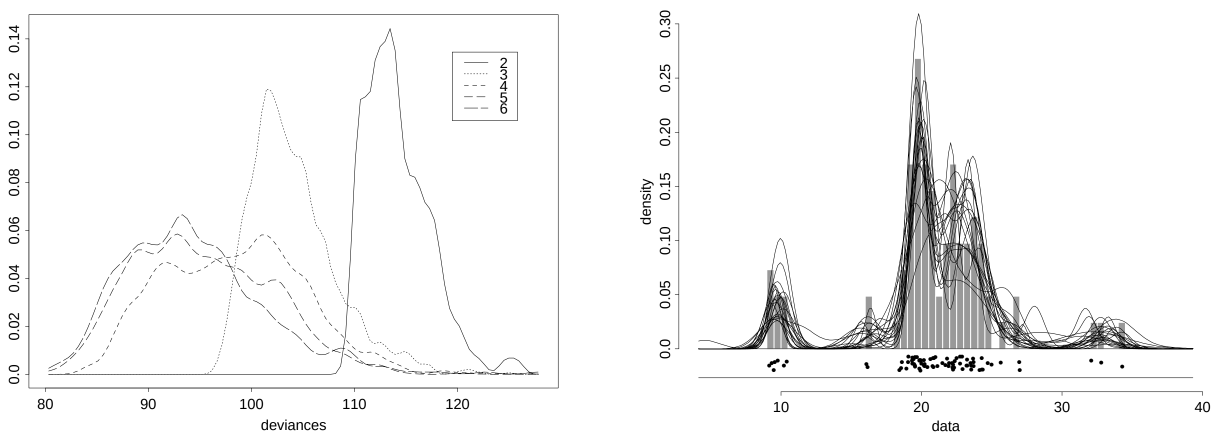

Figure 2: (a) Posterior distributions of deviances ($k = 2$ to 6) for the Enzyme data; (b) sample from the posterior distribution of $f(\cdot|k, w, \theta)$ for the Galaxy data.

As a by product of our implementation, we can also investigate changes in the posterior distribution of "deviances" $-2\log p(y|k,w,\theta)$ for increasing k. For the Enzyme data, there is a marked shift between $k=2$ and 3, whereas from $k=3$ onwards, there is substantial overlap between the deviance distributions; see Figure 2(a). A similar pattern emerges for the other two data sets, with substantial overlap from $k=3$ for the Acidity data and from $k=4$ for the Galaxy data.

4.2 Predictive densities

At each sweep of the algorithm, values (w,θ) for the weights and parameters are produced, from which densities $f(\cdot|k,w,\theta) = \sum_{j=1}^{k} w_j f(\cdot|\theta_j)$ can be computed. Posterior variation among the realised $f(\cdot|k,w,\theta)$ is displayed in Figure 2(b), for the Enzyme data.

Averaging the $f(\cdot|k,w,\theta)$ across the MCMC run, conditional on fixed values of k, gives an estimate of $E(f(\cdot|k,w,\theta)|k,y)$, a Bayesian predictive density estimate of the mixture with k components. Averaging further across values of k gives an estimate of $E(f(\cdot|k,w,\theta)|y)$, the 'overall' Bayesian predictive density estimate of the distribution of y. Note further that these density estimates *do not themselves have the shape of a finite mixture of distributions.* Other density estimates, in particular the mixtures $f(\cdot|k,\hat{w},\hat{\theta})$ for each k, have been considered, where $\hat{w}$ and $\hat{\theta}$ are summary posterior estimates of the weights and parameters of a k components mixture. In the case where the posterior distribution of (w,θ) is fairly spread out or even multimodal, these plug-in estimates would give a poor, oversmooth approximation of the predictive density.

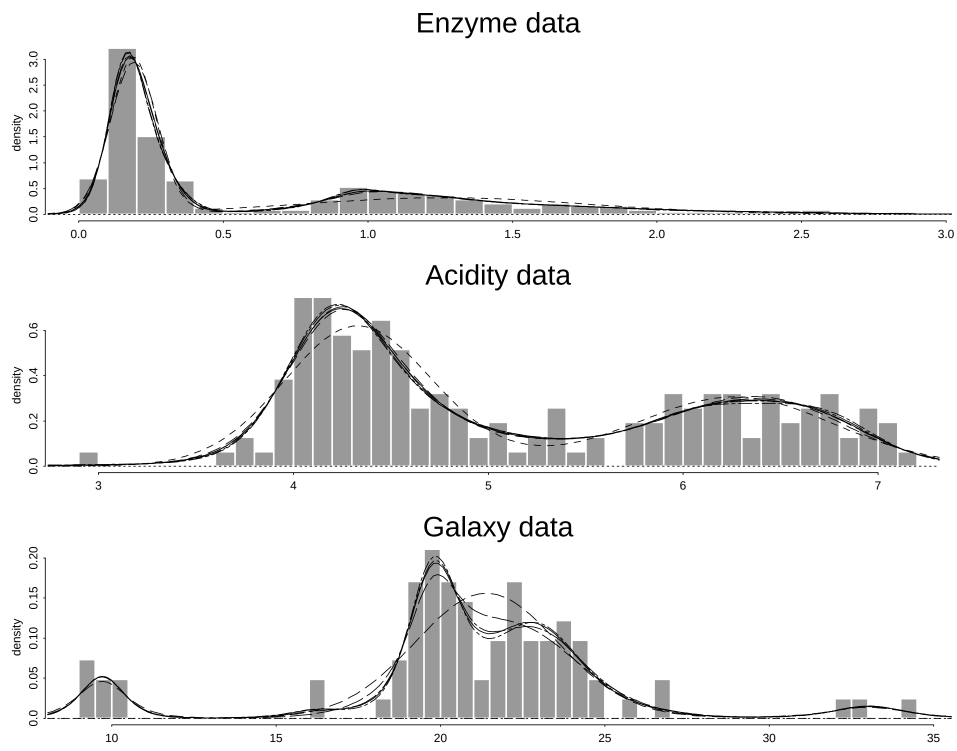

Figure 3: Predictive densities for the 3 data sets, unconditionally (full line), and conditional on various values of k (dotted lines); the curves displayed are for $k=2$ to 6, except for the Galaxy data, where they are for $k=3$ to 6. In each case note that it is only the smallest k shown that gives an appreciably different estimate.

Predictive densities, both conditional on k and unconditional are shown in Figure 3 for the three data sets. Note that the difference between successive predictive densities decreases with increasing values of k and that the overall unconditional plot gives a convincing density estimate of the data distribution. Predictive fit and posterior distribution give complementary evidence on which to draw when assessing the number of components. For the Enzyme data, the predictive plots for 3 or 4 components are very similar. Since one aim of the analysis of this data set is the identification of interpretable subpopulations, we would favour the mixture with 3 components.

4.3 Parameter estimates

4.3.1 Labelling and postprocessing the MCMC output

Although it is in some ways natural to consider the parameters as a set, labelling at each sweep is convenient and becomes necessary when density estimates or other summaries of the posterior distribution of the *parameters* of each component are required. The most appropriate labelling will depend on the example analysed and it is a substantial bonus of the sample-based computation method we use that this can be investigated *after* the run of the algorithm.

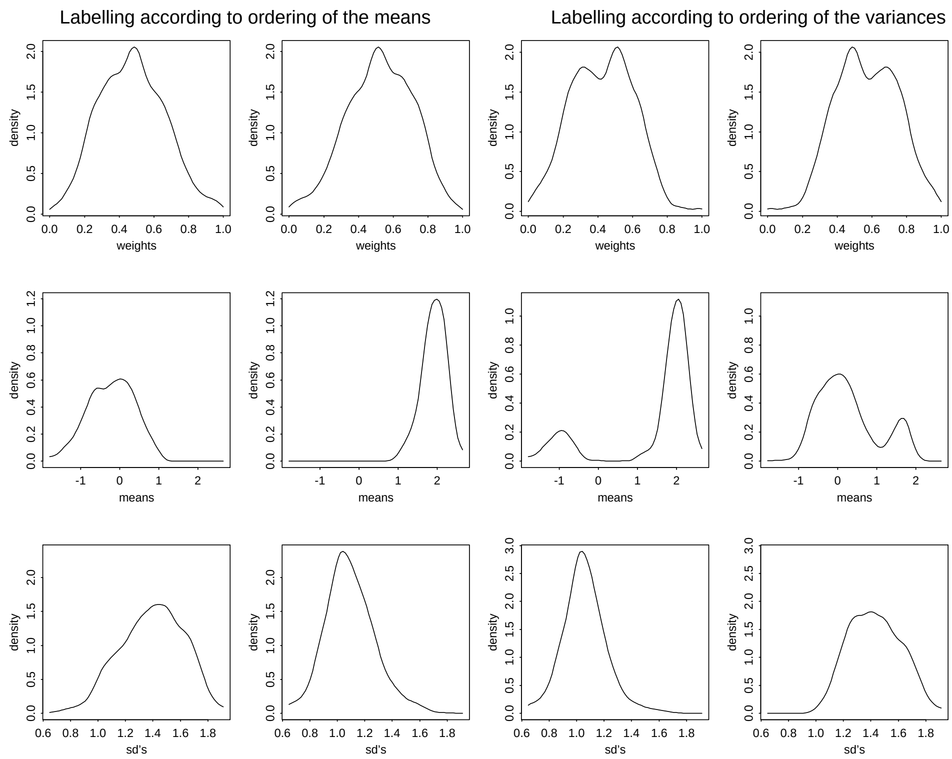

Figure 4: Posterior density of parameter estimates for 2 choices of labelling, simulated data set.

To understand why the issue of labelling is not straightforward, consider the case where the population is really of two normal components, unambiguously labelled. Given a finite sample, the posterior distribution of the two means will overlap, and similarly for the weights and variances; the extent of the overlaps depends on the separation and the sample size. When the means are well separated, labelling of the realisations from the posterior by ordering their means will generally coincide with the population labelling; as the separation reduces, so called 'label-switching' will occur; see also Mengersen and Robert (1996). Depending on the relative separations, label-switching can be minimised by choosing to order on the variances, weights, or some combination of all three parameters.

We illustrate these points in Figure 4 using a simulated data set of $n = 250$ points, drawn from a mixture which gives a skewed unimodal distribution ($w_1 = w_2 = 0.5,\ \mu_1 = 0.0,\ \mu_2 = 1.0,\ \sigma_1 = 1.5,\ \sigma_2 = 1.0$). Figure 4 displays the posterior densities of w_j, μ_j, σ_j for the runs where $k = 2$, on the left hand side for a labelling corresponding to an ordering of the means, and on the right hand side for a labelling corresponding to an ordering of the variances. The labelling according to the variances (right hand side) leads to bimodal densities for $\mu_j,\ j = 1, 2$, which corresponds to label switching for about half the runs. Labelling by ordering the means gives clearer unimodal plots *simultaneously* for all three parameters, with still some evidence of switching. In a real data set, there might not be an obvious choice of labelling. It is then advisable to postprocess the run according to different choices of labels in order to get the clearest picture of the component parameters.

4.3.2 Multimodality of the parameter densities

Quite apart from the labelling problem, there might be cases of genuine multimodality of the posterior distribution of parameter estimates corresponding to different mixture models competing for potential explanations of the data set.

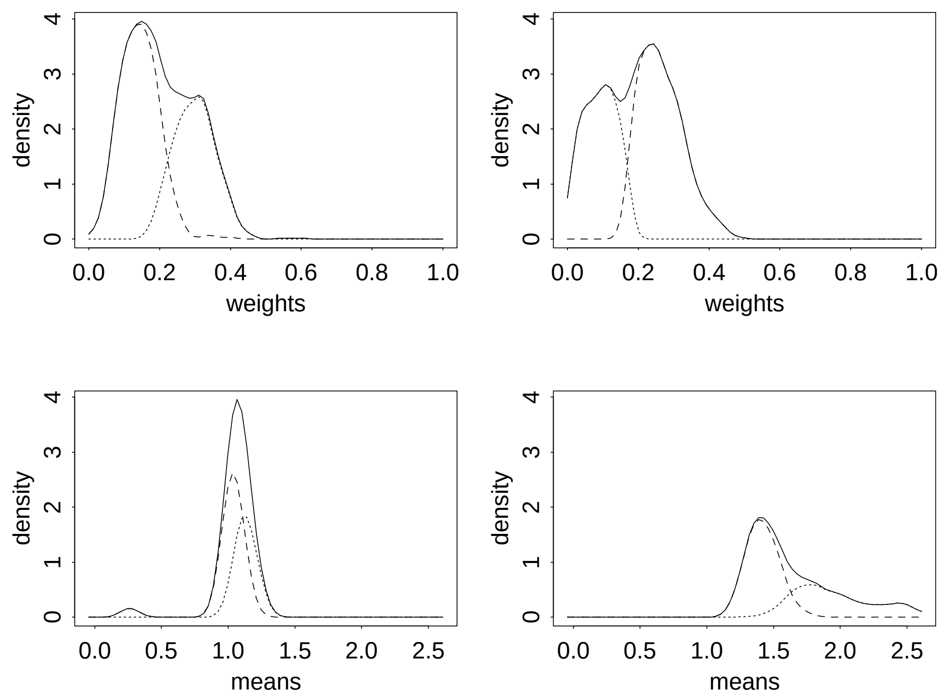

Figure 5: Enzyme data set: posterior densities of weights and means for the second and third component, default prior model, conditioning on $k = 3$ (full line), and conditioning also on $w_3 \leq 0.17$ (dotted line) and on $w_3 > 0.17$ (broken line). (The last two have areas proportional to posterior probability.)

As an example we display in Figure 5 (full line) the posterior densities for the weights and the means of the second and third component for the Enzyme data (with labels according to the ordering of the means). There is some evidence of bimodality in the distributions of the weights for the second and third components. Different labelling does not help to clarify the picture. In an attempt to exhibit possible competing explanations, we separate out in the MCMC output those runs corresponding to $w_3 \leq 0.17$ and plot the parameter densities again (dotted lines in Figure 5).

This produces more clearly peaked posterior densities for all the parameters, and show that low values of w_3 are associated with elevated values of μ_3. In fact, a further separation according to $w_3 \leq 0.05$ would show that this small group corresponds to $\mu_3 \geq 2.0$, indicating that only a small fraction of individuals have high enzymatic activity, as expected.

5 Sensitivity of results to prior assumptions

Our hierarchical approach to mixture modelling involves hyperparameter specifications. We have carried out an in-depth study of their influence which has led us to make the standard default recommendations that we have used in the previous section. In this section, we highlight some important aspects of our sensitivity analysis. We emphasise that we do not recommend that such a study should be performed for a standard implementation of our approach!

5.1 Sensitivity of the posterior distribution of k

5.1.1 Comparison of prior models for the variance: fixed versus random β

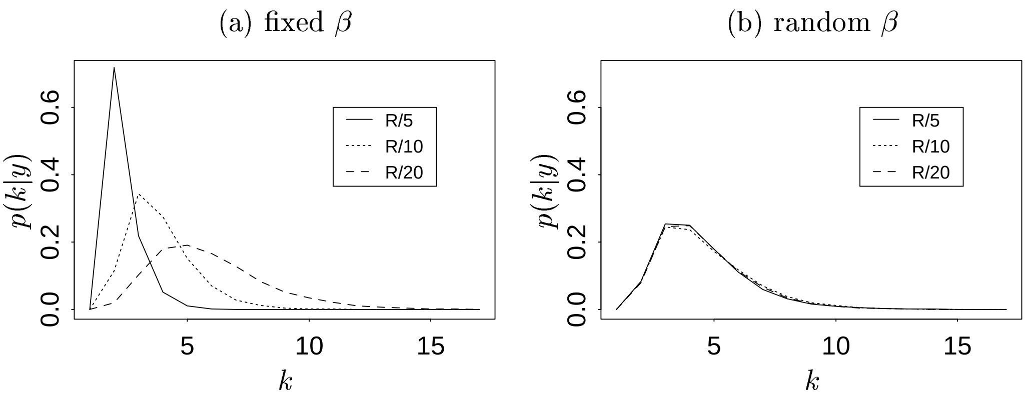

Figure 6: Posterior distributions of k: comparison of sensitivity to hyperparameters between fixed and random β models: (a) $\alpha = 2$, $\sqrt{(\beta/\alpha)}$ varying between $R/5$ and $R/20$, (b) $\alpha = 2$, $g = 0.2$, $\sqrt{(g/h\alpha)}$ varying between $R/5$ and $R/20$.

Our main concern here is to show how, in our model, the number of components is related to the prior information on the variances σ_j^2. In the standard non-hierarchical model with fixed β and α the mean of the gamma distribution, α/β, specifies the typical value of the precision σ_j^{-2}. It is natural to relate σ_j to the range R of the data and increasing values for $\sqrt{(\beta/\alpha)}$ will lead to models with fewer components. In Figure 6(a) we show, on the Acidity data, the substantial change in the posterior distribution of k as $\sqrt{(\beta/\alpha)}$ is varied between $R/5$, $R/10$ and $R/20$. There is little overlap between the posterior distributions of k for the two extreme cases. Hence, in the standard model, the choices of α and β will crucially influence the posterior distribution of k and it is difficult to be weakly informative.

In contrast, the hierarchical model with fixed α but random β that we have implemented, which allows weak information on σ_j to be put in at a higher level, does not exhibit the same behaviour. The posterior distribution of k is quite insensitive over a wide range of values of the ratio g/h (related again to the range R) which all lead to similar posterior mean and standard deviation for β. This is well illustrated in Figure 6(b) which shows strikingly similar posterior distributions for k for three sets of values for α, g and h chosen so that the prior order of magnitude for σ_j *at the higher level*, $\sqrt{(g/h\alpha)}$, ranges again from $R/5$ to $R/20$. Hence our hierarchical formulation for the

Table 2: Influence of prior distribution N(ξ,κ^{-1}) for μ on the posterior distribution of k. Acidity data: mixture model with Poisson (prior $\mathcal{P}(10)$ for k), random β and default* parameter values.

$\kappa^{-1/2}$	R	R/2	R/3	R/4	R/5	R/8	R/10
range of k with							
$p(k\|y) \geq 0.05$	[3–9]	[5–12]	[6–14]	[7–14]	[7–14]	[5–12]	[4–11]
$p(k\|y) \geq 0.001$	[2–13]	[3–17]	[4–19]	[3–20]	[4–20]	[3–17]	[3–16]
k with highest $p(k)$	6	8	9	10	10	8	7

* $\sigma_j^{-2} \sim \Gamma(2, \beta)$, $\beta \sim \Gamma(0.2, h)$ with $\sqrt{(g/h\alpha)} = R/10$

variance distribution in the mixture model allows a high degree of non-informativeness. It is in view of these results that we choose the hierarchical random β model with $\alpha = 2$, $g = 0.2$ and $\sqrt{(g/h\alpha)} = R/10$ as our default option.

5.1.2 Sensitivity to the prior distribution of the means

An important component of the mixture model is the prior model for the means μ_j, which we defined as drawn independently from the normal distribution $N(\xi, \kappa^{-1})$. Using only the extremes of the data, we consider that setting ξ equal to the midrange and the precision κ so that $\kappa^{-1/2}$ is equal to R is a sensible weakly informative prior which places effectively no constraint on the location of the μ_j, but does not encourage the fitting of mixtures with very close μ_j.

There is a subtle interplay between prior information on the location of the means and the number of components. Indeed reducing $\kappa^{-1/2}$ at first will tend to favour a higher number of components. This can be interpreted as the result of defining a prior for the means which is increasingly more *permissive* of components with close means. On the other hand, as $\kappa^{-1/2}$ is further reduced, the number of components will start to decrease, as there is now a shrinkage effect and active prohibition of components with means located towards the extremes of the range. We illustrate these points on the Acidity data. We have used throughout the same hierarchical default option for the variances, but a Poisson prior $\mathcal{P}(10)$ for k as some hyperparameter settings now encourage large k. As the values of $\kappa^{-1/2}$ decrease from R to $R/10$, the number of components with the highest posterior probability first increases to reach a peak value of $k = 10$ for $\kappa^{-1/2}$ between $R/4$ and $R/5$ and then decreases again (Table 2).

We have so far discussed sensitivity to the prior setting of κ with the hierarchical random β model for the component variances, but the same behaviour is observed with the fixed β model. Sensitivity was discussed by Crawford (1994) who computes, for the Acidity data, the posterior distribution of k in three cases: $\alpha = \beta = \kappa = 1$, $\alpha = \beta = \kappa = 5$, and $\alpha = \beta = \kappa = 10$. Note that the restriction imposed on the means is quite severe in the last 2 cases, $\kappa = 1, 5, 10$ corresponding respectively to $\kappa^{-1/2}$ equal to $R/4, R/10, R/13$ for this data set. By simultaneously increasing α, β and κ, the means of the components are restricted, while the standard deviations are tightened around $1 \approx R/4$, a fairly large value, thus creating competing influences on k which are not easy to disentangle. We have fitted our model with the same parameter settings as Crawford. We find more support for 2 components than in our previous analysis with our default priors (cf. Table 1), a posterior distribution mostly concentrated on $k = 2$ or 3, with moderate variation between the 3 cases, $p(k = 2|y)$ being equal to $0.42, 0, 65$ and 0.43 respectively. However, the Laplace approximation estimates for $p(k = 2|y)$ given in Crawford's Table 2 vary over many orders of magnitude.

5.2 Sensitivity of the posterior distributions of parameters

In a complementary way and from the same MCMC runs, we can investigate the sensitivity of the posterior distributions of component parameters for various values of k. We shall briefly summarise some features.

Results concerning the influence of κ are unsurprising. As expected, a reduction of the range of the means μ_j is observed as $\kappa^{-1/2}$ is decreased, a phenomenon which is more noticeable for large k.

It is interesting to compare the influence of prior specifications for the variances $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$ between the fixed β and the random $\beta \sim \Gamma(g, h)$ model. A similar sensitivity pattern to that described for the posterior distribution of k emerges. In the fixed β case the posterior means of σ_j are sensitive to variations of $\sqrt{(\beta/\alpha)}$, the more so when k is larger. For example for the Acidity data with $k = 4$, the posterior means of σ_j become nearly halved as $\sqrt{(\beta/\alpha)}$ is varied from $R/5$ to $R/20$. Whereas for the random β case, all the posterior means of σ_j are remarkably similar. This is further evidence for the value of including an upper hierarchical level in the distribution of the σ_j in a mixture model.

6 Performance of the MCMC sampler

6.1 Mixing over k: performance of the jump moves

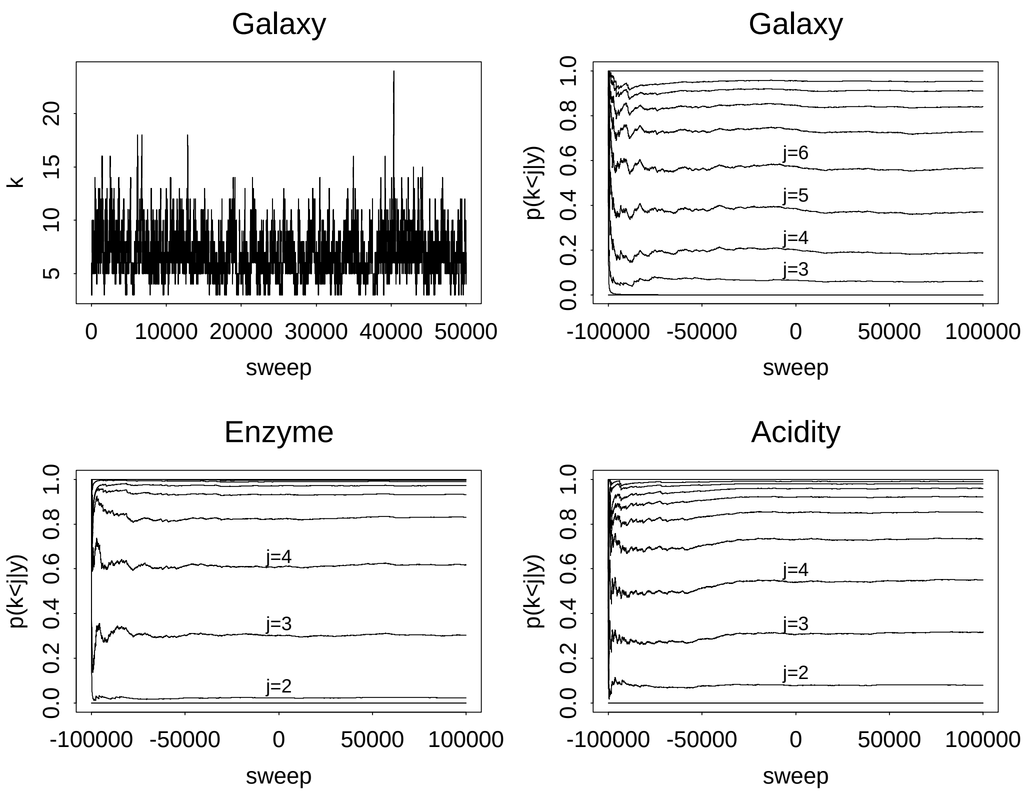

Figure 7: Example of trace of k for the Galaxy data set, for 50 000 sweeps after burn-in, and cumulative occupancy fractions for all three data sets, for complete run including burn-in.

An essential element of the performance of of our MCMC sampler is its ability to move between different values of k. A plot of the changes in k against the number of sweeps for the Galaxy data is presented in Figure 7. It shows that the MCMC algorithm mixes well over k, excursions into very high values being short-lived. Similar plots were obtained for the other data sets. Proportions of accepted 'split or combine' moves vary between 8% and 14% (Table 1). For dimension changing moves, these proportions are satisfactory and show that our proposal based on adjacency is sensible. A useful check on the stationarity is given by the plot of the cumulative occupancy fractions for different values of k against the number of sweeps. These are represented in Figure 7 for the 3 data sets, where it can be seen that the burn-in is more than adequate to achieve stability in the occupancy fractions.

Our model does not preclude the existence of empty components, and they will be included in our count of k. This might cause concern if a high number of them persisted for long times. We have found that including in our algorithm the birth/death moves, which specifically deal with empty components, improves convergence in comparison to that of an algorithm relying only on the split/combine moves, especially when the posteriors are diffuse. The acceptance rate for birth and death moves is highest for the small and multimodal Galaxy data set. The mean number of empty components is equal to 0.10, 0.18 and 0.57 for the Enzyme, Acidity and Galaxy data sets respectively.

We detected no influence of starting values on the distribution of k. For example with the Enzyme data, starting with $k = 1$ typically leads to the acceptance of the first split after less than 5 sweeps, and then $k = 1$ is never accepted again; further, when starting with $k = 20$, and observations ranked then equally allocated between the components, less than one hundred sweeps are needed to reach $k = 10$. Unless specified otherwise, all our runs start with $k = 1$.

The prior distribution of k, $p(k)$, appears in the acceptance ratio for the dimension-changing moves and will influence the mixing behaviour precisely through that ratio. Priors which are highly concentrated on a range of values of k will effectively stop the algorithm from accepting moves which will take it outside this range. However, the Bayes factors $B_{k_1 k_2} = \{p(k_1|y)/p(k_2|y)\} \div \{p(k_1)/p(k_2)\}$, which are theoretically independent of $p(k)$, have MCMC estimates that are not materially affected by it, *within the range of k visited reasonably often.* For example on the Acidity data, B_{34} is estimated as 0.91, 0.99 and 1.01 when the prior for k is a Poisson distribution with mean equal to 1, 3 and 10 respectively and as 1.03 for a uniform prior for k (between 1 and 30).

6.2 Mixing within k

6.2.1 Within-k mixing for parameters

Within the range of weak priors that we have been using, we have observed satisfactorily mixing patterns in all our runs and not encountered any "trapping states" as reported in Robert (1996). We checked that runs with very different initial allocations gave almost identical posterior densities for the parameters. For the Enzyme data and 3 components, Figure 8 displays typical time plots of the sweeps for w, μ, σ. These are based on a run of 100 000 sweeps, which included about 30 000 visits to $k = 3$, but plotted only every 20 for the sake of clarity in these plots. Different pattern of traces for the three components can be seen. The first component (lowest mean and standard deviation, highest weight) is estimated precisely. Very occasionally, much fewer than 60% of the observations are allocated to it, but this creates no problem. Note that when this occurs, the mean of the second component dips as some switching arises between the two components. The weights of the second and third components are more fluctuating. For the third component, competing explanations, as discussed earlier in Section 4.3.2, are clearly visible and the algorithm has no trouble in covering the wider range of values, higher means corresponding to lower weights and standard deviations.

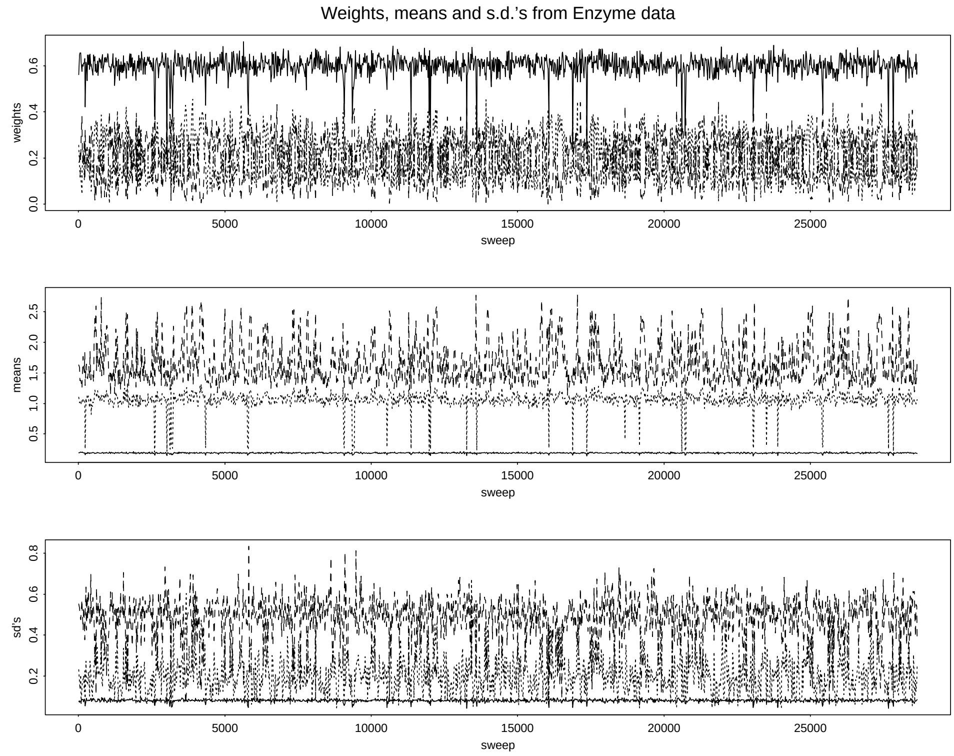

Figure 8: Traces of parameter estimates against visits to $k = 3$, Enzyme data.

6.2.2 Comparisons with fixed-k sampler

Some previous work using MCMC in mixture estimation with fixed k has encountered slow mixing, especially with weak priors. This is usually caused by the presence of two or more modes in the posterior distribution, separated, so far as the available MCMC moves are concerned, by regions of low probability. In statistical terms, there are two or more well-supported explanations for the data with the same k. For example, the data may fall into two rather well-separated clusters, and with k fixed at 3, there may be substantial posterior probability on two components being fitted to the first cluster and one to the second, or *vice-versa.*

It is plausible that in the presence of multimodality, mixing should be improved by the possibility of varying k. In the particular situation described above, the sampler could, at some stage, combine the two components in the first cluster, and subsequently split the component in the second cluster, and so complete a transition from one mode to the other, without visiting regions of low posterior probability. This is an example of what physicists would call "tunnelling" between regions of low "energy" (energy is the negative logarithm of the probability).

Here we present an example of the improved mixing obtained by varying k. The example is somewhat contrived, but we believe that it is qualitatively similar to real problems with clustered data, in the absence of strong prior information. We take 50 observations from $N(2.5, 1)$, 50 from $N(4, 1)$, and assemble a synthetic data set of size 200 by taking these 100 data points, *and their reflections about the origin.* Our default prior is used, except that k is given a Poisson prior, with $\lambda = 4$. We thus contrive a situation in which the joint posterior distribution possesses *exact symmetry* on reflection about 0. We compare results of simulating the joint posterior with variable k, and then conditioning on $k = 3$, with running a fixed-k sampler for $k = 3$ using only moves (a) to (d) of Section 3.2. Run lengths were arranged so that the same numbers of visits to $k = 3$ were made in each case. Some results are displayed in Figure 9. By symmetry, the true posterior

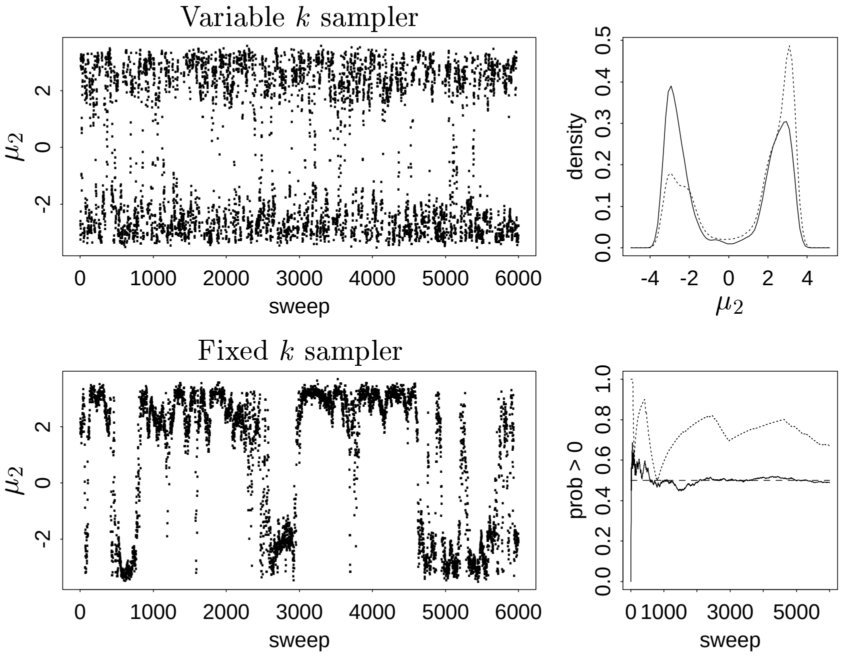

Figure 9: Comparison of mixing of variable-k and fixed-k samplers. Left panels: traces of μ_2 against sweep number. Right panels: (upper) posterior density estimates at the end of the runs, (lower) sequences of estimates of $p(\mu_2 < 0|y, k = 3)$ obtained as the runs proceed; solid lines refer to the variable-k sampler.

$p(\mu_2|y, k = 3)$ is symmetric about 0, and in particular $p(\mu_2 < 0|y, k = 3) = 0.5$. The two panels on the left of the figure show traces of μ_2 against sweep number. The variable-k sampler evidently mixes far better than the fixed-k one. This improvement extends to estimates of the density of μ_2 and the probability that it is positive, illustrated in the two right-hand panels. After more than 6000 sweeps, results for the fixed-k sampler still show severe asymmetry.

7 Bayesian classification

Apart from their role in facilitating computation, the allocation variables z are of interest in their own right, for they form a coherent basis for classification of the observations. Some care is required in interpreting these sensibly. It will rarely be appropriate to discuss classification except conditional on k, and even then the labels $1, 2, \ldots, k$ are only meaningful in the context of some particular declared unambiguous labelling of the k mixture components, for example, by ordering on the μ_j.

Classification can either be done on a within-sample or predictive basis. For the former, the posterior probabilities $\{p(z_i = j|y, k); j = 1, 2, \ldots, k\}$ are appropriate, and these can be directly estimated as empirical averages in the MCMC run. Predictive classification addresses the question of classifying a future observation, y^*, say. If the allocation variable corresponding to this is z^*, then we are in principle interested in $\{p(z^* = j|y, y^*, k)\}$. Unfortunately, inclusion of the additional datum changes all of the posterior distributions, apparently requiring that the MCMC sampler be re-run for each new y^*! This is obviously impractical, so we employ the obvious approximation

$$\begin{aligned} p(z^* = j|y, y^*, k) &= \int p(z^* = j|y, y^*, k, \theta, w) p(\theta, w|y, y^*, k) d\theta\, dw \\ &= \int p(z^* = j|y^*, k, \theta, w) p(\theta, w|y, y^*, k) d\theta\, dw \\ &\approx \int p(z^* = j|y^*, k, \theta, w) p(\theta, w|y, k) d\theta\, dw, \end{aligned}$$

and estimate the last integral, like any other expectation with respect to $p(\theta, w|y, k)$, by a MCMC empirical average, in this case that of $w_j \phi(y^*; \mu_j, \sigma_j) / \sum_{j=1}^{k} w_j \phi(y^*; \mu_j, \sigma_j)$ where $\phi(\cdot; \mu, \sigma)$ is the normal density. In Figure 10, the estimated within-sample and predictive classification probabilities are illustrated for the Enzyme data set. The lower section of each panel shows the *cumulative* classification probabilities, $p(z_i \leq j|y, k)$ and $p(z^* \leq j|y, y^*, k)$ for $j = 1, 2, \ldots, k-1$; differences between adjacent curves indicate the class probabilities. The within-sample probabilities and the predictive ones coincide to within plotting accuracy. This is an effect of the law of large numbers; they are computed separately.

Using the usual "percent correctly classified" loss function, the Bayes classification of an existing observation y_i and a future one y^* are respectively given by

$$\widehat{z_i} = \mathrm{argmax}_j\, p(z_i = j|y, k) \qquad \text{and} \qquad \widehat{z^*} = \mathrm{argmax}_j\, p(z^* = j|y, y^*, k).$$

These are also plotted, in the upper part of each panel, in Figure 10.

Note that there is no monotonicity in the classification with respect to increasing values of the enzyme level. For $k = 3$, data values classified to the third component lie on either side of those assigned to the second component. This is due to the large variance of the third component. Correspondingly, the predictive curve delimiting the second from the third component has a pronounced dip. This phenomenon persists with larger k, indicating that the classifications cannot be simply interpreted in terms of shifts of the mean enzyme level, but take into consideration a combination of mean and spread.

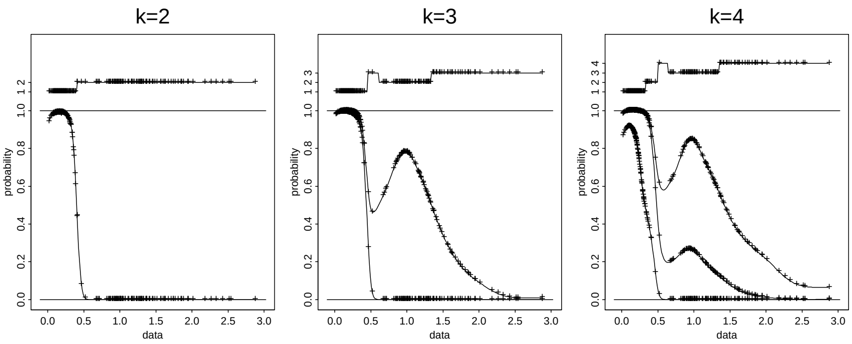

Figure 10: Classifications of the Enzyme data set: within-sample (crosses) and predictive (solid curves) classification probabilities and optimal classifications for $k = 2, 3, 4$.

8 Further work and discussion

8.1 MCMC issues

The key idea in constructing an effective MCMC sampler is to design sensible moves that use current knowledge about the mixture. Basing moves on a notion of adjacency has a generic character, independent of the particular distributional assumptions, and thus these move types could be adapted to a variety of distributions. We believe that they would be suitable for any 2 parameter density $f(\cdot|\theta_1, \theta_2)$, where after a suitable transformation f could be reparametrised in term of a mean and variance parameter with the range of the mean unconstrained.

We could have extended our birth and death moves to include nonempty components, similarly to Phillips and Smith (1996), and then dispensed with our split and combine moves, or defined combine moves with respect to the underlying partition given by the data rather than the parameters, along the line followed by Gruet *et al.*(1996). We felt that these moves would be less efficient, but in further developments of our work, we aim to perform some comparisons. More adventurous moves could be contemplated; for example, moves that combine three components, or take two components and simultaneously update the relevant parameters, weights and allocations. However, there is no evidence from our results that such complications are necessary.

While adjacency and preserving the first two moments are quite natural conditions in a one-dimensional setting, the moves thus created might be too restricted for good mixing in applications to bivariate or multivariate mixtures. There, we feel that the algebra of moves will need to be extended.

In calculating the acceptance ratios for dimension-changing moves, the only term which might be cumbersome is the Jacobian of the transformations. In higher-dimensional parameter space, symbolic computation tools could be useful at this point. These calculations are simplified by preserving as much symmetry as possible in the definition of the moves.

The intricacy of our MCMC sampler may convey an impression that it is computationally demanding. In fact, the burden is not excessive. For the largest of our three real data sets, the Enzyme data, with $n = 245$, and using the default prior setting, our program makes about 160 sweeps per second on a Sun Sparc 4 workstation.

Validation of the MCMC code is an important concern. We have compared our results with analytic calculations on very small data sets and found a good correspondence. We have also checked that *without any data*, our estimate of the joint posterior distribution tallies with the chosen prior.

8.2 Presentation of posterior distributions

Extracting information from such a complex multidimensional posterior distribution is a challenge. Whilst MCMC methods circumvent the restrictions of conventional numerical methods and provide all the raw information, insightful and disciplined summaries are needed, adapted to the particular context. In the paper we have exploited some opportunities, involving relabelling, predictive densities, classification, etc. New summaries, especially regarding joint parameter distributions, should be developed. Postprocessing is especially useful in view of the inherent identifiability problems connected with mixture estimation, which are crystallised in the contrast between the variability of parameter estimates plots, representing competing explanations, and the stability of predictive density plots.

Our approach has revealed clear evidence of multimodality and skewness in posterior distributions, features whose presence is unsurprising in view of the small numbers of observations sometimes allocated to some components. We believe that this is a situation where using analytic approximations such as Laplace can be misleading.

8.3 Other prior structures

The interaction between the model and the number of components, in terms of both structural and functional characteristics of the prior, has been discussed and illustrated extensively. We believe that the hierarchical prior structure that we have introduced will be useful for many examples. Nevertheless, it is certainly not a "black-box" procedure. Our default choice for hyperparameters is aimed at using mixtures for analysing heterogeneity rather than for semiparametric density estimation.

A relationship between the prior distribution of the means and the posterior for k is to be expected, and indeed we have found sensitivity to values of κ. Instead of considering the $\{\mu_j\}$ to be independent, it might be more natural to model the notion of separation of the means explicitly by using *dependent* priors. This is one example of a feature which could be built into a joint prior distribution for $\{\mu_j, \sigma_j^2\}$, for which our computational approach would still be available.

Dependent priors over the component parameters have been considered by several authors when attempting to be noninformative in the mixture context. This essentially entails linking the components via global parameters, to which flat priors can be assigned since all the data points contribute to their estimation. Related but distinct approaches following this line are taken by Robert (1996), Mengersen and Robert (1996) and Gruet *et al.* (1996), scaling with respect to the component with largest variance, and Roeder and Wasserman (1995), placing Markov priors on the means.

8.4 Generalisations of the model

Among related models to which we have extended our MCMC sampling strategy is the Escobar and West (1995) mixture model based on a Dirichlet process prior. The hierarchical structure is now somewhat different from that of equation (5), but the range of moves that are needed is broadly the same, and the elegant algebraic structure of the Dirichlet process model facilitates the evaluations needed to implement the split/combine move.

Our approach has so far been implemented only for *normal* mixtures, and the interpretation of the number of components is conditional on this being an appropriate distribution for all the subpopulations. If we take any phenotypic data like the Enzyme data, the assumption of normality might not be supported by biological considerations. Indeed, the maximum likelihood procedure SKUMIX for analysing mixtures of Maclean *et al.*(1976) does allow for different degrees of skewness through the use of a Box-Cox transformation; and in the original analysis of the Enzyme data, Bechtel *et al.*(1993) concluded that the data were fitted by 2 highly skewed components. One extension that we are considering is the development of a framework where variable number of components and variable skewness in the mixture distribution would be simultaneously considered.

Another straightforward extension is to consider mixtures of discrete distributions, a model which is commonly used in non-parametric estimation and which is usually estimated via the EM algorithm separately for different numbers of mass points.

Finally we emphasise the flexibility of our modelling for incorporating many of the extensions which arise when mixture estimation is used in different application contexts. In particular we aim to consider problems involving constraints on the weights for genetic analysis, modelling component means in terms of covariates, and using mixtures for robust prior modelling in Bayesian analysis and for modelling unknown exposure distributions in measurement error problems.

Acknowledgements

We wish to thank Jim Berger, Ed George, Agostino Nobile, Christian Robert, Kathryn Roeder, Duncan Thomas and Larry Wasserman for stimulating discussions about this work, Catherine Bonaïti for introducing us to the genetic applications of mixture estimation, Pierre Bechtel for providing the Enzyme data set, Christine Monfort for assistance with the computations, and the referees for suggestions which improved the presentation. We acknowledge the financial support of the EPSRC Complex Stochastic Systems Initiative (PJG), INSERM (SR), and the ESF network on Highly Structured Stochastic Systems.

References

Bechtel, Y. C., Bonaïti-Pellié, C., Poisson, N., Magnette, J. and Bechtel, P. R. (1993) A population and family study of N-acetyltransferase using caffeine urinary metabolites, *Clinical pharmacology and therapeutics*, **54**, 134–141.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion), *Statistical Science*, **10**, 3–66.

Celeux, G., Chauveau, D. and Diebolt, J. (1996) Stochastic versions of the EM algorithm: an experimental study in the mixture case, *Journal of Statistical Computation and Simulation*, (to appear).

Crawford, S. L., DeGroot M. H., Kadane, J. B. and Small, M. J. (1992) Modeling lake chemistry distributions: approximate Bayesian methods for estimating a finite mixture model. *Technometrics*, **34**, 441–453.

Crawford, S. L. (1994) An application of the Laplace method to finite mixture distributions, *Journal of the American Statistical Association*, **89**, 259–267.

Dacunha-Castelle, D. and Gassiat, E. (1996) Estimation of the order of a mixture. *Bernoulli*, **2**, to appear in December 1996 issue.

Diebolt, J. and Robert, C. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, B, **56**, 163–175.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixture *Journal of the American Statistical Association*, **90**, 577–588.

Feng, Z. D. and McCulloch, C. E. (1996) Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society*, B, **58**, 609–617.

Green, P. J. (1994) Contribution to the discussion of paper by Grenander and Miller (1994). *Journal of the Royal Statistical Society*, B, **56**, 589–590.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grenander, U. and Miller, M. (1994) Representations of knowledge in complex systems (with Discussion). *Journal of the Royal Statistical Society*, B, **56**, 549–603.

Gruet, M., Robert, C. and Wolpert, R. (1996) Estimating the number of components in a normal mixture. *Technical report, Université de Rouen.*

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to experts systems (with discussion). *Journal of the Royal Statistical Society*, B, **50**, 157–224.

Lindsay, B. G. (1995) *Mixture models: theory, geometry, and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Institute of Mathematical Statistics: Hayward, California.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture models: inference and applications to clustering*, Marcel Dekker: New York.

Maclean, C. J., Morton, N. E., Elston, R. C. and Yee, S. (1976) Skewness in commingled distributions. *Biometrics*, **32**, 695–699.

Mengersen, K. and Robert, C. (1996) Testing for mixtures: a Bayesian entropy approach. In *Bayesian Statistics 5*, J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds. Oxford: Oxford University Press, (in press).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Nobile, A. (1994) *Bayesian analysis of finite mixture distributions*, Ph.D. thesis, Carnegie Mellon University.

Peskun, P. H. (1973) Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

Phillips, D. B. and Smith, A. F. M. (1996) Bayesian model comparison via jump diffusions, chapter 13 (pp. 215–239) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Raftery, A. E. (1996) Hypothesis testing and model selection, chapter 10 (pp. 163–188) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Robert, C. (1996) Mixtures of distributions: inference and estimation, chapter 24 (pp. 441–464) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *Journal of the American Statistical Association*, **85**, 617–624.

Roeder, K. and Wasserman, L. (1995) Practical Bayesian density estimation using mixtures of normals. *Technical report #633, Department of Statistics, Carnegie Mellon University.*

Tierney, L. (1994) Markov chains for exploring posterior distributions, *Annals of Statistics*, **22**. 1701-1762.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.