Multivariate mixtures of normals with unknown number of components

Petros Dellaportas and Ioulia Papageorgiou

Department of Statistics, Athens University of Economics and Business, Greece

Abstract

We present full Bayesian analysis of finite mixtures of multivariate normals with unknown number of components. We adopt reversible jump Markov chain Monte Carlo and we construct, in a manner similar to that of Richardson and Green (1997), split and merge moves that produce good mixing of the Markov chains. The split moves are constructed on the space of eigenvectors and eigenvalues of the current covariance matrix so that the proposed covariance matrices are positive definite. Our proposed methodology has applications in classification and discrimination as well as heterogeneity modelling. We test our algorithm with real and simulated data.

Keywords: Bayesian inference, Classification; Markov chain Monte Carlo; Prediction; Reversible jump.

1 Introduction

A rather influential application of reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995) is the analysis of finite mixtures of normal densities with unknown number of components presented by Richardson and Green (1997). This work has increased the flexibility of mixture models by allowing simultaneous Bayesian estimation of both model parameters and number of mixture components. In this paper we generalize the work of Richardson and Green (1997) to the case of multivariate normal densities. Our motivation is first to enrich the applicability of multivariate mixture modelling by borrowing from the experience of the Bayesian univariate estimation approaches, and second to investigate the power of reversible jump moves in complicated model comparisons; see the recent paper and discussions of Brooks *et al.* (2003).

Mixture models have attracted the attention of statisticians from both a theoretical and a practical perspective even from the end of the 19th century. They primarily serve as means of

modelling heterogeneity for classification and discrimination and of formulating flexible models for density estimation. General background is given in the books by Titteringhton *et al.* (1985), Lindsay (1995) and McLachlan and Basford (1988). When Markov chain Monte Carlo (MCMC) methods entered the area, see Diebolt and Robert (1994) and Robert (1996), and in particular after the paper by Richardson and Green (1997) that used RJMCMC, the power and flexibility of these models has further increased. As a result, a series of new modeling aspects such as model averaging, partial exchangeability and 'multiple-explanations' can be assessed. For some applications of univariate normal mixtures that use RJMCMC see Nobile and Green (2000), Robert *et al.* (2000), Fernandez and Green (2002), Green and Richardson (2001), Bottolo *et al.* (2003).

Richardson and Green (1997) based their RJMCMC algorithm in a series of combine-split moves that rely on moment matching. In particular, a proposed model is generated by randomly choosing a combine or a split move. The combine move picks up randomly two components and proposes to merge them to one, whereas the split move suggests splitting a randomly chosen component into two new components. Birth-death moves are also used to facilitate the mixing of the chain. Clearly, these move types are not typical, since the current and the proposed models are not nested. The key ingredient for their success is the fact that the first two moments are retained after a split or a combine move.

As the dimension p of the normal densities increases, the difference in parameters in split and combine moves increases with order p^2. The main contribution of our paper is the design of multidimensional efficient split and combine moves. To briefly discuss the challenge of this task, notice that the difficulty in the split (and equivalently in the combine) move arises when a covariance matrix needs to split in two matrices so that the overall dispersion remains relatively constant whereas the two new matrices must be positive definite. We achieve this by employing two key ideas. First, the proposed moves operate on the space of the p eigenvectors of the current covariance matrix. Second, the two new sets of p eigenvectors are randomly proposed by permuting the current eigenvectors through randomly chosen permutation matrices P and P' of dimension $p \times p$.

Full Bayesian estimation of mixtures of multivariate normal densities with unknown number of components has been attempted by Stephens (2000) who used continuous time birth and death processes; see Geyer and Møller (1994) and Møller and Waagepetersen (2003). Some limitations of this method, and comparisons with the RJMCMC algorithm can be found in Cappé *et al.* (2003). A recent attempt to deal with the problem using a reversible jump algorithm is the paper by Zhang *et al.* (2004), where all covariance matrices of the mixtures are restricted to share the

same eigenvector matrix.

The rest of the paper is organised as follows. In Section 2 we establish notation and describe the Gibbs sampling algorithm for a given number of components. Sections 3 describes our proposed reversible jump algorithm starting with the two-dimensional case that provides the best insight of the method and then elaborating on the general p-dimensional case. Section 4 discusses various MCMC and modelling aspects, whereas Section 5 has two, three and five dimensional examples with real and simulated data. Section 5 ends the paper with a brief discussion.

2 Known number of components

Let data $y_i,\ i = 1, 2, \ldots, n$ be $p \times 1$ vectors exchangeably distributed according to a known number k of p-dimensional multivariate normal distributions. With probability w_j, $\sum_{j=1}^{k} w_j = 1$, the conditional distribution of y_i given a set of parameters $\theta = (\mu, \Sigma, w)$, denoted $[y_i \mid \theta]$, is $N_p(\mu_j, \Sigma_j)$ where N_p denotes the p-dimensional Normal distribution. (Throughout the paper the usual square-bracket notation is used for joint, conditional and marginal densities). The parameter vector θ consists of the $p \times 1$ mean vectors $\mu = (\mu_1, \ldots, \mu_k)$, the $p \times p$ covariance matrices $\Sigma = (\Sigma_1, \ldots, \Sigma_k)$ and the classification probabilities $w = (w_1, \ldots, w_k)$. To express the mixture model in terms of unobserved data, we introduce indicator parameters $z = (z_{ij};\ i = 1, \ldots, n; j = 1, \ldots, k)$ such that $z_{ij} = 1$ if y_i belongs to the jth component of the mixture, and $z_{ij} = 0$ otherwise. This leads to a $N(\mu_j, \Sigma_j)$ distribution for the conditional distribution $[y_i \mid z_{ij} = 1]$ and to the relation $Pr[z_{ij} = 1 \mid w] = w_j$. Since the conditional density $[z_{ij} \mid w]$ is Multinomial with parameters $(1; w)$, the classification probabilities w can be thought of as hyperparameters determining the distribution of z. The joint distribution of the observed data y and the unobserved indicator parameters z conditional on the model parameters is

$$[y, z \mid \theta] = [y \mid \theta, z][z \mid w] = \prod_{i=1}^{n} \prod_{j=1}^{k} [w_j f(y_i \mid \mu_j, \Sigma_j)]^{z_{ij}} . \tag{1}$$

To facilitate notation and obtain a better insight of the RJMCMC algorithm we use, we will first describe an initial data transformation which reallocates and rescales the data. In particular, let $\bar{y}$ be the p-dimensional sample mean vector and $S = diag(s_1^2, \ldots, s_p^2)$ be a $p \times p$ diagonal matrix with elements the sample variances on each dimension. Then, any inference procedure can be operated on the data $\tilde{y}_i = S^{-1/2}(y_i - \bar{y})$ and the resulting estimates of $\tilde{\mu}_j$ and $\tilde{\Sigma}_j$ from

$$\tilde{y}_i \sim \sum_{j=1}^{k} w_j N(\tilde{\mu}_j, \tilde{\Sigma}_j)$$

are just transformed back to μ_j and Σ_j via $\mu_j = S^{1/2}\tilde{\mu}_j + \bar{y}$ and $\Sigma_j = S^{1/2}\tilde{\Sigma}_j S^{1/2}$. This initial transformation is also useful for computing purposes since it avoids continuous calculations in the RJMCMC algorithm so that proposed moves are adjusted to different scales in each dimension.

Bayesian formulation requires prior distributions for the model parameters θ. We adopt the conjugate priors

- $[\mu_j \mid \Sigma_j] \equiv N_p(\xi_j, \Sigma_j/c_j)$, $j = 1, \ldots, k$, for some means $\xi_j = (\xi_{j1}, \ldots, \xi_{jp})'$ and precision parameters $c_j > 0$. Following Richardson and Green (1997) one choice could be $\xi_{j\ell} = \min[y_{*\ell}] + jR_\ell/(k+1)$ for $\ell = 1, \ldots, p$ where $y_{*\ell}$ denotes the n data points in dimension ℓ and R_ℓ denotes the range of the data in dimension ℓ. Another choice that follows Cappé *et al.* (2003) is to use $\xi_{j\ell} = \bar{y}_{*\ell}$, the sample mean of $y_{*\ell}$, for all j, ℓ, and $c_j = 1$. In our examples in Section 6, since we operate on the transformed data $\tilde{y}$, we just set $\xi_{j\ell} = 0$ for all j, ℓ.

- $[\Sigma_j] \equiv W^{-1}(\zeta, \Xi)$, $j = 1, \ldots, k$, an inverse Wishart distribution with $\zeta_j > 0$ degrees of freedom and scale matrix $\Xi = diag(\gamma_1, \ldots, \gamma_p)$, where the *diag* operator converts a vector to a diagonal matrix. We use $\zeta = p + 1$ in all our examples.

- $\gamma \equiv G(g, \rho)$ a Gamma density with mean g/ρ. In our data analysis we used $g = 2$ and $\rho = 36^{-1}$.

- $[w] \equiv D(\delta, \ldots, \delta)$, a Dirichlet distribution for known δ. Our default choice is $\delta = 1$.

- $k \equiv P$, a discrete density on $\{0, 1, \ldots, k_{max}\}$. We take P to be a discrete uniform density.

Using Bayes theorem, the posterior joint distribution of θ given the observed and the missing data is given by

$$[\theta \mid y, z] \propto [y, z \mid \theta][\theta] \equiv [y \mid \theta, z][z \mid w][w][\mu \mid \Sigma][\Sigma].$$

Inference problems associated with the above posterior density include estimation of the parameter vector θ and investigation of modality. However, conventional Bayesian as well as likelihood-based methods are inhibited by the nature of the sampling density: the complete likelihood (1) is a product of n terms, each one being the product of k distinct elements, and even for moderate values of n considerable computational difficulties arise. For prediction, a useful quantity is the posterior predictive probability that an as yet unrecorded object is actually from one of the k groups. For future data y_f, this probability can be expressed via a future indicator parameter z_f:

$$[z_f = j \mid y_f, \theta] = \frac{[z_f = j \mid \theta][y_f \mid \theta, z_f = j]}{\sum_{j=1}^{k}[z_f = j \mid \theta][y_f \mid \theta, z_f = j]}.$$

Standard Gibbs sampling which generalizes the one-dimensional situation is straighforward for fixed k; see, for example, Lavine and West (1992). The required full conditional distributions of the parameter vector $(\theta, \boldsymbol{\gamma})$ are

- $[z_{ij} = 1 \mid y, \theta] \propto w_j ND_p(x_i \mid \mu_j, \Sigma_j, z_{ij} = 1)$ where ND_p denotes the multivariate normal density of dimension p,
- $[\mu_j \mid \Sigma, w, y, z, \gamma] \equiv N_p(m_j, \Sigma_j/h_j)$ where $m_j = (c_j\xi_j + \sum_{i=1}^n z_{ij}y_i)/h_j$, $h_j = c_j + \sum_{i=1}^n z_{ij} = c_j + n_j$, $n_j = \sum_{i=1}^n z_{ij}$,
- $[\Sigma_j \mid \mu, w, y, z] \equiv W^{-1}(b_j, B_j)$ where $b_j = \zeta + n_j$, $B_j = \Xi + S_j + (\bar{y}_j - m_j)(\bar{y}_j - m_j)' n_j c_j/h_j$, and the sufficient statistics $\bar{y}_j$ and S_j are given by $\sum_{i=1}^n z_{ij}y_i/n_j$ and $\sum_{i=1}^n z_{ij}(y_i - \bar{y}_j)(y_i - \bar{y}_j)'$ respectively,
- $[\gamma_\ell \mid \Sigma] \equiv G(g + \sum_{j=1}^k \zeta_j/2, \rho + \frac{1}{2}\sum_{j=1}^k \Sigma_j^{-1}(\ell, \ell)$,
- $[w \mid \mu, \Sigma, y, z] \equiv D(\delta + n_1, \ldots, \delta + n_k)$.

Note that we make no further assumptions on the prior specifications such as parameter orderings which are motivated by identifiability concerns as in Richardson and Green (1997).

An interesting posterior inference statistic, which also serves as a diagnostic tool, is the predictive density of future data. This is particularly interesting in our modelling setup since, as pointed out by Richardson and Green (1997), the resulting densities do not have the form of mixture of Normals since they are (weighted) averages across k. Samples from these predictive densities, that are invariant to label switching, can be obtained by sampling one, or more, data points for each sampled points of θ selected from the RJMCMC algorithm.

3 Reversible jump moves

We now assume that k is a random variable and $k \leq k_{max}$ for a given value of k_{max}. We follow closely the one-dimensional approach of Richardson and Green (1997), so the required reversible jump transformation requires split, merge and birth-death moves. The major problem that emerges when attempting a split move in higher dimensions is the fact that the covariance matrices of the two new components should be kept positive definite. The key idea to overcome this problem is to use the spectral decomposition of the current covariance matrix and then operate the reversible jump technology on the resulting eigenvalues and eigenvectors. Thus, a covariance matrix is split

into two new covariance matrices by proposing new eigenvalues and eigenvectors following the guidelines given in the rest of this Section.

The number of parameters before and after the split move are $1+p+p(p+1)/2$ and $2+2p+2p(p+1)/2$ respectively, so a jump is proposed to a model with dimension larger by $1+p+p(p+1)/2$ parameters. Thus, the number of extra parameters that are needed for the jump move is increased quadratically with p. Let j_* be the one of the k components chosen to be considered to split, j_1, j_2 be the two proposed components, w_{j*}, w_{j1}, w_{j2} the corresponding weights, $\mu_{j*}, \mu_{j1}, \mu_{j2}$ the means, and $\Sigma_{j*}, \Sigma_{j1}, \Sigma_{j2}$ the covariance matrices. Let also $\Sigma_{j*} = V_{j*}\Lambda_{j*}V'_{j*}$ be the spectral decomposition of Σ_{j*} so that Λ_{j*} is a diagonal matrix $\Lambda_{j*} = diag(\lambda^1_{j*}, \ldots, \lambda^p_{j*})$ with elements the eigenvalues of Σ_{j*} with increasing order, and obtain similarly the spectral decompositions for Σ_{j1} and Σ_{j2}.

To obtain the best insight of our method, it is useful to discuss first the 2-dimensional case and to simplify notation, we will suppress in the following Sections the index j.

3.1 The two-dimensional case

Consider first the split move and let V^1_* and V^2_* be the two (2×1) eigenvectors of V_*. The new weights and mean vectors are specified by the equations

$$\begin{aligned} w_1 &= u_1 w_*, \\ w_2 &= (1-u_1) w_*, \\ \mu_1 &= \mu_* - \left(u^1_2 \sqrt{\lambda^1_*} V^1_* + u^2_2 \sqrt{\lambda^2_*} V^2_*\right) \sqrt{\frac{w_2}{w_1}}, \qquad (2) \\ \mu_2 &= \mu_* + \left(u^1_2 \sqrt{\lambda^1_*} V^1_* + u^1_2 \sqrt{\lambda^2_*} V^2_*\right) \sqrt{\frac{w_1}{w_2}}. \qquad (3) \end{aligned}$$

where the three random components u_1, u^1_2, u^2_2 are generated from beta and uniform distributions

$$u_1 \sim be(2,2), \quad u^1_2 \sim be(1,2), \quad u^2_2 \sim U(-1,1).$$

Note that the expressions for the new weights are identical to the formulas proposed by Richardson and Green (1997) whereas (2)-(3) collapse to their one-dimensional conditions when λ^2_* or V^2_* are zero. The expressions in (2) and (3) imply that the proposed means are obtained by moving across the system of axes formed by the eigenvectors by sizes $u^1_2\sqrt{\lambda^1_*}$ and $u^2_2\sqrt{\lambda^2_*}$ and then multiplying the resulting vector by the quantities under the square roots, see Figure 1.

The necessity to keep u^1_2 positive but let u^2_2 vary in the $(-1,1)$ interval is also evident by inspecting Figure 1: moves in both directions across the large eigenvector axis are achieved via (2) and (3) whereas moves across the shorter axis require both positive and negative u^2_2. A critical

tuning density is that of u_2^1, and we have found that a small bias towards 0 gives, on average, better mixing chains. This bias should increase with dimensions, see the corresponding formulas for general p in the next subsection.

We now proceed to the proposed eigenvalues and eigenvectors of the two new components. The new eigenvalue vectors Λ_1 and Λ_2 are specified by

$$\Lambda_1 = \begin{pmatrix} \lambda_1^1 & 0 \\ 0 & \lambda_1^2 \end{pmatrix} = \begin{pmatrix} u_3^1 & 0 \\ 0 & u_3^2 \end{pmatrix} \begin{pmatrix} 1-(u_2^1)^2 & 0 \\ 0 & 1-(u_2^2)^2 \end{pmatrix} \Lambda_* \frac{w_*}{w_1}, \tag{4}$$

$$\Lambda_2 = \begin{pmatrix} \lambda_2^1 & 0 \\ 0 & \lambda_2^2 \end{pmatrix} = \begin{pmatrix} 1-u_3^1 & 0 \\ 0 & 1-u_3^2 \end{pmatrix} \begin{pmatrix} 1-(u_2^1)^2 & 0 \\ 0 & 1-(u_2^2)^2 \end{pmatrix} \Lambda_* \frac{w_*}{w_2}. \tag{5}$$

which again resemble the one-dimensional variance proposals of Richardson and Green (1997). The random components u_3^1 and u_3^2 are generated via

$$u_3^1 \sim be(1,2), \quad u_3^2 \sim U(0,1).$$

Assume that the new eigenvector matrices V_1 and V_2 have eigenvectors V_1^1, V_1^2 and V_2^1, V_2^2 respectively. We propose the transformations $V_1 = PV_*$ and $V_2 = P'V_*$, where P is a 2×2 matrix with columns orthonormal unit vectors. Since such a matrix satisfies $P'P = I$ and $det(P) = 1$, it is a rotation matrix. It also satisfies $P^{-1} = P'$ and $det(P') = 1$ so both the inverse and the transpose of a rotation matrix are rotation matrices. In two dimensions the entries of rotation matrices can be represented with only one parameter θ taking values in $(0, \pi/2)$, the angle of the rotation, resulting to

$$V_1 = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} V_*, \quad V_2 = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} V_*.$$

Thus, the two new eigenvectors are produced by rotating the old eigenvector V_* by angles θ and $-\theta$, where $\theta \sim U(0, \pi/2)$.

The merge move requires the opposite transformation which is given by

$$\begin{aligned} w_* &= w_1 + w_2 \\ w_*\mu_* &= w_1\mu_1 + w_2\mu_2 \\ w_*\left[\left(\mu_*^{'}V_*^i\right)^2 + \lambda_*^i\right] &= w_1\left[\left(\mu_1^{'}V_*^i\right)^2 + \lambda_1^i\right] + w_2\left[\left(\mu_2^{'}V_*^i\right)^2 + \lambda_2^i\right],\ i = 1,2 \\ V_*^i &= \frac{V_1^i + V_2^i}{\|V_1^i + V_2^i\|},\ i = 1,2. \end{aligned}$$

where $\|.\|$ is the Euclidean norm. Solving for u_1, u_2^i, u_3^i and λ_*^i, $i = 1, 2$, we obtain

$$\begin{aligned}
u_1 &= w_1/w_* \\
u_2^i &= \left(\mu_*' V_*^i - \mu_1' V_*^1\right) / \left(\sqrt{\lambda_*^i w_2/w_1}\right),\ i = 1, 2, \\
u_3^i &= w_1 \lambda_i^1 / \left[w_* \lambda_*^i (1 - (u_2^i)^2)\right],\ i = 1, 2, \\
\lambda_*^i &= w_*^{-1} \left\{ w_1 \left[\left(\mu_1' V_*^i\right)^2 + \lambda_1^i \right] + w_2 \left[\left(\mu_2' V_*^i\right)^2 + \lambda_2^i \right] \right\} - \left(\mu_*' V_*^i\right)^2,\ i = 1, 2
\end{aligned}$$

3.2 General case

We preserve the same notational conventions for the general p-dimensional case. Let u_1, $u_2 = (u_2^1, u_2^2, \ldots, u_2^p)'$ and $u_3 = (u_3^1, u_3^2, \ldots, u_3^p)'$ be $2p + 1$ random variables needed to construct weights, means and eigenvalues for the split move. They are generated as

$$u_1 \sim be(2, 2),\ u_2^1 \sim be(1, 2p),\ u_2^j \sim U(-1, 1),\ u_3^1 \sim be(1, p),\ u_3^j \sim U(0, 1),\ j = 2, \ldots, p.$$

Let also P be a $p \times p$ rotation matrix with columns orthonormal unit vectors which has $p(p-1)/2$ free parameters. We generate P by generating its lower triangular matrix under the diagonal independently from $p(p-1)/2$ uniform $U(0, 1)$ densities. Note that there is a mathematical correspondence with the 2-dimensional case, but the conceptual rotation by a given angle is not any more available (actually there is a geometric understanding of the action of rotation in higher dimension by splitting R^p in subspaces, but it it beyond our scope to describe it here). Then, dropping again the j subscript, the proposed split moves are given by

$$\begin{aligned}
w_1 &= u_1 w_* \\
w_2 &= (1 - u_1) w_* \\
\mu_1 &= \mu_* - \left(\sum_{i=1}^{p} u_2^i \sqrt{\lambda_*^i} V_*^i \right) \sqrt{\frac{w_2}{w_1}} \\
\mu_2 &= \mu_* + \left(\sum_{i=1}^{p} u_2^i \sqrt{\lambda_*^i} V_*^i \right) \sqrt{\frac{w_1}{w_2}} \\
\Lambda_1 &= diag(u_3) diag(\iota - u_2) diag(\iota + u_2) \Lambda_* \frac{w_*}{w_1} \\
\Lambda_2 &= diag(\iota - u_3) diag(\iota - u_2) diag(\iota + u_2) \Lambda_* \frac{w_*}{w_2} \\
V_1 &= P V_* \\
V_2 &= P' V_*
\end{aligned}$$

where ι is a $p \times 1$ vector of ones. The corresponding merge move is specified by the expressions

$$\begin{aligned} w_* &= w_1 + w_2 \\ w_*\mu_* &= w_1\mu_1 + w_2\mu_2 \\ w_*\left[\left(\mu_*^{'}V_*^i\right)^2 + \lambda_*^i\right] &= w_1\left[\left(\mu_1^{'}V_*^i\right)^2 + \lambda_1^i\right] + w_2\left[\left(\mu_2^{'}V_*^i\right)^2 + \lambda_2^i\right],\ i = 1, 2, \ldots, p \\ V_*^i &= \frac{V_1^i + V_2^i}{\|V_1^i + V_2^i\|},\ i = 1, 2, \ldots, p \end{aligned}$$

and the solutions of u_1, u_2, u_3 and λ_*^i are

$$\begin{aligned} u_1 &= w_1/w_* \\ u_2^i &= \left(\mu_*^{'}V_*^i - \mu_1^{'}V_*^1\right) / \left(\sqrt{\lambda_*^i w_2/w_1}\right),\ i = 1, 2, \ldots, p \\ u_3^i &= w_1\lambda_1^i / \left[w_*\lambda_*^i(1 - (u_2^i)^2)\right],\ i = 1, 2, \ldots, p \\ \lambda_*^i &= w_*^{-1}\left\{w_1\left[\left(\mu_1^{'}V_*^i\right)^2 + \lambda_1^i\right] + w_2\left[\left(\mu_2^{'}V_*^i\right)^2 + \lambda_2^i\right]\right\} - \left(\mu_*^{'}V_*^i\right)^2,\ i = 1, 2, \ldots, p. \end{aligned}$$

4 Reversible jump implementation

The algorithm to sample from the joint distribution of model and parameters can be written as follows:

Step 1. Update $(\theta, \boldsymbol{\gamma})$ for known k: Carry out the Gibbs steps outlined in the Section 2.

Step 2. Propose a split or merge move with probabilities b_k and $d_k = 1 - b_k$ respectively.

Step 3. Propose a birth or death move with probabilities b_k and d_k respectively.

In our algorithm, instead of passing through each step deterministically, we choose to randomly select, in each iteration, one of the three steps with some fixed probabilities. This allows some extra tuning that can improve the mixing of the RJMCMC. In our examples we have used $(.2, .7, .1)$ as probabilities of choosing the three steps respectively; note that this implies that direct comparison of our output MCMC statistics with that of Richardson and Green (1997) needs some re-weighting, since one iteration in their algorithm consists of a sweep across all three steps.

Suppressing the j subscript, the acceptance probability for the split move is $\min(1, A)$ where

$$\begin{aligned} A &= (likelihood\ ratio) \times \frac{p(k+1)}{p(k)}(k+1)\frac{w_1^{\delta-1+n_1} w_2^{\delta-1+n_2}}{w_*^{\delta-1+n_1+n_2} B(\delta, k\delta)} \\ &\times \frac{(c_1 c_2)^{p/2}}{(2\pi)^p |\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} exp\left\{-\frac{1}{2}(\mu_1 - \xi_1)' \Sigma_1^{-1}(\mu_1 - \xi_1) - \frac{1}{2}(\mu_2 - \xi_2)' \Sigma_2^{-1}(\mu_2 - \xi_2)\right\} \\ &\times \frac{(2\pi)^{p/2} |\Sigma_*|^{1/2}}{c_*^{p/2}} exp\left\{\frac{1}{2}(\mu_* - \xi_*)' \Sigma_*^{-1}(\mu_* - \xi_*)\right\} \\ &\times \frac{DIW(\Sigma_1; diag(\gamma), \zeta_1)\, DIW(\Sigma_2; diag(\gamma), \zeta_2)}{DIW(\Sigma_*; diag(\gamma), \zeta_*)} \\ &\times \frac{d_{k+1}}{b_k P_{alloc}} \left[\frac{u_1(1-u_1)}{B(2,2)} \frac{(1-u_2^1)^{2p-1}}{B(1,2p)} \left(\frac{1}{2}\right)^{p-1} \frac{(1-u_3^1)^{p-1}}{B(1,p)}\right]^{-1} \left|\frac{\partial \Sigma}{\partial(V, \lambda)}\right| |J| \end{aligned} \tag{6}$$

where B denotes the Beta function and $DIW(X; \alpha, \beta)$ denotes the density of the inverse Wishart with parameters α, β evaluated at X, P_{alloc} is the probability that this particular allocation is made and J is the Jacobian of the transformation. Obvious choices for birth and death probabilities are $d_1 = b_{k_{max}} = 0$ and $d_k = b_k$ for $1 < k < k_{max}$. In fact, the expression for A resembles closely that of Richardson and Green (1997). The term $(k+1)$ in the nominator does not take into account the parameter ordering as in Richardson and Green but arises here from the equivalent ways that the components can produce the same likelihood and prior; see Cappé *et al.* (2003). Moreover, there is an extra Jacobian term that takes into account the transformation from the components of Σ to the eigenvalues and eigenvectors. Note that any two components can be chosen for splitting since no care is taken for the order of the components. The expressions for the Jacobian terms is given in the Appendix.

The corresponding expression for the birth and death moves is given by

$$A = \frac{p(k+1)}{p(k)} \frac{1}{B(k\delta, \delta)} w_*^{\delta-1}(1-w_*)^{n+k\delta-k} \frac{k+1}{k_0+1} \frac{d_{k+1}}{b_k} B(1, k) \tag{7}$$

where k_0 is the number of empty components.

5 Other computational and modelling issues

5.1 Labelling switching

An interesting, challenging problem that arises in the Bayesian analysis of mixture models and more generally in hidden Markov models, is the non-identifiability of the components caused by the invariance of the posterior distribution to the permutations in the parameter labelling. A general background to the solutions suggested in the past can be found in Jasra *et al.* (2003) who categorise them to artificial identifiability constraints (Diebolt and Robert, 1994, Richardson and

Green, 1997), random permutation sampling (Fruhwirth-Schnatter, 2001), relabelling algorithms (Stephens, 2000, Celeux, 1998), and label invariant loss functions methods (Celeux *et al.* , 2000, Hurn *et al.* , 2003).

Unfortunately, the problem is further complicated in higher dimensions, since the number of identifiability constraints on the parameter space is very large and relabelling algorithms will require a lot of computing time. Since this issue is of great complexity, we do not further elaborate on this in this paper; for our illustrative examples that follow, we either use relabelling techniques retrospectively, by post-processing the RJMCMC output sample based on prior information we have (see example of dimension 2), or we focus solely in predictive inferences that are invariant to label switching.

5.2 MCMC with fixed k revisited

The MCMC algorithm we used is based on the initial Gibbs sampler by Diebolt and Robert (1994) generalized in the multivariate setup by Lavine and West (1992). Dellaportas (1998) used this algorithm for a real data classification problem in which the data were both of continuous and discrete mode. This is certainly the easiest algorithm to use, but it is not clear that it is the best. For example, Cappé *et al.* (2001) and Jasra *et al.* (2003) integrate out the latent variables z and use Metropolis-Hastings moves together with tempering mechanism (Neal, 1996) to avoid trapping in local modes. Extension of this issue when dealing with multivariate normal densities needs further research and it is beyond the scope of our paper.

5.3 Convergence of RJMCMC

We have adopted a conservative approach to deal with the convergence of RJMCMC: we run our algorithms for long enough time (much longer that is probably needed), and we plot ergodic posterior estimates of k since we focus here only on inferences about k. Although we realise that this strategy might be deceiving, especially in the Normal mixtures models that may present special convergence problems (Robert, 1996), we felt that by presenting long raw output results, without using burn-in phases, we give the best insight of the behaviour of our RJMCMC. Our predictive image plots that depend not only on k but also on sampled values of all parameters, turn out to be rather insensitive to small departures of θ values. Again, we used many plots of output values (not reported here) rather than formal convergence tests.

5.4 Sensitivity to prior specification

The values of the parameters in the prior specification are only indicative and it is well known that they will affect, somehow, the resulting posterior inferences. Section 5 of Richardson and Green (1997) explores this issue in detail. Since we deal with the same hierarchical modelling framework, we have not encountered any new aspects of this theme so we chose not to replicate any of their findings here.

6 Examples

6.1 2 dimensions: Old Faithful data

We consider the version of Old Faithful data analyzed by Stephens (2000). There are 272 data points consisting of 2 observations each, the duration of a geyser eruption and the waiting time before the next eruption. A scatter plot of the data is shown in Figure 2. The RJMCMC ran for 90000 iterations, visited 8 models, and the resulting convergence graphs are depicted in Figure 3. The posterior model probabilities (without burn-in) are $(0.0002, 0.3035, 0.5854, 0.0941, 0.0146, 0.0016, 0.0005, 0.)$ for $k = 1, \ldots, 8$. The split/merge moves were encountered in 0.53% of the iterations. In this example, we performed a post-processing label switching based on the dimension that corresponds to duration in the mean vectors. Since the most probable model is the one with three components, our inference is conditional on $k = 3$. The mean vactors are $(2.0225, 54.4811)'$, $(3.4421, 70.1888)'$ and $(4.3429, 80.3428)'$ respectively, the covariance matrices are

$$\Sigma_1 = \begin{pmatrix} 0.057 & 0.316 \\ 0.316 & 34.580 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.288 & 3.317 \\ 3.317 & 85.98 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.135 & 0.465 \\ 0.465 & 32.862 \end{pmatrix}$$

and the corresponding weights are $(0.3399, 0.0874, 0.5722)$. It seems that our RJMCMC provides some evidence of a third component on the middle of the data points. Figure 2 provides an image plot of the predictive density that is based on all 90000 output sample, so it is (weighted) averaged across k. Note that these predictions are based, with probability around 0.7, on normal mixtures with more than 2 components. It seems that the predictive density has captured well the higher frequency regions.

6.2 3 dimensions: Simulated data

We report here one out of many simulated data examples we experimented with. We generated 80, 100 and 100 data points from the normal densities with means $(6, 4, 2)'$, $(-11, -4, -1)'$ and

$(-7, -11, -5)$ and covariance matrices

$$\Sigma_1 = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 0 \\ 1 & 0 & 4 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & -1.5 & 1 \\ -1.5 & 5 & 2 \\ 1 & 2 & 3 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 5 & -1 & 1 \\ -1 & 4 & -2 \\ 1 & -2 & 3 \end{pmatrix}$$

respectively. The three components are clearly far apart, see Figure 4, so there is little doubt that the data have been generated from a 3-component mixture. Figure 5 shows the estimates of the posterior model probabilities as the iterations increase. The algorithm visited 6 models and the resulting posterior probabilities for $k = 1, \ldots, 6$ are 0.0075, 0.002, 0.9493, 0.0379, 0.0048 and 0.0002 respectively. There has been label switching in the means in 73.5% of the iterations, and split or merge moves in 0.22% of the iterations.

6.3 5 dimensions: A real data example from Archeometry

We apply our RJMCMC algorithm to a data set on Romano-British pottery published in Tubb *et al.* (1980). The data come from five kiln sites in three regions and refer to 48 observations with 9 variables. The Archeological background suggests that there is a clear chemical compositional grouping that corresponds to the regional grouping. We applied principal component analysis on the original data and the analysis was based on the first 5 principal components that account for the 99% of the total variability. The data points are depicted in Figure 6. The RJMCMC algorithm showed rather quickly a preference to the model with three components and the mixing was very good, achieving split or merge moves in 1.35% of the iterations. The resulting posterior probabilities for $k = 2, \ldots, 8$ (without burn-in) are $(0.023, 0.746, 0.18, 0.0422, 0.008, 0.0005, 0.0002)$ respectively. Figure 6 presents the predictive density based on all iterations. Although these images are just projections of 5-dimensional densities in 2 dimensions, they capture the data points very well.

7 Discussion

We have extended the paper of Richardson and Green (1997) in a multivariate normal densities setting. Our proposed RJMCMC moves exploit the spectral decomposition of a matrix and operate in the space of eigenvectors of the covariance matrices. We briefly discuss here some possible applications of our methodology.

A natural application of the normals mixture problem is classification and discrimination. Since the statistical problem has parameter order p^2, when p is very large it is essential to use some initial principal component analysis, as suggested for example in Liu *et al.* (2003). The

richness of presentation of the posterior summary results, including interesting mixing-within-k data explanations has been discussed in detail by Richardson and Green (1997).

From a rather different perspective, mixture of normal densities can form the basis of specifying prior structures as, for example, in the one-dimensional examples of Nobile and Green (2000) and Bottolo *et al.* (2003). Here, the unknown number of components structure can serve as a way to relax the usual exchangeability assumptions in the prior parameters. Thus, mixture models become computationally feasible mechanisms to approximate partition models studied in, for example, Hartigan (1990) and Consonni and Veronese (1995). Note that this correspondence is not precise since partition models do not exactly correspond to finite mixture of normals, at least as implemented with our RJMCMC algorithm, due to the empty components that are allowed in the latter. A recent application in which random effects were modelled as mixtures of multivariate normal densities, with known number of components, can be found in Lopes *et al.* (2003). Also, in a recent paper, Green and Richardson (2001) advocate the use of mixtures of normal priors in place of Dirichlet process mixtures since the former achieve better computational efficiency and interpretation.

Acknowledgements

The research of the second author has been supported by the National scholarship foundation of Greece. The authors are thankful to Christian Robert for his valuable comments in a conference presentation of this work and to Ajay Jasra for helpful discussions.

Appendix

We first give the derivation for the Jacobian determinant $|J|$ of the split move appeared in (6). Suppressing as in (6) the subscript j, J is given by

$$J = \frac{\partial\left(w_1, w_2, m_1, m_2, \lambda_1, \lambda_2, V_1, V_2\right)}{\partial\left(w_*, u_1, m_*, u_2, \lambda_*, u_3, V_*, P\right)}$$

where P here denotes the $p(p-1)/2$ lower triangular elements of the matrix P. A close look at $|J|$ reveals that it is block diagonal, so it can be written as $|J| = |J_1||J_2||J_3|$, with J_1, J_2 and J_3 given by

$$J_1 = \frac{\partial\left(w_1, w_2\right)}{\partial\left(w_*, u_1\right)},\ J_2 = \frac{\partial\left(m_1, m_2, \lambda_1, \lambda_2\right)}{\partial\left(m_*, u_2, \lambda_*, u_3\right)},\ J_3 = \frac{\partial\left(V_1, V_2\right)}{\partial\left(V_*, P\right)}$$

having dimensions 2×2, $4p \times 4p$ and $p(p-1) \times p(p-1)$ respectively. First note that $|J_1| = w_*$.

Some algebraic elaboration gives

$$|J_2| = \prod_{\ell=1}^{p} \left[\lambda_*^{3/2}\left(1-(u_2^\ell)^2\right)\right] u_1^{-3p/2}(1-u_1)^{-3p/2}$$

and it is easy to show that for $p = 1$, $|J_1||J_2|$ reduces to the exact formula of Richardson and Green (1997). The last ingredient, $|J_3|$, cannot be written in closed form for every p, so we evaluate it element by element for each p. For example, $p = 2$ obtains

$$|J_3| = V_*(1,1)\sqrt{1-(P(1,1))^2}$$

where $V_*(1,1)$ and $P(1,1)$ denote the $(1,1)$ elements of matrices V_* and P respectively.

The Jacobian $|\partial\Sigma/\partial(V,\lambda)|$ can be computed by using the formulae

$$\partial\lambda = V_j'(\partial\Sigma)V_j, \quad \partial V = (\lambda_j I_p - \Sigma)^+(\partial\Sigma)V_j$$

where λ_j and its corresponding V_j, $j = 1,\ldots,p$, are specific eigenvalues and eigenvectors, and $(A)^+$ denotes the Moore-Penrose pesudo-inverse matrix of A; see Magnus and Neudecker (1988, page 157).

References

Bottolo L., Consonni G., Dellaportas P. and Lijoi A. (2003) Bayesian analysis of extreme values by mixture modeling. *Extremes*, **6**, 25-47.

Brooks S.P., Giudici P. and Roberts G.O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with Discussion) *Journal of the Royal Statistical Society, Series B*, **65**, 3-56.

Cappé O., Robert C.P. and Rydén T. (2001) Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *Technical Report.* **65**, 679-700.

Cappé O., Robert C.P. and Rydén T. (2003)Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Socisety, Series B*, **65**, 679-700.

Celeux G., Hurn M., Robert C.P. (2000) Computational and Differential Difficulties With Mixture Posterior Distributions. *Journal of the American Statistical Association*, **95**, 957-70.

Consonni, G. and Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935-944.

Dellaportas P. (1998) Bayesian classification of neolithic tools. *Applied Statistics*, **47**, 279-297.

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

Fernandez C. and Green P.J. (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Socisety, Series B*, **64**, 805-826.

Frühwirth-Schnatter S. (2001) Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, **96**, 194-209.

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

Geyer C. and Møller J. (1994) Simulation procedures and likelihood inference for spatial point processes, *Scandinavian Journal of Statistics*, **21**, 359–373.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, **28**, 355-376.

Hartigan, J. A. (1990). Partition models. *Communication in Statistics, Series A: Theory and Methods*, **19**, 2745-2756.

Hurn M., Justel A. and Robert C.P. (2003) Estimating mixtures of regressions. *Journal of computational and graphical statistics*, **12**, 1-25.

Jasra A., Holmes C.C. and Stephens D.A. (2003) Markov chain Monte Carlo and the label switching problem in Bayesian mixture modelling. *preprint.*

Lavine M. and West M. (1992) A Bayesian method for classification and discrimination. *The Canadian Journal of Statistics*, **20**, 451-461.

Lindsay B. G. (1995) *Mixture models: Theory, Geometry and Applications.* Hayward: Institute of Mathematical Statistics.

Liu J. S., Zhang J. L., Palumbo M.J. and Lawrence C.E. (2003) Bayesian Clustering with Variable and Transformation Selections. *Bayesian Statistics 7*, (eds J.M. Bernardo, Bayarri M.J., Berger J.O., Dawid A.P., Heckerman D., Smith A.F.M. and West M.), pp. 249-275, Oxford University press, UK.

Lopes H.F., Muller P. and Rosner G.L. (2003) Bayesian Meta-analysis for Longitudinal Data Models Using Multivariate Mixture Priors. *Biometrics*, **59**, 66-72.

Magnus J.R. and Neudecker H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons, NY.

McLachlan G. J. and Basford K. E. (1988) *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

Møller, J. and Waagepetersen R. (2003) *Statistical Inference and Simulation for Spatial Point Processes Series*. Monographs on Statistics and Applied Probability, Vol. 100, Chapman and Hall, UK.

Neal R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **4**, 353-366.

Nobile, A. and Green, P.J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, **87**, 15-35.

Richardson, S. and Green, P. J. (1997). On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Socisety, Series B*, **59**, 731-792.

Robert C. P. (1996) Mixtures of distributions: inference and estimation, *Markov chain Monte Carlo in Practice*, (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 441-464, Chapman and Hall, UK.

Robert C.P., Rydén T. and Titterington D.M. (2000)Bayesian inference in hidden Markov models through the reversible jump MArkov chain Monte CArlo method. *Journal of the Royal Statistical Socisety, Series B*, **62**, 57-76.

Stephens M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795-809.

Stephens M. (2000) Bayesian analysis of mixture models with an unknown number of components -an alternative to reversible jump methods. *Journal of the Royal Statistical Society, Series B*, **62**, 795-809.

Titterington D. M., Smith A. F. M. and Makov U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tubb, A., Parker, A. J., and Nickless, G. (1980) The Analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, **22**, 153-171.

Zhang Z., Chan K. L., Wu Y. and Chen C. (2004). Learning a multivariate Gaussian mixture models with the reversible jump MCMC algorithm. *Statistics and Computing*, **14**, 343-355.

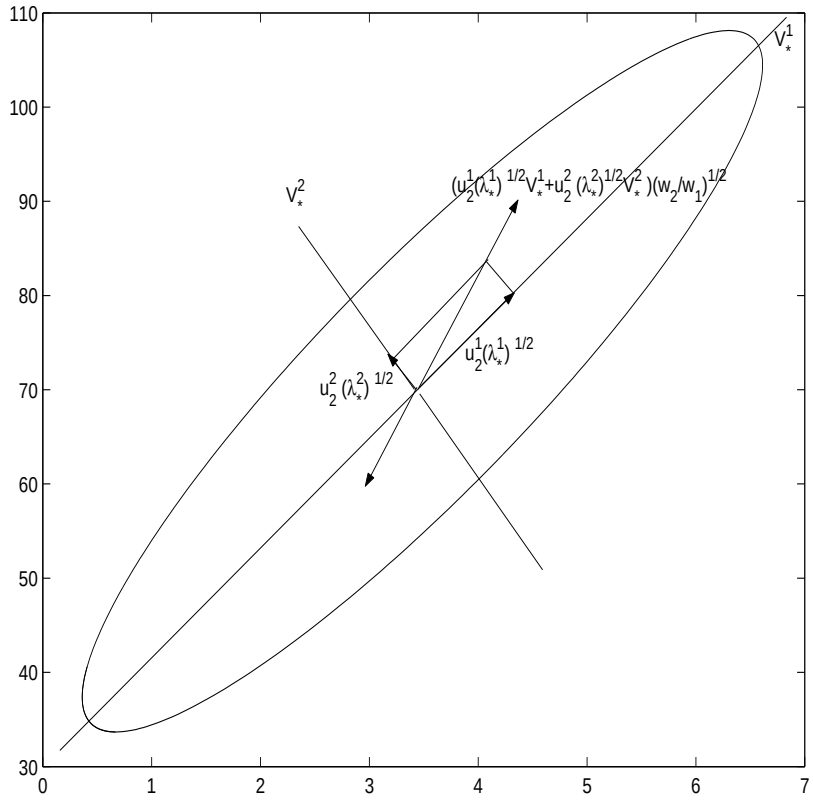

Figure 1: The proposed split move in 2 dimensions

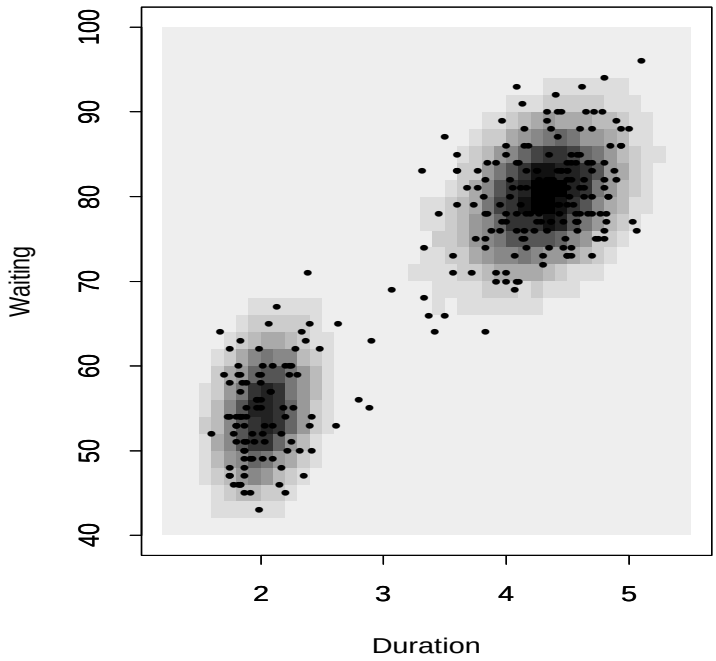

Figure 2: Old Faithful data and image plot of the predictive density

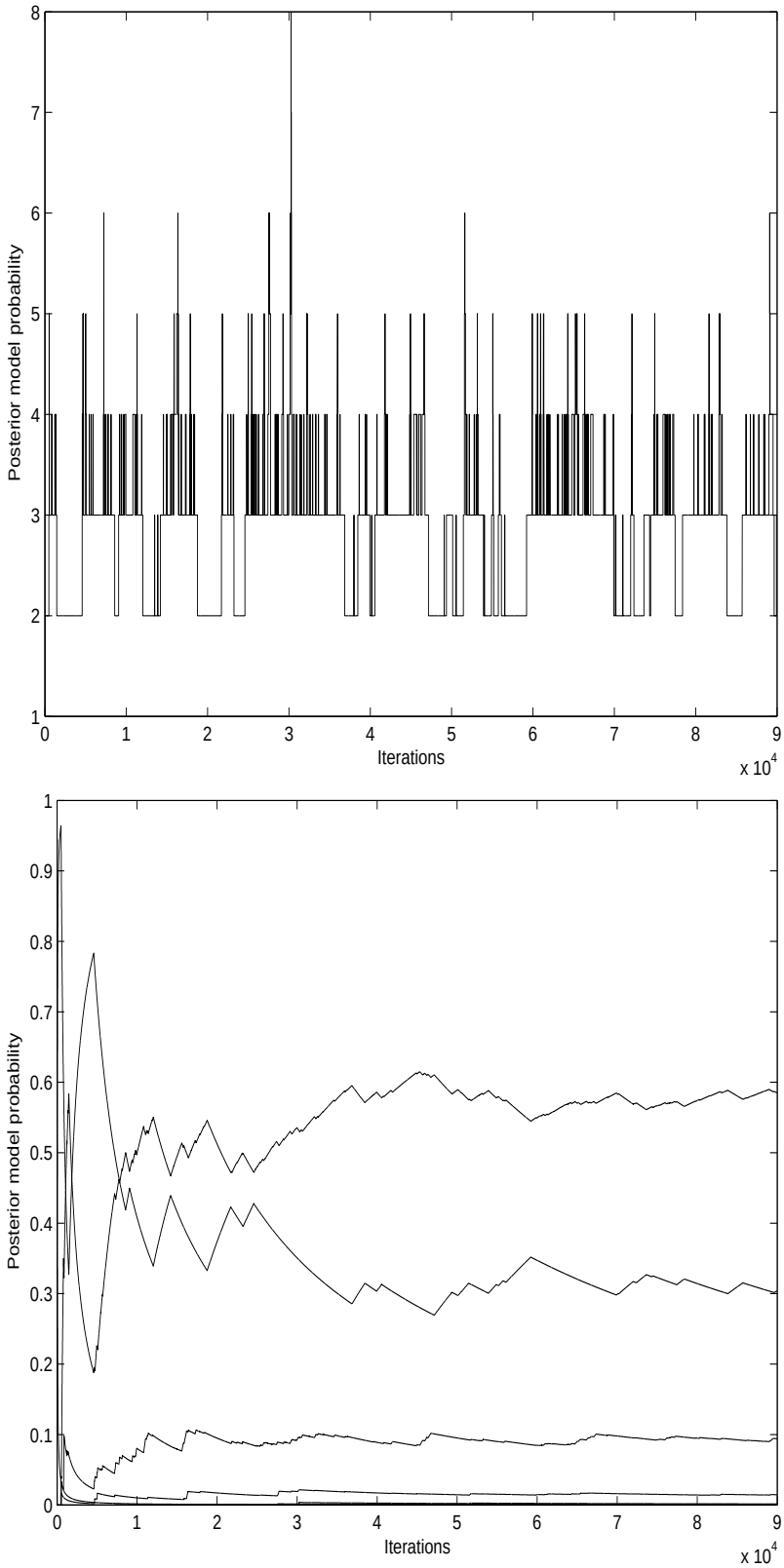

Figure 3: Convergence plots for the Old Faithful data; top: values of k; down: ergodic estimates of k, top line denotes $k = 3$ (probability 0.5854), second line from the top denotes $k = 2$ (probability 0.3035) and third line from the top denotes $k = 4$ (probability 0.0941).

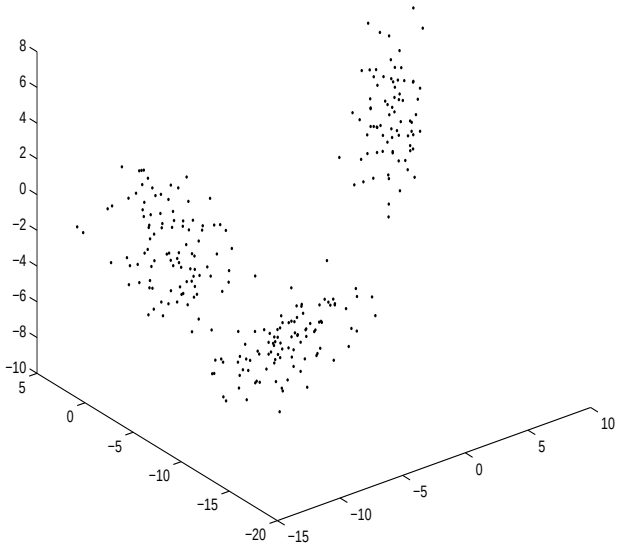

Figure 4: Simulated data in 3 dimensions

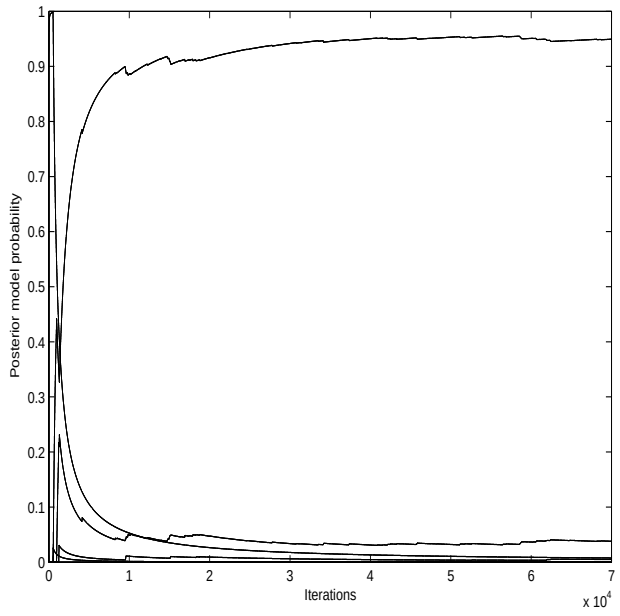

Figure 5: Ergodic estimates of posterior model probabilities for simulated data. Top line represents the model with 3 components with resulting posterior probability 0.9493.

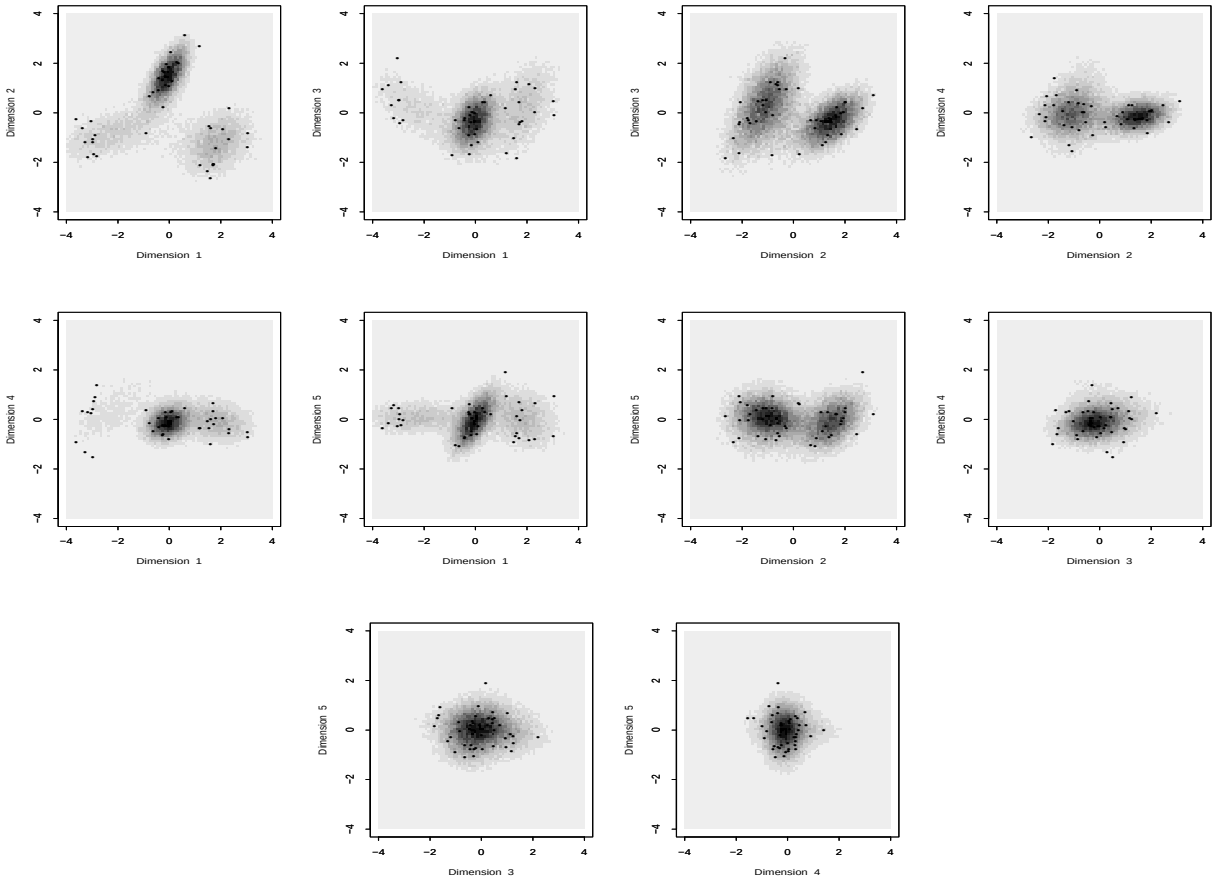

Figure 6: Data and image plots of the predictive densities for pottery data

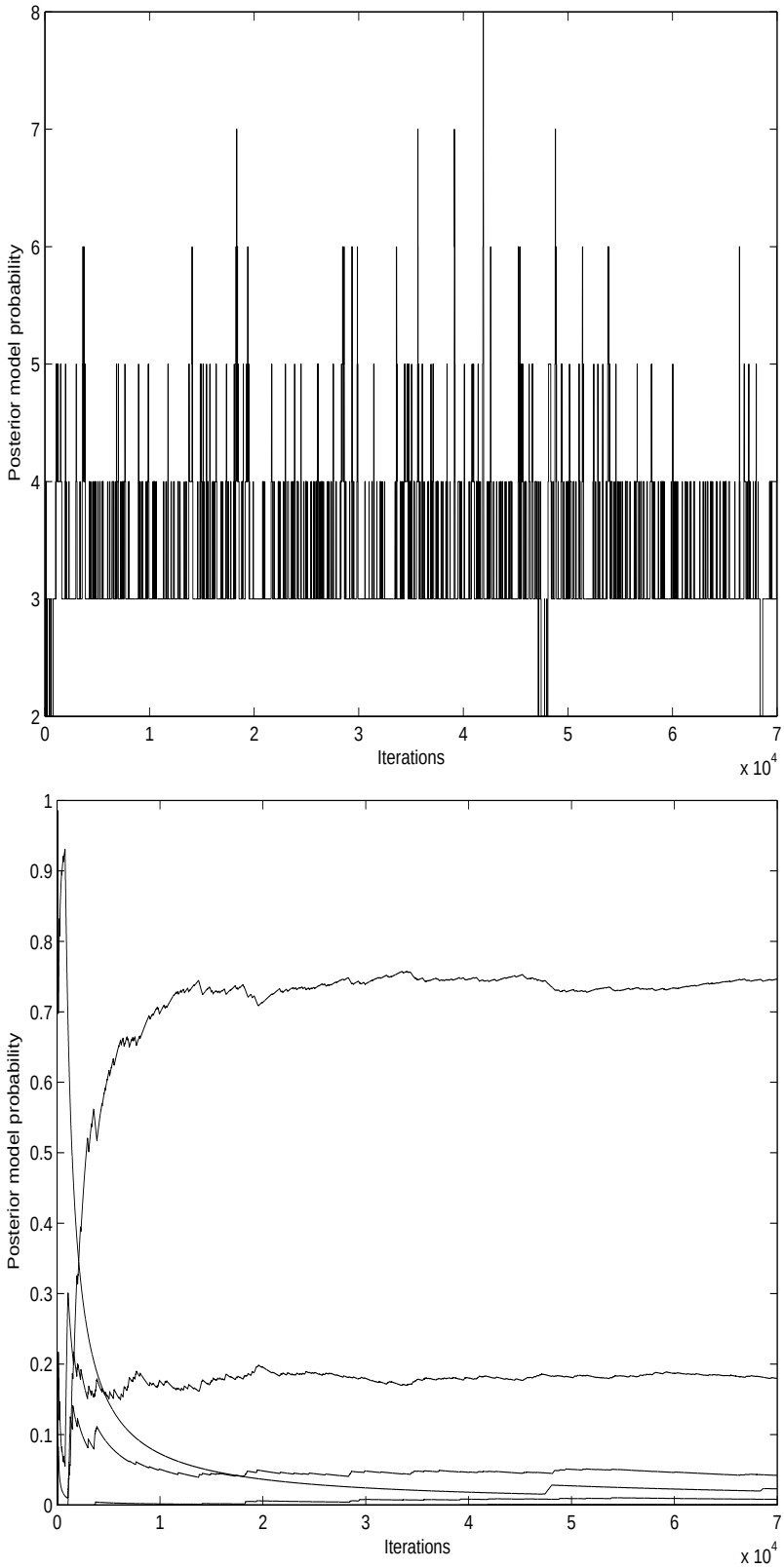

Figure 7: Convergence plots for the British pottery data; top: values of k; down: ergodic estimates of k, top line denotes $k = 3$ (probability 0.746), second line from the top denotes $k = 4$ (probability 0.18).