Birth-death MCMC methods for mixtures with an unknown number of components*

J.Q. Shi,
Depts. of Comp. Science & Statistics.
University of Glasgow,
Glasgow G12 8QQ,
Scotland, UK.
shi@dcs.gla.ac.uk

R. Murray-Smith,
Dept. of Computing Science,
University of Glasgow,
& Hamilton Institute,
National University of Ireland Maynooth.
rod@dcs.gla.ac.uk

D.M. Titterington,
Dept. of Statistics
University of Glasgow.
mike@stats.gla.ac.uk

Abstract

Mixture models have very wide application. However, the problem of determining the number of components, K say, is one of the challenging problems in this area. There are several approaches in the literature for testing K for different values. A recent alternative approach is to treat K as unknown, and to model K and the mixture component parameters jointly; see for example Richardson and Green (1997). In this paper, we propose an approach based on a birth-death process. In contrast with the approach in Stephens(2000), we make use of the latent indicators so that the approach can be used to solve the problems with missing data when calculation of the likelihood requires knowledge of the unobservable latent variables. Specifically, we use the method to analyse hidden Markov models with unknown numbers of states. The model and the algorithm are illustrated by some real examples.

Keywords: Bayesian inference; Birth-death process; Hidden Markov model; Markov chain Monte Carlo method; Mixtures with an unknown number of components.

*Technical Report TR-2002-117, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, UK. August 21, 2002

1 Introduction

Mixture models find many applications in different areas. The mixture model with a known number of components has been studied for several decades; see a comprehensive discussion by Titterington, Smith and Makov (1985), McLachlan and Basford (1988) and McLachlan and Peel (2000). However, a challenging problem has remained – how to select the number of components? A general approach is to test for different values of K, e.g. $K = 2$ against $K = 3$ by methods such as Bayes factors, likelihood ratio tests and penalized likelihood methods; see for example Kass and Raftery (1995), Spiegelhalter *et al.* (2002) and references therein. Those methods are generally quite complicated and it is difficult to use them in practice. An alternative recent approach has been to treat K as an unknown parameter, and to model K and the mixture component parameters jointly. Phillips and Smith (1996) proposed an approach based on an iterative jump-diffusion sampling algorithm. However, this approach requires the calculation of the *jump intensity*, and this involves integrals for which analytical forms do not exist in general. Using Green's (1995) reversible jump samplers, Richardson and Green (1997) proposed a reversible jump MCMC (RJMCMC) method. They used a *split-combine* move as well as a *birth-death* move to change the number of components of the mixtures. This has become a familiar approach in the literature, although it is still a great challenge to extend the approach to multivariate cases. Based on continuous time birth-death processes, a birth-death MCMC (BDMCMC) algorithm has been proposed by Stephens (2000). This method is easily implemented and can be used to deal with multivariate cases. Cappé, Robert and Rydén (2001) proved that RJMCMC converges to BDMCMC under certain conditions.

However, BDMCMC as proposed in Stephens (2000) does not make use of the latent indicators $\boldsymbol{z}$, so it is not easily applied to problems with missing data where calculation of the likelihood requires knowledge of the unobservable latent variable $\boldsymbol{z}$. For example, in the mixture models for spatially correlated data discussed in Green and Richardson (2000), an allocation model is used to model the allocation variables $\boldsymbol{z}$. The conditional likelihood for the observation is easily calculated when the value of $\boldsymbol{z}$ is given. However, it is very difficult to calculate the marginal likelihood, since it is an integral in terms of $\boldsymbol{z}$. Hidden Markov (chain) models (HMMs) provide another example in which the conditional likelihood has a very simple form given z, but it is difficult to calculate the marginal likelihood of the observations, and recursive methods have to be used; see Robert, Rydén and Titterington (1999) and Hurn, Justel and Robert (2001). In this paper, a modified birth-death MCMC method is proposed such that the latent indicators $\boldsymbol{z}$ are used as well as the mixture parameters. In our approach, a transformation of latent indicators is defined; then a birth-death process is constructed such that the stationary distribution is the posterior joint distribution of the latent indicators and the other parameters of interest. The modified BDMCMC can therefore be applied to a very wide class of models and problems. We apply the approach specifically to HMMs in this paper.

The paper is organized as follows: a brief review of the BDMCMC method is given in Section 2.1, and a modified BDMCMC method is proposed in Section 2.2. Section 3 describes HMMs, and applies the approach to this model. Section 4 examines the performance of the method on several numerical examples. Some discussion is given in Section 5. The detailed balance conditions are established and the limit theorem is proved in the appendices.

2 Birth-death MCMC method

2.1 A Birth-death process

A basic mixture model for observations $\boldsymbol{Y} = (y_1, \cdots, y_n)'$ can be defined via a vector of latent indicator variables $\boldsymbol{Z} = (z_1, \cdots, z_n)'$ as follows:

$$y_i | z_i = k \sim f_k(\cdot | \boldsymbol{\theta}_k). \tag{1}$$

The z_i are unobservable, and are supposed independently drawn from the distribution

$$P(z_i = k) = \pi_k, \ \ i = 1, \cdots, K, \tag{2}$$

in ordinary mixture models, or are distributed as a Markov chain in HMMs; see the details in the next section. In a Bayesian framework, inference is based on the posterior distribution of the unknown parameters $(K, \boldsymbol{V})$ and the latent vector $\boldsymbol{Z}$, i.e., $p(K, \boldsymbol{V}, \boldsymbol{Z} | \boldsymbol{Y})$, where $\boldsymbol{V} = (V_1, \cdots, V_K)$ and $V_k = (\pi_k, \boldsymbol{\theta}_k)$.

Stephens (2000) viewed each component of the mixture as a point in the parameter space, and constructed a birth-death process as follows:

Birth-Death Process. Suppose we use a fixed birth rate β, and that the current state is $\boldsymbol{V} = (V_1, \cdots, V_K) = \{(\pi_1, \boldsymbol{\theta}_1), \cdots, (\pi_K, \boldsymbol{\theta}_K)\}$,

1. Calculate the death rate γ_k for the kth component, and the total death rate $\gamma = \gamma_1 + \cdots + \gamma_K$.
2. Simulate the time to the next jump from an exponential distribution with mean $1/(\beta + \gamma)$.
3. Simulate the type of jump as a birth or a death with the birth rate $P_b = \beta/(\beta + \gamma)$ and death rate $1 - P_b$, and adjust $\boldsymbol{V}$ as follows:
   - i. If a birth occurs, then simulate $V = (\pi, \boldsymbol{\theta})$ from the density $b(\boldsymbol{V}; \pi, \boldsymbol{\theta})$. The new state is $\boldsymbol{V} \cup V = (V_1^*, \cdots, V_K^*, V)$, where $V_k^* = (\pi_k(1 - \pi), \boldsymbol{\theta}_k)$.
   - ii. Otherwise, if a death occurs, the kth component is selected to die with probability γ_k/γ. The new state is $\boldsymbol{V} \backslash V_k = (V_1^*, \cdots, V_{k-1}^*, V_{k+1}^*, \cdots, V_K^*)$, where $V_j^* = (\pi_j/(1 - \pi_k), \boldsymbol{\theta}_j)$ for any $j \neq k$.

In the above procedure, $b(\boldsymbol{V}; \pi, \boldsymbol{\theta})$ is the density function for simulating a new sample $(\pi, \boldsymbol{\theta})$ based on the current state $\boldsymbol{V}$. The death rate is given by

$$\gamma_k = \beta \frac{b(\boldsymbol{V} \backslash V_k; V_k)}{K(1 - \pi_k)^{K-2}} \frac{p(K - 1, \boldsymbol{V} \backslash V_k)}{p(K, \boldsymbol{V})} \frac{L(\boldsymbol{V} \backslash V_k)}{L(\boldsymbol{V})}, \tag{3}$$

where $p(K, \boldsymbol{V})$ is the prior density function of $(K, \boldsymbol{V})$. In general, we use the form

$$p(K, \boldsymbol{V}) = p(K)p(\boldsymbol{V}|K) = p(K)\ p(\pi_1, \cdots, \pi_K)\ p(\boldsymbol{\theta}_1) \cdots p(\boldsymbol{\theta}_K). \tag{4}$$

We can take $p(K)$ to be the density of a Poisson distribution with mean λ_0 or a discrete uniform on $\{k = 1, \cdots, K_{max}\}$, for example with $K_{max} = 30$. The prior density of π and $\boldsymbol{\theta}$ can be chosen in the usual way, as in Richardson and Green (1997) and Stephens (2000). $L(\boldsymbol{V})$ is the likelihood

$$L(\boldsymbol{V}) = p(\boldsymbol{Y}|\boldsymbol{V}) = \prod_{i=1}^{n}[\pi_1 f_1(y_i|\boldsymbol{\theta}_1) + \cdots + \pi_K f_K(y_i|\boldsymbol{\theta}_K)]. \tag{5}$$

It is obvious that the prior density and likelihood functions are invariant under permutation of the component labelling of components. By the proof in the appendix of Stephens (2000), the stationary density function of the above birth-death process is the posterior density $p(K, \boldsymbol{V}|\boldsymbol{Y}) = p(K, \boldsymbol{\pi}, \boldsymbol{\Theta}|\boldsymbol{Y})$, where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$.

If we take $b(\boldsymbol{V}; \pi, \boldsymbol{\theta}) = K(1-\pi)^{K-1}p(\boldsymbol{\theta})$, so that $(\pi, \boldsymbol{\theta})$ are simulated from their prior distributions and $\boldsymbol{\pi}$ has a uniform prior distribution, then the death rate (3) simplifies to

$$\gamma_k = \beta \frac{p(K-1)}{Kp(K)} \frac{L(\boldsymbol{V} \backslash V_k)}{L(\boldsymbol{V})}. \tag{6}$$

2.2 A modified birth-death process

The birth-death process discussed in the last subsection does not use the latent indicators $\boldsymbol{Z}$. When the indicators $\boldsymbol{Z}$ are given, the likelihood for $(\boldsymbol{V}, \boldsymbol{Z})$ has the form

$$L(\boldsymbol{V}, \boldsymbol{Z}) = p(\boldsymbol{Y}|\boldsymbol{V}, \boldsymbol{Z}) = \prod_{i=1}^{n} \pi_{z_i} p(y_i|\boldsymbol{\theta}_{z_i}), \tag{7}$$

which is easily calculated for a very wide class of models and problems. The marginal density $p(\boldsymbol{Y}|\boldsymbol{V})$ in (5) involves integrals with respect to $\boldsymbol{Z}$. It is very difficult to calculate this integral for some problems, such as the mixture models for spatially correlated data in Green and Richardson (2000), since $\boldsymbol{Z}$ itself is modelled by an allocation model. For some problems, the likelihood (5) can be calculated but only by an indirect method, as in the case of HMMs, for which a recursive method is used (see Robert, Rydén and Titterington, 1999), that leads to a heavy computational burden. Thus, we will construct a birth-death process that makes use of the latent indicators $\boldsymbol{Z}$. The idea is to treat $(K, \boldsymbol{V}, \boldsymbol{Z})$ as unknown parameters, and construct a modified birth-death process with the posterior $p(K, \boldsymbol{V}, \boldsymbol{Z}|\boldsymbol{Y})$ as its stationary distribution.

The key point of the birth-death process discussed in the last section is to view the posterior distribution of the unknown parameters as a marked point process. Let Ω_k denote the parameter space of the mixture model with k components, ignoring the labelling of the components, and let $\Omega = \cup_{k \geq 1} \Omega_k$. If Ω_k represents the parameter space of $\boldsymbol{V} = \{V_1, \cdots, V_k\}$, then $\boldsymbol{V}$ is a set of k points in Ω, and births and deaths in Ω are defined in Section 2.1. However, since the dimension of $\boldsymbol{Z}$ is n, we cannot extend the above birth-death process directly to deal with parameters $(\boldsymbol{V}, \boldsymbol{Z})$. Fortunately, this problem can be solved by defining a new parameter variable. Let $\boldsymbol{D}$ be a $K \times n$ matrix whose elements d_{ki} are defined as follows:

$$d_{ki} = \begin{cases} 1 & \text{if } z_i = k \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Thus, the ith column of $\boldsymbol{D}$ corresponds to z_i. The value of z_i dictates which element in the column has value 1, the others all being equal to 0. Therefore, the parameter space of $\boldsymbol{D}$ is $d_{ki} \in \{0, 1\}$ with restriction $\sum_k d_{ki} = 1$ for each i. For the kth row, $D_k = (d_{k1}, \cdots, d_{kn})$, the number of elements with value 1 is equal to the number of cases which z_i is equal to k. Thus, if the z_i's are independent, we have

$$p((z_1, \cdots, z_n)|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{c_k} = \prod_{k=1}^{K} \pi_k^{d_{k1}+\cdots+d_{kn}} = \prod_{k=1}^{K} p(D_k|\pi_k). \tag{9}$$

Let $\boldsymbol{W} = (W_1, \cdots, W_K) = ((\pi_1, \boldsymbol{\theta}_1, D_1), \cdots, (\pi_K, \boldsymbol{\theta}_K, D_K)) = ((V_1, D_1), \cdots, (V_K, D_K))$; then $\boldsymbol{W}$ is equivalent to $(\boldsymbol{V}, \boldsymbol{Z})$. If Ω_K represents the parameter space of $\boldsymbol{W} = \{W_1, \cdots, W_K\}$, then $\boldsymbol{W}$ is a set of K points in $\Omega = \cup_{k \geq 1} \Omega_k$. A birth-death process can be constructed for generating $(K, \boldsymbol{W})$ from the posterior distribution $P(K, \boldsymbol{W}|\boldsymbol{Y})$.

Modified Birth-Death Process. Assume that the current state is $\boldsymbol{W} = (W_1, \cdots, W_K)$. The basic procedure is the same as the birth-death process discussed in the last subsection, but Step 3 is modified as follows: 3. Simulate the type of jump as a birth or a death with birth rate $P_b = \beta/(\beta + \gamma)$ and death rate $1 - P_b$, and adjust $\boldsymbol{W}$ as follows,

i. If a birth occurs, then simulate $W = (\pi, \boldsymbol{\theta}, D)$ from the density $b(\boldsymbol{W}; W)$. The new state is defined by $\boldsymbol{W} \cup W = (W_1^*, \cdots, W_K^*, W)$, with $W_i^* = (\pi_i(1 - \pi), \boldsymbol{\theta}_i, D_i^*)$, where $d_{ij}^* = 0$, if $d_j = 1$, and $d_{ij}^* = d_{ij}$, if $d_j = 0$. Here d_{ij}^* and d_j are the elements of D_i^* and D respectively.

ii. If a death occurs, the kth component is selected to die with probability γ_k/γ. The new state is $\boldsymbol{W} \backslash W_k = \{W_1^*, \cdots, W_{k-1}^*, W_{k+1}^*, \cdots, W_K^*\}$ with $W_i^* = (\pi_i/(1 - \pi_k), \boldsymbol{\theta}_i, D_i^*)$ for all $i \neq k$, where $d_{ij}^* = d_{ij}$, if $d_{kj} = 0$, and $d_{ij}^* = 1$ with probability $\pi_i/(1 - \pi_k)$, if $d_{kj} = 1$.

The death rate is different from (3). If the prior density takes the form

$$p(K, \boldsymbol{W}) = p(K)p(\boldsymbol{W}|K) = p(K)p(\boldsymbol{\theta}_1) \cdots p(\boldsymbol{\theta}_K)p(\boldsymbol{\pi}|K)p(\boldsymbol{D}|\boldsymbol{\pi}, K), \tag{10}$$

the death rate is

$$\gamma_k = \gamma(\boldsymbol{W} \backslash W_k; W_k) = \beta \frac{b(\boldsymbol{W} \backslash W_k; W_k)}{K(1 - \pi_k)^{K-2}} \frac{p(K - 1, \boldsymbol{V} \backslash V_k)}{p(K, \boldsymbol{V})} \frac{L(\boldsymbol{W} \backslash W_k)}{L(\boldsymbol{W})} \frac{p(\boldsymbol{D} \backslash D_k|\boldsymbol{V} \backslash V_k)}{p(\boldsymbol{D}|\boldsymbol{V})}, \tag{11}$$

where the likelihood $L(\boldsymbol{W})$ is given by (7), and $p(K, \boldsymbol{V})$ is the prior density given by (4). If $\boldsymbol{\pi}$ takes the uniform prior distribution, $b(\boldsymbol{W}, W)$ takes the prior density of W, i.e. $p(W) = p(D|V)p(V)$ and $p(V) = K(1-\pi)^{K-1}p(\boldsymbol{\theta})$, the death rate simplifies to

$$\gamma_k = \beta \frac{p(K-1)}{Kp(K)} \frac{L(\boldsymbol{W} \backslash W_k)}{L(\boldsymbol{W})} \frac{p(D_k|V_k)p(\boldsymbol{D}\backslash D_k|\boldsymbol{V}\backslash V_k)}{p(\boldsymbol{D}|\boldsymbol{V})}. \tag{12}$$

Theorem 1. The stationary distribution of the birth-death process defined above is the distribution with posterior density $p(K, \boldsymbol{W}|\boldsymbol{Y}) = p(K, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{Y})$ in Ω.

The proof is given in Appendix 1.

A possible problem for the birth-death process above is that the death rate may be very high. To speed up the mixing of the sequence, the birth-death process can be run hybridized with a Gibbs sampler. The basic algorithm is described as follows.

Birth-Death MCMC Algorithm. The current state is denoted by superscript (t) at the tth iteration. Then the $(t+1)$th iteration is as follows:

(a) run a birth-death process for a fixed time t_0 from the starting state $(K^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\Theta}^{(t)})$ to the final state $(K^{(t')}, \boldsymbol{\pi}^{(t')}, \boldsymbol{\Theta}^{(t')})$ $(t' = t + t_0)$, and set $K^{(t+1)} = K^{(t')}$;

(b) sample $(z_1, \cdots, z_n)$ from $p(\boldsymbol{z}|K, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{Y})$;

(c) sample $(\pi_1, \cdots, \pi_K)$ from $p(\boldsymbol{\pi}|K, \boldsymbol{z}, \boldsymbol{\Theta}, \boldsymbol{Y})$;

(d) sample $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ from $p(\boldsymbol{\Theta}|K, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{Y})$.

Convergence of this hybrid birth-death MCMC method should be faster than that of the method based on the birth-death process alone.

3 Application to hidden Markov models

3.1 Hidden Markov models and priors

A general hidden Markov model can be represented by a mixture structure (z_i, y_i) with $z_i \in \{1, \cdots, K\}$, and $y_i|z_i = k \sim f_k(y_i)$, as defined in (1), where the y_i are conditionally independent for $i = 1, \cdots, n$ given the z_i. Here $f_k(y_i)$ is a density function with unknown parameter $\boldsymbol{\theta}_k$. For example, it may be the density function of a normal distribution,

$$y_i|z_i = k \sim N(0, \sigma_k^2), \tag{13}$$

as will be used in this section. Note that the unknown parameter in the kth component density is $\boldsymbol{\theta}_k = \{\sigma_k\}$ for this model. In HMMs, z_i is not observed and is distributed as a finite-state Markov chain with a $K \times K$ transition matrix $\boldsymbol{A} = (a_{kj})$, such that

$$a_{kj} = P(z_{i+1} = j | z_i = k).$$

In this HMM, if K is given, the unknown parameters are $\boldsymbol{A}$ and $\boldsymbol{\theta} = (\sigma_1, \cdots, \sigma_k)'$.

If we assume the number of states K is unknown, the joint density of all variables mentioned so far takes the form

$$p(K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{Y}) = p(\boldsymbol{Y}|K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{Z})p(K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{Z}) = p(\boldsymbol{Y}|K, \boldsymbol{\theta}, \boldsymbol{Z})p(K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{Z}).$$

The prior density takes the form

$$p(K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{Z}) = p(K)p(\boldsymbol{A}|K)p(\boldsymbol{Z}|\boldsymbol{A}, K)p(\boldsymbol{\theta}|K). \tag{14}$$

From the discussion in Robert, Ryden and Titterington (2000), we take the following specific priors. K has a prior uniform on $\{1, 2, \cdots, K_{max}\}$ for a specified K_{max}. Given K, we assume the rows of $\boldsymbol{A}$ are independent, each with a Dirichlet distribution $D_K(\delta, \cdots, \delta)$; we take $\delta = 1$. The σ_k are independent, each with the uniform distribution on $(0, \alpha)$. For robustness, we assume that α has a negative exponential distribution with mean $30 \max |y_i|$.

3.2 A birth-death process

First, we define a $K \times n$ matrix $\boldsymbol{D}$ as in (8). Therefore, the unknown quantities in the HMM are $(K, \boldsymbol{W}) = (K, \boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{D}) = (K, W_1, \cdots, W_K)$, where

$$W_k = (a_{1,k}, \cdots, a_{k-1,k}, \boldsymbol{a}_k, \boldsymbol{\theta}_k, D_k),$$

with $\boldsymbol{a}_k = (a_{k,1}, \cdots, a_{k,k})$ and $D_k = (d_{k,1}, \cdots, d_{k,n})$. The parameter space of $\boldsymbol{\theta}_k$ is R^q, where q is the dimension of $\boldsymbol{\theta}_k$. The matrix $\boldsymbol{A}$ is a transition matrix with parameter space $[0,1]^{K \times K}$ with the restrictions $\sum_j a_{kj} = 1$ for each k. Let Ω_k be the parameter space of $\boldsymbol{W}$ given k. A birth-death process is then constructed in $\Omega = \cup_{k \geq 1} \Omega_k$ such that the stationary density function is the posterior density $p(K, \boldsymbol{W}|\boldsymbol{Y})$.

From (14), the prior density has the form

$$p(K, \boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{D}) = p(K)p(\boldsymbol{\theta}_1) \cdots p(\boldsymbol{\theta}_K)p(\boldsymbol{A}|K)p(\boldsymbol{D}|\boldsymbol{A}, K). \tag{15}$$

We assume that the prior distribution is invariant under permutation of the component labels. The likelihood of $\boldsymbol{W}$ is

$$L(\boldsymbol{W}) = p(\boldsymbol{Y}|\boldsymbol{W}) = \prod_{i=1}^{n} f_{z_i}(y_i|\boldsymbol{\theta}_{z_i}). \tag{16}$$

It is obvious that $L(\boldsymbol{W})$ is invariant under permutation of the component labels.

If the current state is $\boldsymbol{W} \in \Omega_K$, the process is constructed such that it jumps to Ω_{K+1} with a fixed birth rate β, or jumps to Ω_{K-1} with death rate γ. When the process jumps from Ω_K to Ω_{K+1}, we should generate

$$W = (a_{1,K+1}, \cdots, a_{K,K+1}, \boldsymbol{a}_{K+1}, \theta_{K+1}, D_{K+1}), \tag{17}$$

where $\boldsymbol{a}_{K+1} = (a_{K+1,1}, \cdots, a_{K+1,K+1})$ and $D_{K+1} = (d_{K+1,1}, \cdots, d_{K+1,n})$. When W is generated, we define

$$\boldsymbol{W} \cup W = (\theta_1, \cdots, \theta_K, \theta_{K+1}, \boldsymbol{A} \cup A, \boldsymbol{D} \cup D_{K+1}),$$

where $A = (a_{1,K+1}, \cdots, a_{K,K+1}, \boldsymbol{a}_{K+1})$,

$$\boldsymbol{A} \cup A = \begin{pmatrix} & \cdots & & a_{1,K+1} \\ & a^*_{kj} & & \\ & \cdots & & a_{K,K+1} \\ a_{K+1,1} & & a_{K+1,K} & a_{K+1,K+1} \end{pmatrix}$$

with $a^*_{kj} = (1 - a_{k,K+1})a_{kj}$; $\boldsymbol{D} \cup D$ is a $(K+1) \times n$ matrix with D_{K+1} as its last row, and the other elements d^*_{kj} are defined as $d^*_{kj} = 0$, if $d_{K+1,j} = 1$, and $d^*_{kj} = d_{kj}$, if $d_{K+1,j} = 0$.

If the kth component is selected to die, we define

$$\boldsymbol{W} \backslash W_k = (\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_K, \boldsymbol{A} \backslash A_k, \boldsymbol{D} \backslash D_k),$$

where $\boldsymbol{A} \backslash A_k$ is a $(K-1) \times (K-1)$ matrix with elements $a_{k'j}/(1 - a_{kj})$ for all $k' \neq k$ and $j \neq k$. $\boldsymbol{D} \backslash D_k$ is a $(K-1) \times n$ matrix with elements $d^*_{k'j}$, where, for all $k' \neq k$, $d^*_{k'j} = d_{k'j}$, if $d_{kj} = 0$, and $d^*_{k'j} = 1$ with probability $\pi_{k'}/(1 - \pi_k)$, if $d_{kj} = 1$.

Along the lines of the modified birth-death process discussed in Section 2.2, a birth-death process for HMMs can be defined as follows:

Birth-Death Process. Suppose we use a fixed birth rate β and that the current state is $\boldsymbol{W} = (\theta_1, \cdots, \theta_K, \boldsymbol{A}, \boldsymbol{D})$.

Steps 1 and 2 are the same as the first two steps discussed in Section 2.1.

3. Simulate the type of jump as a birth or a death with birth rate $P_b = \beta/(\beta + \gamma)$ and death rate $1 - P_b$, and adjust $\boldsymbol{W}$ as follows:

   i. if a birth occurs, then simulate W from the density $b(\boldsymbol{W}; W)$, so that the new state is $\boldsymbol{W} \cup W$, which belongs to Ω_{K+1};

   ii. otherwise, if a death occurs, the kth component is selected to die with probability γ_k/γ, so that the new state is $\boldsymbol{W} \backslash W_k$ which belongs to Ω_{K-1}.

Theorem 2. If, in the birth-death process defined above, the death rate takes the form

$$\gamma_k = \beta \frac{b(\boldsymbol{W} \backslash W_k : W_k)}{K \prod_{j \neq k} (1 - a_{j,k})^{K-2}} \frac{L(\boldsymbol{W} \backslash W_k)}{L(\boldsymbol{W})} \frac{p(K-1, \boldsymbol{V} \backslash V_k)}{p(K, \boldsymbol{V})} \frac{p(\boldsymbol{D} \backslash D_k | \boldsymbol{V} \backslash V_k)}{p(\boldsymbol{D} | \boldsymbol{V})}, \tag{18}$$

then the stationary distribution of the birth-death process is the distribution with posterior density function $p(K, \boldsymbol{W} | \boldsymbol{Y})$.

The proof is given in Appendix 2.

The W in (17) is generated from the density function $b(\boldsymbol{W}; W)$. One example is

$$b(\boldsymbol{W}; W) = p(a_{1,K+1}) \cdots p(a_{K,K+1}) p(\boldsymbol{a}_{K+1}) p(\theta_{K+1}) p(D_{K+1} | \boldsymbol{A} \cup A).$$

We may generate $d_{K+1,i}$ independently from its stationary probability π_{K+1}, which satisfies

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{A}.$$

The density $p(a_{i,K+1})$ may be chosen to be the density function of a $Beta(1, K)$ distribution, and $p(\boldsymbol{a}_{K+1})$ may be taken to be the density function of the Dirichlet distribution $D_{K+1}(\delta, \cdots, \delta)$.

3.3 Birth-death MCMC algorithm

To speed up the mixing of the algorithm, we combine the birth-death process with the Gibbs sampler. The details of one sweep of the birth-death MCMC algorithm is therefore as follows:

Hybrid BD-MCMC Algorithm. Suppose the current state is denoted by superscript (t) at the tth iteration. Then a sweep of the next iteration is as follows:

(a) run the birth-death process for a fixed time t_0 starting from the state $(K^{(t)}, \boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{D}^{(t)})$, to give the final state $(K^{(t')}, \boldsymbol{A}^{(t')}, \boldsymbol{\theta}^{(t')}, \boldsymbol{D}^{(t')})$, where $t' = t + t_0$, and update K to $K^{(t')}$;

(b) update the transition probability matrix $\boldsymbol{A}$;

(c) update the parameter $\boldsymbol{\theta} = (\sigma_1, \cdots, \sigma_K)$;

(d) update the allocations $\boldsymbol{Z}$;

(e) update the hyperparameter α.

In steps (b) to (e), the number of components is fixed at the value updated in step (a), and the related variables are updated from the distribution conditioned on the current values of all other variables. Thus, steps (b) to (e) are exactly the same as steps (a) to (d) of the MCMC procedure proposed in Robert, Ryden and Titterington (2000). We omit the details here; the related results can be found in there and in Robert, Celeux and Diebolt (1993).

4 Examples

4.1 An example of HMM: Standard and Poors 500 data

The dataset used here consists of 1700 observations of the Standard and Poors 500 stock index during the 1950s. These data were previously analysed in Robert, Ryden and Titterington (2000). The data are presented in Figure 1. We used HMMs to model this dataset, and the Bayesian estimate of its marginal density is presented in Figure 2. For comparison, a kernel estimate is also presented in this Figure. They are quite similar to each other.

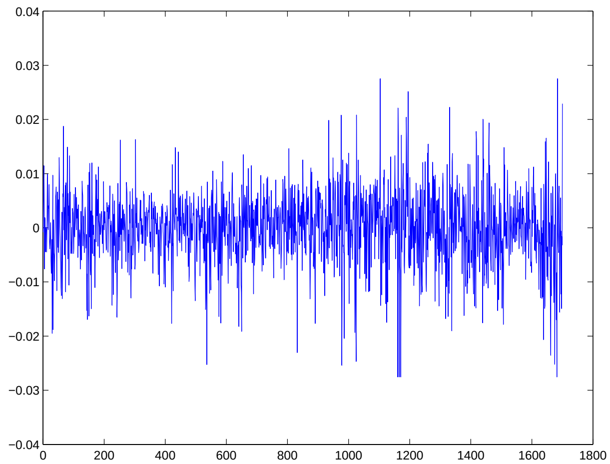

Figure 1: Standard and Poors 500 data.

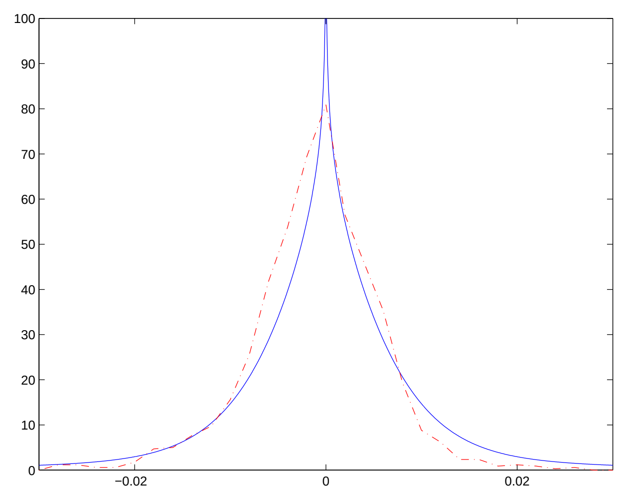

Figure 2: Standard and Poors 500 data: Bayesian estimate (—) and kernel estimate (- - -).

4.2 An example of multivariate case: Old Faithful Data

We now consider, as a multivariate example, Old Faithful data, which has been discussed by Ripley (1977), Stephens (2000) and others. There are data on 272 eruptions of the Old Faithful geyser in the Yellowstone National Park, as plotted in two dimensions in Figure 3. One is the duration in minutes of the eruption, and the other is the waiting time in minutes before the next eruption. There are two clear groups, and each seems have a little bit deviation from normality. A mixture model with three or four normal components should fit the data.

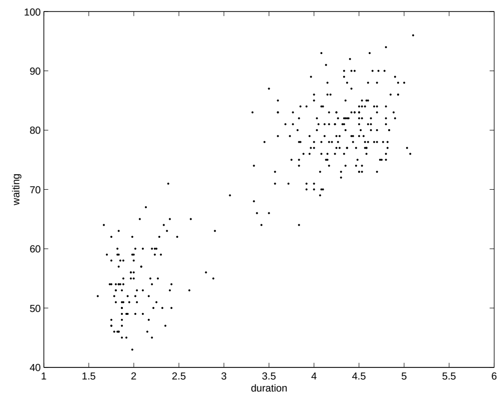

Figure 3: Old Faithful Data: scatter plot.

We use a mixture model with unknown number of components as discussed in Section 2. Each component has normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. We take the same prior distributions as the settings for variable-κ priors in Stephens (2000), and then the conditional distribution in steps (b) to (d) at the end of Subsection 2.2 can be found in Stephens' paper. The difference of the procedure here is that we use the modified Birth-Death process, i.e., we consider indicators $\boldsymbol{z}$ in Birth-Death process as well as the other unknown components. In practice, we find that it will speed up the mixture of the processes if we generate a new D from $p(D|V, \boldsymbol{Y})$.

To study the sensitivity to the choice of priors, we take different priors for the number of mixture components K. In this paper, we present the results with three different priors. One is the uniform distribution ranged from 1 to 30, and the other two are the Poisson distributions with parameter $\lambda = 1$ and $\lambda = 3$. The sampled values of K are presented in Figure 4, and the probability distributions of K are given by Figure 5. We can see that the posterior distributions of K are quite similar for all three different priors. The values 3 and 4 are most often sampled, which matches our previous informal analysis from the original data in Figure 3.

Figure 6 presents contour plots of the bivariate predictive density. Apparently, all of them fit the two groups of data in Figure 3 quite well. The differences between the predictive densities from the three different priors are very minor.

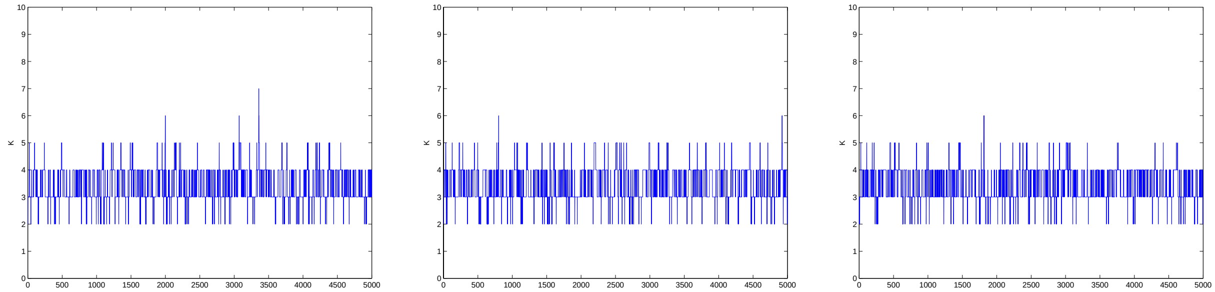

Figure 4: Old Faithful Data: sampled values of K. Left: uniform prior for K; Middle: Poisson prior for K with $\lambda = 1$; Right: Poisson prior for K with $\lambda = 3$.

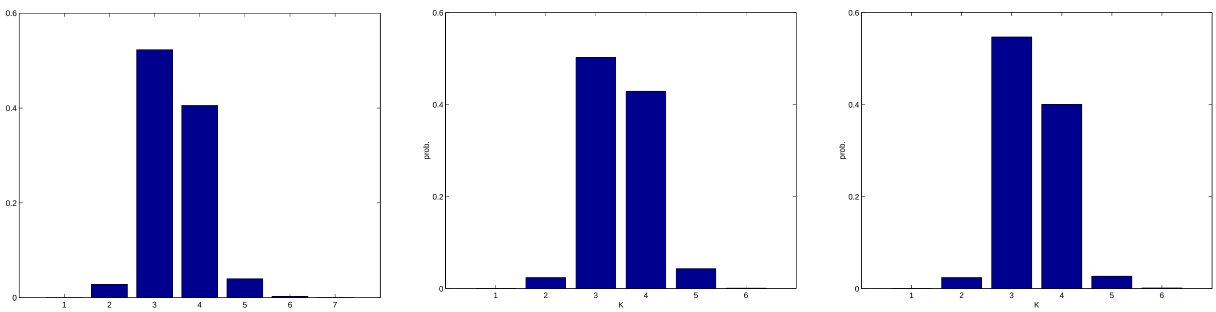

Figure 5: Old Faithful Data: probability distribution for K. Left: uniform prior for K; Middle: Poisson prior for K with $\lambda = 1$; Right: Poisson prior for K with $\lambda = 3$.

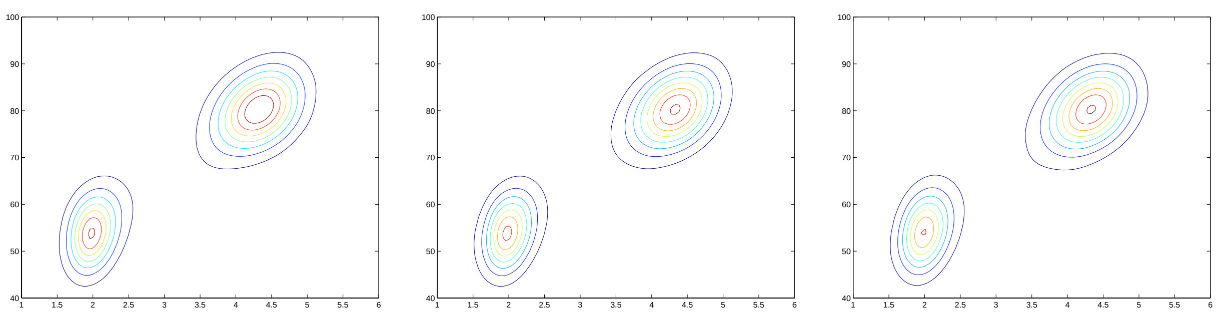

Figure 6: Old Faithful Data: contour plots of predictive density. Left: uniform prior for K; Middle: Poisson prior for K with $\lambda = 1$; Right: Poisson prior for K with $\lambda = 3$.

5 Discussion

The modified BDMCMC proposed in this paper can be viewed as an extension of the original BDMCMC proposed in Stephens (2000). It can be applied to complex problems including problems involving multivariate cases where it is difficult to calculate the likelihood of the observations without making use of the missing latent variable $\boldsymbol{z}$, as in case of the HMMs discussed in this paper. Moreover, for the mixture model with the allocation model, such as the model discussed in Green and Richardson (2000), it is essential to use the modified BDMCMC since the marginal likelihood involves an integral.

If the marginal likelihood can be calculated easily without knowledge of $\boldsymbol{z}$, we agree with the point of view indicated in Cappé, Robert and Rydén (2001) that the original BDMCMC approach without using $\boldsymbol{z}$ may be more effective. As a result of the smaller dimensionality of the parameter space after marginalising over $\boldsymbol{z}$, the mixing of the iterative processes in BDMCMC is faster than by the modified BDMCMC or by RJMCMC.

We applied the approach to HMMs with a normal distribution (13) used for the 'noise' model. The discussion in Section 3 shows that it is easy to extend the approach to the case of normal distributions with nonzero means, i.e., $N(\mu_k, \sigma_k^2)$ in (13). It can also be applied to HMMs with other types of density function, such as the model discussed in Boys, Henderson and Wilkinson (2000). The only difference is that equation (1) in their paper is used to calculate the conditional likelihood in (16).

Appendix 1. The proof of Theorem 1

The parameters are $(K, \boldsymbol{W}) = (K, \boldsymbol{V}, \boldsymbol{D}) = (K, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{D})$. The parameter space of $\boldsymbol{\theta}$ is $\mathcal{R}^Q$. The parameter space of $\boldsymbol{\pi}$ is $[0,1]^K$ with the restriction $\pi_1 + \cdots + \pi_K = 1$. Each d_{ki} belongs to $\{0,1\}$, subject to $\sum_k d_{ki} = 1$. Let Ω_K be the parameter space of the mixture model with K components. Since the likelihood $L(\cdot)$ in (7) and the prior distribution $p(\cdot)$ in (10) are invariant under permutation of the component labels, we ignore the labelling of the components for Ω_k. Moreover, we write $\Omega_K = \Omega_K^V \otimes \Omega_K^D$, where Ω_K^V and Ω_K^D represent the parameter spaces of $\boldsymbol{V}$ and $\boldsymbol{D}$ respectively. Let $\Omega = \cup_{k \geq 1} \Omega_k$.

A birth-death process is defined in Section 2.2. When the current state is $\boldsymbol{W} \in \Omega_K$, the process jumps to Ω_{K+1} with birth rate $\beta(\boldsymbol{W})$, and jumps to Ω_{K-1} with death rate $\delta(\boldsymbol{W})$. When a birth occurs, the birth kernel $T_B^{(K)}(\boldsymbol{W}; G)$ is the probability that the process jumps to the particular set $G \in \Omega_{k+1}$. It satisfies

$$T_B^{(K)}(\boldsymbol{W}; G) = \int_{(\pi, \boldsymbol{\theta}) \in G^V} \sum_{D \in G^D} b(\boldsymbol{W}; (\pi, \boldsymbol{\theta}, D)) d\pi \nu(d\boldsymbol{\theta}),$$

where $\nu(d\boldsymbol{\theta})$ is the probability measure induced by $p(\boldsymbol{\theta})$, and G^V and G^D are the parts of the set $\{(\pi, \boldsymbol{\theta}, D) : \boldsymbol{W} \cup (\pi, \boldsymbol{\theta}, D) \in G\}$ corresponding to $(\pi, \boldsymbol{\theta})$ and D respectively. Similarly, when a death occurs, the death kernel $T_D^{(K)}(\boldsymbol{W}; F)$ is the probability that the process jumps to the particular set $F \in \Omega_{K-1}$, which satisfies

$$T_D^{(K)}(\boldsymbol{W}; F) = \sum_{(\pi, \boldsymbol{\theta}, D): \boldsymbol{W} \backslash (\pi, \boldsymbol{\theta}, D) \in F} \frac{\gamma(\boldsymbol{W} \backslash (\pi, \boldsymbol{\theta}, D); (\pi, \boldsymbol{\theta}, D))}{\gamma(\boldsymbol{W})},$$

where $\gamma(\boldsymbol{W})$ is the total death rate.

It is well known that the stationary distribution of the birth-death process is $p(K, \boldsymbol{W} | \boldsymbol{Y})$ if the following detailed balance conditions are satisfied (Preston, 1976; Stephens, 2000):

$$\int_F \beta(\boldsymbol{W}) d\mu_K(\boldsymbol{W}) = \int_{\Omega_{K+1}} \gamma(\boldsymbol{U}) T_D^{(K+1)}(\boldsymbol{U}; F) d\mu_{K+1}(\boldsymbol{U}), \tag{19}$$

for $K \geq 0$ and $F \subset \Omega_K$, and

$$\int_G \gamma(\boldsymbol{U}) d\mu_{K+1}(\boldsymbol{U}) = \int_{\Omega_K} \beta(\boldsymbol{W}) T_B^{(K)}(\boldsymbol{W}; G) d\mu_K(\boldsymbol{W}), \tag{20}$$

for $K \geq 0$ and $G \subset \Omega_{K+1}$, where μ_K is the probability measure induced on Ω_K by the posterior distribution $p(K, \boldsymbol{W} | \boldsymbol{Y})$. Here $\boldsymbol{U}$ has the same meaning as $\boldsymbol{W}$. We use the different notation to distinguish elements in the space Ω_K from those in Ω_{K+1}.

The left-hand side of the detailed balance condition (19) can be written as

$$\begin{aligned}
& \int_F \beta(\boldsymbol{W}) d\mu_K(\boldsymbol{W}) = \int_F \beta(\boldsymbol{W}) dp(K, \boldsymbol{W} | \boldsymbol{Y}) \\
= \; & c \int_F \beta(\boldsymbol{W}) L(\boldsymbol{W}) dp(K, \boldsymbol{W}) = c \int_{F^V} \sum_{\boldsymbol{D} \in F^D} \beta(\boldsymbol{W}) L(\boldsymbol{W}) p(\boldsymbol{D} | \boldsymbol{V}) dp(K, \boldsymbol{V}) \\
= \; & c \int_{F^V} \int_R \int_0^1 \sum_{d_k \in \{0,1\}} b(\boldsymbol{W}; W) \sum_{\boldsymbol{D} \in F^D} \beta(\boldsymbol{W}) L(\boldsymbol{W}) p(\boldsymbol{D} | \boldsymbol{V}) dp(K, \boldsymbol{V}) d\pi d\boldsymbol{\theta}, \qquad (21)
\end{aligned}$$

where c is the normalising constant of the posterior density. The last equation comes from the fact that the integral of $b(\boldsymbol{W}; W)$ is equal to one. The right-hand side of (19) can be written as

$$\begin{aligned} R &= \int_{\Omega_{K+1}} \gamma(\boldsymbol{U}) T_D^{(K+1)}(\boldsymbol{U}; F) d\mu_{K+1}(\boldsymbol{U}) \\ &= \int_{\Omega_{K+1}} \sum_{W:\boldsymbol{U}\backslash W \in F} \gamma(\boldsymbol{U}\backslash W : W) dp(K+1, \boldsymbol{U}|\boldsymbol{Y}) \\ &= c^* \int_{\Omega_{K+1}} \sum_{i=1}^{K+1} I(\boldsymbol{U}\backslash W_i \in F) \gamma(\boldsymbol{U}\backslash W_i : W_i) L(\boldsymbol{U}) dp(K+1, \boldsymbol{U}). \end{aligned}$$

Since the integrand is invariant under permutation of the component labels, and if we write $\boldsymbol{W} = \boldsymbol{U}\backslash W$, we have

$$R^* = R/(c^*(K+1)) = \int_{\Omega_{K+1}} I(\boldsymbol{W} \in F) \gamma(\boldsymbol{W} : W) L(\boldsymbol{W} \cup W) dp(K+1, \boldsymbol{W} \cup W)$$

$$\begin{aligned} &= \int_{\Omega^V_{K+1}} \sum_{\boldsymbol{D} \cup D \in \Omega^D_{K+1}} I(\boldsymbol{V} \in F^V; \boldsymbol{D} \in F^D) \gamma(\boldsymbol{W} : W) L(\boldsymbol{W} \cup W) \\ &\qquad \cdot p(\boldsymbol{D} \cup D | \boldsymbol{V} \cup V) dp(K+1, \boldsymbol{V} \cup V) \\ &= \int_{\Omega^V_{K+1}} \sum_{d_k \in \{0,1\}} \sum_{\boldsymbol{D} \in F^D} I(\boldsymbol{V} \in F^V) \gamma(\boldsymbol{W} : W) L(\boldsymbol{W} \cup W) p(\boldsymbol{D} \cup D | \boldsymbol{V} \cup V) dp(K+1, \boldsymbol{V} \cup V). \end{aligned}$$

Note that, since $(\boldsymbol{V} \cup V) = [((1-\pi)\pi_1, \boldsymbol{\theta}_1), \cdots, ((1-\pi)\pi_K, \boldsymbol{\theta}_K), (\pi, \boldsymbol{\theta})]$ and $\pi_1 + \cdots + \pi_K = 1$, we have

$$dp(K+1, \boldsymbol{V} \cup V) = p(K+1, \boldsymbol{V} \cup V)(1-\pi)^{K-1} d\boldsymbol{\Theta} d\boldsymbol{\pi} d\pi d\theta = J dp(K, \boldsymbol{V}) d\pi d\boldsymbol{\theta},$$

where

$$J = \frac{p(K+1, \boldsymbol{V} \cup V)}{p(K, \boldsymbol{V})} (1-\pi)^{K-1}.$$

By the above equation, R^* can be expressed as

$$R^* = \int_R \int_0^1 \int_{F^V} \sum_{d_k \in \{0,1\}} \sum_{\boldsymbol{D} \in F^D} \gamma(\boldsymbol{W} : W) L(\boldsymbol{W} \cup W) p(\boldsymbol{D} \cup D | \boldsymbol{V} \cup V) J dp(K, \boldsymbol{V}) d\pi d\boldsymbol{\theta}. \quad (22)$$

If we compare (21) and (22), the detailed balance condition (19) is satisfied when the following equation holds:

$$c^*(K+1) \gamma(\boldsymbol{W} : W) L(\boldsymbol{W} \cup W) p(\boldsymbol{D} \cup D | \boldsymbol{V} \cup V) J = c b(\boldsymbol{W}; W) \beta(\boldsymbol{W}) L(\boldsymbol{W}) p(\boldsymbol{D} | \boldsymbol{V}). \quad (23)$$

This leads to (11). Note that c and c^* are the normalising constants for the related posterior distributions. As a result of the invariance under permutation of the component labels, c and c^* can be ignored when we calculate the death probability (11). The proof that condition (23) implies the detailed balance condition (20) is similar.

Appendix 2: The proof of Theorem 2

Let $\boldsymbol{V} = (\boldsymbol{A}, \boldsymbol{\theta})$. Then the parameters are $(K, \boldsymbol{W}) = (K, \boldsymbol{V}, \boldsymbol{D}) = (K, \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{D})$. As defined in Section 3.2, Ω_K is the parameter space of the mixture model with K components, and $\Omega = \cup_{k\geq 1}\Omega_k$. Since the likelihood $L(\cdot)$ in (16) and the prior distribution $p(\cdot)$ in (15) are invariant under permutation of the component labels, we ignore the labelling of the components for Ω_K. Moreover, we write $\Omega_K = \Omega_K^V \otimes \Omega_K^D$, where Ω_K^V and Ω_K^D represent the parameter spaces of $\boldsymbol{V}$ and $\boldsymbol{D}$ respectively. By a discussion similar to that in Appendix 1, the stationary density function of the birth-death process defined in Section 3.2 is the posterior density $p(k, \boldsymbol{W}|\boldsymbol{Y})$ if the following detailed balance conditions are satisfied (see also equations (8.8) and (8.9) in Preston, 1976):

$$\int_F \beta(\boldsymbol{W}) d\mu_K(\boldsymbol{W}) = \int_{\Omega_{K+1}} \sum_{\substack{k \in \{1, \cdots, K+1\} \\ \boldsymbol{U}\backslash W_k \in F}} \gamma(\boldsymbol{U}\backslash W_k; W_k) d\mu_{K+1}(\boldsymbol{U}), \tag{24}$$

for $K \geq 0$ and $F \subset \Omega_K$, and

$$\int_G \sum_{k=1}^{K+1} \gamma(\boldsymbol{U}\backslash W_k; W_k) d\mu_{K+1}(\boldsymbol{U}) = \int_{\boldsymbol{W} \cup W \in G} \beta(\boldsymbol{W}) b(\boldsymbol{W}; W) d\mu_K(\boldsymbol{W}), \tag{25}$$

for $K \geq 0$ and $G \subset \Omega_{K+1}$, where μ_K is the probability measure induced on Ω_K by the posterior distribution $p(K, \boldsymbol{W}|\boldsymbol{Y})$. This can be expressed by Bayes Theorem as

$$p(K, \boldsymbol{W}|\boldsymbol{Y}) = cL(\boldsymbol{W})p(K, \boldsymbol{W}),$$

where c is the normalising constant. It is obvious that $p(K, \boldsymbol{W}|\boldsymbol{Y})$ is invariant under permutation of the component labels.

We now study the detailed balance conditions. The left-hand side of (24) can be expressed as

$$\begin{aligned} & \int_F \beta(\boldsymbol{W}) d\mu_K(\boldsymbol{W}) = \int_F \beta(\boldsymbol{W}) dp(K, \boldsymbol{W}|\boldsymbol{Y}) \\ = \;& c\int_F \beta(\boldsymbol{W}) L(\boldsymbol{W}) dp(K, \boldsymbol{W}) = c\int_{F^V} \sum_{\boldsymbol{D} \in F^D} \beta(\boldsymbol{W}) L(\boldsymbol{W}) p(\boldsymbol{D}|\boldsymbol{V}) dp(K, \boldsymbol{V}). \end{aligned} \tag{26}$$

Since the posterior density $p(K+1, \boldsymbol{U}|\boldsymbol{Y})$ is invariant under permutation of the component labels, the right-hand side of (24) can be derived as

$$\begin{aligned} R &= \int_{\Omega_{K+1}} \sum_{\substack{k \in \{1, \cdots, K+1\} \\ \boldsymbol{U}\backslash W_i \in F}} \gamma(\boldsymbol{U}\backslash W_k; W_k) dp(K+1, \boldsymbol{U}|\boldsymbol{Y}) \\ &= c^*(K+1) \int_{\mathbf{U}\backslash W_{K+1} \in F} \gamma(\boldsymbol{U}\backslash W_{K+1}; W_{K+1}) dp(K+1, \boldsymbol{U}|\boldsymbol{Y}). \end{aligned}$$

Let $\boldsymbol{W} = \boldsymbol{U} \backslash W_{K+1}$. We then have

$$R^* = R/(c^*(K+1)) = \int_{\Omega_{K+1}} I(\boldsymbol{W} \in F)\gamma(\boldsymbol{W} : W)L(\boldsymbol{W} \cup W)dp(K+1, \boldsymbol{W} \cup W)$$

$$= \int_{\Omega^V_{K+1}} \sum_{\boldsymbol{D} \cup D \in \Omega^D_{K+1}} I(\boldsymbol{V} \in F^V; \boldsymbol{D} \in F^D)\gamma(\boldsymbol{W} : W)L(\boldsymbol{W} \cup W)$$

$$\cdot p(\boldsymbol{D} \cup D | \boldsymbol{V} \cup V)dp(K+1, \boldsymbol{V} \cup V)$$

$$= \int_{\Omega^V_{K+1}} \sum_{D \in \mathcal{R}^D} \sum_{\boldsymbol{D} \in F^D} I(\boldsymbol{V} \in F^V)\gamma(\boldsymbol{W} : W)L(\boldsymbol{W} \cup W)p(\boldsymbol{D} \cup D|\boldsymbol{V} \cup V)dp(K+1, \boldsymbol{V} \cup V),$$

where $\mathcal{R}^D$ is the parameter space of D, corresponding to which each component of D takes a value from $\{0, 1\}$. Note that, since $(\boldsymbol{V} \cup V) = [\theta_1, \cdots, \theta_k, \theta, \boldsymbol{A}^*, A]$ and $a_{k1} + \cdots + a_{kK} = 1$, we have

$$\begin{aligned} dp(K+1, \boldsymbol{V} \cup V) &= p(K+1, \boldsymbol{V} \cup V)[\prod_{k=1}^{K}(1 - a_{k,K+1})^{K-1}]d\boldsymbol{\theta} d\boldsymbol{A} d\theta dA \\ &= J dp(K, \boldsymbol{V}) d\theta dA, \end{aligned}$$

where

$$J = \frac{p(K+1, \boldsymbol{V} \cup V)}{p(K, \boldsymbol{V})}[\prod_{k=1}^{K}(1 - a_{k,K+1})^{K-1}].$$

By the above equation, R^* can be expressed as

$$R^* = \int_{\mathcal{R}^\theta} \int_{\mathcal{R}^A} \int_{F^V} \sum_{D \in \mathcal{R}^D} \sum_{\boldsymbol{D} \in F^D} \gamma(\boldsymbol{W} : W)L(\boldsymbol{W} \cup W)p(\boldsymbol{D} \cup D|\boldsymbol{V} \cup V)J dp(K, \boldsymbol{V}) d\theta dA, \quad (27)$$

where $\mathcal{R}^\theta$ is the parameter space of θ, $\mathcal{R}^A$ is the parameter space of A and each component of A takes a value from $[0, 1]$ with some restrictions. Note that the integral of $b(\boldsymbol{W}; W)$ is equal to one over the space $\mathcal{R}^\theta \times \mathcal{R}^A \times \mathcal{R}^D$. Similarly to the derivation in equation (21), we introduce this fact into (26) and compare it with (27), showing that the detailed balance condition (24) is satisfied when the following equation holds:

$$c^*(K+1)\gamma(\boldsymbol{W} : W)L(\boldsymbol{W} \cup W)p(\boldsymbol{D} \cup D|\boldsymbol{V} \cup V)J = cb(\boldsymbol{W}; W)\beta(\boldsymbol{W})L(\boldsymbol{W})p(\boldsymbol{D}|\boldsymbol{V}). \quad (28)$$

The death rate is therefore expressed by (18).

By a similar proof, (18) implies the other balance condition (25).

References

Boys, R. J., Henderson, A. H. and Wilkinson, D. J. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Appl. Statist.*, **49**, 269-285.

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.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.*, **95**, 957-979.

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

Green, P. J. and Richardson, S. (2000). Spatially correlated models for count data. Technical Report. University of Bristol.

Hurn, M., Justel, A. and Robert, C. P. (2001). Estimating mixtures of regressions. *J. Statist. Comput. Simul.* (to appear).

Kass, R. E. and Raftery, A. (1999). Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773-795.

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

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), London: Chapman and Hall, pp. 215-239.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc. B*, **59**, 731-792.

Ripley, B. D. (1977).Moelling spatial patterns (with discussion). *J. R. Statist. Soc. B*, **39**, 172-212.

Preston, C. J. (1976). Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, **46**, 371-391.

Robert, C. P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statist. Probab. Lett.*, **16**, 77-83.

Robert, C. P., Rydén, T. and Titterington, D. M. (1999). Convergence controls for MCMC algorithms, with application to hidden Markov chains. *J. Statist. Comput. Simul.*, **64**, 327-355.

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. *J. R. Statist. Soc. B*, **62**, 57-75.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R.*

Statist. Soc. B, **64**, to appear.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of componenets – an alternative to reversible jump methods. *Ann. Statist.*, **28**, 40-74.

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