Reversible jump, birth-and-death, and more general continuous time MCMC samplers

Olivier Cappé

CNRS / ENST Département TSI, 75634 Paris cedex 13, France

Christian P. Robert

CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16, France and CREST, INSEE

Tobias Rydén†

Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden

Summary. Reversible jump methods are the most commonly used Markov Chain Monte Carlo tool for exploring variable dimension statistical models. Recently however, an alternative approach based on birth-and-death processes has been proposed by Stephens (2000) in the case of mixtures of distributions. We show that the birth-and-death setting can be generalised to include other types of continuous time jumps like split-and-combine moves in the spirit of Richardson and Green (1997). We illustrate these extensions both for mixtures of distributions and for hidden Markov models. We demonstrate the strong similarity of reversible jump and continuous time methodologies by showing that upon appropriate rescaling of time, the reversible jump chain converges to a limiting continuous time birth-and-death process. A numerical comparison in the setting of mixtures of distributions highlights this similarity.

AMS 1991 classification. Primary 62C05. Secondary 60J05, 62F15, 65D30, 65C60.

Keywords: Birth-and-death process, hidden Markov model, MCMC algorithms, mixture distribution, Rao–Blackwellisation, rescaling

1. Introduction

Markov Chain Monte Carlo (MCMC) methods for statistical inference, in particular Bayesian inference, have become standard during the past ten years (Cappé and Robert, 2000). For variable dimension problems, often arising through model selection, a popular approach is Green's (1995) reversible jump MCMC (RJMCMC) methodology. Recently however, in the context of mixtures of distributions, Stephens (2000) rekindled interest in use of continuous time birth-and-death processes for variable dimension problems, following earlier proposals by Ripley (1977), Geyer and Møller (1994), Grenander and Miller (1994) and Phillips and Smith (1996). We will call this approach birth-and-death MCMC (BDMCMC) and its generalisations continuous time MCMC (CTMCMC).

In this paper, we investigate the similarity between the reversible jump and birth-and-death methodologies. In particular, it is shown in Section 4 that for any BDMCMC process satisfying some weak regularity conditions, there exists a sequence of RJMCMC processes that converges, in a sense specified below, to the BDMCMC process.

In their application of RJMCMC to mixtures of distributions, Richardson and Green (1997) implemented two types of moves that could change the number of components of the mixture: one was *birth-and-death*, in which a new component is created or an existing one is deleted, and the other was *split-and-combine*, in which one component is split in two, or two components are combined into one. In contrast, Stephens (2000) only dealt with birth-and-death moves in order to keep the algorithm within the theory of marked point processes on general spaces, while pointing out that "one can envision a continuous time version of the general reversible jump formulation." We show here that continuous time algorithms are not limited to the birth-and-death structure and that convergence of reversible jump to birth-and-death MCMC is much more general. For

†Supported by the Swedish Research Council.

example, split-and-combine moves could be incorporated, resulting in more general CTMCMC algorithms, and the appropriate theoretical framework is that of Markov jump processes. To complete our study of the connections between RJMCMC and CTMCMC, we implemented a full-scale numerical comparison with moves similar to those in Richardson and Green (1997) used in both algorithms: the outcome is the same for both samplers, with a longer execution time for CTMCMC.

The paper is organised as follows: in Section 2 we review the main features of the BDMCMC methodology, including moves more general than birth-and-death in Section 2.4 and variance reduction techniques in Section 2.5. This technology is exemplified for hidden Markov models in Section 3. Section 4 addresses the comparison of this approach with reversible jump MCMC methodology, recalling the basics of RJMCMC in Section 4.1, establishing convergence of RJMCMC to BDMCMC in Section 4.2 and detailing the numerical comparison of both algorithms in Section 4.4. Section 5 concludes the paper with a discussion.

2. Continuous time MCMC methodologies

In this section we review BDMCMC in the mixture case considered by Stephens (2000) and discuss the extension of the birth-and-death moves to other continuous time moves. While Stephens (2000) provides a full description of the method in the specific set-up of mixtures of distributions, continuous time MCMC is limited neither to birth-and-death moves, nor to mixture models. For example, CTMCMC may be applied to any of the examples in Green (1995). See also Ripley (1977), Geyer and Møller (1994), Grenander and Miller (1994) and Phillips and Smith (1996), where broader descriptions of continuous time approaches can be found. In particular, Ripley (1977) introduces the concept of simulating a birth-and-death process to approximate its limiting distribution, even though he was interested in a problem of fixed dimension, while Geyer and Møller (1994) propose a Metropolis–Hastings algorithm for spatial point processes and argue on the superiority of this scheme compared with a continuous time approach, as do Clifford and Nicholls (1994).

2.1. A reference example: mixture models

Our benchmark is a *mixture model*, with probability density function of the form

$$p(y|k,\mathbf{w},\boldsymbol{\phi}) = \sum_{i=1}^{k} w_i f(y|\phi_i),$$

where k is the unknown number of components, $\mathbf{w} = (w_1, \ldots, w_k)$ are the component weights, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are the component parameters and $f(\cdot|\phi)$ is some parametric class of densities indexed by a parameter ϕ, like the Gaussian, the Gamma, the Beta, or the Poisson family. The component weights are non-negative numbers summing to unity. Mixture models have been extensively considered in the literature, but remain a challenging setting for variable dimension techniques.

The above densities are written as conditional on the parameter ϕ, given the Bayesian perspective of the paper. Hence we need to specify a prior density for $(k, \mathbf{w}, \boldsymbol{\phi})$, denoted by $r(k, \mathbf{w}, \boldsymbol{\phi})$. Here, r is a density with respect to a product measure, made of the counting measure in the k-dimension and of the Lebesgue measure in the $(\mathbf{w}, \boldsymbol{\phi})$-dimension. We make no further assumptions about the prior, except that it is proper and exchangeable for each k, that is, invariant under permutations of the pairs (w_i, ϕ_i). We do not impose any ordering of the ϕ_i, motivated by identifiability reasons, as found in Richardson and Green (1997). We also denote by $L(k, \mathbf{w}, \boldsymbol{\phi})$ the likelihood

$$L(k,\mathbf{w},\boldsymbol{\phi}) = \prod_{i=1}^{m} p(y_i|k,\mathbf{w},\boldsymbol{\phi}),$$

where $\mathbf{y} = (y_1, \ldots, y_m)$ is the observed data. The posterior density, which is our starting point for inference, is thus proportional to $r(k, \mathbf{w}, \boldsymbol{\phi})L(k, \mathbf{w}, \boldsymbol{\phi})$. More realistic models typically involve hyperparameters, which add no further difficulty. Below we set $\boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\phi})$, with k being implicit in this notation, and $\Theta^{(k)}$ denotes the space of k component parameters.

A feature inherent to mixture models is that we may associate with each observation y_i a label or *allocation* $z_i \in \{1, \dots, k\}$ with $P(z_i = j \mid k, \mathbf{w}) = w_j$, that indicates from which component y_i was drawn. Given the data, these labels can be sampled independently according to

$$P(z_i = j \mid k, \mathbf{w}, \boldsymbol{\phi}, y_i) = w_j f(y_i|\phi_j) \Big/ \sum_{\ell=1}^{k} w_\ell f(y_\ell|\phi_\ell). \tag{1}$$

This simulation is called *completing the sample* as, following EM terminology, $(\mathbf{z}, \mathbf{y})$ is referred to as the *complete data*. As detailed in Section 3 and as demonstrated in Celeux *et al.* (2000) for mixtures, completion is not necessary from a simulation point of view. Richardson and Green (1997) devised an algorithm that carries along the complete data through all moves of the sampler. In contrast, the algorithm of Stephens (2000) works with incomplete data, that is, $\mathbf{y}$ alone, in the dimension-changing moves, but completes the data at regular intervals to carry out a resampling of all parameters and hyperparameters but k.

2.2. Birth-and-death MCMC

In Stephens' (2000) form of BDMCMC, new components are created (*born*) in continuous time at a rate $\beta(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ refers to the current state of the sampler. Whenever a new component is born, its weight w and parameter ϕ are drawn from a joint density $h(\boldsymbol{\theta}; (w, \phi))$. In order to include the new component, the old component weights are scaled down proportionally to make all of the weights, including the new one, sum to unity; that is, $w_i := w_i/(1 + w)$. The new component weight-parameter pair (w, ϕ) is then added to $\boldsymbol{\theta}$. We denote these operations by '$\cup$', so that the new state is $\boldsymbol{\theta} \cup (w, \phi)$. Conversely, in a $(k+1)$ component configuration $\boldsymbol{\theta} \cup (w, \phi)$, the component (w, ϕ) is killed at rate

$$\delta(\boldsymbol{\theta}; (w, \phi)) = \frac{L(\boldsymbol{\theta})r(\boldsymbol{\theta})}{L(\boldsymbol{\theta} \cup (w, \phi))r(\boldsymbol{\theta} \cup (w, \phi))} \times \frac{1}{k+1} \times \frac{\beta(\boldsymbol{\theta})h(\boldsymbol{\theta}; (w, \phi))}{(1 - w)^{k-1}}. \tag{2}$$

The factor $(1 - w)^{k-1}$ in (2) results from a change of variable when renormalising the weights. Indeed, when the component (w, ϕ) is removed, the remaining component weights are renormalised to sum to unity. We denote these operations by '$\setminus$', so that $\boldsymbol{\theta} = (\boldsymbol{\theta} \cup (w, \phi)) \setminus (w, \phi)$.

An important feature of BDMCMC is that a continuous time jump process is associated with the birth-and-death rates: whenever a jump occurs, the corresponding move is always accepted. The acceptance probability of usual MCMC methods is replaced by the differential holding times. In particular, implausible configurations, i.e. configurations such that $L(\boldsymbol{\theta})r(\boldsymbol{\theta})$ is small, die quickly.

2.3. The Markov jump process view and local balance

The birth-and-death process described in the previous subsection is a Markov jump process: whenever it reaches state $\boldsymbol{\theta}$, it stays there for an exponentially distributed time with expectation depending on $\boldsymbol{\theta}$, and, after expiration of this holding time, jumps to a new state according to a Markov transition kernel. To ensure that a Markov jump process has an invariant density proportional to $L(\boldsymbol{\theta})r(\boldsymbol{\theta})$, it is sufficient, although not necessary, that the local balance equations

$$L(\boldsymbol{\theta})r(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}') = L(\boldsymbol{\theta}')r(\boldsymbol{\theta}')q(\boldsymbol{\theta}', \boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' , \tag{3}$$

are satisfied (Preston, 1976; Ripley, 1977; Geyer and Møller, 1994). Here $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the rate of moving from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$. Special care is required with such considerations, however, since the transition kernel of the jump chain typically does not have a density with respect to a single dominating measure. For example, after killing a component the new state is completely known given the current one. This problem also occurs for RJMCMC samplers, as exemplified by the measure construction in Green (1995), and we do not detail it further here. Further reading on Markov jump processes is found in, for example, Preston (1976), Ripley (1977, Sections 2 and 4), and Breiman (1992, Chap. 15, Sections 5 and 6).

Let us now derive (2) from (3). In the particular case of birth-and-death moves and a k component configuration $\boldsymbol{\theta}$, (3) takes the form

$$L(\boldsymbol{\theta})r(\boldsymbol{\theta})\times\beta(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))\Big/(k+1)!\times(1-w)^{k-1} = L(\boldsymbol{\theta}\cup(w,\phi))r(\boldsymbol{\theta}\cup(w,\phi))\times\delta(\boldsymbol{\theta};(w,\phi))\Big/k!\,, \quad (4)$$

which indeed leads to (2). The justification for the various factors in (4) is as follows: the factorials $k!$ and $(k+1)!$ arise from the exchangeability assumption on the mixture components. Given that we do not impose an ordering constraint on $\phi_1,\ldots,\phi_k$, there are $k!$ and $(k+1)!$ equivalent ways of writing $\boldsymbol{\theta}$ and $\boldsymbol{\theta}\cup(w,\phi)$, respectively. The equivalence is to be understood as giving the same likelihood, prior and posterior densities. The $1/(k+1)!$ and $1/k!$ terms thus appear as the probabilities of selecting *one* of the $(k+1)!$ and $k!$ possible ways of writing $\boldsymbol{\theta}\cup(w,\phi)$ and $\boldsymbol{\theta}$ in the birth and death moves. This selection is immaterial, since it has no relevance for the posterior distribution. Furthermore, $b(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))$ is the density of proposing a new component (w,ϕ), and $(1-w)^{k-1}$ is again a Jacobian arising from renormalisation of the weights. This determinant should be associated with the density h, as the $(k+1)$ component parameter $\boldsymbol{\theta}\cup(w,\phi)$ is not drawn directly from a density on $\Theta^{(k+1)}$, but rather indirectly through first drawing (w,ϕ) and then renormalising. In order to compute the resulting density on $\Theta^{(k+1)}$ one must then calculate a Jacobian. Thus, $q(\boldsymbol{\theta},\boldsymbol{\theta}\cup(w,\phi))=\beta(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))/(1-w)^{k-1}$.

2.4. Generalisations of birth-and-death MCMC

Stephens (2000) resampled component weights and parameters with fixed k, as well as hyperparameters, at *deterministic* times (as opposed to the *random* occurrences of the birth-and-death moves). This makes the overall process inhomogeneous in time. We can incorporate similar moves into the continuous time sampler by adding a continuous time process in which, in state $\boldsymbol{\theta}$, such moves occur at rate $\gamma(\boldsymbol{\theta})$. Birth-and-death rates stay the same. The rates for resampling the component weights, parameters, and hyperparameters, could also be different.

A further generalisation is to introduce other moves, like the split-and-combine moves of Richardson and Green (1997). We consider here the special case where, as in Green (1995), the combine move is deterministic. For simplicity θ denotes an element of the k component parameter $\boldsymbol{\theta}$. Thus, in a mixture context, typically $\theta=(w,\phi)$.

As for the RJMCMC proposal, the *split* move for a given component θ of the k component vector $\boldsymbol{\theta}$ is to split this component as to give rise to a new parameter vector with $k+1$ components, defined as $((\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon))$ where T is a differentiable one-to-one mapping that outputs two new components and ε is a random variable with density function p. We also assume that the mapping is symmetric in the sense that

$$P\{T(\theta,\varepsilon)\in B'\times B''\}=P\{T(\theta,\varepsilon)\in B''\times B'\} \qquad \text{for all } B',\,B''. \quad (5)$$

We denote the total rate of splitting by $\eta(\boldsymbol{\theta})$ and assume that, in a split move, each component is chosen with equal probability $1/k$. Conversely, the local balance equation (3) provides for any of the $k(k-1)/2$ pairs of components of $\boldsymbol{\theta}$, the rate of *combining* them. In this particular case,

$$\begin{aligned} &L(\boldsymbol{\theta})r(\boldsymbol{\theta})\times\frac{\eta(\boldsymbol{\theta})}{k}\times 2\,p(\varepsilon)\times\left|\frac{\partial T(\theta,\varepsilon)}{\partial(\theta,\varepsilon)}\right|\Big/(k+1)! \\ &\quad= L\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\}\,r\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\}\times q\left[\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\},\boldsymbol{\theta}\right]\Big/k! \end{aligned}$$

As above, the factorials arise as probabilities of selecting particular representations of $\boldsymbol{\theta}$ and $(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)$, and $\eta(\boldsymbol{\theta})/k$ is the rate of splitting a *particular* component as $\eta(\boldsymbol{\theta})$ is the overall split rate. The coefficient 2 is due to the fact that a component can be split in two pairs that are identical apart from the ordering, and that occur with the same probability because of the symmetry assumption (5); otherwise we would have to replace $p(\varepsilon)$ with the average of two terms. Thus, the rate of combining two components, $q\left[\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\},\boldsymbol{\theta}\right]$, is

$$\frac{L(\boldsymbol{\theta})r(\boldsymbol{\theta})}{L\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\}\,r\{(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\}}\times\frac{\eta(\boldsymbol{\theta})}{(k+1)k}\times 2p(\varepsilon)\left|\frac{\partial T(\theta,\varepsilon)}{\partial(\theta,\varepsilon)}\right|. \quad (6)$$

In Section 4.3, we will directly derive this rate from Richardson and Green's (1997) sampler.

2.5. *Sampling in continuous time: a new Rao–Blackwellisation*

For a discrete time RJMCMC sampler, its output is typically monitored after each step, or on regular intervals in order to decrease inter-sample correlation as in Ripley (1977, Section 5) and Richardson and Green (1997).

In continuous time, there are more options. For example, the process may be sampled at regular times, as in Stephens (2000), or at instants given by an independent Poisson process. In either case posterior means $E[g(\boldsymbol{\theta}) \mid \mathbf{y}]$ are estimated by sample means $N^{-1} \sum_1^N g(\boldsymbol{\theta}(\tau_i))$, where $\{\theta(t)\}$ is the CTMCMC process and the τ_i's are the sampling instants. Under the former sampling scheme, if the sampling interval tends to zero, we effectively put a weight on each state visited by $\{\boldsymbol{\theta}(t)\}$ that is equal to the length of the holding time in that state, when computing the sample mean. Before elaborating further on this idea, we introduce some additional notation.

Let T_n be the time of the n-th jump of $\{\boldsymbol{\theta}(t)\}$ with $T_0 = 0$. By the *jump chain* we mean the Markov chain $\{\boldsymbol{\theta}(T_n)\}$ of states visited by $\{\boldsymbol{\theta}(t)\}$. We denote this chain by $\{\widetilde{\boldsymbol{\theta}}_n\}$, that is, $\widetilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(T_n)$. Let $\lambda(\boldsymbol{\theta})$ be the total rate of $\{\boldsymbol{\theta}(t)\}$ leaving state $\boldsymbol{\theta}$, that is, the sum of the birth and all death rates, plus the rates of all other kinds of moves there may be. Then the holding time $T_n - T_{n-1}$ of $\{\boldsymbol{\theta}(t)\}$ in its n-th state $\widetilde{\boldsymbol{\theta}}_{n-1}$ has a conditional exponential $\mathcal{E}xp(\lambda(\widetilde{\boldsymbol{\theta}}_{n-1}))$ distribution.

Returning to the sampling scheme, we can reduce sampling variability by replacing the weight $T_n - T_{n-1}$ by its expectation $1/\lambda(\widetilde{\boldsymbol{\theta}}_{n-1})$. In this way the variances of estimators built from the sampler output are decreased: both numerator and denominator have reduced variance by virtue of the Rao–Blackwell theorem, since

$$\sum_{n=1}^{N} \frac{g(\widetilde{\boldsymbol{\theta}}_{n-1})}{\lambda(\widetilde{\boldsymbol{\theta}}_{n-1})} = \sum_{n=1}^{N} E[T_n - T_{n-1} \mid \widetilde{\boldsymbol{\theta}}_{n-1}]\, g(\widetilde{\boldsymbol{\theta}}_{n-1})$$

and likewise for the denominator. The asymptotic variance of the ratio

$$\sum_{n=1}^{N} g(\widetilde{\boldsymbol{\theta}}_{n-1})/\lambda(\widetilde{\boldsymbol{\theta}}_{n-1}) \Big/ \sum_{n=1}^{N} 1/\lambda(\widetilde{\boldsymbol{\theta}}_{n-1})$$

can then be shown to be smaller than when using $T_n - T_{n-1}$ in place of $1/\lambda(\widetilde{\boldsymbol{\theta}}_{n-1})$, following Geweke (1989).

When sampling $\{\boldsymbol{\theta}(t)\}$ this way, we only simulate the jump chain and store each state it visits and the corresponding expected holding time. Alternatively, the expected holding times may be recomputed when post-processing the sampler output. The transition kernel of the jump chain is as follows: the probability of an event happening is proportional to its rate. For example, the probability of a birth is $\beta(\boldsymbol{\theta})/\lambda(\boldsymbol{\theta})$, and if a birth occurs the new component weight and parameter are drawn from $h(\boldsymbol{\theta}; (w, \phi))$ as before. Thus we need to compute all rates when simulating the jump chain, just as we do when simulating $\{\boldsymbol{\theta}(t)\}$. All *possible* moves are incorporated into the Rao–Blackwellised estimator, not only those that are *selected.*

This reformulation of the continuous time algorithm has more than practical appeal for the approximation of integrals. Indeed it highlights a point that will be made clearer in Section 4, namely that the continuous time structure is paramount neither for the MCMC algorithm nor for the approximation of integrals.

3. An illustration for hidden Markov models

Before moving to the comparison with reversible MCMC, we illustrate the potential of our continuous time extension in the set-up of hidden Markov models (Robert *et al.*, 2000).

3.1. *Setting*

In this generalisation of the mixture model, the observations y_n are such that, conditional on a hidden Markov chain $\{z_n\}$ with finite state space $\{1, \ldots, k\}$, y_n is distributed as a $\mathcal{N}(0, \sigma^2_{z_n})$ variate. Therefore, marginally, y_n is distributed from a mixture of normal distributions.

Unlike previous implementations, we choose to parameterise the transition probability matrix of the Markov chain $\{z_n\}$ by $\mathbf{P} = (\omega_{ij})$ as follows:

$$P(z_{n+1} = j \mid z_n = i) = \omega_{ij} \Big/ \sum_{\ell=1}^{k} \omega_{i\ell}.$$

The ω_{ij}'s are therefore not identified, but this parameterisation should facilitate the MCMC moves, provided a vague proper prior is selected, since it relaxes the constraints on those moves. Further, this reparameterisation allows for a point process representation of the problem (Preston, 1976; Ripley, 1977; Geyer and Møller, 1994). The prior model consists in a uniform prior $\mathcal{U}\{1,\ldots,M\}$ on k, an $\mathcal{E}xp(1)$ prior on the ω_{ij}'s, a uniform $\mathcal{U}(0,\alpha)$ prior on the σ_i's and a data dependent $\mathcal{E}xp(5\max|y_n|)$ prior on the hyperparameter $1/\alpha$; Robert *et al.* (2000) noticed that the factor 5 in the exponential distribution was of little consequence. We stress that we impose no identifiability constraints by ordering the variances, in contrast to Robert *et al.* (2000). Another major difference is that, as in Stephens (2000), we do not use *completion* to run our algorithm. That is to say, the latent Markov chain $\{z_n\}$ is not simulated by the algorithm. This can be avoided due to both the forward recursive representation of the likelihood for a hidden Markov model (Baum *et al.*, 1970), already used in Robert *et al.* (1999), and the random walk proposals as in Hurn *et al.* (2003). While not strictly necessary from an algorithmic point of view (Robert *et al.*, 1999), this choice facilitates the comparison with Stephens (2000).

3.2. The moves of the continuous time MCMC algorithm

Since Robert *et al.* (2000) implemented reversible jump for this model, we focus on the CTMCMC counterpart, extending Stephens (2000) to this framework. In addition to birth-and-death moves, which were enough to provide good mixing in Stephens (2000), we are forced to introduce additional proposals, similar to those in Richardson and Green (1997), because we observed that the birth-and-death moves are not, by themselves, sufficient to ensure fast convergence of the MCMC algorithm. The proposals we add are split-and-combine moves, as described earlier, and fixed-k moves, where the parameters are modified via a regular Metropolis–Hastings step. The latter proposals are essential in ensuring irreducibility and good convergence properties.

The birth-and-death and fixed-k moves are simple to implement, and are equivalent to those given in Hurn *et al.* (2003) with fixed-k moves relying on random walk proposals over the transforms $\log\omega_i$ and $\log(\sigma_i/(\alpha-\sigma_i))$.

The split-and-combine move follows the general framework of Section 2.4 with a combine rate given by (6). We used η^S as an individual split rate which is the same for all components. This means that the overall rate of a split move for a k component vector is $\eta(\boldsymbol{\theta}) = k\eta^S$. In the practical implementation of the algorithm, we chose $\eta^S = \eta^B = 2$ and $\eta^F = 5$, where η^B and η^F correspond to the birth and fixed-k move rates, respectively.

In the case of the above normal hidden Markov model, a split of state i_0 into states i_1 and i_2 involves four different types of actions.

(a) A split move in row $j \neq i_0$ for ω_{j,i_0} as

$$\omega_{j,i_1} = \omega_{j,i_0}\varepsilon_j, \quad \omega_{j,i_2} = \omega_{j,i_0}(1-\varepsilon_j), \qquad \text{where} \ \ \varepsilon_j \sim \mathcal{U}(0,1)\,.$$

This proposal is sensible when thinking that both new states i_1 and i_2 are issued from the state i_0 and the probabilities to reach i_0 are thus distributed between the probabilities to reach the new states i_1 and i_2, respectively.

(b) A split move in column $i \neq i_0$ for $\omega_{i_0,i}$ as

$$\omega_{i_1,i} = \omega_{i_0,i}\xi_i, \quad \omega_{i_2,i} = \omega_{i_0,i}/\xi_i, \qquad \text{where} \ \ \xi_i \sim \mathcal{LN}(0,1)\,.$$

The symmetry constraint (5) is thus satisfied, that is, ξ_i and $1/\xi_i$ have the same lognormal distribution. Before this, we tried a half-Cauchy $\mathcal{C}^+(0,1)$ proposal, which also preserves the distribution under inversion, but this led to very poor mixing properties of the algorithm.

(c) A split move for ω_{i_0,i_0} as

$$\omega_{i_1,i_1} = \omega_{i_0,i_0}\varepsilon_{i_0}\xi_{i_1}, \qquad \omega_{i_1,i_2} = \omega_{i_0,i_0}(1-\varepsilon_{i_0})\xi_{i_2},$$
$$\omega_{i_2,i_1} = \omega_{i_0,i_0}\varepsilon_{i_0}/\xi_{i_1}, \qquad \omega_{i_2,i_2} = \omega_{i_0,i_0}(1-\varepsilon_{i_0})/\xi_{i_2},$$

where ε_{i_0} is uniform on $(0,1)$ and ξ_{i_1}, ξ_{i_2} are $\mathcal{LN}(0,1)$.

(d) A split move for $\sigma^2_{i_0}$ as

$$\sigma^2_{i_1} = \sigma^2_{i_0}\varepsilon_\sigma, \quad \sigma^2_{i_2} = \sigma^2_{i_0}/\varepsilon_\sigma, \qquad \text{where } \varepsilon_\sigma \sim \mathcal{LN}(0,1).$$

The combine move is chosen in a symmetric way, so that states i_1 and i_2 are combined into state i_0 by taking first the geometric average of rows i_1 and i_2 in the unnormalised transition probability matrix and then adding columns i_1 and i_2. One can check that this sequence of moves also applies to the particular case of ω_{i_0,i_0}. The variance $\sigma^2_{i_0}$ is the geometric average of $\sigma^2_{i_1}$ and $\sigma^2_{i_2}$. Appendix B details the computation of the corresponding Jacobian.

3.3. An illustration

For a comparison with Robert *et al.* (2000), we consider a single dataset studied in that paper, namely the wind intensity in Athens (Francq and Roussignol, 1997). Since the prior distribution on the σ's is a uniform $\mathcal{U}(0,\alpha)$, α is an hyperparameter that is estimated from the dataset in a hierarchical way and updated through a slice sampler (see Robert *et al.* (2000) for details) via an additional process with intensity η^α, set equal to 1. The variances σ_i^2, being constrained to be smaller than α^2, are updated via a Gaussian random walk proposal in the α-logit domain, that is using the transform $\log(\sigma/(\alpha-\sigma))$ and its inverse.

Fig. 1 summarises the output for this dataset. As in Robert *et al.* (2000), we obtain a mode of the posterior distribution of k at $k=3$, although the posterior distribution slightly differs in our case since the posterior probabilities for $1,2,3,4$ are $0.0064, 0.1848, 0.7584, 0.0488$, to be compared with Table 1 of Robert *et al.* (2000). Fig. 1 also provides the distribution of the number of moves per time unit (on the continuous time axis). The loglikelihoods cover a wider range than those found in Robert *et al.* (2000), although the highest values are the same. For instance, the largest likelihood for $k=2$ is -688, while it is -675 for $k=3$ and -670 for $k=4$. That we find lower loglikelihoods than with RJMCMC techniques is to be expected since, although both RJMCMC and CTMCMC explore the same target distribution, continuous time algorithms can explore more unlikely regions in the parameter space, like the tails of the target, by down-weighting states with shorter lifetimes.

4. Comparisons of reversible jump MCMC with continuous time algorithms

In this section we provide a comparison of reversible jump and continuous time methodologies, starting with a review of reversible jump MCMC within the framework of mixtures.

4.1. Reversible jump MCMC

In a k component state $\boldsymbol{\theta}$, at each iteration, the simplest version of the reversible jump algorithm proposes with probability $b(\boldsymbol{\theta})$ to create a new component and with probability $d(\boldsymbol{\theta})$ to kill one. Obviously, $b(\boldsymbol{\theta}) + d(\boldsymbol{\theta}) = 1$, if we do not account for fixed-k moves at this level. If an attempt to create a new component is made, its weight and parameter are drawn from $h(\boldsymbol{\theta};(w,\phi))$ as above. If an attempt to kill a component is made then, for instance, in a mixture model, each component is selected with equal probability. A new component is accepted with probability $\min(1,A)$, where

$$\begin{aligned} A = A(\boldsymbol{\theta};\boldsymbol{\theta}\cup(w,\phi)) &= \frac{L(\boldsymbol{\theta}\cup(w,\phi))r(\boldsymbol{\theta}\cup(w,\phi))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})} \times \frac{(k+1)!}{k!} \times \frac{d(\boldsymbol{\theta}\cup(w,\phi))}{(k+1)b(\boldsymbol{\theta})} \times \frac{(1-w)^{k-1}}{h(\boldsymbol{\theta};(w,\phi))} \\ &= \frac{L(\boldsymbol{\theta}\cup(w,\phi))r(\boldsymbol{\theta}\cup(w,\phi))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})} \times \frac{d(\boldsymbol{\theta}\cup(w,\phi))}{b(\boldsymbol{\theta})} \times \frac{(1-w)^{k-1}}{h(\boldsymbol{\theta};(w,\phi))}. \end{aligned} \tag{7}$$

Here the first ratio is the ratio of posterior densities, $b(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))$ is the density corresponding to proposing a new component (w,ϕ), and $d(\boldsymbol{\theta}\cup(w,\phi))/(k+1)$ is the probability of proposing to

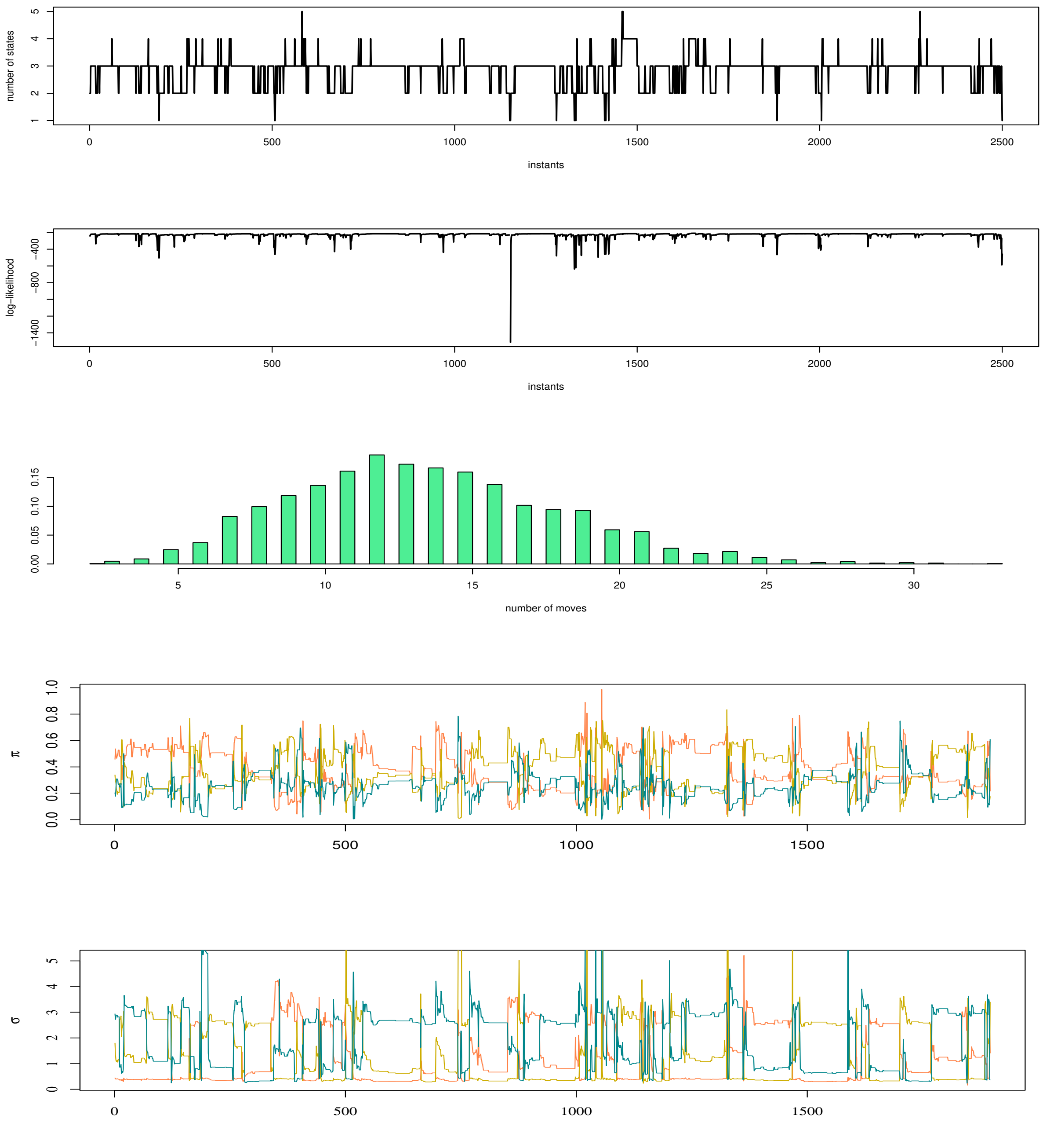

Fig. 1. Continuous time MCMC algorithm output for a sequence of 500 wind intensities in Athens; *(from top to bottom)* (a) plot of an equal time sample of k's; (b) plot of the corresponding loglikelihood values; (c) histogram of the number of moves per time unit; (d) MCMC sequence of the probabilities π_j of the stationary distribution of the three components when conditioning on $k = 3$; (e) same graph for the σ_j's.

kill component (w, ϕ) when in state $\boldsymbol{\theta} \cup (w, \phi)$. Finally $(1-w)^{k-1}$ is the same Jacobian determinant as above, and the factorial ratio arises from the exchangeability assumption. If a proposal to kill a component (w, ϕ) of a $(k+1)$ component state $\boldsymbol{\theta} \cup (w, \phi)$ is made, the acceptance probability is $\min(1, 1/A)$, where $A = A(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi))$ is as above.

RJMCMC typically involves other kinds of moves like fixed-k moves resampling the component weights, parameters ϕ_i and, possibly, hyperparameters—see, e.g., Richardson and Green (1997). A complete *sweep* of the algorithm consists in the composition of a birth-and-death move with these other fixed-k moves. Sampling for a fixed k can be carried out using a Gibbs move after completing the sample according to (1). As noted above, Richardson and Green (1997) designed additional moves for splitting and combining components.

4.2. Convergence to BDMCMC

In this section we construct a sequence of RJMCMC samplers converging to the BDMCMC sampler.

Before proceeding we introduce some additional notation. Let $S^{k-1} = \{(w_1, \ldots, w_k) : w_i > 0, \sum_i w_i = 1\}$ and let Φ denote the space in which each ϕ_i lies. Hence $\Theta^{(k)}$, the space of k-dimensional parameters, is $\Theta^{(k)} = S^{k-1} \times \Phi^k$. Finally let $\Theta = \cup_{k \geq 1} \Theta^{(k)}$ denote the overall parameter space.

For $N \in \mathbb{N}$ we define an RJMCMC sampler by defining birth and death probabilities

$$b_N(\boldsymbol{\theta}) = 1 - \exp\{-\beta(\boldsymbol{\theta})/N\}, \quad d_N(\boldsymbol{\theta}) = 1 - b_N(\boldsymbol{\theta}) = \exp\{-\beta(\boldsymbol{\theta})/N\},$$

where $\beta(\boldsymbol{\theta})$ is the birth rate of the BDMCMC sampler. Then A also depends on N, and we write $A = A_N$. We remark that as $N \to \infty$, $b_N(\boldsymbol{\theta}) \sim \beta(\boldsymbol{\theta})/N$, and if $\beta(\boldsymbol{\theta})$ is bounded we can take instead $b_N(\boldsymbol{\theta}) = \beta(\boldsymbol{\theta})/N$. The state at time $n = 0, 1, \ldots$ of the N-th RJMCMC sampler is denoted by $\boldsymbol{\theta}_n^N$, and for each N we construct a continuous time process $\{\boldsymbol{\theta}^N(t)\}_{t \geq 0}$ as $\boldsymbol{\theta}^N(t) = \boldsymbol{\theta}_{\lfloor Nt \rfloor}^N$, where $\lfloor \cdot \rfloor$ denotes the integer part. The state of the BDMCMC sampler at time $t \geq 0$ is denoted by $\boldsymbol{\theta}(t)$.

We now consider what happens as $N \to \infty$. The probability of proposing a birth in state $\boldsymbol{\theta}$ tends to zero as $\beta(\boldsymbol{\theta})/N$. Hence the acceptance ratio A_N tends to infinity, so that a birth proposal is always accepted. If time is speeded up at scale N, on the nominal time scale the limiting process of accepted births in state $\boldsymbol{\theta}$ is a Poisson process of rate $\beta(\boldsymbol{\theta})$. Furthermore, the scaled probability of deleting component (w, ϕ) in a state $\boldsymbol{\theta} \cup (w, \phi) \in \Theta^{(k+1)}$ is

$$\begin{aligned} &N d_N(\boldsymbol{\theta}) \times \min[1, 1/A_N\{\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)\}]/k + 1 \\ &\quad \longrightarrow \frac{L(\boldsymbol{\theta}) r(\boldsymbol{\theta})}{L(\boldsymbol{\theta} \cup (w, \phi)) r(\boldsymbol{\theta} \cup (w, \phi))} \times \frac{1}{k+1} \times \beta(\boldsymbol{\theta}) \times \frac{h(\boldsymbol{\theta}; (w, \phi))}{(1-w)^{k-1}} \quad \text{as } N \to \infty, \end{aligned}$$

and the right hand side is just $\delta(\boldsymbol{\theta}; (w, \phi))$, given in (2). Considering the rescaled time axis and the independent attempts to create or delete components, in the limit the waiting time until this component is killed has an exponential distribution with rate $\delta(\boldsymbol{\theta}; (w, \phi))$, agreeing with the BDMCMC sampler. Thus, as $N \to \infty$ a birth is rarely proposed but always accepted and a death is almost always proposed but rarely accepted. Both these schemes result in waiting times which are asymptotically exponentially distributed with rates in accordance with the BDMCMC sampler. Thus, one may expect that as $N \to \infty$, the processes $\{\boldsymbol{\theta}^N(t)\}$ and $\{\boldsymbol{\theta}(t)\}$ will become more and more similar.

We will now make this reasoning strict, starting with the following assumptions:

(A0) Φ has a separable topology which can be metrised by a complete metric.

(A1) $\beta(\boldsymbol{\theta})$ is positive and continuous on Θ.

(A2) $r(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ are positive and continuous on Θ.

(A3) For each $(w, \phi) \in (0, 1) \times \Phi$, $h(\cdot; (w, \phi))$ is continuous on Θ and for each $\boldsymbol{\theta} \in \Theta$ there is a neighbourhood G of $\boldsymbol{\theta}$ such that $\sup_{\boldsymbol{\theta}' \in G} h(\boldsymbol{\theta}'; \cdot)$ is integrable.

We first note that, since the standard topology on the open unit interval $(0, 1)$ is separable and can be metrised by a complete metric, for example $d(x, y) = |\log(x/(1-x)) - \log(y/(1-y))|$, S^{k-1}

can be viewed as a complete separable metric space. Then Θ, with the induced natural topology, is a space of the same kind. The process $\{\boldsymbol{\theta}(t)\}$ is a Markov process on Θ which we assume has sample paths in $D_\Theta[0,\infty)$, the space of Θ-valued functions on $[0,\infty)$ which are right-continuous and have left hand limits everywhere.

We then derive the following result (see Appendix A for a proof).

THEOREM 1. *Under assumptions (A0)–(A3) and assuming that $\boldsymbol{\theta}(0)$ and $\boldsymbol{\theta}_0$ are drawn from the same initial distribution, $\{\boldsymbol{\theta}^N(t)\}_{t\geq 0}$ converges weakly to $\{\boldsymbol{\theta}(t)\}_{t\geq 0}$ in the Skorohod topology on $D_\Theta[0,\infty)$ as $N\to\infty$.*

4.3. Convergence to other continuous time processes

Recall again that, in Richardson and Green's (1997) version, the RJMCMC sampler also includes a split-and-combine move. More precisely, using the same notation as in Section 2.4, they propose to split a randomly chosen component of the k component vector $\boldsymbol{\theta}$ with probability $s_N(\boldsymbol{\theta})$ so as to give rise to a new parameter vector with $k+1$ components, defined as $(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)$. Conversely, the probability of proposing to combine a randomly chosen pair of components of $\boldsymbol{\theta}$ (there are $k(k-1)/2$ pairs) is denoted by $c_N(\boldsymbol{\theta}) = 1 - s_N(\boldsymbol{\theta})$.

A split move changing the k component vector $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)$ has acceptance probability

$$\min\left\{1, \frac{L\left((\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\right) r\left((\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)\right)}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})} \times \frac{(k+1)!}{k!} \right.$$
$$\left. \times \frac{c_N((\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon))k}{s_N(\boldsymbol{\theta})k(k+1)/2} \times \frac{1}{2p(\varepsilon)}\left|\frac{\partial T(\theta,\varepsilon)}{\partial(\theta,\varepsilon)}\right|^{-1}\right\}.$$

If, as above, we let $s_N(\boldsymbol{\theta}) = 1-\exp\{-\eta(\boldsymbol{\theta})/N\}$ for some $\eta(\boldsymbol{\theta})$, so that $Ns_N(\boldsymbol{\theta})\to\eta(\boldsymbol{\theta})$, and accordingly scale by N the trajectory of the corresponding discrete time sampler, the limiting continuous time process has a rate of moving from $(\boldsymbol{\theta}\setminus\theta)\cup T(\theta,\varepsilon)$ to $\boldsymbol{\theta}$ by a combine move which is given by (6). Convergence of reversible jump MCMC to continuous time processes thus occurs in a broader context than within the birth-and-death framework of Stephens (2000).

4.4. A numerical comparison of both methodologies

While Theorem 1 establishes a strong connection between RJMCMC and CTMCMC, by showing that CTMCMC can be arbitrarily well approximated by an RJMCMC algorithm, it does not imply that in practice both approaches perform equivalently, for instance in terms of computational cost. We thus carried a numerical comparison of both approaches based on identical moves and identical proposals on both sides. Further implementation details are provided in Appendix C. In this comparison, we chose to remain within the framework of mixtures of distributions, partly because the setting is simpler than hidden Markov models and partly because most of the earlier literature on the topic relates to this area: for instance, we rely on the *Galaxy dataset*, heavily used in the literature since Roeder (1990).

4.4.1. Implementation issues

We first discuss computational aspects of both discrete and continuous time algorithms. In continuous time settings, once a state $\boldsymbol{\theta}$ is visited, it is necessary to compute the rates of all possible moves leading to an exit from that state, that is $\mathrm{O}(k)$ and $\mathrm{O}(k^2)$ computations for birth-and-death and split-and-combine moves, respectively. Discrete time settings do not require this exhaustive checking, as the acceptance ratio of a move is not computed until the move is proposed. This advantage of reversible jump MCMC is however mitigated by three facts.

(i) For continuous time moves such as birth and split, rates are typically very simple (e.g, constant) and it is only the death or combine rates that are expensive to compute.

(ii) Except for small datasets, the cost of evaluating the acceptance probability in RJMCMC mainly lies in computing the loglikelihood at the proposed parameters according to

$$\log L(k, \mathbf{w}, \phi) = \sum_{i=1}^{m} \log \sum_{j=1}^{k} w_j f(y_i|\phi_j)\,, \tag{8}$$

which involves $\mathrm{O}(k \times m)$ computations. For mixture models, the computation associated with RJMCMC thus also increases proportionally to k.

(iii) At the expense of storing all values $f(y_i|\phi_j)$ as in Stephens (2000), it is possible to reduce significantly the cost of repeated evaluations of (8). For instance, in a death proposal the only required new computations are the summations in i and j, omitting the index of the selected component. Although this remark also applies to the RJMCMC sampler, it is most profitable when applied to the implementation of the continuous time sampler.

Thus, when only birth-and-death moves are used, the average computation times for simulating one jump of the continuous time sampler and one step of the reversible jump sampler are comparable. In our implementation, the former is longer by a factor which varies between 1.5 and 2, depending on the dataset. On the other hand, the computation time for continuous time simulation with split-and-combine moves is a factor 3 longer for the Galaxy dataset.

4.4.2. Birth-and-death samplers

We first contrast the performance of the two types of samplers, RJMCMC and CTMCMC, when only birth-and-death moves are used in addition to moves that do not modify the number of components. Except for the fine details of the proposals described in Appendix C and the absence of completion in the fixed-k moves, we are thus in the setting considered by Stephens (2000). Note however that for CTMCMC, we adopted the Rao–Blackwellisation device discussed in Section 2.5 (weighting each visited configuration by the inverse of the overall rate of leaving rather than by the corresponding exponentially distributed holding time). We proposed the fixed-k moves according to an independent Poisson process of rate η^F, which leaves the overall continuous time process Markovian, whereas Stephens (2000) proposed these moves at fixed regular times. By setting the probability P^F of proposing a fixed-k move in RJMCMC equal to the rate $\eta^F = 0.5$ at which fixed-k moves are proposed in CTMCMC, and likewise $P^B = \eta^B = 0.25$ for the birth moves, we guaranteed that the moves were proposed in equal proportions by both samplers. The most important aspect is that both the reversible jump and the continuous time sampler were implemented using exactly the same move proposals to the point of sharing the same routines, which allows for meaningful comparisons. In the following, we compare the performance of both samplers when the number of jumps (number of visited configurations) in CTMCMC is equal to the number of iterations of RJMCMC.

The main message here is conveyed by Fig. 2 which shows that there is no significant difference between the samplers: be it for a small (5,000) or a large (500,000) number of iterations, the accuracy of the estimated posterior probabilities for all allowed values of k is equivalent for both samplers. Other signals like posterior parameter estimates conditional on a fixed k tend to show even less difference; this is not surprising granted that both samplers share the same fixed-k moves.

Another evaluation of the performance of MCMC samplers is provided by the autocovariance function of simulated traces. To implement this idea for the continuous time sampler, the Rao-Blackwellised continuous time path—that is, the path of the continuous time process where the inverse rates are substituted for the corresponding holding times—was sampled regularly, with a number of points equal to the number of jumps. Fig. 3 shows the resulting autocovariance for the posterior simulations of k for both RJMCMC and CTMCMC, estimated on 2 million iterations after discarding a burn-in period of 8 million iterations. Once again, both samplers are seen to perform equivalently: although all moves are accepted in CTMCMC, the mixing is not significantly improved over RJMCMC because of the weighting mechanism. This is well captured by Fig. 4 which shows that only about 30% of the configurations visited by the continuous time sampler are maximally weighted. Conversely, 15% of the configurations have a negligible weight, a situation which occurs when there is at least one death move which has a very large rate.

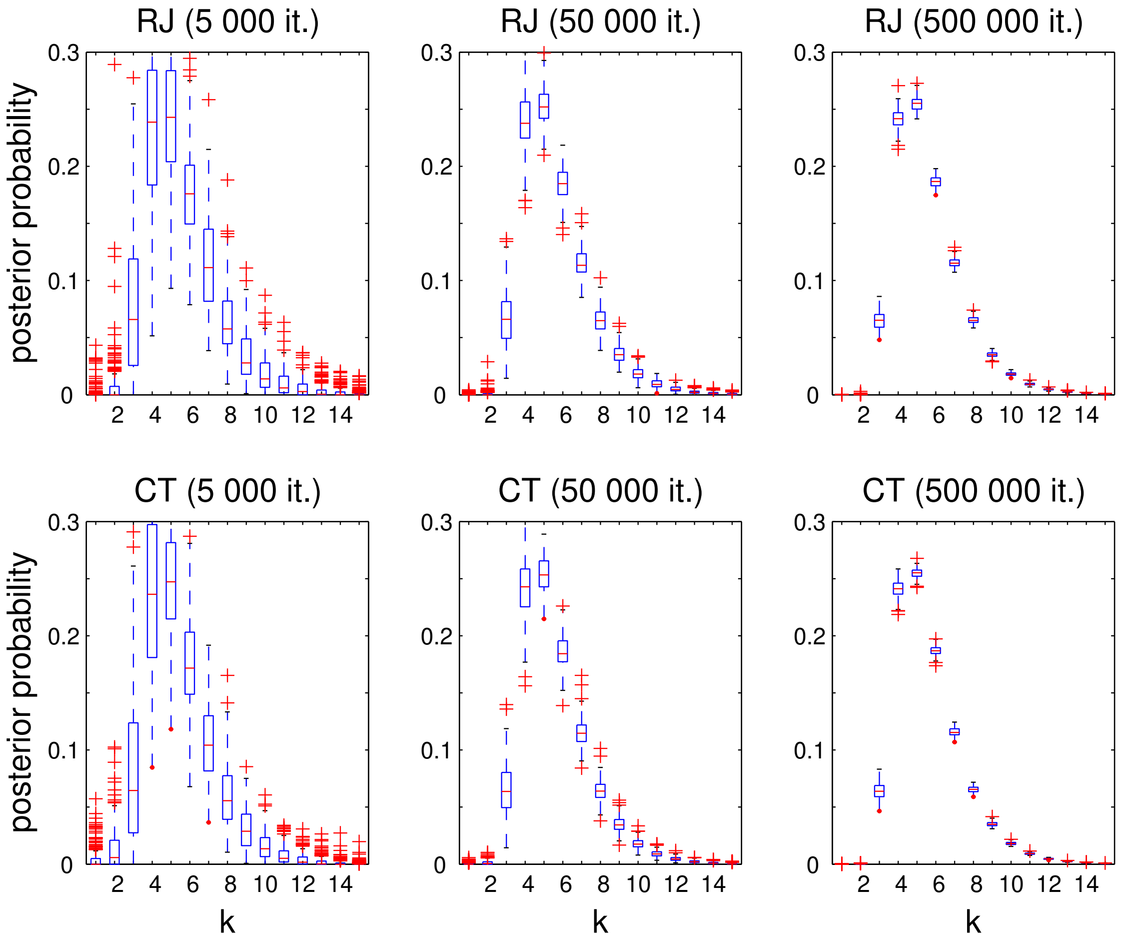

Fig. 2. Galaxy dataset, box plots for the estimated posterior on k obtained from 200 independent runs: RJMCMC (top) and CTMCMC (bottom). The number of iterations varies from 5,000 (left), to 50,000 (middle) and 500,000 (right).

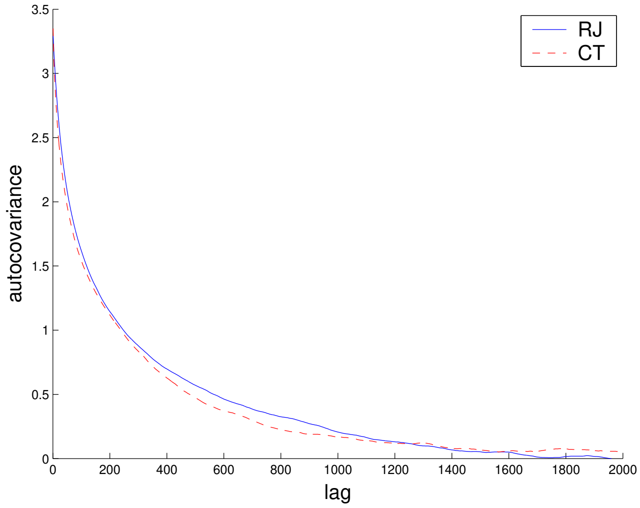

Fig. 3. Estimated autocovariance function for k with RJMCMC and CTMCMC.

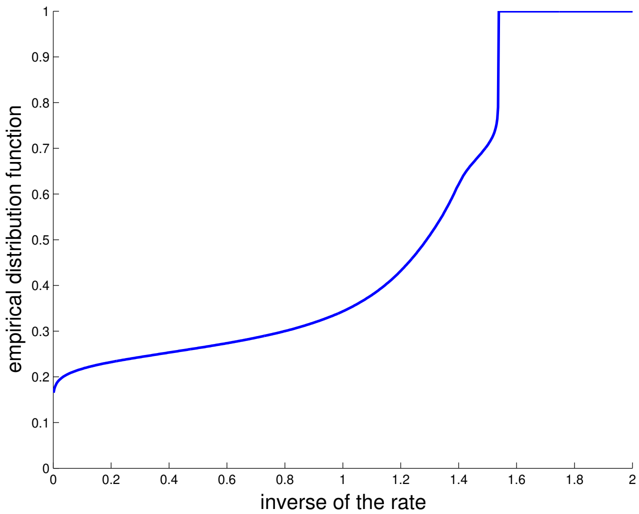

Fig. 4. Empirical distribution function of the inverse rates in CTMCMC. The maximal value corresponds to the addition of the fixed rates: $1/(\eta^F + \eta^B) = 1/(0.3 + 0.35)$ and thus occurs in configurations in which all death rates are negligible.

4.4.3. Samplers with split-and-combine moves

Richardson and Green (1997) suggested that for mixture models, it is profitable to allow moves that can combine two components in a single one or conversely split a component. The inclusion of such moves in the CTMCMC framework is straightforward and has been discussed in Section 4.3.

Fig. 5 is the equivalent of Fig. 2 with all types of moves enabled; here, $P^F = \eta^F = P^B = \eta^B = P^S = \eta^S = 0.2$ is used, where P^S and η^S are the probability of proposing a split move in RJMCMC and the split rate in CTMCMC, respectively. Looking in greater detail at the 5,000 iterations plot, it is possible to see a small advantage for the continuous time sampler: the reversible jump one has a small downward bias for $k = 3$ and its variability is slightly larger at all bins. Part of the explanation is that the weights (inverse rates) in the continuous time sampler have a very similar distribution for the death and combine moves whereas the acceptance probabilities for these are very different in the reversible jump sampler, where deaths are accepted about three times more often. This is due to the fact that even when k is large, there are always at least one or two pairs which have a reasonable rate of being combined and these are selected by the continuous time sampler. In contrast, when k is large, the reversible jump sampler has a low probability of drawing precisely these few pairs.

Another interesting conclusion to be drawn from Fig. 2 and Fig. 5 is that the inclusion of the split-and-combine moves does not significantly improve the accuracy of the results. This is understandable for RJMCMC since split proposals need to be very carefully tuned in order to maintain reasonable acceptance probabilities (see also Appendix C). For CTMCMC however, the same conclusion is also true despite the advantage mentioned above.

In conclusion, were we to rank all techniques on the basis of their computation time, as detailed in Section 4.4.1, the optimal choice would be RJMCMC with birth-and-death only, very closely followed by the equivalent CTMCMC sampler; then, at some distance, RJMCMC with both types of dimension changing moves enabled and finally CTMCMC in the same conditions, which is unattractive because of its high computational cost.

5. Discussion

Our work suggests that there is no clearcut improvement in using continuous time MCMC algorithms: While discrete time moves can also be implemented in continuous time, this alternative implementation does not bring a visible upgrade in the performances of the algorithms. If anything, the continuous time samplers are slower, because they involve consideration of the whole

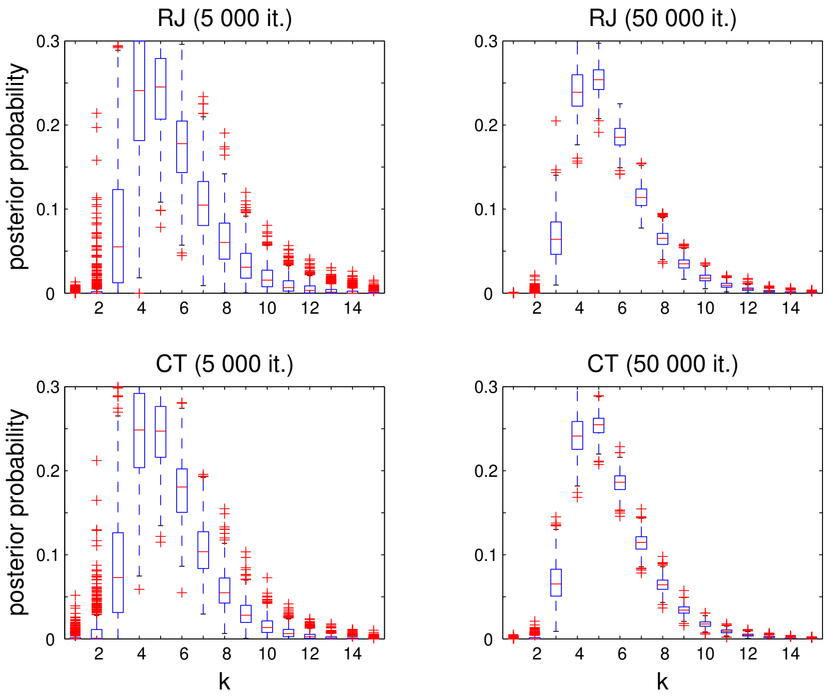

Fig. 5. Galaxy dataset, box plot for the estimated posterior on k obtained from 500 independent runs: Top RJMCMC and bottom, CTMCMC. The number of iterations varies from 5,000 (left plots) to 50,000 (right plots).

range of possible moves and their respective rates after *each* move. Repeated calls to the likelihood function are very costly in computing time and/or memory.

The advantage of continuous time samplers is rather their ability to move to unlikely places: given that the split and birth rates are independent of the data, the algorithm can impose moves to low probability regions of the parameter space. Such regions are of little interest for inference but they can constitute a kind of springboard for the Markov chains, allowing these to move from one mode of the posterior distribution to another one. But this potentially better mixing behaviour can only be achieved when a wide variety of moves are proposed simultaneously, as illustrated in Fig. 5.

A typical set-up of BDMCMC is to let $\beta(\boldsymbol{\theta})$ be constant, say $\beta(\boldsymbol{\theta}) = 1$, since a different constant only rescales time. Likewise, for RJMCMC $b(\boldsymbol{\theta}) = d(\boldsymbol{\theta}) = 1/2$ is typical, except for states $\boldsymbol{\theta}$ with $k = 1$ for which $b(\boldsymbol{\theta}) = 1$. Under these assumptions, equations (2) and (7) relate as $A = (k+1)\delta^{-1}$. Since both samplers have the same stationary distribution, we find that if one of the algorithms performs poorly, so does the other one. For RJMCMC this is manifested as small A's—birth proposals are rarely accepted—while for BDMCMC it is manifested as large δ's—new components are indeed born but die again quickly.

The "attractive alternative" to Richardson and Green (1997) in terms of mixing over the values of k, as reported in Stephens (2000, Section 5.3), is thus not to be sought in the continuous time nature of his algorithm, but rather in the different choices made in the sampler: Stephens (2000) used birth-and-death moves only for modifying the dimension of the model, and these moves did not involve the complete data, that is, the component labels, while Richardson and Green (1997) used split-and-merge moves as well and carried along the component labels through all moves, including the dimension-changing ones. The issue of completion is not directly related to the central theme of this paper, but it may be that the absence of completion explains the different behaviour of the sampler. This was not the case however in the fixed-k mixture setting studied by Celeux *et al.* (2000).

Finally we perceive Rao–Blackwellisation as an advantage of continuous time algorithms; this feature is, as noted above, obtained at no extra cost. Rao–Blackwellisation could in principle be carried out in discrete time as well—holding times have geometric distributions—but, there, the expected holding times cannot be computed easily; see (9) in the proof of Lemma 1 below. See also Casella and Robert (1996) for another Rao–Blackwellisation of Metropolis algorithms.

Acknowledgements

This work was started during the third author's visit to Paris in autumn of 2000, partially supported by CREST, INSEE, and by CNRS (LTCI, ENST Paris). The authors are grateful to the Joint Editor, the Associate Editor and the referee for their detailed, constructive, and encouraging comments that helped to improve this paper, and to Jesper Møller, Gareth Roberts and Matthew Stephens for helpful discussions and suggestions. The wind intensity data were kindly provided by Christian Francq.

A. Proof of Theorem 1

For $\boldsymbol{\theta} \in \Theta^{(k)}$, let

$$\lambda(\boldsymbol{\theta}) = \beta(\boldsymbol{\theta}) + \sum_{i=1}^{k} \delta(\boldsymbol{\theta} \setminus (w_i, \phi_i); (w_i, \phi_i))$$

be the overall rate of leaving state $\boldsymbol{\theta}$ in the BDMCMC sampler and let $\lambda_N(\boldsymbol{\theta})$ be the overall probability of moving away from state $\boldsymbol{\theta}$ (in one step) in the RJMCMC sampler.

Before proving the theorem, we state and prove a lemma.

Lemma 1. *For each $k \geq 1$ and $\boldsymbol{\theta}' \in \Theta^{(k)}$, there is a neighbourhood $G \subseteq \Theta^{(k)}$ of $\boldsymbol{\theta}'$ such that $\sup_{\boldsymbol{\theta} \in G} |N\lambda_N(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| \to 0$ as $N \to \infty$.*

Proof. We first note that for $\boldsymbol{\theta} \in \Theta^{(k)}$, $\lambda_N(\boldsymbol{\theta})$ can be written

$$\begin{aligned} \lambda_N(\boldsymbol{\theta}) = & \int b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} h(\boldsymbol{\theta}; (w,\phi))\, d(w,\phi) \\ & + \sum_{i=1}^{k} d_N(\boldsymbol{\theta}) \frac{1}{k} \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i,\phi_i); \boldsymbol{\theta}), 1\}. \end{aligned} \tag{9}$$

Thus

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in G} |N\lambda_N(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| \\ & \leq \int \sup_{\boldsymbol{\theta} \in G} |N b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} h(\boldsymbol{\theta}; (w,\phi)) - \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w,\phi))|\, d(w,\phi) && (10) \\ & + \sum_{i=1}^{k} \sup_{\boldsymbol{\theta} \in G} |\frac{1}{k} N d_N(\boldsymbol{\theta}) \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i,\phi_i); \boldsymbol{\theta}), 1\} - \delta(\boldsymbol{\theta} \setminus (w_i,\phi_i); (w_i,\phi_i))|. && (11) \end{aligned}$$

We start by looking at the 'birth part' (10) of this expression. We shall prove that it tends to zero by showing that the integrand tends to zero for all (w, ϕ) and showing that the integrand is dominated, for all sufficiently large N, by an integrable function. Bound the integrand as

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in G} |N b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} h(\boldsymbol{\theta}; (w,\phi)) - \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w,\phi))| \\ & \qquad \leq \sup_{\boldsymbol{\theta} \in G} |N b_N(\boldsymbol{\theta}) - \beta(\boldsymbol{\theta})| \times 1 \times \sup_{\boldsymbol{\theta} \in G} h(\boldsymbol{\theta}; (w,\phi)) && (12) \\ & \qquad + \sup_{\boldsymbol{\theta} \in G} \beta(\boldsymbol{\theta}) \times \sup_{\boldsymbol{\theta} \in G} |\min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} h(\boldsymbol{\theta}; (w,\phi)) - h(\boldsymbol{\theta}; (w,\phi))|. && (13) \end{aligned}$$

For $\beta \geq 0$ and $N > \beta$,

$$\frac{\beta}{N} - \frac{1}{2}\frac{\beta^2}{N^2} \leq 1 - e^{-\beta/N} \leq \frac{\beta}{N},$$

so that

$$|N(1-e^{-\beta/N})-\beta| \leq \beta^2/2N.$$

Hence, for sufficiently large N, (12) is bounded by

$$\frac{1}{2N}\sup_{\boldsymbol{\theta}\in G}\beta^2(\boldsymbol{\theta})\times\sup_{\boldsymbol{\theta}\in G}h(\boldsymbol{\theta};(w,\phi)); \tag{14}$$

by (A1) and (A3), for an appropriate G this expression tends to zero as $N\to\infty$ and is dominated by an integrable function.

Regarding (13), it is dominated by an integrable function similar to (14) (remove $1/(2N)$ and the squaring), and it remains to show that it tends to zero as $N\to\infty$. We have

$$\begin{aligned}&|\min\{A_N(\boldsymbol{\theta};\boldsymbol{\theta}\cup(w,\phi)),1\}h(\boldsymbol{\theta};(w,\phi))-h(\boldsymbol{\theta};(w,\phi))| = h(\boldsymbol{\theta};(w,\phi))\\&\quad-\min\left\{\frac{L(\boldsymbol{\theta}\cup(w,\phi))r(\boldsymbol{\theta}\cup(w,\phi))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})}\times\frac{d_N(\boldsymbol{\theta}\cup(w,\phi))}{b_N(\boldsymbol{\theta})}(1-w)^{k-1},h(\boldsymbol{\theta};(w,\phi))\right\}.\end{aligned}$$

By (A2), for each (w,ϕ), $L(\boldsymbol{\theta}\cup(w,\phi))r(\boldsymbol{\theta}\cup(w,\phi))$ and $L(\boldsymbol{\theta})r(\boldsymbol{\theta})$ are bounded away from infinity and zero, respectively, on a sufficiently small G. Likewise, by (A1), $d_N(\boldsymbol{\theta}\cup(w,\phi))$ and $b_N(\boldsymbol{\theta})$ tend to unity and zero, respectively, uniformly over such a G. Finally, by (A3), $h(\boldsymbol{\theta};(w,\phi))$ is bounded on an appropriate G, and we conclude that (13) tends to zero uniformly over G as $N\to\infty$ if G is small enough.

We now turn to the 'death part' (11). By arguments similar to those above, for large N and sufficiently small G it holds that

$$\begin{aligned}&\frac{1}{k}Nd_N(\boldsymbol{\theta})\min\{A_N^{-1}(\boldsymbol{\theta}\setminus(w_i,\phi_i);\boldsymbol{\theta}),1\}\\&\quad=\frac{1}{k}N\min\left\{\frac{L(\boldsymbol{\theta}\setminus(w_i,\phi_i))r(\boldsymbol{\theta}\setminus(w_i,\phi_i))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})}\times\frac{b_N(\boldsymbol{\theta}\setminus(w_i,\phi_i))h(\boldsymbol{\theta}\setminus(w_i,\phi_i);(w_i,\phi_i))}{(1-w_i)^{k-2}},d_N(\boldsymbol{\theta})\right\}\\&\quad=\frac{L(\boldsymbol{\theta}\setminus(w_i,\phi_i))r(\boldsymbol{\theta}\setminus(w_i,\phi_i))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})}\times\frac{1}{k}\times\frac{Nb_N(\boldsymbol{\theta})h(\boldsymbol{\theta}\setminus(w_i,\phi_i);(w_i,\phi_i))}{(1-w_i)^{k-2}}\end{aligned}$$

uniformly over G, and, also using arguments as above, one can show the right hand side of this expression converges to $\delta(\boldsymbol{\theta}\setminus(w_i,\phi_i);(w_i,\phi_i))$ as $N\to\infty$, uniformly over a small enough G. □

Recall the definitions of jump times and the jump chain in Section 2.5. The sequence $\{\widetilde{\boldsymbol{\theta}}_n, T_n-T_{n-1}\}$ of visited states and holding times form a Markov renewal process (MRP). The transition kernel of this MRP is denoted by K, that is, $K(\boldsymbol{\theta};A\times B)=P(\widetilde{\boldsymbol{\theta}}_n\in A, T_n-T_{n-1}\in B\mid\widetilde{\boldsymbol{\theta}}_{n-1}=\boldsymbol{\theta})$. Since $\{\boldsymbol{\theta}(t)\}$ is Markov, the conditional distribution of T_n-T_{n-1} given $\widetilde{\boldsymbol{\theta}}_{n-1}=\boldsymbol{\theta}$ is exponential with rate $\lambda(\boldsymbol{\theta})$. In addition, $\boldsymbol{\theta}(T_n)$ and T_n-T_{n-1} are conditionally independent. Similarly, $\{\boldsymbol{\theta}^N(t)\}$ is a semi-Markov process with jump times $\{T_n^N\}$ in the lattice i/N, and the kernel of the associated MRP is denoted by K_N. Since $\{\boldsymbol{\theta}_n^N\}$ is Markov, $\boldsymbol{\theta}^N(T_n^N)$ and $T_n^N-T_{n-1}^N$ are conditionally independent given $\boldsymbol{\theta}^N(T_{n-1}^N)$.

Proof of Theorem 1. Using results of Karr (1975), it is sufficient to prove that for each real-valued uniformly continuous function g on $\Theta\times[0,\infty)$,

(i) $Kg(\boldsymbol{\theta})$ is continuous on Θ;

(ii) $K_Ng(\boldsymbol{\theta})\to Kg(\boldsymbol{\theta})$ uniformly on compact subsets of Θ as $N\to\infty$.

We start by showing (ii). By the structure of Θ, it is sufficient to show that for each $\boldsymbol{\theta}'\in\Theta^{(k)}$, there is a neighbourhood $G\subseteq\Theta^{(k)}$ of $\boldsymbol{\theta}'$ such that $K_Ng(\boldsymbol{\theta})\to Kg(\boldsymbol{\theta})$ uniformly on G, and this is what we will do. For $\boldsymbol{\theta} \in \Theta^{(k)}$, $K_N g(\boldsymbol{\theta})$ and $Kg(\boldsymbol{\theta})$ can be written

$$
\begin{aligned}
K_N g(\boldsymbol{\theta}) = & \sum_{m=1}^{\infty} \int (1-\lambda_N(\boldsymbol{\theta}))^{m-1} b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} \\
& \qquad h(\boldsymbol{\theta}; (w,\phi)) g\left(\boldsymbol{\theta} \cup (w,\phi), \frac{m}{N}\right) d(w,\phi) \\
& + \sum_{m=1}^{\infty} (1-\lambda_N(\boldsymbol{\theta}))^{m-1} \sum_{i=1}^{k} d_N(\boldsymbol{\theta}) \frac{1}{k} \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i,\phi_i); \boldsymbol{\theta}), 1\} g\left(\boldsymbol{\theta} \setminus (w_i,\phi_i), \frac{m}{N}\right) \\
& = \int_0^{\infty} \int (1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu \rfloor} N b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} \\
& \qquad h(\boldsymbol{\theta}; (w,\phi)) g\left(\boldsymbol{\theta} \cup (w,\phi), \frac{\lceil Nu \rceil}{N}\right) du\, d(w,\phi) \\
& + \int_0^{\infty} (1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu \rfloor} \\
& \quad \sum_{i=1}^{k} N d_N(\boldsymbol{\theta}) \frac{1}{k} \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i,\phi_i); \boldsymbol{\theta}), 1\} g\left(\boldsymbol{\theta} \setminus (w_i,\phi_i), \frac{\lceil Nu \rceil}{N}\right) du;
\end{aligned}
$$

$$
\begin{aligned}
Kg(\boldsymbol{\theta}) = & \int_0^{\infty} \int \lambda(\boldsymbol{\theta}) e^{-\lambda(\boldsymbol{\theta})u} \frac{\beta(\boldsymbol{\theta})}{\lambda(\boldsymbol{\theta})} h(\boldsymbol{\theta}; (w,\phi)) g(\boldsymbol{\theta} \cup (w,\phi), u)\, du\, d(w,\phi) \\
& + \int_0^{\infty} \sum_{i=1}^{k} \lambda(\boldsymbol{\theta}) e^{-\lambda(\boldsymbol{\theta})u} \frac{\delta(\boldsymbol{\theta} \setminus (w_i,\phi_i); (w_i,\phi_i))}{\lambda(\boldsymbol{\theta})} g(\boldsymbol{\theta} \setminus (w_i,\phi_i), u)\, du \\
& = \int_0^{\infty} \int e^{-\lambda(\boldsymbol{\theta})u} \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w,\phi)) g(\boldsymbol{\theta} \cup (w,\phi), u)\, du\, d(w,\phi) \\
& + \int_0^{\infty} \sum_{i=1}^{k} e^{-\lambda(\boldsymbol{\theta})u} \delta(\boldsymbol{\theta} \setminus (w_i,\phi_i); (w_i,\phi_i)) g(\boldsymbol{\theta} \setminus (w_i,\phi_i), u)\, du,
\end{aligned}
$$

where $\lceil x \rceil$ is the smallest integer no smaller than x.

We again start by looking at the 'birth parts' of the kernels, bounding the corresponding part of $|K_N g(\boldsymbol{\theta}) - Kg(\boldsymbol{\theta})|$ as

$$
\begin{aligned}
\int_0^{\infty} \int \sup_{\boldsymbol{\theta} \in G} \Bigg| & (1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu \rfloor} N b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w,\phi)), 1\} h(\boldsymbol{\theta}; (w,\phi)) \\
& \times g\left(\boldsymbol{\theta} \cup (w,\phi), \frac{\lceil Nu \rceil}{N}\right) - e^{-\lambda(\boldsymbol{\theta})u} \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w,\phi)) g(\boldsymbol{\theta} \cup (w,\phi), u) \Bigg| du\, d(w,\phi).
\end{aligned}
$$

We wish to prove that this expression tends to zero as $N \to \infty$. We can do this by showing that the integrand tends to zero for all $u \geq 0$ and all (w,ϕ) and that there exists a dominating (for all sufficiently large N) integrable function.

In order to accomplish this, we add and subtract a number of telescoping terms, giving us

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}\in G} \Big| (1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu\rfloor} N b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta};\boldsymbol{\theta}\cup(w,\phi)),1\} h(\boldsymbol{\theta};(w,\phi)) g\left(\boldsymbol{\theta}\cup(w,\phi),\frac{\lceil Nu\rceil}{N}\right) \\
-e^{-\lambda(\boldsymbol{\theta})u}\beta(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))g(\boldsymbol{\theta}\cup(w,\phi),u)\Big| \\
\le \sup_{\boldsymbol{\theta}\in G}\left|(1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu\rfloor}-e^{-\lambda(\boldsymbol{\theta})u}\right| \times \sup_{\boldsymbol{\theta}\in G} N b_N(\boldsymbol{\theta}) \times 1 \times \overline{h}(w,\phi)\times ||g||_\infty \\
+ \sup_{\boldsymbol{\theta}\in G} e^{-\lambda(\boldsymbol{\theta})u} \times \sup_{\boldsymbol{\theta}\in G} N b_N(\boldsymbol{\theta}) \times 1 \times \overline{h}(w,\phi) \times \delta^g_{1/N} \\
+ \sup_{\boldsymbol{\theta}\in G} e^{-\lambda(\boldsymbol{\theta})u} \times \sup_{\boldsymbol{\theta}\in G} |N b_N(\boldsymbol{\theta})-\beta(\boldsymbol{\theta})| \times 1 \times \overline{h}(w,\phi) \times ||g||_\infty \\
+ \sup_{\boldsymbol{\theta}\in G} e^{-\lambda(\boldsymbol{\theta})u} \times \sup_{\boldsymbol{\theta}\in G} \beta(\boldsymbol{\theta}) \\
\times \sup_{\boldsymbol{\theta}\in G} |\min\{A_N(\boldsymbol{\theta};\boldsymbol{\theta}\cup(w,\phi)),1\}h(\boldsymbol{\theta};(w,\phi)) - h(\boldsymbol{\theta};(w,\phi)| \times ||g||_\infty,
\end{aligned}
$$

where $\overline{h}(w,\phi) = \sup_{\boldsymbol{\theta}\in G} h(\boldsymbol{\theta};(w,\phi))$ and $\delta^g_{1/N} = \sup_{\Delta((\boldsymbol{\theta},u),(\boldsymbol{\theta}',u'))\le 1/N} |g(\boldsymbol{\theta},u)-g(\boldsymbol{\theta}',u')|$ is the modulus of continuity of g; Δ is a metric making $\Theta\times[0,\infty)$ separable and complete. All of the terms on the right hand side but the first one can be treated as in the proof of the lemma, with the extra observation that $\lambda(\boldsymbol{\theta}) \ge \beta(\boldsymbol{\theta})$ is bounded away from zero on compact subsets of Θ. Moreover, since

$$(1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu\rfloor} \le e^{-\lambda_N(\boldsymbol{\theta})\lfloor Nu\rfloor} = e^{-N\lambda_N(\boldsymbol{\theta})(\lfloor Nu\rfloor/N)},$$

the lemma implies that the first term is, for large N's, dominated by an integrable function. Finally

$$
\begin{aligned}
(1-\lambda_N(\boldsymbol{\theta}))^{\lfloor Nu\rfloor} - e^{-\lambda(\boldsymbol{\theta})u} &\le e^{-\lambda_N(\boldsymbol{\theta})\lfloor Nu\rfloor} - e^{\lambda(\boldsymbol{\theta})u} \\
&= e^{-\lambda(\boldsymbol{\theta})u}\left(e^{-\lambda(\boldsymbol{\theta})(\lfloor Nu\rfloor/N-u)+\lfloor Nu\rfloor o(1/N)}-1\right),
\end{aligned}
$$

where, by the lemma, the $o(1/N)$ term is uniform over a small G so that the right hand side tends to zero uniformly over such a G. The inequality $\log(1-x)\ge -x-2x^2$ for $0\le x\le 1/2$ leads to a reverse inequality which is handled similarly.

The 'death parts' of the kernels, that is, bounding the corresponding parts of $|K_N g(\boldsymbol{\theta})-Kg(\boldsymbol{\theta})|$, can be handled combining arguments for the 'birth parts' and arguments used to prove the lemma.

Finally requirement (i) above can be proved using similar techniques. □

B. The Jacobian for the split-and-combine move

The parts of the Jacobian determinant corresponding to the split move in Section 3.2 are

(a) ω_{j,i_0};
(b) $2\omega_{i_0,i}/\xi_i$;
(c)

$$\omega^3_{i_0,i_0}\begin{vmatrix} \varepsilon_{i_0}\xi_{i_1} & \varepsilon_{i_0}/\xi_{i_1} & (1-\varepsilon_{i_0})\xi_{i_2} & (1-\varepsilon_{i_0})/\xi_{i_2} \\ \varepsilon_{i_0} & -\varepsilon_{i_0}/\xi^2_{i_1} & 0 & 0 \\ 0 & 0 & (1-\varepsilon_{i_0}) & -(1-\varepsilon_{i_0})/\xi^2_{i_2} \\ \xi_{i_1} & 1/\xi_{i_1} & -\xi_{i_2} & -1/\xi_{i_2} \end{vmatrix},$$

that is,

$$\omega^3_{i_0,i_0}\begin{vmatrix} \varepsilon_{i_0}\xi_{i_1} & 0 & \xi_{i_2} & 0 \\ \varepsilon_{i_0} & -2\varepsilon_{i_0}/\xi^2_{i_1} & 0 & 0 \\ 0 & 0 & (1-\varepsilon_{i_0}) & -2(1-\varepsilon_{i_0})/\xi^2_{i_2} \\ (1+\xi_{i_1})/2 & 0 & -(1+\xi_{i_2})/2 & 0 \end{vmatrix} = 4\,\omega^3_{i_0,i_0}\frac{\varepsilon_{i_0}(1-\varepsilon_{i_0})}{\xi_{i_1}\xi_{i_2}};$$

(d) this part of the Jacobian can be obtained as

$$4\sigma_{i_1}^2\sigma_{i_2}^2(\alpha-\sigma_{i_1})(\alpha-\sigma_{i_2})\big/\alpha(\alpha-\sigma_{i_0})\sigma_{i_0}^2,$$

where $\sigma_{i_1}=\alpha\text{-logit}^{-1}\left[\alpha\text{-logit}(\sigma_{i_0})+\varepsilon_\sigma\right]$ and $\sigma_{i_2}=\alpha\text{-logit}^{-1}\left[\alpha\text{-logit}(\sigma_{i_0})-\varepsilon_\sigma\right]$ (differentiating with respect to $\sigma_{i_0}^2$).

C. Implementation details for the numerical comparison experiment

C.1. Model

We consider a Gaussian scalar mixture model with parameters $(w_{1:k},\mu_{1:k},\upsilon_{1:k})$, where the υ_i's are the variances. The prior modelling is such that

$$k\sim\mathcal{U}(\{1,\dots,M\}),\quad w_{1:k}\sim\mathcal{D}_k(1,\dots,1),\quad \mu_i\sim\mathcal{N}(0,\kappa),\quad \upsilon_i^{-1}\sim\mathcal{G}a(\alpha,\beta),$$

where $\mathcal{D}$ denotes the Dirichlet distribution, and with the following hyperparameters (scaled for the recentered Galaxy dataset):

$$M=15,\quad \kappa=(\max\{Y_i\}_{1\le i\le n}-\min\{Y_i\}_{1\le i\le n})^2,\quad \alpha=0.5,\quad \beta=10^{-3}.$$

C.2. Sampler

The sampler consists of fixed-k, birth-and-death and split-and-combine moves, for both the reversible jump and the continuous time versions. The fixed-k moves are proposed with probability P^F in RJMCMC and with rate $\eta^F=P^F$ in CTMCMC (for $k=M$ these numbers are both zero). In both cases, it consists of the three Metropolis-Hasting proposals (weights, means, variances) with independent accept/reject decisions. The proposal is a multiplicative lognormal random walk on the w_i's, $\mathcal{LN}(0,\eta)$, an additive normal random walk on the μ_i's, $\mathcal{N}(0,\rho)$, and a multiplicative lognormal random walk on the υ_i's, $\mathcal{LN}(0,\nu)$. These moves can just as easily be carried out globally or one component at a time, but only global moves (i.e. with proposal affecting the parameters of all the components) were used in our simulations. The sampler parameters were tuned in order to achieve acceptance rates that stay in the range 0.3–0.7 for all values of $k\le 15$, and we obtained $\eta=0.05$, $\rho=\kappa/(2000k)$, and $\nu=0.08$. The normalisation of ρ by k tends to stabilise the acceptance rate (with constant ρ the acceptance rate drops for high values of k). Despite good mixing, these moves alone are not sufficient to generate label switching (see Celeux *et al.*, 2000).

The birth-and-death moves are Stephens' (2000), namely such that when in a k component configuration we propose a new component from the prior according to $w\sim\mathcal{B}e(1,k)$, $\mu\sim\mathcal{N}(0,\kappa)$, and $\upsilon^{-1}\sim\mathcal{G}a(\alpha,\beta)$, where $\mathcal{B}e$ is the Beta distribution. For the continuous time version of the move, the birth rate is $\eta^B=P^B$ (again, these numbers are zero for $k=M$) and the death rates are given by

$$\eta^B\,L(\boldsymbol{\theta})/L(\boldsymbol{\theta}\cup(w,\phi))\times k+1,$$

where $\phi=(\mu,\upsilon)$; notice that $h(\boldsymbol{\theta};(w,\phi))/(1-w)^{k-1}$ in (2) cancels with the ratio $r(\boldsymbol{\theta})/r(\boldsymbol{\theta}\cup(w,\phi))$ of prior densities.

The split-and-combine move is inspired by Richardson and Green (1997). If a component i is proposed to be split, this is done according to

(a) $w_i\mapsto(\xi w_i,(1-\xi)w_i)$ with $\xi\sim\mathcal{B}e(\gamma_S,\gamma_S)$,
(b) $\mu_i\mapsto(\mu_i-\xi,\mu_i+\xi)$ with $\xi\sim\mathcal{N}(0,\rho_S)$,
(c) $\upsilon_i\mapsto(\upsilon_i/\xi,\upsilon_i\xi)$ with $\xi\sim\mathcal{LN}(0,\nu_S)$.

In the current implementation P^S is constant except for edge effects ($P^S(M)=0$). On the galaxy data, the choice of parameters that maximises the acceptance rate for the split-and-combine move is $\gamma_S=1$, $\rho_S=0.2$, and $\nu_S=3$. However, the acceptance rate is then only 4.3% (compared to 13.3% for the birth-and-death move).

References

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 164–171.

Breiman, L. (1992) *Probability.* SIAM, Philadelphia, PA.

Cappé, O. and Robert, C.P. (2000) MCMC: Ten years and still running! *J. Am. Statist. Ass.* **95**, 1282–1286.

Casella, G. and Robert, C.P. (1996) Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94.

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

Clifford, P. and Nicholls, G. (1994) Comparison of birth-and-death and Metropolis–Hastings Markov Chain Monte Carlo for the Strauss process. Department of Statistics, Oxford University.

Francq, C. and Roussignol, M. (1997) On white noises driven by hidden Markov chains. *J. Time Ser. Anal.* **18**, 553–578.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339.

Geyer, C.J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21**, 359–373.

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

Grenander, U. and Miller, M. (1994) Representations of knowledge in complex systems (with discussion). *J. R. Statist. Soc.* B **56**, 549–603.

Hurn, M., Justel, A. and Robert, C.P. (2003) Estimating mixtures of regressions *J. Comput. Graph. Statist.* (to appear).

Karr, A.F. (1975) Weak convergence of a sequence of Markov chains. *Z. Wahrsch. ver. Geb.* **33**, 41–48.

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

Preston, C.J. (1976) Spatial birth-and-death processes. *Bull. Inst. Internat. Statist.* **46**, 371–391.

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. Corrigendum in **60** (1998), 661.

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

Robert, C.P., Rydén, T. and Titterington, D.M. (1999) Convergence controls for MCMC algorithms, with applications 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.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Am. Statist. Ass.* **85**, 617–624.

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