Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers†

Olivier Cappé

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

Christian P. Robert

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

Tobias Rydén‡

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

Summary. At present, reversible jump methods are the most common 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 address the comparison of both methods by demonstrating that upon appropriate rescaling of time, the reversible jump chain converges to a limiting continuous time birth-and-death chain. We show in addition that the birth-and-death setting can be generalised to include other types of jumps like split/combine jumps in the spirit of Richardson and Green (1997). We illustrate these extensions in the case of hidden Markov models.

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

Keywords: Bayesian inference, birth-and-death process, completion, hidden Markov model, Jacobian, label switching, MCMC algorithms, mixture distribution, rescaling

Résumé. Les méthodes à sauts réversibles sont à présent la technique la plus utilisée dans l'exploration de modèles statistiques à dimension variable. Une autre approche a cependant été introduite récemment par Stephens (*Annals of Statistics*, 200) dans le cadre des mélanges de lois; elle utilise une représentation par processus ponctuel à sauts. Nous considérons en parallèle les deux méthodes et démontrons que, via un changement d'échelle idoine la chaîne produite par la méthode à sauts réversibles converge vers une chaîne en temps continu correspondant à un processus de vie et de mort. Nous montrons par ailleurs que le processus de vie et de mort peut se généraliser pour inclure des propositions comme les déplacements de fusion et d'éclatement dans l'esprit de Richardson et Green (*J. Royal Statistical Society*, 1997). Nous illustrons ces comparaisons dans le cas de modèles de chaînes de Markov cachées.

Mots-clés: Statistique bayésienne, processus de vie et de mort, complétion, chaîne de Markov cachée, Jacobien, algorithme MCMC, mélange de lois, inidentifiabilité, changement d'échelle

†Work partially supported by EU TMR network ERB–FMRX–CT96–0095 on *'Computational and Statistical Methods for the Analysis of Spatial Data'*. The authors are grateful to Gareth Roberts for helpful comments on the Rao-Blackwellisation improvement.

‡Supported by the Swedish Research Council. Partially supported by CREST, INSEE, and by CNRS (URA 820, ENST) during a visit to Paris in autumn of 2000.

1. Introduction

Markov Chain Monte Carlo (MCMC) methods for statistical inference, in particular Bayesian inference, have undoubtedly 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 (2000a,b) rekindled interest in a different method based on continuous time birth-and-death processes for estimating the number of components of the mixture, following earlier proposals by Geyer and Møller (1994), Grenander and Miller (1994) and Phillips and Smith (1996). We will call this approach birth-and-death MCMC (BDMCMC).

A main question addressed in the present paper is as follows: is there a fundamental difference between the reversible jump and birth-and-death MCMC methodologies, or are these approaches similar? As an answer to this question we show in Section 3 that for any BDMCMC process satisfying some weak regularity conditions there exists a sequence of RJMCMC processes that converges, in a sense to be precised below, to the BDMCMC process.

In their application of reversible jump MCMC to mixtures of distributions, Richardson and Green (1997) involved two types of moves that could change the number of components of the mixture: one was *birth/death*, in which a new component is created or an existing one is deleted, and the other was *split/combine*, in which one component is split in two, or two components are combined in one. On the opposite, Stephens (2000a) only dealt with birth/death moves in order to keep the algorithm within the theory of (marked) point processes on general spaces. We show that convergence of reversible jump to birth-and-death MCMC is not limited to moves of this kind however, but is much more general. For example, the above *split/combine* moves could be incorporated. The approach so obtained could be named continuous time reversible jump MCMC and the appropriate theoretical framework is that of Markov jump processes.

The paper is organised as follows: in Section 2, we provide a review of the main features of reversible jump and birth-and-death MCMC methodologies. The convergence of RJMCMC to BDMCMC is established in Section 3. In Section 4, we discuss the generalisation of moves for continuous time MCMC besides birth/death moves, while in Section 5, we show how sampling can be made more efficient in this approach, introducing a continuous time Rao-Blackwellisation scheme. Section 6 illustrates the general continuous time MCMC methodology for hidden Markov models, in parallel with the RJMCMC approach of Robert, Rydén and Titterington (2000). Section 7 concludes with a discussion of the pros and cons of each method.

2. A quick review of reversible jump and birth-and-death MCMC methodologies

In this section we give a quick review of RJMCMC and BDMCMC in the mixture case considered by Stephens (2000a). We will consider the extension of BDMCMC to hidden Markov models in Section 6. Further reading is provided by Richardson and Green (1997, 1998) and Stephens (2000a,b).

2.1. Mixture models

The model we work with thus has a 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 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. Common examples are the Gaussian family, the Gamma family (in which cases ϕ is typically two-dimensional) and the Poisson family (in which case ϕ is one-dimensional). The component weights are non-negative numbers summing up to unity. Note that we write all densities as conditional ones, as our statistical approach is Bayesian. Hence we need to specify a prior density for $(k, \mathbf{w}, \boldsymbol{\phi})$, denoted by $r(k, \mathbf{w}, \boldsymbol{\phi})$. We do not make any further assumptions about

the prior, except that it is proper and that, for each k, it is exchangeable, that is, invariant under permutations of the pairs (w_i, ϕ_i). We also denote by $L(k, \mathbf{w}, \boldsymbol{\phi})$ the likelihood which is given by

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

where $\mathbf{y} = (y_1, \ldots, y_m)$ is the observed sequence. The posterior density, which is our starting point for inference, is thus proportional to $r(k, \mathbf{w}, \boldsymbol{\phi})L(k, \mathbf{w}, \boldsymbol{\phi})$. A real model typically also involves hyperparameters, which as such do not add any further difficulty. We do not specifically address this issue till Section 6 where hyperparameters are used. Below we put $\boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\phi})$; in this notation k is implicit.

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

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

We call such a simulation *completing the sample* as $(\mathbf{z}, \mathbf{y})$ is often referred to as the complete data. As detailed below in the set-up of hidden Markov models and as demonstrated in Celeux *et al.* (2000) for mixtures, the completion by $\mathbf{z}$ is not necessary from a simulation point of view.

2.2. Birth-and-death MCMC

We now study the following form of BDMCMC: in state $\boldsymbol{\theta}$, new components are created (*born*) in continuous time, at rate $\beta(\boldsymbol{\theta})$. Whenever a new component is born in this state, its weight w and parameter ϕ are drawn from a joint density $h(\boldsymbol{\theta}; (w, \phi))$. In order to make space for the new component, the old component weights are scaled down proportionally as 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 also augmented to $\boldsymbol{\theta}$. We denote these operations by '$\cup$', so that the new state is $\boldsymbol{\theta} \cup (w, \phi)$. Furthermore, 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 Jacobian determinant when renormalising the weights. Indeed, when the component (w, ϕ) is removed, the remaining component weights are renormalised as to sum to unity. We denote these two operations by '$\setminus$', so that $\boldsymbol{\theta} = (\boldsymbol{\theta} \cup (w, \phi)) \setminus (w, \phi)$. An important feature of BDMCMC is that (continuous time) jump processes are associated with the birth and death rates: whenever a jump occurs, the corresponding move is always accepted. What replaces the acceptance probability of classical MCMC methods is the holding time in each state. In particular, implausible states, that is states such that $L(\boldsymbol{\theta})r(\boldsymbol{\theta})$ is small, die quickly.

2.3. Reversible jump MCMC

We now turn to the corresponding reversible jump MCMC sampler. In a k component state $\boldsymbol{\theta}$, at each iteration, the algorithm proposes with probability $b(\boldsymbol{\theta})$ to create a new component and with probability $d(\boldsymbol{\theta})$ it proposes to kill one. Obviously, $b(\boldsymbol{\theta}) + d(\boldsymbol{\theta}) = 1$. 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, each component is selected with equal probability. A new component is accepted with probability $\min(1, A)$, where $A = A(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi))$ is given by

$$\begin{aligned} A(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)) &= \frac{L(\boldsymbol{\theta} \cup (w, \phi))r(\boldsymbol{\theta} \cup (w, \phi))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})} \times (k + 1) \times \frac{d(\boldsymbol{\theta} \cup (w, \phi))}{(k + 1)b(\boldsymbol{\theta})} \times \frac{(1 - w)^{k-1}}{h(\boldsymbol{\theta}; (w, \phi))} \\ &= \frac{L(\boldsymbol{\theta} \cup (w, \phi))r(\boldsymbol{\theta} \cup (w, \phi))}{L(\boldsymbol{\theta})r(\boldsymbol{\theta})} \times \frac{d(\boldsymbol{\theta} \cup (w, \phi))}{b(\boldsymbol{\theta})} \times \frac{(1 - w)^{k-1}}{h(\boldsymbol{\theta}; (w, \phi))}. \end{aligned} \tag{3}$$

Here the first ratio, combined with the first factor $k+1$, is the ratio of posterior densities, $b(\boldsymbol{\theta})h(\boldsymbol{\theta};(w,\phi))$ is the density of proposing a new component (w,ϕ) and $d(\boldsymbol{\theta}\cup(w,\phi))/(k+1)$ is the probability of proposing to kill component (w,ϕ) when in state $\boldsymbol{\theta}\cup(w,\phi)$. Finally $(1-w)^{k-1}$ is the same Jacobian determinant as above. 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.

In (3), the first factor $k+1$ comes from the assumption that in the RJMCMC algorithm we keep $\phi_1,\ldots,\phi_k$ ordered using a predetermined ordering. For example, in the case of Gaussian components we could sort according to the mean. This ordering will, loosely speaking, reduce the size of the space of k-component parameters by a factor $k!$, and the factor $k+1$ is the ratio $(k+1)!/k!$. This factor should thus be associated with the posterior density ratio. We do remark, however, that the assumption of ordered components is a purely technical identifiability device and does not make any practical change to the algorithm; when a new component (w,ϕ) is proposed we keep the components ordered by sorting them. Indeed, if ordering is not imposed, one rather has to work on a quotient space induced by the equivalence relation $\sim$ defined by $\boldsymbol{\theta}\sim\boldsymbol{\theta}'$ if $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are identical up to a permutation of indices. Working with the quotient space also gives rise to a factor $k+1$ as in (2). Both of the above samplers have the posterior density as their stationary distribution. In RJMCMC, one typically includes other kinds of moves such as moves resampling the component weights and the parameters ϕ_i as well as, possibly, the hyperparameters for a fixed k—see, for instance, Richardson and Green (1997). A complete *sweep* of the algorithm consists in the composition of a birth/death move with these other—fixed k—moves. Stephens (2000a) resampled component weights and parameters at regularly separated instants. Sampling for a fixed k can be carried out using a Gibbs move after completing the sample according to (1), but completion was not used by Stephens (2000a) who only considered Metropolis-Hastings updates. As noted above, Richardson and Green (1997) designed, in addition, moves for splitting and combining components (see Section 4 for the generalisation of BDMCMC.)

3. Convergence of reversible jump to birth-and-death MCMC

We shall now, starting from a BDMCMC algorithm as above, construct a sequence of RJMCMC samplers converging, in a certain sense to be defined, to the BDMCMC sampler. Before proceeding we introduce some additional notation. Let $S^{k-1}=\{(w_1,\ldots,w_k):\ w_i>0,\sum w_i=1\}$, denote by Φ the space in which each ϕ_i lies and put $\Theta^{(k)}=S^{k-1}\times\Phi^k$. Thus $\Theta^{(k)}$ is the space of k-dimensional parameters. Finally $\Theta=\cup_{k\geq 1}\Theta^{(k)}$ denotes the overall parameter space.

For $N=1,2,3,\ldots$, we define an RJMCMC sampler by letting

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

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 $n=0,1,2,\ldots$ 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 $\boldsymbol{\theta}^N(t)=\boldsymbol{\theta}^N_{\lfloor Nt\rfloor}$, 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

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

and the right hand side is nothing but $\delta(\boldsymbol{\theta};(w,\phi))$. 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 $\delta(\boldsymbol{\theta};(w,\phi))$, which agrees with the BDMCMC sampler. Thus, summing up, 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. 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. Likewise we assume that Φ has a separable topology which can be metrised by a complete metric. 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 make the following assumptions:

(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.

THEOREM 1. *Under (A1)–(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 $N \to \infty$.*

The proof is given in Appendix A.

4. Generalisations of birth-and-death MCMC

As noted above, Stephens (2000a) resampled component weights and parameters with fixed k, as well as hyperparameters, at equidistant times. This obviously makes the overall process non-Markovian. We can, however, incorporate such moves into the continuous time sampler. Suppose for example that in state $\boldsymbol{\theta}$ of the RJMCMC sampler, a move that resamples component weights and parameters as well as hyperparameters, while keeping k fixed, is proposed with probability $1 - \exp\{-\gamma(\boldsymbol{\theta})/N\}$. Rescaling time as above and passing to the limit produces 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. Of course we can also have different rates for resampling component weights and parameters and hyperparameters, respectively.

A further scope for generalisation is to introduce more complex moves, like the split and combine moves of Richardson and Green (1997). We consider here the case of a split or combine move in the RJMCMC setting where, following Green (1995), the combine move is deterministic. For simplicity, we denote by θ an element of the k component parameter vector $\boldsymbol{\theta}$ and assume that there is no constraint on $\boldsymbol{\theta}$. (In the mixture example considered in Section 2, $\theta = (w,\phi)$ was indeed two or three dimensional and there was a constraint on the set of w's. We will see in Section 6 how the constraint can be effectively removed.)

The RJMCMC sampler proposes 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 $((\boldsymbol{\theta} \setminus \theta) \cup T(\theta,\varepsilon))$ where T is a differentiable one-to-one mapping to Γ^2, where $\theta \in \Gamma$, and ε is a random variable with p.d.f. p. We also assume that the mapping is symmetric in the sense that

$$P(T(\theta,\varepsilon) \in B \times B') = P(T(\theta,\varepsilon) \in B' \times B) \tag{4}$$

for all $B, B' \subseteq \Theta$. For instance if p is a symmetric p.d.f., $T(\theta,\varepsilon) = (\theta - \varepsilon, \theta + \varepsilon)$ is a valid mapping, and likewise, if p is such that ε and ε^{-1} have the same distribution, $T(\theta,\varepsilon) = (\theta\varepsilon, \theta/\varepsilon)$ is a valid mapping. 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 $c_N(\boldsymbol{\theta}) = 1 - s_N(\boldsymbol{\theta})$. The acceptance probability of a split move changing the k component vector $\boldsymbol{\theta}$ to $((\boldsymbol{\theta} \setminus \theta) \cup T(\theta, \varepsilon))$ is given by

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

where, as previously, the factorials in the first ratio stems from the ordering of the components before and after the combine move. The factor 2 is a result of the symmetry assumption (4): in a split move, a component (w, ϕ) can be split into the pair $((w', \phi'), (w'', \phi''))$ as well as into the reversely ordered pair $((w'', \phi''), (w', \phi'))$, but upon sorting the components these configurations are equivalent. However, the two ways of getting there are typically associated with different values of ε and possibly also with different densities $p(\varepsilon)$; the symmetry assumption is precisely what assures that the densities at these two values of ε coincide and hence we may replace the sum of two densities that we would otherwise be required to compute by the factor 2. We could proceed without such symmetry but would then need to consider the densities of ε when combining the pairs (θ', θ'') and (θ'', θ'), respectively, separately.

As in Section 3, we let $s_N(\boldsymbol{\theta}) = 1 - \exp\{-\eta(\boldsymbol{\theta})/N\}$ for some $\eta(\boldsymbol{\theta})$, so that $N s_N(\boldsymbol{\theta}) \to \eta(\boldsymbol{\theta})$, and accordingly scale by N the trajectory of the corresponding discrete time sampler. The limiting continuous time process thus 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

$$\frac{L(\boldsymbol{\theta}) r(\boldsymbol{\theta})}{L\left((\boldsymbol{\theta} \setminus \theta) \cup T(\theta,\varepsilon)\right) r\left((\boldsymbol{\theta} \setminus \theta) \cup T(\theta,\varepsilon)\right)} \times \frac{\eta(\boldsymbol{\theta})}{(k+1)k} \times 2p(\varepsilon) \left| \frac{\partial T(\theta,\varepsilon)}{\partial(\theta,\varepsilon)} \right|. \tag{5}$$

Note that it is not necessary to consider the equivalent discrete time RJMCMC sampler to obtain the above result as it is possible to check directly that the *local balance*

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

holds with $q(((\boldsymbol{\theta} \setminus \theta) \cup T(\theta,\varepsilon)), \boldsymbol{\theta})$ defined by (5). Generally, viewing the continuous time process as a Markov jump process, local balance amounts to requiring

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

where $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the rate of moving from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ (to stress generality we have here absorbed the factorials into the prior densities). Special care is required with such considerations however since the transition kernel of the jump chain (defined in Section 5) typically does not have a density w.r.t. 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, Breiman (1992) and Ripley (1987). Finally we remark that just as continuous time MCMC is not limited to birth-and-death moves, it is not limited to mixture and hidden Markov models either. For example, continuous time MCMC may be applied to any of the examples in Green (1995).

5. Sampling in continuous time

When running a discrete time RJMCMC sampler, its state is typically stored (sampled) after each step or sweep, or on regular intervals in order to decrease inter-sample correlation, as in Richardson and Green (1997) and Robert, Rydén and Titterington (2000), even if convergence assessment for RJMCMC samplers is still in its infancy (Brooks and Giudici, 1999).

In continuous time settings, there are more options. For example, the process may be sampled at regular times, as in Stephens (2000a), or may be sampled using 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 τ_i are the sampling instants. Suppose we adopt the former sampling scheme. If we then let the sampling distance tend 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 $\widetilde{\boldsymbol{\theta}}_n$ has a (conditional) distribution which is exponential with rate $\lambda(\widetilde{\boldsymbol{\theta}}_n)$.

Returning to the sampling scheme, we can then 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 by virtue of the Rao-Blackwell theorem, since

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

When sampling $\{\boldsymbol{\theta}(t)\}$ this way, we only simulate the jump chain and store each state it visits as well as the corresponding expected holding time. Alternatively, the expected holding times may be recomputed later when processing the sampler output. In order to simulate the jump chain we note that its transition law is as follows: the probability of an event happening is proportional to its rate. Hence, for example, the probability of a birth is $\beta(\boldsymbol{\theta})/\lambda(\boldsymbol{\theta})$; if a birth occurs then 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 $\{\boldsymbol{\theta}(t)\}$.

6. An illustration for hidden Markov models

6.1. Setting

We consider in this section an application of the continuous time MCMC methodology to the case of hidden Markov models, as in Robert, Rydén and Titterington (2000). That is, 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 normal variate

$$\mathcal{N}(\mu_{z_n}, \sigma^2_{z_n}).$$

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

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

The ω_{ij}'s are therefore not identified, but this parameterisation is bound to facilitate the MCMC moves (provided a vague proper prior is selected). As in Robert *et al.* (2000), we are interested in estimating the number of hidden states, k. The prior modelling on the parameters is an $\mathcal{E}xp(1)$ distribution on the ω_{ij}'s, a normal $\mathcal{N}(0, 9\sigma_i^2)$ distribution on the μ_i's and an $\mathcal{E}xp(1)$ distribution on the σ_i's.

In Robert *et al.* (2000), the model under consideration consisted of

$$\mathcal{N}(0, \sigma^2_{z_n})$$

for the distribution of y_n conditional on z_n, i.e. did not involve an unknown mean parameter. For this model, we use the same prior, namely a uniform $\mathcal{U}(0, \alpha)$ prior on the σ_i's and an $\mathcal{E}xp(5 \max |y_n|)$ prior on the hyperparameter $1/\alpha$. (Robert *et al.*, 2000, noticed that the factor 5 in the exponential distribution was of little influence on the results.) Note that we do not impose identifiability constraints at the simulation level by ordering the variances, contrary to Robert *et al.* (2000).

A major difference with the above papers is that, as in Stephens (2000a), we will not use completion to run our algorithm. That is to say, the latent Markov chain $\{z_n\}$ is not to be simulated

by the algorithm. This can be avoided thanks to both the forward recursive representation of the likelihood for a hidden Markov model (Baum and Petrie, 1966), already used in Robert *et al.* (1999), and the random walk proposals as in Hurn *et al.* (2001). We believe that this choice is bound to accelerate convergence of the algorithm by a drastic reduction of the dimensionality of the space.

6.2. The moves of the continuous time MCMC algorithm

Since reversible technology was implemented for this model in Robert *et al.* (2000), we now focus on the continuous time MCMC counterpart, extending Stephens (2000a) and Hurn *et al.* (2001) to this framework. In addition to birth-and-death moves, which were enough to provide good mixing in the above papers, we do need to introduce additional proposals, similar to those in Richardson and Green (1997) and Robert *et al.* (2000), 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/combine moves, following the denomination of Richardson and Green (1997), where a given component is broken into two parts, and fixed k moves, where the parameters are modified via a regular MCMC step. The later proposals are quintessential in ensuring good convergence properties.

The birth-and-death and fixed k moves are simple to implement, and are equivalent to those given in Stephens (2000a) and Hurn *et al.* (2001), with fixed k moves relying on random walk proposals over the transforms $\log(\omega_i)$ and $\log(\sigma_i)$—or $\log(\sigma_i/\alpha - \sigma_i)$ in the constrained case of Robert *et al.* (2000). The split/combine move follows the general framework exposed in Section 4 with a combine intensity given by (5). We use η^S as an individual split intensity which is the same for all components. This means that the overall intensity 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 intensities, respectively.

There are many ways of devising a split/combine move but, contrary to Richardson and Green's (1997) observation that their first attempt was successful, we had to try several proposals before obtaining proper mixing behaviour, as detailed now.

In the case of a normal hidden Markov model with means μ_i and variances σ_i^2 all unknown, 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 $j \neq i_0$ of ω_{j,i_0} as

$$\tilde{\omega}_{j,i_1} = \omega_{j,i_0}\varepsilon_j, \quad \tilde{\omega}_{j,i_2} = \omega_{j,i_0}(1-\varepsilon_j),$$

with ε_j uniform on $(0,1)$; this proposal is sensible when thinking that both the 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 $i \neq i_0$ of $\omega_{i_0,i}$ as

$$\tilde{\omega}_{i_1,i} = \omega_{i_0,i}\xi_j, \quad \tilde{\omega}_{i_2,i} = \omega_{i_0,i}/\xi_j$$

where ξ_j is lognormal $\mathcal{LN}(0,1)$. The symmetry constraint (4) is thus satisfied. Note that we first tried this move with a half-Cauchy $\mathcal{C}^+(0,1)$ proposal, which also preserves the distribution by inversion (that is, ξ_j and $1/\xi_j$ have the same distribution), but this led to very poor mixing properties for the algorithm;

(c) a split move for ω_{i_0,i_0} as

$$\begin{aligned} \tilde{\omega}_{i_1,i_1} &= \omega_{i_0,i_0}\varepsilon_{i_0}\xi_{i_1}, & \tilde{\omega}_{i_1,i_2} &= \omega_{i_0,i_0}(1-\varepsilon_{i_0})\xi_{i_2}, \\ \tilde{\omega}_{i_2,i_1} &= \omega_{i_0,i_0}\varepsilon_{i_0}/\xi_{i_1}, & \tilde{\omega}_{i_2,i_2} &= \omega_{i_0,i_0}(1-\varepsilon_{i_0})/\xi_{i_2} \end{aligned}$$

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 on $(\mu_{i_0}, \sigma_{i_0}^2)$ as

$$\tilde{\mu}_{i_1} = \mu_{i_0} + 3\sigma_{i_0}\varepsilon_\mu, \ \tilde{\mu}_{i_2} = \mu_{i_0} - 3\sigma_{i_0}\varepsilon_\mu, \ \tilde{\sigma}_{i_1}^2 = \sigma_{i_0}^2\varepsilon_\sigma, \ \tilde{\sigma}_{i_2}^2 = \sigma_{i_0}^2/\varepsilon_\sigma,$$

where $\varepsilon_\mu \sim \mathcal{N}(0,1)$ and $\varepsilon_\sigma \sim \mathcal{LN}(0,1)$.

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 exponential 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 mean μ_{i_0} is obtained as the arithmetic average of the means $\tilde{\mu}_{i_1}$ and $\tilde{\mu}_{i_2}$, while the variance $\sigma^2_{i_0}$ is the geometric average of the variances $\tilde{\sigma}^2_{i_1}$ and $\tilde{\sigma}^2_{i_2}$. Appendix B details the computation of the corresponding Jacobian.

6.3. Illustrations

First, we consider a simulated dataset of 500 observations, represented on Figure 1(a); this dataset was built by joining stretches of three different normal samples that can be spotted directly on the graph. The most visited value (and posterior mode) of k is 3, as shown in Figure 1(b), with regular visits to 2 and 4. Larger values were hardly visited (although we used a flat prior on $k \in \{1, \ldots, 10\}$). As shown by Figure 1(d), the correspondence between the estimated density, obtained by averaging all the density estimates over the iterations, and the standard nonparametric kernel estimate, is quite satisfactory. Note in addition that the parameter chains, separated component by component, produce a label-switching behaviour that is to be expected from the theory (see Hurn *et al.*, 2001), as well as good mixing properties. (The graphs represented in Figure 1 actually correspond to $50,000$ iterations of the MCMC algorithm, with an average of 25 moves per observation unit.)

Our second dataset corresponds to a transform of the IBM stock over a period of five years, starting in 1992, which represents the volatility of the stock (kindly provided to us by Catalin Starica, Université Libre de Bruxelles). As can be seen from the rawplot of the dataset in Figure 2(a), the states are less clearly identified and, more importantly, there seems to be fewer moves between these states. The resulting inference corroborates this uncertainty: the four values $k = 1, 2, 3, 4$ have similar posterior probabilities and, in opposition to Figure 1(b), the spread of the loglikelihood values is much larger, suggesting that the posterior distribution has several modes that can only be linked by visiting intermediate low likelihood regions. Since the case $k = 3$ is visited less often, the number of simulations in Figure 2(c) is lower than the number of simulations in Figure 1(c), but also exhibits the correct label-switching behaviour and proper mixing features (even though one can spot longer regions when the chain remains invariant). Note also that the fit in Figure 2(d) is just as satisfactory as the nonparametric estimation.

For a comparison with Robert *et al.* (2000), we also consider one dataset studied in this previous paper, namely the wind intensity in Athens (kindly provided to us by Christian Francq, Université du Littoral). As discussed at the beginning of Section 6, the modelling setting slightly differs from the above: the means are now all set to 0 and the prior distribution on the σ's is not an exponential distribution but rather a uniform $\mathcal{U}(0, \alpha)$. Here α is an hyperparameter that is estimated from the dataset in a hierarchical way and updated through a slice sampler (since its full conditional distribution is a truncated gamma—see Robert *et al.* (2000) for details) via an additional process with intensity η^α, 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 $\log((\alpha - \sigma)/\sigma)$ and its inverse. The corresponding modified Jacobian is given in Appendix B.

Figure 3 summarises the output for the dataset corresponding to the wind intensity in Athens. The main point is that, 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 $.0064, .1848, .7584, .0488$, to be compared with Table 1 in Robert *et al.* (2000). Note that Figure 3(b) provides in addition the distribution of the number of moves per unit of time (on the continuous time axis). The loglikelihoods are actually covering a wider ranger 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 $k = 4$. The fit between the nonparametric density and the Bayesian posterior average is quite accurate.

7. Discussion

Considering Theorem 1, one may be tempted to say 'everything that may be done in continuous time can be done in discrete time'. While that might be true from a theoretical point of view,

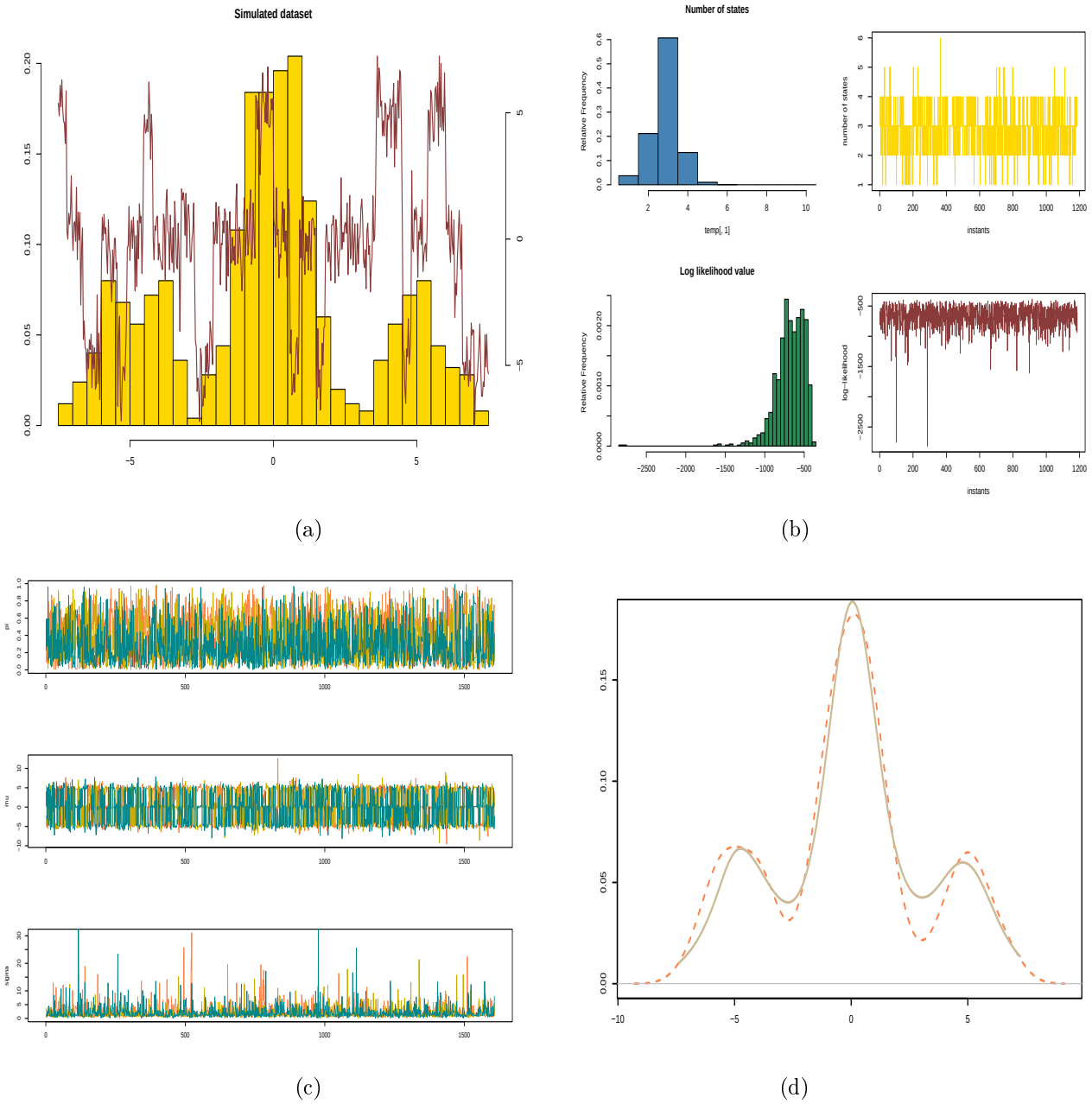

Fig. 1. Continuous time MCMC algorithm output for a simulated dataset of 500 points: (a) histogram and rawplot of the dataset; (b) MCMC output on k (histogram and rawplot), number of states and corresponding likelihood values; (c) MCMC sequence of the parameters of the three components when conditioning on $k = 3$; (d) MCMC evaluation of the marginal density, compared with R nonparametric density estimate.

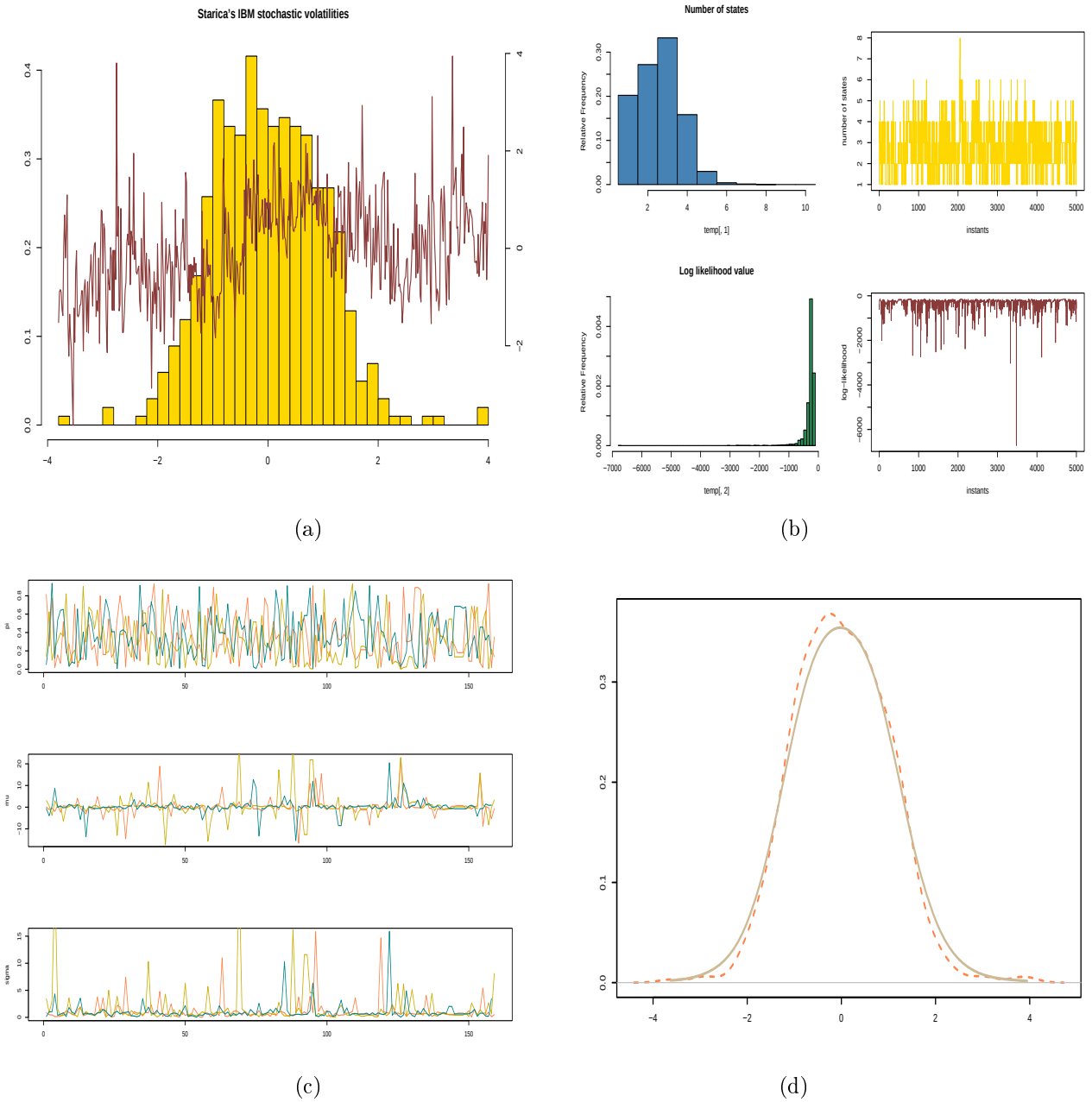

Fig. 2. Continuous time MCMC algorithm output for a transform of 507 IBM stockprices: (a) histogram and rawplot of the dataset; (b) MCMC output on k (histogram and rawplot), number of states, and corresponding likelihood values; (c) MCMC sequence of the parameters of the three components when conditioning on $k = 3$; (d) MCMC evaluation of the marginal density, compared with R nonparametric density estimate.

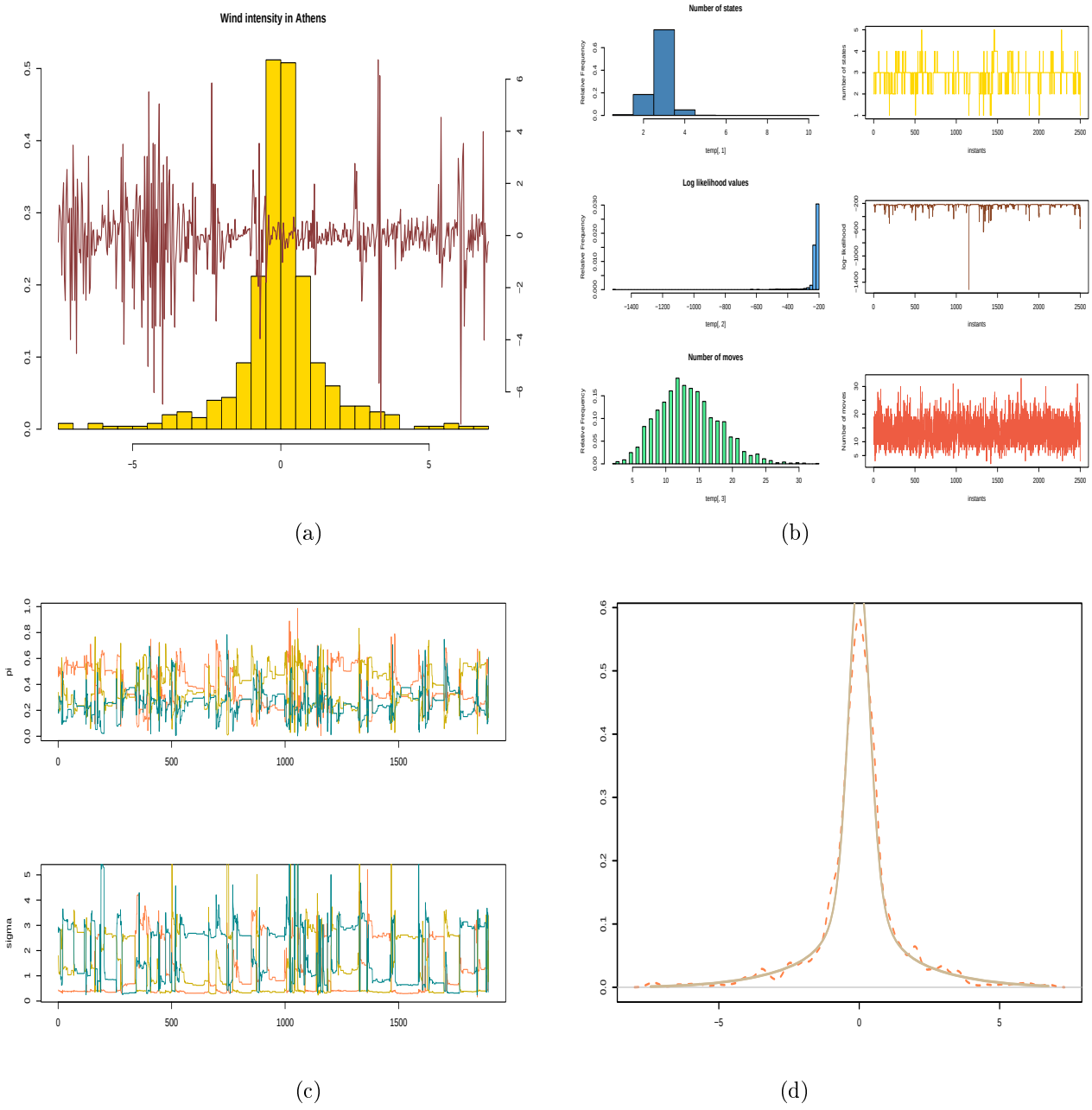

Fig. 3. Continuous time MCMC algorithm output for a sequence of 500 wind intensities in Athens; (a) histogram and rawplot of the dataset; (b) MCMC output on k (histogram and rawplot), number of states, and corresponding likelihood values; (c) MCMC sequence of the parameters of the three components when conditioning on $k = 3$; (d) MCMC evaluation of the marginal density, compared with R nonparametric density estimate.

things are less clearcut when performance considerations are taken into account.

Stephens (2000a) made some comparisons of his algorithm to Richardson and Green's (1997) reversible jump MCMC sampler, which we cite:

A. *Our algorithm works in continuous time, replacing the accept-reject scheme by allowing events to occur at differing rates.*

B. *Our dimension-changing birth and death moves do not make use of the missing data* $\mathbf{z}$*, effectively integrating out over them when calculating the likelihood.*

C. *Our birth and death moves take advantage of the natural nested structure of the models, removing the need for the calculation of a complicated Jacobian, and making implementation more straightforward.*

D. *Our birth and death moves treat the parameters as a point process, and do not make use of any constraint such as* $\mu_1 < \cdots < \mu_k$ *[used by Richardson and Green (1997) in defining their split and combine moves].*

We disagree with point C. since any Jacobian involved does appear in both continuous and discrete time. As we have seen, the Jacobian determinant $(1-w)^{k-1}$ due to renormalising component weights appears in both the death rates (2) and the acceptance ratio (3). Indeed, Stephens (2000a, p. 71) attributes this determinant to a 'simple change of variable formula'. In our view, the determinant should be associated with the proposal 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. (In fact, as noted above, there is no density w.r.t. a fixed reference measure on $\Theta^{(k+1)}$.) We also saw in Section 4 that the Jacobian determinant of the split and combine move does appear in continuous time. The complexity is therefore identical for both methodologies.

Regarding D. above, as noted in Section 2, we find the ordering of the components more a technical device than a practical one. Indeed, a split move usually makes the new set of components unordered but they can be sorted again. Nonetheless, we did not impose ordering when simulating the parameters with fixed k and, more importantly, did not restrict ourselves to implement combine moves only on adjacent components as in Richardson and Green (1997).

Hence, the above item that we find most important is B.; whether the missing data $\mathbf{z}$ is kept track of in all moves or not. It would indeed be interesting to compare the performance of two algorithms, in discrete or continuous time, that are identical except for this aspect. (We recall that Robert *et al.* (2000) did resort to completion in their implementation of RJMCMC.)

We now proceed to discussing computational aspects of discrete and continuous time algorithms. In continuous time, once a state $\boldsymbol{\theta}$ is entered, it is necessary to compute the rates of all possible moves leading to an exit from that state, at the expense of $\mathrm{O}(k)$ for birth/death moves and $\mathrm{O}(k^2)$ for split/combines ones. In discrete time this not necessary, as the acceptance ratio of a move is not computed until the move is proposed. This is an advantage of reversible jump MCMC. On the other hand, for moves such as birth and split in continuous time, rates are typically very simple and it is only the death or combine rates that are expensive to compute. This is an advantage of continuous time algorithms.

What can we say about the mixing performance of the different algorithms? A typical set-up of BDMCMC is to let $\beta(\boldsymbol{\theta})$ be constant, say $\beta(\boldsymbol{\theta}) = 1$ (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 Eqs. (2) and (3) 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.

Finally we again mention 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 as opposed to continuous time, the expected holding times cannot be computed easily; see (6) in the proof of Lemma 1 below.

References

Baum, L.E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37**, 1554–1563.
Breiman, L. (1992) *Probability.* SIAM, Philadelphia.
Brooks, S. and Giudici, P. (1999) Diagnosing convergence of reversible jump MCMC algorithms. In *Bayesian Statistics VI*, Bernardo, J., Berger, J., Dawid, A.P. and Smith, A.F.M. (Eds.), 733–742. Oxford University Press, Oxford.
Cappé, O. and Robert, C.P. (2000) MCMC: Ten years and still running! *J. Amer. Statist. Assoc.* **95**, 1282–1286.
Celeux, G., Hurn, M. and Robert, C.P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–979.
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. Roy. Statist. Soc. Ser. B* **56**, 549–603.
Hurn, M., Justel, A. and Robert, C.P. (2001) Estimating mixtures of regressions *J. Comput. Graphical Statist.* (to appear).
Karr, A.F. (1975) Weak convergence of a sequence of Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **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.T. and Spiegelhalter, D.J. (Eds.), 215–240. Chapman and Hall, London.
Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.
Richardson, S. and Green, P.J. (1998) Corrigendum: "On Bayesian analysis of mixtures with an unknown number of components". *J. Roy. Statist. Soc. Ser. B* **60**, 661.
Robert, C.P., Rydén, T. and Titterington, D.M. (1999) Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *J. Statist. Comput. Simulation* **64**, 327–355.
Robert, C.P., Rydén, T. and Titterington, D.M. (2000) Jump Markov chain Monte Carlo algorithms for Bayesian inference in hidden Markov models *J. Roy. Statist. Soc. Ser. B* **62**, 57–75.
Ripley, B. (1987) *Stochastic Simulation.* Wiley, New York.
Stephens, M. (2000a) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28**, 40–74.
Stephens, M. (2000b) Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B* **62**, 795–809.

A. Proof of Theorem 1

Let for $\boldsymbol{\theta} \in \Theta^{(k)}$,

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

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

$$\lambda_N(\boldsymbol{\theta}) = \int b_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)), 1\} h(\boldsymbol{\theta}; (w, \phi))\, d(w, \phi)$$

$$+ \sum_{i=1}^{k} d_N(\boldsymbol{\theta}) \frac{1}{k} \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i, \phi_i); \boldsymbol{\theta}), 1\}. \tag{6}$$

Thus

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in G} |N\lambda_N(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| \\
&\leq \int \sup_{\boldsymbol{\theta} \in G} |Nb_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)), 1\} h(\boldsymbol{\theta}; (w, \phi)) - \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w, \phi))| \, d(w, \phi) \qquad (7) \\
&+ \sum_{i=1}^{k} \sup_{\boldsymbol{\theta} \in G} |\frac{1}{k} N d_N(\boldsymbol{\theta}) \min\{A_N^{-1}(\boldsymbol{\theta} \setminus (w_i, \phi_i); \boldsymbol{\theta}), 1\} - \delta(\boldsymbol{\theta} \setminus (w_i, \phi_i); (w_i, \phi_i))|. \qquad (8)
\end{aligned}$$

We start by looking at the 'birth part' (7) 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

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in G} |Nb_N(\boldsymbol{\theta}) \min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)), 1\} h(\boldsymbol{\theta}; (w, \phi)) - \beta(\boldsymbol{\theta}) h(\boldsymbol{\theta}; (w, \phi))| \\
&\qquad \leq \sup_{\boldsymbol{\theta} \in G} |Nb_N(\boldsymbol{\theta}) - \beta(\boldsymbol{\theta})| \times 1 \times \sup_{\boldsymbol{\theta} \in G} h(\boldsymbol{\theta}; (w, \phi)) \qquad (9) \\
&\qquad + \sup_{\boldsymbol{\theta} \in G} \beta(\boldsymbol{\theta}) \times \sup_{\boldsymbol{\theta} \in G} |\min\{A_N(\boldsymbol{\theta}; \boldsymbol{\theta} \cup (w, \phi)), 1\} h(\boldsymbol{\theta}; (w, \phi)) - h(\boldsymbol{\theta}; (w, \phi))|. \qquad (10)
\end{aligned}$$

For $\beta \leq 0$ 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

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

Hence, for sufficiently large N (9) is bounded by

$$\frac{1}{2N} \sup_{\boldsymbol{\theta} \in G} \beta^2(\boldsymbol{\theta}) \times \sup_{\boldsymbol{\theta} \in G} h(\boldsymbol{\theta}; (w, \phi)); \tag{11}$$

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 (10), it is dominated by an integrable function similar to (11) (remove $1/2N$ and the square), and it remains to show that it tends to zero as $N \to \infty$. We have

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

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 (10) tends to zero uniformly over G as $N \to \infty$ if G is small enough.

We now turn to the 'death part' (8). By arguments similar to those above, for large N and sufficiently small G it holds that

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

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

We start by showing (ii). By the structure of Θ, it it sufficient to show that for each $\boldsymbol{\theta}' \in \Theta^{(k)}$, there is a neighbourhood $G \subseteq \Theta^{(k)}$ of $\boldsymbol{\theta}'$ such that $K_N g(\boldsymbol{\theta}) \to Kg(\boldsymbol{\theta})$ 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\qquad h(\boldsymbol{\theta}; (w, \phi)) g(\boldsymbol{\theta} \cup (w, \phi), \frac{m}{N})\, 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(\boldsymbol{\theta} \setminus (w_i, \phi_i), \frac{m}{N}) \\
= & \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\qquad h(\boldsymbol{\theta}; (w, \phi)) g(\boldsymbol{\theta} \cup (w, \phi), \frac{\lceil Nu \rceil}{N})\, du\, d(w, \phi) \\
& + \int_0^{\infty} (1 - \lambda_N(\boldsymbol{\theta}))^{\lfloor Nu \rfloor} \\
& \qquad \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(\boldsymbol{\theta} \setminus (w_i, \phi_i), \frac{\lceil Nu \rceil}{N})\, du; \\
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

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

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

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

where $\overline{h}(w,\phi) = \sup_{\boldsymbol{\theta}\in G} h(\boldsymbol{\theta}; (w,\phi))$ and $\delta^g_{1/N} = \sup_{\Delta((\boldsymbol{\theta},u),(\boldsymbol{\theta}',u')) \leq 1/N} |g(\boldsymbol{\theta},u) - g(\boldsymbol{\theta}',u')|$ is g's modulus of continuity; 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 $\lambda(\boldsymbol{\theta}) \geq \beta(\boldsymbol{\theta})$ is bounded away from zero on compact subsets of Θ. Moreover, since

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

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

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

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 $\log(1-x) \geq -x - 2x^2$ for $0 \leq x \leq 1/2$ leads to a reverse inequality which is handled similarly.

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

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

B. The Jacobian for the split-combine move

The parts of the Jacobian determinant corresponding to the split move in §6.2 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,

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

$$= 4\,\omega_{i_0,i_0}^3\,\frac{\varepsilon_{i_0}(1-\varepsilon_{i_0})}{\xi_{i_1}\xi_{i_2}};$$

(d)

$$\begin{vmatrix} 1 & 3\varepsilon_\mu/2\sigma_{i_0} & 3\sigma_{i_0} & 0 \\ 1 & -3\varepsilon_\mu/2\sigma_{i_0} & -3\sigma_{i_0} & 0 \\ 0 & \varepsilon_\sigma & 0 & \sigma_{i_0}^2 \\ 0 & 1/\varepsilon_\sigma & 0 & -\sigma_{i_0}^2/\varepsilon_\sigma^2 \end{vmatrix} = 12\sigma_{i_0}^3/\varepsilon_\sigma,$$

given that we differentiate w.r.t. $\sigma_{i_0}^2$, not σ_{i_0}.

The overall Jacobian determinant for the split move is therefore

$$\left|\frac{\partial T(\theta,\varepsilon)}{\partial(\theta,\varepsilon)}\right| = 3\prod_{i\neq i_0}\frac{\omega_{i,i_0}\omega_{i_0,i}}{\xi_i}\omega_{i_0,i_0}^3\,\frac{\varepsilon_{i_0}(1-\varepsilon_{i_0})}{\xi_{i_1}\xi_{i_2}\varepsilon_\sigma}\sigma_{i_0}^3\,2^{k+3}\,.$$

In the case considered in Robert *et al.* (2000), that is when the means μ_i are set to zero and the variances σ_i^2 are constrained to be less than α^2, part (d) of the Jacobian can be obtained as

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

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