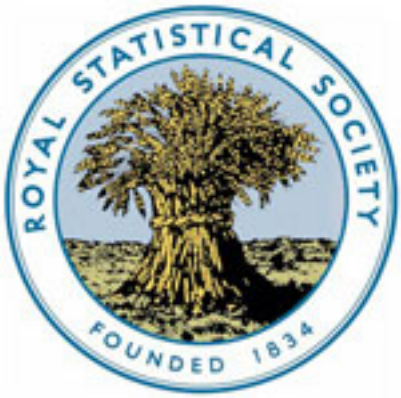

J. R. Statist. Soc. B (2000)
62, *Part* 1, *pp.* 57–75

Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method

Christian P. Robert,

Centre de Recherche en Economie et Statistique, Paris, France

Tobias Rydén

Lund University, Sweden

and D. M. Titterington

University of Glasgow, UK

[Received May 1998. Revised March 1999]

Summary. Hidden Markov models form an extension of mixture models which provides a flexible class of models exhibiting dependence and a possibly large degree of variability. We show how reversible jump Markov chain Monte Carlo techniques can be used to estimate the parameters as well as the number of components of a hidden Markov model in a Bayesian framework. We employ a mixture of zero-mean normal distributions as our main example and apply this model to three sets of data from finance, meteorology and geomagnetism.

Keywords: Bayesian inference; Hidden Markov model; Markov chain Monte Carlo methods; Model selection

1. Introduction

Hidden Markov models (HMMs) have been used in many areas as convenient representations of weakly dependent heterogeneous phenomena; a few examples are econometrics (Hamilton, 1989; Chib, 1996; Krolzig, 1997; Billio *et al.*, 1999), finance (Rydén *et al.*, 1998), biology (Leroux and Puterman, 1992), genetics (Churchill, 1995), neurophysiology (Fredkin and Rice, 1992) and speech processing (Rabiner, 1989). We also refer the reader to MacDonald and Zucchini (1997).

In HMMs, the heterogeneity in the data is represented by a mixture structure, i.e. a pair (z_t, y_t) with $z_t \in \{1, \ldots, k\}$ and $y_t|z_t \sim f_{z_t}(y_t)$. The y_t are assumed to be independent conditional on the z_t. The Markov nomenclature comes from assuming that $\{z_t\}$ is distributed as a (finite state) Markov chain and the hidden part is due to the fact that $\{z_t\}$ is not observed. Usually, the conditional distributions f_i belong to a single parametric family, such as the normal or Poisson family, so that z_t corresponds to the parameter used to generate y_t.

Inference for HMMs was first considered by Baum and Petrie (1966), who treated the case when y_t takes values in a finite set. Petrie (1969) provided identifiability conditions for HMMs whereas Baum *et al.* (1970) considered maximum likelihood further, implementing an

Address for correspondence: Tobias Rydén, Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden.
E-mail: tobias@maths.lth.se

 1369–7412/00/62057

early version of the EM algorithm as well as deriving a general proof of monotonicity for the algorithm. They also produced a procedure which allows for the derivation of the conditional joint distribution of the z_t given the y_t and which has been used ever since (see Appendix A). Later, Leroux (1992) proved consistency of the maximum likelihood estimator (MLE) for general HMMs under mild conditions whereas asymptotic normality has been proved by Bickel *et al.* (1998).

These references apply to the estimation of HMM parameters when the number of states of $\{z_t\}$, k say, is known. The present paper is concerned with inference when k is also unknown. The approach adopted here is to use Bayesian inference and to implement it by the reversible jump Markov chain Monte Carlo (MCMC) techniques first proposed by Green (1995).

Non-Bayesian approaches to the problem include likelihood ratio tests and penalized likelihood methods. With the former method, we first postulate H_0: $k = 1$ as our null hypothesis and H_1: $k = 2$ as the alternative, compute the likelihood ratio and reject H_0 if this statistic is sufficiently large. Standard theory would say that, if H_0 is true, the likelihood ratio is asymptotically distributed as a χ^2 random variable. This result is not true for HMMs or mixtures, however, essentially because, if H_0 is true, then the parameters are not uniquely identified under H_1. For HMMs, the limiting distribution of the likelihood ratio cannot be computed numerically but must be approximated by simulation. McLachlan (1987), in the context of mixtures, proposed to approximate it by the parametric bootstrap, and Rydén *et al.* (1998) indeed applied this idea to HMMs. The approach requires enormous computational effort, however; it is time consuming to compute even one likelihood ratio statistic, and usually 50 or 100 bootstrap replications are simulated. Rydén *et al.* (1998) could not take the approach further than the case of testing $k = 2$ against $k = 3$ because of the computational difficulties. Penalized likelihood methods such as the Akaike and Bayesian information criteria AIC and BIC are less demanding since simulated replications are not involved. However, they produce no number that quantifies the confidence in the result, such as a p-value. If this is desired, we are again confined to the parametric bootstrap.

Bayesian inference through reversible jump MCMC methods is a viable alternative to frequentist analysis that both explores models with different values of k and provides probabilities representing confidence in different models. A most appealing feature is that this approach bypasses the traditional model choice structure, which requires prior modelling and testing for the various models. This alternative is based on standard hierarchical modelling and estimates k as well as the posterior probabilities of the various models. Richardson and Green (1997) developed similar methods for mixture distribution analysis, and our work is in many respects based on, and a development of, their achievements. Our work also continues that of Robert *et al.* (1993), Chib (1996) and Robert and Titterington (1998) on Bayesian estimation for HMMs with a fixed number of components.

As in the case of mixtures, HMMs with an unknown number of components are appealing in two settings: first, a heterogeneous population with an unknown degree of heterogeneity k may come under scrutiny and the value of k matters for the subsequent analysis; secondly, the density of a dependent sample may be estimated and the HMM structure provides a parsimonious alternative to standard nonparametric estimates.

The paper is organized as follows: Section 2 details the prior modelling, Section 3 details the steps in the reversible jump MCMC algorithm, since the Markovian structure together with the reversibility constraint lead to a higher level of complexity than in Richardson and Green (1997), and Section 4 examines the performance of this method on several data sets. The data sets can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. The Bayesian model

We choose to study mixtures of zero-mean normal distributions, partly because we want to analyse data to which this model was previously fitted by using likelihood techniques (see Section 4), and partly to provide some variety over the structure considered by Richardson and Green (1997). A wide range of other mixtures can be dealt with along similar lines. For a given k, the marginal distribution of an observation y_t will thus be

$$\sum_{i=1}^{k} \pi_i \mathcal{N}(0, \sigma_i^2),$$

where the π_i are the components of the stationary vector of the transition matrix of the hidden states, $A = (a_{ij})$, such that $P(z_{t+1} = j | z_t = i) = a_{ij}$. The set of σ_i is denoted by σ.

Along lines similar to Section 2.2 of Richardson and Green (1997), we assume that the joint density of all variables mentioned so far takes the form

$$p(k, A, z, \sigma, y) = p(k)\, p(A|k)\, p(z|A, k)\, p(\sigma|k)\, p(y|\sigma, z).$$

In particular, k is *a priori* uniform on $\{1, 2, \ldots, k_{\max}\}$, where $k_{\max}$ is specified, given k the rows of A are independent, each with a Dirichlet distribution $D(\delta, \delta, \ldots, \delta)$, and given k the σ_i are uniform on $(0, \alpha)$ but sorted in ascending order, so that the parameters and components are identifiable. In addition, we assume that α has a negative exponential distribution with mean $1/\xi$ where $1/\xi = 30 \max |y_t|$. The reason for putting a hyperprior on α is to make the posterior robust; it is not robust if α is fixed but it is with the current model, as we expand on further below. However, we chose a particular value, $\delta = 1$, for δ. Thus, the complete joint density, corresponding to expression (6) in Richardson and Green (1997), is

$$p(\alpha, k, A, z, \sigma, y) = p(\alpha)\, p(k)\, p(A|k, \delta)\, p(z|A, k)\, p(\sigma|k, \alpha)\, p(y|\sigma, z). \tag{1}$$

3. Markov chain Monte Carlo methodology

The general theory and methodology of MCMC procedures is now fairly standard; see for instance Tierney (1994) and Gilks *et al.* (1996). For the particular approach that is used here, we are clearly especially indebted to the formative work on reversible jump MCMC methods in Green (1995) and Richardson and Green (1997); for further background references, Richardson and Green (1997) and the discussion on it are particularly valuable. We simply recall here that the reversible jump techniques were proposed by Green (1995) to overcome the measure theoretic difficulties in implementing standard MCMC algorithms in problems with parameters of varying dimension. The fundamental idea behind Green's (1995) technique is to impose a one-to-one transition in steps when the dimension of the space varies, so that the dominating measure can be chosen to be positive. In a mixture set-up, the moves where the dimension changes can be restricted to so-called *split* and *combine* moves, where a component is broken into two components or, conversely, when two components are reunited into a single component.

Inferences are based on a sample $y_1, \ldots, y_n$, i.e. the sample size is n. Our objective is to generate realizations from the conditional joint density $p(\alpha, k, A, z, \sigma | y)$ defined through equation (1), and the detailed description of one sweep of our MCMC procedure is as follows:

(a) update transition probability matrix A;
(b) update the standard deviations σ;
(c) update the allocations z;

(d) update the hyperparameter α;
(e) consider splitting a component into two or combining two into one;
(f) consider the birth or death of an empty component.

In (f), an empty component is a component to which no observation is allocated, i.e. component i is empty if there is no t with $z_t = i$.

3.1. Gibbs moves

Move types (a)–(d) are Gibbs moves and (a) and (c) follow Robert *et al.* (1993). In (a), the ith row of A is sampled from a Dirichlet distribution $D(\delta + n_{i1}, \ldots, \delta + n_{ik})$, where

$$n_{ij} = \sum_{t=1}^{n-1} I\{z_t = i, z_{t+1} = j\}$$

is the number of jumps from component i to component j; $I\{\cdot\}$ denotes an indicator function. In (b), each σ_i^{-2} is sampled from a truncated gamma distribution with density proportional to

$$u^{(n_i-3)/2} \exp\{-(s_i^2/2)u\}\, I\{u \geqslant 1/\alpha^2\}, \tag{2}$$

where $n_i = \Sigma_{t=1}^n I\{z_t = i\}$ and $s_i^2 = \Sigma_{t=1}^n I\{z_t = i\} y_t^2$ are respectively the number of observations allocated to component i and the corresponding sum of squares. Sampling from expression (2) was carried out by simply rejecting gamma random variates until the condition $u \geqslant 1/\alpha^2$ was fulfilled, except when $n_i = 0$ or $n_i = 1$; more sophisticated approaches include the slice sampler of Damien *et al.* (1999). After sampling each σ_i, the new set of σ_i was accepted if they were ascending; otherwise the update was rejected. In (c), $z_1, \ldots, z_n$ were resampled one at a time from $t = 1$ to $t = n$, with conditional probabilities given by

$$P(z_t = i | \ldots) \propto a_{z_{t-1} i}\, \varphi(y_t; \sigma_i) a_{i z_{t+1}}, \tag{3}$$

where $\varphi(\cdot; \sigma)$ is the density of a normal random variable with mean 0 and standard deviation σ; for $t = 1$, the first factor is replaced by the stationary probability π_i, and, for $t = n$, the last factor is replaced by 1. In (d), α was sampled from a distribution with density proportional to

$$u^{-k} \exp(-\xi u)\, I\{u \geqslant \max_i(\sigma_i)\}. \tag{4}$$

Sampling from this density was carried out by using the method of Damien *et al.* (1999).

3.2. Split and combine moves

Moves (e) and (f) are more involved Metropolis–Hastings steps and allow for increasing or decreasing the number of components by 1. In (e), we choose to split with probability b_k and to combine with probability $d_k = 1 - b_k$. Naturally, $d_1 = b_{k_{\max}} = 0$, and we used $b_k = d_k = \frac{1}{2}$ for $k = 2, 3, \ldots, k_{\max} - 1$. To describe the combine move, suppose that the current state of the MCMC algorithm is $\tilde{x}$, with parameters $\tilde{a}_{ij}$ etc., and that this representation has $k + 1$ components. We randomly select a pair (j_1, j_2) of adjacent components and try to combine them into a single new component j_*, thus creating a new MCMC state x with k components. This is done in several steps. First, for any t with $\tilde{z}_t$ equal to j_1 or j_2, z_t is set to j_* whereas remaining $\tilde{z}_t$ are simply copied. Secondly, the standard deviation of the new state is given by

$$\sigma_{j_*}^2 = \tilde{\pi}_{j_1} \tilde{\sigma}_{j_1}^2 + \tilde{\pi}_{j_2} \tilde{\sigma}_{j_2}^2, \tag{5}$$

whereas remaining $\tilde{\sigma}_i$ are copied. Here, $\tilde{\pi}_j$ is the stationary probability that the hidden chain $\tilde{z}$ is in state j, so the vector $\tilde{\pi}$ satisfies $\tilde{\pi}\tilde{A} = \tilde{\pi}$. Thirdly, the transition probabilities from and to the components involved in the move are set as

$$\begin{aligned} a_{j_* j} &= \frac{\tilde{\pi}_{j_1}}{\tilde{\pi}_{j_1} + \tilde{\pi}_{j_2}} \tilde{a}_{j_1 j} + \frac{\tilde{\pi}_{j_2}}{\tilde{\pi}_{j_1} + \tilde{\pi}_{j_2}} \tilde{a}_{j_2 j} && \text{for } j \neq j_*, \\ a_{i j_*} &= \tilde{a}_{i j_1} + \tilde{a}_{i j_2} && \text{for } i \neq j_*. \end{aligned} \tag{6}$$

Remaining a_{ij} are obtained by equating row sums to 1 or by copying. The expression for $a_{j_* j}$ in equations (6) is the conditional probability of jumping to component j, given that the current component is either j_1 or j_2. Also the new transition probability matrix A has essentially the same stationary probabilities as does $\tilde{A}$; one can readily verify that $\pi_j = \tilde{\pi}_j$ for $j \neq j_*$ and $\pi_{j_*} = \tilde{\pi}_{j_1} + \tilde{\pi}_{j_2}$. This in turn implies that the new HMM (A, σ) has the same second moment as does $(\tilde{A}, \tilde{\sigma})$: $\Sigma\, \pi_j \sigma_j^2 = \Sigma\, \tilde{\pi}_j \tilde{\sigma}_j^2$. Obviously, the first moments of the HMMs that we consider are identically 0.

In the split move, a component j_* is randomly chosen and split into two new components, j_1 and j_2. In describing this, we assume that the old representation is x, with k components, and that the new representation is $\tilde{x}$, with $k+1$ components. In designing the split move, our starting point was the following result. Its proof involves straightforward algebra only and is omitted.

Theorem 1. Let $0 < u_0, u_1 < 1$. Assume that the new transition probabilities are given by

$$\left.\begin{aligned} \tilde{a}_{j_1 j} = a_{j_* j}, \qquad \tilde{a}_{j_2 j} = a_{j_* j} && \text{for } j \neq j_1, j_2, \\ \tilde{a}_{i j_1} = u_0 a_{i j_*}, \qquad \tilde{a}_{i j_2} = (1 - u_0) a_{i j_*} && \text{for } i \neq j_1, j_2, \\ \tilde{a}_{j_1 j_1} = (1 - u_1) a_{j_* j_*}, \qquad \tilde{a}_{j_1 j_2} = u_1 a_{j_* j_*}, \end{aligned}\right\} \tag{7}$$

and that $\tilde{a}_{j_2 j_1}$ and $\tilde{a}_{j_2 j_2}$ are taken so that the new stationary probabilities are $\tilde{\pi}_j = \pi_j$ for $j \neq j_1, j_2$, $\tilde{\pi}_{j_1} = u_0 \pi_{j_*}$ and $\tilde{\pi}_{j_2} = (1 - u_0)\pi_{j_*}$; the remaining $\tilde{a}_{ij}$ are obtained by copying. It is assumed that the pair (u_0, u_1) is chosen so that all $\tilde{a}_{ij}$ are non-negative.

Then, if λ is an eigenvalue of A with right eigenvector h, λ is also an eigenvalue of $\tilde{A}$ with right eigenvector $\tilde{h}$ given by $\tilde{h}_j = h_j$ for $j \neq j_1$, j_2 and $\tilde{h}_{j_1} = \tilde{h}_{j_2} = h_{j_*}$. In addition, $(1 - u_1 - u_0)/(1 - u_0) \times a_{j_* j_*}$ is an eigenvalue of $\tilde{A}$ with right eigenvector $\tilde{h}$ given by $\tilde{h}_j = 0$ for $j \neq j_1, j_2$, $\tilde{h}_{j_1} = 1 - u_0$ and $\tilde{h}_{j_2} = -u_0$. Finally, it holds that $(A^m)_{ij} = (\tilde{A}^m)_{ij}$ for each $m > 0$ and $i, j \neq j_*, j_1, j_2$.

Theorem 1 shows that, by splitting the transition probabilities as in equations (7), the dynamics of the hidden Markov chain are essentially preserved; the difference is found in the dynamics within the new components j_1 and j_2. The variable u_0 determines how much of the stationary probability for j_* is assigned to the new component j_1, and u_1 determines the dynamics within the two new components.

A model (A, σ) with k components has $k(k-1) + k = k^2$ free parameters and a model $(\tilde{A}, \tilde{\sigma})$ with $k+1$ components similarly has $(k+1)^2$ free parameters; the difference in dimensionality is thus $2k+1$. This way of splitting component j_* into two involves only 2 degrees of freedom, u_0 and u_1, and thus cannot be directly used in a reversible jump MCMC algorithm. However, we can take it as a guideline for designing a split move with more degrees of freedom.

Our aim is to split j_* in such a way that stationary probabilities for the hidden chain are preserved as above, i.e. $\tilde{\pi}_j = \pi_j$ for $j \neq j_1, j_2$, $\tilde{\pi}_{j_1} = u_0 \pi_{j_*}$ and $\tilde{\pi}_{j_2} = (1 - u_0)\pi_{j_*}$. We can accomplish this as follows. Let $u_0 \sim \text{Be}(2, 2)$, $u_j \sim \text{Be}(r, s)$ for each $j \neq j_1, j_2$, and $v_i \sim \text{Be}(r, s)$

for each $i \neq j_1, j_2$. The shape parameters r and s are given below. Then set $\tilde{a}_{ij} = a_{ij}$ for $i, j \neq j_1$, j_2, set

$$\left.\begin{aligned} \tilde{a}_{j_1 j} &= \frac{u_j}{u_0} a_{j_* j}, \qquad \tilde{a}_{j_2 j} = \frac{1-u_j}{1-u_0} a_{j_* j} && \text{for } j \neq j_1, j_2, \\ \tilde{a}_{i j_1} &= v_i a_{i j_*}, \qquad \tilde{a}_{i j_2} = (1-v_i) a_{i j_*} && \text{for } i \neq j_1, j_2, \\ \tilde{a}_{j_1 j_2} &= u_1 \left(1 - \sum_{j \neq j_*} \frac{u_j}{u_0} a_{j_* j}\right), \\ \tilde{a}_{j_2 j_1} &= \left\{(1-u_1) \sum_{j \neq j_*} u_j a_{j_* j} + u_0 u_1 - \sum_{i \neq j_*} \kappa_i v_i a_{i j_*}\right\} \Big/ (1-u_0), \end{aligned}\right\} \tag{8}$$

where $\kappa_i = \pi_i / \pi_{j_*}$, and set $\tilde{a}_{j_1 j_1}$ and $\tilde{a}_{j_2 j_2}$ to make rows sum to 1. Since $u_0 \tilde{a}_{j_1 j} + (1-u_0)\tilde{a}_{j_2 j} = a_{j_* j}$ for $j \neq j_1, j_2$, the equilibrium equation $\tilde{\pi}_j = \Sigma_i \tilde{\pi}_i \tilde{a}_{ij}$ holds for $j \neq j_1, j_2$. The expression for $\tilde{a}_{j_2 j_1}$ ensures that it holds also for $j = j_2$. The shape parameters r and s are taken as

$$\begin{aligned} r &= \frac{1-u_0(1+c^2)}{c^2}, \qquad s = r\frac{1-u_0}{u_0} && \text{if } u_0 \leqslant \tfrac{1}{2}, \\ s &= \frac{1-(1-u_0)(1+c^2)}{c^2}, \qquad r = s\frac{u_0}{1-u_0} && \text{if } u_0 > \tfrac{1}{2}. \end{aligned} \tag{9}$$

This produces a beta distribution with mean u_0 and, if $u_0 \leqslant \frac{1}{2}$, squared coefficient of variation c^2; if $u_0 > \frac{1}{2}$ the squared coefficient of variation is not c^2 but rather the distribution is a mirror (around $x = \frac{1}{2}$) version of the distribution obtained for $1 - u_0 \leqslant \frac{1}{2}$. For our numerical results, we used $c^2 = 0.5$. Note that, given u_0, the conditional means of $\tilde{a}_{j_1 j}$ and $\tilde{a}_{j_2 j}, j \neq j_1, j_2$, are $a_{j_* j}$, and the conditional means of $\tilde{a}_{i j_1}$ and $\tilde{a}_{i j_2}, i \neq j_1, j_2$, are $u_0 a_{i j_*}$ and $(1-u_0) a_{i j_*}$ respectively. This is all consistent with theorem 1. We now discuss u_1. Given u_0, the u_j and the v_i, the valid range for u_1, i.e. the range in which the resulting $\tilde{A}$ is stochastic, is $[u_1^{\mathrm{L}}, u_1^{\mathrm{U}}]$, with

$$u_1^{\mathrm{L}} = \max\left(1 - \frac{1 - \sum_{i \neq j_1, j_2} \kappa_i / u_0 \times \tilde{a}_{i j_1}}{1 - \sum_{j \neq j_1, j_2} \tilde{a}_{j_1 j}}, 0\right),$$

$$u_1^{\mathrm{U}} = \min\left\{1 - \frac{1 - \sum_{i \neq j_1, j_2} \kappa_i / u_0 \times \tilde{a}_{i j_1} - (1-u_0)/u_0 \times \left(1 - \sum_{j \neq j_1, j_2} \tilde{a}_{j_2 j}\right)}{1 - \sum_{j \neq j_1, j_2} \tilde{a}_{j_1 j}}, 1\right\}.$$

It may happen that $u_1^{\mathrm{L}} > u_1^{\mathrm{U}}$, in which case there is no valid u_1 and the split move is rejected. If $u_1^{\mathrm{L}} < u_1^{\mathrm{U}}$, we draw u_1 as, written symbolically, $u_1 \sim u_1^{\mathrm{L}} + (u_1^{\mathrm{U}} - u_1^{\mathrm{L}})\,\mathrm{Be}(1, 1)$.

The new standard deviations are created as $\tilde{\sigma}_j = \sigma_j$ for $j \neq j_1, j_2$ and

$$\begin{aligned} \tilde{\sigma}_{j_1}^2 &= \sigma_{j_*}^2 \left\{1 - w\sqrt{\left(\frac{1-u_0}{u_0}\right)}\right\}, \\ \tilde{\sigma}_{j_2}^2 &= \sigma_{j_*}^2 \left\{1 + w\sqrt{\left(\frac{u_0}{1-u_0}\right)}\right\}. \end{aligned} \tag{10}$$

Here the range for w in which the $\tilde{\sigma}_i$ become properly sorted is $[0, w^{\mathrm{U}}]$, where

$$w^{\mathrm{U}} = \max\left\{ \frac{h_1}{\sigma^2_{j_*}} \sqrt{\left(\frac{u_0}{1-u_0} \right)},\ \frac{h_2}{\sigma^2_{j_*}} \sqrt{\left(\frac{1-u_0}{u_0} \right)} \right\}, \tag{11}$$

$h_1 = \sigma^2_{j_*} - \sigma^2_{j_*-1}$ if $j_* > 1$ and $h_1 = \sigma^2_{j_*}$ otherwise, and $h_2 = \sigma^2_{j_*+1} - \sigma^2_{j_*}$ if $j_* < k$ and $h_2 = \alpha^2 - \sigma^2_{j_*}$ otherwise. We draw w as $w \sim w^{\mathrm{U}}$ Be(1, 1). All the above beta variables are independent.

Finally we discuss how the hidden chain z is split, giving $\tilde{z}$. The z_t that are different from j_* are simply copied, but those equal to j_* should be relabelled as j_1 or j_2. Assume that $z_t = j_*$ for $t_1 \leqslant t \leqslant t_2$ with $z_{t_1-1} \neq j_*$ and $z_{t_2+1} \neq j_*$. We then sample $\tilde{z}_{t_1}, \ldots, \tilde{z}_{t_2}$ from the conditional distribution of this vector given the remaining $\tilde{z}_t$, the y_t and the proposed new $\tilde{a}_{ij}$ and $\tilde{\sigma}_i$. To do this, we employ a restricted backward algorithm described in Appendix A.

If we first split component j_* as in equations (8) and (10) and then combine the new components j_1 and j_2 again as in equations (5) and (6), we recover the original representation. This makes the split or combine pair reversible. Also, note that, from equations (5) and (6), we can compute corresponding values of u_0, u_1, the u_j, the v_i and w in the combine move.

The acceptance probabilities for the split and combine moves are min(1, R) and min(1, R^{-1}) respectively, where

$$R = \frac{p(y|\tilde{z}, \tilde{\sigma})}{p(y|z, \sigma)} \frac{p(k+1)}{p(k)} \frac{\prod_{i=1}^{k+1} \dfrac{\Gamma\{(k+1)\delta\}}{\Gamma(\delta)^{k+1}} \prod_{j=1}^{k+1} \tilde{a}_{ij}^{\delta-1}}{\prod_{i=1}^{k} \dfrac{\Gamma(k\delta)}{\Gamma(\delta)^{k}} \prod_{j=1}^{k} a_{ij}^{\delta-1}} \frac{p(\tilde{z}|\tilde{A})}{p(z|A)} \frac{(k+1)!}{k!} \frac{\prod_{i=1}^{k+1} \alpha^{-1} I\{\tilde{\sigma}_i \leqslant \alpha\}}{\prod_{i=1}^{k} \alpha^{-1} I\{\sigma_i \leqslant \alpha\}} \frac{d_{k+1}}{b_k P_{\mathrm{alloc}}}$$

$$\times \left\{ g_{2,2}(u_0) \prod_j g_{r,s}(u_j) \prod_i g_{r,s}(v_i) \frac{1}{u_1^{\mathrm{U}} - u_1^{\mathrm{L}}} g_{1,1}\left(\frac{u_1 - u_1^{\mathrm{L}}}{u_1^{\mathrm{U}} - u_1^{\mathrm{L}}} \right) \frac{1}{w^{\mathrm{U}}} g_{1,1}\left(\frac{w}{w^{\mathrm{U}}} \right) \right\}^{-1} J,$$

where $g_{r,s}$ is the Be(r, s) density and P_{alloc} is the probability of making the particular allocation of the $\tilde{z}_t$ that was made. J is the Jacobian determinant of the transformation given by equations (8) and (10). This determinant, which partly is evaluated numerically, is further discussed in Appendix B. The conditional densities involved in R are simple products: $p(y|z, \sigma) = \prod_{t=1}^{n} \varphi(y_t; \sigma_{z_t})$, $p(z|A) = \pi_{z_1} \prod_{t=1}^{n-1} a_{z_t z_{t+1}}$, etc. When $\delta = 1$, which is the value that we used for the numerical results below, the expression for R simplifies to

$$R = \frac{p(y|\tilde{z}, \tilde{\sigma})}{p(y|z, \sigma)} \frac{p(k+1)}{p(k)} k^k k! \frac{p(\tilde{z}|\tilde{A})}{p(z|A)} (k+1)\, \alpha^{-1} \frac{d_{k+1}}{b_k P_{\mathrm{alloc}}}$$

$$\times \left\{ g_{2,2}(u_0) \prod_j g_{r,s}(u_j) \prod_i g_{r,s}(v_i) \frac{1}{u_1^{\mathrm{U}} - u_1^{\mathrm{L}}} g_{1,1}\left(\frac{u_1 - u_1^{\mathrm{L}}}{u_1^{\mathrm{U}} - u_1^{\mathrm{L}}} \right) \frac{1}{w^{\mathrm{U}}} g_{1,1}\left(\frac{w}{w^{\mathrm{U}}} \right) \right\}^{-1} J. \tag{12}$$

3.3. *Birth and death moves*

We now proceed to the birth and death moves. In move type (f), we first choose at random between birth and death with probabilities b_k and d_k respectively. The death move is accomplished by selecting a component at random among the empty components and deleting it. The remaining rows of A are then renormalized and the z_t are unaltered.

In the birth move, we start with a model with k components and we want to create a new empty component j_*. To do this, we draw the j_*th row $a_{j_*\cdot} = u$ of the new transition probability matrix from the prior $D(\delta, \ldots, \delta)$. Furthermore, we draw v_i, $i \neq j_*$, from Be(1, k) and, when $i \neq j_*$, set

$$\begin{aligned} \tilde{a}_{ij} &= (1 - v_i)a_{ij} \qquad \text{for } j \neq j_*, \\ a_{ij_*} &= v_i. \end{aligned} \tag{13}$$

The new standard deviation $\tilde{\sigma}_{j_*}$ is drawn from the uniform prior $U(0, \alpha)$, the remaining $\tilde{\sigma}_i$ are copied and, again, $\tilde{z} = z$ because the new component is empty. As for the split or combine pair, if we first create a new component by using the birth move and then delete it through the death move, we recover the original state. The acceptance probabilities for the birth and death moves are $\min(1, R)$ and $\min(1, R^{-1})$ respectively, where

$$R = \frac{p(k+1)}{p(k)} \frac{\prod_{i=1}^{k+1} \frac{\Gamma\{(k+1)\delta\}}{\Gamma(\delta)^{k+1}} \prod_{j=1}^{k+1} \tilde{a}_{ij}^{\delta-1}}{\prod_{i=1}^{k} \frac{\Gamma(k\delta)}{\Gamma(\delta)^{k}} \prod_{j=1}^{k} a_{ij}^{\delta-1}} \frac{p(\tilde{z}|\tilde{A})}{p(z|A)} \frac{(k+1)!}{k!} \frac{\prod_{i=1}^{k+1} \alpha^{-1} I\{\tilde{\sigma}_i \leqslant \alpha\}}{\prod_{i=1}^{k} \alpha^{-1} I\{\sigma_i \leqslant \alpha\}} \frac{d_{k+1}}{b_k(k_0+1)}$$

$$\times \left\{ p_{\mathrm{D}}(u) \prod_i g_{1,k}(v_i) \alpha^{-1} \right\}^{-1} J,$$

where k_0 is the number of empty components before the birth and $p_{\mathrm{D}}(u)$ is the $D(\delta, \ldots, \delta)$ density. Furthermore, J is the Jacobian determinant of the transformation from (a_{ij}, u_j), where $i, j \neq j_*$, to $(\tilde{a}_{ij}, a_{j_*j})$, where $j \neq j_*$. Note that we can omit the index j_* because row sums of A and $\tilde{A}$ equal 1. Since the new component is drawn from the prior, cancellations occur in the expression for R, and when $\delta = 1$ it simplifies to

$$R = \frac{p(k+1)}{p(k)} k^k \frac{p(\tilde{z}|\tilde{A})}{p(z|A)} (k+1) \frac{d_{k+1}}{b_k(k_0+1)} \left\{ \prod_i g_{1,k}(v_i) \right\}^{-1} J. \tag{14}$$

Finally,

$$J = \sum_{i \neq j_*} (1 - v_i)^{k-1}; \tag{15}$$

this expression corresponds to the last factor of expression (12) in Richardson and Green (1997), but there is a mistake in their formula, in that exponent k should be $k - 1$; see Richardson and Green (1998). For the death move, the corresponding u is given by the deleted row of $\tilde{A}$ and the v_i are given by equation (13).

That the sampler behaves as desired, in terms of converging to a realization from the joint distribution in equation (1), follows by arguments that are similar to those used at the end of section 3.2 in Richardson and Green (1997) to establish aperiodicity and irreducibility. For instance, irreducibility obtains because the chain can move among the possible values of k by increasing or decreasing its value by 1, with positive probability, in step (c) all allocations z have positive probability, in step (a) the conditional density is positive on the natural parameter space and the same holds true for steps (b) and (d) if we consider two consecutive sweeps.

4. Applications

4.1. Description of examples

To investigate the behaviour and performance of the reversible jump MCMC algorithm, we applied it to three different sets of data. The first set is an extract from the Standard and Poors 500 stock index. It consists of 1700 observations of daily returns during the 1950s. This

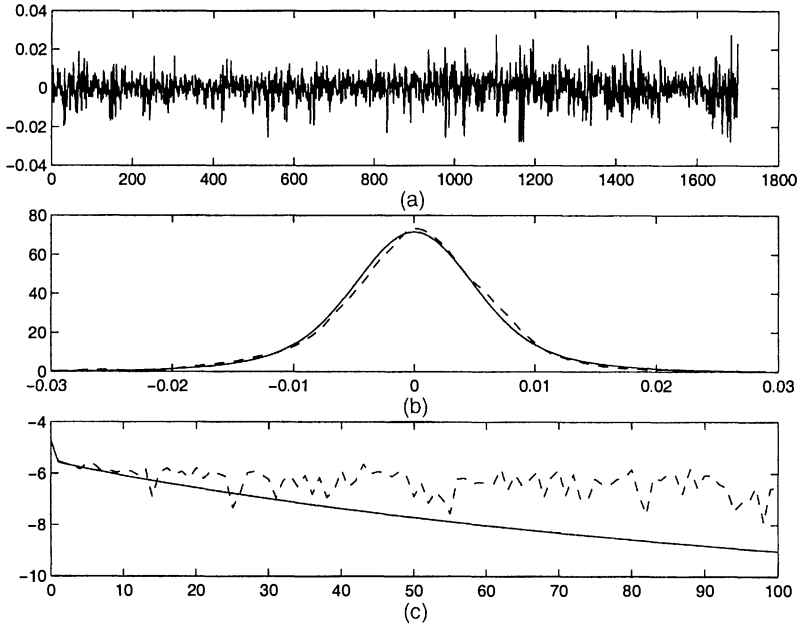

Fig. 1. (a) Standard and Poors 500 data, (b) Bayesian estimate (———) and kernel estimate (– – –) of its marginal density and (c) Bayesian estimate (———) and nonparametric estimate (– – –) of the covariance function of absolute values (base 10 logarithmic scale)

series was previously analysed in Rydén *et al.* (1998); it is the subseries called series E in that paper. The data were preprocessed by shrinking observations outside the range $m \pm 4s$, where m and s are the sample mean and sample standard deviation respectively; see Rydén *et al.* (1998). We also refer readers to Rydén *et al.* (1998) for more information about the data. Our second set of data is a series of 500 hourly wind measurements at Athens in January 1990. It has previously been analysed by Francq and Roussignol (1997) by using a model which was almost identical with ours. The difference is that their hidden Markov chain z cannot jump arbitrarily between states; rather there is a 'base state' that communicates with all other states, but any other jumps are disallowed. Our third set of data consists of 2700 residuals from a fit of an autoregressive moving average model to three-hour planetary geomagnetic activity index measurements. This data set was also analysed by Francq and Roussignol (1997), by using a restricted model as above, and we refer to that paper for further information. The data sets are plotted in Figs 1–3. The wind data are discretized in steps of 0.1 and contain 56 observations that equal 0. The likelihood is then unbounded; indeed, fixing $k > 1$, A and all σ_i except one and letting the remaining σ_i tend to 0 we see that the likelihood grows indefinitely. This of course makes unrestricted maximum likelihood estimation inappropriate and it also has a substantial effect on the result of a Bayesian analysis. Therefore, before using the data, we added an independent uniform $(-0.05, 0.05)$ random variable to each observation.

4.2. Some results

For each set of data our results are based on 1 million sweeps of the MCMC algorithm after a burn-in of 100000 sweeps. Table 1 shows the posterior distribution of the number of components k. For the Standard and Poors 500 series, the results for $k = 2$ and $k = 3$ are almost

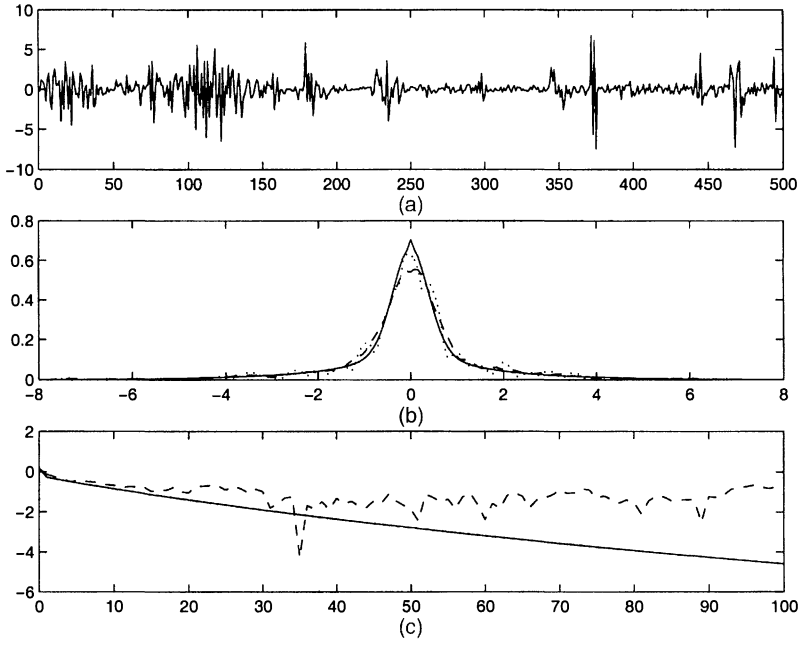

Fig. 2. (a) Wind data, (b) Bayesian estimate (———) and two kernel estimates (– – –, ·····) (with different bandwidths) of its marginal density and (c) Bayesian estimate (———) and nonparametric estimate (– – –) of the covariance function of absolute values (base 10 logarithmic scale)

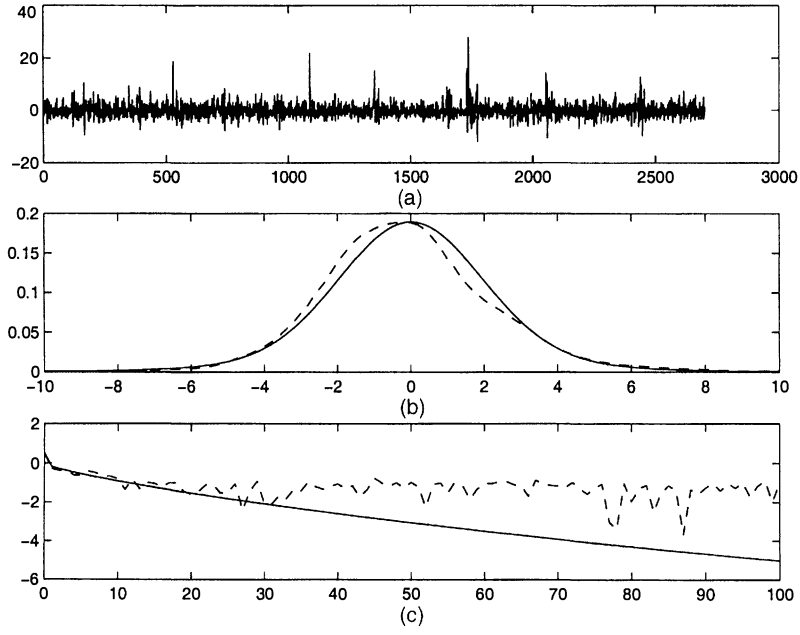

Fig. 3. (a) Magnetic activity data, (b) Bayesian estimate (———) and kernel estimate (– – –) of its marginal density and (c) Bayesian estimate (———) and nonparametric estimate (– – –) of the covariance function of absolute values (base 10 logarithmic scale)

Table 1. Posterior distribution of the number of components k

k	*Distributions for the following data sets:*		
	Standard and Poors 500	*Wind*	*Magnetic activity*
1	0.0000	0.0000	0.0000
2	0.4877	0.0050	0.0000
3	0.4521	0.8725	0.9830
4	0.0550	0.1168	0.0169
5	0.0040	0.0055	0.0002
6	0.0009	0.0001	0.0000
7	0.0003	0.0000	0.0000

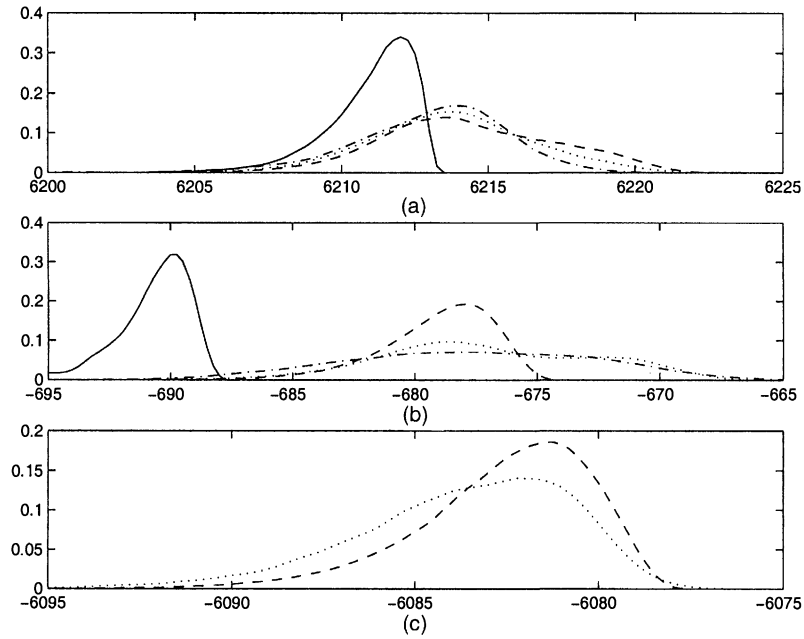

Fig. 4. Density plots for the log-likelihood values sampled at every 10th sweep: (a) Standard and Poors 500 data; (b) wind data; (c) magnetic activity data (———, $k = 2$; — — —, $k = 3$; ·····, $k = 4$; — · — · ·; $k = 5$)

identical. Rydén *et al.* (1998) tested the null hypothesis $k = 2$ *versus* $k = 3$ by using a bootstrapped likelihood ratio test, and they obtained the simulated p-value 0.57. Even if there is no exact correspondence between this figure and the present results, they indicate a similar degree of belief in $k = 2$. Fig. 4 shows density plots of the log-likelihood values sampled at every 10th sweep. There is a clear difference between $k = 2$ and $k = 3$, after which the curves essentially overlap. We also note that the parameter posterior means given $k = 2$ are $\bar{a}_{12} = 0.044$, $\bar{a}_{21} = 0.083$, $\bar{\sigma}_1 = 0.0046$ and $\bar{\sigma}_2 = 0.0093$, which are close to the MLEs 0.037, 0.069, 0.0046 and 0.0092 respectively found in Rydén *et al.* (1998).

For the wind data, Table 1 suggests $k = 3$. Francq and Roussignol (1997) proposed $k = 2$ but, as noted above, they used a somewhat different model. Also, they did not preprocess the data and do not discuss the problem of unbounded likelihoods. The log-likelihood obtained by their MLE is about -689 for both the original and the preprocessed data, which could be compared with the log-likelihood curves in Fig. 4. Fig. 4 shows a clear distinction between

Table 2. Posterior distribution of the number of components k for the Standard and Poors 500 data, $1/\xi = c \max|y_t|$ and various choices of c

k	*Distributions for the following values of c:*						
	$c = 5$	$c = 10$	$c = 20$	$c = 30$	$c = 50$	$c = 70$	$c = 100$
2	0.4607	0.4586	0.4785	0.4793	0.4701	0.4585	0.4831
3	0.4745	0.4756	0.4580	0.4608	0.4723	0.4811	0.4568
4	0.0597	0.0607	0.0583	0.0545	0.0528	0.0550	0.0549
5	0.0047	0.0044	0.0049	0.0048	0.0043	0.0051	0.0050
6	0.0004	0.0007	0.0003	0.0005	0.0005	0.0004	0.0003
7	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000

$k = 2$ and $k = 3$, but also a distinction when going to $k = 4$. For $k = 5$ the curves overlap, however.

For the magnetic activity data set, Table 1 again suggests $k = 3$ components. Francq and Roussignol (1997) proposed $k = 2$ with their MLE yielding a log-likelihood of about -6101. In Fig. 4 curves for $k = 2$ and $k = 5$ are not plotted because there were very few sweeps with such k.

One of the observations that Rydén *et al.* (1998) made was that the dependence structure of the Standard and Poors 500 subseries was difficult to capture with maximum likelihood estimation. First note that the HMMs that we consider are uncorrelated. Therefore, Rydén *et al.* (1998) looked rather at the covariance function of the absolute values of the series and found that it was quite slowly decaying in the Standard and Poors 500 data. However, this property was not present to the same extent in the estimated models. Figs 1–3 show standard nonparametric estimates of the covariance functions of $\{|y_t|\}$, together with a corresponding Bayesian estimate. By Bayesian estimate we mean that in each sweep the covariance function of the current model was computed, and these functions were then averaged over all sweeps. For details of the computation of the covariance function of the HMMs, see Rydén *et al.* (1998). Note that the averaged covariance function is not a covariance function of an HMM. Fig. 1 shows a reasonable agreement between the two estimates up to lags of about 15, but after this the Bayesian estimate decays faster. Similar features emerge for the other sets of data. In contrast, in the Standard and Poors 500 series, for larger lags the nonparametric estimate is of the order of 10^{-6}, from which it is difficult to judge whether the true covariance function is different from 0. Figs 1–3 also show Bayesian estimates of the marginal density together with kernel density estimates, and these agree very well.

4.3. Stability with respect to the prior

The parameter α that determines the range of the prior for the σ_i is obviously a main design parameter. If α is fixed, then we would like to set it larger than $\max|y_t|$ in order not to restrict the range for the σ_i too much. However, it is immediate from equation (12) that the acceptance probability of the split move is inversely proportional to α, so a large α pushes the posterior distribution for k towards $k = 1$. That α affects the acceptance ratio and posterior distribution so directly is of course undesirable, and it is for this reason that α has been made a hyperparameter. Our choice was to let α have a negative exponential distribution with mean $1/\xi$. Table 2 shows, for the Standard and Poors 500 data, the posterior distribution of k for various choices of $1/\xi$ in the range $5 \max|y_t|$–$100 \max|y_t|$. The posteriors virtually agree, showing that this prior structure is much less informative than that with a fixed α. The results in this paper were obtained using $1/\xi = 30 \max|y_t|$.

4.4. Performance of the Markov chain Monte Carlo algorithm

The acceptance rates for the split or combine move were 4.4%, 1.4% and 0.26% respectively for the Standard and Poors 500 data, the wind data and the magnetic activity data. These rates are a little lower than desired, but since the split and combine moves involve a change of model dimension only and all the other parameters are updated in each sweep we feel that they are not too low, except for the magnetic activity data. The results for this series should thus be handled with caution. It should be pointed out that the number of invalid split proposals, i.e. proposals with $\tilde{A}$ being non-stochastic, was very small and hence proposals were rejected mainly because of the acceptance probabilities R.

One should also keep in mind that the degree of precision in the posterior distribution of k and the structure of the jump (dimension changing) move bound the achievable acceptance rate. Suppose, for simplicity, that there are only two possible values 1 and 2 for k, that the posterior probability for $k = 1$ is 0.9 and that the jump move always proposes to change the dimension. Then, in the long run, only one sweep out of 10 may have $k = 2$, a change of dimension may occur only in two sweeps out of 10, and the acceptance rate is thus bounded by 0.2. For the wind and magnetic activity data the posteriors for k are quite concentrated on $k = 3$, but of course we are far from attaining the corresponding bound for the acceptance rates.

The acceptance rates are affected by the parameters of the beta distributions in the split move. We have run the above version and variants of the present algorithm with a large number of choices for the parameters, among which those given in Section 3 were the best. The birth or death move was completely redundant in the sense that its acceptance rates were virtually zero (below 3×10^{-5}). Hence, these moves can just as well be removed from the algorithm. Figs 5–7 show some plots relating to the mixing and convergence of the sampler.

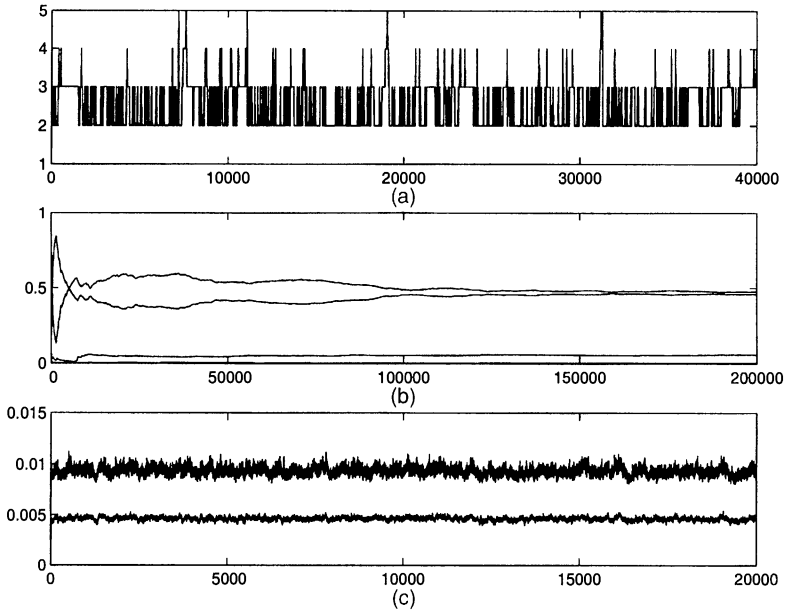

Fig. 5. (a) First 40000 values of k for the Standard and Poors 500 data, plotted every 20th sweep, (b) estimated posterior distribution of k as a function of the number of sweeps and (c) σ_1 and σ_2 in the first 20000 sweeps with $k = 2$

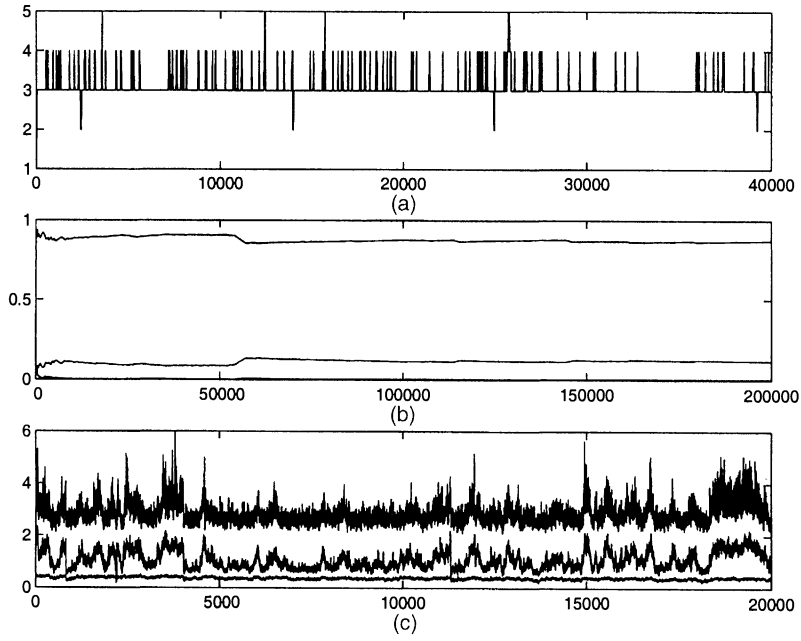

Fig. 6. (a) First 40000 values of k for the wind data, plotted every 20th sweep, (b) estimated posterior distribution of k as a function of the number of sweeps and (c) σ_1, σ_2 and σ_3 in the first 20000 sweeps with $k = 3$

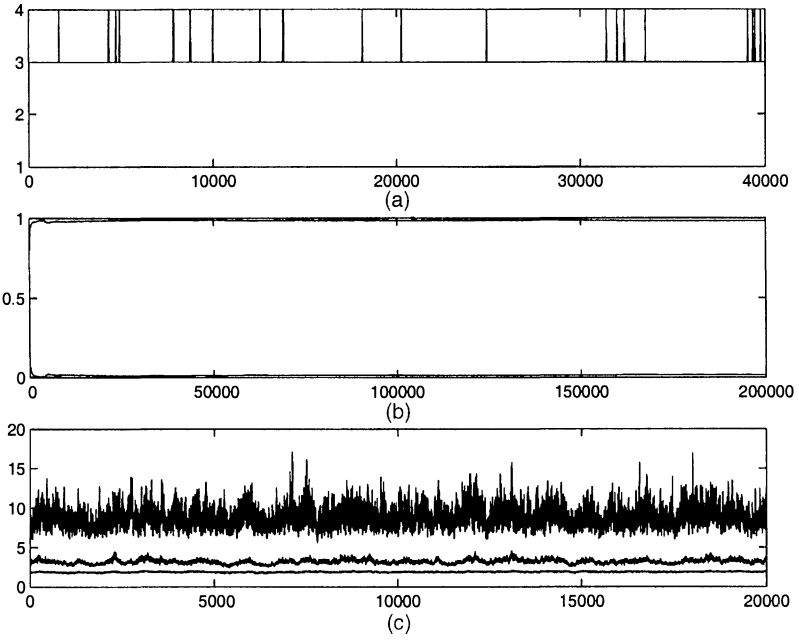

Fig. 7. (a) First 40000 values of k for the magnetic activity data, plotted every 20th sweep, (b) estimated posterior distribution of k as a function of the number of sweeps and (c) σ_1, σ_2 and σ_3 in the first 20000 sweeps with $k = 3$

Figs 5(a), 6(a) and 7(a) show the first 40000 values of k, i.e. the beginning of the burn-in, plotted every 20th sweep, for the three sets of data. Figs 5(b), 6(b) and 7(b) show the settling period of the empirical estimate of the posterior distribution of k. Note that for all sets of data this posterior is virtually zero for most values of k, whence these probabilities are not visible in the figures although all probabilities are indeed plotted. Figs 5(c), 6(c) and 7(c) show the first 20000 values of the σ_i in sweeps with k equalling the value for which the posterior distribution attains its maximum.

4.5. Comparison with an analysis of independent and identically distributed data

We found it interesting to compare our results with what is obtained under assumptions of independent and identically distributed (IID) data, i.e. when the z_t are independent. Obviously, the y_t are then also independent, with each y_t having the mixture distribution $\Sigma_1^k \pi_i \mathcal{N}(0, \sigma_i^2)$ where $\pi_i = P(z_t = i)$. We implemented a reversible jump MCMC algorithm similar to that presented in Richardson and Green (1997) for this case. The prior for k was as above, i.e. uniform on $\{1, 2, \ldots, k_{\max}\}$, the prior for the weights $\pi_1, \ldots, \pi_k$ conditional on k was $D(1, \ldots, 1)$ and the prior structure for the σ_i was as before. In the combine move, the new component weight π_{j_*} was obtained as $\pi_{j_*} = \tilde{\pi}_{j_1} + \tilde{\pi}_{j_2}$ and the new standard deviation was obtained as $\sigma_{j_*} = \tilde{\pi}_{j_1}\tilde{\sigma}_{j_1} + \tilde{\pi}_{j_2}\tilde{\sigma}_{j_2}$. In the split move, the component weight π_{j_*} was split as $\tilde{\pi}_{j_1} = u_0\pi_{j_*}$ and $\tilde{\pi}_{j_2} = (1 - u_0)\pi_{j_*}$ with $u_0 \sim \text{Be}(2, 2)$, and the standard deviation σ_{j_*} was split as

$$\tilde{\sigma}_{j_1} = \sigma_{j_*}[1 - w\sqrt{\{(1 - u_0)/u_0\}}]$$

and

$$\tilde{\sigma}_{j_2} = \sigma_{j_*}[1 + w\sqrt{\{u_0/(1 - u_0)\}}]$$

with $w \sim \text{Be}(1, k)$. Hence, this split or combine pair works in a way that differs slightly from that in Section 3, preserving first absolute moments rather than second moments.

We ran the algorithm as before, i.e. with 1 million sweeps after a 100000-sweep burn-in. The results are quite different from what is obtained by modelling with HMMs. Fig. 8 shows the posterior distributions for k, which are much more spread out than in Table 1. For the Standard and Poors 500 and wind data, there is even mass at $k = 30$, which was our $k_{\max}$. For the magnetic activity data the distribution is less spread out than for the other two data sets and has its mode at $k = 3$, just as in the HMM case. Comparing the posterior means of the σ_i given some k, however, we find similarities. For the Standard and Poors 500 data and $k = 2$, the results are (0.00465, 0.00934) and (0.00464, 0.00995) for the HMM and IID data analysis respectively; for the wind data and $k = 3$ they are (0.375, 1.08, 2.87) and (0.0867, 0.657, 2.59), and for the magnetic activity data and $k = 3$ they are (1.90, 3.28, 8.84) and (1.91, 3.65, 10.1).

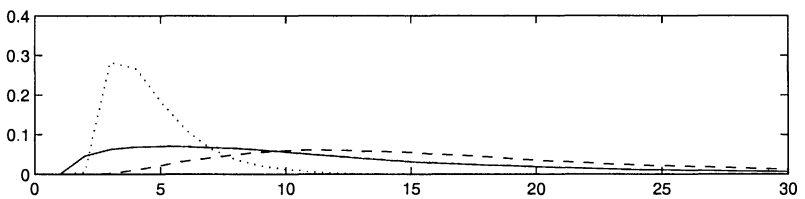

Fig. 8. Posterior distribution of k under assumptions of IID data for the Standard and Poors (———), wind (– – –) and magnetic activity (·····) data

For the wind data, there are only 2488 samples with $k = 3$ in the IID data analysis, which may explain the larger discrepancy.

The acceptance rates for the split or combine move were much larger for the IID data algorithm, despite being less sophisticated in the sense that the valid range of w is not computed. The rates were 31% for the Standard and Poors 500 data, 33% for the wind data and 22% for the magnetic activity data.

5. Conclusions and scope for the future

We have seen that it is possible to construct reversible jump MCMC algorithms for Bayesian analysis of HMMs with an unknown number of components. We believe that such algorithms may be strong competitors to a frequentist approach based on maximum likelihood, not least from a computational point of view, because they combine parameter estimation and model selection in a powerful way. Obviously, it is desirable to find dimension changing moves that achieve larger acceptance rates than those presented here. It may be fruitful to consider moves that preserve the marginal distribution to a lesser extent than do the above moves, but that instead preserve some of the dependence over time, such as the first-order autocovariance of absolute values. We have also experimented with a split move that first splits the standard deviation σ_{j*}, then reallocates the appropriate z_t as if they were independent and finally draws rows j_1 and j_2 of $\tilde{A}$ as in move (a) of Section 3. The acceptance rates obtained by using this split or combine pair were lower than those reported earlier, however.

We find the differences between the results from HMMs and the results based on assumptions of IID data quite striking. On the basis of our experiences with the latter analysis we make two tentative conclusions. First, adding structure (here Markov dependence) to a model pushes the posterior distribution of k towards smaller values. Secondly, adding structure to a model makes it more difficult to design split or combine moves, or more generally dimension changing moves, that yield high acceptance rates.

Acknowledgements

The research of CPR relates to the Training and Mobility of Researchers Network on Spatial and Computational Statistics and was partially supported by CREST, Institut National de la Statistique et des Etudes Economiques. The research of TR was supported by the Swedish Natural Science Research Council (contract M-AA/MA 10538-303) and by the Royal Swedish Academy of Sciences through a grant from the Gustav Sigurd Magnuson Foundation. This work was initiated when CPR and TR visited the Department of Statistics, University of Glasgow, and they wish to thank the members of the department for its friendly environment and financial support.

The density estimates in Figs 1–4 were computed in Matlab by using the kernel density estimation toolbox of C. C. Beardah, Nottingham Trent University, UK.

Appendix A

In this appendix we describe the restricted backward algorithm used to update the hidden Markov chain z in the split move. We are given two time indices $t_1 \leqslant t_2$ such that $z_t = j_*$ for $t_1 \leqslant t \leqslant t_2$ but z_{t_1-1}, $z_{t_2+1} \neq j_*$. We are also given all y_t, the proposed $\tilde{A}$ and $\tilde{\sigma}_i$, and we want to sample $\tilde{z}_{t_1}, \ldots, \tilde{z}_{t_2}$ from the conditional distribution given $\tilde{z}_{t_1-1} = z_{t_1-1}$, $\tilde{z}_{t_2+1} = z_{t_2+1}$ and the above variables and given $\tilde{z}_t \in \{j_1, j_2\}$

Table 3. Table of partial derivatives

	$\tilde{a}_{ij}$	$\tilde{a}_{ij_1}$	$\tilde{a}_{ij_2}$	$\tilde{a}_{j_1 j}$	$\tilde{a}_{j_2 j}$	$\tilde{a}_{j_1 j_2}$	$\tilde{a}_{j_2 j_1}$	$\tilde{\sigma}_{j_1}$	$\tilde{\sigma}_{j_2}$
a_{ij}	I	**0**	**0**	**0**	**0**	**0**	×	**0**	**0**
a_{ij_*}	**0**	diag(v_i)	diag($1 - v_i$)	**0**	**0**	**0**	×	**0**	**0**
v_i	**0**	diag(a_{ij_*})	−diag(a_{ij_*})	**0**	**0**	**0**	×	**0**	**0**
$a_{j_* j}$	**0**	**0**	**0**	×	×	×	×	**0**	**0**
u_j	**0**	**0**	**0**	×	×	×	×	**0**	**0**
u_0	**0**	**0**	**0**	×	×	×	×	×	×
u_1	**0**	**0**	**0**	**0**	**0**	×	×	**0**	**0**
σ_{j_*}	**0**	**0**	**0**	**0**	**0**	**0**	**0**	×	×
w	**0**	**0**	**0**	**0**	**0**	**0**	**0**	×	×

for all $t_1 \leqslant t \leqslant t_2$. It is easy to see that $\tilde{z}_{t_1}, \ldots, \tilde{z}_{t_2}$ is an inhomogeneous Markov chain given the y_t and the parameters, and it still is when restricted to $\{j_1, j_2\}$.

Define the two-dimensional vectors $b_t = (b_t(j_1), b_t(j_2))$, $t_1 \leqslant t \leqslant t_2$, by

$$b_t(i) = P(y_{t+1}, \ldots, y_{t_2}, \tilde{z}_{t+1} \in C, \ldots, \tilde{z}_{t_2} \in C, \tilde{z}_{t_2+1} | \tilde{z}_t = i, \tilde{A}, \tilde{\sigma}),$$

where $C = \{j_1, j_2\}$. These vectors may be computed recursively as

$$b_{t_2}(i) = \tilde{a}_{i\tilde{z}_{t_2+1}} \tag{16}$$

and, for $t = t_2 - 1, t_2 - 2, \ldots, t_1$,

$$b_t(i) = \sum_{j=j_1,j_2} b_{t+1}(j) \tilde{a}_{ij} \varphi(y_{t+1}; \tilde{\sigma}_j). \tag{17}$$

Now for each $t = t_1, \ldots, t_2$, in ascending order, we first compute the two-dimensional vector $c_t = (c_t(j_1), c_t(j_2))$ as

$$c_t(j) = \tilde{a}_{\tilde{z}_{t-1} j} \varphi(y_t; \tilde{\sigma}_j) b_t(j), \tag{18}$$

then we renormalize this vector to make it sum to 1 and finally we draw $\tilde{z}_t$ from the probability distribution so obtained. The normalized c-vector is the conditional distribution $P(\tilde{z}_t = j | \tilde{z}_{t-1}, y_1, \ldots, y_n, \tilde{z}_t \in C, \tilde{z}_{t+1} \in C, \ldots, \tilde{z}_{t_2} \in C, \tilde{z}_{t_2+1}, \tilde{A}, \tilde{\sigma})$, so this procedure provides us with a sample $\tilde{z}_{t_1}, \ldots, \tilde{z}_{t_2}$ from the desired conditional distribution. If $t_2 = n$, then replace the right-hand side of equation (16) by 1, and, if $t_1 = 1$, then replace $\tilde{a}_{\tilde{z}_{t_1-1} j}$ on the right-hand side of equation (18) by the stationary probability $\tilde{\pi}_j$.

Appendix B

In this appendix we evaluate the Jacobian determinant of the split move. We look at the transformation given by equations (8) and (10) from $(a_{ij}, a_{ij_*}, v_i, a_{j_* j}, u_j, u_0, u_1, \sigma_{j_*}, w)$, where $i, j \neq j_*$, to $(\tilde{a}_{ij}, \tilde{a}_{ij_1}, \tilde{a}_{ij_2}, \tilde{a}_{j_1 j}, \tilde{a}_{j_2 j}, \tilde{a}_{j_1 j_2}, \tilde{a}_{j_2 j_1}, \tilde{\sigma}_{j_1}, \tilde{\sigma}_{j_2})$, where $i, j \neq j_1, j_2$. Also, we do not include any diagonal elements of A or $\tilde{A}$ in the tuples above; we can omit these because the row sums equal 1. We obtain the table of partial derivatives, which defines the Jacobian matrix (Table 3). Here I denotes an identity matrix, **0** denotes a suitably sized vector or matrix of 0s and × denotes non-zero entries. Since the Jacobian matrix has a block diagonal structure, it follows that we can disregard the rows and columns corresponding to a_{ij} and $\tilde{a}_{ij}$. What then remains of J is upper block diagonal, whence it follows that we can consider separately the two subdeterminants corresponding to $(a_{ij_*}, v_i) \mapsto (\tilde{a}_{ij_1}, \tilde{a}_{ij_2})$ and $(\sigma_{j_*}, w) \mapsto (\tilde{\sigma}_{j_1}, \tilde{\sigma}_{j_2})$

Since adding a column of a matrix to another column does not change the determinant, the first subdeterminant is given by

$$\begin{vmatrix} \text{diag}(v_i) & \text{diag}(1 - v_i) \\ \text{diag}(a_{ij_*}) & -\text{diag}(a_{ij_*}) \end{vmatrix} = \begin{vmatrix} I & \text{diag}(1 - v_i) \\ \mathbf{0} & -\text{diag}(a_{ij_*}) \end{vmatrix} = \prod_{i \neq j_*} a_{ij_*}.$$

Table 4. Remaining part of the Jacobian J

	$\tilde{a}_{j_1 j}$	$\tilde{a}_{j_2 j}$	$\tilde{a}_{j_1 j_2}$	$\tilde{a}_{j_2 j_1}$
$a_{j_* j}$	$\operatorname{diag}\left(\frac{u_j}{u_0}\right)$	$\operatorname{diag}\left(\frac{1-u_j}{1-u_0}\right)$	$-u_1\frac{u_j}{u_0}$	x
u_j	$\operatorname{diag}\left(\frac{a_{j_* j}}{u_0}\right)$	$-\operatorname{diag}\left(\frac{a_{j_* j}}{1-u_0}\right)$	$-\frac{u_1}{u_0}a_{j_* j}$	$\frac{1-u_1}{1-u_0}a_{j_* j}$
u_0	$-\frac{u_j}{u_0^2}a_{j_* j}$	$\frac{1-u_j}{(1-u_0)^2}a_{j_* j}$	$\frac{u_1}{u_0}(1-\tilde{a}_{j_1 j_1}-\tilde{a}_{j_1 j_2})$	$\frac{u_1+\tilde{a}_{j_2 j_1}}{1-u_0}$
u_1	**0**	**0**	$\tilde{a}_{j_1 j_1}+\tilde{a}_{j_1 j_2}$	$\frac{u_0}{1-u_0}(\tilde{a}_{j_1 j_1}+\tilde{a}_{j_1 j_2})$

The second subdeterminant is given by

$$\begin{vmatrix} \left\{1-w\sqrt{\left(\frac{1-u_0}{u_0}\right)}\right\}^{1/2} & \left\{1+w\sqrt{\left(\frac{u_0}{1-u_0}\right)}\right\}^{1/2} \\ -\frac{1}{2}\sigma_{j_*}\left\{1-w\sqrt{\left(\frac{1-u_0}{u_0}\right)}\right\}^{-1/2}\sqrt{\left(\frac{1-u_0}{u_0}\right)} & \frac{1}{2}\sigma_{j_*}\left\{1+w\sqrt{\left(\frac{u_0}{1-u_0}\right)}\right\}^{-1/2}\sqrt{\left(\frac{u_0}{1-u_0}\right)} \end{vmatrix}$$

$$= \frac{\sigma_{j_*}/2}{([u_0 - w\sqrt{\{u_0(1-u_0)\}}][1-u_0+w\sqrt{\{u_0(1-u_0)\}}])^{1/2}}.$$

What now remains of J is shown in Table 4, where most of the non-zero entries are now explicitly given. The determinant of this matrix is evaluated numerically. In doing this, we must compute the partial derivatives $\partial\tilde{a}_{j_2 j_1}/\partial a_{j_* j}$ for $j \neq j_*$. Looking at equation (8), we see that the main problem here is to calculate $\partial\kappa_i/\partial a_{j_* j}$ for $i, j \neq j_*$ (recall that $\kappa_i = \pi_i/\pi_{j_*}$). We now show how to do this.

Recall that the diagonal entries of A are not considered as independent variables. The equilibrium equations are

$$\pi_l = \pi_{j_*}a_{j_* l} + \sum_{i\neq j_*, l}\pi_i a_{il} + \pi_l a_{ll} = \pi_{j_*}a_{j_* l} + \sum_{i\neq j_*, l}\pi_i a_{il} + \pi_l\left(1-\sum_{i\neq l}a_{li}\right),$$

whence

$$a_{j_* l} + \sum_{i\neq j_*, l}\kappa_i a_{il} - \kappa_l\sum_{i\neq l}a_{li} = 0$$

for each $l \neq j_*$. Differentiating with respect to $a_{j_* j}$ and writing $\kappa_i' = \partial\kappa_i/\partial a_{j_* j}$, we obtain

$$\delta_{lj} + \sum_{i\neq j_*, l}\kappa_i' a_{il} - \kappa_l'\sum_{i\neq l}a_{li} = \delta_{lj} + \sum_{i\neq j_*}\kappa_i' a_{il} - \kappa_l' = 0 \tag{19}$$

for each $l \neq j_*$, where δ_{lj} is Kronecker's δ. The coefficient matrix of this linear system of equations is the restriction to states not equal to j_* of the matrix $A - I$. Our prior for A implies that all its entries are non-negative, whence A is irreducible and aperiodic. Hence 1 is a single eigenvalue of A, so $A - I$ has rank $k-1$. Therefore, its restriction to $k-1$ states has full rank and equation (19) has a unique solution κ_l', $l \neq j_*$.

References

Baum, L. E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

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.

Bickel, P. J., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Billio, M., Monfort, A. and Robert, C. P. (1999) Bayesian estimation of switching ARMA models. *J. Econometr.*, to be published.

Chib, S. (1996) Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometr.*, **75**, 79–97.

Churchill, G. A. (1995) Accurate restoration of DNA sequences (with discussion). In *Case Studies in Bayesian Statistics* (eds C. Gatsonis, J. S. Hodges, R. E. Kass and N. D. Singpurwalla), vol. II, pp. 90–148. New York: Springer.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc.* B, **61**, 331–344.

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

Fredkin, D. R. and Rice, J. A. (1992) Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. R. Soc. Lond.* B, **249**, 125–132.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

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

Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Krolzig, H.-M. (1997) Markov-switching vector autoregressions. *Lect. Notes Econ. Math. Syst.*, **454**.

Leroux, B. G. (1992) Maximum-likelihood estimation for hidden Markov models. *Stoch. Processes Applic.*, **40**, 127–143.

Leroux, B. G. and Puterman, M. L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545–558.

MacDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman and Hall.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318–324.

Petrie, T. (1969) Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **40**, 97–115.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–284.

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.

———(1998) Corrigendum: On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc.* B, **60**, 661.

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

Robert, C. P. and Titterington, D. M. (1998) Resampling schemes for hidden Markov models and their application for maximum likelihood estimation. *Statist. Comput.*, **8**, 145–158.

Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998) Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometr.*, **13**, 217–244.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701–1762.