Scientific Programming 10 (2002) 241–251
IOS Press

On determining the order of Markov dependence of an observed process governed by a hidden Markov model

R.J. Boys[a] and D.A. Henderson[b,*]
[a]*Department of Statistics, University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK*
Tel.: +44 (0)191 222 7297; E-mail: Richard.Boys@ncl.ac.uk
[b]*Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK*
Tel.: +44 (0)1908 652 030; E-mail: d.a.henderson@open.ac.uk

Abstract. This paper describes a Bayesian approach to determining the order of a finite state Markov chain whose transition probabilities are themselves governed by a homogeneous finite state Markov chain. It extends previous work on homogeneous Markov chains to more general and applicable hidden Markov models. The method we describe uses a Markov chain Monte Carlo algorithm to obtain samples from the (posterior) distribution for both the order of Markov dependence in the observed sequence and the other governing model parameters. These samples allow coherent inferences to be made straightforwardly in contrast to those which use information criteria. The methods are illustrated by their application to both simulated and real data sets.

1. Introduction

Markov chains are commonly used as models for data which are observed in discrete time and have a discrete and finite state space. Their application to time series data is widespread, ranging from the analysis of sequences of daily rainfall at a particular location to studying patterns of bases in a deoxyribonucleic acid (DNA) sequence. These data can be described using either a qth order Markov chain with state space $\mathcal{Y}$ or (equivalently) when $q > 1$, a first order Markov chain with a larger state space $\mathcal{Y}^q$ and a constrained transition structure. However, in most practical situations the parameter q is not known, and this leads to fundamental difficulties in making inferences from the data under either model description.

The problem of estimating the order of dependence of a homogeneous Markov chain has a long history dating back to methods based on likelihood ratio tests described by [3,19]. This work was followed by contributions describing procedures based on information (penalized likelihood) criteria such as the AIC [30] and the BIC [21]. Much more recently, a procedure which uses a Bayesian approach to determine the posterior distribution for the order of dependence has been described in [14]. They also show that their method performs favourably when compared to those which use information criteria. In this paper we generalise their method to one which determines the order of dependence in a heterogeneous Markov chain $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ which follows a hidden Markov model (HMM). These models are characterised by r homogeneous Markov chains together with an additional first order homogeneous r-state hidden Markov chain $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$. The hidden chain describes which of the r chains governs the evolution of the observed process at any particular time. HMMs have proved to be very flexible models for describing heterogeneity in time series data and have been applied to a wide variety of problems; [9] provides a comprehensive list of references. For more background on

*Corresponding author.

ISSN 1058-9244/02/$8.00

HMMs see, for example, [26] and [22]. Note that some authors, for example, [6], refer to these models as double chain Markov models (DCMMs).

We shall assume that each of the r homogeneous Markov chains has order q ($\geqslant 0$), that is, the probability of the current observation Y_t depends only on the previous q observations $Y_{t-q}, \ldots, Y_{t-1}$. Also, we shall assume that each chain has state space $\mathcal{Y} = \{1, 2, \ldots, b\}$. Thus the inferential problem is to determine values for q and the other model parameters which are consistent with the data. The key quantities we require for a Bayesian analysis of this problem are the posterior (model) probabilities of q. These describe in simple terms how likely are different values of q in light of the data. Alternatively (and equivalently) we could calculate Bayes factors as is commonplace in Bayesian model choice problems; see [20].

A method for determining the order of dependence in homogeneous sequences ($r = 1$) has been provided by [14]. Their approach is particularly appealing as it is fairly straightforward to use. Moreover, the ease of their approach is a direct consequence of their choice of prior distribution for the transition probabilities governing the evolution of the underlying process; we will return to this point later. Their method can be easily extended to the more general HMM context ($r > 1$) if the configuration of hidden states $\boldsymbol{s}$ is known. In particular, formulae for posterior model probabilities (and thus Bayes factors) are available analytically. However, a fundamental drawback of using HMMs is that the configuration $\boldsymbol{s}$ is unknown and has to be determined from the observed data $\boldsymbol{y}$. This complication precludes a fully analytic treatment of the model and so we resort to computer intensive Markov chain Monte Carlo (MCMC) methods. These methods involve generating samples from a Markov chain which has been constructed so that its stationary distribution is the (posterior) distribution of interest. The resultant dependent samples can be used to approximate posterior quantities of model parameters such as the configuration $\boldsymbol{s}$ and the order of dependence q; see [8] for an overview. These methods also ensure that inferences take due account of uncertainty surrounding the correct configuration, in contrast to many plug-in methods [13].

In many scenarios the number of different hidden states r in the HMM will be unknown a priori. However, for ease of explanation, we restrict our attention to the case where r is known. It is possible to extend the methods described in this paper to also include estimation of r but this would require the use of methods such as those described in [28] which are based on reversible jump MCMC.

The remainder of the paper is organised as follows. The HMM is described in Section 2, followed by details of our Bayesian approach to inference in Section 3. Section 4 outlines an implementation of our MCMC algorithm on both a simulated and a real data set. The paper concludes in Section 5 with a discussion.

2. Model description

We shall assume that the observed data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ are a realisation of a hidden Markov model with observation equations

$$\begin{aligned} & r(Y_t | Y_{1:t-1}, S_{1:t}) \\ &= r(Y_t = j | Y_{t-q} = y_{t-q}, \ldots, Y_{t-1} \\ &\quad = y_{t-1}, S_t = k) \\ &= \rho^{(k)}_{y_{t-q:t-1}, j}, \; y_{t-q}, \ldots, y_{t-1}, \; j \in \mathcal{Y}, \quad k \in \mathcal{S}, \end{aligned}$$

and state equations

$$r(S_t | S_{1:t-1}) = r(S_t = j | S_{t-1} = i) = \lambda_{ij}, \quad i, j \in \mathcal{S}$$

for $t = q+1, q+2, \ldots, n$, where the notation $x_{i:j}$ denotes the sequence $x_i, x_{i+1}, \ldots, x_j$. Note that we have assumed, as is common practice, that the hidden process $\boldsymbol{S}$ follows a first order homogeneous Markov chain. We shall denote its state space by $\mathcal{S} = \{1, 2, \ldots, r\}$ and its transition matrix by $\Lambda = (\lambda_{ij})$, where each row $\boldsymbol{\lambda}_i \in \mathbf{S}_r = \{(x_1, x_2, \ldots, x_r); x_j > 0 \, \forall \, j, \; \sum_{j=1}^{r} x_j = 1\}$, the r-dimensional simplex. Although the order of dependence q is unknown, we shall assume that it can take values in $\mathcal{Q} = \{0, 1, \ldots, q_{\max}\}$, where $q_{\max} \geqslant 1$.

The layout of the qth order transition matrices with elements $\rho^{(k)}_{y_{t-q:t-1}, j}$ is problematic to work with computationally since each increase in order requires an extra dimension in an array. However, we can overcome this problem by reshaping each matrix into its reduced form [25]. Here, for each k, the reduced form $b^q \times b$ matrix $P^{(k)}$ consists of elements $p^{(k)}_{ij}$ where $i \in \mathcal{Y}_q = \{1, 2, \ldots, b^q\}$ indexes the rows of the matrix and $j \in \mathcal{Y}$ indexes the columns. The elements of the reshaped matrix corresponding to transition probabilities $\rho^{(k)}_{y_{t-q:t-1}, j}$ can be found on row

$$\begin{aligned} i &= \mathcal{I}(\boldsymbol{y}, t, q, b) \\ &\equiv 1 + \sum_{\ell=1}^{q} (y_{t-\ell} - 1) b^{\ell - 1}. \end{aligned}$$

In computational terms, this solution results in the algorithm working with fixed (two) dimensional arrays. Thus, each reshaped matrix $P^{(k)}$ has rows $\boldsymbol{p}_i^{(k)} \in \mathbf{S}_b$ and we denote the collection of these transition matrices by $\mathcal{P} = \{P^{(1)}, \ldots, P^{(r)}\}$.

For convenience in the following derivation, we denote the set of unknown hidden and observed state transition matrices by $\boldsymbol{\theta} = \{\Lambda, \mathcal{P}\} \in \mathbf{S}_r^r \times \mathbf{S}_b^{rb^q}$, where the space $\mathbf{S}_r^x$ denotes the product of x simplices, each one r-dimensional.

3. Bayesian inference

The aim of the analysis is to make inferences about the unknown quantities in the model: the order of dependence q, the model transition parameters $\boldsymbol{\theta}$ and the hidden states $\boldsymbol{s}$. We shall adopt a Bayesian approach to inference [24], and begin by quantifying our uncertainty about these unknowns (before observing the data) through a prior distribution.

3.1. Prior specification

We shall assume that our prior distribution takes the form

$$\pi(q, \boldsymbol{\theta}) = \pi(q)\pi(\boldsymbol{\theta}|q) = \pi(q)\pi(\Lambda)\pi(\mathcal{P}|q).$$

The discrete probability distribution $\pi(q)$ defined on $\mathcal{Q}$ describes our prior uncertainty surrounding the value of q. For example, without strong prior beliefs as to likely values, we might take $\pi(q)$ to be a discrete uniform distribution. Alternatively, if this uniform structure were thought inappropriate, for example, because larger values of q were believed to be relatively unlikely, then a truncated Poisson or geometric distribution might be appropriate choices.

The components of $\boldsymbol{\theta}$ are all defined on simplices and therefore there are many choices of priors which could be made. One rich family of distributions is provided by Aitchison's **A**–distribution [2] which has the logistic normal and the Dirichlet distributions as special (limiting) cases. In this paper we shall adopt the same choice as [14] which was fundamental to the simplicity of their method, namely, the Dirichlet distribution: $\boldsymbol{X} = (X_1, X_2, \ldots, X_r) \sim \mathbf{D}(\boldsymbol{\alpha})$ if it has density

$$\pi(\boldsymbol{x}) = \Gamma\left(\sum_{i=1}^{r} \alpha_i\right) \prod_{i=1}^{r} \frac{x_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}, \quad \boldsymbol{x} \in \mathbf{S}_r,$$

where $\Gamma(\cdot)$ is the gamma function [1]. Specifically, we take independent Dirichlet distributions for the rows of each $P^{(k)}$ and Λ, that is

$$\boldsymbol{p}_i^{(k)} = \left(p_{ij}^{(k)}\right) \sim \mathbf{D}\left(\boldsymbol{c}_i^{(k)}\right), \; i \in \mathcal{Y}_q, \; j \in \mathcal{Y}, \; k \in \mathcal{S},$$

$$\boldsymbol{\lambda}_i = (\lambda_{ij}) \sim \mathbf{D}(\boldsymbol{d}_i), \quad i, j \in \mathcal{S}.$$

The Dirichlet parameters $\boldsymbol{c}$ and $\boldsymbol{d}$ should be chosen to reflect the goal of the analysis. In this case, we have little knowledge about the transition structures in the data and so the exchangeable weak specification $\boldsymbol{c}_i^{(k)} = (1, 1, \ldots, 1)$ may be appropriate. The choice of parameters for the transition structure of the hidden chain is more complex. Usually, this is taken to be an off-diagonal exchangeable pattern of the form $(\boldsymbol{d}_i)_j = \alpha\delta_{ij} + \beta(1 - \delta_{ij})$ for some choice of α and β, where δ_{ij} is Kronecker's delta. These parameters are usually chosen to ensure a given prior mean and standard deviation for the length of runs of each state in the hidden chain, that is, for $(1 - \lambda_{kk})^{-1}$; for more details, see [7].

3.2. Likelihood

For this model, the complete-data likelihood is determined as the probability of both the observed and the unobserved data (hidden states) given the parameters, and is given by

$$\begin{aligned} \pi(\boldsymbol{y}, \boldsymbol{s}|q, \boldsymbol{\theta}) &\propto \prod_t \rho_{y_{t-q:t}}^{(s_t)} \lambda_{s_{t-1}s_t} \\ &= \prod_{i \in \mathcal{Y}_q} \prod_{j \in \mathcal{Y}} \prod_{k \in \mathcal{S}} \left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}} \prod_{i \in \mathcal{S}} \prod_{j \in \mathcal{S}} \lambda_{ij}^{m_{ij}}, \end{aligned} \tag{1}$$

where

$$n_{ij}^{(k)} = \sum_t \mathbf{I}\left(\mathcal{I}(\boldsymbol{y}, t, q, b) = i, y_t = j, s_t = k\right)$$

$$\text{and} \quad m_{ij} = \sum_t \mathbf{I}(s_{t-1} = i, s_t = j)$$

denote the observed transition counts and $\mathbf{I}(x)$ is the usual indicator function which equals 1 if x is true and 0 otherwise. Throughout this paper we perform inference conditional on the first $q_{\max}$ observations. This simplifies the solution by removing the need for marginal models to describe the evolution at the beginning of the sequence. Consequently, the range of values for t in the above expressions is $t = q_{\max} + 1, \ldots, n$.

3.3. Posterior inference

In the Bayesian paradigm, inferences are based on the posterior distribution

$$\pi(q,\boldsymbol{\theta},\boldsymbol{s}|\boldsymbol{y}) = \frac{\pi(q,\boldsymbol{\theta},\boldsymbol{s},\boldsymbol{y})}{\pi(\boldsymbol{y})} \propto \pi(q,\boldsymbol{\theta})\pi(\boldsymbol{y},\boldsymbol{s}|q,\boldsymbol{\theta}),$$

and this distribution calibrates our uncertainties about the unknown parameters after observing the data. Although this distribution is highly structured, it does not permit a straightforward analysis. However, the posterior distribution conditional on the hidden states $\boldsymbol{s}$ is much more amenable to analysis. It can be factorised into two component distributions $\pi(\boldsymbol{\theta}|q,\boldsymbol{s},\boldsymbol{y})$ and $\pi(q|\boldsymbol{s},\boldsymbol{y})$, and these distributions can be obtained as follows.

The posterior distribution for $\boldsymbol{\theta}$ given $\boldsymbol{s}$ and q is easily obtained as the Dirichlet structure of the prior distribution is conjugate to the multinomial form of the likelihood (Eq. (1)). Using Bayes' Theorem, it can be shown that this posterior distribution has independent components

$$\boldsymbol{p}_i^{(k)}|q,\boldsymbol{s},\boldsymbol{y} \sim \mathbf{D}\left(\boldsymbol{c}_i^{(k)} + \boldsymbol{n}_i^{(k)}\right), \quad i \in \mathcal{Y}_q;\ k \in \mathcal{S} \tag{2}$$

$$\boldsymbol{\lambda}_i|q,\boldsymbol{s},\boldsymbol{y} \sim \mathbf{D}(\boldsymbol{d}_i + \boldsymbol{m}_i), \quad i \in \mathcal{S}, \tag{3}$$

where $\boldsymbol{n}_i^{(k)} = \left(n_{ij}^{(k)}\right)$ and $\boldsymbol{m}_i = (m_{ij})$.

Inferences about the order of Markov dependence q are based on the posterior distribution of q given $\boldsymbol{s}$

$$\pi(q|\boldsymbol{s},\boldsymbol{y}) = \frac{\pi(q)\pi(\boldsymbol{y}|q,\boldsymbol{s})}{\pi(\boldsymbol{y}|\boldsymbol{s})} = \frac{\pi(q)\pi(\boldsymbol{y}|q,\boldsymbol{s})}{\sum_{q\in\mathcal{Q}}\pi(q)\pi(\boldsymbol{y}|q,\boldsymbol{s})}. \tag{4}$$

In general, computation of the marginal likelihood $\pi(\boldsymbol{y}|q,\boldsymbol{s})$ can be problematic and often is intractable in Bayesian model choice problems such as this. However, the conjugate choice of prior distribution for $\mathcal{P}$ allows us to determine the marginal likelihood using a simple rearrangement of Bayes' Theorem:

$$\pi(\boldsymbol{y}|q,\boldsymbol{s}) = \frac{\pi(\mathcal{P}|q,\boldsymbol{s})\pi(\boldsymbol{y}|\mathcal{P},q,\boldsymbol{s})}{\pi(\mathcal{P}|q,\boldsymbol{s},\boldsymbol{y})}.$$

Substituting the constituent parts produces an exact expression for the marginal likelihood, namely

$$\pi(\boldsymbol{y}|q,\boldsymbol{s}) = \prod_{k=1}^{r}\prod_{i=1}^{b^q} \left[\frac{\Gamma\left(\sum_{j=1}^{b} c_{ij}^{(k)}\right)\prod_{j=1}^{b}\Gamma\left(c_{ij}^{(k)}+n_{ij}^{(k)}\right)}{\prod_{j=1}^{b}\Gamma\left(c_{ij}^{(k)}\right)\Gamma\left\{\sum_{j=1}^{b}\left(c_{ij}^{(k)}+n_{ij}^{(k)}\right)\right\}}\right].$$

This expression is easy to compute and therefore exact calculation of posterior model probabilities/Bayes factors is straightforward. We note that when $r = 1$, this marginal likelihood calculation correctly reproduces the result in [14].

The simplicity of the marginal likelihood calculation is due to the choice of Dirichlet distribution for the prior distribution of the transition probabilities $\mathcal{P}$. There are many other possible choices of prior distribution which allow a more flexible covariance structure than the Dirichlet but, in general, these choices introduce additional complexity into the analysis. Even when using a different conjugate distribution, the Aitchison $\mathbf{A}$ – distribution, no exact expression for the marginal likelihood can be found as the normalising constant for this distribution is algebraically intractable. However, hierarchical generalisations of the Dirichlet distribution, such as a finite mixture or placing a hyper-prior distribution on the Dirichlet parameters, also inherit the simplicity of the marginal likelihood calculation. Both generalisations allow a more general covariance specification for the transition probabilities but would require additional updates on the mixture or hyper-prior parameters.

3.3.1. Posterior inference via MCMC

We have seen that determining posterior distributions is straightforward and exact when the hidden states are assumed known. However, in our model the hidden states are unknown quantities and so we use Markov chain Monte Carlo (MCMC) methods to allow for this uncertainty. Specifically we employ standard Gibbs sampling (data augmentation) procedures for hidden Markov models [11,27]. For our analysis, the MCMC algorithm has two parameter blocks $(q,\boldsymbol{\theta})$ and $\boldsymbol{s}$ in which we simulate from the conditional distributions $\pi(q,\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{y})$ and $\pi(\boldsymbol{s}|q,\boldsymbol{\theta},\boldsymbol{y})$.

In the second block, a sequence of hidden states $\boldsymbol{s}$ is generated from the conditional distribution $\pi(\boldsymbol{s}|q,\boldsymbol{\theta},\boldsymbol{y})$ using a standard forward-backward simulation algorithm. Algorithms of this type originated with the work of [4] and many variants are possible; the algo-

The algorithm has two steps:

1. For $t = q_{\max}+1, q_{\max}+2, \ldots, n$ calculate the 'filtered' probabilities

$$\Pr(S_t = k | y_{1:t}, q, \boldsymbol{\theta}) = \alpha_t(k) = \left\{ \sum_{j \in \mathcal{S}} \lambda_{jk} \alpha_{t-1}(j) \right\} \rho^{(k)}_{y_{t-q:t}} / \xi_t, \quad j, k \in \mathcal{S}$$

where ξ_t is a normalising constant which ensures $\sum_{k \in \mathcal{S}} \alpha_t(k) = 1$.

2. Then, for $t = n, n-1, \ldots, q_{\max}+1$ compute

$$\Pr(S_t = j | s_{t+1:n}, \boldsymbol{y}, q, \boldsymbol{\theta}) = \beta_t(j) = \lambda_{j s_{t+1}} \alpha_t(j) / \eta_t, \quad j \in \mathcal{S}, \tag{5}$$

and at each stage simulate s_t from $\beta_t(\cdot)$.

Fig. 1. Forward-backward algorithm.

Initialise the algorithm with a sequence of hidden states $\boldsymbol{s}^{(0)}$. Then at each iteration $i = 1, 2, \ldots$ perform a fixed sweep of the following steps:

1. simulate $q^{(i)}$ from $\pi(q | \boldsymbol{s}^{(i-1)}, \boldsymbol{y})$;
2. simulate $\boldsymbol{\theta}^{(i)}$ from $\pi(\boldsymbol{\theta} | q^{(i)}, \boldsymbol{s}^{(i-1)}, \boldsymbol{y})$;
3. simulate $\boldsymbol{s}^{(i)}$ from $\pi(\boldsymbol{s} | q^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{y})$.

Fig. 2. MCMC algorithm.

rithm we use is outlined in Fig. 1. We note that the forward sweep is initialised at $t^* = q_{\max} + 1$ with $\alpha_{t^*}(k) = \pi_k \rho^{(k)}_{y_{1:t^*}} / \xi_{t^*}$ where π_k is the stationary probability that $S_{t^*} = k$. In Eq. (5), the scale factor η_t ensures $\beta_t(\cdot)$ is a valid probability distribution; also we use the convention $\lambda_{j s_{n+1}} \equiv 1$.

The overall structure of our MCMC algorithm is outlined in Fig. 2. The joint $(q, \boldsymbol{\theta})$ move is undertaken in steps 1 and 2 using Eqs (2), (3) and (4) and the $\boldsymbol{s}$ move, in step 3, using the algorithm in Fig. 1. We note that this two block scheme should have better convergence properties than a standard three block scheme.

In general, particular care must be taken in the construction of MCMC schemes which incorporate a dimensional parameter, such as the order of Markov dependence q, to ensure that they converge to the correct distribution. The scheme described above does satisfy the necessary convergence conditions since q is simulated from a distribution which is marginalised over the $\boldsymbol{\theta}$ parameters. An alternative verification can be obtained using the pseudo-priors approach suggested by [10] and subsequently modified by [18]. Briefly, this technique provides a convergent scheme in which the dimension parameter is simulated conditionally on the other model parameters, that is, from $\pi(q | \boldsymbol{\theta}, \boldsymbol{s}, \boldsymbol{y})$. This distribution reduces to that in Eq. (4) when the pseudo-priors are chosen appropriately (see [16]). If a more complex prior distribution were thought to be appropriate, updates for the dimension parameter q could be obtained using a different choice of the pseudo-priors or by using reversible jump MCMC techniques [17]. Investigation of this strategy is the subject of on-going work.

3.4. Posterior summaries

Suppose the MCMC algorithm is run until it is thought that convergence has been achieved (the burn-in period) and then for a further N iterations, giving sampled values $(q^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{s}^{(i)})$, $i = 1, 2, \ldots, N$ on which to base our posterior summaries. We can estimate the (marginal) posterior distribution for the order of dependence parameter q using the sampled values $q^{(i)}$ by

$$\widehat{\pi}(q = j | \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}(q^{(i)} = j), \quad j \in \mathcal{Q}. \tag{6}$$

Alternatively, the Rao–Blackwellized estimate [15]

$$\widehat{\pi}_{RB}(q=j|\boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N}\pi(q=j|\boldsymbol{s}^{(i)},\boldsymbol{y}), \quad (7)$$

$$j \in \mathcal{Q}$$

will give a more precise estimate at the expense of additional computing effort.

The posterior distributions of the hidden states $\boldsymbol{s}$ and the model parameters $\boldsymbol{\theta}$ conditional on q are also of interest. However, summarising these distributions is complicated as the parameters are not identifiable. This non-identifiability is caused by the fact that the likelihood is invariant to permutations of the hidden state labels, that is

$$\prod_t \rho^{(s_t)}_{y_{t-q:t}} \lambda_{s_{t-1}s_t} = \prod_t \rho^{(\nu(s_t))}_{y_{t-q:t}} \lambda_{\nu(s_{t-1})\nu(s_t)}$$

for any permutation $\nu(\cdot)$ of the integers $\{1,2,\ldots,r\}$. Consequently, the likelihood has $r!$ symmetric modes corresponding to the permutations of the labels. Combining this likelihood with a symmetric prior distribution (as suggested in Section 3.1) will produce a posterior distribution which also possesses this symmetry and thus the parameters will not be identifiable.

A natural consequence of the symmetry in the posterior distribution is that our posterior sample is subject to label switching in which the hidden state labels randomly permute during the course of the MCMC run; see [29] for a detailed description. As each of the $r!$ label permutations will appear (theoretically) equally often in the posterior sample, naïve summaries which ignore label switching (such as posterior means) will lead to similar values for each of the $k = 1,2,\ldots,r$ hidden states. One solution to this problem is to impose some ordering constraint on the transition parameters for the observed sequence $P^{(k)}$ (for example, using the Fröbenius norm) in order to encourage the algorithm to focus on one of the $r!$ symmetric modes in the posterior distribution. An alternative solution which focuses on the hidden states is to post-process the MCMC output using a relabelling algorithm formulated in the decision-theoretic framework of [29]. The aim of such algorithms is to determine the permutation of $(1,2,\ldots,r)$ (and a relabelling of the sampled values) which minimises posterior expected loss (Monte Carlo risk) for some chosen loss function.

We shall adopt this second alternative and post-process the output using an algorithm whose goal is to obtain the most likely hidden state at each position in the sequence, that is, the marginal posterior mode (MPM) estimate $\hat{\boldsymbol{s}}$. Because of computing storage limitations, we advocate the use of an on-line algorithm as outlined in Fig. 3; the corresponding batch algorithm can be derived easily from this on-line version.

Two ways have been suggested by [29] to obtain a suitable initial best estimate $\hat{\boldsymbol{s}}^{\star(0)}$ which rely on running an initial sample (after a burn-in period). However, in extensive testing we have found that taking $\hat{\boldsymbol{s}}^{\star(1)} = \boldsymbol{s}^{(1)}$ and starting the algorithm at iteration $i = 2$ works well as the algorithm is fairly robust to the choice of starting point. Another advantage of using an algorithm which relabels according to the hidden states rather than the parameters is that its run time does not depend on q.

In the next section we apply our methods to the analysis of both simulated and real data sets. We show how inferences can be made for the order of dependence and use relabelled MCMC output to make inferences about the parameters and the hidden states.

4. Implementation of the algorithm

4.1. Simulated data

We begin by analysing a simulated sequence of length $n = 1000$. The sequence was generated from a hidden Markov model with $r = 2$ hidden states and a $q = 1$ order Markov dependence for a $b = 4$ state observed sequence. The transition matrices $P^{(1)}$ and $P^{(2)}$ were chosen to have roughly similar columns (within each matrix). This ensures that the $q = 0$ and $q = 1$ models are fairly close and so gives the algorithm a reasonable challenge in deciding between them. We give below the transition matrices used for simulating the sequence in order to judge the performance of the estimation procedure:

$$\Lambda = \begin{pmatrix} 0.995 & 0.005 \\ 0.010 & 0.990 \end{pmatrix},$$

$$P^{(1)} = \begin{pmatrix} 0.20 & 0.30 & 0.30 & 0.20 \\ 0.22 & 0.38 & 0.07 & 0.33 \\ 0.23 & 0.27 & 0.32 & 0.18 \\ 0.19 & 0.31 & 0.29 & 0.21 \end{pmatrix}, \quad (8)$$

$$P^{(2)} = \begin{pmatrix} 0.35 & 0.15 & 0.15 & 0.35 \\ 0.37 & 0.13 & 0.13 & 0.37 \\ 0.32 & 0.18 & 0.10 & 0.40 \\ 0.35 & 0.20 & 0.20 & 0.25 \end{pmatrix}.$$

4.1.1. Choice of prior distributions

We restrict our attention to considering dependence structures of order no more than $q_{\max} = 3$. Also we adopt a truncated Poisson distribution to describe our

If at iteration i the current estimate of $\hat{s}^{\star}$ is $\hat{s}^{\star(i-1)}$ then,

1. choose ν_i to minimise

$$-\sum_{t=q_{\max}+1}^{n} \mathbb{I}\left(\nu_i(s_t^{(i)}) = \hat{s}_t^{\star(i-1)}\right);$$

2. apply permutation ν_i to output $s^{(i)}$ and $\boldsymbol{\theta}^{(i)}$;

3. for $t = q_{\max}+1, 2, \ldots, n$, set

$$\hat{s}_t^{\star(i)} = \underset{j\in\mathcal{S}}{\operatorname{argmax}} \sum_{k=1}^{i} \mathbb{I}\left(\nu_k(s_t^{(k)}) = j\right).$$

Fig. 3. Relabelling algorithm.

prior uncertainty about q with

$$\pi(q) = \mathcal{P}r(q = i) \propto a^i/i!, \quad i = 0, 1, 2, 3.$$

As we are particularly interested in whether the algorithm can choose between $q = 0$ or $q = 1$, we take $a = 1$ as the hyperparameter of this distribution. We will see later that the results are fairly robust to changes in this choice of prior. We also take the weak specification for the transition probabilities in $P^{(1)}$ and $P^{(2)}$, that is $\boldsymbol{c}_i^{(k)} = (1, 1, 1, 1)$ for $i \in \mathcal{Y}_q$ and $k = 1, 2$. For the hidden state transition matrix we take $d_{11} = d_{22} = 19$ and $d_{12} = d_{21} = 1$; this is equivalent to the information content of a sequence of length 40 with an expected run length of 20 for both hidden states.

4.1.2. Results

The MCMC algorithm was run for 110000 iterations with the first 10000 being discarded as burn-in. Our results are based on a sample of size $N = 10000$ since we only recorded every 10th iterate to reduce computing overheads. The usual diagnostic checks were made to ensure there was no evidence of lack of convergence. We also confirmed our results using several MCMC runs from different starting points.

Table 1 contains estimates of the marginal posterior distribution for q based on the sampled values of q over the iterations of the sampler (Eq. (6)) together with the alternative Rao–Blackwellized estimate (Eq. (7)). It shows a very high probability for the correct choice $q = 1$. The probability of higher values of q is very low. Repeating the analysis using a uniform prior distribution for q gives similar results. The effect of using other prior distributions can be seen by reweighting these posterior probabilities according to the ratio of the proposed and actual prior probabilities. Such considerations show that the prior odds in favour of $q = 0$ (against $q = 1$) would have to be at least $1500 : 1$ before the analysis favoured the incorrect choice of $q = 0$.

Table 1
Estimates of the marginal posterior distribution for q: simulated sequence $n = 1000$

		Order q			
Prior	Estimate	0	1	2	3
Poisson	$\widehat{\pi}(q\|\mathbf{y})$	0.0005	0.9995	0.0000	0.0000
Poisson	$\widehat{\pi}_{RB}(q\|\mathbf{y})$	0.0006	0.9994	0.0000	0.0000
Uniform	$\widehat{\pi}_{RB}(q\|\mathbf{y})$	0.0006	0.9994	0.0000	0.0000

We now present results for the hidden states $\boldsymbol{s}$ and for the transition parameters $\boldsymbol{\theta}$ conditional on the (posterior) modal value $q = 1$. In general we can estimate the posterior probabilities of the hidden states S_t along the sequence (conditional on a particular $q = q^*$) by

$$\widehat{\mathcal{P}r}(S_t = j|q = q^*, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}(s_t^{(i)} = j, q = q^*),$$

or by its Rao-Blackwellized equivalent. Figure 4 shows the estimate of the probability of hidden state 1 along the sequence. It also indicates the positions of the actual hidden states from which the sequence was simulated. Clearly, the algorithm has uncovered this latent structure very well.

Table 2 contains the posterior means and standard deviations for the model transition probabilities $\boldsymbol{\theta}$. We note that the values shown are not too dissimilar to the values from which the data were simulated and well within sampling error.

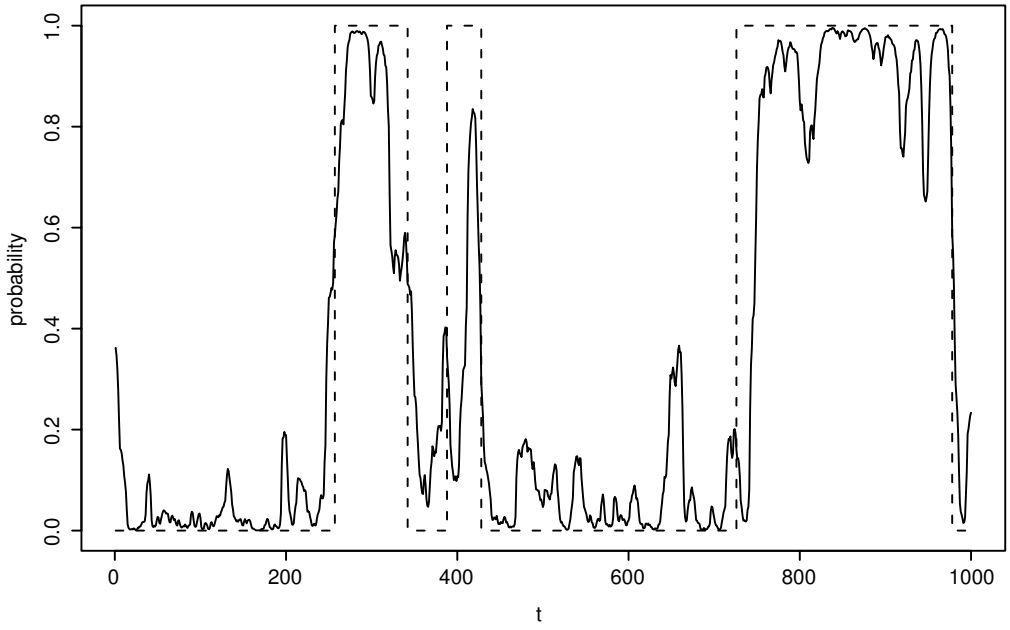

Fig. 4. Estimated posterior probabilities $\widehat{\mathcal{P}r}(S_t = 1|q = 1, \mathbf{y})$ (solid line) together with the true positions $\mathbf{I}(S_t = 1)$ (dashed line): simulated sequence $n = 1000$.

Table 2
Posterior summaries for transition matrices conditional on $q = 1$: simulated sequence $n = 1000$

	means	standard deviations
Λ:	$\begin{pmatrix} 0.978 & 0.022 \\ 0.035 & 0.965 \end{pmatrix}$	$\begin{pmatrix} 0.008 & 0.008 \\ 0.012 & 0.012 \end{pmatrix}$
$P^{(1)}$:	$\begin{pmatrix} 0.217 & 0.328 & 0.311 & 0.144 \\ 0.215 & 0.362 & 0.095 & 0.328 \\ 0.220 & 0.316 & 0.309 & 0.155 \\ 0.160 & 0.330 & 0.293 & 0.217 \end{pmatrix}$	$\begin{pmatrix} 0.038 & 0.043 & 0.044 & 0.033 \\ 0.030 & 0.036 & 0.021 & 0.035 \\ 0.036 & 0.041 & 0.040 & 0.038 \\ 0.032 & 0.042 & 0.043 & 0.038 \end{pmatrix}$
$P^{(2)}$:	$\begin{pmatrix} 0.347 & 0.091 & 0.244 & 0.318 \\ 0.400 & 0.085 & 0.166 & 0.349 \\ 0.225 & 0.122 & 0.093 & 0.561 \\ 0.250 & 0.274 & 0.171 & 0.304 \end{pmatrix}$	$\begin{pmatrix} 0.050 & 0.038 & 0.046 & 0.050 \\ 0.073 & 0.045 & 0.056 & 0.076 \\ 0.064 & 0.052 & 0.043 & 0.075 \\ 0.043 & 0.045 & 0.042 & 0.044 \end{pmatrix}$

4.1.3. The effect of sequence length

We now illustrate the effect of sequence length n on the ability to detect the correct order of dependence q in the sequence. Clearly, longer sequences will reduce uncertainty about q, but to what extent? We now investigate what conclusions can be drawn about the model parameters using only the first half of the earlier sequence ($n = 500$). We have retained the same prior distributions and proceeded with a MCMC algorithm as before. The MCMC algorithm produced a well-mixing chain which traversed the different models (corresponding to the different values of q) regularly, with the value of q changing on approximately 38% of the iterations.

The impact of using a much shorter sequence on the marginal posterior of q is clearly seen in Table 3. Again the various choices of estimate and prior distribution produce very similar results. However, for this shorter sequence, there is considerably more uncertainty surrounding the value of q. As before, $q = 0$ and $q = 1$ receive nearly all of the posterior support with values of $q = 2$ or $q = 3$ highly unlikely. However, there is a high degree of uncertainty about the order of dependence in the data. There needs only to be a shift in prior odds of around $3 : 2$ before the incorrect choice of $q = 0$ is favoured.

This shorter sequence also makes it much more difficult to identify the hidden states. Figure 5 shows the

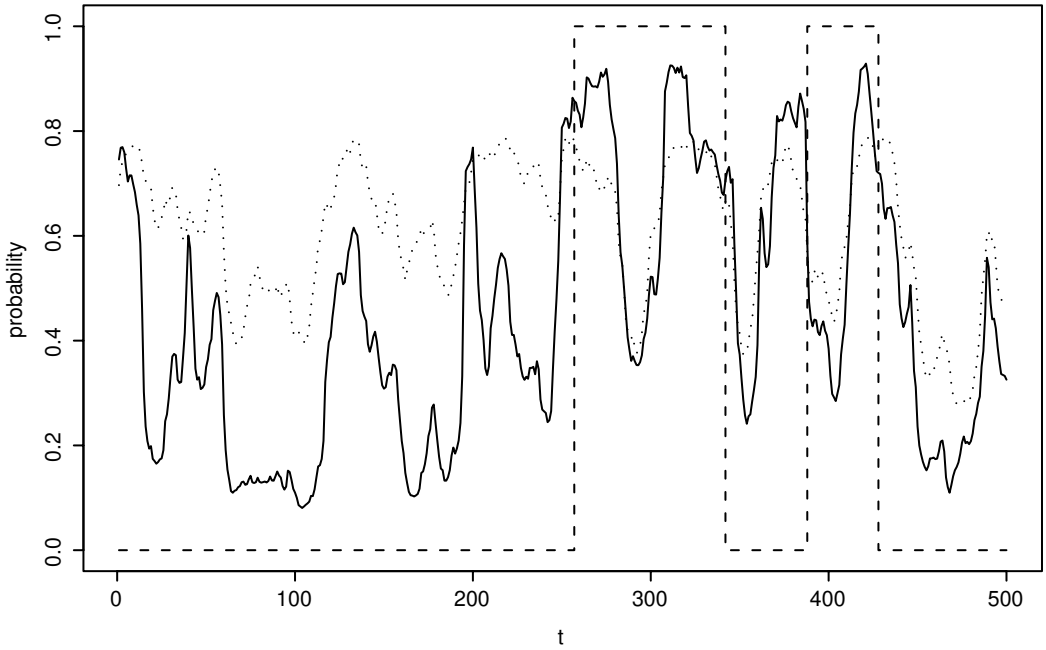

Fig. 5. Estimated posterior probabilities $\widehat{Pr}(S_t = 1|q = 1, \mathbf{y})$ (solid line), $\widehat{Pr}(S_t = 1|q = 0, \mathbf{y})$ (dotted line) together with the true positions $\mathbf{I}(S_t = 1)$ (dashed line): simulated sequence $n = 500$.

Table 3
Estimates of the marginal posterior distribution for q: simulated sequence $n = 500$

		Order q			
Prior	Estimate	0	1	2	3
Poisson	$\widehat{\pi}(q|\mathbf{y})$	0.4007	0.5880	0.0052	0.0061
Poisson	$\widehat{\pi}_{RB}(q|\mathbf{y})$	0.4011	0.5879	0.0057	0.0053
Uniform	$\widehat{\pi}_{RB}(q|\mathbf{y})$	0.4005	0.5855	0.0071	0.0069

plot of estimated (posterior) probabilities for both $q = 0$ and $q = 1$ together with the true sequence. Clearly, the $q = 1$ plot better describes the actual sequence but nevertheless its predictive capacity is much reduced when compared to that obtained using the full sequence. The posterior means and standard deviations for the model transition probabilities $\boldsymbol{\theta}$ (assuming $q = 1$) are given in Table 4. Again these values are consistent with the model parameters in Eq. (8). However, the standard deviations are significantly increased and much more than would be expected due to halving the sequence length. This is due to uncertainty about the hidden states.

4.2. DNA sequence data

For our final example we apply the methods to the analysis of a DNA sequence. These sequences comprise a string of $b = 4$ states (bases) from the alphabet $\mathcal{Y} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. HMMs have been used for some time to model heterogeneity in the composition of DNA sequences with most analyses assuming that $q = 0$, that is, the observed sequence is independent conditional on the hidden states. However, empirical evidence would suggest that this independence model is not sufficiently complex to capture the rich dependence structure in these sequences. In such circumstances, the methods described in this paper can permit inferences to be made about the order of Markov dependence of the bases, thereby ensuring appropriate conclusions are drawn. Some authors have analysed DNA sequences using larger values of q but these analyses have assumed a known fixed value of q (for example [7, 23]) or attempted to choose between various q using information criteria (for example [12]).

We will study the 7th intron of the human α-fetoprotein (AFP) gene which was analysed in [7] using a hidden Markov model. The AFP gene is known be an important factor in embryonic development in mammals and is also thought to play a role in the development of tumors; for further details, see [7]. The intron is $n = 2275$ base pairs in length and is stored in the GenBank sequence database [5] under Accession No. M16110. It can be obtained from the National Center for Biotechnology Information (NCBI) web pages at http://www.ncbi.nlm.nih.gov/.

Table 4
Posterior summaries for transition matrices conditional on $q = 1$: simulated sequence $n = 500$

	means	standard deviations
Λ:	$\begin{pmatrix} 0.952 & 0.048 \\ 0.056 & 0.944 \end{pmatrix}$	$\begin{pmatrix} 0.018 & 0.018 \\ 0.019 & 0.019 \end{pmatrix}$
$P^{(1)}$:	$\begin{pmatrix} 0.281 & 0.258 & 0.308 & 0.153 \\ 0.255 & 0.402 & 0.084 & 0.259 \\ 0.304 & 0.371 & 0.246 & 0.079 \\ 0.197 & 0.318 & 0.321 & 0.164 \end{pmatrix}$	$\begin{pmatrix} 0.081 & 0.078 & 0.078 & 0.055 \\ 0.073 & 0.087 & 0.037 & 0.098 \\ 0.095 & 0.103 & 0.079 & 0.067 \\ 0.081 & 0.102 & 0.102 & 0.076 \end{pmatrix}$
$P^{(2)}$:	$\begin{pmatrix} 0.194 & 0.317 & 0.292 & 0.197 \\ 0.251 & 0.191 & 0.140 & 0.418 \\ 0.175 & 0.202 & 0.170 & 0.453 \\ 0.173 & 0.361 & 0.178 & 0.288 \end{pmatrix}$	$\begin{pmatrix} 0.101 & 0.105 & 0.110 & 0.080 \\ 0.088 & 0.099 & 0.066 & 0.118 \\ 0.118 & 0.102 & 0.092 & 0.167 \\ 0.059 & 0.077 & 0.065 & 0.068 \end{pmatrix}$

Table 5
Estimates of the marginal posterior distribution for q: intron 7 of the human AFP gene conditional on $r = 3$

	Order q			
Estimate	0	1	2	3
$\widehat{\pi}(q\|\mathbf{y})$	0	1	0	0
$\widehat{\pi}_{RB}(q\|\mathbf{y})$	$\simeq 10^{-14}$	$\simeq 1$	$\simeq 10^{-18}$	$\simeq 10^{-51}$

4.2.1. Results

For comparison with the results in [7] we have run the algorithm assuming $r = 3$ hidden states. We chose the maximum complexity of base updates to be order $q_{\max} = 3$ and used the same truncated Poisson prior distribution as in the analysis of the simulated data. We also chose the same prior distribution for the observed state transition matrices but for the hidden state transition matrix we chose $d_{ij} = 99$ for $i = j$ and $d_{ij} = 0.5$ otherwise; this is equivalent to the information content of a sequence of length 300 with an expected run length of 100 for each hidden state.

The MCMC algorithm was again run for 110000 iterations with the first 10000 being discarded as burn-in. The usual checks for evidence on lack of convergence were made using multiple runs from different starting points, and our results are based on a sample of size $N = 10000$ from one such run, recording every 10th iterate. Table 5 contains the sample and Rao-Blackwellized estimates of the marginal posterior distribution for q. It shows that, after convergence, the sampler never moved from the model with $q = 1$ during the course of the simulation. Ordinarily, this may point to the MCMC scheme suffering from poor mixing over q, possibly due to the update conditioning on the hidden states $\boldsymbol{s}$. However, our experience with simulated sequences suggests that this is in fact due to the DNA sequence being sufficiently long to provide overwhelming evidence that $q = 1$.

These results show that the choice of $q = 1$ employed by [7] is well justified.

5. Conclusions

We have seen how inferences can be made about the order of Markov dependence of an observed process governed by a HMM. In many practical examples, the sequence will be sufficiently long that by using these methods it will be straightforward to determine this order, particularly if the transition structures in each hidden state are reasonably different. When sequences are short or the transition structures fairly similar, it can be difficult to determine an appropriate order of dependence. In such circumstances it is important to be able to correctly assess uncertainty about the order q and also the associated transition structures and the pattern of hidden states. The methods presented in this paper do provide this information by adopting a fully Bayesian approach through the use of MCMC techniques.

Acknowledgements

This work was carried out during DAH's tenure of a Lord Adams Fellowship at the University of Newcastle. The authors would like to thank the editor and three referees for their helpful comments.

References

[1] M. Abramowitz and I.A. Stegun, *Pocketbook of Mathematical Functions*, Deutsch, 1984,
[2] J. Aitchison, *The Statistical Analysis of Compositional Data,* Chapman and Hall, London, 1986.

[3] M.S. Bartlett, The frequency goodness of fit test for probability chains, *Proceedings of the Cambridge Philosophical Society* **47** (1951), 86–95.
[4] L.E. Baum, T. Petrie, G. Soules and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics* **41** (1970), 164–171.
[5] D.A. Benson, I. Karsh-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp and D.L. Wheeler, GenBank, *Nucleic Acids Research* **28** (2000), 15–18.
[6] A. Berchtold, The double chain Markov model, *Communications in Statistics – Theory and Methods* **28** (1999) 2569–2589.
[7] R.J. Boys, D.A. Henderson and D.J. Wilkinson, Detecting homogeneous segments in DNA sequences by using hidden Markov models, *Applied Statistics* **49** (2000), 269–285.
[8] S.P. Brooks, Markov chain Monte Carlo method and its application, *Statistician* **47** (1998), 69–100.
[9] O. Cappé, *Ten years of HMMs*, On-line bibliography available from http://tsi.enst.fr/˜cappe/docs/hmmbib.html, 2001.
[10] B.P. Carlin and S. Chib, Bayesian model choice via Markov chain Monte Carlo, *Journal of the Royal Statistical Society* **B 57**(3) (1995), 473–484.
[11] S. Chib, Calculating posterior distributions and modal estimates in Markov mixture models, *Journal of Econometrics* **75** (1996), 79–97.
[12] G.A. Churchill, Stochastic models for heterogeneous DNA sequences, *Bulletin of Mathematical Biology* **51** (1989), 79–94.
[13] J.B. Copas, Regression, prediction and shrinkage, *Journal of the Royal Statistical Society* **B 45** (1983), 311–354.
[14] T.-H. Fan and C.-A. Tsai, A Bayesian method in determining the order of a finite state Markov chain, *Communications in Statistics – Theory and Methods* **28** (1999), 1711–1730.
[15] A.E. Gelfand and A.F.M. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85** (1990), 398–409.
[16] S.J. Godsill, On the relationship between Markov chain Monte Carlo methods for model uncertainty, *Journal of Computational and Graphical Statistics* **10** (2001), 230–248.
[17] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82** (1995), 711–732.
[18] P.J. Green and A. O'Hagan, Model choice with MCMC on product spaces without using pseudo-priors, Nottingham University, Statistics Research Report 98-01, 1998.
[19] P.G. Hoel, A test for Markov chains, *Biometrika* **41** (1954), 430–433.
[20] R.E. Kass and A.E. Raftery, Bayes Factors, *Journal of the American Statistical Association* **90** (1995), 773–795.
[21] R.W. Katz, On some criteria for estimating the order of a Markov chain, *Technometrics* **23** (1981), 243–249.
[22] I.L. MacDonald and W. Zucchini, Hidden Markov and Other Models for Discrete-valued Time Series, Chapman and Hall, London, 1997.
[23] F. Muri, Modelling bacterial genomes using hidden Markov models, in: *COMPSTAT '98 Proceedings in Computational Statistics,* R.W. Payne and P.J. Green, eds, Physica-Verlag, Heidelberg, 1998, pp. 89–100.
[24] A. O'Hagan, *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*, Arnold, 1994.
[25] G.G.S. Pegram, An autoregressive model for multilag Markov chains, *Journal of Applied Probability* **17** (1980), 350–362.
[26] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77** (1989), 257–286.
[27] C.P. Robert, G. Celeux and J. Diebolt, Bayesian estimation of hidden Markov chains: A stochastic implementation, *Statistics and Probability Letters* **16** (1993), 77–83.
[28] C.P. Robert, T. Rydén and D.M. Titterington, Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method, *Journal of the Royal Statistical Society* **B 62** (2000), 57–75.
[29] M. Stephens, Dealing with label switching in mixture models, *Journal of the Royal Statistical Society* **B 62** (2000), 795–809.
[30] H. Tong, Determination of the order of a Markov chain by Akaike's information criterion, *Journal of Applied Probability* **12** (1975), 488–497.