Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 3, 653–665
DOI: 10.1214/16-BJPS328

Second-order autoregressive Hidden Markov Model

Daiane Aparecida Zuanetti and Luis Aparecido Milan

Departamento de Estatística, UFSCar—Brazil

Abstract. We propose an extension of Hidden Markov Model (HMM) to support second-order Markov dependence in the observable random process. We propose a Bayesian method to estimate the parameters of the model and the non-observable sequence of states. We compare and select the best model, including the dependence order and number of states, using model selection criteria like Bayes factor and deviance information criterion (DIC). We apply the procedure to several simulated datasets and verify the good performance of the estimation procedure. Tests with a real dataset show an improved fitting when compared with usual first order HMMs demonstrating the usefulness of the proposed model.

1 Introduction

Hidden Markov models have been applied in many research areas. A few examples are econometrics (Hamilton, 1989, Chib, 1996, Krolzig, 1997, Biblio, Monfort and Robert, 1999), finance (Rydén, Teräsvirta and Asbrink, 1998) and speech recognition (Rabiner, 1989). A growth in the amount of DNA sequence available for analysis and an increasing need to develop efficient computational techniques and statistics to analyze these biological data transformed HMM into an interesting method to analyze DNA sequence. HMM has been used, successively, by Churchill (1989, 1992), Muri (1998) and Boys et al. (2000, 2002, 2004).

This model is useful in molecular biology and genetic, particularly, to find genes, introns, exons, etc, in the DNA sequence or to detect and align remotely homologous sequences which provide information about the protein's function, structure or evolution (Gough et al., 2001, Söding, 2005, Leea et al., 2009).

Baum and Petrie (1966) and Baum et al. (1970) propose maximum likelihood estimators for the case where the number of non-observable states, N, is known. These estimators are obtained using expectation-maximization (EM) algorithm. When N is unknown, estimators are based on model selection methods such as likelihood ratio test, Akaike (AIC) and Bayesian information criteria (BIC). References about these tests are McLachlan (1987), Rydén, Teräsvirta and Asbrink (1998) and Gassiat and Kéribin (2000). Schimert (1992), du Preez (1998), Hadar

Key words and phrases. Hidden Markov model, second-order dependence, Markov chain Monte Carlo (MCMC), gene modeling, bacteriophage *lambda* genome.

Received February 2015; accepted June 2016.

and Messer (2009) and Seifert (2010) describe extensions of the mathematical theory of first-order HMMs to higher-order HMMs. One disadvantage of the maximum likelihood estimators is that they require many training sequences since the autoregressive higher-order HMM can have a large number of parameters, depending on N and dependence order of sequences, and a small dataset could be insufficient to estimate then properly.

The Bayesian approach allows to include *a priori* information about the uncertainty of the parameters and it may help to improve the convergence of the method, especially for complex models. Robert, Celeux and Diebolt (1993), Chib (1996) and Robert and Titterington (1998) propose Bayesian estimators for HMMs with known number of non-observable states. Seifert et al. (2012, 2014) propose the Bayesian Baum-Welch algorithm to estimate higher-order HMM when the observable states have Gaussian distribution. The method combines EM algorithm with *a priori* knowledge of the parameters and locally maximizes the log *a posteriori* distribution by a two-step procedure. Seifert et al. (2014) also include a autoregressive dependence in observable sequence and define the average of the Gaussian distributions as linear combinations of predecessors observations.

In cases where the non-observable states are continuous, particle filter or Kalman filter are used to compute marginal likelihood and characterize the distribution of the states. A review of this topic and applications may be found at Martino et al. (2015), Doucet et al. (2001), Doucet and Johansen (2009), Ristic, Arulampalam and Gordon (2004) and Djurić et al. (2003).

We consider a second-order autoregressive hidden Markov model and propose a Bayesian method to estimate its parameters and the non-observable sequence of states. We propose this model since in the usual model the non-A review of this topic and applications may be found at Martino et al. (2015), Doucet et al. (2001), Doucet and Johansen (2009), Ristic, Arulampalam and Gordon (2004) and Djurić et al. (2003) observable states are a first-order Markov chain and the observable states are conditionally independent (given the non-observable states) and there is no biological reason to restrict ourselves to such a strong supposition (conditional independence). Also, the genetic DNA code is translated to proteins using triplets of nitrogenous bases, the amino acids, suggesting that the second-order dependence in the observable process may fits better than the independence model when analyzing a DNA sequence. Finding homogeneous segments in DNA sequence is our main application. We compare and select the best model, including the dependence order and number of states, using model selection criteria like Bayes factor and deviance information criterion (DIC). Compared with Bayesian EM Baum-Welch algorithm, the proposed Markov chain Monte Carlo (MCMC) method samples from the jointly *a posteriori* distribution in a single step and also allows interval estimates of parameters without using asymptotic properties. The proposed MCMC method avoids usual EM problems as sensibility to starting points and convergence local maximum.

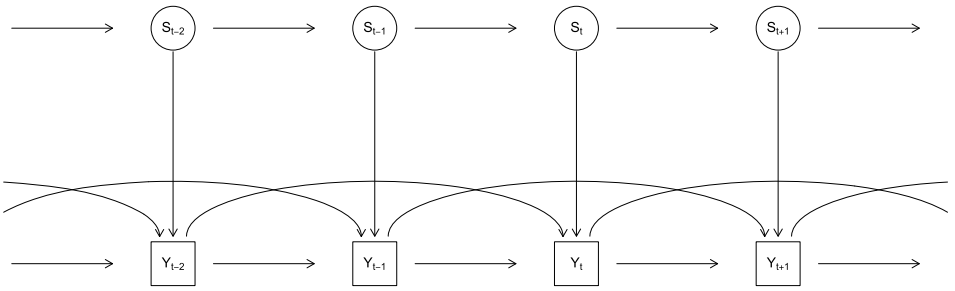

Figure 1 *Directed acyclic graph of a second-order autoregressive HMM in the observable sequence.*

The article is organized as follows. Section 2 presents the model and propose the MCMC schema to estimate the parameters and sequence of states. In Section 3, we apply the proposed model to a simulated and real datasets. We conclude with the discussion in Section 4.

2 The second-order autoregressive HMM

The proposed autoregressive HMM is composed by two random processes with length T, the non-observable first-order Markov chain $\mathbf{S} = \{S_1, S_2, \ldots, S_T\}$, where $S_t \in \{1, 2, \ldots, N\}$, for $t = 1, 2, \ldots, T$, and the observable second-order Markov chain $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_T\}$, where $Y_t|S_t, Y_{t-1}, Y_{t-2} \sim f_{Y_t|S_t,Y_{t-1},Y_{t-2}}(y_t)$, for $t = 3, 4, \ldots, T$.

The relationship between variables is specified by the conditional independence given by

$$S_t \perp \{S_1, Y_1, \ldots, S_{t-2}, Y_{t-2}, Y_{t-1}\}|S_{t-1}, \qquad \text{for } t = 2, 3, \ldots, T \quad \text{and}$$

$$Y_t \perp \{S_1, Y_1, \ldots, S_{t-3}, Y_{t-3}, S_{t-2}, S_{t-1}\}|Y_{t-2}, Y_{t-1}, S_t, \qquad \text{for } t = 3, 4, \ldots, T.$$

This relationship can be visualized by the directed acyclic graph (DAG) in Figure 1.

Assuming $Y_t \in \{1, 2, \ldots, M\}$ as discrete random variables and S_1, Y_1 and Y_2 with independent discrete uniform distributions, this HMM is specified by

1. $A = \{a_{kl}\}$, the transition matrix between non-observable states, where $a_{kl} = \Pr(S_{t+1} = l|S_t = k)$, for $k = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, N$, and
2. $\mathbf{P} = \{P^{(1)}, P^{(2)}, \ldots, P^{(N)}\}$, where $P^{(k)} = \{p^{(k)}_{hij}\}$ the transition matrix between observable states conditioned to non-observable state $S_t = k$ and $p^{(k)}_{hij} = \Pr(Y_t = j|S_t = k, Y_{t-1} = i, Y_{t-2} = h)$, for $k = 1, 2, \ldots, N$, $h = 1, \ldots, M$, $i = 1, \ldots, M$ and $j = 1, \ldots, M$.

The likelihood function for the model parameters $\mathbf{P}$ and A given both the observable sequence $\mathbf{Y} = \mathbf{y}$ and the non-observable sequence $\mathbf{S} = \mathbf{s}$ is

$$\begin{aligned} L(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) &= \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}|A, \mathbf{P}) \\ &= \Pr(S_1 = s_1) \Pr(S_2 = s_2|s_1) \prod_{t=1}^{2} \Pr(Y_t = y_t) \\ &\quad \times \left(\prod_{t=3}^{T} \Pr(S_t = s_t|s_{t-1}) \Pr(Y_t = y_t|y_{t-2}, y_{t-1}, s_t) \right) \\ &= \frac{1}{NM^2} \left(\prod_{k=1}^{N} \prod_{l=1}^{N} a_{kl}^{m_{kl}} \right) \left(\prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \prod_{j=1}^{M} (p_{hij}^{(k)})^{n_{hij}^{(k)}} \right), \end{aligned} \tag{1}$$

where $m_{kl} = \sum_{t=1}^{T-1} I(S_t = k, S_{t+1} = l)$, $n_{hij}^{(k)} = \sum_{t=3}^{T} I(Y_{t-2} = h, Y_{t-1} = i, Y_t = j, S_t = k)$ and $I(A) = 1$ if A is true or $I(A) = 0$ otherwise.

Baum and Petrie (1966) and Baum et al. (1970) propose maximum likelihood estimators for the model in which the observable sequence is independent given the non-observable sequence. These estimators and the non-observable sequence are obtained using EM algorithm. Here, as the autoregressive HMM of order two has a large number of parameters ($N(N + M^3)$ parameters), we propose Bayesian estimators for them and include *a priori* information about the uncertainty of parameters.

2.1 *A priori* distributions

In order to set up *a priori* distributions for the parameters A and $\mathbf{P}$, let $\mathbf{p}_{hi}^{(k)}$ be one row of $P^{(k)}$, $\mathbf{a}_k$ be one row of the transition matrix between non-observable states A, for $k = 1, 2, \ldots, N$, $h, i, j = 1, 2, \ldots, M$ and $\mathbf{p}_{hi}^{(k)}$'s, $\mathbf{a}_k$'s, $\mathbf{P}$ and A are supposed to be independent.

We assume M-vector $\mathbf{p}_{hi}^{(k)}$ and N-vector $\mathbf{a}_k$ has a Dirichlet distribution ($\mathcal{D}$) defined on the simplex with density given, respectively, by

$$\pi(\mathbf{p}_{hi}^{(k)}) \propto \prod_{j=1}^{M} (p_{hij}^{(k)})^{\alpha_{hij}^{(k)} - 1},$$

for $0 < p_{hij}^{(k)} < 1$, $\sum_{j=1}^{M} p_{hij}^{(k)} = 1$, $k = 1, 2, \ldots, N$, $h, i, j = 1, 2, \ldots, M$ and where $\boldsymbol{\alpha}_{hi}^{(k)} = \{\alpha_{hij}^{(k)}\}$ are positive hyperparameters of the distribution, and

$$\pi(\mathbf{a}_k) \propto \prod_{l=1}^{N} a_{kl}^{\beta_{kl} - 1},$$

for $0 < a_{kl} < 1$, $\sum_{l=1}^{n} a_{kl} = 1$, $k = 1, 2, \ldots, N$ and where $\boldsymbol{\beta}_k = \{\beta_{kl}\}$ are positive hyperparameters of the distribution. Therefore, *a priori* distribution for A and $\mathbf{P}$ is given by

$$\begin{aligned}\pi(\mathbf{P}, A) &= \pi(A)\pi(\mathbf{P}) \\ &\propto \left(\prod_{k=1}^{N} \pi(\mathbf{a}_k)\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M} \pi\left(\mathbf{p}_{hi}^{(k)}\right)\right) \\ &= \left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M} \left(p_{hij}^{(k)}\right)^{\alpha_{hij}^{(k)}-1}\right). \end{aligned} \tag{2}$$

2.2 *A posteriori* distributions

Combining the likelihood function given in (1) with the *a priori* information about A and $\mathbf{P}$ in (2) and using the Bayes theorem, the *a posteriori* distribution for A and $\mathbf{P}$ is

$$\begin{aligned}\pi(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) &\propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi(\mathbf{P}, A) \\ &= \frac{1}{NM^2}\left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{m_{kl}+\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M} \left(p_{hij}^{(k)}\right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1}\right). \end{aligned} \tag{3}$$

The conditional *a posteriori* distributions of $\mathbf{p}_{hi}^{(k)}$ and $\mathbf{a}_k$, for $h, i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$, are

$$\pi\left(\mathbf{p}_{hi}^{(k)}|\mathbf{y}, \mathbf{s}, A, \mathbf{P}_{-\mathbf{p}_{hi}^{(k)}}\right) \propto \prod_{j=1}^{M} \left(p_{hij}^{(k)}\right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1},$$

where $\mathbf{P}_{-\mathbf{p}_{hi}^{(k)}}$ denotes $\mathbf{P}$ with row $\mathbf{p}_{hi}^{(k)}$ removed and

$$\pi(\mathbf{a}_k|\mathbf{y}, \mathbf{s}, A_{-\mathbf{a}_k}, \mathbf{P}) \propto \prod_{l=1}^{N} a_{kl}^{m_{kl}+\beta_{kl}-1},$$

where $A_{-\mathbf{a}_k}$ denotes A with $\mathbf{a}_k$ removed, that is, $\mathbf{p}_{hi}^{(k)}|\mathbf{y}, \mathbf{s}, A, \mathbf{P}_{-\mathbf{p}_{hi}^{(k)}} \sim \mathcal{D}(\mathbf{n}_{hi}^{(k)} + \boldsymbol{\alpha}_{hi}^{(k)})$ for $h, i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$, $\mathbf{n}_{hi}^{(k)} = \{n_{hij}^{k}\}$ for $j = 1, 2, \ldots, M$, $\mathbf{a}_k|\mathbf{y}, \mathbf{s}, A_{-\mathbf{a}_k}, \mathbf{P} \sim \mathcal{D}(\mathbf{m}_k + \boldsymbol{\beta}_k)$ for $k = 1, 2, \ldots, N$ and $\mathbf{m}_k = \{m_{kl}\}$ for $l = 1, 2, \ldots, N$.

Therefore, A and $\mathbf{P}$ are updated through Gibbs sampling algorithm using their conditional *a posteriori* distributions described above.

As the sequence $\mathbf{S}$ is non-observable, it must be also simulated and updated in our process through its conditional *a posteriori* distribution, namely $\pi(\mathbf{S} =$

$\mathbf{s}|\mathbf{y}, A, \mathbf{P}) = \Pr(\mathbf{S} = \mathbf{s}|\mathbf{y}, A, \mathbf{P})$. An efficient simulation strategy for obtaining realizations from it is proposed by Boys, Henderson and Wilkinson (2000) for a first-order autoregressive HMM and is based on conditional independence between S_t and Y_i $(i > t)$ given S_{t+1} and $\mathbf{Y}^t = (Y_1, Y_2, \ldots, Y_t)$. Dropping the dependence on $\mathbf{P}$ and A in the notation to simplify,

$$\begin{aligned}\Pr(S_t = s_t|s_{t+1}, \mathbf{y}) &= \Pr\big(S_t = s_t|s_{t+1}, \mathbf{y}^t\big)\\ &= \frac{\Pr(S_t = s_t, S_{t+1} = s_{t+1}|\mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)}\\ &= \frac{\Pr(S_t = s_t|\mathbf{y}^t)\Pr(S_{t+1} = s_{t+1}|s_t, \mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)}\\ &= \frac{a_{s_t s_{t+1}}\Pr(S_t = s_t|\mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)}, \end{aligned} \tag{4}$$

for $t = 1, 2, \ldots, T-1$, $s_t \in \{1, 2, \ldots, N\}$ and where

$$\begin{aligned}\Pr\big(S_t = s_t|\mathbf{y}^t\big) &= \Pr\big(S_t = s_t|y_t, \mathbf{y}^{t-1}\big) = \frac{\Pr(S_t = s_t, Y_t = y_t|\mathbf{y}^{t-1})}{\Pr(Y_t = y_t|\mathbf{y}^{t-1})}\\ &\propto \Pr\big(S_t = s_t, Y_t = y_t|\mathbf{y}^{t-1}\big)\\ &= \sum_{l=1}^{N}\Pr\big(S_t = s_t, Y_t = y_t, S_{t-1} = l|\mathbf{y}^{t-1}\big)\\ &= \sum_{l=1}^{N}\begin{pmatrix}\Pr\big(S_{t-1} = l|\mathbf{y}^{t-1}\big)\Pr\big(S_t = s_t|S_{t-1} = l, \mathbf{y}^{t-1}\big)\\ \times \Pr\big(Y_t = y_t|S_{t-1} = l, s_t, \mathbf{y}^{t-1}\big)\end{pmatrix}\\ &= p^{(s_t)}_{y_{t-2}y_{t-1}y_t}\sum_{l=1}^{N}a_{ls_t}\Pr\big(S_{t-1} = l|\mathbf{y}^{t-1}\big) \quad \text{and} \end{aligned} \tag{5}$$

$$\begin{aligned}\Pr\big(S_{t+1} = s_{t+1}|\mathbf{y}^t\big) &= \sum_{k=1}^{N}\Pr\big(S_{t+1} = s_{t+1}, S_t = k|\mathbf{y}^t\big)\\ &= \sum_{k=1}^{N}\Pr\big(S_t = k|\mathbf{y}^t\big)\Pr\big(S_{t+1} = s_{t+1}|S_t = k, \mathbf{y}^t\big)\\ &= \sum_{k=1}^{N}a_{ks_{t+1}}\Pr\big(S_t = k|\mathbf{y}^t\big). \end{aligned} \tag{6}$$

Equation (5) provides a (forward) iterative scheme to evaluate $\Pr(S_t = s_t|\mathbf{y}^t)$, $t = 1, \ldots, T$. The initial distribution for the iterations is provided by the discrete uniform distribution on (Y_1, S_1) and (Y_2, S_2). Values for $\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)$ and

$\Pr(S_t = s_t | s_{t+1}, \mathbf{y})$ can then be computed by using equations (6) and (4), respectively.

A realization of the non-observable random process can be simulated. First, a value for s_T is obtained by using the distribution $\Pr(S_T = s_T | \mathbf{y}^T)$ and the remaining values are obtained by a backward process $t = T - 1, T - 2, \ldots, 1$, using equation (4).

2.3 Algorithm

The procedure to update and estimate the parameters and non-observable sequence is expressed as an algorithm.

1. Initialize a configuration $\mathbf{s}$;
2. for bth iteration, $b = 1, \ldots, B$ do:

   - sample $\mathbf{a}_k, k = 1, \ldots, N$, from its conditional *a posteriori* distribution;
   - sample $\mathbf{p}_{hi}^{(k)}, k = 1, \ldots, N$ and $i, h = 1, \ldots, M$, from its conditional *a posteriori* distribution;
   - update $\Pr(S_t = s_t | \mathbf{y}^t), t = 1, \ldots, T$;
   - sample S_T from Multinomial$(1, (d_{T1}, \ldots, d_{TN}))$, where $d_{Tk} = \Pr(S_T = k | \mathbf{y}^T), k = 1, \ldots, N$ and;
   - sample $S_t, t = T - 1, \ldots, 1$ from Multinomial$(1, (d_{t1}, \ldots, d_{tN}))$, where $d_{tk} = \Pr(S_t = k | s_{t+1}, \mathbf{y}^t), k = 1, \ldots, N$.

B is the enough number of iterations to ensure algorithm's convergence. The performance of the Gibbs sampling and, in particular, its convergence properties can be checked by using a variety of graphical and numerical diagnostics, for example, Gelman and Rubin (1992). This algorithm is implemented in R language and the codes are available in supplementary information. R is a free software environment for statistical computing and graphics and more details are found in its homepage https://www.r-project.org.

3 Applications

We apply the proposed model to simulated and real datasets and compare it with the first-order autoregressive HMM using model selection methods (Bayes factor and DIC).

3.1 Simulated data

We analyze a simulated sequence of length $T = 3000$. The sequence is generated from a autoregressive HMM of order two with $N = 2$ non-observable states, $M = 4$ observable states and fixed $\mathbf{s}$ as $s_t = 1$ for $t = 1, 2, \ldots, 1000, 2001, 2002, \ldots, 3000$ and $s_t = 2$ for $t = 1001, 1002, \ldots, 2000$, in order to provide DNA sequence characteristics to the simulated data.

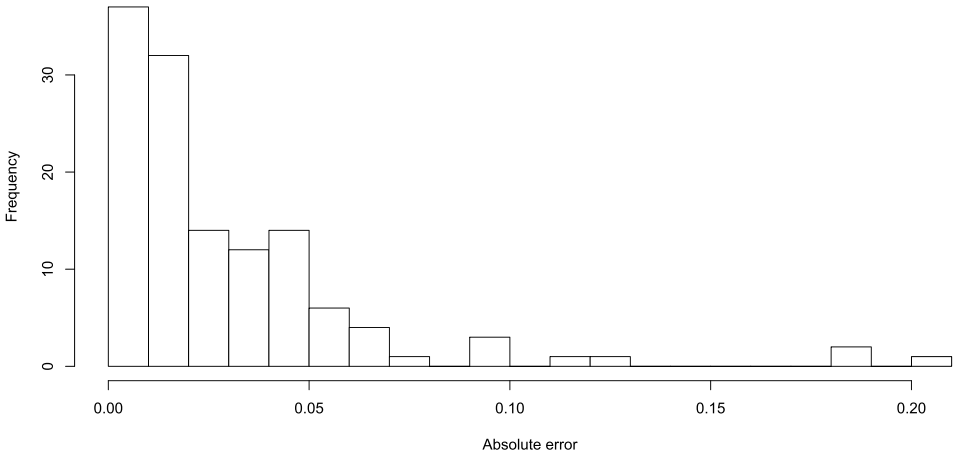

Figure 2 *Histogram of absolute error between the real probabilities* $p_{hij}^{(k)}$ *and their a posteriori estimates* (*average*).

The *a priori* distributions for transition probabilities in $P^{(1)}$ and $P^{(2)}$ are Dirichlet with parameters $\boldsymbol{\alpha}_{hi}^{(k)} = (1, 1, 1, 1)$, for $i, h = 1, \ldots, 4$ and $k = 1, 2$ that provide weak *a priori* information about these parameters.

For transition between non-observable states we choose informative *a priori* distribution. This is based on previous knowledge of these segments length that biologists often have and can improve the convergence of the method. We set $\beta_{11} = \beta_{22} = 99$ and $\beta_{12} = \beta_{12} = 2$, implying that, *a priori*, $E(a_{11}) = E(a_{22}) = 0.98$ and $\text{Var}(a_{11}) = \text{Var}(a_{22}) = 0.0002$.

We run the MCMC algorithm from two different starting points in order to use the Gelman–Rubin diagnostic of convergence. Both sequences produce similar results and the Gelman–Rubin diagnostic is lower than 1.1 for all parameter sequences indicating convergence of the chains. The Gelman–Rubin diagnostic of some parameters (randomly chosen) are available in supplementary information. We report here a run consisting of 11,000 iterations with a *burn-in* of the first 1000 iterations. Estimates are based on $R = 2000$ simulated values since we record one out of 5 values.

Figure 2 shows the histogram of absolute error between the real probabilities $p_{hij}^{(k)}$ and their *a posteriori* estimate (average). We observe the absolute errors are concentrated close to zero and it shows the transition probabilities are well estimated. More than 64% of absolute errors are lower than 0.03 and only 4% of the transition probabilities have absolute error higher than 0.10.

We estimate the *a posteriori* probabilities of non-observable states S_t, for $t = 1, \ldots, T$, by

$$\widehat{\Pr}(S_t = k|\mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} I\left(s_t^{(r)} = k\right), \tag{7}$$

where $I(A) = 1$ if A is true or $I(A) = 0$ otherwise and R is the lenght of the MCMC final sample after *burn-in* and jumps. The change points of the non-observable sequence are also well estimated. The real change points are located at $t = 1001$ and $t = 2001$ in the simulated sequence and their estimates are 1007 and 2002.

3.2 Bacteriophage *lambda* genome

We apply the autoregressive HMM with order two to analyze the genome of bacteriophage *lambda*, a parasite of the intestinal bacterium *Escherichia coli*. This sequence is $T = 48{,}502$ pairs of bases long. Previously, Skalka, Burgi and Hershey (1968), Churchill (1989), da Silva (2003) and Boys and Henderson (2004) analyzed this sequence.

This DNA sequence $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$ can be thought of as a realization of the observable random process $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_T\}$, where $Y_t \in \{a, c, g, t\} \equiv \{1, 2, 3, 4\}$, for $t = 1, 2, \ldots, T$. The letters represent the four nucleic acids or bases: adenine, cytosine, guanine and thymine. Suppose there are N types of homogeneous segments in this DNA sequence. The non-observable sequence of these homogeneous segments is a realization of the non-observable random process $\mathbf{S} = \{S_1, S_2, \ldots, S_T\}$, where $S_t \in \{1, 2, \ldots, N\}$, for $t = 1, 2, \ldots, T$ and N known.

For the non-observable states transition matrix, we choose to use the information available as the *a priori* distribution since biologists usually know approximately the length of segments in a DNA sequence. In this case, we assume that transitions between segment types are rare enough to allow us suppose $E(a_{kk})$ is close to 1. We set $\beta_{11} = \beta_{22} = 2.9$ and $\beta_{12} = \beta_{12} = 0.03$, implying that, *a priori*, $E(a_{11}) = E(a_{22}) = 0.99$ and $\mathrm{Var}(a_{11}) = \mathrm{Var}(a_{22}) = 0.0026$.

We consider autoregressive HMMs of order one and two and values of $N \in \{2, 3, 4\}$. These models are presented in Table 1 where HMM(1, 1) represents the autoregressive HMM of order one and HMM(1, 2) of order two.

We use Bayes factor and DIC (deviance information criterion) as suggested by Spiegelhalter et al. (2002), to identify the best fitted model. Tables 1 and 2 show DIC and Bayes factor, respectively, where B_{ij} represents Bayes factor comparing the models M_a and M_b, for $a = 1, \ldots, 5$ and $b = a + 1, \ldots, 6$.

Table 1 shows the DIC estimates of the six estimated models. It is clear that all second-order models fit better to data than first-order models and, among them, M_4 with $N = 2$, is the best model. We can draw the same conclusion from Bayes factors estimates shown in Table 2. All comparisons between first and second-order result in favor of second-order models, that is, B_{14}, B_{15}, B_{16}, B_{24}, B_{25}, B_{26}, B_{34}, B_{35} and B_{36} are lower than 1 and, among the second-order models, the best model is M_4.

One interesting aspect in HMM is identifying from which non-observable state each observation of the sequence comes from. This information may be gathered by estimating $\Pr(S_t = k|\mathbf{y})$ by (7). Figure 3 displays $\widehat{\Pr}(S_t = 1|\mathbf{y})$ in (a) and

Table 1 *DIC estimates*

Model	N	DIC
M_1:HMM(1, 1)	2	132,568
M_2:HMM(1, 1)	3	132,699
M_3:HMM(1, 1)	4	132,942
M_4:HMM(1, 2)	2	**130,936**
M_5:HMM(1, 2)	3	131,145
M_6:HMM(1, 2)	4	131,017

Table 2 *Bayes factor estimates*

	B_{ij}		B_{ij}		B_{ij}
B_{12}	2.2×10^{11}	B_{23}	2.2×10^{48}	B_{35}	7.52×10^{-302}
B_{13}	4.95×10^{59}	B_{24}	6.04×10^{-263}	B_{36}	1.17×10^{-243}
B_{14}	1.35×10^{-251}	B_{25}	1.66×10^{-253}	B_{45}	3.67×10^{9}
B_{15}	3.72×10^{-242}	B_{26}	2.59×10^{-195}	B_{46}	4.28×10^{67}
B_{16}	5.79×10^{-184}	B_{34}	2.7×10^{-311}	B_{56}	1.16×10^{58}

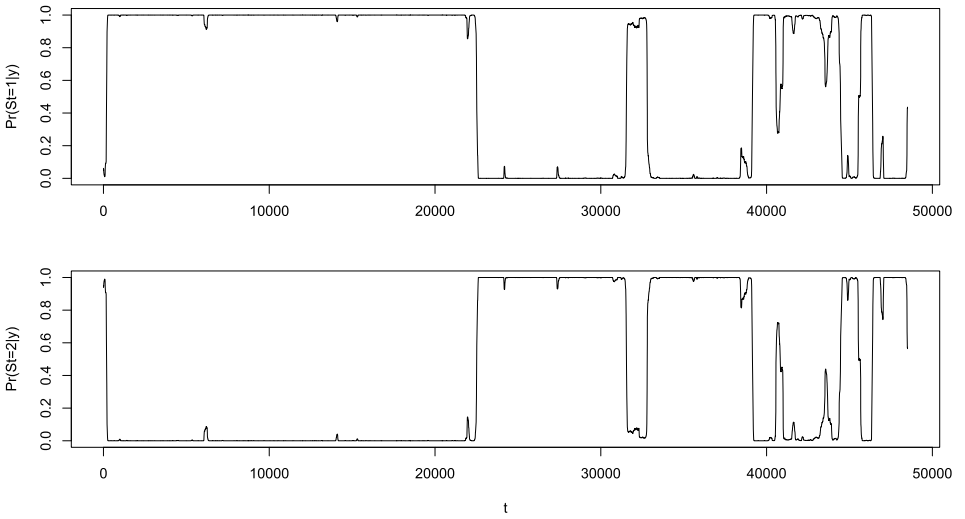

Figure 3 $\widehat{\Pr}(S_t = k|\mathbf{y})$, *for* $k = 1$ (a) *and* $k = 2$ (b).

$\widehat{\Pr}(S_t = 2|\mathbf{y})$ in (b) for bacteriophage *lambda* genome. The first part of the sequence is composed by just one segment and the other part is composed by shorter sub-sequences. We observe 8 change points in this sequence, see also Braun, Braun and Muller (2000).

These results are obtained using *burn-in* of 1000 iterations followed by 10,000 iterations in which only every 5th iterate is recorded, producing $R = 2000$ simulated values. Each run spends 20 hours, approximately, in a personal computer. Convergence is monitored with Gelman–Rubin diagnostics using two sequences with overdispersed starting values. Both sequences produce similar results and the Gelman–Rubin diagnostic is lower than 1.1 for all parameter sequences indicating convergence of the chains. The trace graphics of some parameters (randomly chosen) are available in supplementary information.

4 Discussion

We propose a second-order autoregressive hidden Markov model and its Bayesian estimator using MCMC. Simulations show that estimation of parameters and sequence of states works well and, despite the large number of parameters in this model, the method does not show problem of convergence. The application to real data example presented here, bacteriophage *lambda* genome, illustrates a situation where a second-order dependence fits the data better than a first-order, indicating that second-order autoregressive HMM have a place in model fitting, especially in genetics. Higher order models could be tried, but always respecting the parsimony principle since the number of parameters to be estimated is also larger. The accuracy in estimating the sequence of states and finding the change points, illustrated by Figure 3, is also a plus.

Acknowledgments

This work was partially supported by FAPESP.

References

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* **37**, 1554–1563. MR0202264

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171. MR0287613

Biblio, M., Monfort, A. and Robert, C. P. (1999). Bayesian estimation of switching ARMA models. *Journal of Econometrics* **93**, 229–255. MR1721099

Boys, R. and Henderson, D. (2002). On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientifc Programming* **10**, 241–251.

Boys, R. and Henderson, D. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics* **60**, 573–588. MR2089432

Boys, R., Henderson, D. and Wilkinson, D. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**, 269–285. MR1765825

Braun, J. V., Braun, R. K. and Muller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314. MR1782480

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75**, 79–97. MR1414504

Churchill, G. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94. MR0978904

Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry* **16**, 107–115.

da-Silva, C. Q. (2003). Hidden Markov models applied to a subsequence of the *Xylella fastidiosa* genome. *Genetics and Molecular Biology* **26**, 529–535.

Djurić, P. M., Kotecha, J. H., Zhang, J., Huang, Y., Ghirmai, T., Bugallo, M. F. and Miguez, J. (2003). Particle filtering. *Signal Processing Magazine, IEEE* **20**, 19–38.

Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Media: Springer.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering* **12**, 656–704.

du Preez, J. A. (1998). Efficient higher-order hidden Markov modeling. Ph.D. thesis, University of Stellenbosch. Available: www.ussigbase.org/downloads/jadp_phd.pdf.

Gassiat, E. and Kéribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM. Probabilités Et Statistique* **4**, 25–52. MR1780964

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **17**, 457–472.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology* **313**, 903–919.

Hadar, U. and Messer, H. (2009). High-order hidden Markov models—estimation and implementation. In *Statistical Signal Processing*. *Cardiff, UK*. doi:10.1109/SSP.2009.5278591.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrika* **57**, 357–384. MR0996941

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions. Lecture Notes in Economic and Mathematical Systems 454*. New York: Springer. MR1473720

Leea, S. Y., Leea, J. Y., Jungb, K. S. and Ryu, K. H. (2009). A 9-state hidden Markov model using protein secondary structure information for protein fold recognition. *Computers in Biology and Medicine* **39**, 527–534.

Martino, L., Read, J., Elvira, V. and Louzada, F. (2015). Cooperative parallel particle filters for online model selection and applications to urban mobility. Available at viXra:1512.0420.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C* **36**, 318–324.

Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT'98 Proceedings in Computational Statistics* (R. W. Payne and P. J. Green, eds.) 89–100. Heidelberg: Physica.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–285.

Ristic, B., Arulampalam, S. and Gordon, N. J. (2004). Beyond the Kalman filter: Particle filters for tracking applications. Artech house.

Robert, C. P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statist. Prob. Letters* **16**, 77–83. MR1208503

Robert, C. P. and Titterington, D. M. (1998). Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing* **8**, 145–158.

Rydén, T., Teräsvirta, T. and Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics* **13**, 217–244.

Schimert, J. (1992). A high order hidden Markov model. Ph.D. thesis, University of Washington. MR2688816

Seifert, M. (2010). Extensions of Hidden Markov Models for the analysis of DNA microarray data. Ph.D. thesis, University of Halle-Wittenberg. Available at http://nbn-resolving.de/urn:nbn:de:gbv:3:4-4110.

Seifert, M., Abou-El-Ardat, K., Friedrich, B., Klink, B. and Deutsch, A. (2014). Autoregressive higher-order hidden Markov models: Exploiting local chromosomal dependencies in the analysis of tumor expression profiles. *PLoS ONE* **9**, e100295.

Seifert, M., Gohr, A., Strickert, M. and Grosse, I. (2012). Parsimonious higher-order hidden Markov models for improved array-CGH analysis with applications to Arabidopsis thaliana. *PLoS Computational Biology* **8**, e1002286.

Skalka, A., Burgi, E. and Hershey, A. D. (1968). Segmental distribution of nucleotides in the DNA of Bacteriophage lambda. *Journal of Molecular Biology* **34**, 1–16.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960.

Spiegelhalter, D., et al. (2002). Bayesian measures of model complexity and fit. *Royal Statistical Society* **64**, 583–639. MR1979380

Departamento de Estatística
UFSCar—Brazil
E-mail: dzuanetti@yahoo.com.br
dlam@ufscar.br