A BAYESIAN MODEL FOR BINARY MARKOV CHAINS

SOUAD ASSOUDOU and BELKHEIR ESSEBBAR

This note is concerned with Bayesian estimation of the transition probabilities of a binary Markov chain observed from heterogeneous individuals. The model is founded on Jeffreys' prior which allows for transition probabilities to be correlated. The Bayesian estimator is approximated by means of Monte Carlo Markov chain (MCMC) techniques. The performance of the Bayesian estimates is illustrated by analyzing a small simulated data set.

1. Introduction. Markov chain models have been useful for the analysis of longitudinal data in many areas of research. In ecology, the model was used to study the migration behavior of animal population from capture-recapture data [2]; in pathology, the model was useful to describe the evolution of certain viral or infectious diseases [3, 6]; in sociology, the model was used for the modeling of the behavior of smoking population [5].

In most applications, these models do not take into account possible correlations between different rows of the transition matrix. As the observations are dependent, it seems more reasonable to consider prior distributions which incorporate a certain type of dependence between the components of the parameters.

In this note, we explore the Bayesian model, for binary Markov chains, using Jeffreys' prior which has some advantages: the model has no extra parameters and permits a structure of correlation between the transition probabilities.

In the sequel, $X = (X_0,\ldots,X_n)$ denotes a homogenous and stationary Markov chain with transition probabilities

$$p_{ij} = P(X_{t+1} = j \mid X_t = i), \quad i,j = 0,1. \tag{1.1}$$

The equilibrium probability of observing a 1, which we denote by p, represents the long-run proportion of time when the Markov chain is in state 1. From [1], this probability is given by $p = p_{01}/(p_{01} + p_{10})$.

Letting $x = (x_0,\ldots,x_n)$ denote a fully observed realization of X, conditionally to $X_0 = 1$, the distribution of the observed sequence is then

$$f(x \mid x_0 = 1, \theta) = (1 - p_{01})^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} (1 - p_{10})^{n_{11}}, \tag{1.2}$$

where n_{ij} is the number of one-step transition from state i to state j until time n and $\theta = (p_{01}, p_{10}) \in \,]0,1[^2$ is the unknown parameter which is the aim of this inference.

Using the distribution (1.2), the maximum likelihood estimates (MLEs) of p_{01} and p_{10} are

$$\hat{p}_{01} = \frac{n_{01}}{n_{00} + n_{01}}, \qquad \hat{p}_{10} = \frac{n_{10}}{n_{10} + n_{11}}, \tag{1.3}$$

and the MLE of p is $\hat{p} = \hat{p}_{01}/(\hat{p}_{01} + \hat{p}_{10})$.

The remainder of the note is organized as follows. In the next section, we calculate Jeffreys' prior and the correspondent posterior distribution. Next, we describe the way to approximate the Bayesian estimator via the independent Metropolis-Hasting (IMH) algorithm. Finally, we develop a numerical study by simulation in order to compare the Bayesian estimates with the MLEs.

2. Jeffreys' prior. The goal here is to determine Jeffreys' prior (see, e.g., [4]) and its correspondent posterior distribution. Jeffreys' prior is obtained by taking the determinant of the information matrix which is defined according to Fisher as

$$\mathcal{I}_n(\theta) = E\left[-\frac{\partial^2 l_n(\theta \mid x_0 = 1)}{\partial p_{ij} \partial p_{kh}} \right], \tag{2.1}$$

where $l_n(\theta \mid x_0 = 1)$ is the logarithm function of (1.2). To obtain (2.1), we take the second derivates of $l_n(\theta \mid x_0 = 1)$, and then take the expectation with negative sign to yield

$$\begin{aligned}
E\left[-\frac{\partial^2 l_n(\theta \mid x_0 = 1)}{\partial p_{01}^2} \right] &= \frac{E(n_{01} \mid X_0 = 1)}{p_{01}^2} + \frac{E(n_{00} \mid X_0 = 1)}{(1 - p_{01})^2}, \\
E\left[-\frac{\partial^2 l_n(\theta \mid x_0 = 1)}{\partial p_{10}^2} \right] &= \frac{E(n_{10} \mid X_0 = 1)}{p_{10}^2} + \frac{E(n_{11} \mid X_0 = 1)}{(1 - p_{10})^2}.
\end{aligned} \tag{2.2}$$

Considering the expectation of the sufficient statistics $(n_{00}, n_{01}, n_{10}, n_{11})$, we have

$$\begin{aligned}
E(n_{ij} \mid X_0 = 1) &= \sum_{t=1}^{n} P(X_{t-1} = i, X_t = j \mid X_0 = 1) \\
&= \sum_{t=1}^{n} P(X_t = j \mid X_{t-1} = i) P(X_{t-1} = i \mid X_0 = 1) \\
&= p_{ij} \sum_{t=1}^{n} p_{1i}^{(t-1)},
\end{aligned} \tag{2.3}$$

where $p_{ij}^{(k)}$ denotes the k-step transition probabilities. By the Chapman-Kolmogorov equation [1], these probabilities may be written in terms of the one-step transition probabilities p_{ij} as

$$p_{00}^k = (p_{01} + p_{10})^{-1} \left[p_{10} + p_{01} (1 - p_{01} - p_{10})^k \right],$$

$$p_{01}^k = (p_{01} + p_{10})^{-1} \left[p_{01} - p_{01} (1 - p_{01} - p_{10})^k \right],$$

$$p_{10}^k = (p_{01} + p_{10})^{-1} \left[p_{10} - p_{10} (1 - p_{01} - p_{10})^k \right],$$

$$p_{11}^k = (p_{01} + p_{10})^{-1} \left[p_{01} + p_{10} (1 - p_{01} - p_{10})^k \right]. \tag{2.4}$$

Then, we deduce that

$$E(n_{01} \mid X_0 = 1) = p_{01} (p_{01} + p_{10})^{-1} \left[n p_{10} - p_{10} \frac{1 - (1 - p_{01} - p_{10})^n}{p_{01} + p_{10}} \right],$$

$$E(n_{00} \mid X_0 = 1) = p_{00} (p_{01} + p_{10})^{-1} \left[n p_{10} - p_{10} \frac{1 - (1 - p_{01} - p_{10})^n}{p_{01} + p_{10}} \right],$$

$$E(n_{10} \mid X_0 = 1) = p_{10} (p_{01} + p_{10})^{-1} \left[n p_{01} + p_{10} \frac{1 - (1 - p_{01} - p_{10})^n}{p_{01} + p_{10}} \right],$$

$$E(n_{11} \mid X_0 = 1) = p_{11} (p_{01} + p_{10})^{-1} \left[n p_{01} + p_{10} \frac{1 - (1 - p_{01} - p_{10})^n}{p_{01} + p_{10}} \right]. \tag{2.5}$$

Hence, the Fisher information matrix can be written as

$$\mathcal{I}_n(\theta) = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \tag{2.6}$$

where

$$A_{11} = \frac{p_{10} \left[n(p_{01} + p_{10}) - 1 + (1 - p_{01} - p_{10})^n \right]}{p_{01} (1 - p_{01}) (p_{01} + p_{10})^2},$$

$$A_{22} = \frac{n p_{01} (p_{01} + p_{10}) + p_{10} \left[1 - (1 - p_{10} - p_{01})^n \right]}{p_{10} (1 - p_{10}) (p_{10} + p_{01})^2}. \tag{2.7}$$

Since Jeffreys' prior $\pi(\theta)$ is defined by

$$\pi(\theta) \propto \left[\det (\mathcal{I}_n(\theta)) \right]^{1/2}, \tag{2.8}$$

where $\det(\cdot)$ denotes the determinant, it follows that

$$\pi(\theta) \propto \left[n(p_{01} + p_{10}) - 1 + (1 - p_{01} - p_{10})^n \right]^{1/2}$$
$$\times \left[n p_{01} (p_{01} + p_{10}) + p_{10} (1 - p_{01} - p_{10})^n \right]^{1/2}$$
$$\times p_{01}^{-1/2} (1 - p_{01})^{-1/2} (1 - p_{10})^{-1/2} (p_{01} + p_{10})^{-2}, \tag{2.9}$$

and the posterior density is

$$
\begin{aligned}
\pi(\theta \mid x) \propto & \left[n(p_{01} + p_{10}) - 1 + (1 - p_{01} - p_{10})^n \right]^{1/2} \\
& \times \left[n p_{01}(p_{01} + p_{10}) + p_{10}(1 - p_{01} - p_{10})^n \right]^{1/2} \\
& \times p_{01}^{n_{01} - 1/2}(1 - p_{01})^{n_{00} - 1/2} p_{10}^{n_{10}}(1 - p_{10})^{n_{11} - 1/2}(p_{01} + p_{10})^{-2}.
\end{aligned}
\tag{2.10}
$$

The main advantage of the density (2.9) is that it provides a convenient analysis when the transition probabilities may be correlated. Moreover, this prior has no extra parameters and it is a conjugate distribution for $f(x \mid x_0 = 1, \theta)$ given by (1.2). Also notice that for the particular case $p_{01} + p_{10} = 1$ (independent case), Jeffreys' prior given by (2.9) is just the beta distribution $\mathcal{B}e(1/2, 1/2)$.

3. Bayesian estimation of transition probabilities. Under the squared error loss, we know that the Bayes estimator coincides with the posterior mean, that is,

$$
E(\theta \mid x) = \int \theta \pi(\theta \mid x) d\theta.
\tag{3.1}
$$

In the case of Jeffreys' prior, the above integral is difficult to calculate, so we propose an approximation of it by means of a Monte Carlo Markov chain (MCMC) algorithm; namely, the IMH algorithm (see [7]).

The fundamental idea behind these algorithms is to construct a homogenous and ergodic Markov chain $(\theta^{(l)})$ with stationary measure $\pi(\theta \mid x)$. For m_0 large enough, $\theta^{(m_0)}$ is roughly distributed from $\pi(\theta \mid x)$ and the sample $\theta^{(m_0)}, \theta^{(m_0+1)}, \ldots$ can be used to derive the posterior means. For instance, the Ergodic theorem (cf. [7]) justifies the approximation of the integral (3.1) by the empirical average

$$
\frac{1}{m} \sum_{l=1}^{m} \theta^{(m_0 + l)}
\tag{3.2}
$$

in the sense that (3.2) is converging to the integral (3.1) for almost every realization of the chain $(\theta^{(l)})$ under minimal conditions. Next, we give the description of the IMH algorithm.

Given $\theta^{(0)} = (p_{01}^{(0)}, p_{10}^{(0)})$, the IMH algorithm at step l proceeds as follows.

STEP 1. Generate $y^{(l)} = (y_1^{(l)}, y_2^{(l)}) \sim U_{[0,1]} \times U_{[0,1]}$.
STEP 2. Take

$$
\theta^{(l+1)} = \begin{cases} y^{(l)} & \text{with probability } \sigma(\theta^{(l)}, y^{(l)}), \\ \theta^{(l)} & \text{with probability } 1 - \sigma(\theta^{(l)}, y^{(l)}), \end{cases}
\tag{3.3}
$$

where

   (i) $\sigma(\theta^{(l)}, y^{(l)}) = \min(\pi(y^{(l)} \mid x)/\pi(\theta^{(l)} \mid x), 1)$,
  (ii) $\pi(\cdot \mid x)$ is the posterior density given by (2.10).

As convergence assessments, we use the cumulated sums method (cf. [7]) in the sense that a necessary condition for convergence is the stabilization of the empirical average (3.2). Also, this method of convergence control gives the minimal value m of iterations that provides the approximation of the integral (3.1) by the empirical average (3.2).

4. Numerical study.   In this section, we illustrate the performance of the Bayesian estimation based on Jeffreys' prior by analyzing a small simulated data set.

On the one hand, the simulation study compares the proposed Bayesian estimators with the MLEs. On the other hand, it compares the two estimators for independently-chosen transition probabilities in both cases of the beta distribution $\mathcal{B}e(1/2, 1/2)$ and the uniform distribution.

We recall that under the beta prior distribution

$$\pi(p_{01}, p_{10}) \propto p_{10}^{-1/2}(1 - p_{10})^{-1/2} p_{01}^{-1/2}(1 - p_{01})^{-1/2}, \tag{4.1}$$

the Bayesian estimator is calculated explicitly by

$$\tilde{p}_{01} = \frac{n_{01} + 1/2}{n_{00} + n_{01} + 1}, \qquad \tilde{p}_{10} = \frac{n_{10} + 1/2}{n_{10} + n_{11} + 1}. \tag{4.2}$$

The Bayesian solution, using the uniform prior distribution

$$\pi(p_{01}, p_{10}) = \mathbb{I}_{[0,1]}(p_{10}) \times \mathbb{I}_{[0,1]}(p_{01}), \tag{4.3}$$

is given by

$$\tilde{p}_{01} = \frac{n_{01} + 1}{n_{00} + n_{01} + 2}, \qquad \tilde{p}_{10} = \frac{n_{10} + 1}{n_{10} + n_{11} + 2}. \tag{4.4}$$

4.1. A simulated data set.   Table 4.1 displays a data set consisting of 20 independent Markov chains each with 21 observations; obviously, the chains may be of differing lengths. To generate this data set, transition probabilities for each chain are first drawn from Jeffreys' prior given by (2.9) by using the IMH algorithm (see Section 3). We assume, without loss of generality, that the first state X_0, in each chain, is equal to 1. The remaining observations in each chain are drawn in succession from Bernoulli distribution with successive probabilities given by the appropriate transition probabilities.

We recall that Jeffreys' prior permits a certain type of dependence between the random vectors P_{10} and P_{01}. Indeed, P_{10} and P_{01} are correlated with correlation coefficient

$$\rho = \frac{E(P_{10} P_{01}) - E(P_{10}) E(P_{01})}{\sigma_{10} \sigma_{01}}, \tag{4.5}$$

where σ_{10} (resp., σ_{01}) is the standard deviation of P_{10} (resp., P_{01}).

TABLE 4.1. Estimation of the transition probabilities.

Simulated chains	Actual		MLE		Bayesian estimates					
					Jeffreys' prior		Beta prior		Uniform prior	
	p_{10}	p_{01}	$\hat{p}_{10}$	$\hat{p}_{01}$	$\hat{p}_{10}$	$\hat{p}_{01}$	$\hat{p}_{10}$	$\hat{p}_{01}$	$\hat{p}_{10}$	$\hat{p}_{01}$
1 1010101010101010101	0.9765	0.9853	1.000	1.000	0.9561	0.9551	0.9545	0.9545	0.9167	0.9167
2 1000000000000000000000	0.9842	0.0338	1.000	0.0000	0.9241	0.0325	0.7500	0.0250	0.6667	0.0476
3 1000101010000100000	0.7562	0.1659	0.8333	0.2857	0.7821	0.3074	0.7857	0.3000	0.7500	0.3125
4 1111111111000100000	0.1971	0.3216	0.1538	0.1429	0.1822	0.3320	0.1786	0.1875	0.2000	0.2222
5 1000000000001000010	0.7962	0.0440	1.000	0.1176	0.8151	0.0391	0.8750	0.1389	0.8000	0.1579
6 1100010101011101101	0.3789	0.5509	0.5455	0.6667	0.5448	0.5612	0.5417	0.6500	0.5385	0.6364
7 1101110101100111010	0.5981	0.7597	0.5385	0.8571	0.5617	0.8071	0.5357	0.8125	0.5333	0.7778
8 1111111111111111111	0.0026	0.3017	0.0000	—	0.0356	0.3412	0.0238	0.5000	0.0454	0.5000
9 1000001000000010001000	0.9137	0.2660	1.000	0.1875	0.8905	0.2513	0.9000	0.2059	0.8333	0.2222
10 1010101010101010101	0.9908	0.9931	1.000	1.000	0.9679	0.9743	0.9545	0.9545	0.9167	0.9167
11 1110010101011010011	0.4731	0.7925	0.5000	0.7500	0.4912	0.7316	0.5000	0.7222	0.5000	0.7000
12 1111001111111111111	0.0719	0.6722	0.0556	0.5000	0.0933	0.5937	0.0789	0.5000	0.1000	0.5000
13 1011000011110001011	0.4689	0.4791	0.4000	0.4000	0.4442	0.4518	0.4091	0.4091	0.4167	0.4167
14 1100011001111110111	0.3125	0.3056	0.2143	0.5000	0.4242	0.3901	0.2333	0.5000	0.2500	0.5000
15 1111111100000000001	0.2805	0.1253	0.1111	0.0909	0.2517	0.1245	0.1500	0.1250	0.1818	0.1538
16 1100000000000000000	0.7686	0.0161	0.5000	0.0000	0.6856	0.0341	0.5000	0.0263	0.5000	0.0500
17 1111111101111111111	0.0836	0.7308	0.0526	1.000	0.0916	0.7025	0.0750	0.7500	0.0909	0.6667
18 1010000000001000000	0.6667	0.2376	1.000	0.1176	0.7156	0.1817	0.8750	0.1389	0.8000	0.1579
19 1101111101110101010	0.4444	0.9135	0.4286	1.000	0.4391	0.9292	0.4333	0.9286	0.4375	0.8750
20 1111000110110000000	0.3159	0.1516	0.3750	0.1667	0.3415	0.1697	0.3889	0.1923	0.4000	0.2143
Mean	0.5240	0.4423	0.5354	0.4622	0.5319	0.4455	0.5071	0.4511	0.4939	0.4472
Mean square error (MSE)			0.0160	0.0133	0.0030	0.0024	0.0122	0.0096	0.0131	0.0101

To approximate this coefficient, we use the sample $\theta^{(1)}, \ldots, \theta^{(m)}$, drawn from Jeffreys' prior thanks to the IMH algorithm. Therefore, an approximation of ρ is

$$\tilde{\rho} = \frac{m \sum_{l=1}^{m} p_{10}^{(l)} p_{01}^{(l)} - \sum_{l=1}^{m} p_{10}^{(l)} \sum_{l=1}^{m} p_{01}^{(l)}}{\sqrt{m \sum_{l=1}^{m} \left(p_{10}^{(l)}\right)^2 - \left(\sum_{l=1}^{m} p_{10}^{(l)}\right)^2} \sqrt{m \sum_{l=1}^{m} \left(p_{01}^{(l)}\right)^2 - \left(\sum_{l=1}^{m} p_{01}^{(l)}\right)^2}}. \tag{4.6}$$

Numerically, we have $\tilde{\rho} = 0.36$.

To obtain the Bayesian estimator $\tilde{p}_{ij}$, based on Jeffreys' prior, of the transition probabilities, given each chain, we apply the IMH algorithm to the posterior distribution given by (2.10). The MLE $\hat{p}_{ij}$ is calculated from (1.3). The Bayesian estimator $\tilde{p}_{ij}$ founded on the beta distribution (resp., the uniform distribution) is obtained from (4.2) (resp., from (4.4)). The results of this experiment are provided in Table 4.1.

4.2. Simulation results. Table 4.1 shows the actual transition probabilities p_{ij} from the simulation, the MLE $\hat{p}_{ij}$, and the Bayesian estimates $\tilde{p}_{ij}$ of the transition probabilities for each chain.

Notice that for many chains, the MLE takes extreme values 0 or 1; this is explained by the restricted size of the simulated sample. In addition, for the chain no. 8, $\hat{p}_{01}$ does not exist because the chain never entered state 0, whereas the Bayesian estimates do not suffer from these problems because a common prior distribution is assumed.

Also shown in Table 4.1 are the mean actual and the estimated transition probabilities, as well as mean square errors (MSE) for the estimates. The MSE are calculated by averaging the squared difference between the estimated probability and the actual probability used in simulation. Notice that the Bayes posterior means perform better than the MLEs. In particular, the MSE of the MLEs is clearly higher than that corresponding to the Bayes estimates.

This study also illustrates the usefulness of modeling the dependence among the transition probabilities. In particular, the resulting posterior distributions, under the assumption that P_{10} and P_{01} are independent, may not be accurate. Indeed, using the beta prior distribution (resp., the uniform prior distribution), the resulting changes in the posterior means range from -0.1594 to 0.1909 (resp., from -0.0844 to 0.2574) for $\tilde{p}_{10}$ and from -0.1588 to 0.1445 (resp., from -0.1588 to 0.1098) for $\tilde{p}_{01}$. Moreover, the MSEs corresponding to $\tilde{p}_{10}$ and $\tilde{p}_{01}$ become 0.0122 and 0.0096 (resp., 0.0131 and 0.0101). These are slightly higher than the results obtained by modeling the dependence. All these results lead to privilege the Bayesian solution based on Jeffreys' prior.

For the previous experiment, a Pascal program is written to run the transition probabilities. The Bayesian estimator $\tilde{p}_{ij}$, founded on Jeffreys' prior, is obtained from a single chain including 10^4 iterations.

Figures 4.1 and 4.2 give an example of the convergence evaluation (see Section 3). Figure 4.1 (resp., Figure 4.2) describes the convergence of the estimator $\tilde{p}_{10}$ (resp., $\tilde{p}_{01}$) as the number m of iterations increases. The final values are 0.4073 and 0.1742 for $\tilde{p}_{10}$ and $\tilde{p}_{01}$, respectively.
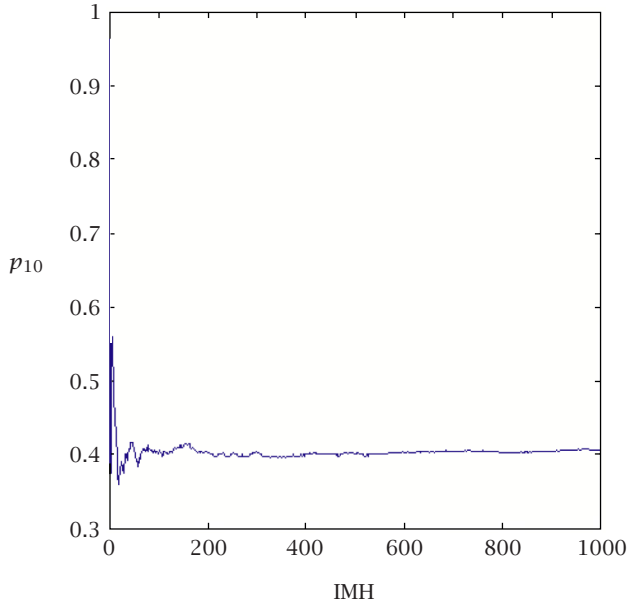
FIGURE 4.1. Convergence of $\tilde{p}_{10}$.
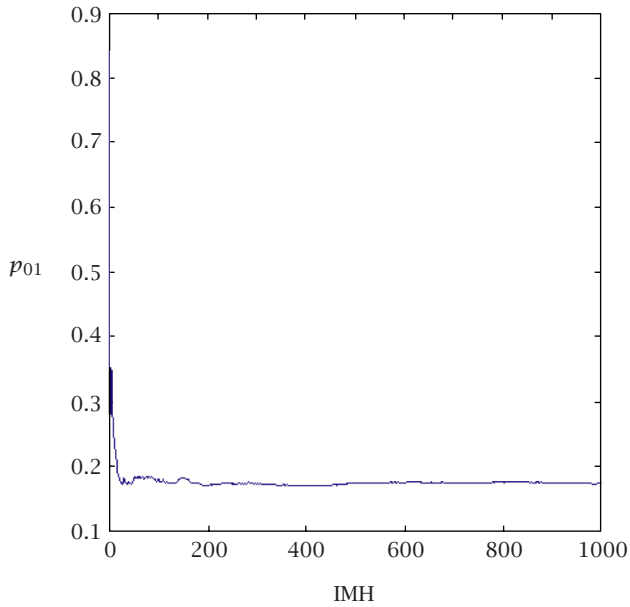
FIGURE 4.2. Convergence of $\tilde{p}_{01}$.

5. Conclusion. In this note, we studied the Bayesian estimation for the transition probabilities of a binary Markov chain under Jeffreys' prior distribution. As shown, this prior has many advantages: it permits a certain type of dependence between the

components of the parameter. The absence of extra parameter in this prior is of great interest because we do not need to do more extra estimation. A numerical study by simulation is also carried out to evaluate the performance of the Bayesian estimates compared to the MLEs. The following stage of this note will be to generalize the suggested method in the case of missing data.

REFERENCES

[1]   D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, John Wiley & Sons, New York, 1965.

[2]   J. A. Dupuis, *Bayesian estimation of movement and survival probabilities from capture-recapture data*, Biometrika **82** (1995), no. 4, 761–772.

[3]   R. C. Gentleman, J. F. Lawless, J. C. Lindsey, and P. Yan, *Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease*, Statistics in Medicine **13** (1994), no. 3, 805–821.

[4]   R. E. Kass and L. Wasserman, *The selection of prior distributions by formal rules*, J. Amer. Statist. Assoc. **91** (1996), no. 435, 1343–1370.

[5]   M. R. Meshkani and L. Billard, *Empirical Bayes estimators for a finite Markov chain*, Biometrika **79** (1992), no. 1, 185–193.

[6]   S. Richardson, I. Deltour, and J. Y. Lehesran, *Stochastic algorithms for Markov models estimation with intermittent missing data*, Biometrics **55** (1999), 565–573.

[7]   C. Robert, *Méthodes de Monte Carlo par Chaînes de Markov* [*Markov Chain Monte Carlo Methods*], Statistique Mathématique et Probabilité, Éditions Économica, Paris, 1996 (French).

Souad Assoudou: Department of Mathematics and Computer Sciences, Faculty of Sciences, Avenue Ibn Batouta, BP 1014, Rabat, Morocco
    E-mail address: souad457@caramail.com

Belkheir Essebbar: Department of Mathematics and Computer Sciences, Faculty of Sciences, Avenue Ibn Batouta, BP 1014, Rabat, Morocco
    E-mail address: essebbar@fsr.ac.ma