Università degli Studi di Macerata

Dipartimento di Istituzioni Economiche e Finanziarie

Bayesian inference for Hidden Markov Models

Rosella Castellano, Luisa Scaccia

Temi di discussione n. 43

Novembre 2007

Bayesian inference for Hidden Markov Models

Rosella Castellano, Luisa Scaccia

Abstract

Hidden Markov Models can be considered an extension of mixture models, allowing for dependent observations. In a hierarchical Bayesian framework, we show how Reversible Jump Markov Chain Monte Carlo techniques can be used to estimate the parameters of a model, as well as the number of regimes. We consider a mixture of normal distributions characterized by different means and variances under each regime, extending the model proposed by Robert *et al.* (2000), based on a mixture of zero mean normal distributions.

Rosella Castellano, Università di Macerata.
E-mail: castellano@unimc.it.
Luisa Scaccia, Università di Macerata.
E-mail: scaccia@unimc.it.

1 Introduction

A Hidden Markov Model (HMM) or Markov Switching Model is a mixture model whose mixing distribution is a finite state Markov Chain. These models provide useful representations of dependent heterogeneous phenomena and, for this reason, they are applied in many different fields, such as econometrics (Hamilton, 1989; Chib, 1996; Krolzig, 1997; Billio *et al.*, 1999), biology (Leroux and Puterman, 1992), genetics (Churchill, 1995), neurophysiology (Fredkin and Rice, 1992), speech processing (Rabiner, 1989). In particular, HMMs have been successfully applied in finance: financial prices usually show non linear dynamics which are often due to the existence of two or more regimes within which returns and/or volatilities display different behavior. Using these models, Rydén *et al.* (1998), reproduce most of the stylized facts about daily series of returns while Rossi and Gallo (2006) provide accurate estimates of stochastic volatility. Engel and Hamilton (1990) model segmented time-trends in the US dollar exchange rates via HMMs. Robert *et al.* (2000) use HMMs to study daily returns of the S&P index, assuming the existence of different regimes characterized by different levels of volatility.

The main problem associated with HMMs is to choose the number of regimes, i.e. the number of generating data processes, which differ one from another just for the value of the parameters. In a classical perspective, choosing the number of regimes would require hypothesis testing with nuisance parameters, identified only under the alternative. Thus, the regularity conditions necessary to apply asymptotic theory do not hold and the limiting distribution of the likelihood ratio test must be approximated by simulation, an approach demanding enormous computational efforts. Penalized likelihood methods such as the Akaike and Bayesian information criteria are less demanding, though, they produce no number quantifying the confidence in the results, such as a p-value.

In a Bayesian context there are different suggestions for choosing the number of regimes in a HMM. For example, Otranto and Gallo (2002) adopt a Bayesian nonparametric approach, based on Dirichlet processes. Since a distribution realised from a Dirichlet process is almost surely discrete, a random sample drawn from it has positive probability of ties, providing a flexible model to cluster the observations into different regimes. The posterior distribution of the number of regimes is estimated through the simulated posterior distribution of the number of clusters. However, a drawback of this approach is that a single parameter controls the variability and the clustering, creating difficulties for the prior specifications. Moreover, the Dirichlet process is well known to favor, *a priori* , unequal allocations (see, for example, Green and Richardson, 2001) and this phenomenon turns to be much more dramatic

with the growth of the number of observations. Often, the unbalance in the allocation distribution persists also a posteriori. Furthermore, it can be easily proved that the predictive distribution of a future observation has a non null probability of being equal to a past observation and this is clearly an unrealistic assumption when the observed data are assumed to be realizations of a continuous distribution. Finally, the posterior distributions obtained by non parametric approaches are very sensitive to the specification of the priors: the number of parameters being infinite, the quantity of information provided by data on such parameters is necessarily limited and the likelihood never dominates over the prior.

A natural alternative to the Dirichlet process model is to use mixtures based on multinomial allocations. Following Robert *et al.* (2000) and Richardson and Green (1997), we use a fully Bayesian analysis, based on the Reversible Jump (RJ) algorithm, developed in Green (1995), which allows for the change of dimension of the parameter space, changing the number of regimes from one iteration to the other. The algorithm allows to estimate the joint posterior distribution of the number of regimes and of all the parameters.

The paper is organized as follows: details of the prior modeling are discussed in Section 2; Section 3 describes the MCMC algorithm used to simulate the joint posterior distribution of all the parameters of the model, including the number of regimes; Section 4 illustrates the Bayesian approach for inference and forecasting; conclusions are reported in Section 5.

2 The model

In this section, we present the proposed hierarchical HMM and, in a bayesian framework, we discuss the prior assumptions on the parameters of the model.

2.1 Hidden Markov Models

Let $\boldsymbol{y} = (y_t)_{t=1}^T$ be the vector of observed variables, indexed by time. HMMs assume that the distribution of each observed data point y_t depends on an unobserved (hidden) variable, denoted s_t, that takes on values from 1 to k. The hidden variable $\boldsymbol{s} = (s_t)_{t=1}^T$ characterizes the "state" or "regime" in which the generating process is at any time t. HMMs further postulate a Markov Chain for the evolution of the unobserved state variable and, hence, the process for s_t is assumed to depend on the past realizations of $\boldsymbol{y}$ and $\boldsymbol{s}$ only through s_{t-1}:

$$p(s_t = j | s_{t-1} = i) = \lambda_{ij}, \tag{1}$$

where λ_{ij} is the generic element of the transition matrix $\mathbf{\Lambda} = (\lambda_{ij})$, with vector of stationary probabilities $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}'\mathbf{\Lambda} = \boldsymbol{\pi}'$. Figure 1 illustrates the dependency structure in a HMM, showing that each observation y_t is conditionally independent of all other unobserved and observed data, given s_t.

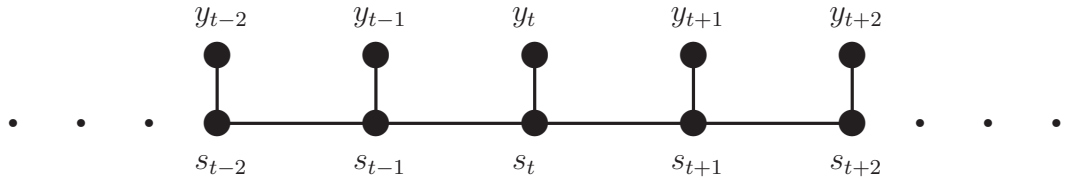

Figure 1: *Graphical representation of HMM dependencies. The conditional distribution of the value at each node, given the values of all the other nodes, depends only on the nodes to which it is connected by an edge.*

As a generating process, we assume that, if $s_t = i$, y_t is a realization from a $N(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 are respectively the mean and the variance of the i-th regime. Thus the marginal distribution for an observation y_t, conditional on weights $\boldsymbol{\pi} = (\pi_i)_{i=1}^k$, means $\boldsymbol{\mu} = (\mu_i)_{i=1}^k$ and standard deviations $\boldsymbol{\sigma} = (\sigma_i)_{i=1}^k$, is

$$y_t|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} \sim \sum_{i=1}^{k} \pi_i \phi(\cdot; \mu_i, \sigma_i^2) \qquad \text{independently for } t = 1, 2, \ldots, T, \quad (2)$$

where $\phi(\cdot; \mu_i, \sigma_i^2)$ is the density of the $N(\mu_i, \sigma_i^2)$. Notice that the model in (2) can be analogously expressed as

$$y_t|\boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\sigma} \sim \phi(\cdot; \mu_{s_t}, \sigma_{s_t}^2). \quad (3)$$

Integrating out s_t in (3), using its stationary distribution, leads back to (2).

Finally, we assume that the number of components k (i.e. the number of regimes) is unknown and subject to inference. Notice that for $k = 1$ the model in (2) reduces to a simple random walk with drift.

2.2 Bayesian approach

As mentioned before, we adopt a Bayesian approach which implies that if $\boldsymbol{\theta}$ is the vector of the parameters describing the model (including k), all the required inference is based on the posterior distribution of $\boldsymbol{\theta}$. Given the data set $\boldsymbol{y}$, from the Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{\theta})$ is the prior distribution of the parameters.

2.3 Prior distributions

In order to get weakly informative priors for the parameters, we introduce an hyperprior structure in order to make only minimal assumptions on the data. In particular, we assume *a priori* :

$$\mu_i|\sigma_i^2 \sim N(\xi, \kappa\sigma_i^2) \quad \text{and} \quad \sigma_i^{-2} \sim \text{Ga}(\eta, \zeta) \qquad \text{independently } \forall i = 1, \ldots, k,$$

where the Gamma distribution is parametrised so that the mean and the variance are η/ζ and η/ζ^2 respectively. For both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ we introduce an additional hierarchical level, implying that their supports are not *a priori* fixed. A good distribution for κ seems to be the one assigning a high probability to the values within the interval $[0, 1]$ (in order to reduce the impact of $\boldsymbol{\sigma}^2$). At the same time, a distribution for κ with infinite variance would not steer results in a specific direction and would guarantee a certain degree of uncertainty. Thus we use an Inverse Gamma with hyperparameters q and r. A possible choice of these two parameters is, for example, $q = 2$ and $r = 2$. The hyperparameter ζ follows a Gamma distribution with parameters f and h, with $\eta > 1 > f$ and h a small multiple of $1/R^2$, where R is the length of the range of data. Finally, the hyperparameter ξ can be set equal to the mid-point of this range.

The rows of the transition matrix are, *a priori* , assumed to be distributed as a Dirichlet:

$$\lambda_{ij} \sim D(\boldsymbol{\delta}), \qquad \text{independently } \forall i = 1, \ldots, k$$

where $\boldsymbol{\delta} = (\delta_j)_{j=1}^k$. In particular, we assume $\delta_j = 1$, $\forall j$, so that each row of the transition matrix is *a priori* uniform on the simplex of dimension k.

The number of components k is *a priori* uniformly distributed on the values $\{1, 2, \ldots, K\}$, where K is a pre-specified integer. As in other mixture model contexts (or indeed in almost all model choice problems), it seems difficult to argue objectively for a specific prior for k. Our choice here, similarly to Richardson and Green (1997), is due to the consideration that a uniform prior allows to adjust the results in order to get posteriors corresponding to other priors, by importance sampling (see, for example, Hammersley and Handscomb, 1964).

2.4 Complete hierarchical model

The joint distribution of all the variables conditional on fixed hyperparameters may be written as:

$$\begin{aligned} &p(k, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{s}, \boldsymbol{y}, \zeta, \kappa|\boldsymbol{\delta}, \xi, \eta, f, h, r, q) = \\ &= p(k)p(\boldsymbol{\Lambda}|k, \boldsymbol{\delta})p(\boldsymbol{s}|\boldsymbol{\Lambda})p(\boldsymbol{\mu}|k, \xi, \kappa, \boldsymbol{\sigma})p(\boldsymbol{\sigma}|k, \eta, \zeta)p(\zeta|f, h)p(\kappa|q, r)p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\sigma}), \end{aligned}$$

where

$$p(\boldsymbol{s}|\boldsymbol{\Lambda}) = p(s_1|\boldsymbol{\Lambda}) \prod_{t=2}^{T} p(s_t|s_{t-1}, \boldsymbol{\Lambda}),$$

with $p(s_t|s_{t-1}, \boldsymbol{\Lambda})$ given by (1) and $p(s_1 = i|\boldsymbol{\Lambda}) = \pi_i$, and

$$p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{t=1}^{T} \phi(y_t; \mu_{s_t}, \sigma_{s_t}^2).$$

The prior distributions $p(k)$, $p(\boldsymbol{\Lambda}|k, \boldsymbol{\delta})$, $p(\boldsymbol{\mu}|k, \xi, \kappa, \boldsymbol{\sigma})$, $p(\boldsymbol{\sigma}|k, \eta, \zeta)$, $p(\zeta|f, h)$ and $p(\kappa|q, r)$ are all given in Section 2.3.

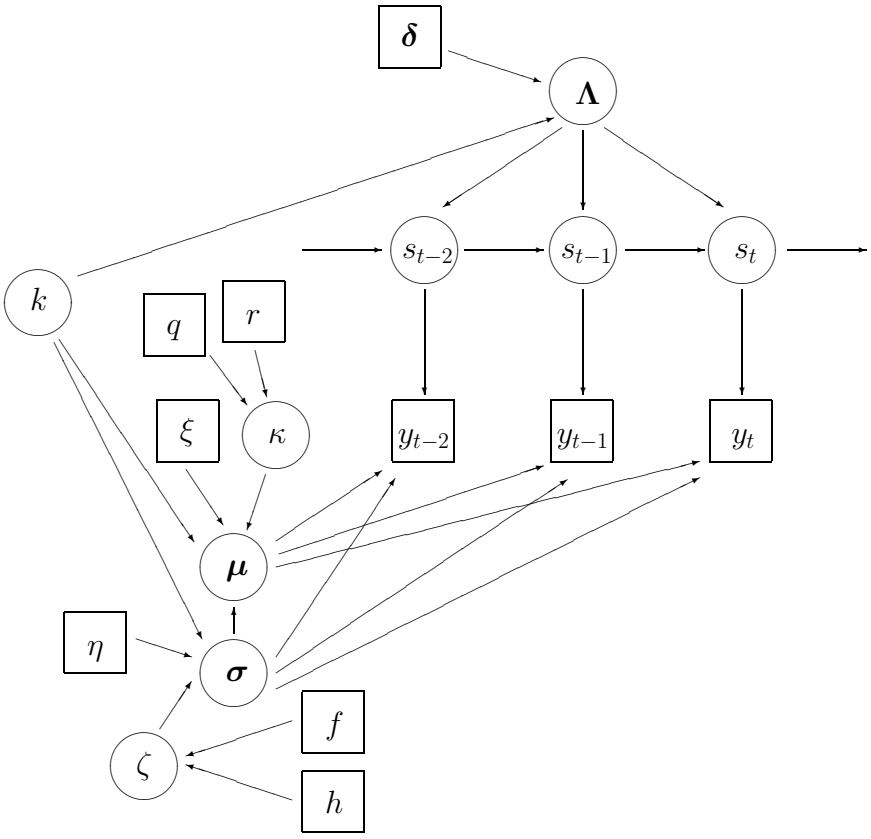

Figure 2: Directed acyclic graph for the complete hierarchical model.

The complete hierarchical model is showed in Figure 2 as a directed acyclic graph (DAG). We follow the usual convention that square boxes represent fixed or observed quantities and circles represent the unknowns.

3 Computational implementation

The complexity of the mixture model presented requires Markov Chain Monte Carlo (MCMC) methods to approximate the posterior distribution. A detailed description of these computational methods can be found in Tierney (1994) and Besag *et al.* (1995).

In order to generate realizations from the posterior joint distribution of all the parameters, we alternate the following moves at each sweep of the MCMC algorithm:

(a) update the transition matrix $\boldsymbol{\Lambda}$,

(b) update the state variables $\boldsymbol{s}$,

(c) update the hyperparameter ζ,

(d) update the standard deviations $\boldsymbol{\sigma}$,

(e) update the hyperparameter κ,

(f) update the means $\boldsymbol{\mu}$,

(g) update the number of regimes k.

The first six moves are fairly simple and all performed through Gibbs sampling. We will go through them rather quickly, while more attention will be devoted to the last move, a variable dimension move, requiring the use of the RJ method (Green, 1995).

3.1 Gibbs moves

Move (a) follows Robert *et al.* (1993): the i-th row of $\boldsymbol{\Lambda}$ is sampled from a Dirichlet distribution $D(\delta_1 + n_{i1}, \ldots, \delta_k + n_{ik})$, where

$$n_{ij} = \sum_{t=1}^{T-1} I\{s_t = i, s_{t+1} = j\}$$

is the number of transitions from regime i to regime j; $I\{\cdot\}$ is the indicator function.

In (b), the standard solution for updating the state variable would be to sample $s_1, \ldots, s_T$, one at a time from $t = 1$ to $t = T$, drawing new values from their full conditional distribution

$$p(s_t = i | \cdots) \propto \lambda_{s_{t-1} i} \phi(y_t; \mu_i, \sigma_i^2) \lambda_{i s_{t+1}}$$

where '$\cdots$' denotes 'all other variables'; for $t = 1$, the first factor is replaced by the stationary probability π_i and, for $t = T$, the last factor is replaced by 1. However, we preferred to sample $\boldsymbol{s}$ directly from $p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Lambda})$, using a stochastic version of the forward-backward recursion (see Scott, 2002). This leads to a faster mixing algorithm because fewer components are introduced into the Gibbs Markov Chain. The forward recursion produces matrices $\boldsymbol{P}_2, \ldots, \boldsymbol{P}_T$, where $\boldsymbol{P}_t = (p_{tij})$ and $p_{tij} = p(s_{t-1} = i, s_t = j|y_1, \ldots, y_t, \boldsymbol{\Lambda})$. In words, $\boldsymbol{P}_t$ is the joint distribution of $(s_{t-1} = i, s_t = j)$ given model parameters and observed data up to time t. $\boldsymbol{P}_t$ can be computed from $\boldsymbol{P}_{t-1}$ as

$$\begin{aligned} p_{tij} &\propto p(s_{t-1} = i, s_t = j, y_t|y_1, \ldots, y_{t-1}, \boldsymbol{\Lambda}) = \\ &= p(s_{t-1} = i|y_1, \ldots, y_{t-1}, \boldsymbol{\Lambda})\lambda_{ij}\phi(y_t; \mu_j, \sigma_j^2) \end{aligned}$$

where proportionality is reconciled by $\sum_i \sum_j p_{tij} = 1$ and where $p(s_{t-1} = i|y_1, \ldots, y_{t-1}, \boldsymbol{\Lambda}) = \sum_j p_{t-1,i,j}$ can be computed once $\boldsymbol{P}_{t-1}$ is known. The recursion starts computing $p(s_1 = i|y_1, \boldsymbol{\Lambda}) \propto \phi(y_1; \mu_i, \sigma_i^2)\pi_i$ and thus $\boldsymbol{P}_2$.

The stochastic backward recursion begins by drawing s_T from $p(s_T|\boldsymbol{y}, \boldsymbol{\Lambda})$, then recursively drawing s_t from the distribution proportional to column s_{t+1} of $\boldsymbol{P}_{t+1}$. In this way, the stochastic backward recursion allows to sample from $p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Lambda})$, factorizing this distribution as

$$p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Lambda}) = p(s_T|\boldsymbol{y}, \boldsymbol{\Lambda}) \prod_{t=1}^{T-1} p(s_{T-t}|s_T, \ldots, s_{T-t+1}, \boldsymbol{y}, \boldsymbol{\Lambda})$$

where

$$\begin{aligned} p(s_{T-t} = i|s_T, \ldots, s_{T-t+1}, \boldsymbol{y}, \boldsymbol{\Lambda}) &= p(s_{T-t} = i|s_{T-t+1}, y_1, \ldots, y_{T-t+1}, \boldsymbol{\Lambda}) \\ &\propto p_{T-t+1,i,s_{T-t+1}}. \end{aligned}$$

In (c) we update ζ by a Gibbs move, sampling from its full conditional:

$$\zeta|\cdots \sim \text{Ga}\left(f + k\eta, h + \sum_{i=1}^{k} \sigma_i^{-2}\right).$$

In (d) we update σ_i^2 independently using a Gibbs move, sampling σ_i^{-2} from its full conditional

$$\sigma_i^{-2}|\cdots \sim \text{Ga}\left(\eta + \frac{1}{2}(n_i + 1), \zeta + \frac{1}{2}\sum_{t:s_t=i}(y_t - \mu_i)^2 + \frac{1}{2\kappa}(\mu_i - \xi)^2\right),$$

where $n_i = \#\{t : s_t = i\}$ is the number of observations currently allocated to the i-th component of the mixture.

In (e) we update κ by a Gibbs move, sampling κ^{-1} from its full conditional:

$$\kappa^{-1}|\cdots \sim \text{Ga}\left(q+\frac{k}{2}, r+\frac{1}{2}\sum_{i=1}^{k}\frac{(\mu_i-\xi)^2}{\sigma_i^2}\right).$$

Before considering the updating of $\boldsymbol{\mu}$ in (f), we comment briefly on the issue of labeling the regimes. The whole model is, in fact, invariant to the permutation of the labels $i = 1, 2, \ldots, k$. For identifiability, Richardson and Green (1997) adopt a unique labeling in which the μ_i are in increasing numerical order. As a consequence the joint prior distribution of the μ_i is $k!$ times the product of the individual normal densities, restricted to the set $\mu_1 < \mu_2 < \ldots < \mu_k$. The μ_i can be updated by means of Gibbs sampler, drawing them independently from the distribution

$$\mu_i|\cdots \sim N\left(\frac{\kappa\sum_{t:s_t=i} y_t+\xi}{1+\kappa n_i}, \frac{\sigma_i^2\kappa}{1+\kappa n_i}\right).$$

In order to preserve the ordering constraints on the μ_i, the move is accepted provided the ordering is unchanged and rejected otherwise.

3.2 Reversible jump move

Updating the value of k implies a change of dimensionality for the components $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and the transition probability matrix $\boldsymbol{\Lambda}$. We follow the approach used by Richardson and Green (1997) consisting in a random choice between splitting an existing regime into two, and merging two existing regimes into one. The probabilities of these alternatives are b_k and $d_k = 1-b_k$, respectively, when there are currently k regimes. Of course $d_1 = 0$ and $b_K = 0$. Otherwise we choose $b_k = d_k = 0.5$, for $k = 2, 3, \ldots, K-1$.

Suppose the current state of the MCMC algorithm is characterized by $k+1$ regimes and parameters which we will indicate with the superscript "˜". For the *combine proposal* we randomly choose a pair of regimes, (i_1, i_2), adjacent in terms of the current value of their means, i.e. $\tilde{\mu}_{i_1} < \tilde{\mu}_{i_2}$, with no other $\tilde{\mu}_i$ in the interval $[\tilde{\mu}_{i_1}, \tilde{\mu}_{i_2}]$. These two regimes are merged into a new one, labeled $i^\star$, reducing the number of regimes by 1. We then reallocate all those observations y_t with $\tilde{s}_t = i_1$ or $\tilde{s}_t = i_2$ to the new regime $i^\star$ and create values for $\mu_{i^\star}, \sigma^2_{i^\star}$ in such a way that:

$$\begin{aligned}
\mu_{i^\star} &= \frac{\tilde{\pi}_{i_1}\tilde{\mu}_{i_1}+\tilde{\pi}_{i_2}\tilde{\mu}_{i_2}}{\tilde{\pi}_{i_1}+\tilde{\pi}_{i_2}} \\
\mu^2_{i^\star}+\sigma^2_{i^\star} &= \frac{\tilde{\pi}_{i_1}(\tilde{\mu}^2_{i_1}+\tilde{\sigma}^2_{i_1})+\tilde{\pi}_{i_2}(\tilde{\mu}^2_{i_2}+\tilde{\sigma}^2_{i_2})}{\tilde{\pi}_{i_1}+\tilde{\pi}_{i_2}},
\end{aligned}$$

while remaining μ_i's and σ_i^2's are copied. Then, the transition probabilities from and to the regimes involved in the move are set as

$$\begin{aligned}
\lambda_{i^\star j} &= \frac{\tilde{\pi}_{i_1}\tilde{\lambda}_{i_1 j} + \tilde{\pi}_{i_2}\tilde{\lambda}_{i_2 j}}{\tilde{\pi}_{i_1} + \tilde{\pi}_{i_2}} \qquad \forall j \neq i^\star \\
\lambda_{i i^\star} &= \tilde{\lambda}_{i i_1} + \tilde{\lambda}_{i i_2} \qquad \forall i \neq i^\star.
\end{aligned}$$

Obviously, $\lambda_{i^\star i^\star} = 1 - \sum_{j \neq i^\star} \lambda_{i^\star j}$, while the remaining λ_{ij} are unchanged. The stationary probability of the new regime $i^\star$ is then obtained as $\pi_{i^\star} = \tilde{\pi}_{i_1} + \tilde{\pi}_{i_2}$. Notice that the new HMM, characterized by k regimes, and the old one, characterized by $k+1$ regimes, both have the same first and second moments.

The *split proposal* starts by choosing a regime $i^\star$ at random. We will indicate the parameters under the new state with $k+1$ regimes using the superscript " ˜ ". The regime $i^\star$ is split into two new ones labeled i_1 and i_2, augmenting k by 1. Then we have to reallocate all those observations y_t with $s_t = i^\star$ between the two new regimes, and create values for $\tilde{\pi}_{i_1}$, $\tilde{\pi}_{i_2}$, $\tilde{\mu}_{i_1}$, $\tilde{\mu}_{i_2}$, $\tilde{\sigma}_{i_1}$, $\tilde{\sigma}_{i_2}$ and the transition probabilities from and to the regimes involved in the move. The aim is to split $i^\star$ in such a way that the dynamics of the Hidden Markov Chain are essentially preserved. We accomplish this in the same manner proposed by Robert *et al.* (2000). We generate vectors $w_0 \sim \text{Be}(2,2)$, $w_1 \sim w_1^L + (w_1^U - w_1^L)\text{Be}(1,1)$, $u_j \sim \text{Be}(a,e)$, for each $j \neq i_1, i_2$, and $v_i \sim \text{Be}(a,e)$, for each $i \neq i_1, i_2$, with shape parameters a, e and lower and upper bounds w_1^L, w_1^U given below. The stationary probabilities of the new regimes i_1 and i_2 are obtained as $\tilde{\pi}_{i_1} = w_0 \pi_{i^\star}$ and $\tilde{\pi}_{i_2} = (1-w_0)\pi_{i^\star}$, respectively, while the transition probabilities are given by

$$\begin{aligned}
&\tilde{\lambda}_{i_1 j} = \frac{u_j}{w_0}\lambda_{i^\star j}, \qquad \tilde{\lambda}_{i_2 j} = \frac{1-u_j}{1-w_0}\lambda_{i^\star j} && \forall j \neq i_1, i_2 \\
&\tilde{\lambda}_{i i_1} = v_i \lambda_{i i^\star}, \qquad \tilde{\lambda}_{i i_2} = (1-v_i)\lambda_{i i^\star}, && \forall i \neq i_1, i_2 \\
&\tilde{\lambda}_{i_1 i_2} = w_1 \left(1 - \sum_{j \neq i^\star} \frac{u_j}{w_0}\lambda_{i^\star j}\right) \\
&\tilde{\lambda}_{i_2 i_1} = \left[(1-w_1)\sum_{j \neq i^\star} u_j \lambda_{i^\star j} + w_0 w_1 - \sum_{i \neq i^\star} \gamma_i v_i \lambda_{i i^\star}\right] / (1-w_0),
\end{aligned}$$

where $\gamma_i = \pi_i / \pi_{i^\star}$, while $\tilde{\lambda}_{i_1 i_1}$ and $\tilde{\lambda}_{i_2 i_2}$ are set to make rows sum to 1. The shape parameters a and e are taken as

$$\begin{aligned}
&a = \frac{1 - w_0(1+c^2)}{c^2}, \qquad e = a\frac{1-w_0}{w_0} && \text{if } w_0 \leq \frac{1}{2}, \\
&e = \frac{1 - (1-w_0)(1+c^2)}{c^2}, \qquad a = e\frac{w_0}{1-w_0} && \text{if } w_0 > \frac{1}{2},
\end{aligned}$$

This produces a beta distribution with mean w_0, if $w_0 \leq 1/2$, and squared coefficient of variation c^2; if $w_0 > 1/2$, the distribution is a mirror (around $x = 1/2$) version of the distribution obtained for $1 - w_0 \leq 1/2$. We used $c^2 = 0.5$. In order to guarantee that the new transition probability matrix is stochastic, the upper and lower bound of w_1 are then computed as:

$$w_1^L = \max\left\{1 - \frac{1 - \sum_{i \neq i_1, i_2} \tilde{\lambda}_{ii_1}\gamma_i/w_0}{1 - \sum_{j \neq i_1, i_2} \tilde{\lambda}_{i_1 j}}\,,\, 0\right\},$$

$$w_1^U = \min\left\{1 - \frac{1 - \sum_{i \neq i_1, i_2} \tilde{\lambda}_{ii_1}\gamma_i/w_0 - \left(1 - \sum_{j \neq i_1, i_2} \tilde{\lambda}_{i_2 j}\right)(1 - w_0)/w_0}{1 - \sum_{j \neq i_1, i_2} \tilde{\lambda}_{i_1 j}}\,,\, 1\right\}.$$

Notice that, if $k = 1$, then $w_1^L = 0$ and $w_1^U = \min\{(1 - w_0)/w_0\ ,\ 1\}$. Clearly, it may happen that $w_1^L > w_1^U$, in this case there is no valid w_1 and the split move is rejected.

In order to split $\mu_{i^\star}$ and $\sigma^2_{i^\star}$, we need to generate a further two-dimensional random vector $\boldsymbol{z}$ to specify the new parameters. We use Beta distributions $z_1 \sim z_1^U \mathrm{Be}(1,1)$ and $z_2 \sim \mathrm{Be}(1,1)$ for this and set

$$\begin{aligned} \tilde{\mu}_{i_1} &= \mu_{i^\star} - z_1\sigma_{i^\star}\sqrt{\tilde{\pi}_{i_2}/\tilde{\pi}_{i_1}}, & \tilde{\mu}_{i_2} &= \mu_{i^\star} + z_1\sigma_{i^\star}\sqrt{\tilde{\pi}_{i_1}/\tilde{\pi}_{i_2}}, \\ \tilde{\sigma}^2_{i_1} &= z_2(1 - z_1^2)\sigma^2_{i^\star}\pi_{i^\star}/\tilde{\pi}_{i_1}, & \tilde{\sigma}^2_{i_2} &= (1 - z_2)(1 - z_1^2)\sigma^2_{i^\star}\pi_{i^\star}/\tilde{\pi}_{i_2}. \end{aligned}$$

where

$$z_1^U = \min\left\{\frac{\mu_{i^\star} - \mu_{i^\star - 1}}{\sigma_{i^\star}}\sqrt{\frac{\tilde{\pi}_{i_1}}{\tilde{\pi}_{i_2}}}\quad,\quad \frac{\mu_{i^\star + 1} - \mu_{i^\star}}{\sigma_{i^\star}}\sqrt{\frac{\tilde{\pi}_{i_2}}{\tilde{\pi}_{i_1}}}\quad,\quad 1\right\}$$

is the upper bound for z_1 in which the $\tilde{\mu}_i$'s are properly sorted. Notice that z_1^U is equal to the minimum between the first and the third elements in the brackets or to the minimum between the second and the third elements in the brackets, respectively, when $i^\star = k$ or $i^\star = 1$. Obviously, if $k = 1$ then $z_1^U = 1$.

Finally, we reallocate those y_t with $s_t = i^\star$ between i_1 and i_2 using the restricted backward algorithm proposed by Robert *et al.* (2000). Obviously, the s_t's different from $i^\star$ are simply copied. Let us suppose that $s_t = i^\star$ for $t_1 \leq t \leq t_2$ with $s_{t_1 - 1} \neq i^\star$ and $s_{t_2 + 1} \neq i^\star$. Then, $\tilde{s}_{t_1}, \ldots, \tilde{s}_{t_2}$ are sampled one at a time from $t = t_1$ to $t = t_2$, with conditional probabilities given by

$$p(\tilde{s}_t = i | \cdots) \propto \tilde{\lambda}_{\tilde{s}_{t-1}, i}\phi(y_t; \tilde{\mu}_i, \tilde{\sigma}_i^2) b_t(i) \qquad \text{for } i = i_1, i_2 \tag{4}$$

where '$\cdots$' denotes $\{\boldsymbol{y}, \tilde{s}_{t_1 - 1}, \tilde{s}_{t_1}, \ldots, \tilde{s}_{t+1} \in [i_1, i_2], \ldots, \tilde{s}_{t_2} \in [i_1, i_2], \tilde{s}_{t_2 + 1}, \tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}\}$, and $b_t(i) = p(y_{t+1}, \ldots, y_{t_2}, \tilde{s}_{t+1} \in [i_1, i_2], \ldots, \tilde{s}_{t_2} \in [i_1, i_2], \tilde{s}_{t_2 + 1} | \tilde{s}_t = i, \tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}})$,

for $i = i_1, i_2$. These elements can be obtained recursively as

$$b_{t_2}(i) = \tilde{\lambda}_{i,\tilde{s}_{t_2+1}} \tag{5}$$

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

$$b_t(i) = \sum_{j=i_1,i_2} b_{t+1}(j)\tilde{\lambda}_{ij}\phi(y_{t+1}; \tilde{\mu}_j, \tilde{\sigma}_j^2).$$

If $t_1 = 1$, then $\tilde{\lambda}_{\tilde{s}_{t_1-1},i}$ on the right hand side of equation (4) is replaced by the stationary probability $\tilde{\pi}_i$ and, if $t_2 = T$, the right end side of equation (5) is replaced by 1.

According to the RJ framework, the acceptance probability for the split move is $\min(1, A)$, where

$$\begin{aligned}
A = & \frac{p(\boldsymbol{y}|\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}})}{p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\sigma})} \times \frac{p(k+1)}{p(k)} \times \frac{p(\tilde{\boldsymbol{\Lambda}}|k+1, \boldsymbol{\delta})}{p(\boldsymbol{\Lambda}|k, \boldsymbol{\delta})} \times \frac{p(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{\Lambda}})}{p(\boldsymbol{s}|\boldsymbol{\Lambda})} \\
& \times (k+1)\frac{1}{\sqrt{2\pi\kappa}}\frac{\sigma_{i^\star}}{\tilde{\sigma}_{i_1}\tilde{\sigma}_{i_2}} \exp\left\{-\frac{1}{2\kappa}\left[\frac{(\tilde{\mu}_{i_1}-\xi)^2}{\tilde{\sigma}_{i_1}^2} + \frac{(\tilde{\mu}_{i_2}-\xi)^2}{\tilde{\sigma}_{i_2}^2} - \frac{(\mu_{i^\star}-\xi)^2}{\sigma_{i^\star}^2}\right]\right\} \\
& \times \frac{\zeta^\eta}{\Gamma(\eta)} \exp\left[-\zeta\left(\frac{1}{\tilde{\sigma}_{i_1}^2} + \frac{1}{\tilde{\sigma}_{i_2}^2} - \frac{1}{\sigma_{i^\star}^2}\right)\right]\left(\frac{\tilde{\sigma}_{i_1}^2\tilde{\sigma}_{i_2}^2}{\sigma_{i^\star}^2}\right)^{-\eta-1} \times \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times J \\
& \times \left[\frac{1}{z_1^U} g_{1,1}\left(\frac{z_1}{z_1^U}\right) g_{1,1}(z_2) g_{2,2}(w_0) \frac{1}{w_1^U - w_1^L} g_{1,1}\left(\frac{w_1 - w_1^L}{w_1^U - w_1^L}\right) \prod_j g_{a,e}(u_j) \prod_i g_{a,e}(v_i)\right]^{-1},
\end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function, P_{alloc} is the probability of making the allocation of the $\tilde{\boldsymbol{s}}$ that was made, $g_{a,e}$ denotes the $\text{Be}(a, e)$ density and J is the Jacobian determinant of the transformation from the parameter vector $(\lambda_{ii^\star}, v_i, \lambda_{i^\star j}, u_j, w_0, w_1, \mu_{i^\star}, z_1, \sigma_{i^\star}^{-2}, z_2)$, with $i, j \neq i^\star$, of the model with k regimes to the parameter vector $(\tilde{\lambda}_{ii_1}, \tilde{\lambda}_{ii_2}, \tilde{\lambda}_{i_1 j}, \tilde{\lambda}_{i_2 j}, \tilde{\lambda}_{i_1 i_2}, \tilde{\lambda}_{i_2 i_1}, \tilde{\mu}_{i_1}, \tilde{\mu}_{i_2}, \tilde{\sigma}_{i_1}^{-2}, \tilde{\sigma}_{i_2}^{-2})$, with $i, j \neq i_1, i_2$, of the model with $k+1$ regimes. Notice that diagonal elements of $\boldsymbol{\Lambda}$ or $\tilde{\boldsymbol{\Lambda}}$ are omitted since the row sums equal 1.

Table 1 shows the table of partial derivatives, which defines the Jacobian matrix, with $\mathbf{0}$ and $\mathbf{x}$ denoting suitable sized vectors or matrices of zero and non-zeros entries, respectively. Since the Jacobian matrix has an upper block diagonal structure, we can evaluate J as the product of the three determinants J_1, J_2 and J_3, corresponding to the three submatrices highlighted along

	$\tilde{\lambda}_{ii_1}$	$\tilde{\lambda}_{ii_2}$	$\tilde{\lambda}_{i_1j}$	$\tilde{\lambda}_{i_2j}$	$\tilde{\lambda}_{i_1i_2}$	$\tilde{\lambda}_{i_2i_1}$	$\tilde{\mu}_{i_1}$	$\tilde{\mu}_{i_2}$	$\tilde{\sigma}_{i_1}^{-2}$	$\tilde{\sigma}_{i_2}^{-2}$
$\lambda_{ii^\star}$	**x**	**x**	**0**	**0**	**0**	**x**	**0**	**0**	**0**	**0**
v_i	**x**	**x**	**0**	**0**	**0**	**x**	**0**	**0**	**0**	**0**
$\lambda_{i^\star j}$	**0**	**0**	**x**	**x**	**x**	**x**	**0**	**0**	**0**	**0**
u_j	**0**	**0**	**x**	**x**	**x**	**x**	**0**	**0**	**0**	**0**
w_0	**0**	**0**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
w_1	**0**	**0**	**0**	**0**	**x**	**x**	**0**	**0**	**0**	**0**
$\mu_{i^\star}$	**0**	**0**	**0**	**0**	**0**	**0**	**x**	**x**	**0**	**0**
z_1	**0**	**0**	**0**	**0**	**0**	**0**	**x**	**x**	**x**	**x**
$\sigma_{i^\star}^{-2}$	**0**	**0**	**0**	**0**	**0**	**0**	**x**	**x**	**x**	**x**
z_2	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**x**	**x**

Table 1: *Table of partial derivatives*

the main diagonal in Table 1. Thus

$$J_1 = \left| \begin{smallmatrix} \mathrm{diag}(v_i) & \mathrm{diag}(1-v_i) \\ \mathrm{diag}(\lambda_{ii^\star}) & -\mathrm{diag}(\lambda_{ii^\star}) \end{smallmatrix} \right| = \left| \begin{smallmatrix} \boldsymbol{I} & \mathrm{diag}(1-v_i) \\ \mathbf{0} & -\mathrm{diag}(\lambda_{ii^\star}) \end{smallmatrix} \right| = \prod_{i \neq i^\star} \lambda_{ii^\star}$$

$$J_2 = \begin{vmatrix} \mathrm{diag}\left(\frac{u_j}{w_0}\right) & \mathrm{diag}\left(\frac{1-u_j}{1-w_0}\right) & -\mathrm{col}\left(\frac{w_1 u_j}{w_0}\right) & \mathrm{col}\left(\frac{(1-w_1)u_j - \sum_{i \neq i^\star} v_i \lambda_{ii^\star} \partial \gamma_i / \partial \lambda_{i^\star j}}{1-w_0}\right) \\ \mathrm{diag}\left(\frac{\lambda_{i^\star j}}{w_0}\right) & -\mathrm{diag}\left(\frac{\lambda_{i^\star j}}{1-w_0}\right) & -\mathrm{col}\left(\frac{w_1 \lambda_{i^\star j}}{w_0}\right) & \mathrm{col}\left(\frac{(1-w_1)\lambda_{i^\star j}}{1-w_0}\right) \\ -\mathrm{row}\left(\frac{u_j \lambda_{i^\star j}}{w_0^2}\right) & \mathrm{row}\left(\frac{(1-u_j)\lambda_{i^\star j}}{(1-w_0)^2}\right) & \frac{w_1(1-\tilde{\lambda}_{i_1 i_1} - \tilde{\lambda}_{i_1 i_2})}{w_0} & \frac{w_1 + \tilde{\lambda}_{i_2 i_1}}{1-w_0} \\ \mathbf{0} & \mathbf{0} & \tilde{\lambda}_{i_1 i_1} + \tilde{\lambda}_{i_1 i_2} & \frac{w_0(\tilde{\lambda}_{i_1 i_1} + \tilde{\lambda}_{i_1 i_2})}{1-w_0} \end{vmatrix}$$

$$J_3 = \begin{vmatrix} -\sigma_{i^\star}\sqrt{\frac{1-w_0}{w_0}} & \sigma_{i^\star}\sqrt{\frac{w_0}{1-w_0}} & \frac{2 z_1 w_0}{\sigma_{i^\star}^2 z_2 (1-z_1^2)^2} & \frac{2 z_1 (1-w_0)}{\sigma_{i^\star}^2 (1-z_2)(1-z_1^2)^2} \\ \frac{z_1 \sigma_{i^\star}^3}{2}\sqrt{\frac{1-w_0}{w_0}} & -\frac{z_1 \sigma_{i^\star}^3}{2}\sqrt{\frac{w_0}{1-w_0}} & \frac{w_0}{z_2(1-z_1^2)} & \frac{(1-w_0)}{(1-z_2)(1-z_1^2)} \\ 0 & 0 & \frac{w_0}{\sigma_{i^\star}^2 z_2^2 (1-z_1^2)} & \frac{1-w_0}{\sigma_{i^\star}^2 (1-z_2)^2 (1-z_1^2)} \end{vmatrix} = \frac{\sqrt{w_0(1-w_0)}}{\sigma_{i^\star} z_2^2 (1-z_2)^2 (1-z_1^2)^3}$$

where the determinant J_2 is evaluated numerically in the same way discussed in Robert *et al.* (2000).

4 Bayesian inference and forecasting

The RJ approach allows joint (or across-model) inferences about the number of regimes and all the other parameters of the model. Given a burn-in period, to guarantee the convergence of the chain to its stationary distribution, the algorithm produces at each sweep a new realization of all the parameters of the model, sampled from their joint posterior distribution. Let us denote by $(k^{(n)}, \boldsymbol{\Lambda}^{(n)}, \boldsymbol{s}^{(n)}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\sigma}^{(n)}, \kappa^{(n)}, \zeta^{(n)})$ for $n = 1, \ldots, N$, the sample obtained after N sweeps of the algorithm. This sample provides the simulated joint posterior distribution of all parameters and can be used to estimate all the quantities of interest.

4.1 Model choice and parameters estimation

From the RJ output, we can easily estimate the posterior distribution of the number of regimes as the proportion of times each model was visited by the algorithm, i.e.:

$$\hat{p}(k = \ell|\boldsymbol{y}) = \sum_{n=1}^{N} I\{k^{(n)} = \ell\}/N = N_\ell/N,$$

where $I\{\cdot\}$ is the indicator function and N_ℓ is the number of times the model with ℓ components was visited by the algorithm.

Conditioning on a particular model (i.e. M_ℓ, the one with ℓ regimes) we can estimate any other parameter of that particular model. Notice that, in the following, we will drop the conditioning on the model in order to have a short hand notation. We will explicitly include the conditioning in the notation, only when required to avoid confusion.

For example, estimating the hidden states $\boldsymbol{s}$ is often the central question in applied problems. All the Bayes estimates of $\boldsymbol{s}$ derive from its posterior distribution $p(\boldsymbol{s}|\boldsymbol{y})$, a high-dimensional distribution that must be summarized to be understood. In general it is sufficient to summarize it through its marginal distributions $p(s_t = i|\boldsymbol{y})$, whose obvious estimates are

$$\hat{p}(s_t = i|\boldsymbol{y}) = \sum_{n:k^{(n)}=\ell} I\{s_t^{(n)} = i\}/N_\ell.$$

More efficient estimates can be obtained through Rao-Blackwellization (Gelfand and Smith, 1990; Casella and Robert, 1996) as:

$$\tilde{p}(s_t = i|\boldsymbol{y}) = \sum_{n:k^{(n)}=\ell} p(s_t = i|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)})/N_\ell.$$

This approach requires that the nonstochastic backward recursion also produces probabilities $p(s_t = i|\boldsymbol{y}, \boldsymbol{\Lambda})$, demanding little effort once the forward recursion has been implemented. These probabilities can be, in fact, computed as the appropriate margin of the matrix $\boldsymbol{P}'_{t+1} = (p'_{t+1,i,j})$, where $p'_{t,i,j} = p(s_{t-1} = i, s_t = j|\boldsymbol{y}, \boldsymbol{\Lambda})$. In practice, $\boldsymbol{P}'_t$ conditions on all of the observed data, whereas $\boldsymbol{P}_t$ conditions only on data observed up to time t. Clearly, $\boldsymbol{P}_T = \boldsymbol{P}'_T$, hence $\boldsymbol{P}'_t$ can be easily obtained from $\boldsymbol{P}_t$ and $\boldsymbol{P}'_{t+1}$, using Bayes's rule,

$$\begin{aligned} p'_{t,i,j} &= p(s_{t-1} = i|s_t = j, \boldsymbol{y}, \boldsymbol{\Lambda}) p(s_t = j|\boldsymbol{y}, \boldsymbol{\Lambda}) \\ &= p(s_{t-1} = i|s_t = j, y_1, \ldots, y_t, \boldsymbol{\Lambda}) \sum_{h=1}^{k} p'_{t+1,j,h} = p_{tij} \frac{\sum_{h=1}^{k} p'_{t+1,j,h}}{\sum_{h=1}^{k} p_{t,h,j}}. \end{aligned}$$

Assigned a quadratic loss function, we can compute the point estimates of any other parameter of the model as the mean of its simulated marginal posterior distribution, conditional on the model with ℓ regimes. Point estimates for the means and the variances of the ℓ regimes and the transition matrix are given respectively by:

$$\hat{\boldsymbol{\mu}} = \sum_{n:k^{(n)}=\ell} \boldsymbol{\mu}^{(n)}/N_\ell, \qquad \hat{\boldsymbol{\sigma}}^2 = \sum_{n:k^{(n)}=\ell} \boldsymbol{\sigma}^{2(n)}/N_\ell, \qquad \hat{\boldsymbol{\Lambda}} = \sum_{n:k^{(n)}=\ell} \boldsymbol{\Lambda}^{(n)}/N_\ell,$$

where the sums of vectors are taken element-wise.

4.2 Bayesian forecasting

Generally, when data indexed by time are analyzed, the main purpose is to forecast the future values of the observed variable, given the information available up to time T. In a Bayesian context, inferences are based on the posterior predictive density of these observations. Suppose we are interested in forecasting the vector $\boldsymbol{Y} = (y_{T+1}, \ldots, y_{T+G})$. Posterior predictive density for this vector can be defined into two different ways, depending on what we consider as "information available at time T". If we believe that data are generated by a specific model, let us say M_ℓ, the information available at time T will encompass the generating model and the observed data up to time T. Then, the posterior predictive density for $\boldsymbol{Y}$ will be defined as:

$$p(\boldsymbol{Y}|\boldsymbol{y}, M_\ell) = \int_{\Theta_{M_\ell}} p(\boldsymbol{Y}|\boldsymbol{y}, \boldsymbol{\theta}_{M_\ell}, M_\ell) p(\boldsymbol{\theta}_{M_\ell}|\boldsymbol{y}, M_\ell) d\boldsymbol{\theta}_{M_\ell}, \tag{6}$$

where $\boldsymbol{\theta}_{M_\ell}$ is the vector of all parameters (except k), including the state variable, under the model with ℓ regimes and Θ_{M_ℓ} is the relative parameter space.

Otherwise, if we are not certain about the true generating model within a set of possible ones, we can take into account this uncertainty and define the posterior predictive density of $\boldsymbol{Y}$ through model averaging as:

$$p(\boldsymbol{Y}|\boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{Y}|\boldsymbol{y}, M_k) p(M_k|\boldsymbol{y}), \tag{7}$$

where $p(\boldsymbol{Y}|\boldsymbol{y}, M_k)$ is defined in (6). Notice that in (7), we are considering as available information at time T only the observed data.

Regardless the definition considered for the posterior predictive density, we cannot compute it analytically but we can easily simulate it as a by-product of the MCMC algorithm. Let us consider the definition in (6), first. We can draw values from $p(\boldsymbol{Y}|\boldsymbol{y}, M_\ell)$ through the following steps:

1. conditioning on the model with ℓ regimes, i.e. for those sweeps n of the MCMC algorithm in which $k^{(n)} = \ell$, draw a vector $(s_{T+1}^{(n)}, \ldots, s_{T+G}^{(n)})$ from $p(s_{T+1}, \ldots, s_{T+G}|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)})$;

2. conditioning on $s_{T+g}^{(n)} = i$, draw a value of $y_{T+g}^{(n)}$ from $N(\mu_i^{(n)}, \sigma_i^{2(n)})$, for $g = 1, \ldots, G$.

To perform step 1, notice that we can use the following decomposition:

$$p(s_{T+1}, \ldots, s_{T+G}|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) = p(s_{T+1}|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) \prod_{g=2}^{G} p(s_{T+g}|s_{T+g-1}, \boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}).$$

Thus, we can draw s_{T+1} from the probability distribution

$$\begin{aligned} p(s_{T+1} = j|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) &= \sum_{i=1}^{k^{(n)}} p(s_{T+1} = j|s_T = i, \boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) p(s_T = i|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) = \\ &= \sum_{i=1}^{k^{(n)}} \lambda_{ij} p(s_T = i|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}), \end{aligned}$$

with $p(s_T = i|\boldsymbol{y}, \boldsymbol{\Lambda}^{(n)})$ given in Section 3.1, and then we iteratively draw s_{T+g} from the probability distribution $p(s_{T+g} = j|s_{T+g-1}^{(n)}, \boldsymbol{y}, \boldsymbol{\Lambda}^{(n)}) = \lambda_{s_{T+g-1},j}^{(n)}$, for $g = 2, \ldots, G$.

Drawing from the posterior predictive distribution in (7) is simply performed repeating steps 1 and 2 at each sweep of the algorithm, regardless the model visited at that sweep.

5 Conclusions

HMMs are a flexible class of models useful to represent dependent heterogeneous phenomena. In this paper we illustrated Bayesian inference for HMMs with an unknown number of regimes. In our formulation, the data are independent realizations, conditional on the state variable, from a normal distribution with mean and variance depending on the value assumed by the state variable. With this respect, we extended the approach of Robert *et al.* (2000), which assume the data as independently generated by a zero mean normal distribution, conditional on the state variable.

We considered a hierarchical model which allowed us to make vague *a priori* assumptions on the parameters of the model. The analytically untractable joint posterior distribution of all the parameters of the model,

including the unknown number of regimes was simulated through MCMC methods and, in particular, making use of the RJ algorithm which allows for the changing dimension of the parameter space. The updating of the unobserved state variable was illustrated in detail and performed through a stochastic forward-backward recursion aimed at improving the mixing of the chain, compared to the standard Gibbs sampling algorithm, and providing, as a by-product, efficient estimates of the posterior marginal distribution of the state variable itself.

We illustrated how to choose an appropriate model for the data and obtain point estimates for its parameters on the basis of the MCMC output. Finally, we showed how the posterior predictive density of future observations can be simulated with a little effort through the MCMC algorithm, either conditioning the forecast on a particular model or averaging over different models, to take into account the uncertainty about the true number of regimes.

References

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science*, **10**, 3–66.

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

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

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

Engel, C. and Hamilton, J. D. (1990). Long swings in the dollar: are they in the data and do markets know it?, *American Economic Review*, **80**, 689–713.

Fredkin, D. R. and Rice, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings, *Proceedings: Biological Sciences*, **249**, 125–132.

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

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process, *Scandinavian Journal of Statistics*, **28**, 355–375.

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

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*, Chapman and Hall, London.

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

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

Otranto, E. and Gallo, G. (2002). A nonparametric Bayesian approach to detect the number of regimes in Markov switching models, *Econometric Reviews*, **21**, 477–496.

Rabiner, L. R. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**, 257–286.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, B*, **59**, 731–792.

Robert, C. P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of Hidden Markov chains: a stochastic implementation, *Statistics and Probability letters*, **16**, 77–83.

Robert, C. P., Rydén, T. and Titterington, D. M. (2000). Bayesian inference in Hidden Markov models through the reversible jump Markov chain Monte Carlo method, *Journal of the Royal Statistical Society, B*, **62**, 57–75.

Rossi, A. and Gallo, G. M. (2006). Volatility estimation via Hidden Markov models, *Journal of Empirical Finance*, **13**, 203–230.

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

Scott, S. L. (2002). Bayesian methods for Hidden Markov models: recursive computing in the 21st century, *Journal of the American Statistical Association*, **97**, 337–351.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics*, **22**, 1701–1762.