An alternative to the Baum-Welch recursions for hidden Markov models

Bartolucci, Francesco

31 December 2011

An alternative to the Baum-Welch recursions for hidden Markov models

Francesco Bartolucci*†

Abstract

We develop a recursion for hidden Markov model of any order h, which allows us to obtain the posterior distribution of the latent state at every occasion, given the previous h states and the observed data. With respect to the well-known Baum-Welch recursions, the proposed recursion has the advantage of being more direct to use and, in particular, of not requiring dummy renormalizations to avoid numerical problems. We also show how this recursion may be expressed in matrix notation, so as to allow for an efficient implementation, and how it may be used to obtain the manifest distribution of the observed data and for parameter estimation within the Expectation-Maximization algorithm. The approach is illustrated by an application to financial data which is focused on the study of the dynamics of the volatility level of log-returns.

KEYWORDS: Expectation-Maximization algorithm, forward-backward recursions, latent Markov model, stochastic volatility

*Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia. E-mail: bart@stat.unipg.it

†I acknowledge the financial support from the "Einaudi for Economics and Finance" (Rome - IT)

1 Introduction

Hidden Markov (HM) models have become a popular statistical tool for the analysis of data having a time-series structure; for an up-to-date review see Zucchini and MacDonald (2009). These models have also found great interest for the analysis of longitudinal data, where independent short time series are observed for typically many statistical units; for a review see Bartolucci et al. (2010). HM models are based on the assumption that the observable random variables, corresponding to the different time occasions, are conditionally independent given an unobservable (or latent) process, which follows a Markov chain. Usually, this Markov chain is assumed to be of first order and time homogenous, so that the transition probabilities are time invariant.

A fundamental tool of inference for HM models is represented by forward-backward recursions of Baum and Welch (see Baum et al., 1970; Welch, 2003). For a first-order HM model, these recursions allow us to compute the manifest probability (or density) of the observed sequence of data and to obtain the posterior distribution of every latent state and of every pair of consecutive latent states given these data. Through this recursion is then possible to implement an Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977) for maximum likelihood estimation of the parameters and to perform local decoding (Juang and Rabiner, 1991), that is to find the most likely state at every occasion, given the observed data. Despite its popularity, the Baum-Welch recursions may suffer from numerical problems due to the fact that certain probabilities may become negligible. This problem typically requires to implement dummy renormalizations; see Scott (2002) for further comments and Lystig and Hughes (2002) for an alternative solution in dealing with the manifest distribution of the observed data.

In a rather recent paper, Bartolucci and Besag (2002) proposed a probabilistic result to obtain the marginal distribution of a random variable in Markov random field model and mentioned that this result may be also used for HM models, providing an example for a first-order and a second-order HM model. Developing the intuition of Bartolucci and Besag (2002), in this paper we propose a general recursion to deal with HM models of any order h. This recursion allows us to obtain the posterior distribution of every latent state given the previous h states and the observed data. With respect to the Baum-Welch recursions, the proposed recursion has the advantage of being more direct to use, especially with higher-order HM models. Moreover, it does not require dummy renormalizations.

We show how the proposed recursion may be used to obtain the manifest distribution of the observed data and the required posterior probabilities to implement the EM algorithm for parameter estimation. Moreover, the recursion may be directly used for local decoding and for prediction. In order to allow for an efficient implementation, we also express the proposed result in matrix notation. Such an implementation in the R language is available to the reader upon request.

The remainder of the paper is organized as follows. In the following section we briefly review HM models and the Baum-Welch recursion. The proposed recursion is illustrated in Section 3, whereas in Section 4 we illustrate its use for maximum likelihood estimation, local decoding, and prediction. Finally, in Section 5 we provide an illustration by an application based on an HM version of the stochastic volatility (SV) model for financial data (Taylor, 2005), in which we assume the existence of discrete levels of volatility.

2 Preliminaries

Consider a sequence of T *manifest random variables* $Y_1, \dots, Y_T$ which are collected in the vector $\boldsymbol{Y}$. A hidden Markov (HM) model assumes that these random variables are conditional independent given the *unobservable random variables* $U_1, \dots, U_T$ which follow a Markov chain with k states. We consider in particular a Markov chain of order h so that

$$p(u_t|u_1,\dots,u_{t-1}) = p(u_t|u_{t-h},\dots,u_{t-1}), \quad t = h+1,\dots,T,$$

where we use the notation $p(u_t|u_1,\dots,u_{t-1}) = P(U_t = u_t|U_1 = u_1,\dots,U_{t-1} = u_{t-1})$. A similar notation will be adopted throughout the paper to denote probability mass functions, in a way that will be clear from the context. It is also assumed that every Y_t depends on the latent process only through U_t and then by $f(y_t|u_t)$ we denote the probability mass (or density) function of this distribution.

The specific HM model adopted in an application is based on assumptions on the above transition probabilities, such as that these probabilities are time homogeneous. These assumptions may also concern the distribution of each response variable given the corresponding latent variable. The specific formulation may also involve covariates, if available. In this section, however, we remain in the general context described above and base most results on the unspecified transition probability function $p(u_t|u_{\max(t-h,1)},\dots,u_{t-1})$ and the conditional response probability (or density) function $f(y_t|u_t)$. Note that, in denoting the transition probabilities, we use the index $\max(t-h,1)$ in order to have a notation that is suitable even for $t < h$. Obviously, when $t = 1$, the conditioning argument in these probabilities vanishes and they reduce to initial probabilities of type $p(u_1)$.

The following example clarifies a possible formulation of an HM model for time-series data. For other examples in the context of longitudinal data see Bartolucci et al. (2010).

Example 1 *Consider an HM version of the SV model for financial data (Taylor, 2005), which is based on the assumption that, given U_t, the log-return Y_t has a normal distribution with mean 0 and variance depending on U_t. In particular, we assume that*

$$f(y_t|u_t) = \frac{1}{\sqrt{2\pi\sigma_{u_t}^2}} \exp\left[-\frac{1}{2}\left(\frac{y_t}{\sigma_{u_t}}\right)^2\right],$$

where σ_v, $v = 1,\dots,k$, are volatility levels associated to the different latent states. We also assume that the underlying Markov chain is of order h and is time-homogenous, so *that, for all $t > h$, we have*

$$p(u_t|u_{t-h},\ldots,u_{t-1}) = \pi_{u_{t-h},\ldots,u_t},$$

where $\pi_{v_1,\ldots,v_{h+1}}$, $v_1,\ldots,v_{h+1} = 1,\ldots,k$, are common transition probabilities to be estimated together with $\sigma_1,\ldots,\sigma_k$. Other parameters to be estimated are the initial and transition probabilities for $t \leqslant h$. These parameters are denoted by

$$\lambda_{t,u_{\max(t-h,1)},\ldots,u_t} = p(u_t|u_{\max(t-h,1)},\ldots,u_{t-1}).$$

Overall, taking into account that the initial probabilities are such that $\sum_{u_1}\lambda_{1,u_1} = 1$ and similar constraints hold for all transition probabilities, the number of free parameters is

$$\#par = \underbrace{k}_{\sigma_v} + \underbrace{(k-1)\sum_{t=1}^{h-1} k^{t-1}}_{\lambda_{t,u_{\max(t-h,1)},\ldots,u_t}} + \underbrace{(k-1)k^h}_{\pi_{v_1,\ldots,v_{h+1}}}. \tag{1}$$

It has to be clear that the same modeling framework described above may be adopted with longitudinal data in which we observe short sequences of data for n sample units, which are usually assumed to be independent. However, we do not explicitly consider the case of longitudinal data since the theory that will be developed easily apply to this case as well.

In order to efficiently compute the probability (or the density) of an observed sequence of T observations, collected in the vector $\boldsymbol{y} = (y_1,\ldots,y_T)$, Baum and Welch (Baum et al., 1970; Welch, 2003) proposed the following forward recursion for a first-order HM model:

$$f(u_t,\boldsymbol{y}_{\leqslant t}) = \sum_{u_{t-1}} f(u_{t-1},\boldsymbol{y}_{\leqslant t-1})p(u_t|u_{t-1})f(y_t|u_t), \quad t = 2,\ldots,T, \tag{2}$$

where $\boldsymbol{y}_{\leqslant t} = (y_1,\ldots,y_t)$. This recursion is initialized with $f(u_1,y_1) = p(u_1)f(y_1|u_1)$ and, in the end, we obtain the manifest probability (or density) function of $\boldsymbol{y}$ as

$$f(\boldsymbol{y}) = \sum_{u_t} f(u_t,\boldsymbol{y}).$$

Moreover, Baum and Welch introduced the backward recursion

$$f(\boldsymbol{y}_{>t}|u_t) = \sum_{u_{t+1}} f(\boldsymbol{y}_{>t+1}|u_{t+1}) p(u_{t+1}|u_t) f(y_{t+1}|u_{t+1}), \quad t = 1, \dots, T-1, \tag{3}$$

where $\boldsymbol{y}_{>t} = (y_{t+1}, \dots, y_T)$, which is initialized with $f(\boldsymbol{y}_{>T}|u_t) = 1$. Using this recursion, we can obtain the posterior probability of every latent state given the observed data, that is $q(u_t|\boldsymbol{y}) = P(U_t = u_t|\boldsymbol{Y} = \boldsymbol{y})$. In particular, we have

$$q(u_t|\boldsymbol{y}) = \frac{f(u_t, \boldsymbol{y}_{\leqslant t}) f(\boldsymbol{y}_{>t}|u_t)}{f(\boldsymbol{y})}, \quad t = 1, \dots, T,$$

whereas for the posterior probability of every pair of consecutive states we have the posterior probability

$$q(u_{t-1}, u_t|\boldsymbol{y}) = \frac{f(u_{t-1}, \boldsymbol{y}_{\leqslant t}) p(u_t|u_{t-1}) f(y_t|u_t) f(\boldsymbol{y}_{>t}|u_t)}{f(\boldsymbol{y})}, \quad t = 2, \dots, T.$$

As mentioned above, the Baum-Welch recursions suffer from the problem of numerical instability due to the fact that, as t increases, the probability in (2) becomes negligible. The problem is evident when T is large and also affects the probabilities in (3). This problem requires suitable renormalizations; see Scott (2002) for a more detailed description.

3 Proposed recursion

Developing a result due to Bartolucci and Besag (2002) for Markov random fields, in this section we propose how to compute the posterior probabilities

$$q(u_t|u_{\max(t-h,1)}, \dots, u_{t-1}, \boldsymbol{y}), \quad t = 1, \dots, T, \tag{4}$$

that is the conditional probability of a certain realization of U_t, given $U_{\max(t-h,1)}, \dots, U_{t-1}$ and a certain configuration of responses collected in the vector $\boldsymbol{y}$.

For last time occasion, that is when $t = T$, the above probability may be simply computed as

$$q(u_T|u_{\max(T-h,1)}, \dots, u_{T-1}, \boldsymbol{y}) = \frac{f(y_T|u_T) p(u_T|u_{\max(T-h,1)}, \dots, u_{T-1})}{c(u_{\max(T-h,1)}, \dots, u_{T-1}, y_T)}, \tag{5}$$

where $c(u_{\max(t-h,1)},\dots,u_{T-1},y_T)$ is the normalizing constant equal to the sum of the numerator of (5) for all the possible values of U_T.

Now consider the following Theorem that allows us to compute the conditional probability in (4) for t smaller than T and is related to Theorem 1 of Bartolucci and Besag (2002).

Theorem 1 *We have that*

$$
\begin{aligned}
& q(u_t|u_{\max(t-h,1)},\dots,u_{t-1},u_{t+1},\dots,u_{t+j},\boldsymbol{y}) = \\
& = \left[\sum_{u_{t+j+1}} \frac{q(u_{t+j+1}|u_{\max(t+j+1-h,1)},\dots,u_{t+j},\boldsymbol{y})}{q(u_t|u_{\max(t-h,1)},\dots,u_{t-1},u_{t+1},\dots,u_{t+j+1},\boldsymbol{y})}\right]^{-1}, \qquad (6)
\end{aligned}
$$

for $t=1,\dots,T-1$ *and* $j=0,\dots,\min(h,T-t)-1$ *and where the conditioning variables* $u_{t+1},\dots,u_{t+j}$ *at lhs vanishes for* $j=0$.

Proof First of all consider that the assumption that the latent Markov process is of order h implies that

$$
q(u_{t+j+1}|u_{\max(t+j+1-h,1)},\dots,u_{t+j},\boldsymbol{y}) = q(u_{t+j+1}|u_{\max(t-h,1)},\dots,u_{t+j},\boldsymbol{y})
$$

and then we have

$$
\frac{q(u_{t+j+1}|u_{\max(t+j+1-h,1)},\dots,u_{t+j},\boldsymbol{y})}{q(u_t|u_{\max(t-h,1)},\dots,u_{t-1},u_{t+1},\dots,u_{t+j+1},\boldsymbol{y})} = \frac{p(u_{\max(t-h,1)},\dots,u_{t-1},u_{t+1},\dots,u_{t+j+1},\boldsymbol{y})}{p(u_{\max(t-h,1)},\dots,u_{t+j},\boldsymbol{y})}.
$$

Consequently, the sum in (6) is equal to

$$
\frac{p(u_{\max(t-h,1)},\dots,u_{t-1},u_{t+1},\dots,u_{t+j},\boldsymbol{y})}{p(u_{\max(t-h,1)},\dots,u_{t+j},\boldsymbol{y})}
$$

and the Theorem holds. □

On the basis of the above result, we implement a backward recursion finalized to computing the probabilities in (4). As already mentioned, for $t=T$ these probabilities may be directly obtained from (5). Then, in reverse order for $t = 1, \dots, T-1$ we first compute the posterior probabilities

$$q(u_t | u_{\max(t-h,1)}, \dots, u_{t-1}, u_{t+1}, \dots, u_{t+j}, \boldsymbol{y}),$$

with $j = \min(T-t, h)$. Since U_t is conditionally independent of $Y_1, \dots, Y_{t-1}, Y_{t+1}, \dots, Y_T$ given $U_{\max(t-h,1)}, \dots, U_{t-1}$, $U_{t+1}, \dots, U_{t+j}$, and Y_t, we have that the above probability is equal to

$$\frac{f(y_t|u_t) \prod_{l=0}^{j} p(u_{t+l}|u_{\max(t+l-h,1)}, \dots, u_{t+l-1})}{c(u_{\max(t-h,1)}, \dots, u_{t-1}, u_{t+1}, \dots, u_{t+j}, y_t)}. \tag{7}$$

The normalizing constant at the denominator is obtained by summing the numerator for all possible values of u_t. Then we apply result (6) from $j = \min(T-t, h)-1$ to $j = 0$, so as to recursively remove the dependence of U_t on U_{t+j+1} and obtaining the target posterior probabilities $q(u_t|u_{\max(t-h,1)}, \dots, u_{t-1}, \boldsymbol{y})$.

In order to clarify the above algorithm, we explicit consider below the case of a first-order and a second-order HM model.

Example 2 *For a first-order model ($h = 1$), the algorithm consists of first computing the probabilities*

$$q(u_T|u_{T-1}, y_T) = \frac{f(y_T|u_T)p(u_T|u_{T-1})}{c(u_{T-1}, y_T)}.$$

Then, we for $t = 1, \dots, T-1$ we apply the rule in (6) in reverse order. In particular, for $T \geqslant 3$, we have

$$q(u_t|u_{t-1}, \boldsymbol{y}) = \left[\sum_{u_{t+1}} \frac{q(u_{t+1}|u_t, \boldsymbol{y})}{q(u_t|u_{t-1}, u_{t+1}, \boldsymbol{y})}\right]^{-1}, \quad t = 2, \dots, T-1,$$

and

$$q(u_1|\boldsymbol{y}) = \left[\sum_{u_2} \frac{q(u_2|u_1, \boldsymbol{y})}{q(u_1|u_2, \boldsymbol{y})}\right]^{-1},$$

where

$$\begin{aligned} q(u_1|u_2, y_1) &= \frac{f(y_1|u_1)p(u_2|u_1)}{c(u_2, y_1)}, \\ q(u_t|u_{t-1}, u_{t+1}, y_t) &= \frac{f(y_t|u_t)p(u_t|u_{t-1})p(u_{t+1}|u_t)}{c(u_{t-1}, u_{t+1}, y_t)}, \quad t = 2, \dots, T-1. \end{aligned}$$

Example 3 *For a second-order model ($h = 2$), the algorithm consists of first computing the probabilities*

$$q(u_T|u_{T-2}, u_{T-1}, y_T) = \frac{f(y_T|u_T)p(u_T|u_{T-2}, u_{T-1})}{c(u_{T-2}, u_{T-1}, y_T)}.$$

Then, for $t = 1, \ldots, T-1$ we apply the rule in (6) for $j = 2$ (provided that $t \leqslant T-2$) and then for $j = 1$. In particular, assuming that $T \geqslant 4$, we first compute

$$q(u_{T-1}|u_{T-3}, u_{T-2}, u_T, y_T) = \frac{f(y_{T-1}|u_{T-1})p(u_{T-1}|u_{T-3}, u_{T-2})p(u_T|u_{T-2}, u_{T-1})}{c(u_{T-3}, u_{T-2}, u_T, y_T)}$$

and consequently

$$q(u_{T-1}|u_{T-3}, u_{T-2}, \boldsymbol{y}) = \left[\sum_{u_T} \frac{q(u_T|u_{T-2}, u_{T-1}, \boldsymbol{y})}{q(u_{T-1}|u_{T-3}, u_{t-2}, u_T, \boldsymbol{y})}\right]^{-1}.$$

Then in reverse order for $T = 3, \ldots, T-2$, we first compute

$$\begin{aligned} q(u_t|u_{t-2}, u_{t-1}, u_{t+1}, u_{t+2}, y_t) &= \frac{f(y_t|u_t)p(u_t|u_{t-2}, u_{t-1})p(u_{t+1}|u_{t-1}, u_t)}{c(u_{t-2}, u_{t-1}, u_{t+1}, u_{t+2}, y_t)} \times \\ &\quad \times p(u_{t+1}|u_{t-1}, u_t)p(u_{t+2}|u_t, u_{t+1}), \end{aligned}$$

we remove the dependence of U_t on U_{t+2} by computing

$$q(u_t|u_{t-2}, u_{t-1}, u_{t+1}, \boldsymbol{y}) = \left[\sum_{u_{t+2}} \frac{q(u_{t+2}|u_t, u_{t+1}, \boldsymbol{y})}{q(u_t|u_{t-2}, u_{t-1}, u_{t+1}, u_{t+2}, \boldsymbol{y})}\right]^{-1},$$

and finally we remove the dependence on U_{t+1} by computing

$$q(u_t|u_{t-2}, u_{t-1}, \boldsymbol{y}) = \left[\sum_{u_{t+1}} \frac{q(u_{t+1}|u_{t-1}, u_t, \boldsymbol{y})}{q(u_t|u_{t-2}, u_{t-1}, u_{t+1}, \boldsymbol{y})}\right]^{-1}.$$

In the end, we use similar rules to obtain $q(u_2|u_1, \boldsymbol{y})$ and consequently $q(u_1|\boldsymbol{y})$ on the basis of

$$\begin{aligned} q(u_1|u_2, u_3, y_1) &= \frac{f(y_1|u_1)p(u_2|u_1)p(u_3|u_1, u_2)}{c(u_2, u_3, y_1)}, \\ q(u_2|u_1, u_3, u_4, y_2) &= \frac{f(y_2|u_2)p(u_2|u_1)p(u_3|u_1, u_2)p(u_4|u_2, u_3)}{c(u_2, u_3, y_2)}. \end{aligned}$$

A crucial point is applying these recursions is the efficient implementation. At this regard, it is worth noting that for the first-order HM model we can express the recursion in matrix notation and then efficiently implement it in languages such as MATLAB and R. Details on this are provided in Appendix.

4 Maximum likelihood estimation using the proposed recursion

Given a sequence of observations $y_1, \ldots, y_T$ collected in $\boldsymbol{y}$, the *model log-likelihood* is

$$\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}) \tag{8}$$

where $\boldsymbol{\theta}$ is vector collecting all model parameters. The structure of $\boldsymbol{\theta}$ depends on the specific parametrization which is adopted for the conditional response distribution $f(y_t|u_t)$ and the initial and transition probabilities $p(u_t|u_{\max(t-h,1)}, \ldots, u_{t-1})$. For instance, for the HM-SV model illustrated in Example 1, $\boldsymbol{\theta}$ includes the initial and transition probabilities $\lambda_{t,u_{\max(t-h,1)},\ldots,u_t}$ and $\pi_{v_1,\ldots,v_{h+1}}$ and the standard deviations σ_v. We recall that, in this case, the probabilities $\pi_{v_1,\ldots,v_{h+1}}$ are common to all $t > h$, being the underlying Markov chain time homogenous.

In the following, we show how to compute the log-likelihood in (8) and implement its maximization by the recursion developed in the previous section. It has to be clear that the same algorithm may be used in with longitudinal data, even in the presence of individual covariates.

First of all, for any sequence of latent states $u_1, \ldots, u_T$ collected in $\boldsymbol{u}$, we simply have that

$$p(\boldsymbol{y}) = \frac{f(\boldsymbol{u}, \boldsymbol{y})}{q(\boldsymbol{u}|\boldsymbol{y})} = \frac{\prod_t f(y_t|u_t)p(u_t|u_{\max(t-h,1)}, \ldots, u_{t-1})}{\prod_t q(u_t|u_{\max(t-h,1)}, \ldots, u_{t-1}, \boldsymbol{y})},$$

where $f(\boldsymbol{u}, \boldsymbol{y})$ refers to the joint distribution of $U_1, \ldots, U_T$ and $Y_1, \ldots, Y_T$ and $q(\boldsymbol{u}|\boldsymbol{y})$ to the posterior distribution of $U_1, \ldots, U_T$ given $Y_1, \ldots, Y_T$. Consequently, given an arbitrary sequence $\boldsymbol{u}$, say that with all states equal to 1, we compute the model log-likelihood as

$$\ell(\boldsymbol{\theta}) = \sum_t \log \frac{f(y_t|u_t)p(u_t|u_{\max(t-h,1)},\ldots,u_{t-1})}{q(u_t|u_{\max(t-h,1)},\ldots,u_{t-1},\boldsymbol{y})}$$

on the basis of the proposed recursion. Note that, in this way, we do not need to use any renormalization, which are instead necessary in the Baum and Welch recursions; see also Lystig and Hughes (2002).

In order to maximize $\ell(\boldsymbol{\theta})$, we can use an Expectation-Maximization algorithm that follows the same principle as that illustrated by Baum et al. (1970). In particular, this algorithm is based on the *complete data log-likelihood*

$$\begin{aligned}
\ell^*(\boldsymbol{\theta}) &= \sum_t \sum_{u_t} w_{t,u_t} \log f(y_t|u_t) + \\
&+ \sum_t \sum_{u_{\max(t-1,h)}} \cdots \sum_{u_t} z_{t,u_{\max(t-h,1)},\ldots,u_t} \log p(u_t|u_{\max(t-h,1)},\ldots,u_{t-1}), \qquad (9)
\end{aligned}$$

where w_{t,u_t} is a dummy variable equal to 1 if the latent state at occasion t is u_t and to 0 otherwise and $z_{t,u_{\max(t-h,1)},\ldots,u_t}$ is a corresponding dummy variable for the sequence of latent states $u_{\max(t-h,1)},\ldots,u_t$, which may be expressed through the product

$$z_{t,u_{\max(t-h,1)},\ldots,u_t} = w_{\max(t-h,1),u_{\max(t-h,1)}} \cdots w_{t,u_t}.$$

At the E-step of the EM algorithm, we need to compute the posterior expected values of the above dummy variables given the observed data and the current value of the parameters. In particular, we have that

$$\begin{aligned}
\hat{w}_{t,u_t} &= E(w_{t,u_t}|\boldsymbol{y}) = q(u_t|\boldsymbol{y}),\\
\hat{z}_{t,u_{\max(t-h,1)},\ldots,u_t} &= E(z_{t,u_{\max(t-h,1)},\ldots,u_t}|\boldsymbol{y}) = q(u_{\max(t-h,1)},\ldots,u_t|\boldsymbol{y}).
\end{aligned}$$

In particular, from the proposed recursion, we directly obtain $q(u_1|\boldsymbol{y})$. Then, for $t>1$, we exploit a trivial forward recursion:

$$\begin{aligned}
&q(u_{\max(t-h,1)},\ldots,u_t|\boldsymbol{y}) = \\
&= \begin{cases} q(u_t|u_{\max(t-h,1)},\ldots,u_{t-1},\boldsymbol{y})q(u_{\max(t-h,1)},\ldots,u_{t-1}|\boldsymbol{y}), & t=2,\ldots,h+1,\\ q(u_t|u_{t-h},\ldots,u_{t-1},\boldsymbol{y})\sum_{u_{t-h-1}} q(u_{t-h-1},\ldots,u_{t-1}|\boldsymbol{y}), & t=h+2,\ldots,T, \end{cases} \qquad (10)
\end{aligned}$$

to be performed for $t = 1, \dots, T$. Then, $q(u_t|\boldsymbol{y})$ is computed by a suitable marginalization. How to formulate the above forward recursion in matrix notation, so as to efficiently implement it, is illustrated in Appendix.

As usual, the M-step of the EM algorithm consists of maximizing $\ell^*(\boldsymbol{\theta})$, once the dummy variables in (9) are substituted by the corresponding expected values obtained as above. The following example clarify how to implement this step for a specific model.

Example 4 *For the HM-SV model illustrated in Example 1, the parameters σ_v are updated at the M-step as follows:*

$$\sigma_v = \sqrt{\frac{\sum_t \hat{w}_{t,v} y_t^2}{\sum_t \hat{w}_{t,v}}}, \quad v = 1, \dots, k.$$

Moreover, for the initial probabilities we have

$$\lambda_{1,u_1} = \hat{w}_{1,u_1}, \quad u_1 = 1, \dots, k,$$

and for the transition probabilities, we have

$$\lambda_{t,u_{\max(t-h,1)},\dots,u_t} = \frac{\hat{z}_{t,u_{\max(t-h,1)},\dots,u_t}}{\sum_v \hat{z}_{t,u_{\max(t-h,1)},\dots,u_{t-1},v}}, \quad u_{\max(t-h,1)}, \dots, u_t = 1, \dots, k,$$

for $t = 1, \dots, h$ and

$$\pi_{v_1,\dots,v_{h+1}} = \frac{\sum_{t>h} \hat{z}_{t,v_1,\dots,v_{h+1}}}{\sum_v \sum_{t>h} \hat{z}_{t,v,\dots,v_h,v}}, \quad v_1, \dots, v_{h+1} = 1, \dots, k.$$

Clearly, the posterior probability obtained by the proposed recursion may also be used for local decoding (Juang and Rabiner, 1991), that is to find the most likely value $\hat{u}_t$ of the latent state U_t, given the observed data. In particular, $\hat{u}_t$ is found as the value that maximize the posterior probability $q(u_t|\boldsymbol{y})$.

Finally, on the basis of a sequence of h latent states of the type $u_{T-h+1}, \dots, u_T$, which may be even fixed by the local decoding method, it is possible to predict the latent state at occasion $T+1$, denoted by $\hat{u}_{T+1}$, as the value which maximizes $q(u_{T+1}|u_{T-h+1}, \dots, u_T, \boldsymbol{y})$;

this posterior probability directly derives from the proposed recursion. We can also predict the manifest distribution of Y_{T+1} through the following finite mixture

$$\sum_{u_{T+1}} f(y_{T+1}|u_{T+1})q(u_{T+1}|u_{T-h+1},\ldots,u_T,\boldsymbol{y}).$$

5 An application

In order to illustrate the proposed approach, we fitted the HM version of the stochastic volatility model described in Example 1 to the SP500 data for the period from the beginning of 2008 to the end 2011. The observed outcome is the percentage log-return with respect to the previous closing day, so that we have $T = 1007$ observations.

For the above data, we estimated the model at issue for different values of h (order of the latent Markov chain) and different values of the number of k (number of latent sates), by the EM algorithm outlined in the previous section. The aim of this preliminary analysis is to check if the assumption that the volatility level follows a first-order process is plausible. This means that the level of volatility in a given day only depends on that of the previous day. This hypothesis may be compared with that of a higher-order dependence, in which the level of volatility in a given day also depends on the volatility of, say, the previous two days.

The results of the preliminary fitting are reported in Table 1 in terms of log-likelihood, number of parameters, computed as in (1), and Bayesian Information Criterion (BIC; Schwarz, 1978). Note that we also include results for the model with $h = 0$, which assumes independence between the volatility levels corresponding to different time occasions.

According to BIC, the observed data supports the hypothesis of a first-order dependence of the stochastic volatility. In fact, the smallest value of the BIC index, among those in Table 1, is observed for $h = 1$ and $k = 3$. For this model, we report in Table 2 the estimates of the parameters of main interest.

	h	k = 1	2	3	4
	0	-2026.60	-1898.73	-1887.46	-1885.57
log-lik.	1	-2026.60	-1819.45	-1778.00	-1764.06
	2	-2026.60	-1807.69	-1768.97	-1746.45
	0	1	3	5	7
#par	1	1	5	11	19
	2	1	9	29	67
	0	4060.12	3818.19	3809.50	3819.54
BIC	1	4060.12	3673.48	3632.05	3659.49
	2	4060.12	3677.61	3738.46	3956.18

Table 1: *Results from the preliminary fitting, in terms of maximum log-likelihood, number of parameters, and BIC, of the HM-SV model for different values of h (latent Markov chain order) and k (number of latent states).*

			$\hat{\pi}_{v_1,v_2}$		
v	$\hat{\sigma}_v$	v_1	$v_2 = 1$	$v_2 = 2$	$v_2 = 3$
1	0.865	1	0.988	0.010	0.002
2	1.609	2	0.013	0.981	0.006
3	3.770	3	0.000	0.025	0.975

Table 2: *Estimates of the parameters σ_v (levels of volatility) and π_{v_1,v_2} (transition probabilities) under the HM-SV model with $h = 1$ and $k = 3$.*

We then observe three distinct levels of stochastic volatility and very high persistence in the volatility level, since the probabilities in the transition matrix in Table 2 are very close to 1. As a comparison, we report in Table 3 the corresponding parameter estimates under the model with $h = 2$ and $k = 3$.

We observe that the estimated levels of volatility under the second-order model are similar to those under the first-order model. Moreover, we again note a high persistence, in the sense that $\hat{\pi}_{v_1,v_2,v_3}$ is very close to 1 whenever $v_1 = v_2 = v_3$. The estimates of these transition probabilities for $v_1 \neq v_2$ seem to be less reasonable, especially when $v_1 = 3$ and

v	$\hat{\sigma}_v$	v_1	v_2	$\hat{\pi}_{v_1,v_2,v_3}$ $v_3 = 1$	$v_3 = 2$	$v_3 = 3$
1	0.842	1	1	0.979	0.021	0.000
2	1.725	1	2	0.909	0.091	0.000
3	4.047	1	3	0.585	0.000	0.415
		2	1	0.113	0.873	0.014
		2	2	0.027	0.966	0.007
		2	3	1.000	0.000	0.000
		3	1	0.000	0.000	1.000
		3	2	0.000	1.000	0.000
		3	3	0.000	0.035	0.965

Table 3: *Estimates of the parameters σ_v (levels of volatility) and π_{v_1,v_2,v_3} (transition probabilities) under the HM-SV model with $h = 2$ and $k = 3$.*

$v_2 = 1$. However, we have to consider that a jump from state 3 to state 1 is very rare and then there is no support from the data to estimate a transition probability given this pair of states. This confirms that the first-order model is preferable for the data at hand and may provide more reliable estimates. In any case, the possibility to estimate a higher order HM model, which is allowed by the proposed recursion, is important in order to have a counterpart against which comparing the more common first-order model.

Appendix: the recursion in matrix notation

First of all let $\boldsymbol{p}_t$ be the column vector of prior probabilities $p(u_t|u_{\max(t-h,1)},\ldots,u_{t-1})$ arranged in lexicographical order so that, for instance, with $h = 2$ and $k = 2$ we have

$$\boldsymbol{p}_t = \begin{pmatrix} p(u_t = 1|u_{t-2} = 1, u_{t-1} = 1) \\ p(u_t = 2|u_{t-2} = 1, u_{t-1} = 1) \\ p(u_t = 1|u_{t-2} = 1, u_{t-1} = 2) \\ p(u_t = 2|u_{t-2} = 1, u_{t-1} = 2) \\ p(u_t = 1|u_{t-2} = 2, u_{t-1} = 1) \\ p(u_t = 2|u_{t-2} = 2, u_{t-1} = 1) \\ p(u_t = 1|u_{t-2} = 2, u_{t-1} = 2) \\ p(u_t = 2|u_{t-2} = 2, u_{t-1} = 2) \end{pmatrix}, \quad t = 3,\ldots,T.$$

Note that for $t=1$ this is a vector of initial probabilities and that the number of elements of $\boldsymbol{p}_t$ is $k^{d_{t,0}}$, where, in general, $d_{t,j} = \min(t-1,h)+j+1$. Also let $\boldsymbol{f}_t$ denote the column vector with elements $f(y_t|u_t)$, $u_t = 1,\ldots,k$, and let $\boldsymbol{q}_{t,j}$ denote the column vector of the posterior probabilities $q(u_t|u_{\max(t-h,1)},\ldots,u_{t-1},u_{t+1},\ldots,u_{t+j},\boldsymbol{y})$ again arranged in lexicographical order. With $h=1$, $k=2$, and $j=1$, for instance, we have

$$\boldsymbol{q}_{t,j} = \begin{pmatrix} q(u_t=1|u_{t-1}=1,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=1,u_{t+1}=2,\boldsymbol{y}) \\ q(u_t=2|u_{t-1}=1,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=2|u_{t-1}=1,u_{t+1}=2,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=2,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=2,u_{t+1}=2,\boldsymbol{y}) \\ q(u_t=2|u_{t-1}=2,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=2|u_{t-1}=2,u_{t+1}=2,\boldsymbol{y}) \end{pmatrix}, \quad t=2,\ldots,T. \tag{11}$$

The dimension of this vector is $k^{d_{t,j}}$; note that for $j=0$ these vectors contain the target posterior probabilities in (4).

Finally, let $\boldsymbol{M}_{a,b}$ be a marginalization matrix such that, given a column vector $\boldsymbol{v}$ with elements indexed by a sequence of b variables assuming k possible values (as the above vectors), $\boldsymbol{M}_{a,b}\boldsymbol{v}$ provides the corresponding vector in which the elements are summed with respect to the a-th of these variables. This matrix may be simply constructed by following Kronecker product $\boldsymbol{M}_{a,b} = \bigotimes_{l=1}^{b} \boldsymbol{M}_l^*$, where

$$\boldsymbol{M}_l^* = \begin{cases} \boldsymbol{1}_k', & l=a, \\ \boldsymbol{I}_k, & l \neq a, \end{cases}$$

with $\boldsymbol{1}_k$ denoting a column vector of k ones and $\boldsymbol{I}_k$ denoting an identity matrix of the same dimension. For instance, in the same context of the example that led to the vector $\boldsymbol{q}_t$ in (11), we have

$$\boldsymbol{M}_{2,3}\boldsymbol{q}_{t,j} = \begin{pmatrix} q(u_t=1|u_{t-1}=1,u_{t+1}=1,\boldsymbol{y}) + q(u_t=2|u_{t-1}=1,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=1,u_{t+1}=2,\boldsymbol{y}) + q(u_t=2|u_{t-1}=1,u_{t+1}=2,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=2,u_{t+1}=1,\boldsymbol{y}) + q(u_t=2|u_{t-1}=2,u_{t+1}=1,\boldsymbol{y}) \\ q(u_t=1|u_{t-1}=2,u_{t+1}=2,\boldsymbol{y}) + q(u_t=2|u_{t-1}=2,u_{t+1}=2,\boldsymbol{y}) \end{pmatrix}.$$

Using the above notation, for $t=T$ we directly obtain the target vector $\boldsymbol{q}_{t,0}$ through the following operations, which directly derive from (5):

$$\begin{aligned}\boldsymbol{a}_{T,0} &= (\boldsymbol{1}_{k^{(d_{T,0}-1)}} \otimes \boldsymbol{f}_T) \times \boldsymbol{p}_T,\\ \boldsymbol{q}_{T,0} &= \boldsymbol{a}_T/(\boldsymbol{M}'_{d_{T,0},d_{T,0}}\boldsymbol{M}_{d_{T,0},d_{T,0}}\boldsymbol{a}_{T,0}),\end{aligned}$$

where $\times$ and $/$ denote, respectively, elementwise product and division. Then, for $t = 1,\dots,T-1$ (in reverse order), we first compute $\boldsymbol{q}_{t,j}$ for $j = \min(T-t,h)$ and then we recursively compute $\boldsymbol{q}_{t,j}$ from $j = \min(T-t,h)-1$ to $j = 0$. In particular, for $j = \min(T-t,h)$ we compute the vector $\boldsymbol{a}_{t,j}$ containing the elements at the numerator of (7) by the following recursion:

$$\begin{aligned}\boldsymbol{a}_{t,0} &= (\boldsymbol{1}_{k^{(d_{t,0}-1)}} \otimes \boldsymbol{f}_t) \times \boldsymbol{p}_t,\\ \boldsymbol{a}_{t,l} &= (\boldsymbol{a}_{t,l-1} \otimes \boldsymbol{1}_k) \times (\boldsymbol{1}_{k^{(d_{t,l}-d_{t+l,0})}} \otimes \boldsymbol{p}_{t+l}), \quad l = 1,\dots,j.\end{aligned}$$

Then, we have

$$\boldsymbol{q}_{t,j} = \boldsymbol{a}_{t,j}/(\boldsymbol{M}'_{d_{t,0},d_{t,j}}\boldsymbol{M}_{d_{t,0},d_{t,j}}\boldsymbol{a}_{t,j}).$$

Finally, from (6) we have the recursion

$$\begin{aligned}\boldsymbol{s}_{t,j+1} &= (\boldsymbol{1}_{k^{(d_{t,j+1}-d_{t+j+1,0})}} \otimes \boldsymbol{q}_{t+j+1,0})/\boldsymbol{q}_{t,j+1},\\ \boldsymbol{q}_{t,j} &= \boldsymbol{1}_{k^{d_{t,j}}}/(\boldsymbol{M}_{d_{t,j+1},d_{t,j+1}}\boldsymbol{s}_{t,j+1}),\end{aligned}$$

to be applied for $j = \min(T-t,h)-1$ until $j = 0$, when we obtain the target vector $\boldsymbol{q}_{t,0}$.

In order to express the forward recursion in (10) using the matrix notation, let $\boldsymbol{q}_t^*$ denote the vector with elements $q(u_{\max(t-h,1)},\dots,u_t|\boldsymbol{y})$ arranged in the usual lexicographical order. Then, we have $\boldsymbol{q}_1^* = \boldsymbol{q}_{1,0}$, whereas for $t > 1$, we have

$$\boldsymbol{q}_t^* = \begin{cases} \boldsymbol{q}_{t,0} \times (\boldsymbol{q}_{t-1}^* \otimes \boldsymbol{1}_k), & t = 2,\dots,h+1,\\ \boldsymbol{q}_{t,0} \times [(\boldsymbol{M}_{1,h+1}\boldsymbol{q}_{t-1}^*) \otimes \boldsymbol{1}_k], & t = h+2,\dots,T.\end{cases}$$

References

Bartolucci, F. and Besag, J. (2002). A recursive algorithm for Markov random fields. *Biometrika*, 89:724–730.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *Technical report available at http://arxiv.org/abs/1003.2804*.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Juang, B. and Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33:251–272.

Lystig, T. C. and Hughes, J. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, 11:678–689.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

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.

Taylor, S. J. (2005). *Asset Price Dynamics, Volatility, and Prediction.* Princeton University Press, Princeton.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov and other Models for Time Series: An Introduction Using R.* Chapman and Hall, CRC.