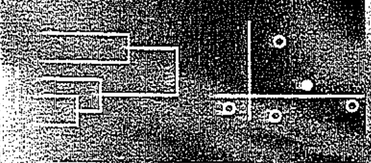
STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION

Francesco Palumbo
Carlo Natale Lauro
Michael J. Greenacre
Editors

Data Analysis and Classification



Springer

Palumbo
Lauro
Greenacre
(Eds.)

Data Analysis and Classification

Data Analysis and Classification

The volume provides results from the latest methodological developments in data analysis and classification and highlights new emerging subjects within the field. It contains articles about statistical models, classification, cluster analysis, multidimensional scaling, multivariate analysis, latent variables, knowledge extraction from temporal data, financial and economic applications, and missing values. Papers cover both theoretical and empirical aspects.

ISSN 1431-8814

ISBN 978-3-642-03738-2

9 783642 037382

springer.com

Bayesian Hidden Markov Models for Financial Data

Rosella Castellano and Luisa Scaccia

Abstract Hidden Markov Models, also known as Markov Switching Models, can be considered an extension of mixture models, allowing for dependent observations. The main problem associated with Hidden Markov Models is represented by the choice of the number of regimes, i.e. the number of generating data processes, which differ one from another just for the value of the parameters. Applying a hierarchical Bayesian framework, we show that Reversible Jump Markov Chain Monte Carlo techniques can be used to estimate the parameters of the model, as well as the number of regimes, and to simulate the posterior predictive densities of future observations. Assuming a mixture of normal distributions, all the parameters of the model are estimated using a well known exchange rate data set.

1 Introduction

A Hidden Markov Model (HMM) or Markov Switching Model is a mixture model whose mixing distribution is a finite state Markov Chain. In practice, given a data set indexed by time, the distribution of each observation is assumed to depend on an unobserved variable, hidden "state" or "regime", whose transition is regulated by a Markov Chain. HMMs have been successfully applied to financial time series: very often financial data show nonlinear dynamics which are possibly due to the existence of two or more regimes, differing one from another only for the value of the parameters. For instance, segmented time-trends in the US dollar exchange rates, Engel and Hamilton (1990), stylized facts about daily returns, Rydén (1998), option prices and stochastic volatilities, Rossi and Gallo (2006), temporal behavior of volatility of daily returns on commodities, Haldrup and Nielsen (2006), have also been modeled via HMMs.

The main problem associated with HMMs is to select the number of regimes (i.e. the number of generating data processes). In a classical perspective, this requires

L. Scaccia (✉)
DIEF, Università di Macerata, Via Crescimbeni, 20, 62100 Macerata, Italy
e-mail: scaccia@unimc.it

F. Palumbo et al. (eds.), *Data Analysis and Classification*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-03739-9_51, © Springer-Verlag Berlin Heidelberg 2010

hypothesis testing with nuisance parameters, identified only under the alternative. Thus, the regularity conditions for the asymptotic theory to hold are not met and the limiting distribution of the likelihood ratio test must be approximated by simulation, an approach demanding enormous computational efforts. Penalized likelihood methods, as the Akaike and Bayesian information criteria, though are less demanding, do not produce a number quantifying the confidence in the results (i.e. p-values).

In a Bayesian context, several approaches to choose the number of regimes can be listed. A Bayesian non-parametric approach, based on a Dirichlet process (DP) with, a priori, infinite number of regimes is described in Otranto and Gallo (2002). Simulations from the posterior distribution of the process are used to estimate the posterior probabilities of the number of regimes. An alternative approach is based on allocation models: a latent variable is explicitly introduced to allocate each observation to a particular regime, Robert et al. (2000). Then, the Reversible Jump (RJ) algorithm, Green (1995), is used to sample from the posterior joint distribution of all the parameters, including the number of regimes.

In this paper, we prefer to deal with the latter approach for several reasons. From a theoretical point of view, the predictive density of a future observation, based on a DP, assigns to this observation a non-null probability of being exactly equal to one of those already observed. Such a behavior is highly unrealistic if data points are assumed to be drawn from a continuous distribution. Moreover, non-parametric approaches are strongly affected by the influence of the prior distribution on the posterior one, so that the likelihood never dominates the prior and the inferential results are particularly sensitive to prior assumptions. Furthermore, in a DP, a single parameter controls the variability and the clustering, making the prior specification difficult. Finally, the DP is well known to favor, a priori, unequal allocations and this phenomenon becomes more dramatic as soon as the number of observations increases. The unbalance in the prior allocation distribution often persists also a posteriori, Green and Richardson (2001). However, the model proposed in Robert et al. (2000) only allows for regimes being different because of their volatilities. We extend this approach to permit the existence of regimes characterized by different means and/or variances.

The paper is organized as follows: the model and prior assumptions are illustrated in Sect. 2; Sect. 3 deals with computational implementation; Sect. 4 discusses Bayesian inference and forecasting; finally, in Sect. 5 an application is considered.

2 The Model

Let $y = (y_t)_{t=1}^T$ be the observed data, indexed by time. In HMMs, the heterogeneity in the data is represented by a mixture structure, that is, a pair (s_t, y_t), with s_t being an unobserved state variable characterizing the regime of the process at any time t and y_t being independent conditional on the s_t's:

$$y_t|s_t \sim f_{s_t}(y_t) \qquad \text{for} \quad t = 1, 2, \ldots, T\ . \tag{1}$$

with $f_{s_t}(\cdot)$ being a specified density function. Assuming $\mathcal{S} = \{1, \ldots, k\}$ to be the set of possible regimes, HMMs further postulate that the dynamics of $s = (s_t)_{t=1}^T$ are described by a Markov Chain with transition matrix $\Lambda = (\lambda_{ij})_{i,j=1}^k$. Accordingly, s_t is presumed to depend on the past realizations of y and s, only through s_{t-1}:

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

We study mixtures of normal distributions, so that the model in (1) becomes

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

conditional on means $\mu = (\mu_i)_{i=1}^k$ and standard deviations $\sigma = (\sigma_i)_{i=1}^k$, where $\phi(\cdot; \mu_i, \sigma_i^2)$ is the density of the $N(\mu_i, \sigma_i^2)$. Thus, if $s_t = i$, y_t is assumed to be drawn from a $N(\mu_i, \sigma_i^2)$. Notice that, if we let π being the stationary vector of the transition matrix, so that $\pi'\Lambda = \pi'$, and we integrate out s_t in (2) using its stationary distribution, the model in (2) can be analogously formalized as

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

In a classical perspective, the model in (2) can be estimated, conditional on k, by means of EM algorithm, Scott (2002). Then, as already mentioned, the main problem is to choose among different models, characterized by a different number of regimes.

In a Bayesian context, we formalize the uncertainty on the parameters of the model, as well as on the number of regimes, k, using appropriate prior distributions. We choose weakly informative priors, introducing an hyperprior structure, so that $\mu_i|\sigma_i^2 \sim N(\xi, \kappa\sigma_i^2)$ and $\sigma_i^{-2} \sim \text{Ga}(\eta, \zeta)$, independently for each $i = 1, \ldots, k$, with the mean and the variance of the Gamma distribution being η/ζ and η/ζ^2. Then we assume κ to follow an Inverse Gamma distribution with parameters q and r, and ζ to follow a Gamma distribution with parameters f and h. Finally, the rows of the transition matrix have a Dirichlet distribution, so that $\lambda_{ij} \sim D(\delta_j)$, for $i = 1, \ldots, k$ where $\delta_j = (\delta_{ij})_{i=1}^k$, while the number of regimes k is a priori uniform on the values $\{1, 2, \ldots, K\}$, with K being a pre-specified integer corresponding to the maximum hypothesized number of regimes. These settings lead up to the hierarchical model in Fig. 1. The choice of the hyperparameters will be briefly discussed in Sect. 5.

3 Computational Implementation

In order to approximate the posterior joint distribution of all the parameters of the above mixture model, Markov Chain Monte Carlo (MCMC) methods are applied (details can be found in Tierney (1994)). To generate realizations from the posterior

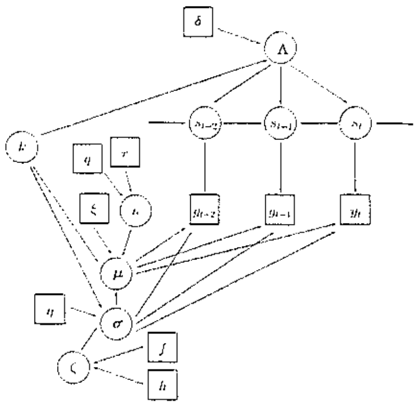


Fig. 1 Directed acyclic graph for the complete hierarchical model

joint distribution, at each sweep of the MCMC algorithm, we update in turn: **(a)** the transition matrix Λ, **(b)** the state variable s, **(c)** the means μ, **(d)** the standard deviations σ, **(e)** the hyperparameter κ, **(f)** the hyperparameter ζ, **(g)** the number of regimes k. The first six moves are fairly standard and all performed through Gibbs sampling.

In particular, in **(a)**, the i-th row of Λ is sampled from $D(\delta_{i1}+n_{i1},\ldots,\delta_{ik}+n_{ik})$, 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 and $I\{\cdot\}$ denotes the indicator function, Robert et al. (1993).

In **(b)**, the standard solution for updating s would be to sample $s_1,\ldots,s_T$ one at a time from $t=1$ to $t=T$, drawing values from their full conditional distribution $p(s_t=i|\cdots)\propto\lambda_{s_{t-1}i}\phi(y_t;\mu_i,\sigma_i^2)\lambda_{is_{t+1}}$ where "$\cdots$" denotes "all other variables". For a faster mixing algorithm, as in Scott (2002), and Castellano and Scaccia (2007), we instead sample s from $p(s|y,\Lambda)$ through a stochastic version of the forward–backward recursion. The forward recursion produces matrices $\boldsymbol{P}_2,\ldots,\boldsymbol{P}_T$, where $\boldsymbol{P}_t=(p_{tij})$ and $p_{tij}=p(s_{t-1}=i,s_t=j|y_1,\ldots,y_t,\Lambda)$. In words, $\boldsymbol{P}_t$ is the joint distribution of $(s_{t-1}=i,s_t=j)$ given parameters and observed data up to time t. $\boldsymbol{P}_t$ is computed from $\boldsymbol{P}_{t-1}$ as $p_{tij}\propto p(s_{t-1}=i,s_t=j,y_t|y_1,\ldots,y_{t-1},\Lambda)=p(s_{t-1}=i|y_1,\ldots,y_{t-1},\Lambda)\lambda_{ij}\phi(y_t;\mu_j,\sigma_j^2)$ with proportionality reconciled by $\sum_i\sum_j p_{tij}=1$, where $p(s_{t-1}=i|y_1,\ldots,y_{t-1},\Lambda)=\sum_j p_{t-1,i,j}$ can be computed once $\boldsymbol{P}_{t-1}$ is known. The recursion starts computing $p(s_1=i|y_1,\Lambda)\propto\phi(y_1;\mu_i,\sigma_i^2)\pi_i$ and thus $\boldsymbol{P}_2$. The stochastic backward recursion begins by drawing s_T from $p(s_T|y,\Lambda)$, 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 $p(s|y,\Lambda)$, factorizing this distribution as $p(s|y,\Lambda)=p(s_T|y,\Lambda)\prod_{t=1}^{T-1}p(s_{T-t}|s_T,\ldots,s_{T-t+1},y,\Lambda)$ where $p(s_{T-t}=i|s_T,\ldots,s_{T-t+1},y,\Lambda)=p(s_{T-t}=i|s_{T-t+1},y_1,\ldots,y_{T-t+1},\Lambda)\propto p_{T-t+1,i,s_{T-t+1}}$.

In **(c)**, for identifiability purpose, we adopt a unique labeling in which the μ_i's are in increasing numerical order, Richardson and Green (1997). Hence, their joint prior distribution 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 drawn independently from $\mu_i|\cdots\sim N\left(\frac{\kappa\sum_{t:s_t=i}y_t+\xi}{1+\kappa n_i},\frac{\sigma_i^2\kappa}{1+\kappa n_i}\right)$, n_i being the number of observations currently allocated in regime i. The move is accepted, provided the invariance of the order.

In **(d)** we update each component of the vector σ^2 independently, drawing σ_i^{-2} from $\sigma_i^{-2}|\cdots\sim\text{Ga}\left(\eta+\frac{1}{2}(n_i+1),\zeta+\frac{1}{2}\sum_{t:s_t=i}(y_t-\mu_i)^2+\frac{1}{2\kappa}(\mu_i-\xi)^2\right)$.

In **(e)** we sample κ^{-1} from $\kappa^{-1}|\cdots\sim\text{Ga}\left(q+\frac{k}{2},r+\frac{1}{2}\sum_{i=1}^{k}\frac{(\mu_i-\xi)^2}{\sigma_i^2}\right)$ and, finally, in **(f)** we sample ζ from $\zeta|\cdots\sim\text{Ga}\left(f+k\eta,h+\sum_{i=1}^{k}\sigma_i^{-2}\right)$.

Updating k implies a change of dimensionality for μ, σ and Λ. We follow the approach used in Richardson and Green (1997) which consists in a random choice between splitting an existing regime into two, and merging two existing regimes into one. For the *combine proposal* we randomly choose a pair of regimes (i_1,i_2) that are adjacent in terms of the current value of their means. These two regimes are merged into a new one, i^*, reducing k by 1. We then reallocate all the y_t, with $s_t=i_1$ or $s_t=i_2$, to the new regime i^* and create values for μ_{i^*}, $\sigma_{i^*}^2$, π_{i^*} and for the transition probabilities from and to the regimes involved in the move. This is performed in such a way to guarantee that the new HMM and the old one both have the same first and second moments. The *split proposal* starts with the random choice of regime i^* which is splitted into two new ones, i_1 and i_2, augmenting k by 1. Accordingly, we reallocate the y_t with $s_t=i^*$ between i_1 and i_2, and create values for $\pi_{i_1},\pi_{i_2},\mu_{i_1},\mu_{i_2},\sigma_{i_1},\sigma_{i_2}$ and the transition probabilities for the regimes involved. The aim is to split i^* in such a way that the dynamics of the Hidden Markov Chain are essentially preserved, Robert et al. (2000). The move is accepted with a probability computed to preserve the *reversibility* between the states of the MCMC algorithm. More details on computational issues can be found in Castellano and Scaccia (2007).

4 Bayesian Inference and Forecasting

After a burn-in period, to guarantee the convergence of the chain to its stationary distribution, the RJ algorithm produces at each sweep n, for $n=1,\ldots,N$, a draw $(k^{(n)},\Lambda^{(n)},s^{(n)},\mu^{(n)},\sigma^{(n)},\kappa^{(n)},\zeta^{(n)})$ from the joint posterior distribution of all the parameters, including k. The sample obtained after N sweeps can be used to estimate all the quantities of interest. For instance, we can easily estimate the posterior distribution of the number of regimes as the proportion of times each model is visited by the algorithm, i.e. $\hat{p}(k=\ell|y)=\sum_{n=1}^{N}I\{k^{(n)}=\ell\}/N=N_\ell/N$, where N_ℓ is the number of times the model with ℓ regimes is visited.

Conditioning on a particular model, say M_ℓ, the one with ℓ regimes, any other parameter of that model can be estimated, Richardson and Green (1997). Estimating the hidden states s often represents a key question in applied problems. Inference on s derives from its posterior $p(s|y)$, a high-dimensional distribution that must be summarized to be understood. In general, it is sufficient to summarize it through its marginal distributions $p(s_t = i|y)$, whose obvious estimates are $\hat{p}(s_t = i|y) = \sum_{n:k^{(n)}=\ell} I\{s_t^{(n)} = i\}/N_\ell$. More efficient estimates demand small additional computational effort, as shown in Castellano and Scaccia (2007).

Finally, when historical series are analysed, the main goal is, generally, to forecast future values of the observed variable, on the basis of the information available up to time T. In a Bayesian context, inferences on future observations, i.e. $Y = (y_{T+1}, \ldots, y_{T+G})$, are based on their posterior predictive density, which 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, say M_ℓ, the information available up to time T will encompass the generating model and the observed data up to time T. Then, the posterior predictive density for Y will be:

$$p(Y|y, M_\ell) = \int_{\Theta_{M_\ell}} p(Y|y, \theta_{M_\ell}, M_\ell) p(\theta_{M_\ell}|y, M_\ell) d\theta_{M_\ell}, \tag{3}$$

where θ_{M_ℓ} 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, the uncertainty about which one is the true generating model, within a set of possible ones, can be expressed through model averaging, defining the posterior predictive density of Y as:

$$p(Y|y) = \sum_{k=1}^{K} p(Y|y, M_k) p(M_k|y) , \tag{4}$$

where $p(Y|y, M_k)$ is defined in (3). Notice that in (4), we consider as available information at time T only the observed data. In both cases, the posterior predictive density can be stimulated as by-product of the MCMC algorithm, Scott (2002).

5 An Application to Financial Data

The proposed model is applied to the exchange rate quarterly returns of the U.S. dollar relative to the French franc, over the period 1973-III to 1988-I, already analysed in Engel and Hamilton (1990) and Otranto and Gallo (2002) through a likelihood ratio and a non-parametric Bayesian approach, respectively. While the approach in Engel and Hamilton (1990) failed to test the model with two regimes against the one with only one regime (more details are reported in Otranto and Gallo (2002)), in Otranto and Gallo (2002) the posterior probability for $k = 2$ was found to be slightly higher than that for $k = 1$. The choice of the data set is motivated by the existence of a benchmark for comparing the results.

For previously unspecified hyperparameters, we let: $\delta_{ij} = 1$, $\forall i, j$; $\xi = 0$; $\eta = 2$; $q = 2$; $r = 2$; $f = 0.95$; $h = 4/R^2$, where R is the data range. All the hyperparameters were chosen in a way that the priors on the parameters would assign large probabilities to a large range of values. Some experimentations highlight that the results are quite robust with respect to reasonable perturbations on the choice of the hyperparameters. Models with a number of regimes up to $K = 5$ were considered. Larger values for K would be unreasonable, given the short time series at hand. Trace plots of the sample parameter values versus iteration, as well as of the sample posterior probability of the number of regimes Fig. 2(b)), were used to control for the stabilization of the simulation. A burn-in of 100,000 sweeps seems sufficient for the convergence to occur. We performed 1,000,000 sweeps of the MCMC algorithm. We observed a quite high acceptance rate (≈17%) for the RJ move, due to the fact that the data set is small and, thus, the posterior distribution of the parameters is not particularly picked, making the algorithm move easily over different models (Fig. 2(a)). After the convergence, the algorithm provides with stable estimates of the posterior model probabilities, given in Fig. 2(b). As in Otranto and Gallo (2002), we get evidence in favor of the two regimes model, the mode of the posterior probability being $k = 2$.

Conditionally on $k = 2$, Fig. 3(b) shows the posterior probability of the U.S. dollar being in a regime of appreciation, with vertical lines representing the switches from one regime to the other. The posterior probability, as well as the switching points, resemble closely the results obtained in Engel and Hamilton (1990). Furthermore, the process seems to stay in the same regime for a while, as confirmed by the estimated transition matrix

$$\hat{\Lambda} = \begin{pmatrix} 0.676 & 0.324 \\ 0.350 & 0.650 \end{pmatrix}$$

showing higher probabilities to persist in the same regime compared to those of switching to the other one (i.e.: the so called *long swings* of the U.S. dollar).

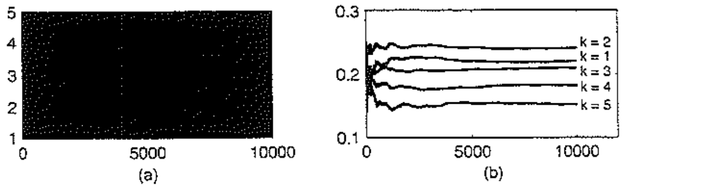


Fig. 2 (a) Last 10,000 values of k. (b) Estimated posterior distribution of k as a function of number of sweeps, plotted every 100*th* sweep

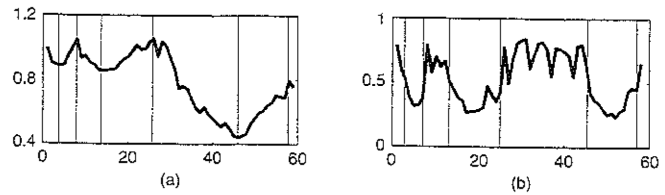


Fig. 3 (a) U.S. dollar/French franc exchange rate. (b) Estimated posterior probability of the U.S. dollar being in a appreciation regime as a function of t

6 Conclusions

In this paper Bayesian inference for HMMs with an unknown number of regimes and its application to financial time series is illustrated. We considered a hierarchical model which allows to make vague *a priori* assumptions on the parameters. The analytically untractable joint posterior distribution of all the parameters and the unknown number of regimes was simulated through MCMC methods and RJ algorithm.

Future developments could encompass the design of RJ moves visiting a larger set of models, in which some regimes may have equal variances but different means or equal means but different variances. The approach can also be adapted to any extension of HMMs, such as time-varying transition probabilities, Markov switching heteroskedasticity, multiple regime Smooth Transition or Threshold AR models.

References

Castellano, R., & Scaccia, L. (2007). Bayesian hidden Markov Models with an unknown number of regimes. Quaderni del Dipartimento di Istituzioni Economiche e Finanziarie, 43. Università di Macerata

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

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

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

Haldrup, N., & Nielsen, M. O. (2006). A regime switching long memory model for electricity prices. *Journal of Econometrics*, *135*, 349–376

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

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

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

Robert, C. P., Rydén, T., & 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 - Series B: Statistical Methodology*, *62*, 57–75

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

Rydén, T., Teräsvirta, T., & Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Economics*, *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*, 1701–1764