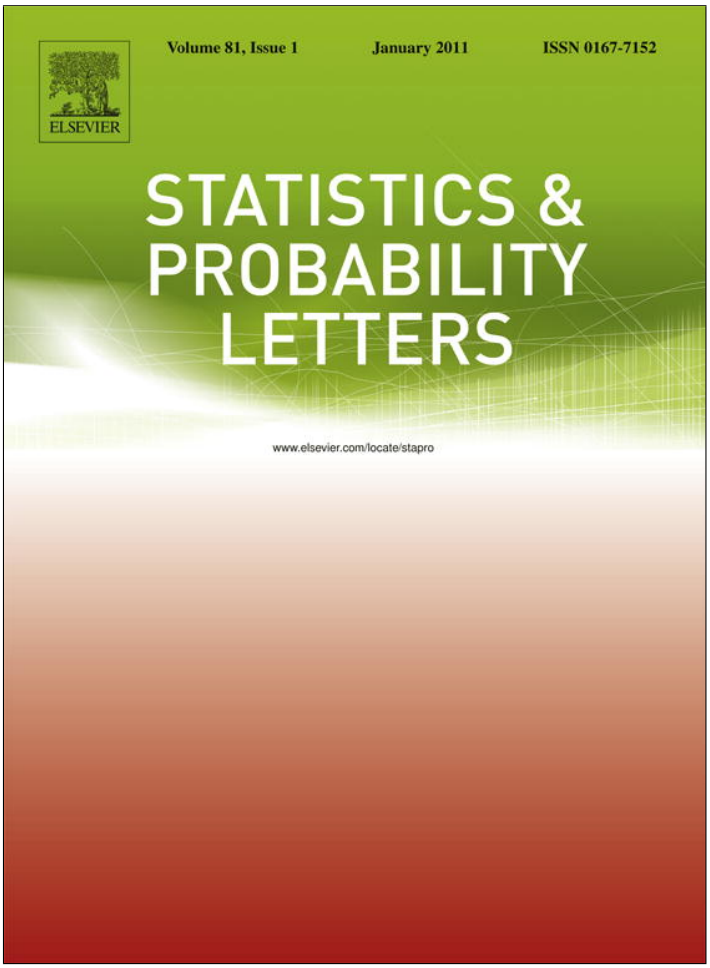

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Statistics and Probability Letters 81 (2011) 1766–1770

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Hidden Markov partition models

Alessio Farcomeni

Sapienza - University of Rome, Italy

ARTICLE INFO

Article history:
Received 9 November 2010
Received in revised form 12 July 2011
Accepted 12 July 2011
Available online 22 July 2011

Keywords:
Hidden Markov model
Partition model
Forward recursion
Backward recursion

ABSTRACT

We describe an extension of the hidden Markov model in which the manifest process conditionally follows a partition model. The assumption of local independence for the manifest random variable is thus relaxed to arbitrary dependence. The proposed class generalizes different existing models for discrete and continuous time series, and allows for the finest trading off between bias and variance. The models are fit through an EM algorithm, with the usual recursions for hidden Markov models extended at no additional computational cost.

1. Introduction

Let X_t, $t = 1, \ldots, T$, be a sequence of random variables and let x_t denote a realization of X_t. In this work we assume that X_t is discrete, even if generalizations to the continuous case are straightforward. Let the state space of X_t be $\{0, \ldots, d-1\}$ without loss of generality.

Many discrete valued time series are generated with high order and/or non-homogeneous Markovian dependencies. The classical tools for modeling such data (i.e., Markov chains) rely usually on a large number of parameters, which grow exponentially with the number of lags. For fixed number of lags, in the class of Markov chains there is just one single model, with a fixed number of parameters. The user is thus not allowed to trade off between bias and variance. There are many proposals of more flexible models, e.g., variable length Markov chains (Bülmann and Wyner, 1999), and the mixture transition distribution model (Raftery, 1985; Berchtold and Raftery, 2002). A particularly general class relies on the assumption that probabilities of transitions from states in a partition of the previous history are all equal (Garcia and Gonzalez-Lopez, 2010; Farcomeni, 2011). Formally, if $(x_{t-1}, \ldots, x_{t-l})$ and $(x'_{t-1}, \ldots, x'_{t-l})$ both belong to L_j in the partition $\{L_1, \ldots, L_p\}$ of $\{0, \ldots, d-1\}^l$, then $p(X_t = x_t | x_{t-1}, \ldots, x_{t-l}) = p(X_t = x_t | x'_{t-1}, \ldots, x'_{t-l})$. All previously mentioned models can be seen as special cases of this general class, which we call the class of partition models (PM).

A limitation of PM is that all the features of the distribution of X_t must be captured by the fixed effects in the model, while in real data situations there could be unobserved heterogeneity which is not explained by previous realizations of the process, even for large l. The assumption of occasion specific parameters (i.e., inhomogeneous PM) would anyway make the model not identifiable. A common approach to the problem of inhomogeneity is given by hidden Markov models (MacDonald and Zucchini, 1997), which assume the existence of a latent process explaining unobserved heterogeneity. While the assumption of first-order dependence for the latent process is seldom restrictive, a limitation of HMM is the assumption of local independence: X_t is assumed to be independent of past observations given the current value of the latent process. This is obviously a limitation in that (i) the latent process may not be able to capture the effects of realizations of X_t at lags far in the past and (ii) the user may prefer to still model the observed heterogeneity through parameters directly related to X_t for interpretability reasons. A generalization of HMM in this spirit is given by the HM-MTD model of Bartolucci and

E-mail address: alessio.farcomeni@uniroma1.it.

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.07.012

Table 1
Comparison of the number of parameters for a discrete valued time series with state space {0, 1, 2, 3} with different models of order l. For the HMM the parameter l does not refer to the lag, but to the number of states of the latent process.

l	0	1	2	3	4	5
Markov chain	3	12	48	192	768	3072
MTD	3	12	13	14	15	16
HM-MTD	3	12	15	20	27	36
HMM_l	–	3	9	17	27	39

Farcomeni (2010). In this work we generalize the HMM in full in this direction, allowing arbitrary dependence in X_t even after conditioning on the latent observations. In the resulting hidden Markov partition models (HMPM) a latent process follows a first-order Markov chain, and conditionally on the current realization of the latent process, the outcome follows a PM. An HM_k-PM_l is specified by fixing a lag l for the memory of the manifest process, a partition $\mathcal{L}$ of $\{0, \ldots, d-1\}^l$, and the number k of latent states for the hidden Markov chain. When $k = 1$, we get back a PM model. When $k > 1$, the model can be fit with an EM-type algorithm (Dempster et al., 1977) with ad hoc recursions which have the same computational complexities as are commonly needed to fit HMM models. HMPM allow for high order dependence with an arbitrary number of parameters between a minimum and a maximum, thus permitting the finest possible trading off between bias and variance.

In Table 1 we report the number of parameters of models for a discrete valued time series with state space {0, 1, 2, 3} (i.e., $d = 4$). PM and HMPM can be specified with *any* number of parameters between the dimension of the independence model (i.e., the model with three parameters) and the dimension of the saturated lth-order Markov chain. HMPM thus allow one to, for instance, summarize long memory with almost as few parameters as wished. For instance, there exists at least one model in the class of HM_2-PM with memory at lag $l = 5$ which is based on 11 parameters.

The rest of the work is as follows: in Section 2 we give a general description of the proposed class of HPM-PM in their most general form; in Section 3 we derive inference for HPM and discuss model choice.

2. Hidden Markov partition models

The class of models that we propose can be formulated by specifying the number of latent states k, a lag l and corresponding partition of the manifest alphabet, $\mathcal{L}^X = \{L_j^X,\ j = 1, \ldots, p\}$. The corresponding HM_k-PM_l model is then as follows:

$$\begin{cases} p(X_t = x_t | (X_{t-1}, \ldots, X_{t-l}) \in L_j^X, U_t = u_t) = \phi_{jx_t}^{(u_t)}, & t > l \\ p(U_{l+1} = u_{l+1}) = \rho_{u_{l+1}} \\ p(U_t = u_t | (U_{t-1})) = \pi_{u_{t-1}u_t}, & t > l+1 \end{cases}$$

with the natural constraints that all parameters are non-negative, $\sum_h \rho_h = 1$; and $\sum_x \phi_{jx}^{(h)} = 1$ for all $h = 1, \ldots, k$ and for all $j = 1, \ldots, p$. Further, $\sum_h \pi_{jh} = 1$ for all $j = 1, \ldots, p$.

The manifest thus follows k PM models indexed at time t by the value of U_t. The number of parameters is $g = pk(d-1) + k - 1 + k(k-1)$. Identifiability is guaranteed as long as $g \leq (d-1)d^l$, that is, the partition is specified so that g is smaller than or equal to the number of parameters of the saturated lth-order Markov chain, which is in one-to-one correspondence with the number of degrees of freedom of the corresponding contingency table for the history counts.

The marginal probabilities are obtained as follows: we have that

$$p(X_{l+1} = x_{l+1} | (X_l, \ldots, X_1) \in L_j^X) = \sum_{h=1}^{k} \rho_h \phi_{jx_{l+1}}^{(h)},$$

while for $t = l+2, \ldots, T$,

$$p(X_t = x_t | (X_{t-1}, \ldots, X_{t-l}) \in L_j^X) = \sum_{h=1}^{k} \lambda_{th}(x_1, \ldots, x_{t-1}) \phi_{jx_t}^{(h)}, \tag{1}$$

where

$$\lambda_{th}(x_1, \ldots, x_{t-1}) = \frac{\sum_{c=1}^{k} p(x_{l+1}, \ldots, x_{t-1}, U_{t-1} = c | x_1, \ldots, x_l) \pi_{ch}}{\sum_{c=1}^{k} p(x_{l+1}, \ldots, x_{t-1}, U_{t-1} = c | x_1, \ldots, x_l)}. \tag{2}$$

Expression (1) clearly indicates that an HMPM is a mixture of PM in which the mixing distribution is time-specific, i.e., weights are time-varying and depend on the previous observations. An explicit expression for $\lambda_{th}(x_1, \ldots, x_{t-1})$ will be given below.

The model can be further generalized by allowing for non-homogeneity, that is, time-specific parameters, both for the manifest and latent processes. Apart from that, no further generalizations relating the manifest distribution are possible and many models result from specific choices of partitions and/or constraints on the parameters. Parametric assumptions (e.g., binomial, negative binomial, etc.) for the manifest distribution are easily seen as constraints on ϕ. Other constraints may be used to reduce the number of minimal parameters, for instance by restricting the hidden transition matrix to be symmetric or banded. The assumption of a diagonal hidden transition matrix leads to a latent class model, i.e., a mixture of partition models with time-fixed weights.

Different kinds of constraints on ϕ can be used to derive MTD and HM-MTD models. An HM-MTD model is specified by setting $l = k > 1$, allowing the manifest to conditionally follow a full lth-order Markov model but with opportune constraints on ϕ. Suppose for instance that $d = l = k = 2$. Then, an HM-MTD is specified by constraining $\phi^{(1)}_{(0,1),x_t} = \phi^{(1)}_{(0,0),x_t} = \phi^{(2)}_{(0,0),x_t} = \phi^{(2)}_{(1,0),x_t}$ and $\phi^{(1)}_{(1,1),x_t} = \phi^{(1)}_{(1,0),x_t} = \phi^{(2)}_{(1,1),x_t} = \phi^{(2)}_{(0,1),x_t}$. Setting $k = 1$ gives back a PM model, with Markov chains, MTD and VLMC as special cases. A VLMC model is in fact a PM model in which partitions are restricted to context trees, i.e., to sets of sequences such that no string in the partition is a suffix of another string in the same partition, e.g., one of the kind $(\{\{000\}, \{100\}, \{010\}, \{110\}\}, \{\{001\}, \{101\}\}, \{\{011\}\}, \{\{111\}\})$ when $l = 3$ and $d = 2$. Setting $p = 0$ gives back an HMM. The new class of models in which a hidden Markov structure and conditional VLMC manifest distributions are assumed is simply obtained by restricting partitions to context trees. The double-chain Markov model of Berchtold (1999) is obtained by setting $l = 1$ and $k > 1$.

We now give a general expression for the equilibrium distribution of any HMPM.

Theorem 1. *Let $X_1, X_2, \ldots$ be distributed according to an HMPM. We have that*

$$\lim_{t\to\infty} p(X_t = x) = \sum_{h=1}^{k} \pi_h \phi^{(h)}_x,$$

where $\phi^{(h)}$ denotes the stationary distribution for the hth transition matrix for the manifest, and π the stationary distribution of the latent transition matrix Π.

Proof. It can be seen that

$$\begin{aligned} p(X_t = x) &= \sum_{j=1}^{p}\sum_{h=1}^{k}\sum_{(x_{t-1},\ldots,x_{t-l})} p(X_t = x|(x_{t-1},\ldots,x_{t-l}) \in L^X_j, U_t = h)P(U_t = h)P((x_{t-1},\ldots,x_{t-l}) \in L^X_j) \\ &= \sum_{j=1}^{p}\sum_{h=1}^{k}\sum_{(x_{t-1},\ldots,x_{t-l})} \phi^{(h)}_{jx} P(U_t = h)P((x_{t-1},\ldots,x_{t-l}) \in L^X_j). \end{aligned}$$

For any h, j and x, $\phi^{(h)}_{jx} = p(X_t = x|(x_{t-1},\ldots,x_{t-l})) \to \phi^{(h)}_x$ and $P(U_t = h) \to \pi_h$, from which the thesis follows. □

Theorem 1 shows that the equilibrium distribution is a finite mixture based on the equilibrium distributions of the manifest and hidden transition matrices. More importantly, it is the same equilibrium distribution as would be obtained from mixtures of first-order Markov chains, i.e., as if the memory for each hth PM model was a first-order one. Theorem 1 directly gives a result for PM models, whose equilibrium distribution is the same as that of the corresponding first-order Markov chain associated with the transition matrix ϕ_{jx}.

3. Inference

For a general HMPM model the likelihood $L(\theta|X)$, with $\theta = (\rho, \pi, \phi)$, can be written as

$$\begin{aligned} L(\theta|X) &= \sum_{(u_1,\ldots,u_T)} p(x_{l+1},\ldots,x_T|x_1,\ldots,x_l,u_{l+1},\ldots,u_T)p(u_{l+1},\ldots,u_T) \\ &= \sum_{(u_1,\ldots,u_T)} p(u_{l+1}) \prod_{t=l+2}^{T} p(u_t|u_{t-1})p(x_t|x_{t-1},\ldots,x_{t-l},u_t) \\ &= \sum_{(u_1,\ldots,u_T)} \rho_{u_{l+1}} \prod_{t=l+2}^{T} \pi_{u_{t-1}u_t} \sum_{j=1}^{p} \phi^{(u_t)}_{jx_t} I((x_{t-1},\ldots,x_{t-l}) \in L^X_j), \end{aligned}$$

where the sum is extended over $(u_1,\ldots,u_T) \in \{1,\ldots,k\}^T$ and $I(\cdot)$ denotes the indicator function. For even moderate values of T it is obviously impractical to evaluate the likelihood directly. We can nevertheless extend a forward recursion from the HMM literature MacDonald and Zucchini (1997, see e.g.). Define $\alpha_t(h) = \Pr(X_{l+1},\ldots,X_t, U_t = h|X_1,\ldots,X_l)$. Due to the Markov assumption, it can be shown that

$$\alpha_{l+1}(h) = \rho_h \sum_{j=1}^{p} \phi^{(h)}_{jx_{l+1}} I((x_1,\ldots,x_l) \in L^X_j),$$

and, for $t = l+2, \ldots, T$,

$$\alpha_t(h) = \sum_{j=1}^{p} \phi_{jx_t}^{(h)} I((x_{t-1}, \ldots, x_{t-l}) \in L_j^X) \sum_{c=1}^{k} \alpha_{t-1}(c)\pi_{ch}.$$

It is straightforward to check that $L(\theta|X) = \sum_{h=1}^{k} \alpha_T(h)$, and that by definition

$$\lambda_{th}(x_1, \ldots, x_{t-1}) = \frac{\sum_{c=1}^{k} \alpha_{t-1}(c)\pi_{ch}}{\sum_{c=1}^{k} \alpha_{t-1}(c)},$$

where $\lambda_{th}(x_1, \ldots, x_{t-1})$ is as in (2).

In order to maximize the likelihood, we set up an EM algorithm. The complete data log-likelihood is given by

$$\begin{aligned} l(\theta|(X,U)) = & \sum_{h=1}^{k} I(u_{l+1} = h) \log(\rho_h) + \sum_{c=1}^{k}\sum_{h=1}^{k}\sum_{t=2}^{T} I(u_t = h, u_{t-1} = c)\pi_{ch} \\ & + \sum_{c=1}^{k}\sum_{t=1}^{T}\sum_{j=1}^{p} \phi_{jx_t}^{(h)} I(u_t = h, (x_{t-1}, \ldots, x_{t-l}) \in L_j^X). \end{aligned} \tag{3}$$

At the E-step, indicators involving the latent variable U_t are replaced by their conditional expected values. In order to compute these, we will compute $\alpha_t(h)$ for $t = l+1, \ldots, T$ and $h = 1, \ldots, k$; and use the following backward recursion in order to compute $\beta_t(h) = \Pr(X_{t+1}, \ldots, X_T | U_t = h, X_1, \ldots, X_l)$ as follows. First, set $\beta_T(h) = 1$. Then, for $t = T-1, \ldots, l+1$, compute

$$\beta_t(h) = \sum_{c=1}^{k}\sum_{j=1}^{p} \phi_{jx_{t+1}}^{(c)} \beta_{t+1}(c)\pi_{hc}.$$

It is straightforward to check that

$$\Pr(U_t = h) = w_t(h) = \frac{\alpha_t(h)\beta_t(h)}{\sum_{c=1}^{k} \alpha_t(c)\beta_t(c)}, \tag{4}$$

and that for $t = l+1, \ldots, T-1$,

$$\Pr(U_{t+1} = h, U_t = c) = w_t(c,h) = \frac{\pi_{ch}\alpha_t(c)\sum_{j=1}^{p} \phi_{jx_{t+1}}^{(h)} \beta_{t+1}(h)}{\sum_{c=1}^{k} \alpha_t(c)\beta_t(c)}. \tag{5}$$

After (4) and (5) have been substituted in (3), the algorithm proceeds with an explicit M-step in which the conditional expected complete likelihood is maximized by setting

$$\hat{\rho}_h = \frac{w_{l+1}(h)}{\sum_{c=1}^{k} w_{l+1}(c)},$$

$$\hat{\pi}_{ch} \propto \sum_{t=l+1}^{T} w_t(c,h);$$

and finally

$$\hat{\phi}_{jx}^{(h)} \propto \sum_{t=l+1}^{T} I(x_t = x, (x_{t-1}, \ldots, x_{t-l}) \in L_j^X) w_t(h).$$

If there are restrictions or constraints on ρ, Π and/or ϕ these can be easily imposed at the M-step. For instance if it is assumed that $\phi_{j_1x}^{(h_1)} = \phi_{j_2x}^{(h_2)}$ then

$$\hat{\phi}_{j_1x}^{(h_1)} = \hat{\phi}_{j_2x}^{(h_2)} \propto \sum_{t=l+1}^{T} I(x_t = x, (x_{t-1}, \ldots, x_{t-l}) \in L_{j_1}^X) w_t(h_1) + I(x_t = x, (x_{t-1}, \ldots, x_{t-l}) \in L_{j_2}^X) w_t(h_2).$$

The E-steps and M-steps are iterated until convergence in $l(\theta) = \log(L(\theta|X))$. In order to reduce the chances of converging to a local maximum for the likelihood, it is recommended that different initial solutions are used.

In order to perform model selection, we propose to use the well known Bayesian information criterion (Schwarz, 1978). Garcia and Gonzalez-Lopez (2010), in a very interesting paper, note that the BIC criterion is consistent in model choice for PM. The BIC of a model is defined as

$$\text{BIC} = -2l(\hat{\theta}) + g\log(T),$$

where g denotes the number of parameters and $l(\hat{\theta})$ the log-likelihood computed at the final estimates. Model choice then proceeds as follows: a set of candidate models is specified, and the model corresponding to the lowest BIC is chosen as the final model.

Acknowledgment

The author acknowledges the financial support of the Einaudi Intitute for Economics and Finance, Rome.

References

Bartolucci, F., Farcomeni, A., 2010. A note on the mixture transition distribution and hidden Markov models. Journal of Time Series Analysis 31, 132–138.
Berchtold, A., 1999. The double chain Markov model. Communications in Statistics—Theory and Methods 28, 2569–2589.
Berchtold, A., Raftery, A.E., 2002. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Statistical Science 17, 328–356.
Bülmann, P., Wyner, A.J., 1999. Variable length Markov chains. Annals of Statistics 27, 480–513.
Dempster, A.P., Laird, N.M., 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.
Farcomeni, A., 2011. Recapture models under equality constraints for the conditional capture probabilities. Biometrika 98, 237–242.
Garcia, J.E., Gonzalez-Lopez, V.A., 2010. Minimal Markov models. arXiv:1002.0729.
MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and other Models for Discrete-Valued Time Series. Chapman and Hall, London.
Raftery, A.E., 1985. A model for high-order Markov chains. Journal of the Royal Statistical Society, Series B 47, 528–539.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.