NONPARAMETRIC HIDDEN MARKOV MODELS

Roland Langrock[1], Thomas Kneib[2] and Alexander Sohn[2]

[1] School of Mathematics and Statistics
University of St Andrews
The Observatory, Buchanan Gardens, St Andrews, KY16 9LZ, UK
(e-mail: roland@mcs.st-and.ac.uk)

[2] Chair of Statistics
Georg August University of Göttingen
Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
(e-mail: tkneib@uni-goettingen.de, asohn@uni-goettingen.de)

ABSTRACT. Hidden Markov models (HMMs) are flexible time series models in which the observed process depends on unobserved serially correlated states. The state-dependent distributions in HMMs are usually taken from some class of parametrically specified distributions. The choice of this class can be difficult, and an unfortunate choice can have serious consequences for example on state estimates and on the interpretation of the underlying system. In a recent paper, Yau *et al.* (2011) suggested a Bayesian approach for estimating the state-dependent distributions in a nonparametric way. We consider an alternative likelihood-based approach, which is based on the idea of representing the densities of the state-dependent distributions as linear combinations of a large number of standardized B-spline basis functions, imposing a penalty term on non-smoothness in order to maintain a good balance between goodness of fit and smoothness.

1 INTRODUCTION

Due to their versatility and mathematical tractability, hidden Markov models (HMMs) have become immensely popular tools for modeling time series in which the observations depend on underlying nonobservable states that are correlated over time. A basic HMM involves two components: 1) an observed time series, which is typically referred to as the state-dependent process, since each of the corresponding observations is assumed to be generated by one of N distributions as determined by the state of 2) an underlying (hidden) N-state Markov chain. In other words, each observation in an HMM is modeled as outcome of a mixture distribution with N components, where the sequence of chosen components is a realization of a Markov chain. Given the current state, each observation is typically assumed to be conditionally independent from previous observations and states. Usually the Markov chain is assumed to be of first order (but see, e.g., Langrock, 2012, for other structures). A key property of HMMs is that dynamic programming algorithms can be used to efficiently evaluate the likelihood and to estimate the unobserved state sequence. HMMs have been successfully applied in a diverse range of fields, including DNA sequence analysis (Durbin *et al.*, 1998), finance (Langrock *et al.*, 2012; Banachewicz *et al.*, 2008), medicine (Langrock *et al.*, in press) and ecology (Zucchini *et al.*, 2008; Langrock and King, in press).

In the existing literature on HMMs, it is usually assumed that each state-dependent distribution is from a class of parametrically specified distributions. Choosing a parametric family

which is sufficiently flexible yet tractable can be difficult, for example if the true (unknown) state-dependent distribution is heavy-tailed, skewed or multi-modal. An unfortunate choice of the parametric family can lead, *inter alia*, to a poor fit, to biased estimates of the state transitions (and hence also to a bad performance of the state decoding), to poor predictive power and also to wrong conclusions regarding the underlying system to be modeled. In a recent paper, Yau *et al.* (2011) suggested a nonparametric specification of the state-dependent distributions of an HMM for a time series of continuous-valued observations. Their approach involves mixtures of Dirichlet processes and uses computationally expensive Markov chain Monte Carlo (MCMC) simulation techniques.

We pursue the same aim as Yau *et al.* (2011), namely to fit the state-dependent distributions in a nonparametric way. However, we consider an alternative non-Bayesian estimation approach and an alternative specification for the state-dependent distributions. We represent these as linear combinations of a large number of standardized B-spline basis functions, while imposing a penalty on non-smoothness in order to arrive at an appropriate balance between goodness of fit and smoothness for the fitted densities. This essentially leads to an overparameterized model since, in fact, each B-spline basis function is still associated with a separate parameter, leading to a finite dimensional model, but the dimensionality is high and the separate parameters are no longer of interest or interpretable. We therefore call our approach nonparametric despite the fact that it relies on a parametric specification with a large number of parameters. This is in line with the standard terminology in the statistical literature on smoothing methods where (penalized) spline approaches are also subsumed under nonparametric approaches (see, for example, Ruppert *et al.*, 2003). We expect our approach to be particularly useful in scenarios where the distribution of observations within states appears to be of a complicated form, making it hard to specify a suitable parametric family. It can also be used as an exploratory tool in cases where it is not immediately clear how to choose a parametric specification. Our approach is fairly easy to implement and exploits the strengths both of likelihood-based HMM machinery and of penalized B-splines (i.e., P-splines).

2 HMMs WITH NONPARAMETRIC STATE-DEPENDENT DISTRIBUTIONS

We begin by reviewing some basics on HMMs, also introducing the required notation. For an HMM, the observable state-dependent process in the following is denoted by $\{X_t\}_{t=1}^T$, and the underlying nonobservable N-state Markov chain by $\{S_t\}_{t=1}^T$. We consider the basic dependence structure where given the current state of S_t, the variable X_t is conditionally independent from previous and future observations and states, and where the Markov chain is of first order. Figure 1 displays the dependence structure of such an HMM in a directed acyclic graph.

For a homogeneous Markov chain, we summarize the probabilities of transitions between the different states in the $N \times N$ transition probability matrix (t.p.m.) $\Gamma = (\gamma_{ij})$, where

$$\gamma_{ij} = \Pr\big(S_{t+1} = j | S_t = i\big), \quad i, j = 1, \ldots, N.$$

The initial state probabilities are summarized in the vector $\delta = (\Pr(S_1 = 1), \ldots, \Pr(S_1 = N))$. It is typically assumed that δ is the stationary distribution of the Markov chain (if that exists).

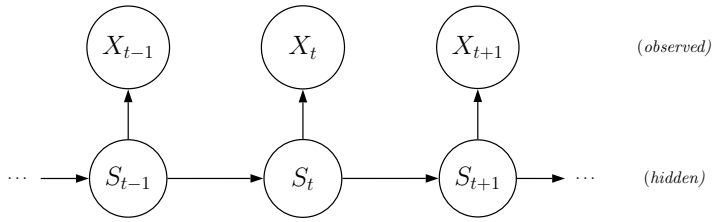

Figure 1. Dependence structure of a basic HMM.

Parameter estimation for HMMs can usually be done by (numerically) maximizing the log-likelihood. For an HMM as described above, with observations given by $x_1,\ldots,x_T$, underlying states denoted by $s_1,\ldots,s_T$ and parameter vector θ, the likelihood is given by

$$\begin{aligned}\mathcal{L}^{\text{HMM}}(\theta) &= f(x_1,\ldots,x_T)\\ &= \sum_{s_1=1}^{N}\cdots\sum_{s_T=1}^{N} f(x_1,\ldots,x_T|s_1,\ldots,s_T)f(s_1,\ldots,s_T)\\ &= \sum_{s_1=1}^{N}\cdots\sum_{s_T=1}^{N}\prod_{i=t}^{T} f(x_t|s_t)\delta_{s_1}\prod_{t=2}^{T}\gamma_{s_{t-1},s_t}\,.\end{aligned}$$

Here f was used as a general symbol for a density, either discrete or continuous. In this form the likelihood involves N^T summands, which would make numerical maximization infeasible even for a small number of states N and a moderate number of observations T. Fortunately, there is a much more efficient way of calculating the likelihood $\mathcal{L}^{\text{HMM}}(\theta)$, given by a recursive scheme called the *forward algorithm*. To see this, we consider the vectors of forward variables, defined as

$$\begin{gathered}\alpha_t = \big(\alpha_t(1),\ldots,\alpha_t(N)\big)\,,\ t=1,\ldots,N,\\ \text{where}\ \ \alpha_t(j) = f(x_1,\ldots,x_t,S_t=j)\ \ \text{for}\ \ j=1,\ldots,N.\end{gathered}$$

Then we can apply the recursive scheme:

$$\begin{aligned}\alpha_1 &= \delta Q(x_1)\,,\\ \alpha_{t+1} &= \alpha_t\Gamma Q(x_{t+1})\,,\end{aligned}\tag{1}$$

where $Q(x_t) = \text{diag}\big(f_1(x_t),\ldots,f_N(x_t)\big)$, with $f_i(x_t) = f(x_t|S_t=i)$ denoting the density of the i-th state-dependent distribution. The recursion (1) can easily be derived using the HMM dependence assumptions and some straightforward matrix algebra (see, e.g., the proof of Proposition 2 in Zucchini and MacDonald, 2009). The likelihood can then be written as a matrix product:

$$\mathcal{L}^{\text{HMM}}(\theta) = \sum_{i=1}^{N}\alpha_T(i) = \delta Q(x_1)\Gamma Q(x_2)\ldots\Gamma Q(x_T)1\,,\tag{2}$$

where $1 \in R^N$ is a column vector of ones. The computational cost of evaluating (2) is *linear* in the number of observations, T, such that a numerical maximization of the likelihood becomes feasible in most cases.

We are concerned with the nonparametric estimation of the densities $f_1, \ldots, f_N$, which we conduct following ideas from Schellhase and Kauermann (2012). More specifically, we suggest to represent each of these densities as a finite linear combination of B-spline densities $\phi_{-K}, \ldots, \phi_K$ (B-splines standardized such that they integrate to 1):

$$f_i(x) = \sum_{k=-K}^{K} a_{ik}\phi_k(x), \quad i = 1, \ldots, N. \tag{3}$$

Clearly, $f_i(x)$ is a probability density function if $\sum_{k=-K}^{K} a_{ik} = 1$ and $a_{ij} \geq 0$ for all $j = -K, \ldots, K$. To enforce these constraints, the coefficients to be estimated, $a_{i,-K}, \ldots, a_{iK}$, are transformed using the multinomial logit link

$$a_{ik} = \frac{\exp(\beta_{ik})}{\sum_{j=-K}^{K} \exp(\beta_{ij})}, \tag{4}$$

where we set $\beta_{i0} = 0$ for identifiability. The number of B-splines employed in the mixture specification (3), $2K+1$, determines the potential flexibility, where a larger number of basis elements will yield estimates that follow the data very closely but may be too wiggly while a small number of basis elements yields very smooth estimates that may, however, be severely biased. To overcome the problem of selecting an optimal number of basis elements, we follow the penalized spline approach by Eilers and Marx (1996) and modify the log-likelihood by including a penalty on the sum of squared (m-th order) differences between coefficients associated with adjacent B-splines. We hence do not maximize the likelihood given in (2), but instead a penalized log-likelihood, given by

$$l_p^{\mathrm{HMM}}(\theta, \lambda) = \log\big(\mathcal{L}^{\mathrm{HMM}}(\theta)\big) - \left[\sum_{i=1}^{N} \frac{\lambda_i}{2} \sum_{k=-K+m}^{K} \big(\Delta^m a_{ik}\big)^2\right], \tag{5}$$

where $\Delta a_k = a_k - a_{k-1}$ and $\Delta^m a_k = \Delta(\Delta^{m-1} a_k)$. The parameter vector θ comprises the state transition probabilities and the parameters β_{ki} ($-K \leq k \leq K$, $k \neq 0$, $i = 1, \ldots, N$), and $\lambda = (\lambda_1, \ldots, \lambda_N)$ is a vector of smoothing parameters. The smoothing parameters may be chosen to be different across states, e.g., if for some state-dependent distributions the (true) densities are much more wiggly than for others, or if some states of the Markov chain are visited much less frequently than others, potentially requiring higher penalties on roughness due to a decreased number of observations being available. We choose the smoothing parameters using a computer-intensive cross-validation procedure, comparing various values from a prespecified grid using likelihood scores.

3 Some simulation experiments

To demonstrate the feasibility of the suggested approach, we conduct a simple simulation experiment. We consider a two-state HMM where the state-dependent densities substantially

overlap, with one of the two distributions being bimodal (see illustration in Figure 2). In practice these features would make it difficult to specify an adequate parametric HMM based on a visual inspection of the (bimodal) marginal distribution of the data (e.g., by a histogram). Also it may intuitively not be clear that our nonparametric approach is able to allocate the observations associated with the smaller peak of the bimodal distribution (at about $x = -5$) to the right state, since the marginal distribution suggests its association with the unimodal state; without the correlation over time there clearly would be no chance of correctly identifying the underlying structure. The states of the Markov chain where generated using the t.p.m.

$$\Gamma = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}.$$

In each of 20 simulation runs, $T = 1000$ observations were generated. We then fitted one-, two- and three-state models in order to additionally illustrate that we are able to do a model selection on the number of states, N. This model selection, which was based on the likelihood scores obtained in the cross-validation procedure, led to correct identification of the underlying two-state model in all 20 cases (in each case the scores obtained for the two-state model were significantly better than for the one-state model, and there was no significant difference in the scores for the two- and the three-state model, respectively, in which case the simpler model is to be preferred). For the 20 simulation runs, the sample mean estimates of the transition probabilities γ_{11} and γ_{22} are 0.953 and 0.952, respectively, with sample standard deviations of 0.007 and 0.011, respectively. The fitted densities of the state-dependent distributions are illustrated in Figure 2. The results demonstrate that the nonparametric approach works very well in this fairly difficult scenario.

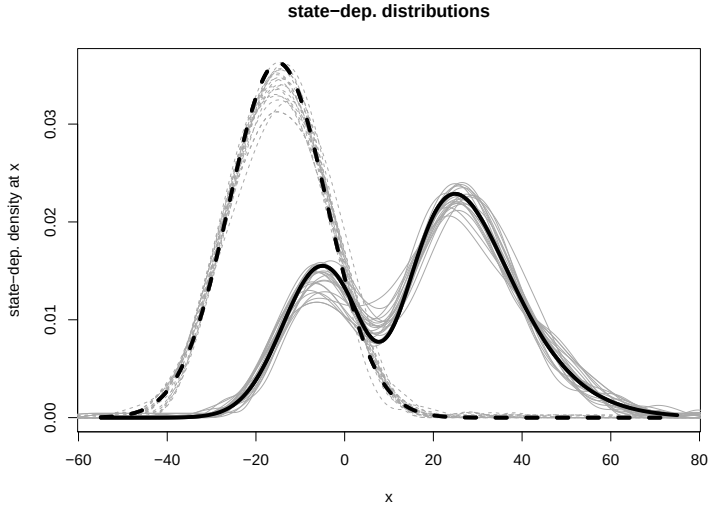

Figure 2. True (black) and estimated (grey) densities of the state-dependent distributions in the simulation experiments.

REFERENCES

BANACHEWICZ, K., LUCAS, A., VAART, A. (2008): Modelling portfolio defaults using hidden Markov models with covariates. *Econometrics Journal*, *11*, 155–171.

DURBIN, R., EDDY, S., KROGH, A., MITCHISON, G.J. (1998): *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

EILERS, P.H.C, MARX, B.D. (1996): Flexible smoothing with *B*-splines and penalties. *Statistical Science*, *11*, 89–121.

LANGROCK, R. (2012): Flexible latent-state modelling of Old Faithful's eruption interarrival times in 2009. *Australian and New Zealand Journal of Statistics*, *54*, 261–279.

LANGROCK, R., MACDONALD, I.L., ZUCCHINI, W. (2012): Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, *19*, 147–161.

LANGROCK, R., KING, R. (in press): Maximum likelihood estimation of mark-recapture-recovery models in the presence of continuous covariates. *Annals of Applied Statistics*.

LANGROCK, R., SWIHART, B., CAFFO, B., CRAINICEANU, C., PUNJABI, N. (in press): Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine*.

RUPPERT, D., WAND, M.P., CARROLL, R.J. (2003): *Semiparametric Regression*. Cambridge University Press, Cambridge.

SCHELLHASE, C., KAUERMANN, G. (2012): Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, *27*, 757–777.

YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G.O., HOLMES, C. (2011): Bayesian nonparametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B*, *73*, 37–57.

ZUCCHINI, W., MACDONALD, I.L. (2009): *Hidden Markov models for time series: an introduction using R*. Chapman & Hall, London.

ZUCCHINI, W., RAUBENHEIMER, D., MACDONALD, I.L. (2008): Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, *64*, 807–815.