Computational Statistics and Data Analysis 55 (2011) 715–724

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Hidden Markov models with arbitrary state dwell-time distributions

R. Langrock*, W. Zucchini

Institute for Statistics and Econometrics, University of Goettingen, Platz der Goettinger Sieben 5, 37073 Goettingen, Germany

ARTICLE INFO

Article history:
Received 6 July 2009
Received in revised form 14 June 2010
Accepted 14 June 2010
Available online 25 June 2010

Keywords:
Daily rainfall occurrence
Dwell-time distribution
Hidden Markov model
Hidden semi-Markov model
Numerical likelihood maximization

ABSTRACT

A hidden Markov model (HMM) with a special structure that captures the 'semi'-property of hidden semi-Markov models (HSMMs) is considered. The proposed model allows arbitrary dwell-time distributions in the states of the Markov chain. For dwell-time distributions with finite support the HMM formulation is exact while for those that have infinite support, e.g. the Poisson, the distribution can be approximated with arbitrary accuracy. A benefit of using the HMM formulation is that it is easy to incorporate covariates, trend and seasonal variation particularly in the hidden component of the model. In addition, the formulae and methods for forecasting, state prediction, decoding and model checking that exist for ordinary HMMs are applicable to the proposed class of models. An HMM with explicitly modeled dwell-time distributions involving seasonality is used to model daily rainfall occurrence for sites in Bulgaria.

1. Introduction

Hidden Markov models (HMMs), or Markov-dependent mixture models, have been successfully applied to a wide range of types of time series: continuous-valued, circular, multivariate, as well as binary data, bounded and unbounded counts and categorical observations (see for example Zucchini and MacDonald, 2009). HMMs comprise two components: an unobserved (hidden) Markov chain and a state-dependent process. Each realization is assumed to be generated by one of N distributions as determined by the state of an N-state Markov chain. The realizations are assumed to be conditionally independent, given the states. For comprehensive accounts of the theory of HMMs see, e.g., Ephraim and Merhav (2002) or Cappé et al. (2005).

The number of consecutive time points that the Markov chain spends in a given state (the dwell time) follows a geometric distribution. Thus, for example, the modal dwell time for every state of an HMM is one. Hidden semi-Markov models (HSMMs) are designed to relax this restrictive condition; the dwell time in each state can follow any discrete distribution on the natural numbers. HSMMs and their applications are discussed, inter alia, in Ferguson (1980), Sansom and Thomson (2001) and Yu and Kobayashi (2003).

The additional generality offered by HSMMs carries a computational cost; they are more demanding to apply than are HMMs. Furthermore, in HSMMs state changes and state dwell-time distributions are modeled separately, meaning that the embedded Markov chain operates on a non-uniform time scale, and consequently the Markov property is lost. In some applications this can be regarded as natural, e.g. in the modeling of breakpoint rainfall data in Sansom and Thomson (2001). However, in general it can be unnatural and it leads to difficulties if one wants to do prediction or if one wishes to include covariates, trend or seasonality in the model. Covariate modeling in HMMs, in the state-dependent process as well as in the Markov chain, has been broadly explored and is fairly standard (see, e.g., Bartolucci et al., 2009 and Part Two in Zucchini and MacDonald, 2009). For HMMs, random effects can also be included in both components of the model; parameters then can be estimated using a deterministic approximation or a Monte Carlo expectation-maximization (MCEM) algorithm (see Altman, 2007). Chaubert-Pereira et al. (2010) proposed an MCEM-like algorithm for including random effects in the

* Corresponding author. Tel.: +49 551 394605; fax: +49 551 397279.
E-mail address: rsoleck@uni-goettingen.de (R. Langrock).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.015

semi-Markovian framework. However, the incorporation of covariates or random effects in the latent part of the HSMM, i.e. in the semi-Markov chain, is much more difficult due to the timing problems arising from the separate modeling of state changes and state sojourns. This problem can be circumvented by using HMMs with arbitrary state dwell-time distributions; the existing HMM methodology then becomes available. In particular and in contrast to the HSMM case, simple expressions for the forecasts are available for HMMs. Furthermore, in the HSMM literature it is generally assumed that the start of the series coincides with a state switch (see, e.g. Sansom and Thomson, 2001; Guédon, 2003; Bulla et al., 2009). This assumption in general makes the HSMM non-stationary. Indeed it turns out that fitting stationary (as opposed to only homogeneous) HSMMs is not as easy as in the case of HMMs. To the best of our knowledge, this still is an unsolved problem. Finally, the elaborateness of the EM algorithm, which usually is applied to fit HSMMs, increases unless one makes the simplifying assumption that there is a state switch at the end of the series (see Guédon, 2003).

We consider specially structured HMMs that capture the 'semi'-property of HSMMs, i.e., that can fit – at least approximately – any desired state dwell-time distribution, both parametric and nonparametric ones. The idea is to use so-called "state aggregates" which in similar, but not identical, ways have also been discussed in, e.g., Russell and Cook (1987) and Guédon (2005). A good overview of the various existing approaches can be found in Johnson (2005) who also argues that the usage of this kind of models is "almost certainly a much better practical choice for duration modeling than development and implementation of more complex and computationally expensive models with explicit modifications to handle duration probabilities". Our HMM formulation is designed to fit, at least approximately, *any* given dwell-time distribution. It has exactly the same number of parameters as the corresponding HSMM and, in theory, the approximation of any dwell-time distribution can be made arbitrarily accurate. Furthermore, there are important subclasses of distributions for which the representation of the dwell-time distributions in the HMM formulation is exact. The topology of the state aggregates in our model is convenient since it is the same whatever dwell-time distribution we want to represent. Using our HMM formulation and direct likelihood maximization it is straightforward to fit stationary HMMs (with arbitrary state dwell-time distributions), or to incorporate trend, seasonality and covariates in the models, either in the hidden process or in the observed process. Furthermore, the whole standard HMM methodology including state prediction, local and global decoding, forecasting and model checking as described in, e.g., Zucchini and MacDonald (2009) is applicable.

In Section 2 we describe how to structure an HMM so that it has the desired state dwell-time distribution. We illustrate the HMM formulation using a number of examples. Section 3 is about parameter estimation by numerical maximization of the likelihood function and the numerical complexity of its evaluation. In Section 4 we describe an application, namely the modeling of daily rainfall occurrence for sites in Bulgaria. We illustrate how it is possible to incorporate seasonality in both the hidden and the observed process.

2. The model

An HSMM comprises an observable output process $(X_t)_{t=1,\dots,T}$, where the distribution of X_t is determined by the state, S_t, of an unobserved (hidden) N-state semi-Markov process $(S_t)_{t=1,\dots,T}$ (for a general reference about semi-Markov processes see Kulkarni, 1995). Conditional on the states the observations of the output process are assumed to be independent.

Let p_k denote the probability mass function (p.m.f.) of the dwell time in state $k \in \{1, \dots, N\}$ and let F_k denote its distribution function. The support of p_k is $\mathbb{N}$, the set of natural numbers, or some subset of $\mathbb{N}$.

Consider the subsequence of $(S_t)_{t=1,\dots,T}$ comprising the first occurrences of states in each run. (For example the subsequence corresponding to 1, 1, 2, 2, 2, 1, 3, 3, 3, 3 is 1, 2, 1, 3.) We assume that this subsequence is generated by an irreducible and ergodic Markov chain (the 'embedded Markov chain') with the transition probability matrix (t.p.m.) $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} = \Pr\big(S_{t+1} = j | S_t = i, S_{t+1} \neq i\big), \quad i, j = 1, \dots, N,$$

$\sum_j a_{ij} = 1$, $a_{ii} = 0$, and that the initial probabilities for this Markov chain are given by δ_i, $i = 1, \dots, N$. An HMM is the special case of an HSMM for which p_k is the p.m.f. of the geometric distribution.

Our aim now is to show how one can structure an HMM such that it is a reformulation of any given HSMM, i.e., such that it represents any desired dwell-time distribution. Let $m_1, m_2, \dots, m_N \in \mathbb{N}$, $m_0 := 0$, and consider an HMM with state-dependent process $(X_t^\star)_{t=1,\dots,T}$ (observable) and Markov chain $(S_t^\star)_{t=1,\dots,T}$ (unobservable) with states $\{1, 2, \dots, \sum_{i=1}^N m_i\}$. We refer to the sets

$$I_k := \left\{ n \mid \sum_{i=0}^{k-1} m_i < n \leq \sum_{i=0}^{k} m_i \right\}, \quad k = 1, \dots, N,$$

as *state aggregates*, and define $i_k^- := \min(I_k)$ and $i_k^+ := \max(I_k)$. We assume that each state of I_k is associated with the same distribution of the state-dependent process, namely the distribution associated with the kth state in the HSMM defined above. In other words the distribution of X_t^* given a state $S_t^* = l$ is the same for all $l \in I_k$ and is also the same as that of X_t given $S_t = k$, i.e.

$$\Pr^{X_t^\star | S_t^\star \in I_k} = \Pr^{X_t | S_t = k}, \quad t = 1, \dots, T,\ k = 1, \dots, N. \tag{1}$$

We denote the t.p.m. of $(S_t^*)_{t=1,\ldots,T}$ by $\mathbf{B} = \{b_{ij}\}$, where $b_{ij} = \Pr\left(S_{t+1}^\star = j | S_t^\star = i\right)$, $i, j = 1, \ldots, \sum_{i=1}^N m_i$, and structure it as follows:

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \ldots & \mathbf{B}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{B}_{N1} & \ldots & \mathbf{B}_{NN} \end{pmatrix}, \tag{2}$$

where the $m_i \times m_i$ diagonal matrices $\mathbf{B}_{ii}$ $(i = 1, \ldots, N)$ are defined, for $m_i \geq 2$, as

$$\mathbf{B}_{ii} := \left(\begin{array}{c|ccccc} 0 & 1 - c_i(1) & 0 & \ldots & & 0 \\ \vdots & 0 & \ddots & & & \vdots \\ & \vdots & & & & 0 \\ 0 & 0 & \ldots & 0 & & 1 - c_i(m_i - 1) \\ \hline 0 & 0 & \ldots & 0 & & 1 - c_i(m_i) \end{array}\right), \tag{3}$$

$\mathbf{B}_{ii} := 1 - c_i(1)$ for $m_i = 1$, and the $m_i \times m_j$ off-diagonal matrices $\mathbf{B}_{ij}$ $(i, j = 1, \ldots, N, i \neq j)$ as

$$\mathbf{B}_{ij} := \begin{pmatrix} a_{ij}c_i(1) & 0 & \ldots & 0 \\ a_{ij}c_i(2) & 0 & \ldots & 0 \\ \vdots & & & \\ a_{ij}c_i(m_i) & 0 & \ldots & 0 \end{pmatrix}. \tag{4}$$

(In case of $m_j = 1$ the zeros disappear.) Here, for $r = 1, 2, 3, \ldots$,

$$c_k(r) := \begin{cases} \dfrac{p_k(r)}{1 - F_k(r-1)} & \text{for } F_k(r-1) < 1, \\ 1 & \text{for } F_k(r-1) = 1. \end{cases}$$

Note first that the functions c_k – the *hazard rates* of the dwell time distributions – play the key role in our HMM formulation since they are responsible for rendering the desired dwell-time distributions p_k. Different dwell-time distributions lead to different c_k's, while the a_{ij}'s and the structure of the t.p.m. remain unaffected. The c_k's are generated solely from the parameters of the dwell-time distributions, no additional parameters or constraints occur in comparison to the HSMM.

Note also that the matrix $\mathbf{B}$ indeed constitutes a t.p.m. since the entries all lie in the interval [0, 1] and the row sums are equal to one. Although $\mathbf{B}$ appears to be somewhat complicated, its structure is not difficult to interpret: All transitions within state aggregate I_k are governed by the diagonal matrix $\mathbf{B}_{kk}$ which thus determines the dwell-time distribution of that state aggregate. The off-diagonal matrices determine the probabilities of transitions between different state aggregates. For example, for $k \neq l$, the matrix $\mathbf{B}_{kl}$ contains the probabilities of all possible transitions between the state aggregates I_k and I_l. Note that in this construction a transition from I_k to I_l must enter I_l in state i_l^-. We will now show that this choice of $\mathbf{B}$ yields an HMM that is a reformulation of the given HSMM.

We denote the probability of a transition from the state aggregate I_i to I_j by $a_{ij}^\star := \Pr\left(S_{t+1} \in I_j | S_t \in I_i, S_{t+1} \notin I_i\right)$. This is analogous to the transition probability a_{ij} in the HSMM. It is shown in the Appendix that:

Proposition 1. *For $i \neq j$, $1 \leq i, j \leq N$,*

$$a_{ij}^\star = a_{ij}.$$

We focus now on the accuracy of the representation of the dwell-time distributions p_k in our HMM formulation. We denote the p.m.f. of the dwell-time distribution in the state aggregate I_k by $p_k^\star$ $(k = 1, \ldots, N)$. With the possible exception of the state aggregate that is active at $t = 1$, the stay in a given aggregate I_k begins in state i_k^-. By $p_k^\star$ we refer to the distributions of those dwell times that do start in state i_k^-.

Proposition 2. *For any $k \in \{1, \ldots, N\}$*

$$p_k^\star(r) = \begin{cases} p_k(r) & \text{for } r \leq m_k, \\ p_k(m_k)(1 - c_k(m_k))^{r - m_k} & \text{for } r > m_k. \end{cases}$$

The two p.m.f.'s thus differ only for $r > m_k$, i.e. in the right tail. Clearly, the difference between p_k and $p_k^\star$ can be made arbitrarily small by choosing m_k sufficiently large. It also follows that, for *any dwell-time distribution with finite support*, we can ensure that $p_k^\star(r) = p_k(r)$ $\forall r$ by choosing m_k to be the maximum dwell time in state k having a non-zero probability.

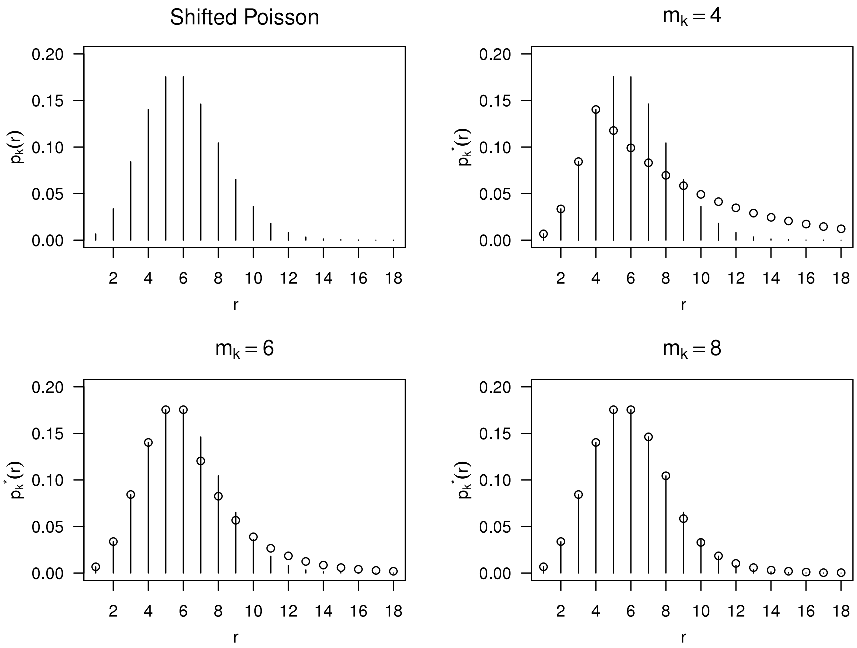

Fig. 1. The functions $p_k(r)$ (vertical lines), and $p_k^\star(r)$ (circles) for $m_k = 4, 6, 8$, from Example 1.

The use of state aggregates to allow for non-geometric dwell-time distributions while preserving the Markovian unit time scale has been suggested before, but in ways that are different to that proposed here. Well-known is the so-called *method of stages*, which can be divided into the *stages in series* and the *stages in parallel* (see Cox and Miller, 1965). The former method leads to a distribution which is a convolution of geometric dwell-time distributions and can be used to fit HSMMs with negative binomial dwell-time distributions (see Guédon, 2005). The method of stages in parallel on the other hand corresponds to mixtures of geometric distributions which involves greater dispersion than that of a single geometric. Using combinations of both methods, one can approximate any distribution on the natural numbers (see Cox and Miller, 1965). However, as pointed out by Kleinrock (1975), a good approximation can require a "horribly complicated" state-aggregate structure. Consequently this theoretical result is likely to be of limited usefulness in applications except in some special cases, e.g. where the dwell-time distribution is negative binomial, a mixture of geometric distributions, or some simple combination of these two possibilities. In contrast the HMM formulation presented here has the same simple state-transition diagram whatever distribution we want to approximate. Another approach using state aggregates utilizes the Fergusson topology and is designed to allow for arbitrary dwell-time distributions having finite support (see Russell and Cook, 1987). This is equivalent to the special case of our model in which no self-transition in the last visited state of the aggregates is allowed. An overview of the existing approaches can be found in Johnson (2005).

The following examples illustrate the approximation using our HMM formulation.

Example 1. Let p_k be the p.m.f. of a shifted Poisson distribution:

$$p_k(r) = \exp(-\lambda_k)\frac{\lambda_k^{r-1}}{(r-1)!}, \quad r = 1, 2, \ldots .$$

Then, according to Proposition 2,

$$p_k^\star(r) = \begin{cases} p_k(r) & \text{for } r \le m_k, \\ p_k(m_k) z^{r-m_k} & \text{for } r > m_k, \end{cases}$$

where $z = 1 - p_k(m_k)/\big(1 - F_k(m_k - 1)\big)$ is independent of r. Although the functions $p_k^\star(r)$ and $p_k(r)$ differ for $r > m_k$, the discrepancy between them becomes small as m_k increases. This is illustrated in Fig. 1, which displays a shifted Poisson distribution with parameter $\lambda = 5$ and the corresponding $p_k^\star(r)$ for $m_k = 4, 6, 8$.

Example 2. Let p_k be the p.m.f. of the geometric distribution:

$$p_k(r) = \pi_k(1 - \pi_k)^{r-1}, \quad r = 1, 2, \ldots .$$

Then, for $r \ge 2$,

$$c_k(r) = \frac{p_k(r)}{1 - \sum_{s=1}^{r-1} p_k(s)} = \frac{\pi_k(1-\pi_k)^{r-1}}{\sum_{s=r}^{\infty} \pi_k(1-\pi_k)^{s-1}}$$
$$= \frac{\pi_k}{\sum_{s=0}^{\infty} \pi_k(1-\pi_k)^{s}} = \pi_k.$$

By Proposition 2 we have that $p_k^\star(r) = p_k(r)$ for $r \le m_k$ and, for $r > m_k$, that

$$p_k^\star(r) = p_k(m_k)\big(1 - c_k(m_k)\big)^{r-m_k} = \pi_k(1-\pi_k)^{m_k-1}(1-\pi_k)^{r-m_k}$$
$$= \pi_k(1-\pi_k)^{r-1} = p_k(r).$$

In this example the choice of m_k thus does not play any role; the HSMM reduces to an HMM.

Example 3. Consider a shifted binomial distribution with p.m.f.

$$p_k(r) = \binom{n_k}{r-1} \pi_k^{r-1}(1-\pi_k)^{n_k-(r-1)}, \quad r = 1, \ldots, n_k + 1.$$

Since $\sum_{i=1}^{n_k+1} p_k(i) = 1$, we have, for $m_k = n_k + 1$, that $c_k(m_k) = p_k(n_k+1)/(1 - \sum_{i=1}^{n_k} p_k(i)) = 1$. For $r > m_k$, application of Proposition 2 yields

$$p_k^\star(r) = p_k(m_k)\big(1 - c_k(m_k)\big)^{r-m_k} = 0 = p_k(r).$$

Proposition 2 also guarantees that $p_k^\star(r) = p_k(r)$ for $r \le m_k$. Thus choosing $m_k = n_k + 1$ ensures that $p_k^\star(r) = p_k(r)\ \forall r$.

In summary, Propositions 1 and 2, together with (1), imply that our HMM formulation is capable of representing, at least approximately, any given dwell-time distribution. In other words, with respect to HSMMs, the distribution of $(X_t)_{t=1,\ldots,T}$ can be approximated by that of $(X_t^\star)_{t=1,\ldots,T}$, where the approximation can be made arbitrarily accurate by choosing the m_k sufficiently large.

We now consider situations in which it can be beneficial to use the HMM formulation instead of HSMMs. First, the separate modeling of state changes and state dwell-time distributions in the hidden part of HSMMs can lead to difficulties when introducing covariates, including trend and seasonality, in the latent process. Up to now hardly any work on seasonal HSMMs has been published, except the contributions of Sansom and Thomson, e.g. in 2007. On the other hand covariate modeling in HMMs was considered numerous times in the literature (see, e.g., Altman, 2007; Bartolucci et al., 2009 and Part Two of Zucchini and MacDonald, 2009). In principle it is straightforward to fit HMMs involving covariates or random effects, both in the state-dependent process and in the latent process. A benefit of using HMMs with arbitrary state dwell-time distributions is that it enables the transfer of the very general MHMM approach of Altman (2007) to the HSMM setting. In addition, standard HMM techniques, e.g. for forecasting and model checking, are available.

Another question of interest concerns the choice of the initial distribution of the process. Suppose that the HSMM is in state k at time $t = 1$. Now unless the state process *entered* state k at $t = 1$ (i.e. unless $S_0 \neq k$) the distribution of the first dwell time will differ from p_k in general. It is to circumvent the difficulties of taking this difference into account that Sansom and Thomson (2001), as well as Guédon (2003) and Bulla et al. (2009), assume that the time instant $t = 1$ is a state boundary. The initial distribution δ then is explicitly modeled and can feasibly be estimated (see Guédon, 2003). This assumption is unlikely to have much effect on parameter estimation except for short series. However, despite being natural in many applications (see, e.g. Sansom and Thomson, 2001), it can be unrealistic in others. More serious, the enforced state switch at the start of the series in general impedes stationarity of the HSMM. In fact, there is no simple procedure yet available to fit a stationary HSMM. In contrast, the assumption of a state switch at the start of the series is easily avoidable using our model by allowing the initial probabilities, δ_i^*, to be nonzero also for $i \notin \{i_1^-, i_2^-, \ldots, i_N^-\}$. Moreover, it is straightforward to fit a stationary HMM: The initial distribution δ^* of a stationary HMM with t.p.m. $\mathbf{B}$ is the solution to the linear system $\delta^{*t}\mathbf{B} = \delta^{*t}$, and thus can be expressed in terms of the parameters that determine $\mathbf{B}$. Thus, in effect, the model formulation considered here allows one to fit stationary HSMMs. On the other hand this formulation can also deal with the case where the user wishes to assume that there is a state boundary at time $t = 1$.

The assumption that there is a state change at the end of the series impacts on the forecast distribution as well as on the likelihood. Furthermore, it excludes the possibility of absorbing states in the semi-Markov chain (Guédon, 2003). Both theory (Guédon, 2003) and software (Bulla et al., 2009) have been developed to fit HSMMs that avoid such an assumption. On the other hand it is a convenient feature of the HMM formulation that no such modification of the standard algorithm is required.

The following two examples illustrate other possible variants of the model.

Example 4. To model rainfall data Sansom and Thomson (2001) proposed an HSMM having dwell-time p.m.f. of the following form:

$$p(r)=\begin{cases}\theta_r & \text{for } r\le d,\\ \theta(1-\theta)^{r-d} & \text{for } r>d,\end{cases}$$

where $\theta := \theta_1+\cdots+\theta_d$. This distribution on the positive integers has an unstructured start and a geometric tail. We say that the start is *of order* $d-1$, motivated by the fact that in comparison to the geometric distribution $d-1$ additional parameters are considered. (Note that the case $d = 1$ (order 0) reduces to a geometric distribution.) This HSMM can be formulated as an HMM – in the manner of (2) – which has the identical probability structure if we choose the corresponding m_i equal to d.

Example 5. Guédon (2005) studied *hybrid models* in which some of the states are Markovian and others are semi-Markovian. These can easily be described using our model by choosing $m_i = 1$ and thus

$$\mathbf{B}_{ii}=1-c_i(1)=1-\pi_i,\qquad \mathbf{B}_{ij}=a_{ij}c_i(1)=a_{ij}\pi_i,$$

for each Markovian state i, where π_i is the parameter of the corresponding geometric distribution, and by defining $\mathbf{B}_{kk}$ and $\mathbf{B}_{kl}$ according to (3) and (4) respectively for each semi-Markovian state k.

Clearly, the HMM formulation can also render dwell-time distributions from *different families* in one model. We just define the block matrices according to (3) and (4). If all the dwell-time distributions have either finite support or geometric tails then the HMM formulation is equivalent, otherwise it is approximate. In the latter case the approximation can be made arbitrarily accurate.

3. Estimation of the parameters

An important benefit of using the HMM formulation is that this enables us to apply all the well-established methods available for HMMs. In particular the likelihood function of the resulting HMM is available in a form that is easy to compute, the parameters thus can be estimated by direct numerical likelihood maximization. Cappé et al. (2005, Chapter 12) show that, under certain regularity conditions, the MLEs of HMM parameters are consistent, asymptotically normal and efficient.

For convenience we restrict our attention to the case in which the state-dependent process is discrete-valued; the results for the continuous-valued case are similar. We denote the p.m.f. of the state-dependent process by

$$q_{\boldsymbol{\Theta}_k}(x)=\Pr\big(X_t=x|S_t\in I_k\big)$$

where $\boldsymbol{\Theta}_k$ is the set of parameters for the state-dependent process corresponding to the state aggregate I_k. Being an HMM the likelihood of our model can be written as the following product (see, e.g., Zucchini and MacDonald, 2009):

$$L_T=\delta^{*}\mathbf{P}(x_1)\mathbf{B}\mathbf{P}(x_2)\mathbf{B}\cdots\mathbf{B}\mathbf{P}(x_{T-1})\mathbf{B}\mathbf{P}(x_T)\mathbf{1},$$

where

$$\mathbf{P}(x)=\operatorname{diag}\big(\underbrace{q_{\Theta_1}(x),\ldots,q_{\Theta_1}(x)}_{m_1\text{ times}},\ldots,\underbrace{q_{\Theta_N}(x),\ldots,q_{\Theta_N}(x)}_{m_N\text{ times}}\big)$$

and $x_1,\ldots,x_T$ denote the observations. The above expression applies to the homogeneous case. In some applications, such as that described in Section 4, either $\mathbf{B}$ or $\mathbf{P}(x)$ can depend on t.

In order to numerically maximize the likelihood with respect to the parameters one needs to take care of some technical problems, such as avoiding numerical underflow. For details on how to deal with this problem see, e.g., Chapter 3 of Zucchini and MacDonald (2009). Second, if an unconstrained maximization algorithm, e.g. `nlm()` in **R**, is used, then it is necessary to reparameterize the model in terms of unconstrained parameters. However, since the parameters in the HMM formulation are exactly the same as those of the corresponding HSMM, there are therefore precisely as many constrained parameters in our model as there are constrained parameters in the HSMM model. If, say, a Poisson dwell-time distribution is to be fitted, then the parameter λ_k has to be positive. We would then numerically maximize the likelihood with respect to the unconstrained parameter, $\eta_k = \log\lambda_k$, and afterwards obtain the estimator of the constrained parameter by $\lambda_k = \exp\eta_k$. (Note that no further constraints concerning the entries of the t.p.m. $\mathbf{B}$ stem from the Poisson distribution since the entries are completely determined by the transition probabilities of the embedded Markov chain and the values $c_k(r)$, the latter being solely generated from λ_k, as pointed out in Section 2.) Hence, there is no additional difficulty regarding parameter constraints in the HMM formulation in comparison to the HSMM methodology. For more details on how to deal with parameter constraints see, e.g., Chapter 3 of Zucchini and MacDonald (2009). Finally, one should be aware that the maximization algorithm might converge to a local rather than the global maximum. The same problem can occur when applying the EM algorithm.

The computational effort required to evaluate the likelihood is linear in the number of observations, T, and quadratic in the number of HMM states, $M=\sum_{i=1}^{N}m_i$. Likelihood evaluation for HSMMs on the other hand is quadratic in T in the worst case if one uses standard HSMM methodology (see Guédon (2003), for reference). The computational effort due to the explicit modeling of the dwell-time distributions hence in the HSMM and in the HMM formulation is merely expressed in different terms; the inflation of the t.p.m. effectively does not lead to an increase in time complexity. Furthermore, the

Table 1
Minus the log likelihood and AIC for Models 1–3 with $q = 1$.

Model	Mllk	AIC
1	9468.44	18948.87
2	9445.84	18907.67
3	9423.27	**18862.55**

efficiency of the computation can be enhanced by taking advantage of the special structure of $\mathbf{B}$ – at most $N \cdot M$ out of the M^2 entries of $\mathbf{B}$ are non-zero – and the fact that $\mathbf{P}$ is a diagonal matrix.

We believe that, at least in applications where the modes of the dwell-time distributions are rather small, a simple generalization from geometric dwell-time distributions to dwell-time distributions with unstructured start and geometric tail is likely to provide sufficient additional flexibility in comparison to HMMs. (Examples include the applications in Sansom and Thomson (2001) as well as our application in Section 4.) Clearly, the HMM formulation will be advantageous when the number of states in the state aggregates is not large, i.e. when the modes of the dwell-time distributions are small. The method becomes less efficient as the size of the state aggregates increases since the matrices to be multiplied become larger and also require more memory space.

In simulations with $T > 2000$, Poisson distributed dwell-times and $M = 80$, a straightforward **R** implementation for estimating the parameters of the HMM was faster than the **R** function "hsmm" (Bulla et al., 2009), which is specifically designed to fit HSMMs, and which makes extensive use of **C** code.

4. Application: daily rainfall occurrence in Bulgaria

There exists a substantial literature on the stochastic modeling of daily precipitation (for reviews see Woolhiser, 1992 or Srikanthan and McMahon, 2001). Many of the proposed models are constructed from two submodels: the first describes rainfall occurrence (whether a particular day is wet or dry) and the second the rainfall amount on wet days. We restrict our attention to rainfall occurrence and fit a variety of models, in particular HMMs with non-geometric dwell-time distributions, to the binary sequence of dry and wet days. A similar application can be found in MacDonald and Zucchini (1997) in which Markov Chains and standard HMMs with geometric dwell-time distributions were considered.

Our main objective is to illustrate the application of the class of HMMs described in the preceding sections. As do Sansom and Thomson (2007) we include seasonality in the models. We show how this can be done in two different ways and thereby illustrate that it is easy to incorporate covariates, in the latent process as well as in the observed process, in the considered class of HMMs.

The data considered here comprise binary series of dry and wet days over a period of about 47 years at 5 sites in Bulgaria, namely Zlatograd, Plovdiv, Kurdjali, Ithiman and Ivailo. As is usually done with hydrological series in order to avoid the complication arising from having 366 days on leap years, we discarded observations for February 29. All computations were carried out using **R** (Ihaka and Gentleman, 1996).

We begin by considering a simple model for the daily rainfall series from Zlatograd (a town in the Rhodope mountains in Bulgaria), namely a two-state Markov chain in which the transition probabilities are allowed to vary seasonally (*Model* 1). Let S_t ($t = 1, 2, \dots, 16951$) be a Markov chain representing rainfall occurrence, where $S_t = 1$ if day t is dry, and $S_t = 2$ if day t is wet. The t.p.m. for *Model* 1 depends on t:

$$\mathbf{B}_t = \begin{pmatrix} \mathbf{P}(S_{t+1} = 1|S_t = 1) & \mathbf{P}(S_{t+1} = 2|S_t = 1) \\ \mathbf{P}(S_{t+1} = 1|S_t = 2) & \mathbf{P}(S_{t+1} = 2|S_t = 2) \end{pmatrix}.$$

To take care of seasonality the logit transforms of the diagonal elements of the t.p.m. are modeled as linear combinations of trigonometric functions:

$$\alpha_0 + \sum_{i=1}^{q} \left(\alpha_i \cos\left(\frac{2\pi \mathrm{i}(t-1)}{365}\right) + \beta_i \cos\left(\frac{2\pi \mathrm{i}(t-1)}{365}\right)\right), \tag{5}$$

where q is usually chosen using a model selection criterion, and is seldom greater than 2. We return to the choice of q later.

As a next step we fitted an HMM incorporating seasonality in the transition probabilities to the series. Now the Markov chain S_t is no longer taken to be an observation; it represents the unobserved state on day t. The observation on day t is regarded as a realization of a Bernoulli random variable whose parameter is determined by the state: The probability that day t is wet is π_1 if day t is in state 1, and is π_2 if day t is in state 2 (*Model* 2). *Model* 1 is the special case of *Model* 2 with $\pi_1 = 0$ and $\pi_2 = 1$.

An alternative way of introducing seasonality in the HMM is to assume that the entries of the t.p.m. are constant (i.e. do not depend on t) but that the Bernoulli parameters depend on t instead. Suppose now that the logit transforms of $\pi_1(t)$ and of $\pi_2(t)$ each have the form given in (5) (*Model* 3).

Table 1 gives minus the log likelihood and the AIC for *Models* 1–3 with $q = 1$. In terms of the AIC, the HMM with seasonality in the Bernoulli parameters provides the best fit by a substantial margin.

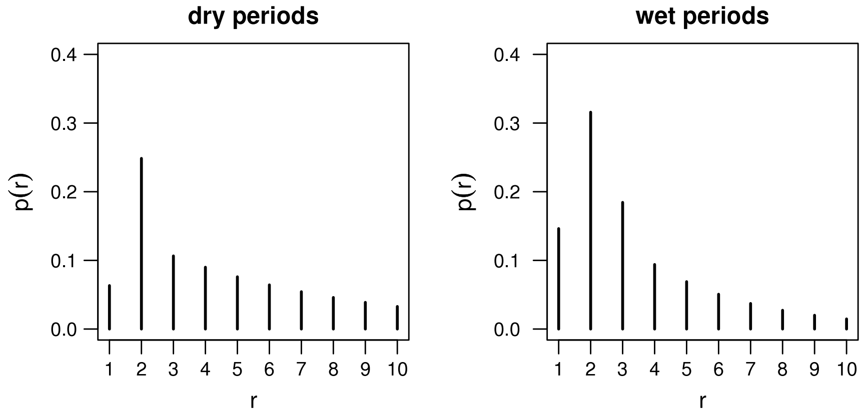

Fig. 2. Estimated p.m.f.'s for dry and wet periods.

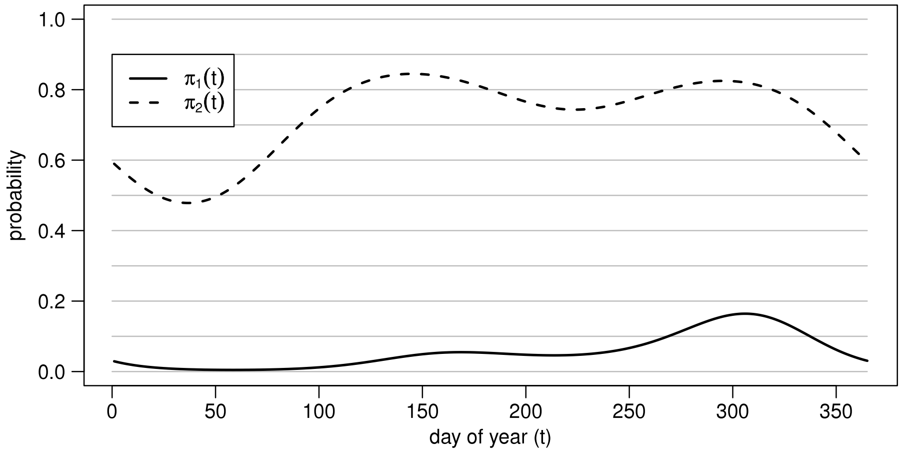

Fig. 3. Estimated Bernoulli parameter functions.

As a next step, we extended *Model* 3 by allowing the dwell-time distribution in each state to have an unstructured start and a geometric tail, as in Example 4. For the orders of the unstructured starts we tried the values 0 (Markovian), 1, 2 and 3 for both the state belonging to the dry periods and that belonging to the wet periods. We also tried different orders of seasonality, i.e. $q = 1, 2, 3$. Considering all possible combinations of orders of the unstructured starts and the seasonality another 47 different models emerge (one of the combinations is *Model* 3 above).

Out of these models the AIC would select the HMM which has a dwell-time distribution with an unstructured start of order 2 for the dry periods, unstructured start of order 3 for the wet periods and order of seasonality $q = 2$ in the Bernoulli parameters. The fitted transition probability matrix is given by

$$\left(\begin{array}{ccc|cccc} 0.00 & 0.94 & 0.00 & 0.06 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.73 & 0.27 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.85 & 0.15 & 0.00 & 0.00 & 0.00 \\ \hline 0.15 & 0.00 & 0.00 & 0.00 & 0.85 & 0.00 & 0.00 \\ 0.37 & 0.00 & 0.00 & 0.00 & 0.00 & 0.63 & 0.00 \\ 0.34 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.66 \\ 0.27 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \end{array}\right). \qquad (6)$$

Here the state aggregate $I_1 = \{1, 2, 3\}$ is associated with a low probability of precipitation (dry periods), and $I_2 = \{4, 5, 6, 7\}$ with a high probability of precipitation (wet periods). The upper left block matrix determines the dwell-time distribution in the dry periods, the lower right that in the wet periods. The estimated p.m.f.'s of the dwell-time distributions in the two state aggregates are displayed in Fig. 2. The deviation from the p.m.f. of a geometric distribution is evident, in particular the modal dwell time is not one in both cases. The estimated Bernoulli parameter functions $\pi_1(t)$ or $\pi_2(t)$, which have a period of 365, are displayed in Fig. 3. State aggregate I_1 can be regarded as the "dry" HSMM state; the probability of rain is generally low but peaks slightly in early November. In state aggregate I_2, the "wet" HSMM state, the probability of rain is generally high but drops (from 0.8 to 0.5) in January/Febrary. The stationary probabilities for state aggregates I_1 and I_2 are 0.63 and 0.37 respectively. In other words the system is in the "dry" state 63% of the time, and 37% in the "wet" state.

Table 2
Selected models for five sites.

Site	Order dry periods	Order wet periods	Order of seasonality
Zlatograd	2	3	2
Plovdiv	2	1	2
Kurdjali	2	1	3
Ihtiman	0	1	2
Ivailo	2	1	2

The AIC value for the chosen model is 18806.04, while it is 18832.25 for the HMM with geometric dwell-time distributions and the same order of seasonality ($q = 2$). The additional flexibility provided by allowing the dwell-time distributions to be non-geometric leads to a significantly improved fit.

The model selection exercise was repeated using the data from sites in Plovdiv, Kurdjali, Ithiman and Ivailo with all the above models. Table 2 lists the models that led to the lowest AIC in each case. Recall that order 0 means a Markovian state; the selected model for Ihtiman hence is a hidden hybrid HMM/HSMM (see Example 5). The selected models are not the same for all sites but then there is no reason to suppose that there exists a single model that is appropriate for all sites.

5. Summary

A restrictive feature of standard HMMs is that the state dwell-time distributions are necessarily geometric. We have shown how this restriction can be relaxed to allow for arbitrary dwell-time distributions while preserving the Markov property of the latent process. This is done by implementing an existing idea, the use of state aggregates, in a new way. The resulting models are closely related to HSMMs. A benefit of the proposed models is that one can use the well-established methodology that is available for HMMs. In particular one can easily incorporate covariates, in the state-dependent process as well as in the latent process. The models are easy to implement and add flexibility to the family of HMMs.

Acknowledgement

We thank the referees for valuable comments and suggestions.

Appendix

Proof of Proposition 1.

$$\begin{aligned}
a^{\star}_{ij} &= \Pr\left(S_{t+1} \in I_j | S_t \in I_i, S_{t+1} \notin I_i\right) \\
&= \frac{\Pr\left(S_{t+1} \in I_j, S_t \in I_i\right)}{\Pr\left(S_{t+1} \notin I_i, S_t \in I_i\right)} \quad \text{(since, by assumption, } i \neq j\text{)} \\
&= \frac{\sum_{k\in I_i} \Pr\left(S_{t+1} \in I_j | S_t = k\right) \Pr\left(S_t = k\right)}{\sum_{k\in I_i} \Pr\left(S_{t+1} \notin I_j | S_t = k\right) \Pr\left(S_t = k\right)} \\
&= \frac{\sum_{k\in I_i} a_{ij} c_i(k) \Pr\left(S_t = k\right)}{\sum_{k\in I_i} \sum_{l\neq i} a_{il} c_i(k) \Pr\left(S_t = k\right)} \\
&= a_{ij},
\end{aligned}$$

since $\sum_{l\neq i} a_{il} = 1$. □

Lemma 1. *Let $k \in \{1, \ldots, N\}$ and $n \in \mathbb{N}$. Then*

$$\prod_{i=1}^{n} (1 - c_k(i)) = 1 - F_k(n). \tag{7}$$

Proof. Straightforward using induction. □

Proof of Proposition 2. We consider the case $m_k \geq 2$ and $2 \leq r \leq m_k$ (the case $m_k = 1$ is trivial and the cases $r = 1$ and $r > m_k$ respectively follow with analogous arguments).

The structure of $\mathbf{B}$ is such that the dwell time in state aggregate I_k is of length r if and only if the state sequence successively runs through the states $i_k^-, i_k^- + 1, \ldots, i_k^- + r - 1$ and then immediately switches from state $i_k^- + r - 1$ to a different state aggregate.

If $F_k(r-1) < 1$ then, by (7), it follows that

$$\begin{aligned} p_k^\star(r) &= \prod_{i=1}^{r-1} \big(1 - c_k(i)\big) \sum_{1 \le s \le N,\, s \ne k} a_{ks} c_k(r) \\ &= \big(1 - F_k(r-1)\big) \frac{p_k(r)}{1 - F_k(r-1)} = p_k(r), \end{aligned}$$

and if $F_k(r-1) = 1$ then $p_k^\star(r) = 0 = p_k(r)$. $\square$

References

Altman, R., 2007. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. Journal of the American Statistical Association 102, 201–210.

Bartolucci, F., Lupparelli, M., Montanari, G.E., 2009. Latent Markov model for longitudinal binary data: an application to the performance evaluation of nursing homes. Annals of Applied Statistics 3 (2), 611–636.

Bulla, J., Bulla, I., Nenadić, O., 2009. An R package for analyzing hidden semi-Markov models. Computational Statistics and Data Analysis 54 (3), 611–619.

Cappé, O., Moulines, E., Rydén, T., 2005. Inference in Hidden Markov Models. Springer, New York.

Chaubert-Pereira, F., Guédon, Y., Lavergne, C., Trottier, C., 2010. Markov and semi-Markov switching linear mixed models used to identify forest tree growth components. Biometrics doi:10.1111/j.1541-0420.2009.01338.x. (published online).

Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. Chapman and Hall, London.

Ephraim, Y., Merhav, N., 2002. Hidden Markov processes. IEEE Transactions on Information Theory 48, 1518–1569.

Ferguson, J.D., (1980). Variable duration models for speech. In: Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech, Princeton, New Jersey, pp. 143–179.

Guédon, Y., 2003. Estimating hidden semi-Markov chains from discrete sequences. Journal of Computational and Graphical Statistics 12 (3), 604–639.

Guédon, Y., 2005. Hidden hybrid Markov/semi-Markov chains. Computational Statistics and Data Analysis 49 (3), 663–688.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

Johnson, M.T., 2005. Capacity and complexity of HMM duration modeling techniques. IEEE Signal Processing Letters 12 (5), 407–410.

Kleinrock, L., 1975. Queueing Systems, Vol. 1: Theory. Wiley, New York.

Kulkarni, V.G., 1995. Modeling and Analysis of Stochastic Systems. Chapman and Hall, London.

MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and other Models for Discrete-valued Time Series. Chapman and Hall, London.

Russell, M.J., Cook, A.E., 1987. Experimental evaluation of duration modelling techniques for automatic speech recognition. IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2376–2379.

Sansom, J., Thomson, P., 2001. Fitting hidden semi-Markov models to breakpoint rainfall data. Journal of Applied Probability 38A, 142–157.

Sansom, J., Thomson, P., 2007. On rainfall seasonality using a hidden semi-Markov model. Journal of Geophysical Research 112, D15105.

Srikanthan, R., McMahon, T.A., 2001. Stochastic generation of annual, monthly and daily climate data: a review. Hydrology and Earth System Sciences 5, 653–670.

Woolhiser, D.A., 1992. Modeling daily precipitation — progress and problems. In: Walden, A.T., Guttorp, P. (Eds.), Statistics in the Environmental and Earth Sciences. Edward Arnold, pp. 71–89.

Yu, S.-Z., Kobayashi, H., 2003. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. IEEE Signal Processing Letters 10 (1), 11–14.

Zucchini, W., MacDonald, I.L., 2009. Hidden Markov Models for Time Series: An Introduction Using R. Chapman & Hall, London.