CHAPTER 8

Extensions of the basic hidden Markov model

> A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives.
>
> McCullagh and Nelder
> *Generalized Linear Models* (1989, p. 8)

8.1 Introduction

A notable advantage of HMMs is the ease with which the basic model can be modified or generalized, in several different directions, in order to provide flexible models for a wide range of types of observation.

We begin this chapter by describing the use in the basic HMM of univariate state-dependent distributions other than the Poisson (Section 8.2), and then show how the basic HMM can be extended in other ways. The first such extension (Section 8.3) adds flexibility by generalizing the underlying parameter process; the assumption that the parameter process is a first-order Markov chain is relaxed by allowing it to be a second-order Markov chain. This extension can be applied not only to the basic model but also to most of the other models to be discussed.

We then illustrate how the basic model can be generalized to construct HMMs for a number of different and more complex types of observation, including the following.

- Series of multinomial-like observations (Section 8.4.1): An example of a multinomial-like series would be daily sales of a particular item categorized into the four consumer categories: adult female, adult male, juvenile female, juvenile male.
- Categorical series (Section 8.4.2): An important special case of the multinomial-like series is that in which there is exactly one observation at each time, classified into one of q possible mutually exclusive categories: that is, a categorical time series. An example is an hourly series of wind directions in the form of the conventional 16 categories, i.e. the 16 points of the compass.
- Other multivariate series (Section 8.4.3): An example of a bivariate

discrete-valued series is the number of sales of each of two related items. A key feature of multivariate time series is that, in addition to serial dependence within each series, there may be dependence across the series.

- Series that depend on covariates (Section 8.5): Many, if not most, time series studied in practice exhibit time trend, seasonal variation, or both. Examples include monthly sales of items, daily number of shares traded, insurance claims received, and so on. One can regard such series as depending on time as a covariate. In some cases covariates other than time are relevant. For example, one might wish to model the number of sales of an item as a function of the price, advertising expenditure and sales promotions, and also to allow for trend and seasonal fluctuations.
- Models with additional dependencies (Section 8.6): One way to produce useful generalizations of the basic HMM as depicted in Figure 2.2 is to add extra dependencies between some of the random variables that make up the model. There are several different ways to do this. For instance, if one suspects that the continuous observations X_{t-1} and X_t are not conditionally independent given the Markov chain, one might wish to use a model that switches between AR(1) processes according to a Markov chain, that is, a Markov-switching AR(1).

8.2 HMMs with general univariate state-dependent distribution

In this book we have introduced the basic (univariate) HMM by concentrating on Poisson state-dependent distributions. One may, however, use any distribution — discrete, continuous, or a mixture of the two — as the state-dependent distribution; in fact there is nothing preventing one from using a different family of distributions for each state. One simply redefines the diagonal matrices containing the state-dependent probabilities, and in the estimation process takes note of whatever constraints the state-dependent parameters must observe, either by transforming the constraints away or by explicitly constrained optimization.

In what follows we describe HMMs with various univariate state-dependent distributions without going into much detail.

- HMMs for unbounded counts

  The Poisson distribution is the canonical model for unbounded counts. However, a popular alternative, especially for overdispersed data, is the negative binomial distribution. One can therefore consider replacing the Poisson state-dependent distribution in a Poisson–HMM by

the negative binomial, which is given, for all nonnegative integers x, by the probability function

$$p_i(x) = \frac{\Gamma\left(x + \frac{1}{\eta_i}\right)}{\Gamma\left(\frac{1}{\eta_i}\right)\Gamma(x+1)} \left(\frac{1}{1+\eta_i\mu_i}\right)^{\frac{1}{\eta_i}} \left(\frac{\eta_i\mu_i}{1+\eta_i\mu_i}\right)^{x},$$

where the parameters μ_i (the mean) and η_i are positive. (But note that this is only one of several possible parametrizations of the negative binomial.) A negative binomial–HMM may sensibly be used if even a Poisson–HMM seems unable to accommodate the observed overdispersion. Conceivable examples for the application of Poisson– or negative binomial–HMMs include series of counts of stoppages or breakdowns of technical equipment, earthquakes, sales, insurance claims, accidents reported, defective items and stock trades.

- HMMs for binary data

  The Bernoulli–HMM for binary time series is the simplest HMM. Its state-dependent probabilities for the two possible outcomes are, for some probabilities π_i, just

  $$\begin{aligned} p_i(0) &= \Pr(X_t = 0 \mid C_t = i) = 1 - \pi_i \qquad \text{(failure)}, \\ p_i(1) &= \Pr(X_t = 1 \mid C_t = i) = \pi_i \qquad \text{(success)}. \end{aligned}$$

  An example of a Bernoulli–HMM appears in Section 2.3.1. Possible applications of Bernoulli–HMMs are to daily rainfall occurrence (rain or no rain), daily trading of a share (traded or not traded), and consecutive departures of aeroplanes from an airport (on time, not on time).

- HMMs for bounded counts

  Binomial–HMMs may be used to model series of bounded counts. The state-dependent binomial probabilities are given by

  $${}_tp_i(x_t) = \binom{n_t}{x_t} \pi_i^{x_t} (1-\pi_i)^{n_t - x_t},$$

  where n_t is the number of trials at time t and x_t the number of successes. (We use the prefix t as far as possible to indicate time-dependence.)

  Possible examples for series of bounded counts that may be described by a binomial–HMM are series of:

  – purchasing preferences, e.g. n_t = number of purchases of all brands on day t, x_t = number of purchases of brand A on day t;
  – sales of newspapers or magazines, e.g. n_t = number available on day t, x_t = number purchased on day t.

Notice, however, that there is a complication when one computes the forecast distribution of a binomial–HMM. Either n_{T+h}, the number of trials at time $T + h$, must be known, or one has to fit a separate model to forecast n_{T+h}. Alternatively, by setting $n_{T+h} = 1$ one can simply compute the forecast distribution of the 'success proportion'.

- HMMs for continuous-valued series

  So far, we have primarily considered HMMs with discrete-valued state-dependent component distributions. However, we have also mentioned that it is possible to use continuous-valued component distributions. One simply has to replace the probability functions by the corresponding state-dependent probability density functions.

  Important state-dependent distributions for continuous-valued time series are the exponential, Gamma and normal distributions.

  Normal–HMMs are sometimes used for modelling share returns series because the observed kurtosis of most such series is greater than 3, the kurtosis of a normal distribution. See Section 13.2 for a multivariate model for returns on four shares. Note that the (continuous) likelihood of a normal–HMM is unbounded; it is possible to increase the likelihood without bound by fixing a state-dependent mean μ_i at one of the observations and letting the corresponding variance σ_i approach zero. In practice, this may or may not lead to problems in parameter estimation. If it does, using the discrete likelihood is advisable; see Section 1.2.3.

8.3 HMMs based on a second-order Markov chain

One generalization of the basic HMM is that which results if one replaces the underlying first-order Markov chain in the basic model (or in the extensions to follow later in this chapter) by a higher-order chain. Here we describe only the model that has as parameter process a stationary second-order chain. Such a second-order chain is characterized by the transition probabilities

$$\gamma(i, j, k) = \mathrm{P}(C_t = k \mid C_{t-1} = j, C_{t-2} = i),$$

and has stationary bivariate distribution $u(j, k) = \mathrm{P}(C_{t-1} = j, C_t = k)$. (See Section 1.3.6 for a more detailed description of higher-order Markov chains.)

We mention here two important aspects of such a second-order HMM, which is depicted in Figure 8.1. The first is that it is possible to evaluate the likelihood of a second-order HMM in very similar fashion to that of the basic model; the computational effort is in this case cubic in m, the number of states, and, as before, linear in T, the number of observations.

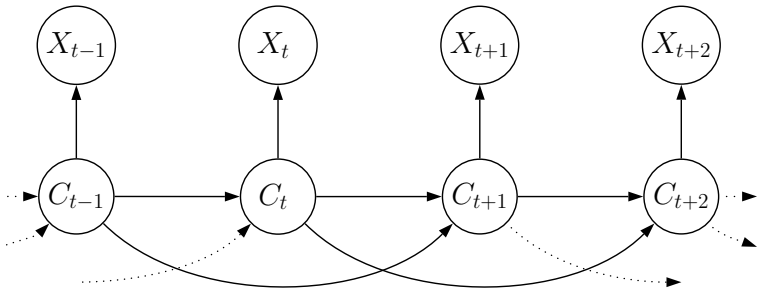

Figure 8.1 *Directed graph of second-order HMM.*

(See Exercise 3.) This then enables one to estimate the parameters by direct maximization of the likelihood, and also to compute the forecast distributions.

The second aspect is the number of free parameters of the (second-order) Markov chain component of the model. In general this number is $m^2(m-1)$, which rapidly becomes prohibitively large as m increases. This overparametrization can be circumvented by using some restricted subclass of second-order Markov chain models, for example those of Pegram (1980) or those of Raftery (1985a). Such models are necessarily less flexible than the general class of second-order Markov chains. They maintain the second-order structure of the chain, but trade some flexibility in return for a reduction in the number of parameters.

Having pointed out that it is possible to increase the order of the Markov chain in an HMM from one to two (or higher), in what follows we shall restrict our attention almost entirely to the simpler case of a first-order Markov chain. The exception is Section 10.2.2, where we present (among other models) an example of a second-order HMM for a binary time series. We note in passing that the complications caused by the constraints on the parameters in a Raftery model have been discussed by Schimert (1992) and by Raftery and Tavaré (1994), who describe how one can reduce the number of constraints.

8.4 HMMs for multivariate series

8.4.1 Series of multinomial-like observations

One of the variations of the basic model mentioned in Section 8.2 is the binomial–HMM, in which, conditional on the Markov chain, the observations $\{x_t : t = 1, \ldots, T\}$ are the numbers of successes in n_1, n_2, ...,

n_T independent Bernoulli trials. The m-state model has m probabilities of success π_i, one for each state i.

A **multinomial–HMM** is the obvious generalization thereof to the situation in which there are $q \geq 2$, rather than two, mutually exclusive and exhaustive possible outcomes to each trial. The observations are then q series of counts, $\{x_{tj} : t = 1, \ldots, T,\ j = 1, \ldots, q\}$ with $x_{t1} + x_{t2} + \ldots + x_{tq} = n_t$ where n_t is the (known) number of trials at time t. Thus for example x_{23} represents the number of outcomes at time $t = 2$ that were of type 3. The counts x_{tj} at time t can be combined in a vector $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tq})$.

We shall suppose that, conditional on the Markov chain $\mathbf{C}^{(T)}$, the T random vectors $\{\mathbf{X}_t = (X_{t1}, X_{t2}, \ldots, X_{tq}) : t = 1, \ldots, T\}$ are mutually independent.

The parameters of the model are as follows. As in the basic model, the matrix $\mathbf{\Gamma}$ has $m(m-1)$ free parameters. With each of the m states of the Markov chain there is associated a multinomial distribution with parameters n_t (known) and q unknown probabilities which, for state i, we shall denote by $\pi_{i1}, \pi_{i2}, \ldots, \pi_{iq}$. These probabilities are constrained by $\sum_{j=1}^{q} \pi_{ij} = 1$ for each state i. This component of the model therefore has $m(q-1)$ free parameters, and the entire model has $m^2 - m + m(q-1) = m^2 + m(q-2)$.

The likelihood of observations $\mathbf{x}_1, \ldots, \mathbf{x}_T$ from a general multinomial–HMM differs little from that of a binomial–HMM; the only difference is that the binomial probabilities

$$ {}_tp_i(x_t) = \binom{n_t}{x_t} \pi_i^{x_t} (1 - \pi_i)^{n_t - x_t} $$

are replaced by multinomial probabilities

$$ {}_tp_i(\mathbf{x}_t) = \mathrm{P}(\mathbf{X}_t = \mathbf{x}_t \mid C_t = i) = \binom{n_t}{x_{t1}, x_{t2}, \ldots, x_{tq}} \pi_{i1}^{x_{t1}} \pi_{i2}^{x_{t2}} \cdots \pi_{iq}^{x_{tq}}. $$

Note that these probabilities are indexed by the time t because the number of trials n_t is permitted to be time-dependent. We assume that the mq state-dependent probabilities π_{ij} are constant over time, but that is an assumption that can if necessary be relaxed.

The likelihood is given by

$$ L_T = \boldsymbol{\delta}\, {}_1\mathbf{P}(\mathbf{x}_1) \mathbf{\Gamma}\, {}_2\mathbf{P}(\mathbf{x}_2) \cdots \mathbf{\Gamma}\, {}_T\mathbf{P}(\mathbf{x}_T) \mathbf{1}', $$

where ${}_t\mathbf{P}(\mathbf{x}_t) = \mathrm{diag}\left({}_tp_1(\mathbf{x}_t), \ldots, {}_tp_m(\mathbf{x}_t)\right)$. Parameters can then be estimated by maximizing the likelihood as a function of $m(q-1)$ of the 'success probabilities', e.g. π_{ij} for $j = 1, 2, \ldots, q-1$, and of the $m^2 - m$ off-diagonal transition probabilities. If one does so, one must observe not only the usual 'generalized upper bound' constraints $\sum_{j \neq i} \gamma_{ij} \leq 1$ on the

transition probabilities, but also the m similar constraints $\sum_{j=1}^{q-1} \pi_{ij} \leq 1$ on the probabilities π_{ij}, one constraint for each state i — as well as, of course, the lower bound of 0 on all these probabilities.

Once the parameters have been estimated, these can be used to estimate various forecast distributions. There is, however, the same complication to such forecasts as described already in the case of the binomial–HMM. We have assumed that n_t, the number of trials at time t, is known. This number, being the sum of the q observed counts at time t, is certainly known at times $t = 1, 2, \ldots, T$. But in order to compute the one-step-ahead forecast distribution, one needs to know n_{T+1}, the number of trials that will take place at time $T+1$. This will be known in some applications, for instance when the number of trials is prescribed by a sampling scheme. But there are also applications in which n_{T+1} is a random variable whose value remains unknown until time $T+1$. For the latter it is not possible to compute the forecast distribution of the counts at time $T+1$. As before, by setting $n_{T+1} = 1$ it is possible to compute the forecast distribution of the count-proportions.

An alternative approach is to fit a separate model to the series $\{n_t\}$, to use that model to compute the forecast distribution of n_{T+1} and then, finally, to use that to compute the required forecast distribution for the counts of the multinomial–HMM.

8.4.2 A model for categorical series

A simple but important special case of the multinomial–HMM is that in which $n_t = 1$ for all t. This provides a model for categorical series, e.g. DNA base sequences or amino-acid sequences, in which there is exactly one symbol at each position in the sequence: one of A, C, G, T in the former example, one of 20 amino-acids in the latter. In this case the state-dependent probabilities ${}_tp_i(\mathbf{x})$ and the matrix expression for the likelihood simplify somewhat.

Because n_t is constant, the prefix t is no longer necessary, and because $\sum_{k=1}^{q} x_{tk} = 1$, the q-vector $\mathbf{x}_t$ has one entry equal to 1 and the others equal to zero. It follows that, if

$$\mathbf{x} = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{q-j}),$$

then $p_i(\mathbf{x}) = \pi_{ij}$ and

$$\mathbf{P}(\mathbf{x}) = \operatorname{diag}(\pi_{1j}, \ldots, \pi_{mj}).$$

For convenience we denote $\mathbf{P}(\mathbf{x})$, for $\mathbf{x} = (0, \ldots, 0, 1, 0, \ldots, 0)$ as above, by $\mathbf{\Pi}(j)$. In this notation the likelihood of observing categories j_1, j_2,

$\ldots$, j_T at times 1, 2, $\ldots$, T is given by

$$L_T = \boldsymbol{\delta}\mathbf{\Pi}(j_1)\mathbf{\Gamma}\mathbf{\Pi}(j_2)\mathbf{\Gamma}\cdots\mathbf{\Pi}(j_T)\mathbf{1}'.$$

If we assume the Markov chain is stationary, it is implied, for instance, that the probability of observing category l at time $t+1$, given that category k is observed at time t, is

$$\frac{\boldsymbol{\delta}\mathbf{\Pi}(k)\mathbf{\Gamma}\mathbf{\Pi}(l)\mathbf{1}'}{\boldsymbol{\delta}\mathbf{\Pi}(k)\mathbf{1}'}, \tag{8.1}$$

and similarly, that of observing l at time $t+1$, given k at time t and j at time $t-1$, is

$$\frac{\boldsymbol{\delta}\mathbf{\Pi}(j)\mathbf{\Gamma}\mathbf{\Pi}(k)\mathbf{\Gamma}\mathbf{\Pi}(l)\mathbf{1}'}{\boldsymbol{\delta}\mathbf{\Pi}(j)\mathbf{\Gamma}\mathbf{\Pi}(k)\mathbf{1}'}.$$

The above two expressions can be used to compute forecast distributions. An example of a forecast using Equation (8.1) is given in Section 12.2.2 (see p. 172).

8.4.3 Other multivariate models

The series of multinomial-like counts discussed in the last section are, of course, examples of multivariate series, but with a specific structure. In this section we illustrate how it is possible to develop HMMs for different and more complex types of multivariate series.

Consider q time series $\{(X_{t1}, X_{t2}, \ldots, X_{tq}) : t = 1, \ldots, T\}$ which we shall represent compactly as $\{\mathbf{X}_t, : t = 1, \ldots, T\}$. As we did for the basic HMM, we shall assume that, conditional on $\mathbf{C}^{(T)} = \{C_t : t = 1, \ldots, T\}$, the above random vectors are mutually independent. We shall refer to this property as **longitudinal conditional independence** in order to distinguish it from a different conditional independence that we shall describe later.

To specify an HMM for such a series it is necessary to postulate a model for the distribution of the random vector $\mathbf{X}_t$ in each of the m states of the parameter process. That is, one requires the following probabilities to be specified for $t = 1, 2, \ldots, T$, $i = 1, 2, \ldots, m$, and all relevant x:

$$_tp_i(\mathbf{x}) = \Pr(\mathbf{X}_t = \mathbf{x} \mid C_t = i).$$

(For generality, we keep the time index t here, i.e. we allow the state-dependent probabilities to change over time.) In the case of multinomial–HMMs these probabilities are supplied by m multinomial distributions.

We note that it is not required that each of the q component series have a distribution of the same type. For example, in the bivariate model discussed in Section 11.5, the state-dependent distributions of X_{t1} are gamma distributions, and those of X_{t2} von Mises distributions; the first

component is linear-valued and the second circular-valued. Secondly, it is not assumed that the m state-dependent distributions of any one series belong to the same family of distributions. In principle one could use a gamma distribution in state 1, an extreme-value distribution in state 2, and so on. However, we have not yet encountered applications in which this feature could be usefully exploited.

What is necessary is to specify models for m joint distributions, a task that can be anything but trivial. For example there is no single bivariate Poisson distribution; different versions are available and they have different properties. One has to select a version that is appropriate in the context of the particular application being investigated. (In contrast, one can reasonably speak of *the* bivariate normal distribution because, for many or most practical purposes, there is only one.)

Once the required joint distributions have been selected, i.e. once one has specified the state-dependent probabilities ${}_tp_i(\mathbf{x}_t)$, the likelihood of a general multivariate HMM is easy to write down. It has the same form as that of the basic model, namely

$$L_T = \boldsymbol{\delta}\, {}_1\mathbf{P}(\mathbf{x}_1)\boldsymbol{\Gamma}\, {}_2\mathbf{P}(\mathbf{x}_2)\cdots\boldsymbol{\Gamma}\, {}_T\mathbf{P}(\mathbf{x}_T)\mathbf{1}',$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are the observations and, as before,

$${}_t\mathbf{P}(\mathbf{x}_t) = \text{diag}\left({}_tp_1(\mathbf{x}_t), \ldots, {}_tp_m(\mathbf{x}_t)\right).$$

The above expression for the likelihood also holds if some of the series are continuous-valued, provided that where necessary probabilities are interpreted as densities.

The task of finding suitable joint distributions is greatly simplified if one can assume **contemporaneous conditional independence**. We illustrate the meaning of this term by means of the multisite precipitation model discussed by Zucchini and Guttorp (1991). In their work there are five binary time series representing the presence or absence of rain at each of five sites which are regarded as being linked by a common weather process $\{C_t\}$. There the random variables X_{tj} are binary. Let ${}_t\pi_{ij}$ be defined as

$${}_t\pi_{ij} = \Pr(X_{tj} = 1 \mid C_t = i) = 1 - \Pr(X_{tj} = 0 \mid C_t = i).$$

The assumption of contemporaneous conditional independence is that the state-dependent joint probability ${}_tp_i(\mathbf{x}_t)$ is just the product of the corresponding marginal probabilities:

$${}_tp_i(\mathbf{x}_t) = \prod_{j=1}^{q} {}_t\pi_{ij}^{x_{tj}}(1 - {}_t\pi_{ij})^{1-x_{tj}}. \tag{8.2}$$

Thus, for example, given weather state i, the probability that on day t it will rain at sites 1, 2, and 4, but not at sites 3 and 5, is the product

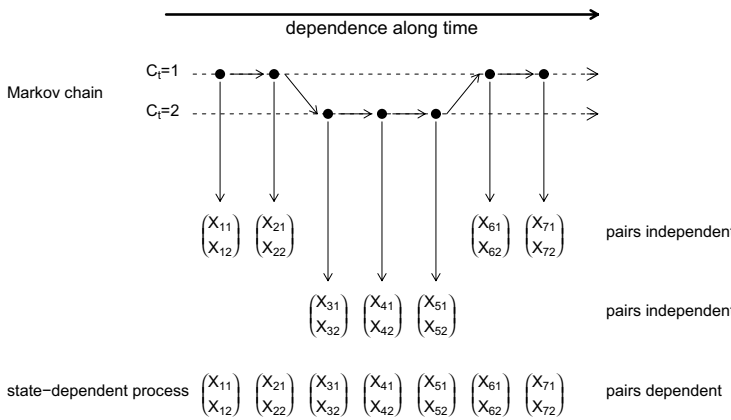

Figure 8.2 *Contemporaneous conditional independence.*

of the (marginal) probabilities that these events occur: ${}_t\pi_{i1}\,{}_t\pi_{i2}(1 - {}_t\pi_{i3})\,{}_t\pi_{i4}(1 - {}_t\pi_{i5})$.

For general multivariate HMMs that are contemporaneously conditionally independent, the state-dependent probabilities are given by a product of the corresponding q marginal probabilities:

$$ {}_tp_i(\mathbf{x}_t) = \prod_{j=1}^{q} \Pr(X_{tj} = x_{tj} \mid C_t = i). $$

We wish to emphasize that the above two conditional independence assumptions, namely longitudinal conditional independence and contemporaneous conditional independence, do not imply that

- the individual component series are serially independent; or that
- the component series are mutually independent.

The parameter process, namely the Markov chain, induces both serial dependence and cross-dependence in the component series, even when these are assumed to have both of the above conditional independence properties. This is illustrated in Figure 8.2.

Details of the serial- and cross-correlation functions of these models are given in Section 3.4 of MacDonald and Zucchini (1997), as are other general classes of models for multivariate HMMs, such as models with time lags and multivariate models in which some of the components are discrete and others continuous. See also Exercise 8.

Multivariate HMMs (with continuous state-dependent distributions)

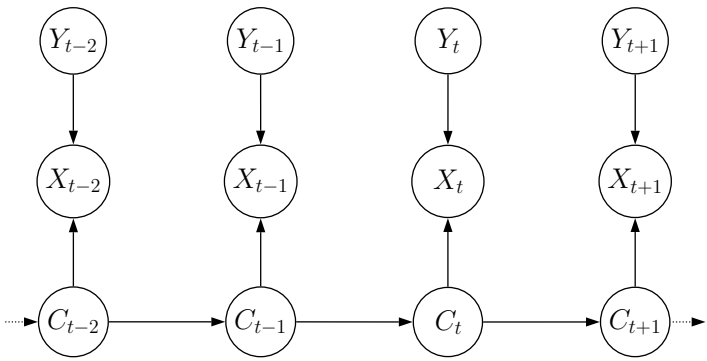

Figure 8.3 *Structure of HMM with covariates Y_t in the state-dependent probabilities.*

might for instance be used for modelling multivariate financial time series. For example, one could fit a two-state multivariate normal HMM to a multivariate series of daily returns on a number of shares where, as in the univariate case, the states of the Markov chain might correspond to calm and turbulent phases of the stock market; see Section 13.2 for such a model.

8.5 Series that depend on covariates

We now discuss two ways in which covariates can be introduced into HMMs: via the state-dependent probabilities, and via the transition probabilities of the Markov chain.

8.5.1 Covariates in the state-dependent distributions

HMMs can be modified to allow for the influence of covariates by postulating dependence of the state-dependent probabilities ${}_tp_i(x_t)$ on those covariates, as demonstrated in Figure 8.3. This opens the way for such models to incorporate time trend and seasonality, for instance.

Here we take $\{C_t\}$ to be the usual Markov chain, and suppose, in the case of Poisson–HMMs, that the conditional mean ${}_t\lambda_i = \mathrm{E}(X_t \mid C_t = i)$ depends on the (row) vector $\mathbf{y}_t$ of q covariates, for instance as follows:

$$\log {}_t\lambda_i = \boldsymbol{\beta}_i \mathbf{y}_t'.$$

In the case of binomial–HMMs, the corresponding assumption is that

$$\text{logit}\, {}_tp_i = \boldsymbol{\beta}_i \mathbf{y}_t'.$$

The elements of $\mathbf{y}_t$ could include a constant, time (t), sinusoidal components expressing seasonality (for example $\cos(2\pi t/r)$ and $\sin(2\pi t/r)$ for some positive integer r), and any other relevant covariates. For example, a binomial–HMM with

$$\text{logit}\, {}_tp_i = \beta_{i1} + \beta_{i2}t + \beta_{i3}\cos(2\pi t/r) + \beta_{i4}\sin(2\pi t/r) + \beta_{i5}z_t + \beta_{i6}w_t$$

allows for a (logit-linear) time trend, r-period seasonality and the influence of covariates z_t and w_t, in the state-dependent 'success probabilities' ${}_tp_i$. Additional sine-cosine pairs can if necessary be included, to model more complex seasonal patterns. Similar models for the log of the conditional mean ${}_t\lambda_i$ are possible in the Poisson–HMM case. Clearly link functions other than the canonical ones used here could instead be used. The expression for the likelihood of T consecutive observations $x_1, \ldots, x_T$ for such a model involving covariates is similar to that of the basic model:

$$L_T = \boldsymbol{\delta}\, {}_1\mathbf{P}(x_1, y_1)\boldsymbol{\Gamma}\, {}_2\mathbf{P}(x_2, y_2) \cdots \boldsymbol{\Gamma}\, {}_T\mathbf{P}(x_T, y_T)\mathbf{1}',$$

the only difference being the allowance for covariates y_t in the state-dependent probabilities ${}_tp_i(x_t, y_t)$ and in the corresponding matrices

$${}_t\mathbf{P}(x_t, y_t) = \text{diag}({}_tp_1(x_t, y_t), \ldots, {}_tp_m(x_t, y_t)).$$

Examples of models that incorporate a time trend can be found in Sections 13.1.1, 14.2 and 15.2. In Section 14.3 there are models incorporating both a time trend and a seasonal component.

It is worth noting that the binomial– and Poisson–HMMs which allow for covariates in this way provide important generalizations of logistic regression and Poisson regression respectively, generalizations that drop the independence assumption of such regression models and allow serial dependence.

8.5.2 Covariates in the transition probabilities

An alternative way of modelling time trend and seasonality in HMMs is to drop the assumption that the Markov chain is homogeneous, and assume instead that the transition probabilities are functions of time, denoted for two states as follows:

$${}_t\boldsymbol{\Gamma} = \begin{pmatrix} {}_t\gamma_{11} & {}_t\gamma_{12} \\ {}_t\gamma_{21} & {}_t\gamma_{22} \end{pmatrix}.$$

More generally, the transition probabilities can be modelled as depending on one or more covariates, not necessarily time but any variables considered relevant.

Incorporation of covariates into the Markov chain is not as straightforward as incorporating them into the state-dependent probabilities. One reason why it could nevertheless be worthwhile is that the resulting Markov chain may have a useful substantive interpretation, e.g. as a weather process which is itself complex but determines rainfall probabilities at several sites in fairly simple fashion. We illustrate one way in which it is possible to modify the transition probabilities of the Markov chain in order to represent time trend and seasonality.

Consider a model based on a two-state Markov chain $\{C_t\}$ with

$$\Pr(C_t = 2 \mid C_{t-1} = 1) = {}_t\gamma_{12}, \quad \Pr(C_t = 1 \mid C_{t-1} = 2) = {}_t\gamma_{21}\,,$$

and, for $i = 1, 2$,

$$\operatorname{logit} {}_t\gamma_{i,3-i} = \boldsymbol{\beta}_i \mathbf{y}_t'.$$

For example, a model incorporating r-period seasonality is that with

$$\operatorname{logit} {}_t\gamma_{i,3-i} = \beta_{i1} + \beta_{i2}\cos(2\pi t/r) + \beta_{i3}\sin(2\pi t/r).$$

In general the above assumption on $\operatorname{logit} {}_t\gamma_{i,3-i}$ implies that the transition probability matrix, for transitions between times $t-1$ and t, is given by

$${}_t\mathbf{\Gamma} = \begin{pmatrix} \dfrac{1}{1+\exp(\boldsymbol{\beta}_1\mathbf{y}_t')} & \dfrac{\exp(\boldsymbol{\beta}_1\mathbf{y}_t')}{1+\exp(\boldsymbol{\beta}_1\mathbf{y}_t')} \\ \dfrac{\exp(\boldsymbol{\beta}_2\mathbf{y}_t')}{1+\exp(\boldsymbol{\beta}_2\mathbf{y}_t')} & \dfrac{1}{1+\exp(\boldsymbol{\beta}_2\mathbf{y}_t')} \end{pmatrix}.$$

Extension of this model to the case $m > 2$ presents some difficulties, but they are not insuperable.

One important difference between the class of models proposed here and other HMMs (and a consequence of the nonhomogeneity of the Markov chain) is that we cannot always assume that there is a stationary distribution for the Markov chain. This problem arises when one or more of the covariates are functions of time, as in models with trend or seasonality. If necessary we therefore assume instead some initial distribution $\boldsymbol{\delta}$, i.e. a distribution for C_1.

A very general class of models in which the Markov chain is nonhomogeneous and which allows for the influence of covariates is that of Hughes (1993). This model, and additional details relating to the models outlined above, are discussed in Chapter 3 of MacDonald and Zucchini (1997).

8.6 Models with additional dependencies

In Section 8.3 we described a class of models which have dependencies in addition to those found in the basic HMM as depicted by Figure 2.2 on p. 30: second-order HMMs. In that case the additional dependencies are entirely at latent process level. Here we briefly describe three further classes of models with additional dependencies.

In the basic model of Figure 2.2 there are no edges directly connecting earlier observations to X_t; the only dependence between observations arises from the latent process $\{C_t\}$. There may well be applications, however, where extra dependencies at observation level are suspected and should be allowed for in the model. The additional computational features of such models are that the 'state-dependent' probabilities needed for the likelihood computation depend on previous observations as well as on the current state, and that the likelihood maximized is conveniently taken to be that which is conditional on the first few observations. Figure 8.4 depicts two models with such extra dependencies at observation level. In Section 14.3 we present some examples of models with additional dependencies at observation level; the likelihood evaluation and maximization proceed with little additional complication.

Some models with additional dependencies at observation level that have appeared in the literature are the double-chain Markov model of Berchtold (1999), the M1–Mk models of Nicolas *et al.* (2002) for DNA sequences, those of Boys and Henderson (2004), and hidden Markov AR(k) models, usually termed Markov-switching autoregressions.

There is another type of extra dependency which may be useful. In the basic model, the distribution of an observation depends on the current state only. It is not difficult, however, to make the minor generalization needed for that distribution to depend also on the previous state. Figure 8.5 depicts the resulting model. If we define the $m \times m$ matrix $\mathbf{Q}(x)$ to have as its (i, j) element the product $\gamma_{ij} \Pr(X_t = x \mid C_{t-1} = i, C_t = j)$, and denote by $\boldsymbol{\delta}$ the distribution of C_1, the likelihood is given by

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{Q}(x_2)\mathbf{Q}(x_3)\cdots\mathbf{Q}(x_T)\mathbf{1}'. \tag{8.3}$$

If instead we denote by $\boldsymbol{\delta}$ the distribution of C_0 (not C_1), the result is somewhat neater:

$$L_T = \boldsymbol{\delta}\mathbf{Q}(x_1)\mathbf{Q}(x_2)\cdots\mathbf{Q}(x_T)\mathbf{1}'. \tag{8.4}$$

Extra dependencies of yet another kind appear in Section 13.3.3 and Chapter 16, which present respectively a discrete state-space stochastic volatility model with leverage (see Figure 13.4), and a model for animal behaviour which incorporates feedback from observation level to motivational state (see Figure 16.2). In both of these applications the extra dependencies are from observation level to the latent process.

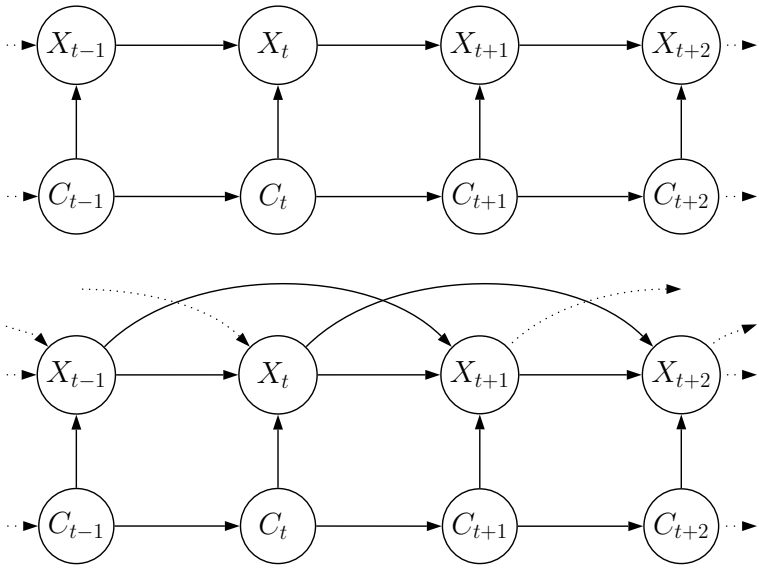

Figure 8.4 *Models with additional dependencies at observation level; the upper graph represents (e.g.) a Markov-switching AR(1) model, and the lower one a Markov-switching AR(2).*

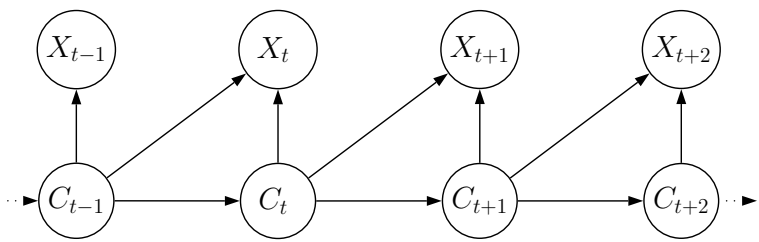

Figure 8.5 *Additional dependencies from latent process to observation level.*

Exercises

1. Give an expression for the likelihood function for each of the following HMMs. The list below refers to the state-dependent distributions.

   (a) Geometric with parameters $\theta_i \in (0,\ 1),\ i = 1, 2, \ldots, m$.

   (b) Exponential with parameters $\lambda_i \geq 0,\ i = 1, 2, \ldots, m$.

(c) Bivariate normal with parameters

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \qquad \boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{12i} \\ \sigma_{12i} & \sigma_{2i}^2 \end{pmatrix},$$

for $i = 1, 2, \ldots, m$.

2. Let $\{X_t\}$ be a second-order HMM, based on a stationary second-order Markov chain $\{C_t\}$ on m states. How many parameters does the model have in the following cases? (Assume for simplicity that one parameter is needed to specify each of the m state-dependent distributions.)

(a) $\{C_t\}$ is a general second-order Markov chain.

(b) $\{C_t\}$ is a Pegram model.

(c) $\{C_t\}$ is a Raftery model.

3. Let $\{X_t\}$ be a second-order HMM, based on a stationary second-order Markov chain $\{C_t\}$ on m states. For integers $t \geq 2$, and integers i and j from 1 to m, define

$$\nu_t(i, j; \mathbf{x}^{(t)}) = \Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_{t-1} = i, C_t = j).$$

Note that these probabilities are just an extension to two dimensions of the forward probabilities $\alpha_t(i)$, which could more explicitly be denoted by $\alpha_t(i; \mathbf{x}^{(t)})$. For the case $t = 2$ we have

$$\nu_2(i, j; \mathbf{x}^{(t)}) = u(i, j) p_i(x_1) p_j(x_2).$$

(a) Show that, for integers $t \geq 3$,

$$\nu_t(j, k; \mathbf{x}^{(t)}) = \left(\sum_{i=1}^{m} \nu_{t-1}(i, j; \mathbf{x}^{(t-1)}) \, \gamma(i, j, k) \right) p_k(x_t). \qquad (8.5)$$

(b) Show how the recursion (8.5) can be used to compute the likelihood of a series of T observations, $\Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})$.

(c) Show that the computational effort required to find the likelihood thus is $O(Tm^3)$.

4. Consider the generalization of the basic HMM which allows the distribution of an observation to depend on the current state of a Markov chain *and* the previous one; see Section 8.6. Prove the results given in Equations (8.3) and (8.4) for the likelihood of such a model.

5. Find the autocorrelation functions for stationary HMMs with (a) normal and (b) binomial state-dependent distributions.

6. (Runlengths in Bernoulli–HMMs) Let $\{X_t\}$ be a Bernoulli–HMM in which the underlying stationary irreducible Markov chain has the states 1, 2, ..., m, transition probability matrix $\boldsymbol{\Gamma}$ and stationary

distribution $\boldsymbol{\delta}$. The probability of an observation being 1 in state i is denoted by p_i. Define a run of ones as follows: such a run is initiated by the sequence 01, and is said to be of length $k \in \mathbb{N}$ if that sequence is followed by a further $k-1$ ones and a zero (in that order).

(a) Let K denote the length of a run of ones. Show that

$$\Pr(K=k) = \frac{\boldsymbol{\delta}\mathbf{P}(0)(\boldsymbol{\Gamma}\mathbf{P}(1))^k\boldsymbol{\Gamma}\mathbf{P}(0)\mathbf{1}'}{\delta\mathbf{P}(0)\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}'},$$

where $\mathbf{P}(1) = \text{diag}(p_1, \ldots, p_m)$ and $\mathbf{P}(0) = \mathbf{I}_m - \mathbf{P}(1)$.

(b) Suppose that $\mathbf{B} = \boldsymbol{\Gamma}\mathbf{P}(1)$ has distinct eigenvalues ω_i. Show that the probability $\Pr(K=k)$ is, as a function of k, a linear combination of the kth powers of these eigenvalues, and hence of $\omega_i^{k-1}(1-\omega_i)$, $i=1,\ldots,m$.

(c) Does this imply that K is a mixture of geometric random variables? (Hint: will the eigenvalues ω_i always lie between zero and one?)

(d) Assume for the rest of this exercise that there are only two states, with the t.p.m. given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1-\gamma_{12} & \gamma_{12} \\ \gamma_{21} & 1-\gamma_{21} \end{pmatrix},$$

and that $p_1=0$ and $p_2 \in (0,1)$.

Show that the distribution of K is as follows, for all $k \in \mathbb{N}$:

$$\Pr(K=k) = ((1-\gamma_{21})p_2)^{k-1}(1-(1-\gamma_{21})p_2).$$

(So although not itself a Markov chain, this HMM has a geometric distribution for the length of a run of ones.)

(e) For such a model, will the length of a run of zeros also be geometrically distributed?

7. Consider the three two-state stationary Bernoulli–HMMs specified below. For instance, in model (a), $\Pr(X_t = 1 \mid C_t = 1) = 0.1$ and $\Pr(X_t = 1 \mid C_t = 2) = 1$, and X_t is either one or zero. The states are determined in accordance with the stationary Markov chain with t.p.m. $\boldsymbol{\Gamma}$. (Actually, (c) is a Markov chain; there is in that case a one-to-one correspondence between states and observations.)

Let K denote the length of a run of ones. In each of the three cases, determine the following: $\Pr(K = k)$, $\Pr(K \leq 10)$, $\mathrm{E}(K)$, σ_K and $\text{corr}(X_t, X_{t+k})$.

(a)

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.99 & 0.01 \\ 0.08 & 0.92 \end{pmatrix}, \qquad \mathbf{p} = (0.1, 1).$$

(b)
$$\mathbf{\Gamma} = \begin{pmatrix} 0.98 & 0.02 \\ 0.07 & 0.93 \end{pmatrix}, \qquad \mathbf{p} = (0, 0.9).$$

(c)
$$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}, \qquad \mathbf{p} = (0, 1).$$

Notice that these three models are comparable in that (i) the unconditional probability of an observation being one is in all cases 0.2; and (ii) all autocorrelations are positive and decrease geometrically.

8. Consider a two-state bivariate HMM $\{(X_{t1}, X_{t2}) : t \in \mathbb{N}\}$, based on a stationary Markov chain with transition probability matrix $\mathbf{\Gamma}$ and stationary distribution $\boldsymbol{\delta} = (\delta_1, \delta_2)$. Let μ_{i1} and μ_{i2} denote the means of X_{t1} and X_{t2} in state i, and similarly σ^2_{i1} and σ^2_{i2} the variances. Assume both contemporaneous conditional independence and longitudinal conditional independence.

   (a) Show that, for all nonnegative integers k,
   $$\mathrm{Cov}(X_{t1}, X_{t+k,2}) = \delta_1\delta_2(\mu_{11} - \mu_{21})(\mu_{12} - \mu_{22})(1 - \gamma_{12} - \gamma_{21})^k.$$
   (b) Hence find the cross-correlations $\mathrm{Corr}(X_{t1}, X_{t+k,2})$.
   (c) Does contemporaneous conditional independence imply independence of X_{t1} and X_{t2}?
   (d) State a sufficient condition for X_{t1} and X_{t2} to be uncorrelated.
   (e) Generalize the results of (a) and (b) to any number (m) of states.

9. (Irreversibility of a binomial–HMM) In Section 1.3.3 we defined reversibility for a random process, and showed that the stationary Markov chain with the t.p.m. $\mathbf{\Gamma}$ given below is not reversible.
$$\mathbf{\Gamma} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$
Let $\{X_t\}$ be the stationary HMM with $\mathbf{\Gamma}$ as above, and having binomial state-dependent distributions with parameters 2 and 0/0.5/1; e.g. in state 1 the observation X_t is distributed Binomial(2, 0).
By finding the probabilities $\Pr(X_t = 0, X_{t+1} = 1)$ and $\Pr(X_t = 1, X_{t+1} = 0)$, or otherwise, show that $\{X_t\}$ is irreversible.