CHAPTER 2

Hidden Markov models: definition and properties

2.1 A simple hidden Markov model

Consider again the observed earthquake series displayed in Figure 1.1 on p. 4. The observations are unbounded counts, making the Poisson distribution a natural choice to describe them, but their distribution is clearly overdispersed relative to the Poisson. We saw in Chapter 1 that this feature can be accommodated by using a mixture of Poisson distributions with means λ_1, λ_2, …, λ_m. The choice of mean is made by a second random process, the parameter process. The mean λ_i is selected with probability δ_i, where $i = 1, 2, \ldots, m$ and $\sum_{i=1}^m \delta_i = 1$.

An independent mixture model will not do for the earthquake series because – by definition – it does not allow for the serial dependence in the observations. The sample autocorrelation function (ACF), displayed in Figure 2.1, clearly indicates that the observations are serially dependent. One way of allowing for serial dependence in the observations is to relax the assumption that the parameter process is serially independent. A simple and mathematically convenient way to do so is to assume that it is a Markov chain. The resulting model for the observations is called a Poisson–hidden Markov model, a simple example of the class of

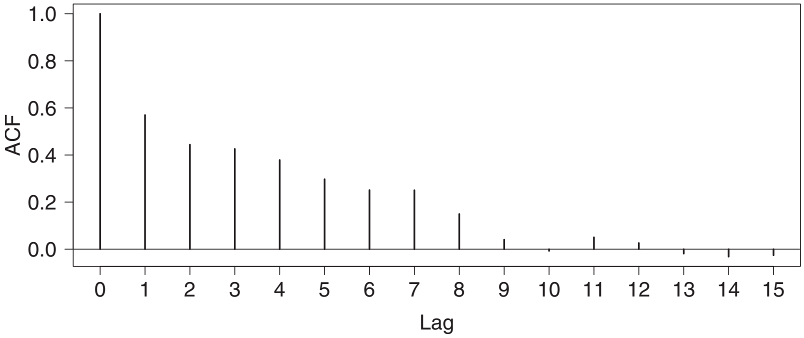

Figure 2.1 *Earthquakes series: sample autocorrelation function.*

models discussed in the rest of this book, namely hidden Markov models (HMMs).

We shall not give an account here of the (interesting) history of such models, but two valuable sources of information on HMMs that include accounts of the history are Ephraim and Merhav (2002) and Cappé *et al.* (2005).

2.2 The basics

2.2.1 Definition and notation

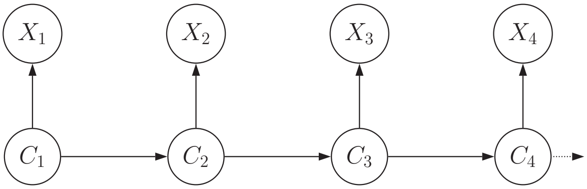

Figure 2.2 *Directed graph of basic HMM.*

A **hidden Markov model** $\{X_t : t \in \mathbb{N}\}$ is a particular kind of dependent mixture. With $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$ representing the histories from time 1 to time t, one can summarize the simplest model of this kind by:

$$\Pr(C_t \mid \mathbf{C}^{(t-1)}) = \Pr(C_t \mid C_{t-1}), \quad t = 2, 3, \dots \tag{2.1}$$

$$\Pr(X_t \mid \mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = \Pr(X_t \mid C_t), \quad t \in \mathbb{N}. \tag{2.2}$$

The model consists of two parts: firstly, an unobserved 'parameter process' $\{C_t : t = 1, 2, \dots\}$ satisfying the Markov property; and secondly, the 'state-dependent process' $\{X_t : t = 1, 2, \dots\}$, in which the distribution of X_t depends only on the current state C_t and not on previous states or observations. This structure is represented by the directed graph in Figure 2.2.

If the Markov chain $\{C_t\}$ has m states, we call $\{X_t\}$ an m-state HMM. Although it is the usual terminology in speech-processing applications, the name 'hidden Markov model' is by no means the only one used for such models or similar ones. For instance, Ephraim and Merhav (2002) argue for 'hidden Markov process', Leroux and Puterman (1992) use 'Markov-dependent mixture', and others use 'Markov-switching model' (especially for models with extra dependencies at the level of the observations X_t), 'models subject to Markov regime', 'Markov mixture model', or 'latent Markov model'. Bartolucci, Farcomeni and Pennoni (2013) use the term 'latent Markov model' specifically for models for longitudinal data, as opposed to single time series.

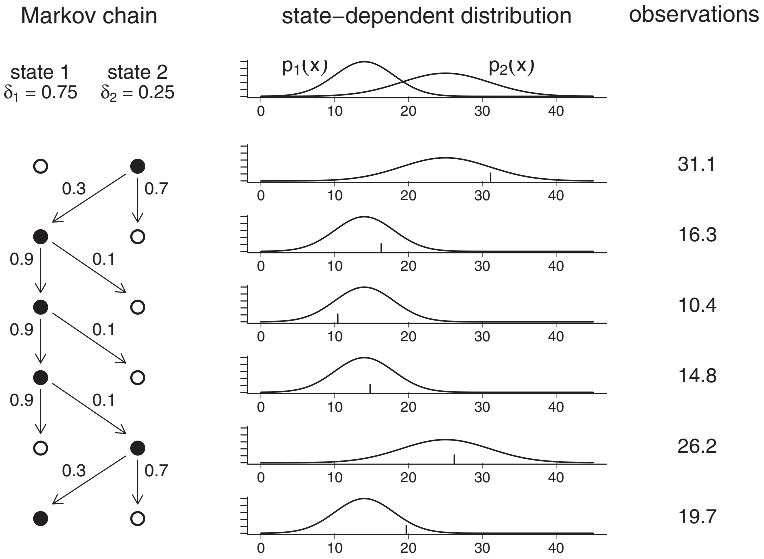

Figure 2.3 *Process generating the observations in a two-state HMM. The chain followed the path* $2, 1, 1, 1, 2, 1$*, as indicated on the left. The corresponding state-dependent distributions are shown in the middle. The observations are generated from the corresponding active distributions.*

The process generating the observations is demonstrated again in Figure 2.3, for state-dependent distributions p_1 and p_2, stationary distribution $\boldsymbol{\delta} = (0.75, 0.25)$, and t.p.m. $\boldsymbol{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$. In contrast to the case of an independent mixture, here the distribution of C_t, the state at time t, does depend on C_{t-1}. As is also true of independent mixtures, there is for each state a different distribution, discrete or continuous.

We now introduce some notation which will cover both discrete- and continuous-valued observations. In the case of discrete observations we define, for $i = 1, 2, \ldots, m$,

$$p_i(x) = \Pr(X_t = x \mid C_t = i).$$

That is, p_i is the probability mass function of X_t if the Markov chain is in state i at time t. The continuous case is treated similarly: there we define p_i to be the probability *density* function of X_t associated with state i. We refer to the m distributions p_i as the **state-dependent distributions** of the model. Many of our results are stated only for the discrete case, but, if probabilities are replaced by densities, apply also to the continuous case.

2.2.2 Marginal distributions

We shall often need the marginal distribution of X_t and also higher-order marginal distributions, such as that of (X_t, X_{t+k}). We shall derive the results for the case in which the Markov chain is homogeneous but not necessarily stationary, and then give them also for the special case in which the Markov chain is stationary. For convenience the derivation is given only for discrete state-dependent distributions; the continuous case can be derived analogously.

Univariate distributions

For discrete-valued observations X_t, defining $u_i(t) = \Pr(C_t = i)$ for $t = 1, \dots, T$, we have

$$\begin{aligned}\Pr(X_t = x) &= \sum_{i=1}^{m} \Pr(C_t = i) \Pr(X_t = x \mid C_t = i) \\ &= \sum_{i=1}^{m} u_i(t) p_i(x).\end{aligned}$$

This expression can conveniently be rewritten in matrix notation:

$$\begin{aligned}\Pr(X_t = x) &= (u_1(t), \dots, u_m(t)) \begin{pmatrix} p_1(x) & & 0 \\ & \ddots & \\ 0 & & p_m(x) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \mathbf{u}(t)\mathbf{P}(x)\mathbf{1}',\end{aligned}$$

where $\mathbf{P}(x)$ is defined as the diagonal matrix with ith diagonal element $p_i(x)$. It follows from equation (1.3) that $\mathbf{u}(t) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}$, and hence that

$$\Pr(X_t = x) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}\mathbf{P}(x)\mathbf{1}'. \tag{2.3}$$

Equation (2.3) holds if the Markov chain is merely homogeneous, and not necessarily stationary. If, as we shall often assume, the Markov chain is stationary, with stationary distribution $\boldsymbol{\delta}$, then the result is simpler: in that case $\boldsymbol{\delta}\mathbf{\Gamma}^{t-1} = \boldsymbol{\delta}$ for all $t \in \mathbb{N}$, and so

$$\Pr(X_t = x) = \boldsymbol{\delta}\mathbf{P}(x)\mathbf{1}'. \tag{2.4}$$

Bivariate distributions

The calculation of many of the distributions relating to an HMM is most easily done by first noting that, in any directed graphical model, the joint distribution of a set of random variables V_i is given by

$$\Pr(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i \mid \mathrm{pa}(V_i)), \tag{2.5}$$

where $\mathrm{pa}(V_i)$ denotes the set of all 'parents' of V_i in the set V_1, V_2, $\ldots$, V_n; see, for example, Davison (2003, p. 250) or Jordan (2004).

In the directed graph of the four random variables X_t, X_{t+k}, C_t, C_{t+k} (for positive integer k), C_t has no parents, $\mathrm{pa}(X_t) = \{C_t\}$, $\mathrm{pa}(C_{t+k}) = \{C_t\}$ and $\mathrm{pa}(X_{t+k}) = \{C_{t+k}\}$. It therefore follows that

$$\Pr(X_t, X_{t+k}, C_t, C_{t+k}) = \Pr(C_t)\Pr(X_t|C_t)\Pr(C_{t+k}|C_t)\Pr(X_{t+k}|C_{t+k}),$$

and hence that

$$\begin{aligned}
&\Pr(X_t = v, X_{t+k} = w) \\
&\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr(X_t = v, X_{t+k} = w, C_t = i, C_{t+k} = j) \\
&\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} \underbrace{\Pr(C_t = i)}_{u_i(t)}\, p_i(v)\, \underbrace{\Pr(C_{t+k} = j \mid C_t = i)}_{\gamma_{ij}(k)}\, p_j(w) \\
&\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t) p_i(v) \gamma_{ij}(k) p_j(w).
\end{aligned}$$

(Here and elsewhere, $\gamma_{ij}(k)$ denotes the (i,j) element of $\boldsymbol{\Gamma}^k$.) Writing the above double sum as a product of matrices yields

$$\Pr(X_t = v, X_{t+k} = w) \;=\; \mathbf{u}(t)\mathbf{P}(v)\boldsymbol{\Gamma}^k\mathbf{P}(w)\mathbf{1}'. \tag{2.6}$$

If the Markov chain is stationary, this reduces to

$$\Pr(X_t = v, X_{t+k} = w) \;=\; \boldsymbol{\delta}\mathbf{P}(v)\boldsymbol{\Gamma}^k\mathbf{P}(w)\mathbf{1}'. \tag{2.7}$$

Similarly, one can obtain expressions for the higher-order marginal distributions; in the stationary case, the formula for a trivariate distribution is, for positive integers k and l,

$$\Pr(X_t = v, X_{t+k} = w, X_{t+k+l} = z) = \boldsymbol{\delta}\mathbf{P}(v)\boldsymbol{\Gamma}^k\mathbf{P}(w)\boldsymbol{\Gamma}^l\mathbf{P}(z)\mathbf{1}'.$$

2.2.3 Moments

First, we note that

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \mathrm{E}(X_t \mid C_t = i)\Pr(C_t = i) = \sum_{i=1}^{m} u_i(t)\mathrm{E}(X_t \mid C_t = i),$$

which, in the stationary case, reduces to

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \delta_i \mathrm{E}(X_t \mid C_t = i).$$

More generally, analogous results hold for $\mathrm{E}(g(X_t))$ and $\mathrm{E}(g(X_t, X_{t+k}))$, for any functions g for which the relevant state-dependent expectations exist. In the stationary case

$$\mathrm{E}(g(X_t)) = \sum_{i=1}^{m} \delta_i \mathrm{E}(g(X_t) \mid C_t = i); \tag{2.8}$$

and

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g(X_t, X_{t+k}) \mid C_t = i, C_{t+k} = j)\, \delta_i \gamma_{ij}(k). \tag{2.9}$$

Often we shall be interested in a function g which factorizes as

$$g(X_t, X_{t+k}) = g_1(X_t) g_2(X_{t+k}),$$

in which case equation (2.9) becomes

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g_1(X_t) \mid C_t = i)\, \mathrm{E}(g_2(X_{t+k}) \mid C_{t+k} = j)\, \delta_i \gamma_{ij}(k). \tag{2.10}$$

These expressions enable us to find covariances and correlations; convenient explicit expressions exist in many cases. For the case of a stationary two-state Poisson–HMM:

- $\mathrm{E}(X_t) = \delta_1 \lambda_1 + \delta_2 \lambda_2$;
- $\mathrm{Var}(X_t) = \mathrm{E}(X_t) + \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 \geq \mathrm{E}(X_t)$;
- $\mathrm{Cov}(X_t, X_{t+k}) = \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 (1 - \gamma_{12} - \gamma_{21})^k$, for $k \in \mathbb{N}$.

Notice that the resulting formula for the correlation of X_t and X_{t+k} is of the form $\rho(k) = A(1 - \gamma_{12} - \gamma_{21})^k$ with $A \in [0, 1)$, and that $A = 0$ if $\lambda_1 = \lambda_2$. For more details, and for more general results, see Exercises 3 and 4.

2.3 The likelihood

The aim of this section is to develop a convenient formula for the likelihood L_T of T consecutive observations $x_1, x_2, \ldots, x_T$ assumed to be generated by an m-state HMM. That such a formula exists is indeed fortunate, but by no means obvious. We shall see that the computation of the likelihood, consisting as it does of a sum of m^T terms, each of which is a product of $2T$ factors, appears to require $O(Tm^T)$ operations. However, it has long been known in several contexts that the likelihood is easily computable; see, for example, Baum (1972), Lange and Boehnke (1983), and Cosslett and Lee (1985). What we describe here is in fact a special case of a much more general theory; see Smyth, Heckerman and Jordan (1997) or Jordan (2004).

It is our purpose here to demonstrate that L_T can in general be computed relatively simply in $O(Tm^2)$ operations. The way will then be open to estimate parameters by numerical maximization of the likelihood. First the likelihood of a two-state model will be explored, and then the general formula will be presented.

2.3.1 The likelihood of a two-state Bernoulli–HMM

Example. Consider the stationary two-state HMM with t.p.m.

$$\boldsymbol{\Gamma} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

and state-dependent distributions given by

$$\begin{aligned} \Pr(X_t = x \mid C_t = 1) &= \frac{1}{2} \quad (\text{for } x = 0, 1), \\ \Pr(X_t = 1 \mid C_t = 2) &= 1. \end{aligned}$$

We call a model of this kind a **Bernoulli–HMM**. The stationary distribution of the Markov chain is $\boldsymbol{\delta} = \frac{1}{3}(1, 2)$. Consider the probability $\Pr(X_1 = X_2 = X_3 = 1)$. First, note that, by equation (2.5),

$$\begin{aligned} &\Pr(X_1, X_2, X_3, C_1, C_2, C_3) \\ &= \Pr(C_1)\Pr(X_1 \mid C_1)\Pr(C_2 \mid C_1)\Pr(X_2 \mid C_2)\Pr(C_3 \mid C_2)\Pr(X_3 \mid C_3); \end{aligned}$$

and then sum over the values assumed by C_1, C_2, C_3. The result is

$$\begin{aligned} &\Pr(X_1 = 1, X_2 = 1, X_3 = 1) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \Pr(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \delta_i p_i(1) \gamma_{ij} p_j(1) \gamma_{jk} p_k(1). \qquad (2.11) \end{aligned}$$

Notice that the triple sum (2.11) has $m^T = 2^3$ terms, each of which is a product of $2T = 2 \times 3$ factors. To evaluate the required probability, the different possibilities for the values of i, j and k can be listed and the sum (2.11) calculated as in Table 2.1.

Summation of the last column of Table 2.1 tells us that $\Pr(X_1 =$

Table 2.1 *Example of a likelihood computation.*

i	j	k	$p_i(1)$	$p_j(1)$	$p_k(1)$	δ_i	γ_{ij}	γ_{jk}	Product
1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{96}$
1	1	2	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
1	2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{96}$
1	2	2	$\frac{1}{2}$	1	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
2	1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
2	1	2	1	$\frac{1}{2}$	1	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{24}$
2	2	1	1	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{16}$
2	2	2	1	1	1	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{8}$
									$\frac{29}{48}$

$1, X_2 = 1, X_3 = 1) = \frac{29}{48}$. In passing we note that the largest element in that column is $\frac{3}{8}$; the state sequence that maximizes the joint probability

$$\Pr(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k)$$

is therefore the sequence $i = 2$, $j = 2$, $k = 2$. Equivalently, it maximizes the conditional probability $\Pr(C_1 = i, C_2 = j, C_3 = k \mid X_1 = 1, X_2 = 1, X_3 = 1)$. This is an example of 'global decoding', which will be discussed in Section 5.4.2.

But a more convenient way to present the sum is to use matrix notation. Let $\mathbf{P}(u)$ be defined (as before) as $\text{diag}(p_1(u), p_2(u))$. Then

$$\mathbf{P}(0) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{P}(1) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix},$$

and the triple sum (2.11) can be written as a matrix product:

$$\boldsymbol{\delta}\mathbf{P}(1)\boldsymbol{\Gamma}\mathbf{P}(1)\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}'.$$

More generally, the likelihood turns out to be a T-fold sum which can also be written as a matrix product.

2.3.2 The likelihood in general

Here we consider the likelihood of an HMM in general. We suppose there is an observation sequence $x_1, x_2, \ldots, x_T$ generated by such a model. We seek the probability L_T of observing that sequence, as calculated under an m-state HMM which has *initial* distribution $\boldsymbol{\delta}$ and t.p.m. $\boldsymbol{\Gamma}$ for the Markov chain, and state-dependent probability (density) functions p_i. In many of our applications we shall assume that $\boldsymbol{\delta}$ is the stationary distribution implied by $\boldsymbol{\Gamma}$, but it is not necessary to make that assumption in general.

Proposition 1 *The likelihood is given by*

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'. \tag{2.12}$$

If $\boldsymbol{\delta}$, the distribution of C_1, is the stationary distribution of the Markov chain, then in addition

$$L_T = \boldsymbol{\delta}\boldsymbol{\Gamma}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'. \tag{2.13}$$

Proof. We present only the case of discrete observations. First, note that

$$L_T = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \sum_{c_1,c_2,\ldots,c_T=1}^{m} \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, \mathbf{C}^{(T)} = \mathbf{c}^{(T)}),$$

and that, by equation (2.5),

$$\Pr(\mathbf{X}^{(T)}, \mathbf{C}^{(T)}) = \Pr(C_1)\prod_{k=2}^{T}\Pr(C_k \mid C_{k-1})\prod_{k=1}^{T}\Pr(X_k \mid C_k). \tag{2.14}$$

It follows that

$$\begin{aligned} L_T &= \sum_{c_1,\ldots,c_T=1}^{m} \left(\delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}\right)\left(p_{c_1}(x_1)p_{c_2}(x_2)\cdots p_{c_T}(x_T)\right) \\ &= \sum_{c_1,\ldots,c_T=1}^{m} \delta_{c_1}p_{c_1}(x_1)\gamma_{c_1,c_2}p_{c_2}(x_2)\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}p_{c_T}(x_T) \\ &= \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}', \end{aligned}$$

which is equation (2.12). The last equality above exploits the fact that a multiple sum of terms having a certain simple multiplicative form can in general be written as a matrix product. Exercise 7(b) provides the detailed justification.

If $\boldsymbol{\delta}$ is the stationary distribution of the Markov chain, we have

$$\boldsymbol{\delta}\mathbf{P}(x_1) = \boldsymbol{\delta}\boldsymbol{\Gamma}\mathbf{P}(x_1),$$

hence equation (2.13), which involves an extra factor of $\boldsymbol{\Gamma}$ but may be slightly simpler to code. □

A very simple but crucial consequence of the matrix expression for the likelihood is the 'forward algorithm' for recursive computation of the likelihood. Such recursive computation plays a key role, not only in likelihood evaluation and hence parameter estimation, but also in forecasting, decoding and model checking. The recursive nature of likelihood evaluation via either (2.12) or (2.13) is computationally much more efficient than brute-force summation over all possible state sequences. The fact that such computationally inexpensive recursive schemes can be used to address various questions of interest is a key feature of HMMs. Recursive evaluation of such multiple sums has been discussed by Lange and Boehnke (1983) and Lange (2002, p. 120).

To state the forward algorithm we define the vector $\boldsymbol{\alpha}_t$, for $t = 1, 2, \ldots, T$, by

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_t) = \boldsymbol{\delta}\mathbf{P}(x_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\mathbf{P}(x_s), \quad (2.15)$$

with the convention that an empty product is the identity matrix. It follows immediately from this definition that

$$L_T = \boldsymbol{\alpha}_T\mathbf{1}', \quad \text{and} \quad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t \geq 2.$$

Accordingly, we can conveniently set out as follows the computations involved in the likelihood formula (2.12):

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\mathbf{P}(x_1); \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 2, 3, \ldots, T; \\ L_T &= \boldsymbol{\alpha}_T\mathbf{1}'. \end{aligned}$$

That the number of operations involved is of order Tm^2 can be deduced thus. For each of the values of t in the loop, there are m elements of $\boldsymbol{\alpha}_t$ to be computed, and each of those elements is a sum of m products of three quantities: an element of $\boldsymbol{\alpha}_{t-1}$, a transition probability γ_{ij}, and a state-dependent probability (or density) $p_j(x_t)$.

The corresponding scheme for computation of (2.13) (i.e. if $\boldsymbol{\delta}$, the distribution of C_1, is the stationary distribution of the Markov chain) is

$$\begin{aligned} \boldsymbol{\alpha}_0 &= \boldsymbol{\delta}; \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 1, 2, \ldots, T; \\ L_T &= \boldsymbol{\alpha}_T\mathbf{1}'. \end{aligned}$$

The elements of the vector $\boldsymbol{\alpha}_t$ are usually referred to as **forward probabilities**; the reason for this name will be seen in Section 4.1.1, where we show that the jth element of $\boldsymbol{\alpha}_t$ is $\Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j)$.

We show here **R** code that uses the forward algorithm to evaluate the likelihood of observations $x_1, \ldots, x_T$ under a Poisson–HMM with at least two states, t.p.m. $\boldsymbol{\Gamma}$, vector of state-dependent means $\boldsymbol{\lambda}$, and initial distribution $\boldsymbol{\delta}$ (not necessarily the stationary distribution). Note, however, that, unless the series is short, one needs to guard against underflow and evaluate the log-likelihood rather than the likelihood; see p. 49 for code that does so.

```
alpha                  <- delta*dpois(x[1],lambda)
for (i in 2:T) alpha <- alpha %*% Gamma*dpois(x[i],lambda)
sum(alpha)
```

In the above discussion we have used the multiple-sum expression for the likelihood in order to arrive at the matrix expression, and then used the matrix expression to arrive at the forward recursion. An alternative route, which anticipates some of the material of Chapter 4, is to *define* the vector of forward probabilities $\boldsymbol{\alpha}_t$ by

$$\alpha_t(j) = \Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j), \quad j = 1, 2, \dots, m,$$

and then to deduce the forward recursion:

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t).$$

The matrix expression is then a simple consequence of the forward recursion. This alternative route is described in Exercise 8.

2.3.3 HMMs are not Markov processes

HMMs do not in general satisfy the Markov property. This we can now establish via a simple counterexample. Let X_t and C_t be as defined in the example in Section 2.3.1. We already know that

$$\Pr(X_1 = 1, X_2 = 1, X_3 = 1) = \frac{29}{48},$$

and from equations (2.4) and (2.7) it can be established that $\Pr(X_2 = 1) = \frac{5}{6}$ and that

$$\Pr(X_1 = 1, X_2 = 1) = \Pr(X_2 = 1, X_3 = 1) = \frac{17}{24}.$$

It therefore follows that

$$\begin{aligned} \Pr(X_3 = 1 \mid X_1 = 1, X_2 = 1) &= \frac{\Pr(X_1 = 1, X_2 = 1, X_3 = 1)}{\Pr(X_1 = 1, X_2 = 1)} \\ &= \frac{29/48}{17/24} = \frac{29}{34} \end{aligned}$$

and that

$$\begin{aligned} \Pr(X_3 = 1 \mid X_2 = 1) &= \frac{\Pr(X_2 = 1, X_3 = 1)}{\Pr(X_2 = 1)} \\ &= \frac{17/24}{5/6} = \frac{17}{20}. \end{aligned}$$

Hence $\Pr(X_3 = 1 \mid X_2 = 1) \neq \Pr(X_3 = 1 \mid X_1 = 1, X_2 = 1)$; this HMM does not satisfy the Markov property. That some HMMs do satisfy the property, however, is clear. For instance, a two-state Bernoulli–HMM can degenerate in obvious fashion to the underlying Markov chain; one simply identifies each of the two observable values with one of the two underlying states. For the conditions under which an HMM will itself satisfy the Markov property, see Spreij (2001).

2.3.4 The likelihood when data are missing

In a time series context it is potentially awkward if some of the data are missing. In the case of hidden Markov time series models, however, the adjustment that needs to be made to the likelihood computation if data are missing turns out to be a simple one.

Suppose, for example, that one has available the observations x_1, x_2, x_4, x_7, x_8, ..., x_T of an HMM, but x_3, x_5 and x_6 are missing. Then the likelihood of the observations is given by

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, X_4 = x_4, X_7 = x_7, \ldots, X_T = x_T)\\ = \sum \delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_4}(2)\gamma_{c_4,c_7}(3)\gamma_{c_7,c_8}\cdots\gamma_{c_{T-1},c_T}\\ \times p_{c_1}(x_1)p_{c_2}(x_2)p_{c_4}(x_4)p_{c_7}(x_7)\cdots p_{c_T}(x_T),\end{aligned}$$

where (as before) $\gamma_{ij}(k)$ denotes a k-step transition probability, and the sum is taken over all indices c_t other than c_3, c_5 and c_6. But this is just

$$\begin{aligned}\sum \delta_{c_1}p_{c_1}(x_1)\gamma_{c_1,c_2}p_{c_2}(x_2)\gamma_{c_2,c_4}(2)p_{c_4}(x_4)\gamma_{c_4,c_7}(3)p_{c_7}(x_7)\\ \times\cdots\times\gamma_{c_{T-1},c_T}p_{c_T}(x_T)\\ = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}^2\mathbf{P}(x_4)\boldsymbol{\Gamma}^3\mathbf{P}(x_7)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'.\end{aligned}$$

With $L_T^{-(3,5,6)}$ denoting the likelihood of the observations (other than x_3, x_5 and x_6), it follows that

$$L_T^{-(3,5,6)} = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}^2\mathbf{P}(x_4)\boldsymbol{\Gamma}^3\mathbf{P}(x_7)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'.$$

In general, in the expression for the likelihood the diagonal matrices $\mathbf{P}(x_t)$ corresponding to missing observations x_t are replaced by the identity matrix; that is, the corresponding state-dependent probabilities $p_i(x_t)$ are replaced by 1 for all states i. If one can assume that the missingness is ignorable, this 'ignorable likelihood' is a reasonable basis for estimating parameters (Little, 2009, p. 411).

The fact that, even if some observations are missing, the likelihood of an HMM can be computed easily is especially useful in the derivation of conditional distributions, as will be shown in Section 5.2.

2.3.5 The likelihood when observations are interval-censored

Suppose that we wish to fit a Poisson–HMM to a series of counts, some of which are interval-censored. For instance, the value of x_t may be known only for $4 \leq t \leq T$, with the information $x_1 \leq 5$, $2 \leq x_2 \leq 3$ and $x_3 > 10$ available about the remaining observations. For simplicity, let us first assume that the Markov chain has only two states. In that case, one replaces the diagonal matrices $\mathbf{P}(x_i)$ $(i = 1, 2, 3)$ in the likelihood expression (2.12) by the matrices

$$\begin{aligned}&\text{diag}(\Pr(X_1 \leq 5 \mid C_1 = 1), \Pr(X_1 \leq 5 \mid C_1 = 2)),\\&\text{diag}(\Pr(2 \leq X_2 \leq 3 \mid C_2 = 1), \Pr(2 \leq X_2 \leq 3 \mid C_2 = 2)), \text{and}\\&\text{diag}(\Pr(X_3 > 10 \mid C_3 = 1), \Pr(X_3 > 10 \mid C_3 = 2)).\end{aligned}$$

More generally, suppose that $a \leq x_t \leq b$, where a may be $-\infty$ (although that is not relevant to the Poisson case), b may be ∞, and the Markov chain has m states. One replaces $\mathbf{P}(x_t)$ in the likelihood by the $m \times m$ diagonal matrix of which the ith diagonal element is $\Pr(a \leq X_t \leq b \mid C_t = i)$. See Exercise 12. It is worth noting that missing data can be regarded as an extreme case of such interval-censoring.

Exercises

1. Consider a *stationary* two-state Poisson–HMM with parameters

$$\mathbf{\Gamma} = \begin{pmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda} = (1, 3).$$

In each of the following ways, compute the probability that the first three observations from this model are 0, 2, 1.

(a) Consider all possible sequences of states of the Markov chain that could have occurred. Compute the probability of each sequence, and the probability of the observations given each sequence.

(b) Apply the formula

$$\Pr(X_1 = 0, X_2 = 2, X_3 = 1) = \boldsymbol{\delta}\mathbf{P}(0)\mathbf{\Gamma}\mathbf{P}(2)\mathbf{\Gamma}\mathbf{P}(1)\mathbf{1}',$$

where

$$\mathbf{P}(s) = \begin{pmatrix} \lambda_1^s e^{-\lambda_1}/s! & 0 \\ 0 & \lambda_2^s e^{-\lambda_2}/s! \end{pmatrix} = \begin{pmatrix} 1^s e^{-1}/s! & 0 \\ 0 & 3^s e^{-3}/s! \end{pmatrix}.$$

2. Consider again the model defined in Exercise 1. In that question you were asked to compute $\Pr(X_1 = 0, X_2 = 2, X_3 = 1)$. Now compute $\Pr(X_1 = 0, X_3 = 1)$ in each of the following ways.

(a) Consider all possible sequences of states of the Markov chain that could have occurred. Compute the probability of each sequence, and the probability of the observations given each sequence.

(b) Apply the formula

$$\Pr(X_1 = 0, X_3 = 1) = \boldsymbol{\delta}\mathbf{P}(0)\boldsymbol{\Gamma}\mathbf{I}_2\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}' = \boldsymbol{\delta}\mathbf{P}(0)\boldsymbol{\Gamma}^2\mathbf{P}(1)\mathbf{1}',$$

and check that this probability is equal to your answer in (a).

3. Consider an m-state HMM $\{X_t : t = 1, 2, \ldots\}$, based on a stationary Markov chain with transition probability matrix $\boldsymbol{\Gamma}$ and stationary distribution $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$, and having (univariate) state-dependent distributions $p_i(x)$. Let μ_i and σ_i^2 denote the mean and variance of the distribution p_i, $\boldsymbol{\mu}$ the vector $(\mu_1, \mu_2, \ldots, \mu_m)$, and $\mathbf{M}$ the matrix diag($\boldsymbol{\mu}$).

Derive the following results for the moments of $\{X_t\}$.

(a) $\mathrm{E}(X_t) = \sum_{i=1}^m \delta_i \mu_i = \boldsymbol{\delta}\boldsymbol{\mu}'$.

(b) $\mathrm{E}(X_t^2) = \sum_{i=1}^m \delta_i(\sigma_i^2 + \mu_i^2)$.

(c) $\mathrm{Var}(X_t) = \sum_{i=1}^m \delta_i(\sigma_i^2 + \mu_i^2) - (\boldsymbol{\delta}\boldsymbol{\mu}')^2$.

(d) If $m = 2$, $\mathrm{Var}(X_t) = \delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2$.

(e) For $k \in \mathbb{N}$, i.e. for positive integers k,

$$\mathrm{E}(X_t X_{t+k}) = \sum_{i=1}^m \sum_{j=1}^m \delta_i \mu_i \gamma_{ij}(k) \mu_j = \boldsymbol{\delta}\mathbf{M}\boldsymbol{\Gamma}^k\boldsymbol{\mu}'.$$

(f) For $k \in \mathbb{N}$,

$$\rho(k) = \mathrm{Corr}(X_t, X_{t+k}) = \frac{\boldsymbol{\delta}\mathbf{M}\boldsymbol{\Gamma}^k\boldsymbol{\mu}' - (\boldsymbol{\delta}\boldsymbol{\mu}')^2}{\mathrm{Var}(X_t)}.$$

Note that, if the eigenvalues of $\boldsymbol{\Gamma}$ are distinct, this is a linear combination of the kth powers of those eigenvalues.

(g) If the state-dependent means μ_i are all equal, X_t and X_{t+k} are uncorrelated for $k \in \mathbb{N}$.

Timmermann (2000) and Frühwirth-Schnatter (2006, pp. 308–312) are useful references for moments. See also Exercise 1 of Chapter 1.

4. (Marginal moments and autocorrelation function of a Poisson–HMM: special case of Exercise 3.) Consider a stationary m-state Poisson–HMM $\{X_t : t = 1, 2, \ldots\}$ with transition probability matrix $\boldsymbol{\Gamma}$ and state-dependent means $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$. Let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$ be the stationary distribution of the Markov chain. Let $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$.

Derive the following results.

(a) $\mathrm{E}(X_t) = \boldsymbol{\delta}\boldsymbol{\lambda}'$.

(b) $\mathrm{E}(X_t^2) = \sum_{i=1}^m (\lambda_i^2 + \lambda_i)\delta_i = \boldsymbol{\delta}\Lambda\boldsymbol{\lambda}' + \boldsymbol{\delta}\boldsymbol{\lambda}'$.

(c) $\mathrm{Var}(X_t) = \boldsymbol{\delta\Lambda\lambda}' + \boldsymbol{\delta\lambda}' - (\boldsymbol{\delta\lambda}')^2 = \mathrm{E}(X_t) + \boldsymbol{\delta\Lambda\lambda}' - (\boldsymbol{\delta\lambda}')^2 \geq \mathrm{E}(X_t)$.

(d) For $k \in \mathbb{N}$, $\mathrm{E}(X_t X_{t+k}) = \boldsymbol{\delta\Lambda\Gamma}^k \boldsymbol{\lambda}'$.

(e) For $k \in \mathbb{N}$,

$$\rho(k) = \mathrm{Corr}(X_t, X_{t+k}) = \frac{\boldsymbol{\delta\Lambda\Gamma}^k\boldsymbol{\lambda}' - (\boldsymbol{\delta\lambda}')^2}{\boldsymbol{\delta\Lambda\lambda}' + \boldsymbol{\delta\lambda}' - (\boldsymbol{\delta\lambda}')^2}.$$

(f) In the case $m = 2$, $\rho(k) = Aw^k$ for $k \in \mathbb{N}$, where

$$A = \frac{\delta_1\delta_2(\lambda_2 - \lambda_1)^2}{\delta_1\delta_2(\lambda_2 - \lambda_1)^2 + \boldsymbol{\delta\lambda}'}$$

and $w = 1 - \gamma_{12} - \gamma_{21}$. Notice that the extra level of randomness in the HMM, as compared with the underlying Markov chain, has reduced the autocorrelations by the factor $A \in [0, 1)$.

5. (A serially dependent process with zero autocorrelation.) In finance, time-series models consisting of serially uncorrelated but dependent random variables are often of interest. We consider here a stationary HMM $\{X_t\}$, with normal state-dependent distributions, that is such a process. Suppose that

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.990 & 0.005 & 0.005 \\ 0.010 & 0.980 & 0.010 \\ 0.015 & 0.015 & 0.970 \end{pmatrix}$$

and that, given $C_t = i$, $X_t \sim \mathrm{N}(1, \sigma_i^2)$, with $(\sigma_1, \sigma_2, \sigma_3) = (1, 10, 20)$. By Exercise 3(g), X_t and X_{t+k} are uncorrelated for $k \in \mathbb{N}$.

(a) Simulate (say) 10 000 observations $\{x_t\}$ from this model. One way of doing so is to modify the code in Section A.1.5, which applies to the case of Poisson state-dependent distributions.

(b) Using the **R** function `acf`, plot the sample ACF of:

i. $\{x_t\}$;
ii. $\{|x_t|\}$;
iii. $\{x_t^2\}$.

What can you conclude from these three sample ACFs?

(c) Find the ACF of $\{X_t^2\}$ under the model, and superimpose it on your plot of the ACF of $\{x_t^2\}$. You should get a plot similar to that shown in Figure 2.4.

6. We have the general expression

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'$$

for the likelihood of an HMM, e.g. of Poisson type.

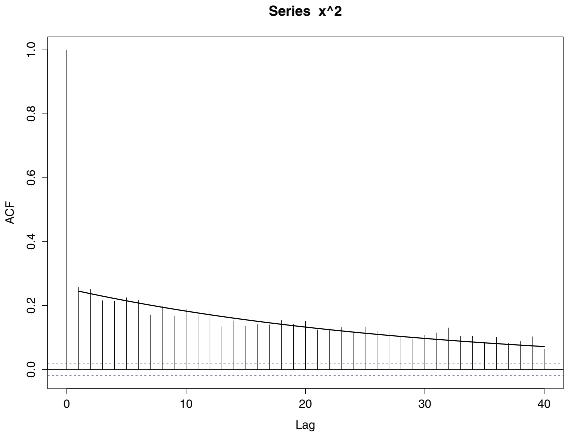

Figure 2.4 *Exercise 5: Sample autocorrelation function of the squares of 10 000 simulated observations, plus ACF of* $\{X_t^2\}$ *under the model (continuous line).*

(a) Consider the special case in which the Markov chain degenerates to a sequence of independent, identically-distributed random variables, i.e. an independent mixture model. Show that, in this case, the likelihood simplifies to the expression given in equation (1.1) for the likelihood of an *independent* mixture.

(b) Suppose instead that, for all i and x, $p_i(x) = p(x)$. What does the likelihood expression

$$\boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'$$

now reduce to, and what do you conclude?

7. This exercise shows that a sum of m^T terms of a certain simple multiplicative form can (perhaps surprisingly) be computed efficiently, in $O(Tm^2)$ operations.

   Consider a multiple sum S of the following general form:

$$S = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_T=1}^{m} h(i_1) \prod_{t=2}^{T} f_t(i_{t-1}, i_t).$$

   For $i_1 = 1, 2, \ldots, m$, define the (row) vector $\boldsymbol{\alpha}_1$ by

$$\alpha_1(i_1) \equiv h(i_1);$$

and for $r = 1, 2, \ldots, T-1$ and $i_{r+1} = 1, 2, \ldots, m$, define

$$\alpha_{r+1}(i_{r+1}) \equiv \sum_{i_r=1}^{m} \alpha_r(i_r)\, f_{r+1}(i_r, i_{r+1}).$$

That is, the vector $\boldsymbol{\alpha}_{r+1}$ is defined by, and can be computed as, $\boldsymbol{\alpha}_{r+1} = \boldsymbol{\alpha}_r\, \mathbf{F}_{r+1}$, where the $m \times m$ matrix $\mathbf{F}_t$ has (i, j) element equal to $f_t(i, j)$.

(a) Show by induction on T that $\alpha_T(i_T)$ is precisely the sum over all but i_T, i.e. that

$$\alpha_T(i_T) = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_{T-1}} h(i_1)\prod_{t=2}^{T} f_t(i_{t-1}, i_t).$$

(b) Hence show that $S = \sum_{i_T} \alpha_T(i_T) = \boldsymbol{\alpha}_T \mathbf{1}' = \boldsymbol{\alpha}_1 \mathbf{F}_2 \mathbf{F}_3 \cdots \mathbf{F}_T \mathbf{1}'$.

8. Consider an m-state HMM with the basic dependence structure as depicted in Figure 2.2.

(a) Consider the vector $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(m))$ defined by

$$\alpha_t(j) = \Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j), \quad j = 1, 2, \ldots, m.$$

Use conditional probability and the conditional independence assumptions to show that

$$\alpha_t(j) = \sum_{i=1}^{m} \alpha_{t-1}(i)\gamma_{ij}p_j(x_t).$$

(b) Verify for yourself that the result from (a), written in matrix notation, yields the forward recursion

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t), \quad t = 2, \ldots, T.$$

(c) Hence derive the matrix expression for the likelihood.

9. Write a function `pois-HMM.moments(m,lambda,gamma,lag.max=10)` that computes the expectation, variance and autocorrelation function (for lags 0 to `lag.max`) of an `m`-state stationary Poisson–HMM with t.p.m. `gamma` and state-dependent means `lambda`.

Hint: when finding the autocorrelation function, use the **R** package `expm` to compute the necessary powers of the t.p.m.

10. Write the three functions listed below, relating to the marginal distribution of an `m`-state Poisson–HMM with parameters `lambda`, `gamma`, and possibly `delta`. In each case, if `delta` is specified as `NULL`, the stationary distribution should be used. You can use your function `statdist` (see Exercise 9(b) of Chapter 1) to provide the stationary distribution.

```
dpois.HMM(x, m, lambda, gamma, delta=NULL)
ppois.HMM(x, m, lambda, gamma, delta=NULL)
qpois.HMM(p, m, lambda, gamma, delta=NULL)
```

The function `dpois.HMM` computes the probability function at the arguments specified by the vector `x`, `ppois.HMM` the distribution function, and `qpois.HMM` the inverse distribution function.

11. Consider the function `pois.HMM.generate_sample` in Section A.1.5 that generates observations from a stationary m-state Poisson–HMM. Test the function by generating a long sequence of observations (10 000, say), and then check whether the sample mean, variance, ACF and relative frequencies correspond to what you expect.

12. (Interval-censored observations.)

    (a) Suppose that, in a series of unbounded counts $x_1, \dots, x_T$, only the observation x_t is interval-censored, and $a \leq x_t \leq b$, where b may be ∞. Prove the statement made in Section 2.3.5 that the likelihood of a Poisson–HMM with m states is obtained by replacing $\mathbf{P}(x_t)$ in the expression (2.12) by the $m \times m$ diagonal matrix of which the ith diagonal element is $\Pr(a \leq X_t \leq b \mid C_t = i)$.

    (b) Extend part (a) to allow for any number of interval-censored observations.