CHAPTER 7

Bayesian inference for Poisson–hidden Markov models

As alternative to the frequentist approach, one can also consider Bayesian estimation. There are several approaches to Bayesian inference in hidden Markov models: see for instance Chib (1996), Robert and Titterington (1998), Robert, Rydén and Titterington (2000), Scott (2002), Cappé *et al.* (2005) and Frühwirth-Schnatter (2006). Here we follow Scott (2002) and Congdon (2006).

Our purpose is to demonstrate an application of Bayesian inference to Poisson–HMMs. There are obstacles to be overcome, e.g. label switching and the difficulty of estimating m, the number of states, and some of these are model specific.

7.1 Applying the Gibbs sampler to Poisson–HMMs

We consider here a Poisson–HMM $\{X_t\}$ on m states, with underlying Markov chain $\{C_t\}$. We denote the state-dependent means, as usual, by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$, and the transition probability matrix of the Markov chain by $\boldsymbol{\Gamma}$.

Given a sequence of observations $x_1, x_2, \ldots, x_T$, a fixed m, and prior distributions on the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$, our objective in this section is to estimate the posterior distribution of these parameters by means of the Gibbs sampler. We shall later (in Section 7.2) drop the assumption that m is known, and also consider the Bayesian estimation thereof.

The prior distributions we assume for the parameters are of the following forms. The r th row $\boldsymbol{\Gamma}_r$ of the t.p.m. $\boldsymbol{\Gamma}$ is assumed to have a Dirichlet distribution with parameter vector $\boldsymbol{\nu}_r$, and the increment $\tau_j = \lambda_j - \lambda_{j-1}$ (with $\lambda_0 \equiv 0$) to have a gamma distribution with shape parameter a_j and rate parameter b_j. Furthermore, the rows of $\boldsymbol{\Gamma}$ and the quantities τ_j are assumed mutually independent in their prior distributions.

Our notation and terminology are as follows. Random variables $Y_1, \ldots, Y_m$ are here said to have a Dirichlet distribution with parameter vector $(\nu_1, \ldots, \nu_m)$ if their joint density is proportional to

$$y_1^{\nu_1 - 1} y_2^{\nu_2 - 1} \cdots \; y_m^{\nu_m - 1}.$$

More precisely, this expression, with y_m replaced by $1 - \sum_{i=1}^{m-1} y_i$, is (up to proportionality) the joint density of $Y_1, \ldots, Y_{m-1}$ on the unit simplex in dimension $m-1$, i.e. on the subspace of $\mathbb{R}^{m-1}$ defined by $\sum_{i=1}^{m-1} y_i \leq 1$, $y_i \geq 0$. A random variable X is said to have a gamma distribution with shape parameter a and rate parameter b if its density is (for positive x)

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}.$$

With this parametrization, X has mean a/b, variance a/b^2 and coefficient of variation (c.v.) $1/\sqrt{a}$.

If we were to observe the Markov chain, updating the transition probabilities $\mathbf{\Gamma}$ would be straightforward. Here, however, we have to generate sample paths of the Markov chain in order to update $\mathbf{\Gamma}$.

An important part of Scott's model structure, which we copy, is this. Each observed count x_t is considered to be the sum $\sum_j x_{jt}$ of contributions from up to m regimes, the contribution of regime j to x_t being x_{jt}. Note that, if the Markov chain is in state i at a given time, regimes 1 to i are *all* said to be active at that time, and regimes $i+1$ to m to be inactive. This is an unusual use of the word 'regime', but convenient here.

Instead of parametrizing the model in terms of the m state-dependent means (in our notation, the quantities λ_i), Scott parametrizes it in terms of nonnegative increments $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$, where $\tau_j = \lambda_j - \lambda_{j-1}$ (with $\lambda_0 \equiv 0$). Equivalently,

$$\lambda_i = \sum_{j=1}^{i} \tau_j.$$

This has the effect of placing the λ_js in increasing order, which is useful in order to prevent the technical problem known as label switching. For an account of this problem, see e.g. Frühwirth-Schnatter (2006, Section 3.5.5). The random variable τ_j can be described as the mean contribution of regime j, if active, to the count observed at a given time.

In outline, we proceed as follows.

- Given the observed counts $\mathbf{x}^{(T)}$ and the current values of the parameters $\mathbf{\Gamma}$ and $\boldsymbol{\lambda}$, we generate a sample path of the Markov chain.
- We use this sample path to decompose the observed counts into (simulated) regime contributions.
- With the MC sample path available, and the regime contributions, we can now update $\mathbf{\Gamma}$ and $\boldsymbol{\tau}$, hence $\boldsymbol{\lambda}$.

The above steps are repeated a large number of times and, after a 'burn-in period', the resulting samples of values of $\mathbf{\Gamma}$ and $\boldsymbol{\lambda}$ provide the required

estimates of their posterior distributions. In what follows, we use $\boldsymbol{\theta}$ to represent both $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$.

7.1.1 Generating sample paths of the Markov chain

Given the observations $\mathbf{x}^{(T)}$ and the current values of the parameters $\boldsymbol{\theta}$, we wish to simulate a sample path $\mathbf{C}^{(T)}$ of the Markov chain, from its conditional distribution

$$\Pr(\mathbf{C}^{(T)} \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}) = \Pr(C_T \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}) \times \prod_{t=1}^{T-1} \Pr(C_t \mid \mathbf{x}^{(T)}, \mathbf{C}_{t+1}^{T}, \boldsymbol{\theta}).$$

We shall be drawing values for C_T, C_{T-1}, ..., C_1, in that order, and quantities that we shall need in order to do so are the probabilities

$$\Pr(C_t \mid \mathbf{x}^{(t)}, \boldsymbol{\theta}) = \frac{\Pr(C_t, \mathbf{x}^{(t)} \mid \boldsymbol{\theta})}{\Pr(\mathbf{x}^{(t)} \mid \boldsymbol{\theta})} = \frac{\alpha_t(C_t)}{L_t} \propto \alpha_t(C_t), \text{ for } t = 1, \ldots, T. \tag{7.1}$$

As before (see p. 59), $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(m))$ denotes the vector of forward probabilities

$$\alpha_t(i) = \Pr(\mathbf{x}^{(t)}, C_t = i),$$

which can be computed from the recursion $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t)$ ($t = 2, \ldots, T$), with $\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(x_1)$; L_t is the likelihood of the first t observations.

We start the simulation by drawing C_T, the state of the Markov chain at the final time T, from $\Pr(C_T \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}) \propto \alpha_T(C_T)$, (i.e. case $t = T$ of Equation (7.1)). We then simulate the states C_t (in the order $t = T-1$, $T-2$, ..., 1) by making use of the following proportionality argument, as in Chib (1996):

$$\begin{aligned}
&\Pr(C_t \mid \mathbf{x}^{(T)}, \mathbf{C}_{t+1}^{T}, \boldsymbol{\theta}) \\
&\quad \propto \Pr(C_t \mid \mathbf{x}^{(t)}, \boldsymbol{\theta}) \Pr(\mathbf{x}_{t+1}^{T}, \mathbf{C}_{t+1}^{T} \mid \mathbf{x}^{(t)}, C_t, \boldsymbol{\theta}) \\
&\quad \propto \Pr(C_t \mid \mathbf{x}^{(t)}, \boldsymbol{\theta}) \Pr(C_{t+1} \mid C_t, \boldsymbol{\theta}) \Pr(\mathbf{x}_{t+1}^{T}, \mathbf{C}_{t+2}^{T} \mid \mathbf{x}^{(t)}, C_t, C_{t+1}, \boldsymbol{\theta}) \\
&\quad \propto \alpha_t(C_t) \Pr(C_{t+1} \mid C_t, \boldsymbol{\theta}).
\end{aligned} \tag{7.2}$$

The third factor appearing in the second-last line is independent of C_t, hence the simplification. (See Exercise 4.) The expression (7.2) is easily available, since the second factor in it is simply a one-step transition probability in the Markov chain. We are therefore in a position to simulate sample paths of the Markov chain, given observations $\mathbf{x}^{(T)}$ and parameters $\boldsymbol{\theta}$.

7.1.2 Decomposing the observed counts into regime contributions

Suppose we have a sample path $\mathbf{C}^{(T)}$ of the Markov chain, generated as described in Section 7.1.1, and suppose that $C_t = i$, so that regimes 1 to i are active at time t. Our next step is to decompose each observation x_t ($t = 1, 2, \ldots, T$) into regime contributions $x_{1t}, \ldots, x_{it}$ such that $\sum_{j=1}^{i} x_{jt} = x_t$. We therefore need the joint distribution of $X_{1t}, \ldots, X_{it}$, given $C_t = i$ and $X_t = x_t$ (and given $\boldsymbol{\theta}$). This is multinomial with total x_t and probability vector proportional to $(\tau_1, \ldots, \tau_i)$; see Exercise 1.

7.1.3 Updating the parameters

The transition probability matrix $\boldsymbol{\Gamma}$ can now be updated, i.e. new estimates produced. This we do by drawing $\boldsymbol{\Gamma}_r$, the r th row of the t.p.m. $\boldsymbol{\Gamma}$, from the Dirichlet distribution with parameter vector $\boldsymbol{\nu}_r + \mathbf{T}_r$, where $\mathbf{T}_r$ is the r th row of the (simulated) matrix of transition counts; see Section 7.1.1. (Recall that the prior for $\boldsymbol{\Gamma}_r$ is Dirichlet($\boldsymbol{\nu}_r$), and see Exercise 2.)

Similarly, the vector $\boldsymbol{\lambda}$ of state-dependent means is updated by drawing τ_j ($j = 1, \ldots, m$) from a gamma distribution with parameters $a_j + \sum_{t=1}^{T} x_{jt}$ and $b_j + N_j$; here N_j denotes the number of times regime j was active in the simulated sample path of the Markov chain, and x_{jt} the contribution of regime j to x_t. (Recall that the prior for τ_j is a gamma distribution with shape parameter a_j and rate parameter b_j, and see Exercise 3.)

7.2 Bayesian estimation of the number of states

In the Bayesian approach to model selection, the number of states, m, is a parameter whose value is assessed from its posterior distribution, $p(m \mid \mathbf{x}^{(T)})$. Computing this posterior distribution is, however, not an easy problem; indeed it has been described as 'notoriously difficult to calculate' (Scott, James and Sugar, 2005).

Using p as a general symbol for probability mass or density functions, one has

$$p(m \mid \mathbf{x}^{(T)}) = p(m)\, p(\mathbf{x}^{(T)} \mid m)/p(\mathbf{x}^{(T)}) \propto p(m)\, p(\mathbf{x}^{(T)} \mid m), \qquad (7.3)$$

where $p(\mathbf{x}^{(T)} \mid m)$ is called the integrated likelihood. If only two models are being compared, the posterior odds are equal to the product of the 'Bayes factor' and the prior odds:

$$\frac{p(m_2 \mid \mathbf{x}^{(T)})}{p(m_1 \mid \mathbf{x}^{(T)})} = \frac{p(\mathbf{x}^{(T)} \mid m_2)}{p(\mathbf{x}^{(T)} \mid m_1)} \times \frac{p(m_2)}{p(m_1)}. \qquad (7.4)$$

7.2.1 Use of the integrated likelihood

In order to use (7.3) or (7.4) we need to estimate the integrated likelihood

$$p(\mathbf{x}^{(T)} \mid m) = \int p(\boldsymbol{\theta}_m, \mathbf{x}^{(T)} \mid m)\, \mathrm{d}\boldsymbol{\theta}_m = \int p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m)\, p(\boldsymbol{\theta}_m \mid m)\, \mathrm{d}\boldsymbol{\theta}_m.$$

One way of doing so would be to simulate from $p(\boldsymbol{\theta}_m \mid m)$, the prior distribution of the parameters $\boldsymbol{\theta}_m$ of the m-state model. But it is convenient and — especially if the prior is diffuse — more efficient to use a method that requires instead a sample from the posterior distribution, $p(\boldsymbol{\theta}_m \mid \mathbf{x}^{(T)}, m)$. Such a method is as follows.

Write the integrated likelihood as

$$\int p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m)\, \frac{p(\boldsymbol{\theta}_m \mid m)}{p^*(\boldsymbol{\theta}_m)}\, p^*(\boldsymbol{\theta}_m)\, \mathrm{d}\boldsymbol{\theta}_m;$$

i.e. write it in a form suitable for the use of a sample from some convenient density $p^*(\boldsymbol{\theta}_m)$ for the parameters $\boldsymbol{\theta}_m$. Since we have available a sample $\boldsymbol{\theta}_k^{(j)}$ $(j = 1, 2, \ldots, B)$ from the posterior distribution, we can use that sample; i.e. we can take $p^*(\boldsymbol{\theta}_m) = p(\boldsymbol{\theta}_m \mid \mathbf{x}^{(T)}, m)$. Newton and Raftery (1994) therefore suggest *inter alia* that the integrated likelihood can be estimated by

$$\hat{I} = \sum_{j=1}^{B} w_j p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m^{(j)}) \Big/ \sum_{j=1}^{B} w_j, \tag{7.5}$$

where

$$w_j = \frac{p(\boldsymbol{\theta}_m^{(j)} \mid m)}{p(\boldsymbol{\theta}_m^{(j)} \mid \mathbf{x}^{(T)}, m)}.$$

After some manipulation this simplifies to the harmonic mean of the likelihood values of a sample from the posterior:

$$\hat{I} = \left(B^{-1} \sum_{j=1}^{B} p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m^{(j)}) \right)^{-1}; \tag{7.6}$$

see Exercise 5 for the details. This is, however, not the only route one can follow in deriving the estimator (7.6); see Exercise 6 for another possibility.

Newton and Raftery state that, under quite general conditions, $\hat{I}$ is a simulation-consistent estimator of $p(\mathbf{x}^{(T)} \mid m)$. But there is a major drawback to this harmonic mean estimator, its infinite variance, and the question of which estimator to use for $p(\mathbf{x}^{(T)} \mid m)$ does not seem to have been settled. Raftery *et al.* (2007) suggest two alternatives to the harmonic mean estimator, but no clear recommendation emerges, and one of the discussants of that paper (Draper, 2007) bemoans the

disheartening 'ad-hockery' of the many proposals that have over the years been made for coping with the instability of expectations with respect to (often diffuse) priors.

7.2.2 Model selection by parallel sampling

However, it is possible to estimate $p(m \mid \mathbf{x}^{(T)})$ relatively simply by 'parallel sampling' of the competing models, provided that the set of competing models is sufficiently small; see Congdon (2006) and Scott (2002). Denote by $\boldsymbol{\theta}$ the vector $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K)$, and similarly $\boldsymbol{\theta}^{(j)}$; K is the maximum number of states. Make the assumption that

$$p(m, \boldsymbol{\theta}) = p(\boldsymbol{\theta}_m \mid m)\, p(m);$$

that is, assume that model m is (for $j \neq m$) indifferent to values taken by $\boldsymbol{\theta}_j$.

We wish to estimate $p(m \mid \mathbf{x}^{(T)})$ (for $m = 1, \ldots, K$) by

$$B^{-1} \sum_{j=1}^{B} p(m \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}^{(j)}). \tag{7.7}$$

We use the fact that, with the above assumption,

$$p(m \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}^{(j)}) \propto G_m^{(j)},$$

where

$$G_m^{(j)} \equiv p(\mathbf{x}^{(T)} \mid \boldsymbol{\theta}_m^{(j)}, m)\, p(\boldsymbol{\theta}_m^{(j)} \mid m)\, p(m). \tag{7.8}$$

(See Appendix 1 of Congdon (2006).) Hence

$$p(m \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}^{(j)}) = G_m^{(j)} / \sum_{k=1}^{K} G_k^{(j)}.$$

This expression for $p(m \mid \mathbf{x}^{(T)}, \boldsymbol{\theta}^{(j)})$ can then be inserted in (7.7) to complete the estimate of $p(m \mid \mathbf{x}^{(T)})$.

7.3 Example: earthquakes

We apply the techniques described above to the series of annual counts of major earthquakes. The prior distributions used are as follows. The gamma distributions used as priors for the λ-increments (i.e. for the quantities τ_j) all have mean $50m/(m+1)$ and c.v. 1 in one analysis, and 2 in a second. The Dirichlet distributions used as priors for the rows of $\boldsymbol{\Gamma}$ all have all parameters equal to 1. The prior distribution for m, the number of states, assigns probability $\frac{1}{6}$ to each of the values 1, 2, ..., 6. The number of iterations used was $B = 100\,000$, with a burn-in period of 5000.

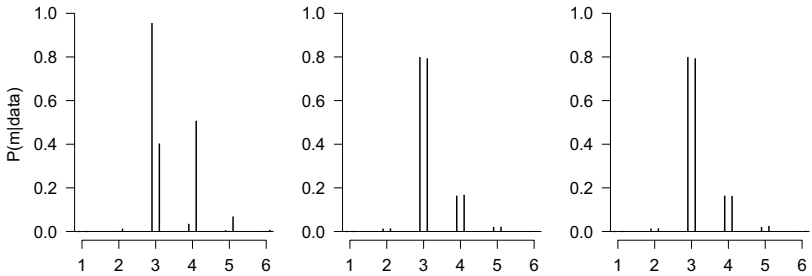

Figure 7.1 *Earthquakes data, posterior distributions of m given a uniform prior on $\{1, 2, \ldots, 6\}$. Each panel shows two posterior distributions from independent runs. In the left and centre panels the c.v. in the gamma prior is 1. Left panel: harmonic mean estimator. Centre panel: parallel sampling estimator. Right panel: parallel sampling estimator with c.v. = 1 (left bars) and c.v. = 2 (right bars).*

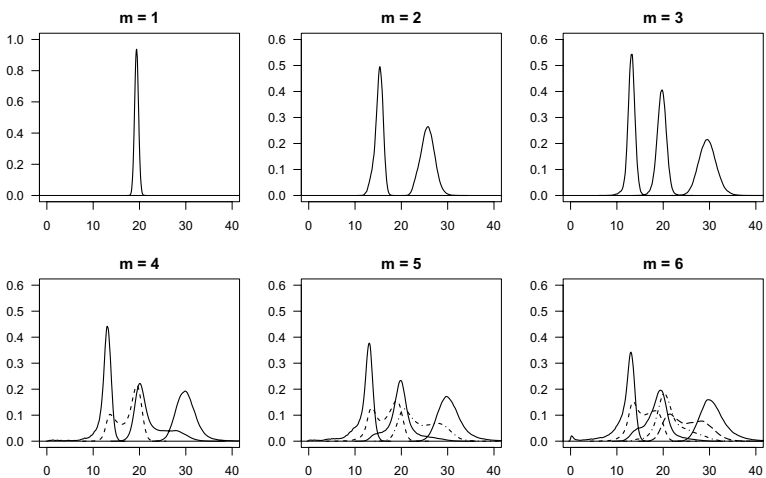

Figure 7.2 *Earthquakes data: posterior distributions of the state-dependent means for one- to six-state Poisson–HMMs.*

Figure 7.1 displays three comparisons of estimates of the posterior distribution of m. It is clear from the comparison of two independent runs of the harmonic mean estimator (left panel) that it is indeed very unstable. In the first run it would have chosen a three-state model by a large margin, and in the second run a four-state model. In contrast, the parallel sampling estimator (centre and right panels) produces very

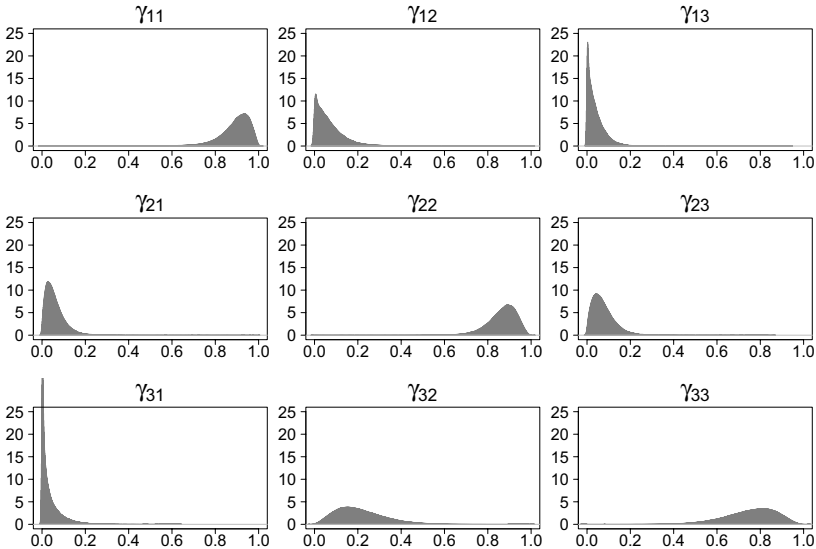

Figure 7.3 *Three-state Poisson–HMM for earthquakes data: posterior distribution of transition probability matrix* $\mathbf{\Gamma}$.

consistent results even if the c.v. of the prior distributions for the λ-increments is changed from 1 to 2, and clearly identifies $m = 3$ as the posterior mode.

The posterior distributions of the Poisson means for $m = 1, \ldots, 6$ are displayed in Figure 7.2. The posterior distributions of the entries of $\mathbf{\Gamma}$ for the three-state model are displayed in Figure 7.3. Table 7.1 lists posterior statistics for the three-state model. The posterior modes are generally quite close to the maximum likelihood estimates given on p. 51. In particular the values $\boldsymbol{\lambda}$ are almost the same, but the posterior modes for the entries of $\mathbf{\Gamma}$ are mostly a little closer to 0.5 than are the corresponding MLEs.

7.4 Discussion

From the above example it seems clear that the Bayesian approach is demanding computationally and in certain other respects. The model needs to be parametrized in a way that avoids label switching. In HMMs the labels associated with the states are arbitrary; the model is invariant under permutation of the labels. This is irrelevant to maximum likelihood estimation, but it is a problem in the context of MCMC estimation of posterior distributions. One must ensure that only one of the $m!$ per-

Table 7.1 *Earthquakes data: posterior statistics for the three-state model.*

parameter	min	Q1	mode	median	mean	Q3	max
λ_1	6.21	12.62	13.12	13.15	13.12	13.68	16.85
λ_2	13.53	19.05	19.79	19.74	19.71	20.42	27.12
λ_3	22.08	28.33	29.87	29.59	29.64	30.88	43.88
γ_{11}	0.001	0.803	0.882	0.861	0.843	0.905	0.998
γ_{12}	0.000	0.047	0.056	0.085	0.104	0.139	0.964
γ_{13}	0.000	0.020	0.011	0.042	0.053	0.075	0.848
γ_{21}	0.000	0.043	0.050	0.070	0.083	0.108	0.979
γ_{22}	0.009	0.784	0.858	0.837	0.824	0.880	0.992
γ_{23}	0.000	0.052	0.060	0.082	0.093	0.122	0.943
γ_{31}	0.000	0.021	0.011	0.049	0.068	0.096	0.758
γ_{32}	0.000	0.144	0.180	0.213	0.229	0.296	0.918
γ_{33}	0.010	0.627	0.757	0.718	0.703	0.795	0.986

mutations of the labels is used in the simulation. In the algorithm for Poisson–HMMs discussed above, label switching was avoided by ordering the states according to the means of the state-dependent distributions. An analogous reparametrization has been used for the normal case by Robert and Titterington (1998). Prior distributions need to be specified for the parameters and, as a rule, the choice of prior distribution is driven by mathematical convenience rather than by prior information. The above difficulties are model specific; the derivations, priors and hence the computer code all change substantially if the state-dependent distribution changes.

Although the computational demands of MCMC are high in comparison with those of ML *point estimation*, this is not a fair comparison. Interval estimation using the parametric bootstrap, though easy to implement (see Section 3.6.2), is comparably time-consuming. Interval estimates based instead on approximate standard errors obtained from the Hessian (as in Section 3.6.1) require potentially demanding parametrization-specific derivations.

A warning note regarding MCMC has been sounded by Celeux, Hurn and Robert (2000), and echoed by Chopin (2007):

> [. . .] we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!

This statement was made of independent mixture models but is presumably also applicable to HMMs.

Another Bayesian approach to model selection in HMMs is the use of

reversible jump Markov chain Monte Carlo techniques (RJMCMC), in which m, the number of states, is treated as a parameter of the model and, like the other parameters, updated in each iteration. This has the advantage over parallel sampling that relatively few iterations are 'wasted' on unpromising values of m, thereby reducing the total number of iterations needed to achieve a given degree of accuracy. Such an advantage might be telling for very long time series and a potentially large number of states, but for typical applications and sample sizes, models with $m > 5$ are rarely feasible. The disadvantage for users who have to write their own code is the complexity of the algorithm.

Robert, Rydén and Titterington (2000) describe in detail the use of that approach in HMMs with normal distributions as the state-dependent distributions, and provide several examples of its application. As far as we are aware, this approach has not been extended to HMMs with other possible state-dependent distributions, e.g. Poisson, and we refer the reader interested in RJMCMC to Robert *et al.* and the references therein, especially Green (1995) and Richardson and Green (1997).

Exercises

1. Consider u defined by $u = \sum_{j=1}^{i} u_j$, where the variables u_j are independent Poisson random variables with means τ_j.

   Show that, conditional on u, the joint distribution of $u_1, u_2, \ldots, u_i$ is multinomial with total u and probability vector $(\tau_1, \ldots, \tau_i)/\sum_{j=1}^{i} \tau_j$.

2. (Updating of Dirichlet distributions) Let $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ be an observation from a multinomial distribution with probability vector $\mathbf{y}$, which has a Dirichlet distribution with parameter vector $\mathbf{d} = (d_1, d_2, \ldots, d_m)$.

   Show that the posterior distribution of $\mathbf{y}$, i.e. the distribution of $\mathbf{y}$ given $\mathbf{w}$, is the Dirichlet distribution with parameters $\mathbf{d} + \mathbf{w}$.

3. (Updating of gamma distributions) Let $y_1, y_2, \ldots, y_n$ be a random sample from the Poisson distribution with mean τ, which is gamma-distributed with parameters a and b.

   Show that the posterior distribution of τ, i.e. the distribution of τ given $y_1, \ldots, y_n$, is the gamma distribution with parameters $a + \sum_{i=1}^{n} y_i$ and $b + n$.

4. Show that, in the basic HMM,

$$\Pr(\mathbf{X}_{t+1}^{T}, \mathbf{C}_{t+2}^{T} \mid \mathbf{X}^{(t)}, C_t, C_{t+1}) = \Pr(\mathbf{X}_{t+1}^{T}, \mathbf{C}_{t+2}^{T} \mid C_{t+1}).$$

   (Hint: either use the methods of Appendix B or invoke d-separation, for which see e.g. Pearl (2000), pp. 16–18.)

5. Consider the estimator $\hat{I}$ as defined by Equation (7.5):

$$\hat{I} = \sum_j w_j \, p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m^{(j)}) \Big/ \sum_j w_j .$$

   (a) Show that the weight $w_j = p(\boldsymbol{\theta}_m^{(j)} \mid m)/p(\boldsymbol{\theta}_m^{(j)} \mid \mathbf{x}^{(T)}, m)$ can be written as $p(\mathbf{x}^{(T)} \mid m)/p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m^{(j)})$.

   (b) Deduce that the summand $w_j \, p(\mathbf{x}^{(T)} \mid m, \boldsymbol{\theta}_m^{(j)})$ is equal to $p(\mathbf{x}^{(T)} \mid m)$ (and is therefore independent of j).

   (c) Hence show that $\hat{I}$ is the harmonic mean displayed in Equation (7.6).

6. Let the observations $\mathbf{x}$ be distributed with parameter (vector) $\boldsymbol{\theta}$. Prove that

$$\frac{1}{p(\mathbf{x})} = \mathrm{E}\left(\frac{1}{p(\mathbf{x} \mid \boldsymbol{\theta})} \;\middle|\; \mathbf{x} \right);$$

   i.e. the integrated likelihood $p(\mathbf{x})$ is the harmonic mean of the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ computed under the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$ for $\boldsymbol{\theta}$. (This is the 'harmonic mean identity', as in Equation (1) of Raftery *et al.* (2007), and suggests the use of the harmonic mean of likelihoods sampled from the posterior as estimator of the integrated likelihood.)