

HMM III: Some examples of HMMs

In practice, we usually do not know the parameters of the HMM and they must be estimated from data. To do this, we find the joint distribution of two consecutive states conditional on the observation sequence \mathbf{A} . Note that

$$\begin{aligned}
 \xi_t(i, j) &= \mathbb{P}(X_t = i, X_{t+1} = j \mid \mathbf{A}) \\
 &\propto \mathbb{P}(X_t = i, X_{t+1} = j, A_1, \dots, A_n) \\
 &= \mathbb{P}(X_{t+1} = j, A_{t+1}, \dots, A_n \mid X_t = i, A_t, \dots, A_1) \mathbb{P}(X_t = i, A_1, \dots, A_t) \\
 &= \alpha_t(i) \mathbb{P}(X_{t+1} = j, A_{t+1}, \dots, A_n \mid X_t = i) \\
 &= \alpha_t(i) \mathbb{P}(X_{t+1} = j, A_{t+1} \mid X_t = i) \mathbb{P}(A_{t+2}, \dots, A_n \mid X_{t+1} = j) \\
 &= \alpha_t(i) p(i, j) e(j, A_{t+1}) \beta_{t+1}(j).
 \end{aligned}$$

Thus

$$\xi_t(i, j) = \frac{\alpha_t(i) p(i, j) e(j, A_{t+1}) \beta_{t+1}(j)}{\sum_i \sum_j \alpha_t(i) p(i, j) e(j, A_{t+1}) \beta_{t+1}(j)}. \quad (1)$$

Next, note that

$$n(r, s) = \sum_{t=1}^{n-1} \xi_t(r, s) \quad (2)$$

gives the expected number of times the hidden chain moves from r to s conditional on \mathbf{A} , and that

$$n(r, +) = \sum_{s \in \mathcal{S}} n(r, s) = \sum_{t=1}^{n-1} \gamma_t(r)$$

gives the expected number of times the hidden chain moves from state r conditional on \mathbf{A} . To estimate the emission probabilities, we also need

$$m(j, a) = \sum_{t=1}^n \gamma_t(j) \mathbb{1}(A_t = a) \quad (3)$$

where $\mathbb{1}(A_t = a) = 1$ if $A_t = a$ and 0 otherwise. $m(j, a)$ is the expected number of times state j is visited and letter a is emitted. We also need

$$m(j, +) = \sum_{a \in \mathcal{A}} m(j, a) = \sum_{t=1}^n \gamma_t(j).$$

EM algorithm.

The idea of the EM algorithm (EM stands for Expectation-Maximization) is to assume you know the parameters, then re-estimate them via the updating formulae given below. This process is repeated until convergence. So suppose we have a current estimate $P^{(l)}$ of the transition matrix P and $e^{(l)}$ of the emission probabilities e . Use these parameters to compute the quantities in (2) and (3), and then update via

$$P^{(l+1)}(i, j) = \frac{n(i, j)}{n(i, +)}, \quad (4)$$

$$e^{(l+1)}(i, a) = \frac{m(i, a)}{m(i, +)}. \quad (5)$$

It can be shown that this iterative scheme converges to a critical point of the likelihood function (which is just $\mathbb{P}(\mathbf{A})$ viewed as a function of the parameters P and e .) The likelihood surface can be very complex, and it has many local maxima (we found some of them in our computer lab examples!).

Examples.

In these examples, taken from Churchill (1989), the underlying hidden chain has 2 states, $\mathcal{S} = \{0, 1\}$, with transition matrix

$$P = \begin{pmatrix} 1 - \lambda & \lambda \\ \tau & 1 - \tau \end{pmatrix}$$

The alphabet is $\mathcal{A} = \{0, 1\}$ and $e(0, 1) = p_0, e(1, 1) = p_1$.

When the chain starts in state 0 and $\tau = 0$ the model corresponds to a change point problem; the chain waits in state 0 for a while then switches to state 1 and stays there. In that case, we would calculate the posterior probability of a change at time t as

$$\mathbb{P}(X_t = 0, X_{t+1} = 1 \mid \mathbf{A}),$$

and the probability of no change point is

$$\mathbb{P}(X_n = 0 \mid \mathbf{A}).$$

The human X-chromosome fragment Xrep is 2352bp was isolated because it stimulates replication in bacterial plasmids. It contains features similar to

those found in eukaryotic viral origins of replication. The overall $G + C$ content is 57.8%, with marked heterogeneity across the region. Local $G + C$ content is around 50% in the 5' end of the fragment, and near 70% near the 3' end; the change occurs somewhere between base 1500 and 1800. In this case he coded the sequence as $\mathcal{A} = \{AT, GC\}$, and set $p_0 = 0.5, p_1 = 0.7, \lambda = 0.0005$. Plots of the smoothed estimate of the state process are given in Figure 2 below. Note the sharp transition in the region of base 1600. The posterior density of the change point has a mode at base 1590, and a 90% maximum posterior density region from base 1571 to 1632.

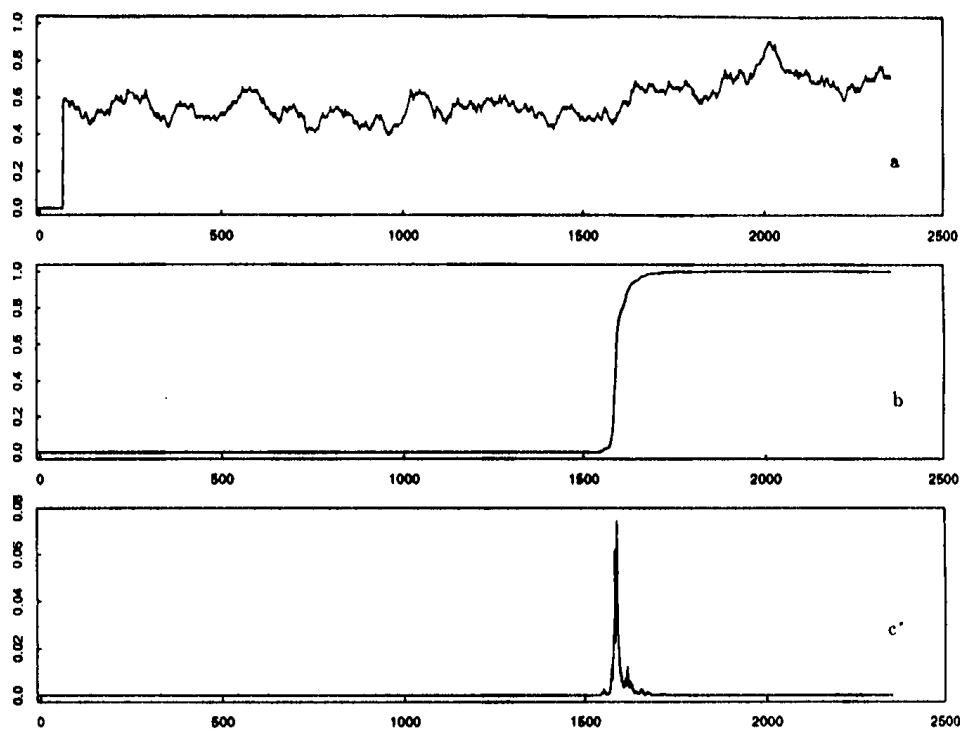


Figure 2. Xrep change-point model. The GC content averaged over a window of size 64 (a); smoothed estimates of the hydrogen bonding process (b); the posterior density of the change point (c); are plotted against the sequence index.

The second example concerns the purine-pyrimidine sequence of mammalian mitochondrial DNAs. The alphabet is $\mathcal{A} = \{AG, CT\}$, the states being as in the earlier example. The data comprises the L-strand of the mtDNA for both mouse and man. In this case the algorithms are modified slightly to account for the fact that the molecules are circular. The estimated

parameters for human sequence were

$$\hat{p}_0 = 0.425, \hat{p}_1 = 0.525, \hat{\lambda} = 0.000115, \hat{\tau} = 0.000610,$$

whereas for mouse he got

$$\hat{p}_0 = 0.454, \hat{p}_1 = 0.541, \hat{\lambda} = 0.000090, \hat{\tau} = 0.000434.$$

Plots of $\mathbb{P}(X_t = 0 \mid \mathbf{A})$ are given in Figure 3 below.

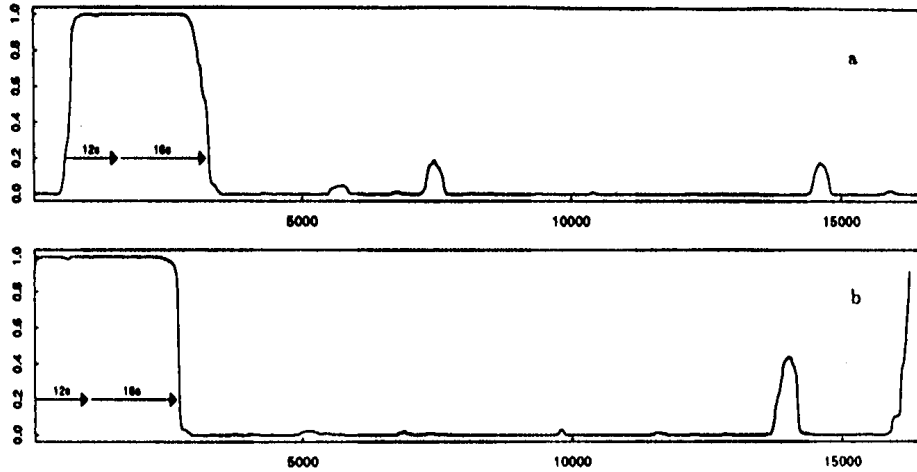


Figure 3. Mammalian mitochondrial DNA. The two-state smoothed estimates from maximum likelihood fits to the purine-pyrimidine processes of human (a) and mouse (b) mtDNA are plotted against the sequence index. Ribosomal RNA coding sequences are indicated by arrows.

Extensions.

The first modification we make is to allow sequences generated by the HMM to have random lengths. To do this, we imagine a start state, a series of backbone states M_1, \dots, M_N , insert states I_1, \dots, I_{N+1} , and delete states D_1, \dots, D_N , and an end state. The chain starts in the start state, moves through a series of I, M, D states, then ends. Symbols are emitted from any of the I or M states, but no symbols are emitted from a D state. The typical structure of the model is shown in the figure below (taken from P147 of Baldi and Brunak, 1998). This model can be used to describe the structure of a collection of motifs (in either DNA or amino-acid alphabets). These motifs do not have to be the same length (hence the delete states). The parameter N is usually taken to be the average length of the motifs.

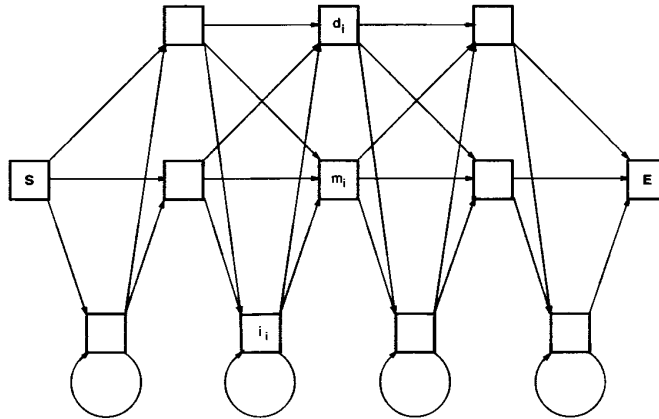


Figure 7.2: The Standard HMM Architecture. S is the start state, E is the end state, and d_i, m_i , and i_i denote delete, main, and insert states, respectively.

The model has $9N + 3$ transition probability parameters, and $(2N + 1)|\mathcal{A}|$ emission parameters. This is the architecture of one of the models analyzed in the HMMpro software we will use in the Computer lab. Our illustration will be an application to finding DNA-binding sites in *E. coli*.

Changing the emission laws

In the examples we have looked at so far, the emission model is independent letters, conditional on the hidden sequence \mathbf{X} . This can be generalized in a number of ways. In particular we can have Markov emissions, in the sense that

$$\mathbb{P}(A_{t+1} = b | X_{t+1} = j, A_t = a, X_t = i) = e(j; a, b)$$

where for each j , $e(j; a, b), a, b \in \mathcal{A}$ is a transition matrix. The forward-backward equations and the Viterbi algorithm can readily be modified to deal with this modification. See Churchill (1989) for example.

Changing the hidden chain.

It is also possible to change the model in such a way that the underlying hidden chain is a semi-Markov chain. In this case the chain can have durations in states that are not geometrically distributed (as they are for a regular Markov chain). The basics are described in Rabiner (1989). They are used in the world of gene finding; see Burge and Karlin (1997) and Lukashin and Borodovsky (1998) for example.

References

- Baldi P, Brunak S (1998) *Bioinformatics. The machine learning approach*. MIT Press, Chapter 8.
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78–94.
- Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**: 79–94.
- Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research* **26**: 1107–1115.