In what follows [Dav] refers to Davison's book.

1. For the six-parameter model given at the end of Example 6.8 in [Dav] (p. 237) and the data in Table 6.2 (p. 227), write down the log likelihood function, and calculate the maximum likelihood estimates and the maximized log likelihood. Compare the AIC value of this model with those of the zeroth-order and first-order Markov models based on the same data.

Solution: For the six-parameter model, denote

$$\theta_{1\beta} = p_{\alpha\beta}, \quad \text{where} \ \ \alpha \in \{\mathtt{A},\mathtt{G},\mathtt{T}\}, \ \ \beta \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\},$$

and

$$\theta_{2\beta} = p_{\mathtt{C}\beta}, \quad \text{where} \ \ \beta \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}.$$

For each $\beta \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$, we obtain from Table 6.2 the following counts $n_{1\beta}$ of all pairs of bases which start from some base in $\{\mathtt{A},\mathtt{G},\mathtt{T}\}$ and end at base β:

$$n_{1\mathtt{A}} = 516 - 101 = 415, \quad n_{1\mathtt{C}} = 263 - 41 = 222, \quad n_{1\mathtt{G}} = 226 - 6 = 220, \quad n_{1\mathtt{T}} = 566 - 115 = 451.$$

The corresponding frequency of β as the second base among all pairs that start from some base in $\{\mathtt{A},\mathtt{G},\mathtt{T}\}$ is given for each β by

$$q_{1\mathtt{A}} = 0.317, \quad q_{1\mathtt{C}} = 0.170, \quad q_{1\mathtt{G}} = 0.168, \quad q_{1\mathtt{T}} = 0.345.$$

For pairs of the form $(\mathtt{C},\beta)$, Table 6.2 gives the counts $n_{\mathtt{C}\beta}$ and frequencies $q_{\mathtt{C}\beta}$ as

$$n_{\mathtt{CA}} = 101, \quad n_{\mathtt{CC}} = 41, \quad n_{\mathtt{CG}} = 6, \quad n_{\mathtt{CT}} = 115,$$

$$q_{\mathtt{CA}} = 0.384, \quad q_{\mathtt{CC}} = 0.156, \quad q_{\mathtt{CG}} = 0.023, \quad q_{\mathtt{CT}} = 0.437.$$

Let θ denote the collection of $\theta_{1\beta}, \theta_{2\beta}, \beta \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. The log likelihood function is

$$\begin{aligned}\ell(\theta) &= \sum_{\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} n_{1\beta} \ln\theta_{1\beta} + \sum_{\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} n_{\mathtt{C}\beta} \ln\theta_{2\beta} \\ &= 1308 \sum_{\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} q_{1\beta} \ln\theta_{1\beta} + 263 \sum_{\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} q_{\mathtt{C}\beta} \ln\theta_{2\beta}.\end{aligned}$$

By the information inequality, the expression in the first term, $\sum_{\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} q_{1\beta}\ln\theta_{1\beta}$, is maximized by taking $\theta_{1\beta} = q_{1\beta}$, and similarly for the second term. So the maximum likelihood estimates are

$$\hat\theta_{1\beta} = q_{1\beta}, \quad \hat\theta_{2\beta} = q_{\mathtt{C}\beta}, \quad \beta \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}.$$

The maximized log likelihood is

$$\ell(\hat\theta) = -2033.4.$$

Since this model has 6 free parameters, its AIC value is

$$\text{AIC} = 2(-\ell(\hat\theta) + 6) = 4078.9.$$

The maximized log likelihoods of the zero-th order and first-order Markov chains are calculated similarly. The zero-th order Markov chain has 3 free parameters, and its maximized log likelihood and AIC value are

$$\ell(\hat\theta) = -2060.6, \quad \text{AIC} = 2(-\ell(\hat\theta) + 3) = 4127.1.$$

The first-order Markov chain has 12 free parameters, and its maximized log likelihood and AIC value are

$$\ell(\hat\theta) = -2028.3, \quad \text{AIC} = 2(-\ell(\hat\theta) + 12) = 4080.7. \quad \square$$

2. Do Exercise 6.12 in [Dav] (p. 244) with "stationary distribution π" replaced by "initial distribution μ."
(About this exercise: the notation $p_{rs}(n)$ refers to the n-step transition probability from state r to s; notice the hint on the margin "Look carefully at the data.")

Solution: As we have shown in the class, being the subsequence of a Markov chain at times $0, t_1, \ldots, t_k$, the sequence of variables $S_0, S_1, \ldots, S_k$ is also a Markov chain. The likelihood is

$$L = p(s_0, s_1, \ldots, s_k) = p(s_0) \cdot \prod_{j=1}^{k} p(s_j \,|\, s_{j-1}) = p(s_0) p_{s_{j-1} s_j}(t_j - t_{j-1}).$$

In the case of a sequence of states '12311' observed at times $0, 1, 3, 4, 6$,

$$L = \mu(1) \cdot p_{12} \cdot p_{23}(2) \cdot p_{31} \cdot p_{11}(2).$$

For Example 6.2 with the data in Table 6.3, there are 37 independent Markov chains, one for each patient. Corresponding to the ith patient, we can calculate the likelihood L_i using the above formula. As the Markov chains are independent of each other, we multiply L_i together to obtain the likelihood L:

$$L = \prod_{i=1}^{37} L_i.$$

In this case $L = 0$ because $L_{19} = 0$: we have $p_{13} = 0$ according to our model, and the state sequence of the 19th patient is '113' observed at times $0, 3, 6$ months, so $L_i = \mu(1) \cdot p_{11} \cdot p_{13} = 0$. This shows that the model would be more plausible if $p_{13} > 0$. □

3. Suppose Θ is our model for a discrete random variable Y_0 whose true distribution is Q^* and PMF q^*. Let

$$\theta^* \in \operatorname*{arg\,min}_{\theta \in \Theta} \mathrm{KL}\big(q^*, p_\theta\big) \neq \emptyset,$$

where p_θ denotes the PMF of the distribution associated with θ. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample of size n from the distribution Q^*. For each observation $y = (y_1, y_2, \ldots, y_n)$ of Y, let Q_y be the empirical distribution of Y_0, i.e., $Q_y(Y_0 = i)$ equals the frequency of i in y, and let q_y denote the corresponding PMF. Let $\hat{\theta}(y)$ be the maximum likelihood estimate of θ based on the data y (assume $\hat{\theta}(y)$ exists). Show that for all possible observations y,

$$\mathrm{KL}\big(q_y, p_{\hat{\theta}(y)}\big) \leq \mathrm{KL}\big(q_y, p_{\theta^*}\big).$$

Discuss what this means in practice.

Solution: The likelihood for θ based on y is

$$L(\theta; y) = \prod_{k=1}^{n} P(Y_0 = y_k; \theta) = \prod_{k=1}^{n} p_\theta(y_k).$$

Assume that Y_0 takes m possible values, $1, 2, \ldots, m$. Then, we can write the log likelihood as

$$\ell(\theta; y) = \sum_{k=1}^{n} \ln p_\theta(y_k) = \sum_{k=1}^{n} \sum_{i=1}^{m} I(y_k = i) \cdot \ln p_\theta(y_k),$$

where $I(\cdots)$ denotes the indicator function:

$$I(y_k = i) = 1 \quad \text{if} \;\; y_k = i; \quad I(y_k = i) = 0 \quad \text{otherwise.}$$

It follows then

$$\begin{aligned} \frac{1}{n}\ell(\theta; y) &= \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{m} I(y_k = i) \cdot \ln p_\theta(i) \\ &= \sum_{i=1}^{m} \frac{1}{n} \Big(\sum_{k=1}^{n} I(y_k = i) \Big) \cdot \ln p_\theta(i) \\ &= \sum_{i=1}^{m} q_y(i) \cdot \ln p_\theta(i). \end{aligned} \tag{1}$$

We can relate $\ell(\theta; y)$ to the KL-divergence between q_y and p_θ: by the definition of KL-divergence,

$$\mathrm{KL}(q_y, p_\theta) = \sum_{i=1}^{m} q_y(i) \cdot \ln\left(\frac{q_y(i)}{p_\theta(i)}\right) = \sum_{i=1}^{m} q_y(i) \cdot \ln q_y(i) - \sum_{i=1}^{m} q_y(i) \cdot \ln p_\theta(i),$$

where the first term is minus the entropy of Y_0, $-H(Y_0)$, and it does not depend on θ. So Eq. (1) is equivalent to

$$\frac{1}{n}\ell(\theta; y) = -\mathrm{KL}(q_y, p_\theta) - H(Y_0). \tag{2}$$

Since $\hat{\theta}(y)$ is the maximum likelihood estimate of θ,

$$\ell(\hat{\theta}(y); y) \geq \ell(\theta; y), \quad \forall \theta \in \Theta,$$

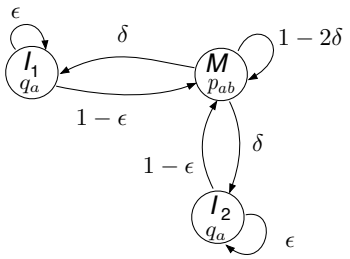
and by Eq. (2) this is equivalent to

$$\mathrm{KL}\big(q_y, p_{\hat{\theta}(y)}\big) = \min_{\theta \in \Theta} \mathrm{KL}\big(q_y, p_{\hat{\theta}(y)}\big) \leq \mathrm{KL}\big(q_y, p_{\theta^*}\big).$$

Contrasting the above inequality with the definition of θ^*, we see that if the empirical distribution q_y is a poor approximation of the true distribution q^*, (which can happen when the sample size n is too small relatively to m and the dimension of θ), then $\hat{\theta}(y)$ can be far away from the best distribution θ^* in the model. And furthermore, from Eq. (2), we see that the maximized log likelihood $\ell(\hat{\theta}; y)$ can be unreliable for assessing the fitness of the model. □

4. HMM is widely applied in biological sequence alignment. For aligning two sequences with possibly different lengths, the model is often described concisely in terms of a probabilistic finite state automaton (PFSA) like the one shown below.

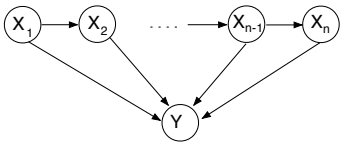
   The PFSA has three states: state M, corresponding to a match, and states I_1, I_2, corresponding to inserting a gap in the first and second sequence, respectively. State M emits an aligned pair of symbols 'a:b' with probability p_{ab}; state I_1 emits a symbol 'a' against a gap with probability q_a; and similarly, state I_2 emits a gap against a symbol 'a' with probability q_a. Possible transitions between the three states are indicated by the arcs with the corresponding probabilities. For example, with '-' representing a gap, the PFSA can generate an aligned pair of sequences




```
ACG-
A-GT
```

   For our alignment problem, we assume that the pairs of sequences to be aligned are generated by this PFSA. What are the latent and observable random variables in this model? Specify the form of their joint distribution and draw the corresponding graphical model.

Solution: We can define latent random variables $X = \{X_i\}$ and observable random variables Y as follows. We let $Y = (Y_1, Y_2)$, where Y_1, Y_2 corresponds to the first and the second sequence, respectively, in the alignment problem. We let X_i correspond to the state and the symbols emitted by the PFSA at time i, in particular, $x_i = (s_{\mathrm{PFSA},i}, (\alpha_i : \beta_i))$, where $s_{\mathrm{PFSA},i}$ is the state that the PFSA is in and $(\alpha_i : \beta_i)$ the pair of symbols the PFSA emits at time i. The graphical model is



The length n is also a random variable. The sequence $X_1, \ldots, X_n$ is a Markov chain whose transitions are described by the PFSA model. The relation between Y and X is deterministic: if $X = x$, then Y_1 is the sequence $\alpha_1 \ldots \alpha_n$ with the gap symbols removed, and Y_2 is the sequence $\beta_1 \ldots \beta_n$ with the gap symbols removed. □

5*. Prove Eq. (6.13) in [Dav] (p. 246) (i.e., Besag's theorem in the slides of Lecture 4) under the positivity condition.
Hint: for any two possible values $(y_1, y_2, \ldots, y_n)$ and $(x_1, x_2, \ldots, x_n)$ of $(Y_1, Y_2, \ldots, Y_n)$, use the identity

$$p(y_1, \ldots, y_n) = p(y_n \,|\, y_1, \ldots, y_{n-1})\, p(y_1, \ldots, y_{n-1})$$

to show that

$$p(y_1, \ldots, y_n) = \frac{p(y_n \,|\, y_1, \ldots, y_{n-1})}{p(x_n \,|\, y_1, \ldots, y_{n-1})}\, p(y_1, \ldots, y_{n-1}, x_n),$$

and then that

$$p(y_1, \ldots, y_{n-1}, x_n) = \frac{p(y_{n-1} \,|\, y_1, \ldots, y_{n-2}, x_n)}{p(x_{n-1} \,|\, y_1, \ldots, y_{n-2}, x_n)}\, p(y_1, \ldots, y_{n-2}, x_{n-1}, x_n),$$

and so on.

Solution: We want to prove Eq. (6.13), which states that for any two possible values $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ of $(Y_1, \ldots, Y_n)$,

$$\frac{p(y_1, \ldots, y_n)}{p(x_1, \ldots, x_n)} = \prod_{i=1}^{n} \frac{p(y_i \,|\, y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n)}{p(x_i \,|\, y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n)}. \tag{3}$$

To this end, consider

$$(z_1, \ldots, z_n), \quad \text{where} \;\; z_i \in \{y_i, x_i\}, \;\; i = 1, \ldots, n.$$

Since $p(z_i) > 0$ for all i, we have by the positivity condition,

$$p(z_1, \ldots, z_n) > 0. \tag{4}$$

Let us define

$$z^i = (y_1, \ldots, y_{i-1}, x_i, \ldots, x_n), \quad i = 1, \ldots, n; \qquad z^{n+1} = (y_1, \ldots, y_n).$$

Let z^i_{-i} denote the collection of the components of z^i except for the ith component. For $i = 1, \ldots, n$, $p(z^{i+1})$ and $p(z^i)$ differ only in the ith component, namely

$$z^{i+1}_{-i} = z^i_{-i}, \quad z^{i+1}_i = y_i, \quad z^i_i = x_i; \tag{5}$$

and $p(z^{i+1})$, $p(z^i)$ satisfy

$$p(z^{i+1}) = p(z^{i+1}_i \,|\, z^{i+1}_{-i})\, p(z^{i+1}_{-i}) = p(y_i \,|\, z^{i+1}_{-i})\, p(z^{i+1}_{-i}),$$
$$p(z^i) = p(z^i_i \,|\, z^{i+1}_{-i})\, p(z^i_{-i}) = p(x_i \,|\, z^{i+1}_{-i})\, p(z^i_{-i}).$$

By Eq. (4), $p(z^i) > 0$, so dividing the first equation by the second and using also the fact that $z^{i+1}_{-i} = z^i_{-i}$, we obtain

$$\frac{p(z^{i+1})}{p(z^i)} = \frac{p(y_i \,|\, z^i_{-i})}{p(x_i \,|\, z^i_{-i})}, \quad i = 1, \ldots, n.$$

This implies

$$\prod_{i=1}^{n} \frac{p(z^{i+1})}{p(z^i)} = \prod_{i=1}^{n} \frac{p(y_i \,|\, z^i_{-i})}{p(x_i \,|\, z^i_{-i})}. \tag{6}$$

Since

$$\prod_{i=1}^{n} \frac{p(z^{i+1})}{p(z^i)} = \frac{p(z^{n+1})}{p(z^1)} = \frac{p(y_1, \ldots, y_n)}{p(x_1, \ldots, x_n)}$$

and $z^i_{-i} = (y_1, \ldots, y_{i-1}, x_{i+1}, \ldots x_n)$ by the definition of z^i, Eq. (6) is identical to Eq. (3). □