Bioinformatics Research Center
Aarhus University

Statistical Methods in Bioinformatics
Asger Hobolth
January 4, 2017

Exam

General information

A printed version of your report must be handed in before noon (12.00 am) Friday January 6, 2017, to me (office 324, Building 1110) or Ellen Noer (office 319, Building 1110).

Writing and organization of the report

The report should be written in danish or english. Part of the exam is to write clear and documented R code. In the report you should include the documented R code in an Appendix.

Project description

The project falls in three parts:

1. In the first part of the project we consider the global alignment problem for a pair of sequences and with a scoring scheme that corresponds to the Felsenstein model (this model is described in Ewens and Grant, 2005, Section 14.3.3 page 492).
2. In the second part of the project we analyse a pair of homologous sequences using the Felsenstein model. We estimate the parameters of the model and determine the evolutionary distance (in terms of expected number of substitutions) between the sequences.
3. In the third part of the project we analyse a 3-way alignment using a symmetric 2-state continuous time Markov chain. In particular we estimate the parameters of the model using the EM algorithm and estimating equations.

Part 1: Global alignment

Consider the global alignment problem for a pair of nucleotide sequences. A scoring scheme arising from the Felsensenstein model (recall Ewens and Grant, 2005, Section 14.3.3 page 492) is given by

$$S(j,k) = \begin{cases} 2 & \text{if } j = k = a \text{ (matching pair of } a\text{'s)} \\ 6 & \text{if } j = k = g \text{ (matching pair of } g\text{'s)} \\ 4 & \text{if } j = k = c \text{ (matching pair of } c\text{'s)} \\ 3 & \text{if } j = k = t \text{ (matching pair of } t\text{'s)} \\ -3 & \text{if } j \neq k \text{ (mismatch)} \end{cases}$$

for pairs of nucleotides (j,k) and -5 for a gap.

(i) Show that the score for the alignment

$$\begin{array}{cccccccc} a & t & a & g & - & c & c & g \\ a & t & - & c & c & c & c & g \end{array}$$

is 6.

(ii) Consider the pairwise global alignment problem with sequences $x = gatc$ and $y = gtac$ and the scoring scheme from above.
Determine the alignment matrix (called B in Ewens and Grant, 2005, Section 6.4.2), and find the global alignment(s) with the highest score.

(iii) Consider the pairwise global alignment problem with sequences

$$x = atgaactcgaataactcccaaattgagtaaaatttaacactcactatgggaaaaa$$

and

$$y = attacttttaaacactctgaagtagaaacgtagcccttcactataggaaaaa$$

and the scoring scheme from above.
Determine a global alignment with the highest score.

Part 2: Felsenstein model for a pairwise alignment

Consider the F81 model from Section 14.3.3 in Ewens and Grant (2005) with rate matrix

$$Q = u \begin{bmatrix} \varphi_a - 1 & \varphi_g & \varphi_c & \varphi_t \\ \varphi_a & \varphi_g - 1 & \varphi_c & \varphi_t \\ \varphi_a & \varphi_g & \varphi_c - 1 & \varphi_t \\ \varphi_a & \varphi_g & \varphi_c & \varphi_t - 1 \end{bmatrix}$$

where $0 < \varphi_i < 1$ for $i \in \{a, g, c, t\}$, $\varphi_a + \varphi_g + \varphi_c + \varphi_t = 1$, and u is a scaling parameter.

(i) Show that the stationary distribution is $\varphi = (\varphi_a, \varphi_g, \varphi_c, \varphi_t)$ and that the process is reversible.

Note that in matrix notation we have

$$Q = u(\epsilon'\varphi - I)$$

where $\epsilon = (1, 1, 1, 1)$ is a row vector with one in every entry, $'$ denotes vector transpose, and I is the 4×4 identity matrix.

(ii) Use the series expansion of the matrix exponential to show that the transition probability matrix is given by

$$P(u) = \exp(Q) = e^Q = e^{-u} I + (1 - e^{-u})\epsilon'\varphi.$$

Now consider two homologous sequences of length $n = 1000$ summarized in the usual table $\{n_{ij}\}$ where e.g. $n_{ag} = 27$ means that in 27 sites we have observed the pair (a, g) in the sequence alignment (see Table 1).

We want to estimate the parameters φ and u in the F81 model. We estimate the parameters φ from

$$\hat{\varphi} = (\hat{\varphi}_a, \hat{\varphi}_g, \hat{\varphi}_c, \hat{\varphi}_t) = (n_a, n_g, n_c, n_t)/n = (187, 312, 294, 207)/1000 = (0.187, 0.312, 0.294, 0.207),$$

where (n_a, n_g, n_c, n_t) are the row sums in the table.

(iii) Argue that the likelihood function for the parameter u is given by

$$L(u; \{n_{ij}\}) = \prod_{i=1}^{4} \prod_{j=1}^{4} P_{ij}(u)^{n_{ij}},$$

where $P_{ij}(u)$ is entry (i, j) in the transition probability matrix. (The likelihood only depends on the parameters through the rate matrix QT where T is the time between the two sequences. We therefore without loss of generality here and in the following let $T = 1$.)

	a	g	c	t	$\sum$
a	106	27	28	26	187
g	38	190	50	34	312
c	23	56	190	25	294
t	22	37	47	101	207
$\sum$	189	310	315	186	1000

Table 1: Summary of pairwise sequence alignment.

(iv) Plot the log-likelihood as a function of u, and determine the maximum likelihood estimate.

(v) What is the expected number of substitutions between the two sequences?

We now consider the EM algorithm for maximum likelihood estimation of the parameter u. Consider a fully observed process $\{x(s) : 0 \leq s \leq T\}$ from the F81 model for a single site.

(vi) Show that the full likelihood is given by

$$\begin{aligned} L(u; \{x(s) : 0 \leq s \leq T\}) &= \Big(\prod_{i=1}^{4} e^{-q_i T(i)}\Big)\Big(\prod_{i=1}^{4}\prod_{j\neq i} q_{ij}^{N(i,j)}\Big) \\ &\equiv e^{-u\{\sum_{i=1}^{4}(1-\hat{\varphi}_i)T(i)\}} u^{N_{\text{total}}} \end{aligned}$$

where $\equiv$ means equal up to a multiplicative constant, $T(i)$ is the time spent in state i, $N(i,j)$ is the number of jumps from state i to state j, and $N_{\text{total}} = \sum_{i=1}^{4}\sum_{j\neq i} N(i,j)$.

(vii) Show that the maximum of the full likelihood for a single site is given by

$$\hat{u} = \frac{N_{\text{total}}}{\sum_{i=1}^{4}(1-\hat{\varphi}_i)T(i)}.$$

(viii) Use the EM-algorithm for discretely observed continuous-time Markov chains to determine the maximum likelihood estimate for u for the data in Table 1. Describe the E-step and M-step in the algorithm in detail. The number of iterations should be 25 and the starting value of the EM should be $u = 1$. Plot the values of the parameter u in the iterative procedure.

Part 3: Branch lengths for a 3-way alignment

Consider the symmetric 2-state continuous time Markov chain (CTMC) with rate matrix

$$Q = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

(i) Show that the stationary distribution is $\varphi = (1/2, 1/2)$ and that the process is reversible.

We want to analyse a 3-way alignment of length $n = 20000$. The data is summarized in Table 2 where we have observed the column (x_1, x_2, x_3) a number $n_{x_1x_2x_3}$ times.

(ii) Argue that the data should be analysed using an *unrooted* tree.

(iii) Determine the maximum likelihood estimates for the branch lengths (t_1, t_2, t_3) using the EM-algorithm with starting values $(t_1, t_2, t_3) = (0.2, 0.2, 1.5)$ and 2000 iterations. Plot the values of the parameters and the log-likelihood for the data in the iterative procedure.

x_1	x_2	x_3	$n_{x_1 x_2 x_3}$
0	0	0	3816
0	0	1	3514
0	1	0	1367
0	1	1	1372
1	0	0	1407
1	0	1	1402
1	1	0	3355
1	1	1	3767

Table 2: Summary of a 3-way alignment. For example the alignment column $(0,0,1)$ has been observed $n_{001} = 3514$ times in the 3-way alignment.

We now consider an alternative method for parameter estimation. Let

$$\hat{p}_{12} = \frac{n_{010} + n_{011} + n_{100} + n_{101}}{n}$$

be the observed frequency that the states in the first and second sequences are different. Similarly let

$$\hat{p}_{13} = \frac{n_{001} + n_{011} + n_{100} + n_{110}}{n}$$

be the observed frequency that the first and third sequences are different and

$$\hat{p}_{23} = \frac{n_{001} + n_{101} + n_{010} + n_{110}}{n}$$

the observed frequency that the second and third sequences have different states.

(iv) Argue that

$$\begin{aligned}
\hat{p}_{12} &= \frac{1}{2}\left(1 - e^{-2(\hat{t}_1 + \hat{t}_2)}\right) \\
\hat{p}_{13} &= \frac{1}{2}\left(1 - e^{-2(\hat{t}_1 + \hat{t}_3)}\right) \\
\hat{p}_{23} &= \frac{1}{2}\left(1 - e^{-2(\hat{t}_2 + \hat{t}_3)}\right)
\end{aligned}$$

are natural estimating equations for the branch lengths (t_1, t_2, t_3).

(v) Solve the estimating equations in terms of $(\hat{t}_1, \hat{t}_2, \hat{t}_3)$ for the data in Table 2.

(vi) The maximum likelihood estimates obtained from the EM algorithm and the parameter estimates determined from the estimating equations are actually identical! Show that in general the estimating equations give the same parameter values as maximum likelihood estimation for the branch lengths of a 3-way alignment from the symmetric 2-state continuous time Markov chain.