MATH50010 Coursework

Chris Hallsworth

03/12/2020

This coursework task concerns the dataset that is read in below.

```
library(seqinr)
dat<-read.fasta("sequences.fasta")
```

The data object contain the complete genome sequence for the bacterium mycoplasma pneumoniae. You will first load in the data and fit the simplest possible statistical model, although only with the aim of rejecting it. You will then model the sequence as a Markov chain. Unusually, here X_n represents the nth position in the genome sequence, and **not** time. The point is that nearby positions are correlated. No knowledge of biology is necessary here, although some reading around might help you to develop insight into shortcomings of the Markov model for the final part. What you do need to know is that DNA is encoded in a four-letter chemical alphabet, A,C,G,T. This alphabet is our state space.

Some commands to help you are given at the end of the file. You will need to install the package `seqinr`. If you find code runs slow, you might want to work with just a smaller subset of the data while you get everything working, to save time.

Your solution should be as an RMarkdown document, knitted to make a PDF. Where you have to write mathematics, you are welcome to handwrite your solution, although you are encouraged to learn how to do LaTeX in RMarkdown. There are examples below. All your code should be commented, so that the marker can understand your approach.

Submit your solution to the Turnitin box on Blackboard by Thursday 17 December, 1pm UK time.

1. (2 marks) For the data given, compute the overall proportion of the four different bases in the sequence.
2. (3 marks) In the DNA double helix, the bases G and C are always bound together on opposite strands. Produce a plot to illustrate how the proportion of the sequence that is a G or a C varies across the genome using a sliding window of size 1000 bases. More precisely,

- choose a region of the genome of length (say) 100000 bases
- For each window of length 1000 bases, compute the proportion of G+C.
- Plot the proportion G+C against position.

3. (15 marks) A very naive model of the DNA sequence would be to regard each base pair as being chosen independently of its surroundings.

a) Using the base proportions already calculated, determine the proportions q_{ij} of each pair of consecutive bases $i, j \in \{A, C, G, T\}$ that would be expected under an independence model.

b) Now, we'll assess the strength of the evidence against this simple null model. To do this, you'll need to compute the log likelihood ratio statistic. This compares the (log of) the likelihood under the null model and the more general model where we estimate the probabilities for each pair $i, j \in \{A, C, G, T\}$ as $\widehat{p}_{ij} = \frac{n_{ij}}{n-1}$, where n_{ij} is the number of times the consecutive pair ij is seen, and n is the length of the sequence, so that there are $n-1$ (overlapping) pairs.

We assume under both hypotheses that pairs are sampled from a multinomial distribution (omitting a multinomial coefficient that cancels in the ratio)

$$L(\widehat{p}_{ij}) \propto \prod_{i,j} \widehat{p}_{ij}^{n_{ij}}, \qquad L(\widehat{q}_{ij}) \propto \prod_{ij} \widehat{q}_{ij}^{n_{ij}},$$

Note that this is not quite right - consecutive pairs overlap, so it is not entirely reasonable to regard the sampling as multinomial. This turns out not to matter - see d).

The log likelihood ratio statistic is then

$$\log L(\widehat{p}_{ij}) - L(\widehat{q}_{ij}).$$

Note: when computing the log likelihood ratio statistic, be mindful of the underflow errors - careful use of logs avoids this.

In a sense, the likelihood ratio is a measure of whether the data are better explained by the alternative model or the null model. But we have to make the comparison carefully - under the alternative model we are estimating more free parameters - this gives us more wiggle room to accommodate the data. So even if the data *were* generated under the null model, it is possible that the alternative model would fit better. This line of thinking tells us how to calibrate the test. We generate many data sets from the null model. For each data set, we estimate the parameters for the null model and the alternative model, and compute the likelihood ratio just as we did for our real dataset.

Sure enough, we will find that the likelihood ratio statistic is often greater than one: data simulated under the null model will actually have higher probability under the alternative model. But if the likelihood ratio is much greater than we would expect for data simulated under the null model, then we have evidence against the null hypothesis.

c) Use the `permutation` function to generate random permutations of the DNA sequence. These shuffled sequences are samples from the null distribution, because their ordering is entirely random. For each permuted sequence, compute the likelihood ratio as in b) This should allow you to reject the null hypothesis very comfortably: check that this is the case.

d) When, as here, we have lots of data, the null distribution of the likelihood ratio test statistic has a nice form. *Wilks' theorem* tell us that (twice) the likelihood ratio test statistic should have a chi-square distribution, with r degrees of freedom, where r is the difference between the number of free parameters estimated under the alternative hypothesis and the number of free parameters estimated under the null hypothesis. Use e.g. a QQ-plot to compare the distribution of (twice) the likelihood ratio to a chi-square distribution, specifying the degrees of freedom carefully. Comment on the result.

4. (10 marks) a) Show that if $(x_0, x_1, x_2, \ldots x_n)$ is a realization from the Markov chain with transition matrix P, then the log likelihood for P has the form

$$l(P) = \Pr(X_0 = x_0) + \sum_{i,j \in \mathcal{E}} n_{ij} \log p_{ij}.$$

b) Show that the maximum likelihood estimators are given by $\widehat{p}_{ij} = \frac{n_{ij}}{n_{i\cdot}}$, where

$$n_{i\cdot} = \sum_{j \in \mathcal{E}} n_{ij}.$$

Hint: you know that for each $i \in \mathcal{E}$, $\sum_{j \in \mathcal{E}} p_{ij} = 1$ - you can respect these constraints using Lagrange multipliers.

c) Compute the maximum likelihood estimates based on the DNA data, assuming the sequence is a first order Markov chain.

5. (15 marks)

a) Write code to simulate n independent realizations of length m using the transition probabilities from the data by maximum likelihood.

b) For each of the n realizations of the chain, compute the estimates of the transition probabilities. Produce plots to show that the estimators are approximately unbiased, with roughly normal sampling distribution. (You need only plot one entry).

c) Are the estimates of different transition probabilities correlated?

6. (5 marks) Suggest (briefly) how the suitability of the Markov model could be evaluated. Are there any obvious incompabitilities between the data and the model? How might the model be improved?

Helpful functions

Recall that the command `?some_function` gives you the help page for `some_function`, and will include a working example that you can modify.

- `count` from the `seqinr` package is useful for getting counts of each base, and pairs of bases.
- `getLength` extracts the length of the sequence
- `sample` can be used to choose at random from a vector, according to a specified probability distribution.