Exercises – Lecture 2
Stochastic Processes and Markov Chains, Part 2

Question 1 *(without* `R`*)*

a) The transition matrix of Markov chain is:

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

Find the stationary distribution of this Markov chain in terms of a and b.

b) For which a and b is the Markov chain reversible?

c) For which a and b is the Markov chain periodic?

Question 2

The size (n) of a cell population over time ($t = 1, 2, 3, \ldots$) can be described by a first order Markov process. During each time step one cell from the population may (but need not) die. This happens with probability λ. Simultaneously, one cell may (but need not) divide into two cells, an event with probability μ. The time steps have been chosen small enough that the probability of simultaneous dying (or dividing) of multiple cells is negligible. In particular, within a time step it may happen that neither cell division nor death takes place. Assume a maximum population size (n_{max}) of three cells. When this maximum is achieved, the cells stop with dying and dividing.

a) Give the state space and the transition matrix of this process. In this specify the restrictions on λ and μ. Also draw the *state diagram* and include the probabilities of each transition.

b) Does this process have a stationary distribution? Motivate your answer.

c) Now set the maximum population size equal to two ($n_{max} = 2$). Assume that $n = 1$ at $t = 1$. Calculate the probability that the cell population goes extinct. Similarly, the probability that the population lives forever.

d) Still assume $n_{max} = 2$. For which values of λ and μ does the probability of extinction exceed that of eternal life?

*e)** Back to $n_{max} = 3$. Let $\lambda = 0.1 = \mu$. Is the probability of extinction larger than that of eternal life?

Question 3 *(with* `R`*)*

Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.1500 & 0.3500 & 0.3500 & 0.1500 \\ 0.1660 & 0.3340 & 0.3340 & 0.1660 \\ 0.1875 & 0.3125 & 0.3125 & 0.1875 \\ 0.2000 & 0.3000 & 0.3000 & 0.2000 \end{pmatrix}.$$

a) Calculate the stationary distribution of the transition matrix $\mathbf{P}$ analytically and through matrix multiplication.

b) The GC content of a DNA sequence is defined as the percentage of C's and G's on the total number of bases of the sequence: $(\#\texttt{C} + \#\texttt{G})/(\#\texttt{A} + \#\texttt{C} + \#\texttt{G} + \#\texttt{T}) * 100\%$. Calculate the GC content for an infinitely long DNA sequence generated in accordance with the sampling model above (the first order Markov dependent stochastic process described by transition matrix above). Compare this to your answer to Question 4a of the previous set of exercises (corresponding to Lecture 1).

Question 4* *(without* R)

a) Prove that irreducible, aperiodic first-order Markov chains, defined by a transition matrix with a nonsymmetric off-diagonal zero entry, are irreversible. Note that a nonsymmetric zero means $p_{ij} = 0$ while $p_{ji} \neq 0$ for some i and j (for $i \neq j$).

b) Show that all first order Markov chains with an associated symmetric transition matrix are reversible.

Question 5

Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.40 & 0.10 & 0.30 & 0.20 \\ 0.00 & 0.20 & 0.80 & 0.00 \\ 0.00 & 0.60 & 0.40 & 0.00 \\ 0.30 & 0.30 & 0.10 & 0.30 \end{pmatrix}.$$

a) (without R*)* Draw the state diagram that corresponds to the transition matrix **P** above.

b) (without R*)* Is this Markov chain associated with the transition matrix above irreducible?

c) (without R*)* On the basis of parts *a)* and *b)*, which limiting behavior do you expect?

d) (with R*)* Perform a spectral decomposition of **P**, using R's `eigen`-function, and store the results in an object called `Pdecomp`. Verify that all eigenvalues are smaller or equal than 1 (in an absolute sense).

e) (with R*)* Having performed the spectral decomposition, then `Pdecomp$vectors` contains the right eigenvectors. Verify that the left eigenvector are given by the inverse of `Pdecomp$vectors`. Reconstruct **P** from the matrices of left and right eigenvectors and the diagonal matrix with eigenvalues on the diagonal.

f) (with or without R*)* Using the spectral decomposition, investigate how fast the influence of the initial value washes out (i.e., how fast do you reach the stationary distribution).

Question 6

a) (without R*)* Give the definition of stationary distribution of a Markov chain.

On the basis of their chemical properties the nucleotides are divided into two groups, the so-called purines (A en G) and pyrimidines (C en T). Assume the DNA may be modeled by a first order Markov process. This Markov process has a multinomial initial distribution with probability 0.4 for a purine and 0.1 for a pyrimidine. With respect to the elements of the transition matrix:

- the probability of a purine at position $j+1$ in the DNA, if position j is occupied by a *different* purine, is equal to 0.15.
- the probability of a pyrimidine at position $j+1$ in the DNA, if position j is occupied by a purine, is equal to 0.25.
- the probability of a pyrimidine at position $j+1$ in the DNA, if position j is occupied by a *different* pyrimidine, is equal to 0.35.

- the probability of a purine at position $j+1$ in the DNA, if position j is occupied by a pyrimidine, is equal to 0.20.

b) (without R*)* Give the transition matrix and its corresponding stationary distribution.

The DNA is not one long, uniform sequence of nucleotides, but can (very) crudely be divided in non-coding and coding pieces of DNA. Coding DNA contains the information for the formation of a protein, non-coding DNA does not. Assume the non-coding and coding sequences alternate every 1000th nucleotide. Given that a particular piece of the DNA encodes for a protein or not, the sequence of nucleotides may be described by a first order Markov process. The transition probabilities of coding DNA is given by the transition matrix of question 6b, while in non-coding DNA every nucleotide has a probability of 0.6 to be succeeded by a *different* nucleotide, with equal probability for the different nucleotides to be drawn.

c) (without R*)* Give the transition matrix and its corresponding stationary distribution of non-coding DNA.

d) (with R*, but also without* R *when using the answers to parts b) and c)* Give the proportion of time spent in a particular state of DNA regularly composed of the coding and non-coding sequences (as described above).

It is nonsense to presume that non-coding and coding DNA alternate so regularly. Replace this assumption. Let the length of both non-coding and coding DNA be Poisson distributed with $\lambda_{non-coding} = 1000$ and $\lambda_{coding} = 1500$, respectively.

e) (with R*, but also without* R *when using the answers to parts b) and c))* Keep in mind the definition given under Question 6d, and give the proportion of time spent in a particular state of DNA composed of the coding and non-coding sequences with random lengths (as described above). *Hint:* calculate the expected length of both non-coding and coding DNA sequences, and combine this with the answers to parts *b* and *c*.

Question 7

Throughout assume a stationary, 1st order, discrete time Markov process. Consider a DNA sequence of 1000 bases long. The nucleotide frequencies of this sequences are:

$$\begin{pmatrix} \#\texttt{A} & \#\texttt{C} & \#\texttt{G} & \#\texttt{T} \end{pmatrix} = \begin{pmatrix} 186 & 242 & 265 & 307 \end{pmatrix},$$

and the di-nucleotide frequencies:

$$\begin{pmatrix} \#\texttt{AA} & \#\texttt{AC} & \#\texttt{AG} & \#\texttt{AT} \\ \#\texttt{CA} & \#\texttt{CC} & \#\texttt{CG} & \#\texttt{CT} \\ \#\texttt{GA} & \#\texttt{GC} & \#\texttt{GG} & \#\texttt{GT} \\ \#\texttt{TA} & \#\texttt{TC} & \#\texttt{TG} & \#\texttt{TT} \end{pmatrix} = \begin{pmatrix} 58 & 60 & 0 & 68 \\ 39 & 77 & 83 & 43 \\ 43 & 0 & 87 & 134 \\ 46 & 105 & 95 & 61 \end{pmatrix}.$$

a) (without R*)* How many times do you expect to observe the motif TAGA in the sequence? Hereto obtain the ML estimates of the transition matrix from the mono- and di-nucleotide frequencies given above.

b) (with R*)* The motif TACA appears 17 times in the sequence. How exceptional is this (still using the above data)?