Assignments II - Markovian modeling and Bayesian learning

6. Define a stationary discrete time 2nd order Markov chain where each state takes values in the set $\mathcal{X} = \{A, C, G, T\}$, such that all transition probabilities $p(x_i|\mathbf{x} = (x_{i-1}, x_{i-2}))$ with $\mathbf{x} \in \{\mathcal{X} \times \mathcal{X}\}$ are larger than zero. Show explicitly how the model may also be defined as a 1st order Markov chain. Notice that this property of higher order Markov chain can be utilized in the remainder of this assignment! Simulate a sequence of 500 states from the model and perform both maximum likelihood (ML) and Bayesian estimation of the model parameters (transition probabilities). The parameters $\boldsymbol{\pi}_0$ of the initial distribution of the chain can be ignored in this assignment. In the Bayesian estimation you can use a default conjugate prior equal to Dirichlet distribution with hyperparameters $\lambda = (1, ..., 1)$. Compare the errors of ML and Bayesian estimates as a function of sequence length when the estimation is performed using the first 100, 200, 300, 400 and all the 500 observations.

7. Examine how Bayes factor behaves in learning of the order of Markov chain for the model you defined and simulated in assignment #2. Consider the comparison of the two models: Markov(0) and Markov(1) and calculate the Bayes factor for the batches of data consisting of the first 100, 200, 300, 400 and all the 500 observations. Notice that Markov(0) model corresponds to independent sequential draws from a multinomial distribution *given* an underlying vector of probabilities over $\mathcal{X} = \{A, C, G, T\}$. The Bayes factor can be calculated analytically if the conjugate prior from assignment #6 is used.

8. Compare the results from assignment #7 with those obtained using: a) an asymptotic approximation to the Bayes factor based on Bayesian Information Criterion (BIC), see, e.g. Kass & Raftery (1995) in the course bibliography, b) log-likelihood ratio test where $\chi^2(\nu)$-distribution with ν degrees of freedom is used for the test statistic under the null model (Markov(0)).

9. Learn variable order (i.e. length) Markov chain model for the DNA sequence downloadable here: http://web.abo.fi/fak/mnf//mate/jc/miscFiles/DNAsequence.txt, using e.g. the VLMC package for R, or some other software or your liking. Report a summary of your findings (learned contexts, tree). In case you are curious to know what sequence this is, you may BLAST for it in the NCBI nucleotide collection.

10. Use the data introduced in assignment #9 and learn order for ordinary Markov chain model using methods considered in assigments #7 & #8. In the learning process restrict the attention to the Markov chain orders 0,1,...,4. Compare the results with findings about the order in the VLMC analysis of assignment #9.