**Bioinformatics Research Center**
**Aarhus University**

**Advanced Statistical Methods in Bioinformatics**
**Asger Hobolth**
**October 23, 2018**

# Poisson and Posterior HMM analysis: Project

## General information

The project should be handed in before noon Monday November 12, 2018, in Noors mailbox in the BiRC building.

## Writing and organization of the report

Report your analysis in a clear, complete, and concise language. Organise the report in a logical order with descriptive headings and subheadings. This means that the headings should not be e.g. '1(a)', but rather 'Getting the evolutionary distance from the probability of genomic difference'.

The report should be written in english.

A main part of the project is to write R code. In the report you should include in an Appendix the R code that you have written. The R code should be documented. This means that you should not just include your R commands, but also briefly describe what the R commands do and also describe the structure of the R code.

## Project description

Recall the fetal lamb movement data from Handout 5. We model the data using a HMM with two hidden states corresponding to low activity (state 1) and high activity (state 2). The transition matrix for the hidden states is then determined by the probability for staying in the low activity state $a = p_{11}$ and the probability for staying in the high activity state $b = p_{22}$. Furthermore we let the initial distribution be $\pi = (1/2, 1/2)$. The number of movements in the low activity state follows a Poisson distribution with rate $\lambda_1$ and the number of movements in the high activity state is assumed Poisson with rate $\lambda_2$.

We use the notation from Ewens and Grant (2005) Section 12, where $\mathcal{O} = (\mathcal{O}_1, \dots, \mathcal{O}_T)$ is the observed data and $Q = (q_1, \dots, q_T)$ is the hidden state sequence.

1. Determine the maximum likelihood estimates of the four parameters $(a, b, \lambda_1, \lambda_2)$ using the EM-algorithm. Derive and describe the EM-algorithm in detail.

2. Determine the Viterbi and Posterior Decoding state sequences.

3. In the following the parameters are fixed at their maximum likelihood estimates and assumed known. Calculate the posterior transition probabilities

$$a_t = \text{Prob}(q_t = 1 | q_{t-1} = 1, \mathcal{O}) \text{ and } b_t = \text{Prob}(q_t = 2 | q_{t-1} = 2, \mathcal{O}) \text{ for } t = 2, \dots, T.$$

4. Plot the transition probabilities $a = (a_2, \dots, a_T)$ and $b = (b_2, \dots, b_T)$ and the original data. Discuss the plots: Do the posterior transition probabilities behave as expected?

5. What is the posterior probability of the Viterbi sequence? Do you find that the Viterbi sequence is a good representative for the posterior distribution of hidden state sequences?

6. Describe and implement a procedure for simulating a hidden state sequence from the posterior distribution.

7. Calculate the empirical distribution (from a reasonable number of simulations) of the fraction of time, the fetal lamb is in the high activity state.

8. Supply the empirical distribution with a numerical calculation of the distribution. Plot the two distributions and add the fraction from Viterbi and Posterior Decoding.

9. Do you prefer the simulation-based or the numerical procedure for calculating the fraction of time spent in the high activity state? Why?

10. What is the role of parameter uncertainty? Clearly, the original parameters $(a, b, \lambda_1, \lambda_2)$ are estimated with some uncertainty. But does that matter for the distribution of the fraction of time spent in a state?

11. We finally consider three statements in Aston and Martin (2007); the full paper is available on BlackBoard.

    (a) On page 586 in the Introduction they write:

    *Currently, inference on patterns in the hidden state sequence of an HMM usually proceeds as follows. The HMM is determined and the Viterbi algorithm is used to find the most probable state sequence among all possible ones, conditional on the observations. This state sequence is then treated as if it is deterministically correct and patterns are found by examining it. However, the conditional distribution (given the observations) of patterns over all state sequences is more relevant. If, for example, the number of genes present in a DNA sequence is of interest and the Viterbi sequence of an HMM is used [as in methods based on Krogh, Mian and Haussler (1994)], then counting genes from the Viterbi sequence cannot be guaranteed to even give a good estimate of the number of genes. This is because there could be gene counts that correspond to many state sequences, and when accumulating probabilities over those sequences, one could find that those counts are much more likely than the count corresponding to the Viterbi sequence. This could especially be true if there are many different sequences all with likelihood close to that of the Viterbi sequence. If a single choice of gene count is needed, then the mean of the conditional distribution over state sequences, given the observations, would seem to be a more reasonable choice. Thus, a method to compute pattern distributions in state sequences modeled as HMMs would be helpful.*

    (b) On page 606 in the Discussion they write:

    *An alternative method for calculating the distributions is to sample from the conditional distribution of the states through the conditional transition probabilities given by Lemma 3.1 and then empirically calculate the distributions of pattern lengths and number of patterns from these samples. For especially complex patterns, this could well lead to improvements in the computation time, although the computed distribution would then be approximate. However, given the inhomogeneous nature of the transition probabilities and the large number of samples needed to get accurate distributions for long sequences, unless the patterns are quite complicated, this method is likely to be less efficient.*

    (c) On page 607 in the Discussion they write:

    *In this work, it is assumed that appropriate values for the transition probabilities have been estimated, with the estimated parameters being used as input to compute the desired waiting time distributions. There could, however, be an appreciable effect of the estimation process on the probabilities that are eventually determined. An area of future work is to determine the effect of small errors in parameter estimation on the output waiting time distributions for patterns.*

    Discuss the three paragraphs in relation to your analysis of the fetal lamb movement data.

# References

Aston, J.A.D. and Martin, D.E.K. (2007). Distributions associated with general runs and patterns in hidden Markov models. *The Annals of Applied Statistics*, Vol. 1, No. 2, 585–611.

Ewens, W.J. and Grant, G. (2005). *Statistical Methods in Bioinformatics.* 2nd Edition. Springer, USA.