Bioinformatics Research Center
Aarhus University

Statistical Methods in Bioinformatics
Asger Hobolth
November 1, 2018

Handout 9

Second mandatory project

The 2nd mandatory project is available. The project description is the HMMPoisson file in the HMM folder. Deadline for handing in the project is November 12, 2018.

Lectures in Week 44 and 45

The plan has changed: On Monday, November 5, we have the presentation on the Cone Projection algorithm, and on Wednesday, November 7, we have the presentation on the EM algorithm. On Monday, November 12, we will discuss the Alexandrov *et al* (2013) paper (an application of NMF to a data set on mutation counts from human cancer).

Exercises in Week 44

1. **The objective function for NMF**
   Let V be the observed $M \times N$ matrix, W a $M \times K$ matrix of free parameters, and H a $K \times N$ matrix of free parameters.

   Suppose $V_{mn} \sim \mathrm{N}\big((WH)_{mn}, \sigma^2\big)$ and that all observations are independent. Show that maximum likelihood estimation of W and H in this model corresponds to minimizing the objective function

$$||V - WH||^2 = \sum_{m=1}^{M} \sum_{n=1}^{N} \big(V_{mn} - (WH)_{mn}\big)^2.$$

   Now assume $V_{mn} \sim \mathrm{Pois}\big((WH)_{mn}\big)$. Show that maximum likelihood estimation in this model corresponds to minimizing the objective function

$$\sum_{m=1}^{M} \sum_{n=1}^{N} \Big\{(WH)_{mn} - V_{mn} \log\big((WH)_{mn}\big)\Big\}.$$

   On page 21 in Hobolth, Guo, Kousholt and Jensen (2018) we have a discussion about the choice of objective function. Argue that the Kullback-Leibler divergence corresponds to the Poisson model, and that the Kullback-Leibler divergence is approximately equal to half the Goodness of Fit statistic.

   Which model (Gaussian or Possion) do you prefer for learning the mutational processes in human cancer genomics? Or, in other words, which objective function do you think is the most appropriate? Can you suggest a procedure for checking the model assumptions? Do you think that it matters which of the two objective functions are chosen?