Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
November 13, 2018

NMF analysis: Project

General information

The project should be handed in before noon Wednesday November 28, 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. 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

The purpose of this project is to reproduce the NMF analysis of the mutation count data from Alexandrov *et al.* (2013). We follow Alexandrov *et al.* (2013) and fix the number of signatures to four.

On BlackBoard in the NMF folder you can find the mutation count data from Alexandrov *et al.* (2013) in the BRCA21.Rdata file. Download the file, and load the data into R using the command `load("BRCA21.Rdata")`. The data is now available in the matrix V. The matrix is of dimension 96 (the mutation types) times 21 (the number of patients), as can be checked with the command `dim(V)`. The mutation types are specified in the row names of V; see them using the command `rownames(V)`, and the patients can be identified from the column names (the command is `colnames(V)`).

1. Use the alternating non-negative least square (NLS) algorithm to determine the exposures (loadings) and mutational signatures for the 21 breast cancer patients. Use your preferred NLS algorithm, but make sure that you describe the choices that you have made in your design of the alternating NLS algorithm. Recall that you should fix the number of signatures to four.

2. Argue that your algorithm has converged.

3. Do you find the same four signatures as in Alexandrov *et al.* (2013)?

4. The *total* expected and observed mutation count for each patient should be close to each other; make a plot where you demonstrate that this is indeed the case.

5. Two of the signatures have a high loading in many of the patients. What are these two signatures?

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. and Stratton, M.J. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3, 246–259.