**Assignment Multivariate Component Analysis**

**(Data Science 2, BFVM19DATASC2**)

*July 5, 2022*

Instructor: Dr. Tsjerk A. Wassenaar (+316 1703 1440)

This assignment is to be made by the candidate in the form of a single, well-formatted (sectioned) **Jupyter notebook**, to be handed in through Blackboard before **Friday July 8th** **17:00**. Any material available may be used. The work handed in must be performed and written by the candidate, and in case of doubt, the candidate may be invited to an interview about the work.

For the assignment there are several datasets available in the assignment folder. These are described below. The assignment is to **perform meaningful, complete, and correct analyses of two or more datasets and present these**. Demonstrate (at least) three different methods. **PCA and/or FA** must be included, and **regression (MLR, PCR, or PLSR)** must be included. A third method may be selected as desired (PA, MDS, LDA, SGP, ...).

For each method write a section in the Jupyter notebook in which you:

- give the rationale for using the method for the data selected and state expectations (20%)

- justify the selections (rows/columns) and/or treatment of the data (15%)

- explain/justify the choices made in the application of the method (15%)

- plot the results in a clear and logical manner (20%)

- interpret and reflect on the results (30%)

It is allowed to make selections of columns or rows if deemed appropriate or required, e.g., due to memory constraints. Doing so will require justification of the approach and reflection on the consequences for the results.

It is allowed to chain methods, i.e., use the results from one method as input for another method.

**Datasets:**

*The file* **metabolom\_data.csv** *contains* *measurements of monoterpenes in certain plant leaves in AUC/mg. These data come from the 2021 paper 'Unifying community detection across scales from genomes to landscapes' from Hudon and coworkers.*

*The file* **SourceData2.xlsx** *contains infrared spectra from saliva from people that tested negative and people that tested positive for SARS nCoV2019. For each subject, three spectra were determined. The data accompany the study 'Infrared based saliva screening test for COVID-19' from Wood et al.*

*The file* **randompoisson.dat.gz** *contains 1000 by 10000 Poisson distributed random numbers with means per column taken from an exponential distribution. This can be considered simplified fictive/dummy count data, e.g., from genomics/transcriptomics, with 1000 observations in 10000 variables.*

*The file* **lysozyme\_rmsds.csv** *contains a distance matrix of a selection of lysozyme structures from the Protein DataBank (PDB). Lysozyme is one of the most used/studied proteins, with >5400 structures available in the database. Distances were determined as root-mean-square deviation (RMSD) after least-squares alignment using the CE align algorithm.*