**Assignment Multivariate Component Analysis**

**(Data Science 2, BFVM19DATASC2**)

*April 8, 2022*

Instructor: Dr. Tsjerk A. Wassenaar (06 1703 1440)

This assignment is to be made by the candidate in the form of a **Jupyter notebook**, to be handed in through Blackboard before **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.

There are several datasets available in the assignment folder:

*The file* **galton.dat** *contains the original data from Galton, which was used to formulate the notion of* regression*. It comprises measurements of lengths of children and their parents, and the sex of the subjects.*

*Use multilinear regression (Height is dependent on gender and on the height of parents),m at the end split by category gender and see how well it explains it.*

*The file* **connectome\_vectorized.csv.gz** *comprises brain connectivity data for men and women. Each row consists of boolean values indicating whether a connection between two brain areas is present or absent. Only connections with variance among the subjects are included, so there are no connections that are present or absent in all subjects. The last column indicates whether the subject is male or female. A relevant research question is whether there are differences in brain connection patterns between men and women.*

*Multilinear discriminant analysis? We know 2 groups and believe there should be differences based on that.*

*The file* **transhydrogenase.csv.gz** *contains molecular dynamics simulation data of wildtype and mutant forms of the protein transhydrogenase (TH). TH is a dimeric integral membrane enzyme in mitochondria and prokaryotes that couples proton translocation across a membrane with hydride transfer between NAD(H) and NADP(H) in soluble domains. This process involves conformational changes in the protein. Mutations may also affect the conformation and can thus modify the efficiency of the protein. This dataset contains conformations of TH sampled over time using molecular dynamics simulations. Each row corresponds to one conformation. The first field of a row contains a label indicating what mutation is present (WT for wild-type). The rest of the columns contain the coordinates of the atoms in vectorized form. All structures have already been aligned to a common frame of reference using Procrustes analysis. The coordinates are thus suitable for direct further analysis. A relevant question is whether the different variants can be distinguished in terms of conformation and dynamics (covariance structure).*

*PA has already been performed to decrease the individual differences, thus now a PCA would be logical to verify the difference between the molecules*

*The file***reflectance and transmission spec measurements.xlsx***contains Near-InfraRed (NIR) spectra measured on white feathers from a variety of birds (50 species). These data come from the paper*[*'The microstructure of white feathers predicts their visible and near-infrared reflectance properties'*](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199129)*. The conclusion of the paper was that the 'results highlight the potential for adaptive variation in NIR reflectance mediated by feather structure, which may inform the design of functional materials to control light and heat.' As such, these measurements are a starting point for bio-inspired design of materials. NIR spectra (like other spectra) can be considered 1D images.****Note*** *that 1. the file can be read using the* read.excel *function from pandas and 2. the data are transposed: the first column contains the wavelengths measured and the other columns contain the observations.*

*????*

**Assignment**

Perform correct and meaningful analyses of two or more datasets and present this in a Jupyter notebook. 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 (MDS, LDA, SGP, ...).

For each method

- 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%)