# Multivariate data analysis project summary

## Purpose of MDS

Multidimensional scaling (MDS) serves 2 purposes:

1. Dimensionality reduction

As real-world data comes with many features and is often collected at large scales, data from various fields such as finance/environmental/biology can become very hard to read and understand. Just like PCA/LDA/Factor analysis algorithms(2), MDS can also be used to reduce the data dimensionality from a higher dimension to a lower dimension to obtain a set of principle variables (1).

However, there are differences between MDS and PCA of how to measure the similarity or dissimilarity between variables or variable clusters.

1. Visual representation technology

MDS can be used as a visual representation technology to display the similarity or dissimilarity between the data points. Given pairwise dissimilarities, reconstruct a map that preserves distances is the goal of MDS (3).

A famous example is: MDS reproducing the well-known two-dimensional colour circle.

There are two main types of MDS; metric MDS and nonmetric MDS. Metric scaling uses the actual values of the dissimilarities of the data, whereas the dissimilarities in nonmetric scaling are essentially interpreted as ranks. Because of this, nonmetric MDS is more robust to non-linear relationships between the calculated dissimilarities and the projected distance between objects.

## How to measure distance/dissimilarity (3)

To measure the distance, dissimilarity and similarity is the fundamental part of MDS. Euclidean or non-Euclidean distance can be chosen to measure the distance, what MDS does is to find an optimal configuration xi that gives di,j ≈ ||xi – xj|| as closely as possible. For our data, this will be done by:

* Using the dataset, select a few variables and calculate their distance.
* Stating the process of minimizing the stress/dissimilarity with examples.
* Plotting the dissimilarity with R and interpreting the results.

## Choose a method and explain how to implement it

There are a few methods which can be used by MDS, such as INDSCAL, SINDSCAL, POLYCON, ALSCAL, KYST.

For our project, we will choose an interesting one or two, explain the process of how the method works, and explain the distance /dissimilarity calculation and result. We may also analyse the difference between the two methods if possible.

## Data clustering analysis and interpretation

The visual representation that comes out of MDS can be used as a tool to discover structure in the data such as clusters. Using MDS we can explore and compare inner-cluster distances with intra-cluster differences.

## Proposed dataset

We propose to use a dataset on mice protein expression from the UCI Machine Learning Repository (see <http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression>). This is a dataset of 82 variables:

* 1: Mouse ID
* 2-78: Values of expression levels of 77 proteins; the names of proteins are followed by \_n indicating that they were measured in the nuclear fraction. For example: DYRK1A\_n
* 79: Genotype: control (c) or trisomy (t)
* 80: Treatment type: memantine (m) or saline (s)
* 81: Behaviour: context-shock (CS) or shock-context (SC)
* 82: Class: c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

The variables of interest are the protein expression levels and the class.

The experiment that this dataset is sourced from took 15 measurements of each protein per mouse. As there were 72 mice, this means that the dataset is comprised of 72x15=1080 observations. If this proves to be too large for our analysis, a smaller random sample can be taken from the dataset.

## References

1. <https://medium.com/datadriveninvestor/the-multidimensional-scaling-mds-algorithm-for-dimensionality-reduction-9211f7fa5345>
2. <https://www.statisticshowto.com/multidimensional-scaling/>
3. <https://www.stat.pitt.edu/sungkyu/course/2221Fall13/lec8_mds_combined.pdf>
4. <http://www.stat.yale.edu/~lc436/papers/JCGS-mds.pdf>
5. <https://papers.nips.cc/paper/1008-multidimensional-scaling-and-data-clustering.pdf>