## Introduction

Multidimensional scaling (MDS) is a technique used in multivariate statistics to represent similarities between high-dimensional data. This is done by using the distance between points as a dissimilarity measure, with the aim of being able to display the relative distances between all points in reduced dimensions.

The technique of MDS will be demonstrated by analysing a real dataset. The dataset that will be used is of protein expression in control and Down syndrome mice with different exposures to a drug and learning methods. This data will be analysed using variations of MDS to present how the methodology affects results.

## Purpose of MDS

The primary purpose of MDS is to arrange points representing objects in a way so that the geometrical distance between points reflects relationships between the objects. This is done through two objectives, reducing dimensionality and visual representation.

The first main objective is to reduce the dimensionality of data. This is significant in multivariate statistics, as datasets with large numbers of variables and observations can be difficult to interpret without a high level of analysis. MDS addresses this by working with the dissimilarities between data, which can reduce the overall number of dimensions needed to represent the analysis. The dissimilarity data (represented by distances in space) can be scaled to find a lower dimensional configuration that maintains pairwise distances as well as possible.

The second main objective of MDS is as a method of visually representing the similarity or dissimilarity between data points. The results of MDS analysis can be graphed in two- or three-dimensional space so that observations that are more similar are visually closer together. This is a very intuitive and interpretable way of representing similarities between observations. Visually representing these similarities can help to identify overall patterns or clusters in the data.

*Something about similarity to PCA(?)*

A basic example of the application of MDS is a problem presented by Krabbe (2017). In this problem, the aim is to construct a map of airport locations only knowing the distances between all the airports. With only a few airports this can be simple to deduce, but as the number of airports and therefore distances increases, the problem becomes more complex. With MDS these distances can be interpreted to almost perfect locations of the airports. This can also be done using only the ranking of the distances by magnitude, as shown in Figure 1 below.

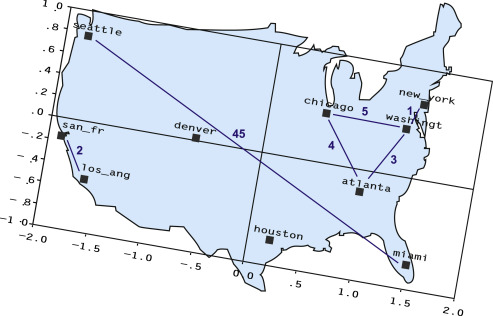


Figure - The two-dimensional solution of multidimensional scaling on the ranks of the 45 airline mileages (with distance ranks 1–5 and 45 depicted) between 10 cities in the United States (Krabbe, 2017)

## Measuring similarity

The input to MDS analysis is a matrix that indicates relationships among a set of items. This is most commonly represented as proximity data given by a proximity matrix **D** that represents the dissimilarity between objects. The way this proximity matrix is calculated can significantly affect the analysis. There are various ways to create the proximity matrix **D**, that depend on the structure of the data used. Examples of ways of obtaining proximity data include:

* Euclidean distance – the straight-line distance between two points in space
* Human evaluated similarity – using data based on people’s perceived similarity of objects
* Confusion data – the frequency at which objects are categorised to be the same, or how objects are categorised

These similarities/dissimilarities are used to map objects so that Di j approximates

In the analysis for this report, BLANK will be used

## Types of scaling

To scale an object is to change its dimensions by a certain factor. There are several different types of scaling that can be used in MDS, which differ primarily in the loss function used to scale the proximity data. The most common way that scaling used in MDS is described is as metric or non-metric. Metric MDS uses the actual values of the dissimilarities of the data, whereas non-metric MDS uses the rankings of these dissimilarities.

## Data used

The mice protein expression dataset that will be analysed is based on the protein expression measured in control and Down syndrome (trisomy) mice. The data was collected to understand the effects of treatments on the mice’s learning abilities, which were measured through the expression levels of gene proteins that are believed to contribute to learning ability. The mice also underwent various treatments; administering a drug called memantine or a saline placebo, and administering an electrical shock after being exposed to the novel learning environment (context-shock) or an electrical shock straight away (shock-context). 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.

In total, 82 variables were available in the dataset:

* 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.

*Sampling maybe?*

## Methodology

## Results

## Clustering analysis(?)

## References

Data reference

<https://www.sciencedirect.com/science/article/pii/B9780128015049000143>