**Fundamentals of Data Science**  
Unsupervised Learning (2 classes)

**Introduction – dimensionality reduction**

There are many difficulties associated with multidimensional data, mainly related to:  
a) the efficiency of EDA algorithms for which computational complexity linearly - or even worse exponentially - depends on the dimensionality of the problem  
b) unfavorable phenomena occurring in multidimensional sets  
An example for the latter is illustrated by the figure below showing how the distance between the nearest and furthest elements of the set disappears with increasing dimensionality - regardless of the size of the sample (different line designations):



Source: K. Beyer et al., „When Is «Nearest Neighbor» meaningful?”, in: Proc. ICDT, 1999.

Unfortunately, most of today's data sets have this type of character (documents - the number of features is the size of the dictionary, social research - number of indicators examined, DNA - thousands of genes, and often much less cases).

Therefore, advanced dimensional reduction techniques are used to provide the analysis of multi-dimensional data.

They include:

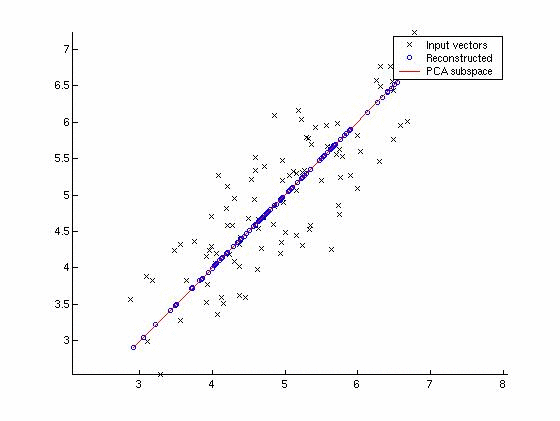
a) selection of important features (feature selection)

b) creating a new, reduced set of features through the transformation of the previous (feature extraction)

Selection methods can be treated as a subset of the extraction methods.

**Principal Component Analysis and Multidimensional Scaling**

The first relies on projecting the data set into new axes - maximizing the variance. The second - to find a configuration of points that will keep distances from the space of original dimensionality.



**Assignment**

Please import (in any program) a set of seeds (google: seeds dataset). This collection represents measurements of several geometrical features of wheat grains - made using X-ray techniques. The last column (values: 1, 2, 3) means the grade of the tested grain. The purpose of this data collection was to verify whether the grain grade can be determined based on the measurement of geometrical features. Unfortunately, the dimensionality of the problem does not allow for its direct analysis.

The course of the task:

1. Please reduce the data dimension to N = 2 using the PCA method. For example, you can use princomp from the Octave package or similar MATLAB function (<http://octave.sourceforge.net/statistics/function/princomp.html)>.

or Python’s sklearn PCA:

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

Note: you must transform the data using the function, then use the first two columns from the score matrix. What are the elements of the latent array?

1. Please plot the data in two-dimensional space by coloring the grain species. In the case of Octave/MATLAB, you can use gscatter (requires two dimensions of the file and then a label - allowing the coloring of its elements). For Python use Matplotlib scatter plot (<https://kite.com/python/answers/how-to-color-a-scatter-plot-by-category-using-matplotlib-in-python>). Do such classes are distinguishable?
2. Please examine the result of dimensionality reduction – while looking at scatter plot – for more recent non-linear method of dimensionality reduction: t-SNE (Matlab: <https://www.mathworks.com/help/stats/tsne.html>, Python: <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>). Remember: we reduce the dimensionality to N=2.