DATA.STAT.770 Dimensionality Reduction and Visualization, Spring 2023, Exercise set 3

Part B: Linear Dimensionality Reduction

Problem B1: Mathematics of Principal Component Analysis

Derive the solution for Principal Component Analysis (PCA) using maximization of variance in the projected space as an optimization criterion. That is, given data $X_{d\times N}$ with N samples in d dimensions, find projection matrix $W_{d\times k}$, $1 \leq k \leq d$, such that $\mathrm{Var}(Z)$ is maximized for projected data $Z = W^{\mathrm{T}}X$.

You may assume that X has zero mean. Derive the solution for the first principal component w_1 by maximizing $Var(w_1^{\mathrm{T}}X)$, and using the constraint that $w_1^{\mathrm{T}}w_1 = 1$.

Problem B2: Principal Component Analysis of Wines

The Wine Quality data set "winequality-red.txt" and "winequality-white.txt" provided in this archive is a data set which can be used to predict quality of white and red wine varieties based on their chemical characteristics. However, in this exercise we do not try to predict the wine quality but simply analyze the rest of the input variables. The data set comes from the UCI Machine Learning Repository and is available there at `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`. Each wine is described by 12 feature values, in order: 1-fixed acidity, 2-volatile acidity, 3-citric acid, 4-residual sugar, 5-chlorides, 6-free sulfur dioxide, 7-total sulfur dioxide, 8-density, 9-pH, 10-sulphates, 11-alcohol, 12-quality.

- For the red wines, find the two first principal components, project the data onto them, and plot the data along them as a scatter plot. Compute the amount of variance explained by the two first components.
- Then repeat the same analysis for the white wines.
- Principal component analysis finds orthogonal projections (orthogonal axes) that contain the maximal variance. The original variables are also orthogonal axes; if PCA was restricted to use only one variable per projection, it would reduce to variable ranking of the original variables based on the variance. Perform such variable ranking for the white wines, take the top-two variables, and compare the resulting picture to the "real" (unrestricted) PCA result.

Hint: in Matlab, consider the Matlab functions "cov" and "eig". R and Python have corresponding functions. Try to plot the wines with low quality

(quality value 5 or below), medium quality (quality value 6) and high quality (quality value 7 or above) with different colors.