Responsible TA:
Jiwoo Oh
jiwoo.oh@chem.ethz.ch

Statistical and Numerical Methods
Exercise 12, HS23

19/12/2023

1 Principal Component Analysis

One of the first multivariate data sets was introduced by Sir Ronal Fisher in 1936. This data quantifies the morphologic variation of Iris flowers of three related species.

1. Read in the dataset `Fisher Iris.xlsx`.
2. Standardize the four numerical X variables using the function `zscore`.
3. Perform a Principal Component Analysis using the function `pca` on the standardized data.
4. Plot the specific and cumulative variance explained by the principal components normalizing it by the maximal variance explained such that $R^2_{max} = 1$.
   - What can you interpret regarding the general correlation structure of the variables?
   - How many 'main effects' seem to be present in the data?
5. Plot the loadings and scores of the first two principal components using the function `biplot`.
   - Can you observe different groups? Distinguish those groups by showing the different flower species in different colors and labeling them with their species name using `text` and an additional scatter plot.
   - What can you conclude regarding the correlation of the variables and the 'main effects' in the data?
6. Investigate the correlation matrix (using the function `corrcoef`) to support your first conclusion.
7. Which species features most of the abnormal observations?
   - Use the Hotelling's T2 distance in the plane of the first two PCs (t_1 and t_2) for each observation *i* comparing it to a critical value of 6.3 (corresponds to 95 % level)
   - Using the variances of t_1 and t_2, s_1^2 and s_2^2, it can be defined as

$$T_i^2 = \frac{t_{i1}^2}{s_1^2} + \frac{t_{i2}^2}{s_2^2}$$

8. Which species can be worst explained by the model with two PCs? How could you improve this deviation from the model plane?
   - Calculate the SPE value for each observation comparing it to critical level of 0.6 (corresponds to 95 % level).
   - You can access the residuals using the function `pcares`.
9. What would happen without the initial standardization step?