Principal Component Analysis

Application and Discussion

Lena Teresa Will

WiSo Faculty,
University of Cologne

Multivariate Statistics
Winter Term 2022/2023

Motivation

Principal Component Analysis:

- *Objective:* Dimensionality Reduction.
- *Technique:* Find linear combinations that capture as much variance of the original data as possible.
- *Application:* Input for prediction models (here: classification).

PCA Recap I

Starting point:

- p-dimensional random vector,

$$y' = (y_1, \ldots, y_p)$$

- New variables (principal components) are linear combinations of the original data y,

$$z_j = a_{1j}y_1 + \cdots + a_{pj}y_p = a_j'y$$

PCA Recap II

To calculate the first principal component:

- Lagrangian

$$\mathcal{L}(a_1) = \underbrace{a_1' \Sigma a_1}_{\mathbb{V}\{z_1\}} - \lambda \underbrace{(a_1' a_1 - 1)}_{\text{constraint}}$$

- Taking the FOC w.r.t a_1 and setting it to zero

$$(\Sigma - \lambda I_p) a_1 = 0$$

- Choose λ such that this holds

$$\det(\Sigma - \lambda I_p) = 0$$

- First PC

$$z_1 = a_1' y$$

Overview

1. A first look at the data: What is the relation between the variables?
2. Principal Component Analysis.
3. Using the principal components for classification.
4. Discussion and comparison to autoencoders.

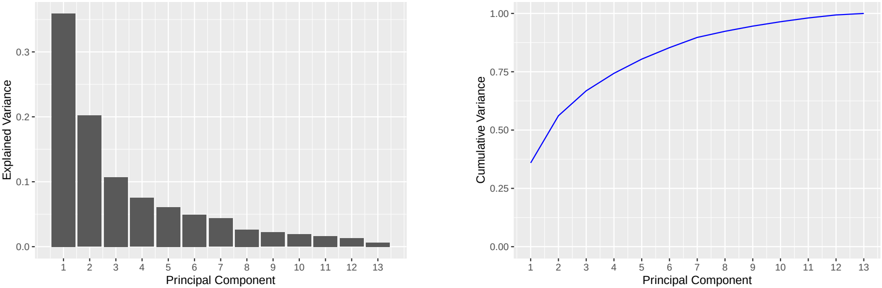
The Dataset: Quality of Wine

Classification of wines based on their cultivars (varieties):

- y is the cultivar class, $y \in \{1, 2, 3\}$.
- X includes 13 different variables (chemicals) describing different types of wine:
  - Alcohol
  - Magnesium
  - Colour intensity
  - ...

$\Rightarrow$ What is the relation between the chemicals?

Choosing the Number of Principal Components I



(a) Explained Variance

(b) Cumulative Variance

Figure 1: Variance of the original data explained by the principal components

Choosing the Number of Principal Components II

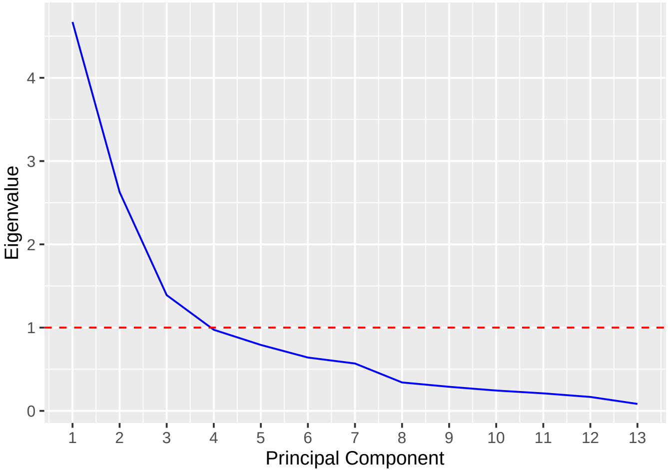


Figure 2: Scree Plot

Interpretation of the Principal Components I

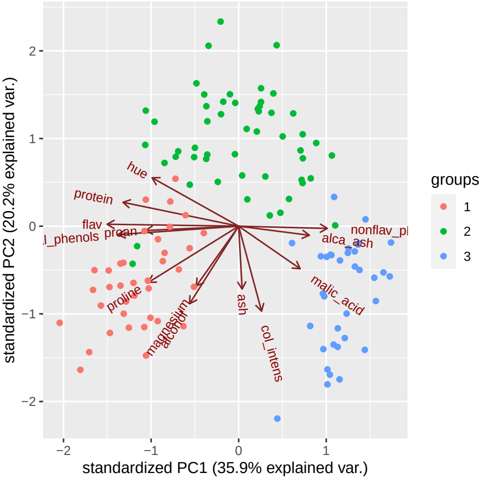


Figure 3: Biplot of the first two principal components

Interpretation of the Principal Components II

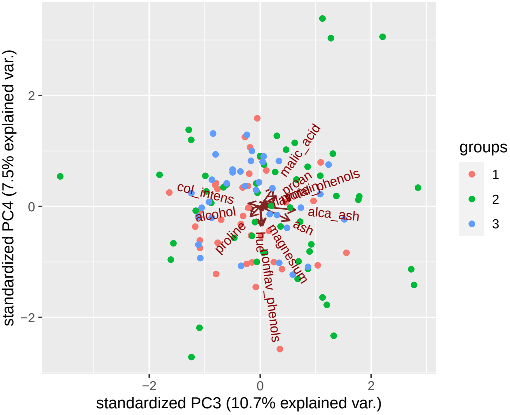


Figure 4: Biplot of the third and fourth principal component

Discussion

- Principal component analysis captures linear relations between the variables in the feature space.
- Even with non-linear relations, PCA can perform really well.
- There might be cases where non-linearity in the feature space is important to be accounted for.
- Dimension Reduction Techniques that account for non-linearity:
  - Kernel PCA
  - Autoencoders[1]

[1] Code for direct comparison is on my github: https://github.com/lena-will/pca

References I

Gribisch, B. (2022). *Lecture in multivariate statistics*.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis*. John Wiley & Sons.

UCI Machine Learning Repository. (1991). *Wine data set*. Retrieved November 12, 2022, from https://archive.ics.uci.edu/ml/datasets/wine

Autoencoder