Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

# Business Analytics & Machine Learning
# Homework sheet 10: Principle Component Analysis

**Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami**
**Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao**

## Exercise H10.1 *PCA Calculation*

You are given the following dataset:

| $i$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 0 | 2 | 3 | 4 |
| 1 | -2 | 0 | 2 |
| 2 | 0 | 1 | 2 |
| 3 | 4 | 2 | 0 |
| 4 | 1 | 1 | 1 |

**Table 1** Dataset

You are further given the covariance matrix $\Sigma_x$:

$$\Sigma_x = \begin{bmatrix} 5 & 2 & -1 \\ 2 & 1.3 & 0.6 \\ -1 & 0.6 & 2.2 \end{bmatrix}$$

You compute the characteristic polynomial of $\Sigma_x$ to be:

$$f(\lambda) = \lambda^3 - 8.5\lambda^2 + 15\lambda$$

a) What percentage of the variance does each principal component explain? How many principal components are necessary to explain all of the variance in the data?

b) Compute the principal component that explains most of the variance and scale it to be a unit vector. Using this principal component, calculate the one-dimensional projection of the data (as coordinates along the principal component).

c) A colleague plans to run a linear regression model using the three features $x_1$, $x_2$, and $x_3$. Discuss what problem occurs and how this problem can be mitigated.

d) Now another datapoint $(4, 2.6, 1.2)$ is added to the dataset. Do the eigenvalues, the principal components, and the ratio of explained variance of each component change? Discuss your reasoning.

## Exercise H10.2 *PCA implementation in Python*

In this exercise, we will conduct PCAs using primitive Python functions, trying to mimic the PCA computation algorithm. We will use the built-in *iris* dataset.
*from sklearn.datasets import load_iris*
*iris = load_iris()*
*data = iris.data*
*Note: you may use functions from the* numpy *library throughout this exercise.*

a) Check the structure of the dataset. Which attributes are numerical? Compute a correlation matrix for them.

b) Construct a new matrix consisting only of numerical attributes with mean values subtracted.

c) Calculate covariance matrix, eigenvectors and eigenvalues.

d) Compute the PCA scores by multiplying the transposed eigenvectors matrix and the transposed zero-mean matrix. Check the first six scores.

## Exercise H10.3 *PCA in sklearn:* `sklearn.decomposition.PCA`

The code from the previous exercise is cumbersome and long. In this exercise, we will use sklearn.decomposition.PCA to get the same result with less effort.

a) Use sklearn.decomposition.PCA in order to compute PCA's for the *iris* dataset.

b) How much standard deviation does each component have? How much variance is explained by the first component?

c) Check the computed eigenvectors. Are they equal to what was computed in the previous exercise?

d) Check the first six PCA scores for the dataset. Compare them with the scores computed in the previous exercise.

e) Plot the projection of the dataset points on a space consisting of two principal components only.