Multivariate Verfahren – Summer Semester 2024 **Exercise sheet 5**

Hannah Schulz-Kümpel 02.07.2024

Principal Component Analysis

Question 1: PCA by hand

Consider a data matrix given by

$$\boldsymbol{X} = \begin{pmatrix} 24 & 22 & 24 \\ 24 & 21 & 25 \\ 24 & 22 & 20 \\ 24 & 23 & 21 \end{pmatrix}.$$

a) Derive the principal components via eigen decomposition of the sample covariance matrix.

b) Let us assume that we want to reduce the data's dimension to $k = 2$. Calculate the new data points in $\mathbb{R}^2$.

Question 2: Invariance of PCA w.r.t. transform

Given a PCA of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, consider the matrix of scores

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & \cdots & y_{n1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{1m} & \cdots & \cdots & y_{nm} \end{pmatrix} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top \in \mathbb{R}^{m\times n},$$

where each columns gives the coordinates $\boldsymbol{y}_i$ of observation i, $i = 1, \dots, n$, in the m-dimensional space with the principal component (vectors) as axes.

a) Show that the sample covariance of $\boldsymbol{Y}$ is equal to $\boldsymbol{\Lambda}_{\text{ord}}$, i.e. the diagonal matrix of ordered eigenvalues of either the sample covariance matrix $\boldsymbol{S}$.

b) In the lecture, we have learned that PCA is not scale-invariant when we solve the optimization problem $\boldsymbol{a}_p^\top \boldsymbol{S} \boldsymbol{a}_p \to \max$, only when we solve $\boldsymbol{a}_p^\top \boldsymbol{R} \boldsymbol{a}_p \to \max$.

Can you reason why this is the case, using a diagonal matrix $\boldsymbol{T} \in \mathbb{R}^{m\times m}$ which transforms the varible scales by replacing each observation $\boldsymbol{x}_i$ with $\boldsymbol{T}\boldsymbol{x}_i$?

c) Next, consider shifting each data point by a constant $c \in \mathbb{R}$. Is PCA invariant w.r.t. a shift of each data point by a constant?

d) Lastly, consider an orthogonal matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$. How does PCA behave w.r.t. orthogonal transformation, i.e. w.r.t. replacement of each observation $\boldsymbol{x}_i$ with $\boldsymbol{A}\boldsymbol{x}_i$?

Question 3: Interpreting PCA output in R

There are two main ways to perform PCA in R:

- the `princomp()` function - based on eigen decomposition and
- the `prcomp()` function - based on singular value decomposition (SVD).

According to the R help, `prcomp()` via SVD has slightly better numerical accuracy. Here you can use the option `scale=TRUE` to perform standardized PCA, i.e. the version that iteratively solves $\boldsymbol{a}_p^\top \boldsymbol{R} \boldsymbol{a}_p \rightarrow \max$.

For visualization of PCA results, the `factoextra` package is very popular; except for biplots, for which the `ggfortify` package is standard.

a) Perform PCA on the `iris` data set excluding the variable `Species` and interpret the output.

b) Plot the scree plot and select the number of PCs that should be selected for dimension reduction according to each of the criteria on lecture-slide 67.

c) Plot the Biplot and interpret it.