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 $\mathbb{R}^2$.

Solution:

a) 1. **Compute the sample covariance matrix:**

Recall from the lecture, that the following holds for the sample covariance matrix:

$$\boldsymbol{S} \;=\; \frac{1}{n-1}\boldsymbol{X}_C^\top \boldsymbol{X}_C \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)^\top.$$

In this case, we have

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \frac{1}{4}\Big(24+24+24+24\Big) \\ \frac{1}{4}\Big(22+21+22+23\Big) \\ \frac{1}{4}\Big(24+25+20+21\Big) \end{pmatrix} = (24,\, 22,\, 22.5)^\top.$$

It follows that

$$
\begin{aligned}
\boldsymbol{S} = \frac{1}{3}\Bigg(& \begin{pmatrix} (24-24) \\ (22-22) \\ (24-22.5) \end{pmatrix} \Big((24-24),\,(22-22),\,(24-22.5)\Big) + \\
& \begin{pmatrix} (24-24) \\ (21-22) \\ (25-22.5) \end{pmatrix} \Big((24-24),\,(21-22),\,(25-22.5)\Big) + \\
& \begin{pmatrix} (24-24) \\ (22-22) \\ (20-22.5) \end{pmatrix} \Big((24-24),\,(22-22),\,(20-22.5)\Big) + \\
& \begin{pmatrix} (24-24) \\ (23-22) \\ (21-22.5) \end{pmatrix} \Big((24-24),\,(23-22),\,(21-22.5)\Big) \Bigg) \\
= \frac{1}{3}\Bigg(& \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2.25 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -2.50 \\ 0 & -2.5 & 6.25 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6.25 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1.50 \\ 0 & -1.5 & 2.25 \end{pmatrix} \Bigg) \\
= \frac{1}{3} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & -4 \\ 0 & -4 & 17 \end{pmatrix}.
\end{aligned}
$$

2. **Perform an eigen decomposition of $\boldsymbol{S}$:**

First, we need to compute the eigenvalues via the characteristic polynom

$$\det(\boldsymbol{S} - \lambda \boldsymbol{I}_3) \overset{!}{=} 0\,.$$

$$
\begin{aligned}
\Rightarrow \begin{vmatrix} -\lambda & 0 & 0 \\ 0 & \left(\frac{2}{3} - \lambda\right) & -\frac{4}{3} \\ 0 & -\frac{4}{3} & \left(\frac{17}{3} - \lambda\right) \end{vmatrix} &= -\lambda \cdot \left(\frac{2}{3} - \lambda\right) \cdot \left(\frac{17}{3} - \lambda\right) - \frac{16}{9}\lambda \\
&= \frac{1}{3}\Big(3\lambda^3 + 19\lambda^2 + 6\alpha\Big) \overset{!}{=} 0\,.
\end{aligned}
$$

$\Rightarrow \lambda_1 = 6, \quad \lambda_2 = \frac{1}{3}, \quad \lambda_3 = 0\,.$

Eigenvector corresponding to λ_1

$$\begin{pmatrix} -6x_1 & 0 & 0 \\ 0 & -\frac{16}{3}x_2 & -\frac{4}{3}x_3 \\ 0 & -\frac{4}{3}x_2 & -\frac{1}{3}x_3 \end{pmatrix} \overset{!}{=} 0 \quad \Leftrightarrow \quad v_1 = \begin{pmatrix} 0 \\ -\frac{1}{4} \\ 1 \end{pmatrix}.$$

v_1 needs to be normalized: $\boldsymbol{v_1} = (0.0000000,\, -0.2425356,\, 0.9701425)^\top$

Eigenvector corresponding to λ_2

$$\begin{pmatrix} -\frac{1}{3}x_1 & 0 & 0 \\ 0 & -\frac{1}{3}x_2 & -\frac{4}{3}x_3 \\ 0 & -\frac{4}{3}x_2 & -\frac{16}{3}x_3 \end{pmatrix} \overset{!}{=} 0 \quad \Leftrightarrow \quad v_2 = \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}.$$

v_2 needs to be normalized: $\boldsymbol{v_2} = (0.0000000,\, 0.9701425,\, 0.2425356)^\top$

Eigenvector corresponding to λ_3

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{2}{3}x_2 & -\frac{4}{3}x_3 \\ 0 & -\frac{4}{3}x_2 & \frac{17}{3}x_3 \end{pmatrix} \overset{!}{=} 0 \quad \Leftrightarrow \quad v_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} .$$

v_3 does not need to be normalized.

Finally, the eigen decomposition of $\boldsymbol{S}$ is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3} \end{pmatrix} \begin{pmatrix} 6 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v_1}^\top \\ \boldsymbol{v_2}^\top \\ \boldsymbol{v_3}^\top \end{pmatrix}$$

and the PCs are $\boldsymbol{a_1} = \boldsymbol{v_1}$, $\boldsymbol{a_2} = \boldsymbol{v_2}$, and $\boldsymbol{a_3} = \boldsymbol{v_3}$.

b) To use the PCs for dimension reduction, we multiply the original data with the matrix of the first k columns of eigenvectors.

In our case, $k = 2$, so we achieve dimension reduction via

$$\boldsymbol{X} \begin{pmatrix} \boldsymbol{v_1}, \boldsymbol{v_2} \end{pmatrix} = \begin{pmatrix} 24 & 22 & 24 \\ 24 & 21 & 25 \\ 24 & 22 & 20 \\ 24 & 23 & 21 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ -0.2425356 & 0.9701425 \\ 0.9701425 & 0.2425356 \end{pmatrix} = \begin{pmatrix} 17.94764 & 27.16399 \\ 19.16031 & 26.43638 \\ 14.06707 & 26.19385 \\ 14.79467 & 27.40653 \end{pmatrix} .$$

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

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

where each columns gives the coordinates $\boldsymbol{y}_i$ of observation i, $i = 1, \ldots, n$, 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 $\boldsymbol{A}\boldsymbol{x}_i$?

Solution:

a) Let $\boldsymbol{S} \in \mathbb{R}^{m \times m}$ again denote the sample covariance matrix for the following.

We recall that

1. For $\boldsymbol{V}$ denoting the matrix whose columns are the eigenvectors of $\boldsymbol{S}$, ordered in descending order according to the corresponding eigenvalues and $\boldsymbol{X}_C$ denoting the centered data matrix, we have
   - $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Lambda}_{\text{ord}}\boldsymbol{V}^\top$ and
   - $\boldsymbol{X}_C = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.
2. For $\boldsymbol{a}_p$ denoting the pth PC (i.e. pth column of $\boldsymbol{V}$), $p = 1, \ldots, m$, the pth entry of $\boldsymbol{y_i}$ is given by

$$y_{ip} = \boldsymbol{a}_p^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}}), \qquad i = 1, \ldots, n$$

$$\Leftrightarrow \mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\top = \mathbf{V}^\top(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad i = 1, \ldots, n.$$

It immediately follows that the sample covariance matrix of $\boldsymbol{Y}$ is given by

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\top &= \frac{1}{n-1}\sum \boldsymbol{V}^\top(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\boldsymbol{V} \\ &= \boldsymbol{V}^\top\boldsymbol{S}\boldsymbol{V} \\ &= \boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Lambda}_{\text{ord}}\boldsymbol{V}^\top\boldsymbol{V} \text{ substituting the eigen decomposition for } \boldsymbol{S} \\ &= \boldsymbol{\Lambda}_{\text{ord}}\,.\end{aligned}$$

b) Just as in the first exercise, that the following holds for the sample covariance matrix:

$$\boldsymbol{S} \;=\; \frac{1}{n-1}\boldsymbol{X}_C^\top\boldsymbol{X}_C \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^\top.$$

Now, if we change the scale of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ by replacing each observation $\boldsymbol{x}_i$ with $\boldsymbol{T}\boldsymbol{x}_i$, the new data's arithmetic mean is given by $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{T}\boldsymbol{x}_i = \boldsymbol{T}\bar{\boldsymbol{x}}$ and the new data's sample covariance matrix by

$$\begin{aligned}&\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{T}\boldsymbol{x_i} - \boldsymbol{T}\bar{\boldsymbol{x}})(\boldsymbol{T}\boldsymbol{x_i} - \boldsymbol{T}\bar{\boldsymbol{x}})^\top \\ =&\frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{T}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^\top\boldsymbol{T}^\top \\ =&\,\boldsymbol{T}\boldsymbol{S}\boldsymbol{T}^\top \underset{\text{because } \boldsymbol{T} \text{ is diagonal}}{=} \boldsymbol{T}\boldsymbol{S}\boldsymbol{T}\,.\end{aligned}$$

Clearly, the eigenvalues and eigenvectors of $\boldsymbol{TST}$ will not be identical to those of $\boldsymbol{S}$ unless all diagonal entries are equal to 1.

⇒ This is what is meant by "PCA is not scale-invariant".

However, since the sample correlation has standardized entries, this is not an issue when considering $\boldsymbol{R}$ instead of $\boldsymbol{S}$.

c) PCA is invariant w.r.t. shifting by a constant, even when using the sample covariance matrix $\boldsymbol{S}$.

This holds because, each point of the shifted data is given by $\boldsymbol{x_i}+c$, the arithmetic mean by $\bar{\boldsymbol{x}}+c$, and the shifted data's sample covariance matrix by

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=1}^{n}\Big((\boldsymbol{x_i}+c)-(\bar{\boldsymbol{x}}+c)\Big)\Big((\boldsymbol{x_i}+c)-(\bar{\boldsymbol{x}}+c)\Big)^\top\\
=&\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^\top\\
=&\boldsymbol{S}.
\end{aligned}$$

Clearly, it immediately follows that we get the same score vectors.

This is also easily shown:

$$\mathbf{y}_i \;=\; \mathbf{V}^\top(\mathrm{x}_i+\mathbf{c}-(\bar{\mathrm{x}}+\mathbf{c})) \;=\; \mathbf{V}^\top(\mathrm{x}_i-\bar{\mathrm{x}}) \quad \forall l\in\{1,\dots,n\}\,.$$

d) Equivalently to subtask a), the sample covariance matrix of the orthogonally transformed data, i.e. data with new observations $\boldsymbol{Ax_i}$, is given by

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{Ax_i}-\boldsymbol{A\bar{x}}\right)\left(\boldsymbol{Ax_i}-\boldsymbol{A\bar{x}}\right)^\top\\
=&\frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{A}\left(\boldsymbol{x_i}-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x_i}-\bar{\boldsymbol{x}}\right)^\top\boldsymbol{A}^\top\\
=&\;\boldsymbol{ASA}^\top\,.
\end{aligned}$$

Thereby, PCA is definitely not invariant w.r.t. orthogonal transformation.

However, the following also holds for the above sample covariance matrix:

$$\boldsymbol{ASA}^\top = \mathbf{AV\Lambda V}^\top\mathbf{A}^\top = \mathbf{B\Lambda B}^\top\,,$$

for $\boldsymbol{B}:=\boldsymbol{AV}$. Since $\boldsymbol{B}$ is also orthogonal by definition, it follows that the eigenvalues of the sample covariance matrix **are not changed by the orthogonal transformation!**

For this reason, PCA is sometimes called *equivariant with respect to orthogonal transformations.*

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 $\boldsymbol{a}_p^\top \boldsymbol{R}\boldsymbol{a}_p \to \max$.

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.

Solution:

a) ***See R code.***

The output is

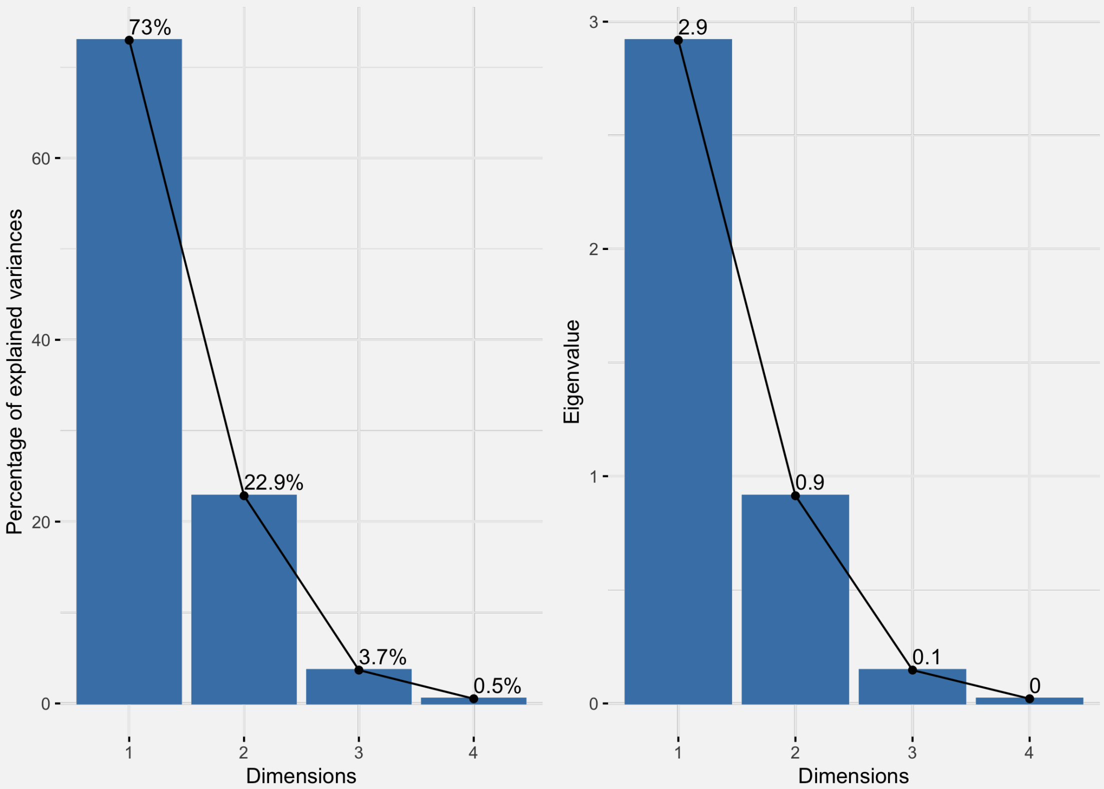
```
Standard deviations (1, .., p=4):
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Interpretation:

- The standard deviations are the standard deviations of the principal components, which are equal to the square roots of the eigenvalues of the covariance/correlation matrix.
- The Rotation columns are equal to the principal component (vectors).
  Meanwhile, the Rotation rows correspond to the loadings of each variable.

b) ***See R code.***



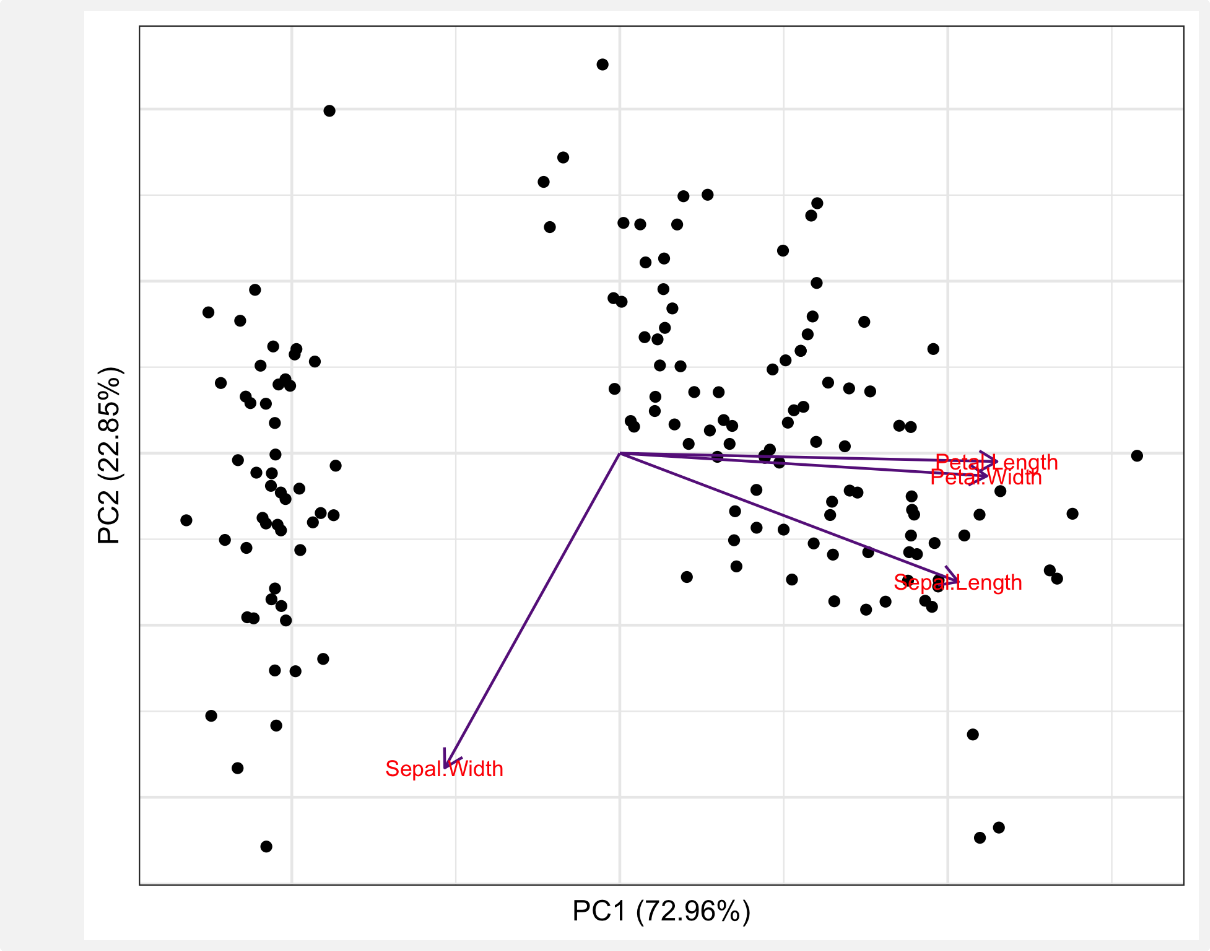
- *Kaiser criterion:* Principal components with eigenvalue greater than 1. (I.e. the maximal k s.t. $\lambda_k > 1$) – The choice would be 1.
- All principal components needed to get a total of 80% of the variance. (I.e. the minimal k s.t. $\mathrm{tr}(\boldsymbol{\Lambda}_{\mathrm{ord}})^{-1} \cdot \sum_{i=1}^{k} \lambda_k \geq 0.8$) – The choice would again be 2.
- *Scree Plot:* Consider a graphical representation of the eigenvalues. Use as many principal components up to the bend of the graph (elbow).
  - Here, we might decide to go with 3.
- Simply choose k so that it is convenient (e.g. for a planned visualization).
  - For visualization, one would often choose 2.

c) ***See R code.***

The Biplot overlays scoreplots, i.e. dots with coordinates given by the first k entries of the score vectors $\boldsymbol{y}_i$, $i = 1, \ldots, n$, with loading plots, i.e. arrows that point towards the coordinates given by the first k entries of the columns of the Rotation matrix from subtask a).

In the plot below, we can observe a few things:

- The data may vaguely be divided into two clusters on the first component.
- Since the loadings for `Petal.Length` and `Petal.Width` mostly contribute to the variability along PC1.
- `Petal.Length` and `Petal.Width` are highly positively correlated with each other, but both negatively correlated with `Sepal.Width`.


PC2 (22.85%)
PC1 (72.96%)
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width