Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 6 (December, 16, 2024)

Exercise 1 (Principal Component Analysis)
Given a data set with covariance matrix

$$\Sigma = \begin{pmatrix} 4 & -\sqrt{3} \\ -\sqrt{3} & 2 \end{pmatrix}$$

1. How many features and observations are in the data set?
2. Identify the corresponding principal components by hand.
3. How much variance is explained by the first component?
4. Is it possible to reduce the dimensionality by using the principal components, if we want to describe at least 90% of the variance?
5. Prove that your principal components are orthogonal.

Example solution:

1. The data set consists of 2 features. There are at least two observations in the data. A more specific number can't be derived from the given data
2. First, one needs to compute the eigenvalues:

$$\begin{aligned} 0 \overset{!}{=} \det(\Sigma - \lambda \cdot I) &= \begin{vmatrix} 4-\lambda & -\sqrt{3} \\ -\sqrt{3} & 2-\lambda \end{vmatrix} \\ &= 8 - 6\lambda + \lambda^2 - 3 \\ &= \lambda^2 - 6\lambda + 5 \\ &= (\lambda - 5) \cdot (\lambda - 1) \end{aligned}$$

Thus, the eigenvalues of Σ are $\lambda_1 = 5$ and $\lambda_2 = 1$.

- Eigenvector of $\lambda_1 = 5$:

$$\left(\begin{array}{cc|c} 4-5 & -\sqrt{3} & 0 \\ -\sqrt{3} & 2-5 & 0 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 1 & \sqrt{3} & 0 \\ 0 & 0 & 0 \end{array}\right)$$

$$\Rightarrow \qquad x_2 := \alpha \text{ and } x_1 = -\sqrt{3} \cdot x_2 \qquad \Rightarrow \frac{1}{2} \cdot \begin{pmatrix} -\sqrt{3} \\ 1 \end{pmatrix}$$

- Eigenvector of $\lambda_2 = 1$:

$$\left(\begin{array}{cc|c} 4-1 & -\sqrt{3} & 0 \\ -\sqrt{3} & 2-1 & 0 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 3 & -\sqrt{3} & 0 \\ 0 & 0 & 0 \end{array}\right)$$

$$\Rightarrow \qquad x_2 := \alpha \text{ and } x_1 = \frac{\sqrt{3}}{3} \cdot x_2 \quad \Rightarrow \frac{1}{2} \cdot \begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix}$$

Therefore, the principal components are

$$y_1 = -\frac{\sqrt{3}}{2} \cdot x_1 + \frac{1}{2} \cdot x_2 \text{ and } y_2 = \frac{1}{2} \cdot x_1 + \frac{\sqrt{3}}{2} \cdot x_2.$$

3. The first component explains $\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{5}{5+1} = 83.\overline{3}\%$ of the variance.
4. We only have two dimensions, thus a reduction of dimensions would be equal to referring to only one principal component. However, as the first component only explains $83.\overline{3}\% < 90\%$ of the variance, a dimensionality reduction is not possible.
5. We can show the orthogonality of the components as follows:

$$\left\langle \frac{1}{2} \cdot \begin{pmatrix} -\sqrt{3} \\ 1 \end{pmatrix}, \frac{1}{2} \cdot \begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix} \right\rangle = \frac{1}{2} \cdot \frac{1}{2} \cdot \begin{pmatrix} -\sqrt{3} & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix} = \frac{1}{4} \cdot \left(-\sqrt{3} \cdot 1 + 1 \cdot \sqrt{3}\right) = \frac{1}{4} \cdot 0 = 0$$

Exercise 2 (Properties of the trace of a matrix)
The *trace* of a square matrix $A \sim (p,p)$ is defined as the sum of its diagonal elements, i.e., $\text{tr}(A) = \sum_{i=1}^{p} a_{ii}$. Show the following properties that we used to prove that principal components are indeed uncorrelated:

1. For two matrices $A, B \sim (p,p)$ it holds that $\text{tr}(A \cdot B) = \text{tr}(B \cdot A)$.
2. For matrices $A, B, C \sim (p,p)$ the trace is invariant given cyclic swaps, i.e.,

$$\text{tr}(A \cdot B \cdot C) = \text{tr}(C \cdot A \cdot B) = \text{tr}(B \cdot C \cdot A).$$

3. Conclude that for a covariance matrix $\Sigma \sim (p,p)$ with eigenvalue-decomposition (EVD) $\Sigma = A^T \cdot D \cdot A$ it holds that $\text{tr}(\Sigma) = \text{tr}(D)$.

Example solution:

1. By definition of matrix multiplication the diagonal elements of the matrix products are given by

$$(A \cdot B)_{ii} = \sum_{j=1}^{p} a_{ij} \cdot b_{ji} \quad \text{and likewise} \quad (B \cdot A)_{ii} = \sum_{j=1}^{p} b_{ij} \cdot a_{ji}$$

for all $1 \leq i \leq p$. Now the trace is by definition

$$\text{tr}(A \cdot B) = \sum_{i=1}^{p} (A \cdot B)_{ii} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} \cdot b_{ji} = \sum_{j=1}^{p} \sum_{i=1}^{p} b_{ji} \cdot a_{ij} = \text{tr}(B \cdot A).$$

2. Essentially we need to use the results from (1) and leverage that matrix multiplication is an associative operation. It follows that:

$$\mathrm{tr}(A \cdot B \cdot C) = \mathrm{tr}((A \cdot B) \cdot C) \overset{(1)}{=} \mathrm{tr}(C \cdot (A \cdot B)) = \mathrm{tr}((C \cdot A) \cdot B) \overset{(1)}{=} \mathrm{tr}(B \cdot (C \cdot A)) = \mathrm{tr}(B \cdot C \cdot A)$$

3. Using (2) we get:

$$\mathrm{tr}(\Sigma) \overset{(\mathrm{EVD})}{=} \mathrm{tr}(A^T \cdot D \cdot A) \overset{(2)}{=} \mathrm{tr}(\underbrace{A \cdot A^T}_{=I_p} \cdot D) = \mathrm{tr}(D)$$

where $A \cdot A^T = I_p$ due to A being orthogonal. Here, I_p is the $(p \times p)$ identity matrix.

Exercise 3 (Influence of different scales in PCA)
The results of a PCA is highly influenced by scales of each feature when we perform a PCA on a covariance matrix. This exercise illustrates these effects.

1. Load the data set `mtcars` and calculate the variance of each feature. Use some variant of `summarize` of the package `dplyr`. Which features have such a high variance that they differ significantly from the others?
2. Now perform a PCA using `prcomp` based on the covariance matrix. Which variables of the original data set contribute to the first two principal components.
3. How many principal components are necessary to explain at least 90% of the variance of the data set?
4. Perform step 2. and 3. for a PCA which is based on a correlation matrix (if you call to `prcomp` set `scale = TRUE`).
5. Which of both approaches seem more reasonable and why?

Example solution:
See Jupyter notebook.

Exercise 4 (*t*-distributed Stochastic Neighbour Embedding)
As we have seen during the lectures, PCA does not always produce good results depending on the data it is used on, where alternative approaches such as *t*-SNE might work better. In this exercise, you will examine a number of different datasets and compare the performance of PCA.

- Using R, create 3 3-dimensional clusters using multivariate normal distributions. Each cluster should have a different mean and different covariance (and thus a different size). Run both PCA and *t*-SNE on this data and report the results. Which approach performed better? Multivariate normal distributions can be sampled from using the `mvrnorm` function from the `MASS` package.
- Download the fashion MNIST dataset. You can use either the training or the testing data, although the testing set is recommended due to the smaller size. Use both PCA and *t*-SNE on the data and report the results. Which approach works better? Note that due to the size of the dataset, t-sne might run quite slowly. In this case, feel free to just take a random sample of 1000 points or use a smaller number of iterations.

Example solution:

See Jupyter notebook.