Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 3 (November, 11, 2024)

Exercise 1 (Recap: normal distribution)
Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be independent and identically distributed random variables. Show that for

$$Y := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

it holds that $E(Y) = 0$ and $\mathrm{Var}(Y) = 1$.

Example solution:
We first show the that $E(Y) = 0$. To this end

$$\begin{aligned}
E(Y) &= E\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)\right) \\
&= \frac{1}{\sigma\sqrt{n}} E\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{\sigma\sqrt{n}} \underbrace{E\left(\sum_{i=1}^{n} \mu\right)}_{=n\mu} && \text{(Linearity of } E(\cdot)\text{)} \\
&= \frac{1}{\sigma\sqrt{n}} \underbrace{\sum_{i=1}^{n} E(X_i)}_{=nE(X_1)=n\mu} - \frac{1}{\sigma\sqrt{n}} n\mu && \text{(Linearity of } E(\cdot)\text{)} \\
&= \frac{1}{\sigma\sqrt{n}} n\mu - \frac{1}{\sigma\sqrt{n}} n\mu \\
&= 0 .
\end{aligned}$$

For the variance we use $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$ and the independence of the random variables to obtain

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)\right) \\
&= \frac{1}{\sigma^2 n} \mathrm{Var}\left(\sum_{i=1}^{n} (X_i - \mu)\right) \\
&= \frac{1}{\sigma^2 n} \sum_{i=1}^{n} \underbrace{\mathrm{Var}(X_i - \mu)}_{=\mathrm{Var}(X_1)=\sigma^2 \text{ due to independence of the } X_i} \\
&= \frac{1}{\sigma^2 n} n\sigma^2 = 1.
\end{aligned}$$

Thus, $E(Y) = 0$ and $\mathrm{Var}(Y) = 1$ as claimed.

Exercise 2 (QQ-Plots)

Consider the *penguins* dataset from the package *palmerpenguins* which provides various measurements for a group of adult penguins in Antarctica. Below you are given 10 observations of this data set from which `NA` values have been removed (you can use the function `complete.cases()` for subsetting). Your task is to check those data of the variable `flipper_length_mm` for normality.

```
   species island    bill_length_mm bill_depth_mm flipper_length_mm

 1 Adelie  Torgersen           39.1          18.7               181
 2 Adelie  Torgersen           39.5          17.4               186
 3 Adelie  Torgersen           40.3          18                 195
 4 Adelie  Torgersen           36.7          19.3               193
 5 Adelie  Torgersen           39.3          20.6               190
 6 Adelie  Torgersen           38.9          17.8               181
 7 Adelie  Torgersen           39.2          19.6               195
 8 Adelie  Torgersen           41.1          17.6               182
 9 Adelie  Torgersen           38.6          21.2               191
10 Adelie  Torgersen           34.6          21.1               198
```

a) Normalise the variable appropriately so that you can check for standard normal distribution (you can use R for this purpose). Provide the values of this variable `flipper_length_mm_norm`.
Example solution:
mean: 189.2, sd: 6.32
-1.2972547, -0.5062458, 0.9175704, 0.6011668
0.1265614, -1.2972547, 0.9175704, -1.1390529
0.2847632, 1.3921758

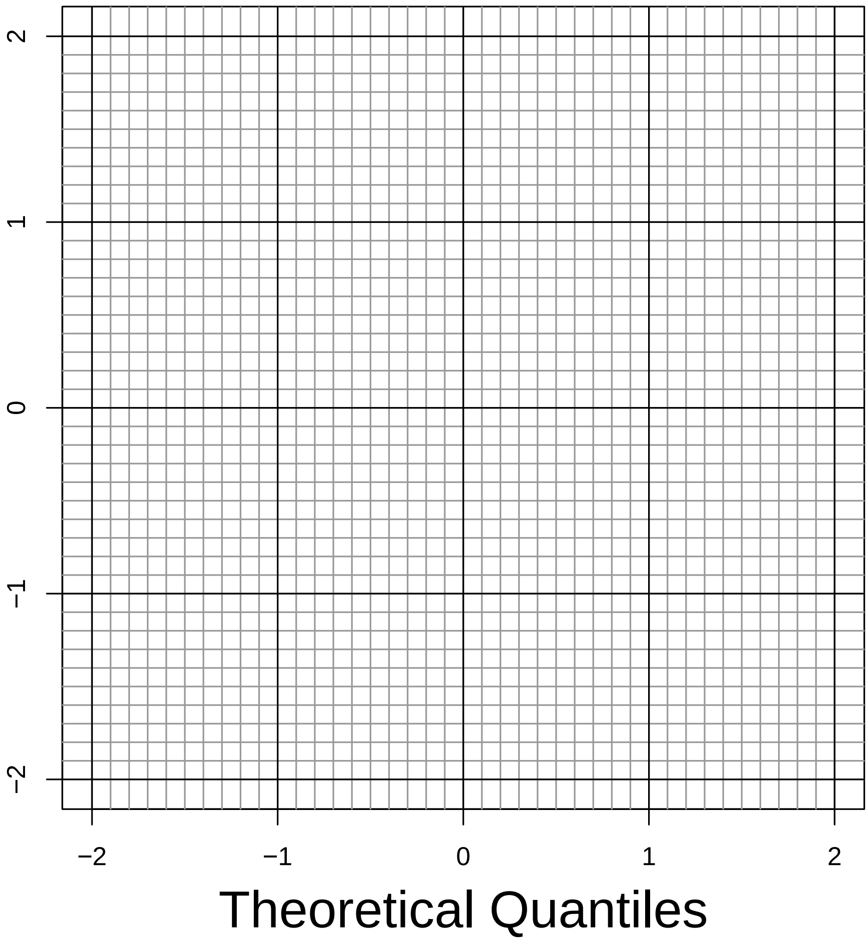
b) Fill the following table as the basis for generating the required data for the QQ-plot below. For identical values, randomly assign them to the related adjacent ranks (e.g., two identical values can have 3rd and 4th rank). Use R to find the required values for q (normal):

$flip_n$	ranks	j^*	q (normal)
	1		
	2		
	3		
	4		
	5		
	6		
	7		
	8		
	9		
	10		

Example solution:

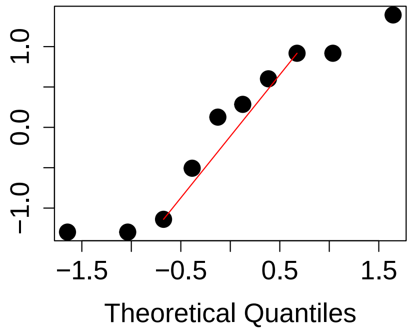
$flip_n$	ranks	j^*	q (normal)
-1.297	1	0.05	-1.64
-1.297	2	0.15	-1.04
-1.139	3	0.25	-0.67
-0.506	4	0.35	-0.39
0.1266	5	0.45	-0.13
0.2848	6	0.55	0.13
0.6012	7	0.65	0.39
0.9176	8	0.75	0.67
0.9176	9	0.85	1.04
1.3922	10	0.95	1.64

c) Complete the QQ-plot below and insert the qq-line. Are you deciding for or against a possible normal distribution? Explain your reasoning.



Example solution:

Decision more towards assuming normality. Also confirmed by Shapiro-Wilk Test. (p-value 0.2515)



Exercise 3 (Shapiro-Wilk Test Outlier Sensitivity)
Reproduce the box-plots from the lecture slides on the sensitivity of the Shapiro-Wilk normality test to a single outlier. To this end for each sample size $n \in \{100, 1\,000, 2\,500\}$ and each outlier $o \in \{4, 4.2, 4.4, \ldots, 5.8, 6\}$ repeat the following experiment 30 times:

- Sample n random values from an $\mathcal{N}(0, 1)$-distribution.
- Add the outlier o to the sample.
- Apply the Shapiro-Wilk test and store the p-value.

Plot the distribution of the p-values for each outlier split by the sample size n. Interpret the results.

Exercise 4 (χ^2-distribution properties)

1. Let $X_1, \ldots, X_p$ be independent identically $\mathcal{N}(\mu, \sigma^2)$-distributed random variables. Show that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2(p)$$

   where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. Now let $U_i \sim \chi^2_{p_i}, i = 1, \ldots, l$ be l independent random variables. Show that

$$\sum_{i=1}^{l} U_i \sim \chi^2_{p_1+\ldots+p_l}.$$

Example solution:

1. We keep in mind the definition of a χ^2-distribution (sum of squared i.i.d. standard normal random variables). First of all we observe

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu$$

   by the weak law of large numbers. With this we obtain:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu)^2 \\ &= \sum_{i=1}^{n}\left(\frac{\sigma}{\sigma}\cdot(X_i - \mu)\right)^2 \\ &= \sigma^2\sum_{i=1}^{n}\underbrace{\left(\frac{X_i - \mu}{\sigma}\right)^2}_{\sim\mathcal{N}(0,1)}.\end{aligned}$$

   Now we have a sum of p i.i.d. $\mathcal{N}(0, 1)$ random variables that we know is χ^2_p-distributed and scaled by a factor of σ^2.

2. The proof is very straight forward. We first observe that

$$\sum_{i=1}^{l} U_i = \sum_{i=1}^{l} \sum_{j=1}^{p_i} X_j.$$

We know that all the X_j are standard normal random variables. The claim thus follows directly.

Exercise 5 (Outlier Detection Study)
Load the `heptathlon` data set from the `HSAUR3` R package. Familiarise yourself with the data set, look for possible outliers in the data and interpret your findings

Example solution:
For the solution, please see the notebook file on PANDA.

Exercise 6 ((Bi-variate) Normal Distribution ★)
Let the density of a bi-variate variable $Z = (X_1,\ X_2)^T$ be given by the following expression:

$$f_Z(x_1, x_2) = \frac{1}{4\pi \cdot \sqrt{1-\rho^2}} \cdot \left(\exp\Big(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\Big) + \exp\Big(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\Big)\right)$$

Proof that the marginal distributions of $f_Z(x_1, x_2)$ are $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ with

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x_i^2}{2}\right), \quad i = 1, 2.$$

Hints:

- The marginal density is defined as follows: $f_{X_2}(x_2) = \int_{-\infty}^{+\infty} f_{(X_1,X_2)}(x_1, x_2)\, dx_1$ (analogous for the marginal density $f_{X_1}(x_1)$)
- Split the bivariate density into two terms and integrate each term on its own (or better: get rid of the two terms within the brackets by simplifying the density)
- Split the exponential terms into a product of two terms by making use of artificially adding $+\ (x_2\rho)^2 - (x_2\rho)^2$ in the numerator
- Also, try to use the properties of a normal distribution and densities in general!
- Don't be frustrated if you fail, this is not an easy task, but try to do your best!

Example solution:
Given a bivariate variable $Z = (X_1,\ X_2)^T$ with density function

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{1}{4\pi \cdot \sqrt{(1-\rho^2)}} \cdot \left(\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) + \ \exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\right).$$

Actually the density can be simplified as follows:

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{1}{2\pi \cdot \sqrt{(1-\rho^2)}} \cdot \left(\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\right).$$

First, have a look at the exponential term:

$$\begin{aligned}
\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) &= \exp\left(-\frac{x_1^2 - 2x_1x_2\rho + (x_2\rho)^2 - (x_2\rho)^2 + x_2^2}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{x_1^2 - 2x_1x_2\rho \ + \ (x_2\rho)^2}{2\cdot(1-\rho^2)} - \frac{x_2^2 - (x_2\rho)^2}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{(x_1 - x_2\rho)^2}{2\cdot(1-\rho^2)} - \frac{x_2^2\cdot(1-\rho^2)}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right) \qquad (*)
\end{aligned}$$

Now, one can integrate the bivariate density w.r.t. X_1 (and thus, derive the marginal distribution of X_2):

$$\begin{aligned}
f_{X_2}(x_2) &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\, dx_1\\
&\overset{(*)}{=} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\cdot\sqrt{2\pi}\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)\, dx_1\\
&= \frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)\cdot\underbrace{\int_{-\infty}^{+\infty}\underbrace{\frac{1}{\sqrt{2\pi}\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)}_{\text{density function of normal dist. with exp. value } x_2\rho \text{ and variance } 1-\rho^2}\, dx_1}_{=1}\\
&= \frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)
\end{aligned}$$

As X_1 and X_2 are symmetric (within the bivariate function), the marginal distribution of X_1 is:

$$f_{X_1}(x_1) = \ldots = \frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_1^2\right)$$

Longer and cumbersome solution sticking to two terms:

First, have a look at the first exponential term:

$$\begin{aligned}
\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) &= \exp\left(-\frac{x_1^2 - 2x_1x_2\rho + (x_2\rho)^2 - (x_2\rho)^2 + x_2^2}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{x_1^2 - 2x_1x_2\rho \ + \ (x_2\rho)^2}{2\cdot(1-\rho^2)} - \frac{x_2^2 - (x_2\rho)^2}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{(x_1 - x_2\rho)^2}{2\cdot(1-\rho^2)} - \frac{x_2^2\cdot(1-\rho^2)}{2\cdot(1-\rho^2)}\right)\\
&= \exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right) \qquad (*)
\end{aligned}$$

Analogously, the second exponential term can be written as:

$$\begin{aligned}
\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) &= \exp\left(-\frac{x_1^2 + 2x_1x_2\rho + (x_2\rho)^2 - (x_2\rho)^2 + x_2^2}{2\cdot(1-\rho^2)}\right)\\
&= \dots\\
&= \exp\left(-\frac{1}{2}\cdot\frac{(x_1+x_2\rho)^2}{1-\rho^2}\right)\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right) \qquad (**)
\end{aligned}$$

Furthermore, one can split the bivariate density function into two parts:

$$\begin{aligned}
& f_{(X_1,X_2)}(x_1,x_2)\\
=\ & \frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\left(\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) + \exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\right)\\
=\ & \frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right) + \frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)
\end{aligned}$$

Now, one can integrate the bivariate density w.r.t. X_1 (and thus, derive the marginal distribution of X_2):

$$\begin{aligned}
f_{X_2}(x_2) &= \int_{-\infty}^{+\infty} f_{(X_1,X_2)}(x_1,x_2)\,dx_1\\
&= \underbrace{\int_{-\infty}^{+\infty}\frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)dx_1}_{(i)} + \underbrace{\int_{-\infty}^{+\infty}\frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)dx_1}_{(ii)}
\end{aligned}$$

First, have a closer look at (i):

$$\begin{aligned}
& \int_{-\infty}^{+\infty}\frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)dx_1\\
\overset{(*)}{=}\ & \int_{-\infty}^{+\infty}\frac{1}{2\cdot\sqrt{2\pi}\cdot\sqrt{2\pi}\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)dx_1\\
=\ & \frac{1}{2\cdot\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)\cdot\underbrace{\int_{-\infty}^{+\infty}\underbrace{\frac{1}{\sqrt{2\pi}\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{1}{2}\cdot\frac{(x_1-x_2\rho)^2}{1-\rho^2}\right)}_{\text{density function of normal dist. with exp. value } x_2\rho \text{ and variance } 1-\rho^2}dx_1}_{=1}\\
=\ & \frac{1}{2\cdot\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)
\end{aligned}$$

Analogous, (ii) can be written as:

$$\int_{-\infty}^{+\infty}\frac{1}{4\pi\cdot\sqrt{(1-\rho^2)}}\cdot\exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)dx_1 = \dots = \frac{1}{2\cdot\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\cdot x_2^2\right)$$

Thus, the marginal distribution of X_2 can be summarised by:

$$\begin{aligned}
f_{X_2}(x_2) &= \int_{-\infty}^{+\infty} f_{(X_1,X_2)}(x_1, x_2)\ dx_1 \\
&= \int_{-\infty}^{+\infty} \frac{1}{4\pi \cdot \sqrt{(1-\rho^2)}} \cdot \exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\ dx_1 + \int_{-\infty}^{+\infty} \frac{1}{4\pi \cdot \sqrt{(1-\rho^2)}} \cdot \exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2\cdot(1-\rho^2)}\right)\ dx_1 \\
&= \frac{1}{2\cdot\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\cdot x_2^2\right) + \frac{1}{2\cdot\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\cdot x_2^2\right) \\
&= 2 \cdot \frac{1}{2\cdot\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\cdot x_2^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\cdot x_2^2\right)
\end{aligned}$$

As X_1 and X_2 are symmetric (within the bivariate function), the marginal distribution of X_1 is:

$$f_{X_1}(x_1) = \ldots = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\cdot x_1^2\right)$$