MS-E2112 Multivariate Statistical Analysis – 19.4.2024

Materials: In this exam you may have your pens and pencils, a ruler and an eraser. On top of that you may have one A4 of notes. The rules for the note are: size A4, text on one side only, it must be hand-written, your name has to be on the top right corner of the note.

Answer to all the questions.
In problem 1, you do not have to justify your answer. In all the other problems, **justify your solutions and write down all your calculations.**

---

1. True or False (6 p.)

   Determine whether the statement is true or false. In this problem, you do not have to justify your answers. Simply state whether the statement is true or false.(Every correct answer +1 p., every wrong answer -1 p., no answer 0 p.)

   (a) The asymptotical breakdown point of the univariate sample mean is 0.

   (b) The empirical influence function of the sample mean is bounded.

   (c) Attraction-repulsion indices can not be negative.

   (d) Multiple correspondence analysis (MCA) is a method that can not be applied for continuous variables.

   (e) Assume that we have two groups of variables and that we analyse the relationship between the groups of variables by applying canonical correlation analysis. Assume that in the first group, we have 2 variables, and in the second group, we have 4 variables. We now obtain 6 nonzero canonical correlations.

   (f) The initial K centers do not have an effect on the results of the moving centers clustering methods (K-means clustering methods).

2. Depth based classification (6 p.)

Consider the following bivariate samples:

$$A = \{(1.5, -1.1), (1.3, 1.2), (2.2, 1.0), (1.0, 0.2), (2.1, 0.3), (0.0, -0.6), (1.6, -1.3)\}$$

and

$$B = \{(-0.4, 0.2), (0.3, -2.0), (1.5, -0.5), (-1.6, -1.2), (-2.5, 0.1)\}.$$

(a) (1 p.)

Draw a scatter plot where both samples are in the same figure, but use different marker types for points from sample A and for points from sample B.

(b) (5 p.)

Perform half-space depth based classification to allocate new points, $c_1 = (0.5, -0.5)$, $c_2 = (-0.7, -0.7)$ and $c_3 = (2.5, 0.5)$, into these sets. If there is a tie in the classification, use the information given by the scatter plot in the allocation.

i. What is the half-space depth of the point $c_1 = (0.5, -0.5)$ with respect to the sample A and with respect to the sample B? Based on that (and the scatter plot), would you classify the point c_1 into sample A or into sample B? Justify your answer.

ii. Calculate the half-space depth of the point $c_2 = (-0.7, -0.7)$ with respect to the sample A and with respect to the sample B? Based on that (and the scatter plot), would you classify the point c_2 into sample A or into sample B? Justify your answer.

iii. What is the half-space depth of the point $c_3 = (2.5, 0.5)$ with respect to the sample A and with respect to the sample B? Based on that (and the scatter plot), would you classify the point c_3 into sample A or into sample B? Justify your answer.

3. Scatter functionals (6 p.)

Let $x = (u, v)^T$ denote a bivariate random vector with a cumulative distribution function F_x. Assume that x is spherically distributed. That is

$$x = z + \mu,$$

where $\mu \in \mathbb{R}^2$, and $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{2\times 2}$. Let S be an affine equivariant scatter functional and assume that $S(F_x)$ exists as a finite quantity. Prove that $S(F_x)$ is proportional to the 2×2 identity matrix.

4. Principal component analysis (6 p.)

Let x denote a p-variate random vector with finite mean $E[x] = \mu$, and finite covariance matrix $E[(x - \mu)(x - \mu)^T] = \Sigma$.

(a) (1 p.)
Describe (in 2-3 sentences) the core idea behind principal component transformation for the random vector x.

(b) (1 p.)
Explain how the theoretical principal components are calculated.

(c) (2 p.)
Assume now that you have a p-variate sample $x_1, x_2, ..., x_{688}$. You conduct sample principal component analysis and you obtain the score vectors. Explain how you can measure the quality of the representation of x_{72} by the first principal axis.

(d) (2 p.)
Explain how and under what conditions you can robustify principal component analysis.

BONUS QUESTION (2 p.): Choosing minimum linkage in agglomerative hierarchical clustering may lead to chaining. Explain, in 2-5 sentences and giving an example figure, what that means and why it may happen.