MS-E2112 Multivariate Statistical Analysis – 7.6.2023

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) If the influence function of a functional Q is bounded (with respect to L_2 norm), then the asymptotical breakdown point of Q can not be 0.

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

   (c) In MCA, rare modalities have neglible/small effect on the analysis.

   (d) 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 7 variables, and in the second group, we have 3 variables. We now obtain at least 7 nonzero canonical correlations.

   (e) Fisher's linear discriminant analysis is based on maximizing the ratio of between groups dispersions and within group dispersions.

   (f) According to Zuo and Serfling, depth functions should be invariant under affine transformations.

2. Clustering (6 p.)

Consider the following bivariate sample:

$A = (1, -1.2), B = (-3, 3), C = (1, -2), D = (2, -2), E = (0, 0), F = (-1.5, -1.5)$.

(a) (1 p.)

Draw a scatter plot of the data.

(b) (3 p.)

Perform agglomerative hierarchical clustering on the data. Use Euclidian distance as the distance measure and in clustering, select **maximum distance** as the linkage function. Draw the corresponding classification tree. If you choose the number of the final clusters to be two, what are the two clusters? (Note that you do not have to calculate all the pairwise Euclidean distances. You can use the scatter plot from part a to assess which points are closest to each other.)

(c) (2 p.)

Perform agglomerative hierarchical clustering on the data. Use Euclidian distance as the distance measure and in clustering, select **minimum distance** as the linkage function. If you choose the number of the final clusters to be two, what are the two clusters?

3. Scatter functionals (6 p.)

Let x denote a p-variate random vector with a cumulative distribution function F_x. Assume that

$$x = \Omega z + \mu,$$

where $\mu \in \mathbb{R}^p$, $\Omega \in \mathbb{R}^{p \times p}$, Ω is full rank, and $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{p \times p}$. Let S be an affine equivariant scatter functional and assume that $S(F_x)$ exists as finite quantity. Show that $S(F_x)$ is proportional to $\Omega\Omega^T$.

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 the idea behind principal component transformation for the random vector x using 2-3 sentences.

(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 an individual r by the first principal axis.

(d) (2 p.)

Explain how and under what conditions you can robustify principal component analysis.

BONUS QUESTION (2 p.):

Consider the following bivariate sample:

$S = \{(0.4, 1.5), (-2.0, 1.3), (-2.4, -0.4), (-0.9, -1.8), (-0.8, 1.2), (-2.5, 0.8), (-2.1, 2.3)\}$.

What is the half-space depth of the point $(-1.2, -0.8)$ with respect to the sample S?