MS-E2112 Multivariate Statistical Analysis – 2021
Online Exam
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 (4 p.)

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

   (a) All affine equivariant scatter estimators estimate the same population quantity even when the data comes from a skew distribution.

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

   (c) 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 6 variables, and in the second group, we have 4 variables. We now obtain max 4 pairs of canonical variables.

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

Table 1: Cookie tasting data (observed frequencies):

	below the average	average	above the average	
brand A	180	20	270	470
brand B	30	200	200	430
	210	220	470	900

2. Attraction-repulsion Indices (6 p.)

   A group of 900 high-school students were asked to taste cookies. Each student was given a cookie that was either of brand A or brand B. All the cookies looked the same. Students were asked to rate the taste of the cookie as "below the average", "average" or "above the average". Table 1 above displays the collected data as a two-way contingency table.

   (a) Display the data as a relative frequency table. How many percentages of the students tasted the cookie brand A? How many percentages of the students tasted the cookie brand A and rated the cookie as "above the average"?

   (2 p.)

   (b) Form the corresponding attraction-repulsion matrix. What is the attraction repulsion index that corresponds to brand A and category "above the average"? What is the attraction repulsion index that corresponds to brand A and category "below the average"? Interpret this finding.

   (4 p.)

3. Robustness (4 p.)

   (a) Derive the finite sample breakdown point and the asymptotic breakdown point of the sample median.

   (2 p.)

   (b) Derive the empirical influence function of the sample mean.

   (2 p.)

4. Principal Component Analysis (2 p.)

   Let $x \in \mathbb{R}^{p\times 1}$ be a p-variate random vector with mean μ and covariance matrix Σ. Let

$$\Sigma = \Gamma\Lambda\Gamma^T,$$

where the column vectors of Γ are orthonormal eigenvectors of Σ, and $\Lambda = diag(\lambda_1, ..., \lambda_p)$ is a diagonal matrix having its diagonal elements in decreasing order. Let $y = (y_1, ..., y_p)^T = \Gamma^T(x - \mu)$. Show that

(a) $E[y_i] = 0$

(b) $Var[y_i] = \lambda_i$

(c) $Cov[y_i, y_j] = 0, i \neq j$

(d) $Var[y_1] \geq Var[y_2] \geq \cdots \geq Var[y_p] \geq 0$.

5. Depth functions (4 p.)

   According to Zuo and Serfling, depth functions should fulfill four general properties. State the four properties and explain (using 2-3 sentences) what they mean.

6. Clustering (4 p.)

   Explain what is Agglomerative hierarchical clustering method. (Describe the algorithm, explain how the number of clusters can be chosen, and comment shortly how to choose the used distance and how to measure the distance between two groups.)

BONUS QUESTION (2 p.):

Consider the following bivariate sample:

$$S = \{(4.5, 1.5), (-1.5, -2.5), (2.5, -1.5), (1.0, 1.0), (-0.5, 1.5), (0.0, 4.5)\}.$$

What is the half-space depth of the point $(-1.0, -1.0)$ with respect to the sample S?