MS-E2112 Multivariate Statistical Analysis – 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) PCA transformation is invariant under affine transformations.

   (b) Classical PCA is a robust method.

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

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

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

   (f) Consider the half-space depth of a point x with respect to normal distribution with expected value vector μ and full rank covariance matrix Σ. Then the maximum value of the half-space depth is attained at $x = \mu$.

Table 1: Cookie tasting data (observed frequencies):

	Below the average	Average	Above the average	
Brand A	230	20	320	570
Brand B	50	300	80	430
	280	320	400	1000

2. Attraction-repulsion Indices (6 p.)

A group of 1000 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) (1 p.)

Display the data as a relative frequency table.

(b) (1 p.)

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”?

(c) (4 p.)

Calculate the attraction repulsion index that corresponds to brand A and category ”above the average” and the attraction repulsion index that corresponds to brand A and category ”below the average”. Interpret this finding.

3. Robustness (6 p.)

(a) (3 p.)

Derive the finite sample breakdown point and the asymptotic breakdown point of the sample median.

(b) (3 p.)

Derive the empirical influence function of the sample mean.

4. Canonical correlation analysis (6 p.)

Let x be a 3-variate random vector, let y be a 4-variate random vector. You conduct canonical correlation analysis to examine the relationships between the two sets of variables (given by x and y).

(a) (1 p.)

Describe the theoretical maximization problem in canonical correlation analysis.

(b) (1 p.)

Explain how the theoretical canonical vectors and canonical correlations are calculated.

(c) (1 p.)

Assume now that you have a sample $(x_1, y_1), (x_2, y_2), ..., (x_{785}, y_{785})$ and you conduct sample canonical correlation analysis. The obtained sample canonical correlations are $\hat{\rho}_1, \hat{\rho}_2$ and $\hat{\rho}_3$.
Explain how the canonical vectors and canonical correlations are estimated from the sample.

(d) (3 p.)

You continue analyzing the sample mentioned in part c. You decide to test the null hypothesis "all the canonical correlations are equal to zero" against the alternative hypothesis "at least one of the canonical correlations is not equal to zero" by using the following test statistic:

$$T = \ln(\prod_{k=1}^{3}(1 - \hat{\rho}_k^2)).$$

Explain, step by step, how you can be approximate the p-value of the test statistic by using permutations.

BONUS QUESTION (2 p.):

Consider the following bivariate sample:

$S = \{(0.4, 1.5), (-2.5, 1.0), (-2.4, -0.4), (-0.9, -1.8), (-0.5, 1.6), (-1.5, 0.8), (-2.1, 2.3)\}$.

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