

MS-E2112 Multivariate Statistical Analysis – Fall 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 influence function of a functional Q is bounded (with respect to L_2 norm) if and only if the asymptotical breakdown point of Q is equal to 0.
- (b) All affine equivariant scatter estimators estimate the same population quantity even when the data comes from a skew distribution.
- (c) Attraction-repulsion indices are always larger than or equal to 1.
- (d) Multiple correspondence analysis (MCA) is a method that can be applied only for ordinal variables.
- (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 $x = \mu$.

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. Principal component analysis (6 p.)

Let $x = (x_1, x_2)^T$ denote a bivariate random vector with finite mean vector μ , and finite covariance matrix Σ . Let $y = (y_1, y_2)^T = \Gamma^T(x - \mu)$, where $\Gamma \in \mathbb{R}^{2 \times 2}$ is orthogonal, $\Gamma^T \Sigma \Gamma = \Lambda = \text{diag}(\lambda_1, \lambda_2)$ and $\lambda_1 \geq \lambda_2$. Let σ_{11} denote the 1st diagonal element of Σ , let γ_{12} denote the 12 element of Γ and let ρ_{12} denote the Pearson correlation coefficient between the random variables x_1 and y_2 .

Show that

$$\rho_{12} = \frac{\gamma_{12} \lambda_2}{\sqrt{\sigma_{11} \lambda_2}}.$$

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 (in 2-3 sentences) the core idea behind canonical correlation analysis for the random vectors x and y , and state the corresponding mathematical optimization problem that is solved in canonical correlation analysis.

(b) (1 p.)

Explain how the solution to the maximization problem, that is described in part a, is solved. That is, explain how the theoretical canonical vectors and canonical correlations are calculated.

(c) (1 p.)

Assume now that you have a sample $(x_1, y_1), (x_2, y_2), \dots, (x_{352}, y_{352})$ and you conduct sample canonical correlation analysis. 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. Let $\hat{\rho}_1, \hat{\rho}_2$ and $\hat{\rho}_3$ denote the obtained sample canonical correlations. 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:

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

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

BONUS QUESTION (2 p.): Let x denote a p -variate random vector with a cumulative distribution function F_x . Assume that x is centrally symmetric about 0. That is $x \sim -x$. Let T be an affine equivariant location functional and assume that $T(F_x)$ exists as a finite quantity. Prove that $T(F_x) = 0$.