

MS-E2112 Multivariate Statistical Analysis – 10.4.2025

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 median is  $\frac{1}{2}$ .
- (b) The componentwise multivariate median is affine equivariant.
- (c) The value of a sample attraction-repulsion index is always smaller than or equal to 1.
- (d) Multiple correspondence analysis (MCA) is a method that can be applied for both qualitative categorical variables and categorized quantitative variables.
- (e) In MCA, modalities with small probability mass have negligible/small effect on the analysis.
- (f) Fisher's linear discriminant analysis is based on maximizing the ratio of between groups dispersions and within group dispersions.

2. Depth based classification (6 p.)

Consider the following bivariate samples:

$$A = \{(-1.0, 2.0), (1.5, 1.0), (-1.5, 0.5), (1.0, 1.5), (0.0, -0.5)\}$$

and

$$B = \{(0.0, -2.0), (-1.0, -1.0), (2.0, -0.5), (1.0, -1.0), (-2.5, 0.0), (-2.0, 0.5), (-1.5, -2.0)\}$$

(a) (0 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) (6 p.)

Perform half-space depth based classification to allocate new points,  $c_1 = (0.0, 0.4)$ ,  $c_2 = (0.0, -0.4)$  and  $c_3 = (0.0, -2.4)$ , 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  $c_1 = (0.0, 0.4)$  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  $c_2 = (0.0, -0.4)$  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  $c_3 = (0.0, -2.3)$  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. Principal component analysis (6 p.)

Let  $x = (x_1, x_2)^T$  denote a bivariate random vector with finite mean vector  $\mu$ , and finite covariance matrix  $\Sigma$ . 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  $\lambda_1 > \lambda_2$ . Let  $\sigma_{11}$  denote the 1st diagonal element of  $\Sigma$ , let  $\gamma_{12}$  denote the element that is on the row 1 and column 2 of the matrix of  $\Gamma$ , and let  $\rho_{12}$  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 5-variate random vector. you have a sample  $(x_1, y_1), (x_2, y_2), \dots, (x_{532}, y_{532})$  and you conduct sample canonical correlation analysis to examine the relationships between the two sets of variables (given by  $x$  and  $y$ ).

(a) (3 p.)

Describe the theoretical maximization problem in canonical correlation analysis, explain how the theoretical canonical vectors and canonical correlations are calculated, and explain how the canonical vectors and canonical correlations are estimated from the sample.

(b) (3 p.)

You obtain sample canonical correlations  $\hat{\rho}_1, \hat{\rho}_2$  and  $\hat{\rho}_3$ . 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:

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

The random vectors  $x$  and  $y$  are not normally distributed. Explain, step by step, how you can approximate the  $p$ -value of the test statistic by using permutations.

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