MS-E2112 Multivariate Statistical Analysis – 2017 24.5.2017
Exam
Answer to all the questions.

You are allowed to have pens and pencils, an eraser and a ruler, a calculator and one size A4 note (handwritten, text on one side only, name on the top right corner).

1. True or False (6 p.)

   Determine whether the statement is true or false. A statement is true if it is always true — otherwise it is false. (Every correct answer +3/8 p., every wrong answer -3/8 p., no answer 0 p.)

   (a) PCA transformation is invariant under affine transformations.

   (b) PCA is sensitive to heterogenous scaling of the variables.

   (c) Classical PCA is a robust method.

   (d) 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.

   (e) The empirical influence function of the sample mean vector is bounded.

   (f) In bivariate correspondence analysis, PCA is applied to scaled and shifted contingency tables of relative frequencies.

   (g) Multiple correspondence analysis (MCA) is based on applying bivariate correspondence analysis on the so called complete disjunctive table.

   (h) Whereas PCA relies on euclidian distances, MCA relies on chi-square distances.

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

   (j) Canonical correlation analysis focuses on relationships between groups of variables.

   (k) Assume that we perform canonical correlation analysis to two groups of variables. Assume that in the first group, we have 6 variables, and in the second group, we have 4 variables. We now obtain max 6 pairs of canonical variables.

(l) Discriminant analysis is a method for splitting a set of individuals into unknown homogenous groups. F

(m) In discriminant analysis, sample misclassification rates grossly ~~over-~~ under-estimate the true misclassification rates.

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

(o) The initial K centers do not have an effect on the results of the Moving centers clustering methods (K-means clustering methods).

(p) If the sample size n is large, clustering is usually performed by considering all the possible partitions of the n data points into K classes, $K = 1, 2, ..., n$.

Statement	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
True		X		X		X	X	X		X				X		
False	X		X		X				X		X	X	X		X	X

2. Multiple Correspondence Analysis (6 p.)

A survey was administered to 201 first year engineering students. The goal of the survey was to understand what are the factors influencing the success of a student in her/his first year as a university student. Variables considered are given below:

- Attendance to class: *low, average, high*
- Average time spent studying for the exams: *less than 4h., between 4h. and 8h., more than 8h.*
- Activity in class: *low, high*
- Participation to additional guidance: *low, high*

Use the picture and the eigenvalues (next page) to justify your answers.

(a) What is the total variance of the variables?

(b) How much of the total variance do the first three MCA components explain?

(c) Why do you think that the modality *less than 4h.* is far away from the center?

(d) Based on the picture, does it seem that students with average attendance to class are active in class? Justify!

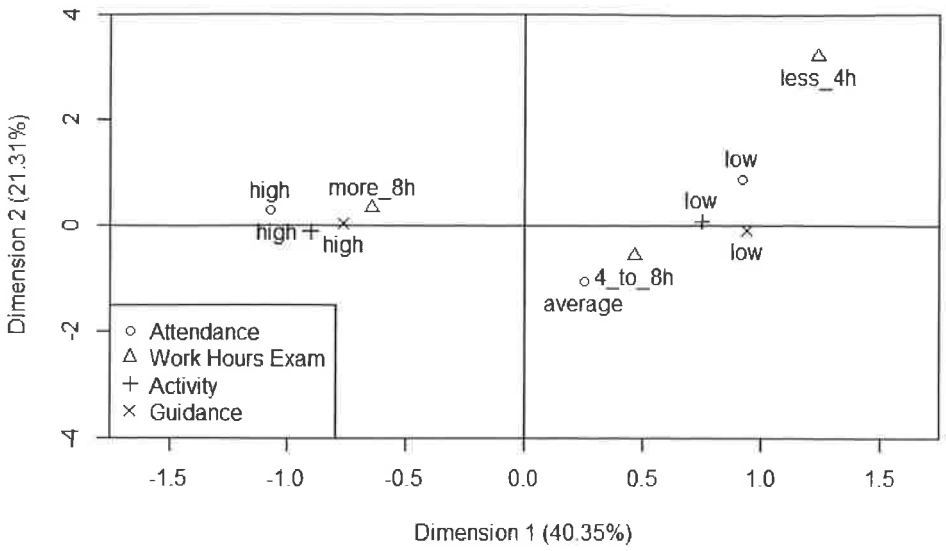

Table 1: Eigenvalues (rounded) associated with the MCA transformation:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.61	0.32	0.21	0.17	0.11	0.08

3. Multivariate Location and Scatter (6 p.)

   (a) Let μ_x denote the mean vector, and Σ_x denote the covariance matrix of a p-variate random vector x. Let $A \in \mathbb{R}^{p \times p}$ be nonsingular, and let $b \in \mathbb{R}^p$. Let $y = Ax + b$. Let μ_y denote the mean vector, and Σ_y denote the covariance matrix of the random vector y. Show that $\mu_y = A\mu_x + b$ and that $\Sigma_y = A\Sigma_x A^T$.

   (b) Sow that, under the assumption of central symmetry, all affine equivariant location functionals measure the same population quantity.

4. Clustering (6 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:

$$\{(2.1, 1.3), (-1.5, -0.3), (1.1, -1.4), (1.1, 1.1), (-0.2, 1.4)\}.$$

What is the half-space depth of the data point $(1.1, 1.1)$?