MS-E2112 Multivariate Statistical Analysis – 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 is sensitive to heterogenous scaling of the variables.

   (b) The influence function can be seen as a measure of global robustness, and the breakdown point can be seen as a measure of local robustness.

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

   (d) The componentwise multivariate median is affine equivariant.

   (e) All affine equivariant scatter estimates do estimate the same population quantity.

   (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 within 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.

(m) In discriminant analysis, sample misclassification rates grossly overestimate the true misclassification rates.

(n) All existing classification methods rely heavily on the assumption of multivariate normality.

(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																
False																

2. Clustering (6 p.)

Consider the following bivariate sample:

$$A = (2.0, 1.0), B = (-1.5, 0.0), C = (1.0, -1.5), D = (1.0, 1.0), E = (0.0, 1.5).$$

- Draw a scatter plot of the data. (1 p.)
- Perform agglomerative hierarchical clustering on the data. Use Euclidian distance and maximum distance to measure the distance between groups. Draw the corresponding classification tree. If you choose the number of the final clusters to be two, what are the two clusters? (2.5 p.)
- Perform agglomerative hierarchical clustering on the data. Use Euclidian distance and minimum distance to measure the distance between groups. Draw the corresponding classification tree. If you choose the number of the final clusters to be two, what are the two clusters? (2.5 p.)

3. Principal Component Analysis (6 p.)

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

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

where the column vectors of Γ are the orthogonal 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) $Var[y_1] \geq Var[y_2] \geq \cdots \geq Var[y_p] \geq 0$.

4. Depth functions (6 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.

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 $(1.1, 1.1)$?