TÉCNICO LISBOA

Departamento de Matemática

Multivariate Analysis

Master in Mathematics and Applications

2^{nd} Exam — 1^{st} Semester – 2012/2013

Duration: 3 hours — 28/01/2013 – 3 pm

Please justify conveniently your answers

Group I — 6.0 points

1. Supose $\boldsymbol{X} = (X_1, X_2, X_3)^t$ has a multivariate normal distribution. Show that if: (i) X_1 and $X_2 + X_3$ are independent, (ii) X_2 and $X_1 + X_3$ are independent, and (iii) X_3 and $X_1 + X_2$ are independent, then X_1, X_2, and X_3 are independent random variables. (1.5)

2. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be two independent random vectors, where $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Consider also two independent random samples, with sizes n_1 e n_2 from each population. Let $\boldsymbol{C}$ be a $k \times p$ full-rank matrix of constants.

   (a) Prove that: $\boldsymbol{C}\bar{\boldsymbol{X}}_i \sim \mathcal{N}_k(\boldsymbol{C}\boldsymbol{\mu}_i, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^t/n_i)$, $i = 1, 2$. (1.0)

   (b) Prove that: (1.5)

   $$\frac{n_1 n_2}{n_1 + n_2}(\boldsymbol{C}\bar{\boldsymbol{X}}_1 - \boldsymbol{C}\bar{\boldsymbol{X}}_2 - (\boldsymbol{C}\boldsymbol{\mu}_1 - \boldsymbol{C}\boldsymbol{\mu}_2))^t (\boldsymbol{C}\boldsymbol{S}_p\boldsymbol{C}^t)^{-1}(\boldsymbol{C}\bar{\boldsymbol{X}}_1 - \boldsymbol{C}\bar{\boldsymbol{X}}_2 - (\boldsymbol{C}\boldsymbol{\mu}_1 - \boldsymbol{C}\boldsymbol{\mu}_2)) \sim T^2_{(k, n1+n2-2)},$$

   where $\boldsymbol{S}_p$ is the centered estimator of $\boldsymbol{\Sigma}$.

   (c) A biologist measured the following four variables: (2.0)

   x_1 – the distance of the transverse groove from the posterior border of the prothorax (μm),

   x_2 – the length of the elytra (in 0.01 mm),

   x_3 – the length of the second antennal joint (μm),

   x_4 – the length of the third antennal joint (μm),

   on two species of flea beetles. He observed $n_1 = 19$ *haltica oleraceas* and $n_2 = 20$ *haltica carduorum*, obtaining the following results:

   $$\bar{\boldsymbol{x}}_1 = (194.5, 267.0, 137.4, 186.0)^t, \quad \bar{\boldsymbol{x}}_2 = (179.4, 292.0, 157.6, 209.4)^t, \text{ and}$$

   $$\boldsymbol{S}_p^{-1} = \begin{pmatrix} 1.35 & -0.51 & 0.15 & -0.23 \\ -0.51 & 0.67 & -0.50 & -0.07 \\ 0.15 & -0.50 & 1.42 & -0.08 \\ -0.23 & -0.07 & -0.08 & 0.65 \end{pmatrix} \times 10^{-2}.$$

   Use (2b) to test $H_0 : \boldsymbol{C}\boldsymbol{\mu}_1 = \boldsymbol{C}\boldsymbol{\mu}_2$, with

   $$\boldsymbol{C} = \begin{pmatrix} 1 & -3 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \text{ and } \quad (\boldsymbol{C}\boldsymbol{S}_p\boldsymbol{C}^t)^{-1} = \begin{pmatrix} 0.06 & 0.06 \\ 0.06 & 0.57 \end{pmatrix} \times 10^{-2},$$

   at a 0.05 significance level. State the hypotheses, test statistic, decision rule, and conclusion.

Group II — 4.5 points

A team of investigators has the objective to form a measure of the Consumer Price Index (CPI), that summarizes how expensive or cheap are a given city's food items. To build this index they registered the average price, in cents per pound, of five food items (bread, burger, milk, oranges, and tomatoes) in 24 US cities. See the Appendix for details about this dataset.

1. How many principal components are of interest? Decide based on the original and standardized variables. (1.0)

2. Write and interpret the chosen principal components. Does the interpretation change if you consider the original or the standardized principal components? (1.5)

3. In your opinion, which type of data should be use to reach the objective of the research team? Why? (1.0)

4. Can principal component analysis be used to define CPI? In what way? What is the most and the least expensive city among Chicago, New York, and San Francisco? (1.0)

Group III **5.5 points**

Consider the following set of 6 objects, characterized by 5 binary variables:

$$\mathcal{X} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

1. Let $d_{ij} = 1 - s_{ij}$, where s_{ij} is the Jaccard's similarity coefficient between object i and j. Show that d_{ij} is a dissimilarity coefficient. (1.5)

2. Let $\boldsymbol{D}$ be the dissimilarity matrix associated with $\mathcal{X}$, obtained based on the Jaccard's coefficient. Calculate a, b, and c. (1.0)

$$\mathbf{D} = \begin{pmatrix} 0 & & & & & \\ a & 0 & & & & \\ 0.333 & 0.750 & 0 & & & \\ 0.750 & 0.333 & 0.667 & 0 & & \\ 0.500 & 0.800 & b & 1.000 & 0 & \\ 0.750 & 0.750 & 0.667 & c & 0.750 & 0 \end{pmatrix}$$

3. Apply agglomerative hierarchical clustering using complete linkage. (2.0)
Hint: If necessary, consider $a = 1/2$, $b = 3/4$, and $c = 2/3$.

4. Plot the dendrogram for the solution of the previous question. How many clusters do you recommend to consider. Justify your choice. Indicate the chosen partition. (1.0)

Group IV **4.0 points**

An observation $\boldsymbol{x}$ comes from one of two populations with prior probabilities $P(Y = 1) = P(Y = 2)$, $\boldsymbol{X}|Y = j \sim \mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$, and

$$f_{\boldsymbol{X}|Y=j}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{p/2}} \exp\left\{-(\boldsymbol{x} - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)/2\right\}.$$

1. Show that the classification rule which minimizes the total probability of misclassification is: Assign $\boldsymbol{x}_0$ as $\{Y = 1\}$ if $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. (2.0)

2. Use the beetle data, of exercise (I.2.c), to estimate the optimal classification rule. (1.5)

3. Classify an beetle characterized by $(189, 245, 137, 163)^t$. (0.5)

FORMULAE

- $T^2(p,n) = \frac{np}{n-p+1}F(p, n-p+1)$.
- $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(p)$
- If $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$ then

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^t \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \sim T^2(p, n_1 + n_2 - 2)$$

.