DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Multivariate Statistical Methods for Engineering and Management

Master in Industrial Engineering and Management

1st Semester – 2016/2017 **1st Exam**

13/01/2017 – 1:00 PM – Room: 0.9 **Duration: 3h**

Please justify conveniently your answers!

Group I 8.0 valores

A study examining differences in life satisfaction between young, middle and older women was conducted. Each woman who participated in the study completed a life satisfaction questionnaire. A high score on the test indicates a higher level of life satisfaction. Test scores are recorded below:

Young	Middle	Older
4	10	7
2	5	7
3	7	9
7	10	8
7	10	9
	7	12
$y_{1.} = 23$	$y_{2.} = 49$	$y_{3.} = 52$

with $\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 = 1018$.

(a) Describe the model that you consider more convenient for this situation, indicating the assumptions associated with the chosen model. (0.5)

(b) Obtain the analysis of variance table. (2.0)

(c) Test, at a 5% significance level, if the woman age has an effect in life satisfaction. State the hypotheses, test statistic, decision rule and conclusions. (1.5)

(d) Find a 95% confidence interval estimate for the difference between the mean value of life satisfaction for middle aged women and older women. Is it possible to conclude, at a 5% significance level, that the mean value of life satisfaction is the same for those two groups of women? (2.0)

(e) Test, at a 5% significance level, if the mean value of life satisfaction of young aged women is less or equal to 5. State the hypotheses, test statistic, decision rule and conclusions. What is the p-value of the test? (2.0)

Group II

5.0 valores

Consider a data set of examinations grades (0-20) in five exams (mechanics, physics, algebra, analysis and statistics) for 100 students. The three exams were "open-books" (algebra, analysis and statistics). The two remaining, mechanics and physics, were "closed-book". Based on R output given below answer to the following questions:

(a) Complete the values (`a` and `b`) missing in the R results. (1.0)

(b) How many principal components should be retained? Give your answer based one the percentage of the total sample variability explained by each sample principal component. (1.0)

(c) Write the first two sample principal components and interpret then. (1.0)

(d) Find the sample correlation between the first sample principal component and each variable. Compare the first sample principal component interpretation with your findings in part (c). (2.0)

```
> x<-princomp(data, cor = TRUE, scores = TRUE)
> (x$sdev)^2
Comp.1    Comp.2    Comp.3    Comp.4       Comp.5
3.1810    0.7396       a      0.3879       0.2465
> Loadings:
         Comp.1  Comp.2  Comp.3    Comp.4   Comp.5
mec      -0.3996  0.6454    b       -0.1458  -0.1307
phy      -0.4314  0.4415  -0.7050    0.2981  -0.1817
alg      -0.5033 -0.1291  -0.0370   -0.1086   0.8467
ana      -0.4570 -0.3879  -0.1362   -0.6662  -0.4222
sta      -0.4382 -0.4704   0.3125    0.6589  -0.2340
```

Group III

3.0 valores

Consider that $\mathbf{X} \in I\!R^3$ with $E(\mathbf{X}) = \mathbf{0}$ and correlation matrix $\boldsymbol{\rho} = \begin{pmatrix} 1.00 & 0.63 & 0.45 \\ & 1.00 & 0.35 \\ & & 1.00 \end{pmatrix}$ can be generate by the one-factor model. Admit that the correlation of the common factor f with the manifest variables are: $\rho_{X_1,f} = 0.9$, $\rho_{X_2,f} = 0.7$ and $\rho_{X_3,f} = 0.5$.

(a) State the assumptions associated with the factor analysis model. Compute the communalities and the specific variances. (1.5)

(b) The eigenvectors ($\boldsymbol{\gamma}_i$) and the eigenvalues (a_i), for $i = 1, 2, 3$, of the correlation matrix $\boldsymbol{\rho}$ are: (1.5)

	γ_1	γ_2	γ_3
X_1	0.63	0.22	0.75
X_2	0.59	0.50	-0.64
X_3	0.51	-0.84	-0.18
	$a_1 = 1.96$	$a_2 = 0.68$	$a_3 = 0.36$

Estimate the loading matrix and the specific variances using the principal component estimation method. Compute the estimated correlation matrix and the residual matrix. Comment.

Group IV 4.0 valores

Suppose we measure two variables X_1 and X_2 for each of eight objects and that the data mining task is clustering. The data set are given in the following table:

objects	x_1	x_2
O_1	2	10
O_2	2	5
O_3	8	4
O_4	5	8
O_5	7	5
O_6	6	4
O_7	1	2
O_8	4	9

The distance function is the square of the euclidean distance.

(a) The Ward method uses as a criterion for merger two groups A and B together in a group $C = A \cup B$: (2.0)

$$SSW_C - (SSWA + SSW_B),$$

where $SSW_A = \sum_{i \in A} \sum_{j=1}^{p} (x_{ijA} - \bar{x}_{jA})^2$ is the A group sum of squares. Similar expressions are used for the sum of squares of groups B and C.

Using the Ward method, assign the objects O_1, O_2, O_3 and O_4 into two clusters.

(b) Assume that initially the objects was assigned to $k = 3$ clusters as follows: (2.0)

$$C_1 = \{O_1, O_2, O_3\}, C_2 = \{O_4, O_5, O_6\}, C_3 = \{O_7, O_8\}.$$

Apply a non-hierarchical algorithm to obtain the final clusters.