DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Multivariate Statistical Methods for Engineering and Management

Master in Industrial Engineering and Management

1st Semester – 2019/2020 **2nd Exam**

31/01/2020 – 3:00 PM – Room: 1-2 **Duration: 3h**

Justify your answers

Group I 7.5 points

In a comparison of the cleaning action of four detergents, 20 pieces of white cloth were first soiled with India ink. The cloths were then washed under controlled conditions with 5 pieces for each of the detergents. Unfortunately three pieces of cloth were lost in the course of the experiment. Whiteness readings, made on the 17 remaining pieces of cloth, are shown below.

Detergent	Whiteness					$\sum_{j=1}^{n_i} y_{ij}$	$\sum_{j=1}^{n_i} y_{ij}^2$
A	87	81	85	86	79	418	34992
B	74	66	58			198	13196
C	73	78	57	69	63	340	23392
D	76	85	77	64		302	23026

The aim of this study is to check if there is no difference between the four brands as regards mean whiteness readings after washing.

(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. (1.5)

(c) Test, at a 5% significance level, the hypothesis of no difference between the four brands as regards mean whiteness readings after washing. State the hypotheses, test statistic, decision rule and conclusions. (1.5)

(d) Derive a 90% confidence interval estimate for the difference between the mean value of whiteness for detergents A and C. Is it possible to conclude, at a 10% significance level, that the mean value of of whiteness is the same for those two brands of detergents? (2.0)

(e) Evaluate, at a 1% significance level, the hypothesis of the mean value of whiteness for detergent A be at least 83. State the hypotheses, test statistic, decision rule and conclusions. Calculate the p-value and comment the result. (2.0)

Group II 5.5 points

In the Sibley's Bird Database of North American birds to gather data on a simple random sample of 100 bird species three factors were measured: length (**L**, in inches), wingspan (**W**, in inches), and weight (**WE**, in ounces). The sample covariance matrix of the data set (**S**) and its orthonormal eigenvectors ($\hat{\boldsymbol{\gamma}}_i$) and eigenvalues ($\hat{\lambda}_i$) are:

$$\mathbf{S} = \begin{pmatrix} 91.43 & 171.92 & 297.99 \\ & 373.92 & 545.21 \\ & & 1297.26 \end{pmatrix};$$

$\hat{\boldsymbol{\gamma}}_1$	$\hat{\boldsymbol{\gamma}}_2$	$\hat{\boldsymbol{\gamma}}_3$
0.22	0.25	0.94
0.41	0.85	-0.32
0.88	**a**	-0.08
$\hat{\lambda}_1$=1626.52	$\hat{\lambda}_2$= 128.99	$\hat{\lambda}_3$= **b**

(a) Compute the missing values **a** and **b**. (1.0)

(b) Compute the percentage of the total sample variability explained by each sample principal component. How many sample principal components should be retained? Justify your answer. (1.5)

(c) Write the first sample principal component and interpret it. (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)

Group III 3.0 points

A software company is planning to develop a new software package to support sophisticated accounting tasks and performed an initial study into the variables of importance for their new package. Suppose that for the four variables, $X_1, \ldots, X_4$, the correlation matrix is

$$\boldsymbol{\rho} = \begin{pmatrix} 1.00 & & & \\ 0.25 & 1.00 & & \\ 0.10 & 0.10 & 1.00 & \\ 0.40 & 0.28 & 0.12 & 1.00 \end{pmatrix}.$$

Admit that a one-factor model, f, was applied to generate the correlation matrix, with the following sample correlation with the four manifest variables: $r_{X_1,f} = 0.6$, $r_{X_2,f} = 0.4$, $r_{X_3,f} = 0.2$ and $r_{X_4,f} = 0.8$.

(a) Calculate the communalities and the unique variances associated with each manifest variable. Obtain the proportion of total variance that is explained by the common factor f. Comment the results. (1.5)

(b) Compute the residual matrix. Is the common factor f the best explanation for the observed correlations between the manisfest variables? (1.5)

Group IV 4.0 points

Suppose that we want to cluster the following 5 objects based on the bivariate observations (x_1, x_2):

object	1	2	3	4	5
x_1	1.0	1.0	6.0	8.0	8.0
x_2	1.0	2.0	3.0	2.0	0.0

(a) Build the euclidean distance matrix $\mathbf{D}$ between the objects. (1.0)

(b) Use the matrix $\mathbf{D}$ to cluster the objects with the complete linkage method. Draw the corresponding dendrogram. (2.0)

(c) Using the rule $\bar{h} + 1.25 s_h$, where $\bar{h}$ and s_h are the threshold distances mean and standard deviation, how many clusters would you recommend? (1.0)