INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

Multivariate Statistics Test

June 11, 2021 Duration: 2h30

I [16 points]

A study of the chemical composition of ice cores in the Arctic region was carried out[1] and gave rise to a set of observations on 64 ice cores (from different locations). Measurements were made of 16 elements or compounds (`Al`, `Ti`, `Fe`, `Mn`, `Ca`, `Mg`, `Na`, `K`, `P` - all in weight (%) - and `Ba`, `Sr`, `Cr`, `Ni`, `Sc`, `V` and `Zr` - in parts per million). Summary statistics, and the correlation matrix, for this dataset are given below.

```
    Min. 1st Qu.    Mean 3rd Qu.     Max.  St.Dev
Al  2.260   6.250   7.223   8.805   10.450   2.349
Ti  0.098   0.372   0.429   0.522    0.712   0.161
Fe  2.390   4.760   5.419   6.168    9.050   1.302
Mn  0.012   0.036   0.161   0.154    1.097   0.243
Ca  0.190   0.308   0.449   0.460    4.040   0.469
Mg  0.400   0.785   0.906   1.040    1.710   0.243
Na  0.370   1.315   1.450   1.693    1.890   0.349
K   0.450   1.070   1.679   2.055    2.910   0.614
P   0.021   0.045   0.080   0.100    0.239   0.040
Ba 78.000 485.000 632.484 693.000 2506.000 456.340
Sr 46.000  89.000 129.844 161.250  231.000  46.641
Cr 18.000  70.750  83.203 101.000  195.000  31.232
Ni 23.800  40.625  51.230  51.400  191.900  25.130
Sc  4.700  12.350  14.833  16.925   38.300   4.739
V  56.000 161.250 195.984 234.250  291.000  52.937
Zr 43.000 131.250 154.953 184.250  264.000  51.885
```

```
> round(cor(arctic[,3:18]),d=2)
      Al    Ti    Fe    Mn    Ca   Mg    Na     K     P    Ba    Sr   Cr    Ni   Sc     V    Zr
Al  1.00  0.93 -0.03  0.25  0.05 0.71  0.41  0.94  0.38  0.57  0.83 0.75  0.32 0.73  0.70  0.91
Ti  0.93  1.00 -0.11  0.19  0.07 0.68  0.40  0.89  0.38  0.60  0.76 0.76  0.26 0.62  0.64  0.97
Fe -0.03 -0.11  1.00  0.16 -0.02 0.18 -0.16 -0.13  0.22 -0.16 -0.08 0.02  0.11 0.07 -0.08 -0.12
Mn  0.25  0.19  0.16  1.00  0.14 0.30  0.25  0.17  0.44 -0.01  0.36 0.10  0.11 0.12  0.14  0.18
Ca  0.05  0.07 -0.02  0.14  1.00 0.43  0.16  0.12  0.13 -0.06  0.24 0.17  0.14 0.02 -0.08  0.03
Mg  0.71  0.68  0.18  0.30  0.43 1.00  0.59  0.71  0.33  0.27  0.76 0.63  0.28 0.41  0.46  0.58
Na  0.41  0.40 -0.16  0.25  0.16 0.59  1.00  0.44  0.32  0.02  0.63 0.40  0.05 0.17  0.40  0.36
K   0.94  0.89 -0.13  0.17  0.12 0.71  0.44  1.00  0.37  0.49  0.83 0.69  0.30 0.64  0.68  0.87
P   0.38  0.38  0.22  0.44  0.13 0.33  0.32  0.37  1.00 -0.02  0.38 0.25 -0.04 0.16  0.12  0.39
Ba  0.57  0.60 -0.16 -0.01 -0.06 0.27  0.02  0.49 -0.02  1.00  0.47 0.40  0.47 0.69  0.36  0.65
Sr  0.83  0.76 -0.08  0.36  0.24 0.76  0.63  0.83  0.38  0.47  1.00 0.65  0.45 0.67  0.64  0.74
Cr  0.75  0.76  0.02  0.10  0.17 0.63  0.40  0.69  0.25  0.40  0.65 1.00  0.23 0.53  0.52  0.73
Ni  0.32  0.26  0.11  0.11  0.14 0.28  0.05  0.30 -0.04  0.47  0.45 0.23  1.00 0.77  0.38  0.34
Sc  0.73  0.62  0.07  0.12  0.02 0.41  0.17  0.64  0.16  0.69  0.67 0.53  0.77 1.00  0.65  0.71
V   0.70  0.64 -0.08  0.14 -0.08 0.46  0.40  0.68  0.12  0.36  0.64 0.52  0.38 0.65  1.00  0.65
Zr  0.91  0.97 -0.12  0.18  0.03 0.58  0.36  0.87  0.39  0.65  0.74 0.73  0.34 0.71  0.65  1.00
```

1. A preliminary analysis of the data involved a *correlation matrix* Principal Component Analysis. Here are some partial results:

   ```
   > summary(arctic.acp1)
   Importance of components:
   ```

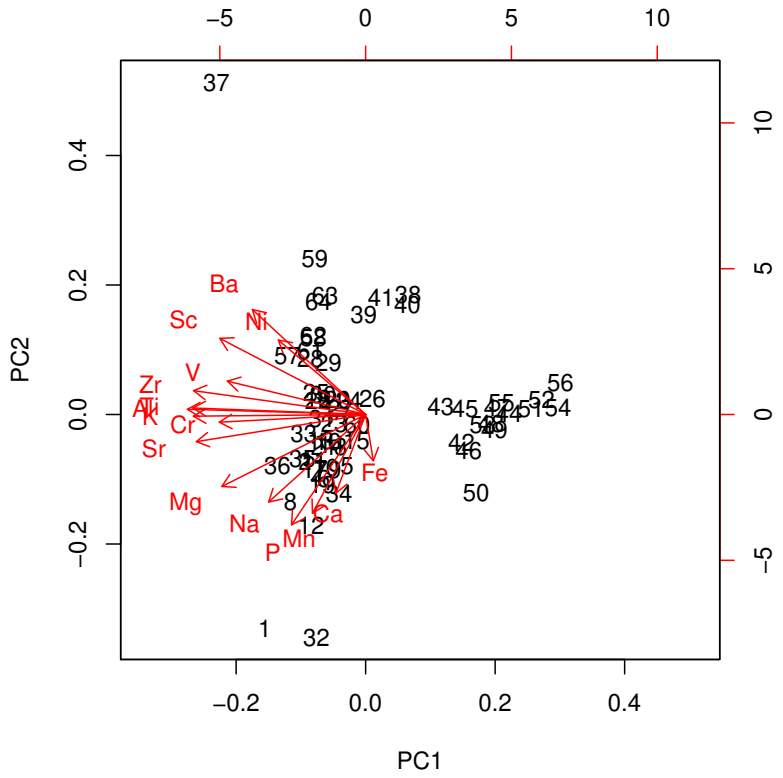
[1] N.C. Martinez, R.W. Murray, G.R. Dickens, and M. Molling (2009). *Discrimination of Sources of Terrigenous Sediment Deposited in the Central Arctic Ocean Through the Cenozoic*, Paleoceonography, Vol. 24, PA1210, doi:10.1029/2007PA001567, 2009

```
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9  [...]
Standard deviation     2.7899 1.3972 1.17937 1.07633 0.94507 0.89816 0.71258 0.67580 0.58056
Proportion of Variance 0.4865 0.1220 0.08693 0.07241 0.05582 0.05042 0.03174 0.02854 0.02107
Cumulative Proportion  0.4865 0.6085 0.69540 0.76781 0.82363 0.87405 0.90579 0.93433 0.95540

> round(cor(arctic.acp1$x[,1:3], arctic[,3:18]) , d=2)
       Al    Ti    Fe    Mn    Ca    Mg    Na     K     P    Ba    Sr    Cr    Ni    Sc     V    Zr
PC1 -0.95 -0.93  0.04 -0.28 -0.16 -0.77 -0.52 -0.92 -0.40 -0.61 -0.91 -0.78 -0.47 -0.78 -0.74 -0.92
PC2  0.03  0.03 -0.25 -0.53 -0.42 -0.39 -0.47 -0.01 -0.59  0.56 -0.15 -0.04  0.40  0.41  0.18  0.13
PC3 -0.07 -0.18  0.73  0.34  0.17  0.08 -0.31 -0.17  0.12  0.06  0.01 -0.12  0.60  0.36 -0.07 -0.13
```

(a) Comment the decision to carry out a PCA on the normalized data and discuss the results shown.

(b) Here is the corresponding two-dimensional biplot. Discuss it.



(c) What is the sum of the squared correlations between the fourth PC and all 16 original variables?

(d) Indicate an upper bound (smaller than 1) for the correlation of any original variable with the ninth Principal Component. Can that upper bound be attained? Justify your answers.

2. Someone who was not involved in collecting the data noticed that the label for each of the 64 observations began with a two-symbol code, as follows:

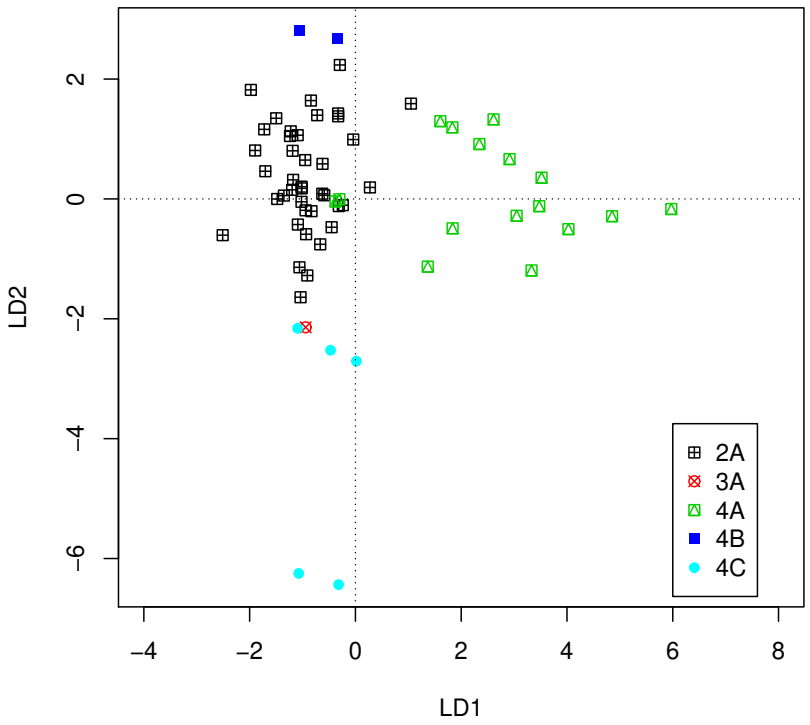
```
> arctic$groups
 [1] 4C 4C 4C 3A 4C 2A 4C 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A
[33] 2A 2A 2A 2A 2A 2A 2A 4B 4B 2A 2A 2A 2A 2A 2A 2A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A
> summary(arctic$groups)
2A 3A 4A 4B 4C
40  1 16  2  5
```

Suspecting that identical labels indicate observations collected in similar conditions, it was decided to discriminate the five groups of observations through a Linear Discriminant Analysis of the 16 numerical variables.

(a) Below are the discriminant capacities of the axes obtained using the `adl` function of Exercise 17, which produces results in accordance with the definitions in the slides. Comment these results.

```
> arctic.adl$val
[1] 2.50842423 1.79961417 0.79033773 0.06527797
```

(b) The figure below shows the scatterplot of the 64 points on the first two discriminant axes (with information from the `lda` command in the `MASS` package). Comment it.



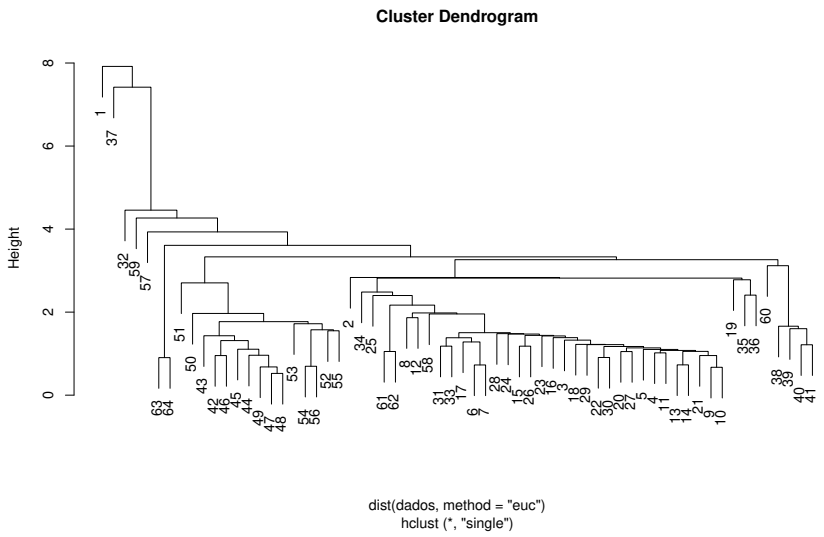
(c) Two new observations were created, one by taking the mean value, and the other by taking the third quartile, of all 64 observations on each of the 16 variables. Below are the coefficients on the linear discriminant axes for these two new observations, as created by the `lda` command:

```
> predict(arctic.lda, new=novos)$x
            LD1           LD2           LD3           LD4
1 -1.016129e-15 1.585299e-16 -3.785922e-16 -4.574619e-16
2 -1.489254e+00 5.195185e-01  4.977837e-02 -8.053507e-02
```

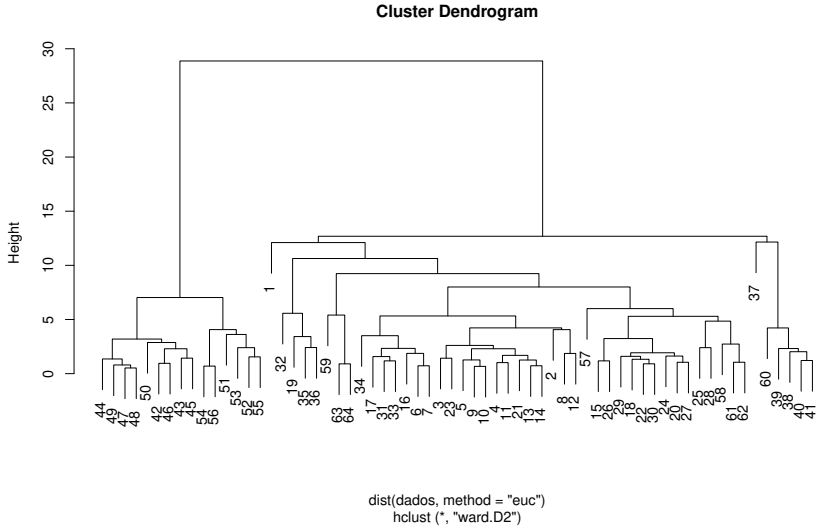
i. Identify, justifying your answer, which row corresponds to the observation created by taking all the mean values.

ii. To which group would you associate the other observation? What does that tell us about the individuals in the group that you chose?

3. A clustering analysis was performed on the set of 64 standardized observations using the hierarchical single-linkage method with the euclidean distance, yielding the dendrogram below. It is known that the cophenetic distances matrix associated with this dendrogram contains the values 7.42 and 7.92.



Later, it was performed a clustering analysis on the set of the 64 standardized observations using the hierarchical Ward's method with the euclidean distance, yielding the next dendrogram and a partition of the dataset into 5 groups.



The information on the pairwise distances between the groups is presented in the next table (rounded to 2 decimal places), where the designation of the groups is in accordance with the dendrogram order (from left to right):

	C_1	C_2	C_3	C_4
C_2	15.58			
C_3	28.88	12.10		
C_4	16.81	12.13	12.28	
C_5	14.57	13.04	12.48	12.15

(a) According to the available information, justify which of the following sentences you can assure that are correct:

i. The distance between observations 1 and 37 is equal to 7.92.

ii. The distance between observations 1 and 37 is greater than or equal to 7.92.

iii. The distance between observations 1 and 37 is inferior to 7.92

iv. The distance between observations 1 and 32 is greater than or equal to 15.34

(b) Determine the cophenetic distance between observations 1 and 37 defined by the clustering analysis with Ward's method. What is the meaning of the result from the point of view of the aggregation method that was used ?

(c) It is known that the RAND index between the classification into 5 groups given by the tags **2A**, **3A**, **4A**, **4B** and **4C**, and the classification into 5 groups obtained using Ward's method has the value of 0.5729167, and that 642 pairs of observations are assigned to distinct groups by the two classification procedures. Determine the number of pairs that are assigned in the same group by both classification procedures.

(d) It turned out later that one of the observations with a tag **4B** was misclassified, being reassigned with a tag **4D**. Determine the RAND index between the new classification into 6 groups given by the tags and the previous classification into 5 groups given by Ward's method.

(e) Posteriorly, a consolidation procedure of the partition into 5 groups of the set of 64 observations was performed, applying the k-means method with initial seeds given by the centers of gravity of these groups, and it turned out that we get the classes previously obtained with Ward's method. What do you conclude? This fact is enough to assure that the partition into 5 groups given by Ward's method minimizes the total intra-clusters inertia among all partitions into 5 groups of the set of 64 observations ? Justify.

II [4 points]

1. (a) Show that it is *not*, in general, true that the product of two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ is also symmetric. Give a necessary and sufficient condition for the product $\mathbf{AB}$ to be symmetric.

   (b) Consider a covariance matrix Principal Component Analysis. Assume there is also a group structure of the individuals, defined by the levels of some factor, with corresponding within-group variability matrix $\mathbf{W}$. Find a formula for the *discriminant* capacity of any given (centred) Principal Component, $\mathbf{X}^c\vec{\mathbf{a}}$, which depends only on the variance of that PC and on its within-group variability, $\vec{\mathbf{a}}^t\mathbf{W}\vec{\mathbf{a}}$. Using that formula, provide an upper bound for the within-group variability of that PC.

2. (a) Prove that if in Lance-Williams's formula,

$$d_{ij,k} = \alpha_i d_{i,k} + \alpha_j d_{j,k} + \beta d_{i,j} + \gamma |d_{i,k} - d_{j,k}|,$$

   the parameters $\alpha_i, \alpha_j, \gamma$ are nonnegative and verify $\alpha_i + \alpha_j + \beta \geq 1$, then $d_{ij,k} \geq d_{i,j}$ for every group $C_k \neq C_i, C_j, C_{ij} (= C_i \cup C_j)$.

   (b) Deduce from the previous question that Ward's method cannot have inversions.