**Sandra Marion Kam Tsemo**

**ASSIGNMENT 1:**

**Answer questions here and check the table from the attached pdf.**

The data analyzed are based on measurements made on a sample of 31 men during a sporting event. They represent age (in years), weight (WEIGHT in kg), oxygen consumed per kg per minute (OXYGEN), time taken to cover 2.4km (RUNTIME, in minutes), resting heart rate (RESTPULSE), heart rate during the race (RUNPULSE) and the maximum of this rate (MAXPULSE).

The age variable has been classified as follows: C1= between 38 and 44 years old, C2= between 45 and 50 years old, C3= between 51 and 57 years old.

To highlight the links between the different measures, a standardized Principal Component Analysis was performed on these data.

1) By analyzing the first table on page 3 explain why it was decided to carry out a second standard PCA on only 28 individuals.

The following questions concern the second standard PCA. The weight and age class variables are supplementary variables.

2) Justify why a standard PCA is used.

3) Give the inertia of the cloud of points. Justify the answer.

4) Specify the possible alternative to the standard PCA and its effects on the results of the analysis? Give the value of inertia in this case.

5) Recalling the usual criteria, how many components do you select?

6) Interpret (internal) the first two principal components and in particular the position of individuals 2, 4, 9, 21, and 28.

7) Explain how the additional variables weight and age class were added to the PCA results. Give the respective formulas.

8) Interpret (external) the results using supplementary variables. Check the consistency between the internal and external interpretation of the age and age-class variables.

9) A hierarchical clustering by Ward's method has been performed on the PCA results. The results in the appendices (pages 9 to 12) are identical to those of the direct classification on the original variables.

a. Justify this result.

b. Justify the choice of partitioning into 4 classes.

c. What do the level indices represent?

d. How much is the sum of the level indices? Justify the answer.

e. Recall what mixed classification is all about. The usefulness of each of its steps should be explained.

f. Calculate the inertia of cluster 3.

g. What is the value of the between inertia of the partition? Justify the answers (formulas can be given).

h. Interpret the clusters.

As for the age variable, it was decided to classify all the other variables: 7 classes for restpulse and maxpulse, 8 classes for runpulse, 2 classes for runtime, 4 classes for oxygen and 3 classes for weight. A multiple correspondence analysis was then carried out on all the qualitative variables. Answer the following questions:

10) Briefly recall the broad outlines of the method.

11) Regarding the number of categories of the variables, which preliminary treatments should be carried out?

12) How many non-trivial eigenvalues do we have? The first eigenvalue is denoted.

13) What are the rules for the choice of factorial components to retain?

14) The MCA could have been performed on the BURT table accordingly to give the value of the first eigenvalue that we would have obtained.

**ANSWER QUESTIONS**

1. Looking at the first table on page 3, it was decided to perform a second standard PCA on only 28 individuals because of two large contributions in the data. Indeed, there are 3 individuals (12 and 24 for axis 1) and (10 for axis 2) that have a high construct value on these axes.
2. A standard PCA was used because of the problems posed by the different units of measurement of the variables and the different variances.
3. The sum of the eigenvalues is the sum of the inertias in the orthogonal directions, and this sum is equal to the global inertia of the cloud of points. So,

Where is the number of variables.

Therefore, in this case,

1. The possible alternative to standard PCA is to perform PCA on the covariance matrix (rather than correlation matrix) (non standardized variables) . The effects on the results of the analysis are that variables with low variance are not taken into account, and mixing of units that does not make sense are also not taken into account.

Then, in this case,

1. We have two usual criteria for selecting the principal components :

(i) - Kaiser's rule: eigenvalues greater than 1

(ii) - Scree test: kink or large drop between two successive eigenvalues, starting with the second, i.e. always at least 2 components.

With these two rules we select 2 components and have 75.51% inertia.

1. Interpretation (internal) of the first two principal components and in particular the position of individuals 2, 4, 9, 21 and 28.

(i) **Interpretation of the first two principal components.**

We have the following information:

- Axis 1: 43.53% of information

- Axis 2: 31.98% of information

To interpret two principal components, we focus on the individuals and variables that contribute the most to the principal axis. Concretely,

Firstly, we observe that the variables RUNPULSE (0.82), MAXPULSE (0.79), RESTPULSE (0.68), OXYGEN (-0.66) and RUNTIME (0.59) have a significant correlation with axis 1, i.e. individuals who have a strong contribution on axis 1 will be characterised by a high value of these variables. Thus it is these variables and individuals that have contributed to the formation of axis 1.

Secondly, the variables RUNTIME (0.72), AGE (0.71), OXYGEN (-0.67) and MAXPULSE (-0.50) have a significant correlation to axis 2, i.e. individuals who have a strong contribution on axis 2 will be characterised by a high value of these variables. So it is these variables and individuals there that have contributed to the formation of axis 1.

In general, individuals in the 1st quadrant of the factorial plane will be characterized by RUNTIME, those in the 2nd quadrant by AGE, those in the 3rd quadrant by OXYGEN and those in the 4th quadrant by MAXPULSE and RUNPULSE.

(ii) **Interpretation particular the position of individuals 2, 4, 9, 21, and 28**

* Individual 2 is characterised by a high average value of MAXPULSE and RUNPULSE.
* Individual 4 is characterised by a high value of the mean of OXYGEN.
* Individual 9 is characterised by a value just above the mean of MAXPULSE, RUNPULSE and RUNTIME.
* Individual 21 will be characterised by a value just above the mean of RUNTIME.
* Individual 28 will be characterised by a value just above the mean of AGE.

1. Explanations

(i) **For variable weight** :

As the weight variable is quantitative and additional, its addition to the correlation circle is achieved through the correlations with the axes, which leads to the following formulas:

Where is standard deviation of height, is the standard deviation of component1 which is equal to the square root of the first eigenvalue which is the variance of component1.

(ii) **For age class** :

As the variable age class is qualitative, its addition to the factorial design will be the projection of each modality to the barycentre of the individuals taking the modality. So, we just compute the gravity center of each category.

1. Interpretation ( external) of results using additional variables. Let us check the consistency between the internal and external interpretation of the age and age class variables.

External interpretation consist of using supplementary variables :

Correlation for the quantitative : the correlation must be high to be interpretable, here it isn’t the case. V-test for the categories of the qualitative variable : the v-test must be less than -2 or greater than 2 to be interpretable, here non interpretable on axis 1, opposition on axis 2 between C1( less than -2 ) and C3 ( greater than 2), consistency with the interpretation of the second axis it can be said that is opposition between young people with low RUNTIME high OXYGEN and oldest people with high RUNTIME low OXYGEN.

1. A hierarchical clustering by Ward's method has been performed on the PCA results. The results in the appendices (pages 9 to 12) are identical to those of the direct classification on the original variables.

a. **Justification of this result**

This result is justified by the fact that by using all the components, we have all the inertia, the ith clustering uses the same information. The distance between individuals calculated using all components is equal to the distance between individuals calculated using all initial variables.

b. **Justification of the choice of partitioning into 4 classes**

There is a big drop between the indices of cluster 52 and 53, so we will have 4 clusters.

However, it is also possible to have 3 clusters because there is also a drop between cluster 53 and 54. Moreover, it is also possible to have 2 clusters because there is a drop between cluster 54 and 55 but this case is not interesting because 2 clusters is generally not enough, we will have too big clusters.

c. **We will explain what the level indices represent**

The level index is the Ward index which represents the (minimal) loss of inertia between objects resulting from the aggregation of the two clusters: when two clusters are aggregated, we gain internal inertia and lose inertia between objects, Ward's strategy consists in aggregating the two objects corresponding to the lowest loss.

d. **Amount of the sum of the level indices**

Summing up all the lost inertia, we obtain the total inertia, which means that the sum is equal to 6

e. **Recall what mixed classification is all about. The usefulness of each of its steps should be explained.**

Mixed clustering consists of 3 steps

(i) perform a K-means method with K = k’ around few hundreds

(ii) perform hierarchical ascendant clustering (HAC) on the results obtained from step 1, we’ll get k clusters

(iii) perform a K-means method with K = k

**step 1 :** is useful in case of large number of objects, hierarchical ascendant clustering is slow and the first aggregations are not useful, interesting so using K-means allows to accelerate the process, the main drawback of K-means which is the number of cluster to fix a priori has no effect here because k’ will not be the final number of cluster

**step 2** : takes benefit of HAC (no need to fix a priori the number of cluster) without its drawback (slow) since it is performed on a reduced number of objects

**step 3** : to consolidate the partition from HAC, to improve the between inertia.

f. Calculate the inertia of cluster 3

Inertia of a cluster is the mean of the distances between the individuals and the center So,

g. What is the value of the between inertia of the partition? Justify the answers (formulas can be given).

The partition between inertia is the weighted mean of the distances between the cluster center and the gravity center. The formula of between inertia is :

As we are in a normalized PCA, this means that the center of gravity of the scatterplot is 0

Then

h. Interpret the clusters

Use the v-test >2 or <-2

Cluster 1 is characterized by runtime mean value (for the 5 individuals of cluster 1) greater than the global runtime mean value

(for all the 28 individuals) and oxygen mean value lower than the global oxygen mean value

For cluster 3 it is the opposite : runtime mean value lower than the global runtime mean value and oxygen mean value greater than

the global oxygen mean value

Cluster 2 is characterized by maxpluse and runpulse mean values greater than their global mean values

For cluster 4 it is the opposite : maxpluse and runpulse mean values lower than their global mean values



The multiple correspondence analysis is equivalent to a simple correspondence analysis of the

of the disjunctive table.

Step 1: The contingency table is transformed into a disjunctive table;

Step 2: The margins are calculated;

Step 3: Find the BURT table, with K disjunctive table;

Step 4: Apply PCA to the BURT table.

11)

If we notice that there is a categorical variable that has more categories than the others, we will aggregate its categories so that they all have the same number of categories.

12)

If we remove k categories in variable j which has many categories than the others, we will have k non-trivial eigenvalues.

13)

We have two rules:

The first, if the variables are mutually independent, then any eigenvalue λ = p 1 .

The second is the Scree-test, as all eigenvalues are equal to p 1 , so an interesting eigenvalue will be the one that is strictly greater than p 1 with p the number of qualitative variables.

14)

Any eigenvalue λ of the disjunctive array, will be a Burt array eigenvalue ( ) with

K complete disjunctive array.