Transcript of audio of PCA video

In this video, we will see how to perform PCA with Factoshiny, the graphical interface of FactoMineR. To do this, we will use a dataset called decathlon. The dataset has the following form: there are 41 rows corresponding to 41 athletes, and 13 variables. Among these 13 variables, we have 10 variables that relate to performance in the various decathlon events. For example, 100 meters, long jump, shot put, high jump, 400 meters, etc. Then we have two more quantitative variables which are the athlete's rank in the competition, and the number of points scored by the athlete. In addition, there is also a competition variable with 2 categories: the decastar category and the Olympic Games category. These were 2 major competitions that took place in 2004.

We will perform PCA on this dataset considering the 10 trials as the active variables. In turn, the variables that have to do with the rank, the number of points, and the type of competition, will be considered as additional or supplementary variables, also refer to as illustrative. In other words, we will compare the athletes only from the point of view of their performances in the 10 events.

Now let's see how to launch PCA with FactoMineR, and more specifically using Factoshiny as its graphical interface. This interface launches the commands of FactoMineR and it is not necessary to know the syntax of R. This interface also improves the readability of the graphics. Let's load the Factoshiny package. To start PCA, like any other data analysis method available in the interface, simply use the Factoshiny function on the dataset. This function can be run on a data set, on an object of class PCA, or also on an object from the Factoshiny function. Let's run the function on the decathlon dataset.

The graphical user interface opens in the default browser. The window is divided into 2 parts. On the left side, there is a brief description of the dataset, then the methods that can be applied to that dataset, and a link to a video that helps choosing which method to use. On the right hand side we have the different methods. Clicking on a method's help button provides a quick description of the method as well as links to course videos about the method. If you then click on "Run", the analysis is executed and a new window opens in the browser. This new window is divided into 2 parts. On the left is the menu that will allow you to parameterize the method or the graphs, on the right are the results. In the left menu we have several tabs. The first one will be used to parameterize the method, i.e. to choose the variables that will be active and the variables that will be illustrative, the active or additional individuals and also the management of missing data if missing data are present in the dataset.

We then have a tab that allows us to enhance the graphs, a tab that allows us to perform a clustering at the end of the PCA. A tab for an automatic report to interpret the main results of the PCA. And finally 2 buttons. A button to retrieve the lines of code from the analysis so that you can re-implement the PCA and rebuild the graphics as they are in the interface and a button to exit the application.

First of all, let's set up the method. We will click on the "PCA Settings" tab. So I'm going to choose the quantitative variables that will be illustrative or supplementary. The other quantitative variables will be active. In the example, the variables ranking and number of points are illustrative quantitative variables. The competition variable is an illustrative qualitative variable. By default, all qualitative variables will be used as additional variables. If you don't want certain variables to be used as extras, you have to delete them. And I don't have any extra people, so I'm going to leave that box blank. By default, the variables are standardized. If we didn't want to standardize the variables, we would have to uncheck this box. And finally, there is the possibility to manage missing data.

In this example, there is no missing data. If the dataset contained missing data, here we would have the possibility to choose different imputation methods. Here's what we would get. The possibility of imputing by the mean of the variable. This method is very fast but not recommended. The possibility of imputing through a 2-dimensional PCA model, which is rather a good compromise in most situations. And finally the possibility of imputing by a k-dimensional PCA model. The number k is initially estimated by cross-validation, which can be a bit long on large datasets. But the number of dimensions is then optimal for the PCA model that will be used to impute the missing data in the dataset. Once the dataset is imputed, we end up with a complete dataset on which we will be able to implement PCA.

Let's go back to our dataset that has no missing data. We have finished setting up the method. We are going to submit to be able to take into account all the modifications we have made, either by clicking on this button or by clicking on this box to exit the settings tab. Thus, the new settings of the method are taken into account and the PCA results are updated with the two additional variables *ranking* and *number of points* and the *competition* variable as an additional qualitative variable.

On the right we have the main results grouped in five tabs: a tab on the graphs, a tab with the main quantitative results, a tab on the automatic description of the dimensions and then a summary of the dataset and the data table.

Let us first look at the summaries of the main results. We have here a listing with the main outputs. The first one shows the results on the percentages of inertia associated with each dimension. We see that the first dimension summarizes 33% of the information while the second dimension summarizes about 17% of the information. We then have the results on the individuals, by default for the first 10 individuals. If we want the results for all individuals, just put the number of individuals here or a greater number than the number of individuals in the dataset. If I put 100, I will have the results on all the individuals, because there are 41 individuals. And so in the results on the individuals, we have the name of the individuals then the distance of the individual to the center of gravity of the cloud then the results on the first dimension with the coordinate of the individual on the 1st dimension, his contribution to the construction of the dimension and its quality of representation on this dimension measured by the cosine squared, which is worth for example for this individual 0.695. So we have this result on the first dimension and we have the same results on the second dimension and then the third dimension. So if we wanted to see the quality of representation of this individual for the plan, it would be enough to sum 0.695 and 0.080. So here are the results for the individuals.

We then have the results for the variables; with the name of the variables, then the coordinates on the first dimension, the contribution of the variable to the construction of the dimension and the quality of representation measured by the cosine squared. So much for the first dimension and then the second dimension and then the third dimension. If additional quantitative variables are present in the analysis, we have a table with the coordinate of each variable on the first dimension, its quality of representation, of course we do not have its contribution since the additional variables did not contribute to the construction of the dimensions. For the additional qualitative variables, we have the results for each category of all the qualitative variables, with the distance to the barycentre, the coordinate on the first dimension, the quality of representation and a v-test that will measure if the coordinate is significantly different from 0. The v-test follows a Normal law and therefore the extreme values will tell us which are the values that are significantly different from 0 on a dimension. And so here we have the results again on dimensions one, two and three.

We also have more details of each of these results. For example, in the table of eigenvalues, we can look at the percentages of inertia and a graph with the percentages of inertia associated with each dimension. And then the same thing, the results on the variables, on the individuals, the additional

variables and the additional qualitative variables. The same results as those summarized in the first tab can be found.

Next, we have an automatic description of the dimensions, more precisely of the first three dimensions, according to quantitative or qualitative variables. For the quantitative variables, the correlation coefficient between the coordinates of the individuals on the dimension and the variable was calculated. For example, the variable number of points is positively correlated with the first dimension, the correlation is 0.96 which means that there is a very high correlation between the coordinate and the number of points. So individuals with a low coordinate on the first dimension scored a low number of points. And this correlation is significantly different from 0 since we have a p-value less than 5% here. Only correlations that are significantly different from 0 are retained, so the variables are sorted from the most significantly related to the most negatively related at the bottom. I can change here the threshold for keeping variables, and keep variables that have a p-value less than 20% rather than 5%. We'll keep a few more variables. I'm just doing this to show that you can see the links with quantitative variables but also with qualitative variables. For example, here, for the 1st dimension: we see that the competition variable is linked to the 1st dimension and this is significant at the 20% threshold. So the R-square is the correlation ratio. 5% of the variability of the coordinates is explained by the competition variable, which is significant at the 20% threshold (but not at the classic 5% threshold).

Now let's go back to the graphics tab. There are two default graphs: one for the individuals and another for the variables. The graph for the individuals depicts observations in black and additional categories in pink. The graph for the variables displays active variables in black and illustrative or additional variables in blue with a dashed pattern. Often, it’s very useful to work on your graphs to better highlight the information. So we can choose different graphic options here. First we can modify the choice of the dimensions we are going to draw. By default, dimensions 1 and 2 are drawn, but we can, for example, decide to draw dimensions 3 and 4. Simply edit here and choose 3 and 4 to draw the 3-4 plot for the individuals and variables graph. Both graphs are modified simultaneously since they are commented on at the same time. Let's put back dimensions 1 and 2 and then we can work on the graph of individuals or the graph of variables.

Let's start with the graph of individuals by changing the title. One can also choose the points to be drawn and remove the additional categories to draw only active individuals. This is not very useful in our example, but when you have a lot of qualitative variables, it prevents the graph from being crowded by too many categories. If you have a lot of individuals, you can also delete the labels of the individuals and keep only the labels of the additional terms. You can increase the font size or choose to put the labels only for certain individuals. For example, keep the labels only for those individuals who are well represented at the principal level, those with a quality of representation greater than 0.5 at the principal level will have a label. We can also select individuals according to their contribution to the construction of the plan: by choosing the ten individuals who have contributed the most to the construction of the main plan, i.e. the first two dimensions. We can also colour individuals according to the quality of representation (of the cos2). Or colour individuals according to a quantitative variable such as, for example, the number of points variable. If I plot all the points, I will see that the variable number of points is positively correlated to the first dimension. We can see it very well with the color code: on the left we find individuals in blue that take low values on the number of points, in the middle we find individuals in purple that take average values and on the right individuals in red that take high values on the number of points. It is the correlation between the coordinates and the variable number of points that we illustrate here with the colors.

We can also colour individuals according to a qualitative variable, more precisely according to the categories they take for a qualitative variable. In our example, there is only one qualitative variable, so

we will depict individuals according to the type of competition. We see that the individuals coloured in black participated in the decastar and the individuals coloured in red participated in the Olympic Games.

We can also build confidence ellipses around the barycentres of the decastar and Olympic Games. These ellipses should be seen as confidence regions of a barycentre. In other words, if we had had other athletes participating in the decastar, the average would have ended up, with a confidence level of 95%, in this ellipse. It can be seen that the decastar and Olympic Games confidence regions overlap and therefore there is no significant difference in the position of these categories in the main plan.

We can now look at the graph of variables. Of course, change the title, increase the size of the labels, and select the variables according to their quality of representation, for example to keep only those with a representation quality higher than 0.6.

You can increase the size of the graphics here, and of course you can download the graphics in jpeg, png or pdf format.

At the end of the PCA, it is possible to perform a cluster analysis. Just check this box here and choose the number of dimensions you want to keep to get the clusters. If only the first dimensions of the PCA are retained, it is equivalent to retaining the dimensions that contain the signal, the information, and removing the last dimensions that contain noise. That way, we'll have a more stable clustering configuration. The main idea is to keep the first dimensions, i.e. those that will allow 70 or 80% of the information to be retrieved. However, it is also possible to keep all the dimensions of the PCA, which amounts to perform clustering on the initial data, or more precisely the standardized data. I'm not going to do the clustering here because this is explained in another video.

It is also possible to obtain a report on the results of the PCA, i.e. a simple automatic interpretation of the results of the analysis in English or French. This automatic report can use graphs suggested by the analysis or use the graphs we just worked on. First, it is interesting to see the graphs suggested by the method. We will be able to retrieve this automatic report in different formats: in Rmarkdown format, in html format or in word format. I'll get the report in html format. It takes a little time to write the report. Here is the report that begins by specifying on which dataset the analysis was carried out: here a dataset with 13 variables, 2 of which are considered additional quantitative and one additional qualitative variable. First, there is a search for extreme individuals. Extreme individuals are the individuals who would contribute enormously to the construction of the dimensions and without whom the dimensions would be very different. There are no extreme individuals here. Then the percentages of inertia associated with each dimension are commented. The graph gives the decomposition of inertia with the percentage of inertia on each dimension and comments on these percentages of inertia compared to what would have been obtained for data sets of the same size (same number of individuals and same number of active variables) if the variables were not structured. This allows you to see how well the first dimensions summarize the information. The report also suggests the number of dimensions that should be interpreted. Then we have a description of the individuals' plot, with some individuals having a label and others not, to avoid having cluttered graphs. A graph with the individuals colored according to a qualitative variable. Here we have a single qualitative variable, but if several variables were available, the Wilks test would allow us to choose the variable for which the categories are most separated in terms of design. Thus, individuals would be coloured according to this variable. Then we have the graph of the variables again with some labels only for some variables, and then a graph with the additional categories. Finally, there is an interpretation, or an attempt at interpretation, of each of the dimensions through the graph of individuals and variables. This interpretation is very limited and only allows us to see the connections and characteristics of individuals. It is then up to each one to try to go beyond this interpretation in

the explanation of the dimensions. In this case 3 dimensions are proposed, there is also a description of the third dimension. The automatic report has suggested some graphs, it is also possible to use the graphs just made with the interface.

Finally, there is a “get the PCA code" button that retrieves lines of code from the analysis to implement the method and rebuild the graphs identically. By clicking on “get the PCA code”, 3 lines of code appear: one to parameterize the method, one to build the graph of variables and one to build the graph of individuals here with ellipses and the plotellipses function.

Finally, you can quit the application by clicking on the "quit the App" button. My result object contains the lines of code to parameterize the method and build the graph of individuals and the graph of variables. I can also find the application exactly as I left it before. I just have to do Factoshiny(result). You can see that all the settings are there. So I can change my graphs again. And I can quit again and close.

You've seen the main features of the Factoshiny feature, feel free to test the available features. Now it's up to you to implement PCAs with FactoMineR and Factoshiny.