Transcript of audio of MCA video

We're now going to see how to do MCA using FactoMineR. The dataset we're going to use, deals with tea consumption. A questionnaire was answered by three hundred people, and there were three types of variable in it. Eighteen variables dealt with "how" tea is consumed; for instance: do you take tea with breakfast? Yes or no? At afternoon tea time? Yes or no? In what form? In a teabag? As loose tea? So, that's the first eighteen variables. Then there were the questions about the image people had about tea. And lastly, questions about the people themselves.

These are the questions in green here. For example, sex, socio-professional category, whether they do sport or not. Age, coded as a qualitative variable Age_Q, in age brackets. As for their image of tea, there were questions like: to you, is tea good for health? Is it a diuretic? And so on. Then, one last quantitative variable: actual age.

Now, let's get into the practical details of running a MCA using FactoMineR and Factoshiny. We load the package Factoshiny and the dataset tea that is available in the package FactoMineR.

Now let's see how to run MCA on the tea dataset using Factoshiny. We write Factoshiny(tea) which opens the graphical user interface in the browser. 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. You can click on the chosen method, i.e. Multiple Correspondence Analysis. A new window opens. 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 side 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.

First of all, let's set up the method. We will click on the "MCA Settings" tab. The quantitative variable "age" is considered as additional since it is quantitative; if I remove it here it will disappear from the analysis. I will specify that the variables concerning the image of tea and the questions about people will be supplementary (I have to select them one by one). Here there are no missing data in the dataset. If there were, missing data could be managed according to several options: by adding an NA category for each variable containing missing values (this is the default option); by imputing the missing values of the indicator matrix by the proportion of the category calculated on the observed values. This method is very fast but is not recommended. It is possible to impute the indicator matrix using a 2-dimensional MCA model, which is a rather good compromise in most situations. Finally, it is possible to impute by a k-dimensional MCA model. The number k is first estimated by cross-validation, which can be time consuming on large datasets. But the number of dimensions is then optimal for the MCA model that will be used to impute the indicator matrix. Once the indicator matrix has been imputed, the MCA is then done from this new indicator matrix. Once the settings are complete, I submit or exit this settings tab. This submits the method with the new settings.

Some plots are output. One shows the individuals, and all the categories, with the active ones in red, and the supplementary ones in green. The plot is really quite overloaded for the moment; we'll see later how to make this better. Another plot shows the active and supplementary qualitative variables, giving the squared of the correlation ratio between each of them and the axis. We also have the age variable; here, we have the R-squared of the regression of the age variable in terms of each axis. Next

up, we also have a plot showing the supplementary quantitative variables - here, there's only one, and the correlation circle.

Before we see how to make better the graphs, let's see the results of the MCA in the second tab. The output of this function first reminds us of the line of code we used to run it, followed by a table with the eigenvalues and percentages of inertia associated with each axis, and then the results on the individuals. By default the results are given for the first 10 elements. If we set a higher value we will see more results.

So, for the individuals, we have: their coordinate values on the first axis, their contributions to the construction of the axes, which are quite small, but don't forget there are 300 individuals. Followed by the quality of representation, measured by the squared cosine. We have these results for the first, second, and third dimensions. Next, we have the results for the categories. This means: all categories, their coordinate values, their contributions, their quality of representation as defined by squared cosine, and a test statistic.

The test statistics here follow a Gaussian distribution, so those below minus 2 or above 2 can be considered significant. This means that the category in question has a coordinate significantly different to zero. This is useful for knowing which categories have large positive or negative values in each dimension. For example, here, breakfast has a large positive value in the first dimension. We have the results for the first, second, and third dimensions. We also get the results on the variables; that is, the correlation ratio of each variable for each of the first three dimensions.

This is simply the correlation ratio between the axis and the variable. These are the coordinate values that help us to build the plot of the squared relationships. Next, if we have any supplementary categories, we obtain results on them, including coordinate values, representation quality, and a v-test. However, no contributions, because they did not contribute to axis construction.

Also, we get a table showing the correlation ratios for the supplementary qualitative variables. Next up, we have results for the supplementary quantitative variables. Here, there's only one of them, so we have its coordinate value in each dimension, that is, the correlation coefficient between each axis and the quantitative variable.

We can also give a description of the axes, which can be quite useful when we have a lot of variables. We can see for example that the first dimension isn't characterized by quantitative variables; none of them. Instead, it is significantly linked to various qualitative variables. The strongest link is with "where do you buy your tea", then "do you drink it in a tea room", etc. The variables are ranked from most to least related, and only the ones with a correlation ratio significantly different to zero are shown in the results.

Next, we have tables for the categories, showing each category and its coordinate values on the axis, along with the test: "is this coordinate value significantly different to zero, or not?" We can see that we have both the active and supplementary variables used to describe the axes, as well as the active and supplementary categories. So, we have the results in the first dimension, then the second dimension, where we can see for example that the "age" variable is significantly correlated with it. And we also have the results in the third dimension.

We went a bit fast on these default plots, because they are quite useful when we don't have many individuals or categories, but if we do have many of either, we have some work to do to make them more easily interpretable. First, we can put labels only on active and supplementary categories, and not on the individuals.

This gives us the biplot with the active individuals -- there are no supplementary individuals but if there were, we could show them -- the active categories, the supplementary categories, and as we said, labels only on the active and supplementary categories. Another thing we can do to the plot is to make certain things invisible. For example, we can make the active and supplementary categories invisible, like this. This gives a plot that only shows the individuals. As another example, here we make the individuals and the supplementary categories invisible. This plot therefore shows only the active categories.

If I want to retrieve the lines of code corresponding to a graph, I will click on the button line of code of the MCA and retrieve the line of code.

We can then, for instance, indicate that we want the font size to be smaller than the default. We can add a title: "Active categories".

And lastly, we can make the active categories invisible, as well as the individuals, leaving us with a plot with only the supplementary categories on it. We can then select certain categories to highlight. Here, we select the ones we want, only putting labels on the best-represented categories. Here is that plot, showing the active categories with a quality of representation, a cosine squared, higher than 0.5. Instead of using the quality of representation, we could instead use contributions.

If we want to do a similar kind of thing but for the individuals, for instance, and put names only on the twenty which have the highest-quality projection onto the plane of the first two axes, we first make the categories invisible, and then put label for the twenty individuals with the highest contribution.

We can also colour the individuals according to their quality of representation or their contribution. For example, according to their contribution, we see that the most extreme individuals contribute the most. You can also draw individuals according to a variable. For example, according to the variable where. We have a different color for each of the 3 categories of this variable. We find that this variable is linked to the first 2 dimensions since the sub-clouds are well separated.

We can also build confidence ellipses around each of the 3 categories. We can see that the confidence ellipses are very small. Indeed the number of individuals is important here and the sub-populations are quite separated. The ellipses do not overlap, which indicates that the sub-populations are significantly separated.

Finally, we can colour all categories of the same variable with different colours. This makes it easier to see the categories of the same variable. This graph is very useful when there are few variables and the variables have a large number of categories. The colors used on the category graph are simultaneously used on the variable graph. This graph of variables gives the squares of the links and indicates which variables are globally linked to the different dimensions. You can also improve this graph by clicking on Variables. This plot is overloaded with information. We can, for instance, make the qualitative and quantitative supplementary variables invisible. All that remains are the active variables.

The last plot is the one that shows the supplementary quantitative variables - here, there's only one, and the correlation circle.

Of course, we can see the third and fourth axes, not just the first and second.

At the end of the MCA, 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. Thus MCA is used as a preprocessing to transform qualitative variables in quantitative variables. Moreover, If only the first dimensions of the MCA are retained, it is equivalent to retaining the dimensions that contain the signal 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. 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 MCA, 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.

Finally, as I said before, there is a “get the MCA code" button that retrieves lines of code from the analysis to implement the method and rebuild the graphs identically. By clicking on “get the MCA code”, several lines of code appear: one to parameterize the method, the others to build the plots.

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.

So, there you have it. Now it's your turn to use MCA with Factoshiny. Good luck!