Transcript of audio of MFA video

The aim of this video is to show you how to do multiple factor analysis, that is, MFA, with FactoMineR and Factoshiny, and how to improve the default output graphs.

We're going to use the following dataset in which twenty one wines were described by a panel of experts using sensory descriptors. Two qualitative variables, the origin of the wines and the soil type, were also recorded. We have therefore 21 rows of wines, and 31 columns of variables: two qualitative and 29 quantitative, which can be grouped into 5 groups of variables: one group of smells, one group of visual perceptions, one group on taste variables, one group on smell after swirling, and a group on tasting appreciation. The groups for smell, visuals, tasting, and smell after swirling are considered active groups, while the groups for the wine's origin, and for tasting appreciation, are considered as supplementary groups.

So, now we can go and use R or Rstudio. So, let's run the first command and load the Factoshiny library. To begin, we load the "wine" dataset which is available in the package FactoMineR. And then we can run the Factoshiny function on the wine dataset using Factoshiny (wine).

A window opens in the default browser. Here, we choose the multiple factor analysis method. The first thing to do is to define each group of variables. There are two ways to define the groups. A first way is to construct groups one by one, variable by variable. This is useful when there are few variables and when the variables of the dataset are not sorted by group. However, when there is a lot of variables or a lot of groups, this way of doing things takes too much time. It is thus preferable to define the groups by giving the number of variables belonging to each group, and specifying the nature of each group.

Let's see how to define the groups from the interface on a simple example. I will build a first group of variables with the qualitative variables, so I select qualitative and then choose the variables, and I can give a name to this group. I can specify that this group is illustrative. I can then construct a group of quantitative variables with 3 quantitative variables. Then a 3rd group of quantitative variables that will contain 4 variables. If I don't want to add any more groups, then I have to validate the groups.

Let's go back to our full example which contains more variables. We will now use the second strategy to define the groups. First, we specify the number of variables per group, separated by spaces or commas. In the wine's example, the first group of variables consists of the first 2 variables, the second group contains the following 5, then the 3rd the next 3, the 4th the next 10, the 5th the next 9, and the last group contains the last 2 variables. The nature of the variables in a group is specified by the "type" option. We specify the type of each group of variables with "c" for continuous or quantitative, "s" for continuous or quantitative but with "scaled" variables, "n" for groups of nominal (or qualitative) variables, and "f" for frequency tables. Thus, in the example, the first group is composed of qualitative variables, while all subsequent groups contain quantitative variables with an "s", which here means that we will standardize the variables in each group.

It is possible to name the groups of variables: a group for the origin of the wines (including "label" and "soil type" variables), a group of smell description variables, a group of visual variables, a group for smell after swirling, a group on tasting, and a group of overall preference. If we do not label the groups, we will have the names group1, group2, etc.

Then we specify that the groups of variables number 1 and 6, that is, the group for the "origin" and the group of the overall preference, are supplementary groups which do not participate in the construction of the dimensions. Then, we validate the definition of the groups.

I can then go to setting the parameters and specify if some individuals are supplementary, or specify how to manage missing data if missing data are present in the dataset. We can impute by the mean of the variable for quantitative variables, or by the proportion for qualitative variables, or by using a 2-dimensional MFA model.

You can start by opening the Values tab which contains the numerical results. First of all, we see the line of code that was run; Then, we have a table with the eigenvalues and the percentages of inertia associated with each dimension. The first dimension recovers 49% of the information, that is, 49% of the inertia, and the second dimension, 19%. Next, we have results on the groups of variables, so, the active groups first with the coordinates of the groups, the contributions of each of the groups to the construction of the first dimension, and their representation quality on the first dimension. Then, the same results on the second dimension (coordinates, contributions, and cosine squared), then the 3rd dimension.

Next, we have the results on the supplementary groups, giving the coordinates and the cosine squared. There are no contributions here, since these are groups that did not contribute to the construction of the dimensions. The following table provides the results on the individuals, by default the first ten individuals. If we want the results on all the individuals, we modify the value by setting a high value. We will get the results for everything: all the individuals, all the variables, and so on. Here again we get the coordinates, the contributions, and the cosine squared, first on the first dimension, then on the second, then the third.

Next up, we have the results for the active quantitative variables, again with coordinates, contributions, and cosine squared. For the supplementary quantitative variables, we have just the coordinates and the cosine squared; here again, these are variables that did not contribute to the construction of the dimensions.

There are no qualitative variables that are active here, so we have results only for the supplementary qualitative variables, and more specifically for the categories of the supplementary qualitative variables. So we have the coordinates, the cos squared, no contributions of course, and a test-value denoted v.test. The test value indicates whether a category's coordinate value is significantly different to 0, or not. More precisely, the test value corresponds to the transformation of a p-value into a quantile of the normal distribution. If the p-value is less than 5%, then the absolute value of the test value will be greater than 1.96. The sign of the test value indicates whether the coordinate value of the category is lower (negative sign) or greater (positive sign) than 0. For example, the Env4 category has a coordinate value significantly different to 0 on the second dimension; and greater than 0 on this second dimension.

So, that's it for the main outputs obtained from an MFA.

Now let’s have a look to the plots. Several graphs appear by default. The number of graphs and the ones that appear depend on the presence or absence of groups of quantitative variables and groups of qualitative variables. Let's now look at, one by one, the graphs obtained in our example.

We have a plot showing the mean individuals and the categories, that is, the individuals as seen by all the groups of active variables. For example, wines T1 and T2 have very similar sensory profiles when all sensory points of view (visual, smell, taste, etc.) are taken into account. Each category is at the barycentre of the points that take it into account. We can make the categories invisible for example.

We can also represent the partial points. Partial points are given the colors used in the groups plot. For the 1VAU wine, the red point represents how it is seen in terms of smell variables only, and the green point shows how it is seen in relation to the visual variables only.

It's possible not to draw all the points, and to just select certain individuals. We can for example select only the wines that are well projected on the plane, those which have a $\cos^2$ greater than 0.4. In black, with a label, we have individuals with a cos squared greater than 0.4, and gray for those with a $\cos^2$less than 0.4. We can also color individuals according to a variable, as for instance variable “Soil”. The Saumur wines are in red, Bourgueils in green, and Chinons in blue. We have still highlighted those with $\cos^2$ greater than 0.4. We can also make selections according to contributions, and show only the 5 individuals who have contributed the most to the construction of the dimensions. So here, only the individuals who have contributed most to the construction of the plane have colored points. We can play around with transparent points using the unselect option.

If I want to retrieve the lines of code corresponding to a graph, I will click on “get the MFA code” button and retrieve the lines of code.

It’s possible to draw a plot with the categories. In our analysis, there are no active categories, so only the supplementary ones are shown. For each category, the mean point, and its partial points, are shown. Mean and partial points can be interpreted here like for the individuals plot. We can say for example that the wines grown on the reference soil type, and those on the Env4 soil type, are visually perceived in the same way, but are very different in terms of smell and taste.

As, in our analysis, some groups are made up of quantitative variables, we also get a variables plot with the correlation circle. In this plot, the variables are colored according to the group they are in. So, one color for each group. This variables plot can be interpreted simultaneously with the individuals plot, like in PCA. Thus, for example, the wines to the right of the individuals plot are intense wines in terms of aroma and taste. And the wines at the top of the plot are spicy.

As with the individuals, variables can be selected according to their contribution, for example by keeping the five variables that contributed most to the construction of the dimensions. The five variables that contributed the most will get a label, and the other variables will be shown but partially transparent. We can see which group they belong to, and where they're projected. This makes it possible to have plots with labels that overlap less. Which means more readable plots when there are many variables. Often, it's interesting to highlight only a few of them, because the variables close to the center of the circle, that is, those with very short arrows, are not of great interest to us, because they're poorly projected. We'll often focus on the best projected variables, which therefore have large-valued coordinates or have strongly contributed to the construction of the dimensions.

We then get the plot showing the groups of variables, with solid triangles for the active groups, and empty triangles for the supplementary groups. If a group has a large coordinate value on a certain dimension, then that dimension is also present in the group of variables. In other words, this group divides the individuals like the MFA does with this dimension. If two groups have large and close-together coordinates on several dimensions, then they induce similar structures on the individuals.

Lastly, we get the graph showing the partial axes. For each group of variables, one analysis was carried out: for the quantitative variables, a PCA, for the qualitative variables an MCA, and the dimensions of these PCA and MCA were projected as supplementary information onto the dimensions of the MFA. So, for example, for the visual group, the first dimension is closely related to the first dimension of the MFA, while the second dimension is somewhat less related to the second dimension of the MFA. For

the "origin" group of qualitative variables, the dimensions of the MCA are projected. Here, the first 2 dimensions of each group are represented, but we can choose to represent 3 of them.

By default, the 1-2 planes are provided, but the 3-4 planes can be also constructed. Here are the plots with dimensions 3 and 4. So, that's it for the main graphical outputs.

There is also a tab that provides an automatic description of the factor dimensions exactly as it does with PCA and MCA results. The first dimension is therefore characterized by the most closely related variables with the first dimension of the MFA, and then by the qualitative variables and the most related categories.

At the end of the MFA, 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 MFA are retained, it is equivalent to retaining the dimensions that contain the signal, the information, and removing the last dimensions that contain noise. The main idea is to keep the first dimensions, i.e. those that will allow 70 or 80% of the information to be retrieved. A clustering done on MFA results will take into account the weightings of MFA and thus the balance between groups. I'm not going to do the clustering here because this is explained in another video.

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

You've now seen the main plots and main indices for MFA. Now it's up to you to implement MFAs with FactoMineR and Factoshiny.