Audio transcription of the data analysis methods summary video

In this final video, we are going to make a summary of all the methods we've seen during the course. Each method is considered in terms of the questions it can be used to answer, when doing data analysis.

To start, are there several groups of variables? Thus, does the data set contains different information sources, each source containing several variables. If yes, we consider methods that take into account these groups and we can perform Multiple Factor Analysis.

The second question is: what kind of data and information have we been given? Two possibilities: either it's a contingency table or several contingency tables, in which case we are going to think about doing correspondence analysis if there is only one table and multiple factor analysis on contingency tables if we want to compare several contingency tables. Or is it a table with individuals lined up against variables? Individuals are described by several variables and thus according to the type of the variables there are different methods. We can think about doing PCA if variables are quantitative, MCA if variables are qualitative, FAMD if variables are of both type or MFA if variables are structured in groups. The functions in FactoMineR, written in orange, are CA for Correspondence Analysis, PCA for Principal Component Analysis, MCA for Multiple Correspondence Analysis, FAMD for Multiple Factor Analysis on Mixed Data, and MFA for Multiple Factor Analysis. We can describe briefly the method named Factor Analysis on Mixed Data. This method deals with both quantitative and qualitative variables and consider all these variables as active, thus all of them participate in the construction of the axes. FAMD balances the influence of each variable whatever their type. The graphs and indicators are the same as in PCA and MCA: we draw a graph with the individuals and the categories as in MCA, the correlation circle as in PCA for the continuous variables.

Another question is: which elements are active? By this, we mean: which elements should we use to construct the factor axes? Which elements should we use to calculate distances between individuals or between rows? By making this decision, we end up with a table with active elements in blue, supplementary columns in green, and supplementary rows in pink. Thus, depending on the situation, we are going to have some or all of the following terms in the code: ind.sup to indicate supplementary individuals, quanti.sup to indicate supplementary quantitative variables, quali.sup for supplementary qualitative variables, row.sup and col.sup if we are doing CA in the presence of supplementary rows and/or columns, and lastly group.sup for supplementary groups of variables in MFA.

Once the active elements have been defined, we need to know whether variables are quantitative or qualitative. If the variables are quantitative, we should do a PCA using the PCA function in FactoMineR. If the variables are qualitative, two things: if there are only two variables, we suggest constructing a table for them with the categories of one variable as rows, and the categories of the other as columns, then filling in the contingency table and analyzing it using CA. And in the case where there are more than two qualitative variables, do MCA instead, this time with the MCA function in FactoMineR. If the

active variables are both quantitative and qualitative, we use multiple factor analysis on mixed data. And with groups of variables, we perform multiple factor analysis.

Fifth question: if we decide to go with PCA or MFA, should we standardize the variables? We have seen that if variables are given in different units, it is necessary to standardize them. Now if all the active variables are in the same units, we can decide whether to standardize them or not. Standardizing the variables will go hand in hand with giving the same importance to each variable, whereas keeping all variables goes with assigning more importance to variables with higher variance. So, if we reduce the variable set with PCA, we must put: scale.unit=TRUE in FactoMineR when using the PCA function. The same question must be answered for each group of variables in MFA. If we standardize the variables of one group, we use the type "s" (for scaled) and if we do not want to standardize we use the type "c" (for continuous).

Sixth question: is there any missing data? If we have missing data, it is helpful to use the missMDA package alongside FactoMineR. The missMDA package helps us to impute missing data. Then, with the resulting full data table, we can use standard PCA, MCA, FAMD or MFA on it.

So, these are six important questions that need to be answered before running analyses. Now that this has been done, we can run the principal component method we desire. Depending on the type of data we have, this means doing PCA, correspondence analysis, multiple correspondence analysis, multiple factor analysis on mixed data or multiple factor analysis.

Next, we obtain the results and output plots. This means looking at the results using the summary functions, like summary.PCA, summary.CA, summary.MCA and summary.MFA. Then, if we want to improve the automatically output plots, we can use the functions plot.PCA, plot.CA, plot.MCA, and plot.MFA.

Then, it is always interesting to characterize the principal dimensions in terms of the initial variables. In order to do this, we can use the dimdesc function on the principal component method results.

Then, perhaps, we can do a clustering of the individuals, or a clustering of the rows or columns in the CA case, after running the principal component method. After doing the clustering, we can try and characterize the clusters found, using the initial variables. To do the clustering, we can use the HCPC function in FactoMineR.

And there you have it: the methods seen in the course.

Thanks for listening. Bye for now, and see you soon!