Audio transcription of the Multiple Correspondence Analysis course

Part 1. Data - issues

(Slides 1 to 9)

Slide 1: This week, we have four videos for you on multiple correspondence analysis, MCA for short. We'll have a look at the main features of the method, using a specific example to guide us along the way.

Slide 2 (outline): The videos look at the following things. First, we describe the types of data MCA can be used for. With this data in mind, we will look at what our goals are, and what issues we may have. This will lead us to ways to manipulate the data table. In multiple correspondence analysis, as in any principal component methods, we are going to build point clouds, including point clouds of the rows, and point clouds of the columns.

In the MCA context, we are going to have a point cloud of individuals, and a point clouds of categories. We will see how to visualize the point cloud of individuals, and how to interpret it using the categories. We'll then see how to directly visualize the point cloud of categories. We'll also show how in MCA, the point cloud of individuals, and that of the categories, can be shown simultaneously on the same graph. This is called the simultaneous representation of the point clouds. In the final video, we will go through various classical interpretation aids in multiple correspondence analysis. A running example, a survey by INSEE on leisure in France, will be used to illustrate the course each step of the way.

Slide 3: The data we are going to be working with is a rectangular table with I individuals, each with a row, and J qualitative variables as columns. Be careful to make the distinction between a qualitative variable like color, and categories of that variable, such as the color blue. The entry corresponding to the intersection of row i and column j holds a value Vij, which is the category of the j-th variable for the i-th individual.

For example. The most well-known example for MCA is: surveys. We have talked to I people, and the survey has J multiple-choice questions. Questions like, for example: what is your socio-professional class? Possible categories of this include working class, white-collar worker, etc. Another example question: what is your family situation: single? married? etc. Here, the variable is family situation. The categories of this variable are: single, married, and so on. We use all of this data to build the data table, which we can then feed to multiple correspondence analysis software.

Slide 4: Starting from this table, we construct an indicator matrix called the complete disjunctive table, or C D T. In this indicator matrix, rows are individuals, but now the columns are the categories of the qualitative variables. So, if the j-th column of the original table has Kj categories, there will be Kj columns for this variable in the indicator matrix, each corresponding to one category of the j-th variable. The entry at the intersection of the i-th row and the k-th column, which we call yik, will be equal to 1 if the i-th individual has category k of the j-th variable, and 0 otherwise.

Let's take an example. The marital status variable has Kj categories, and the individual i-prime is in the "married" category, which is the second one. In the complete disjunctive table, in the i-th row and in the columns corresponding to the j-th variable, we see that there is a 1 in the second column, because this is the category that this individual is in. Clearly, they have zeros in the other columns for the j-th variable. Here, we see why this table is called disjunctive and complete. It's disjunctive because in each block of columns, there is only one 1 and all the rest are zeros; and complete because there has to be at least one 1.

The complete disjunctive table plays a central role in MCA, as it's this that is analyzed by software packages. However, you, the user, never actually write it down or type it into the computer - the computer builds it for you, starting with the original table.

To start to understand the structure of the complete disjunctive table, we're going to calculate its margins. Let's start with the column margin. We calculate the sum of the terms of row i, and remember that yik equals 1 if the individual is in category k, and 0 otherwise. So, for each block of columns corresponding to one variable, we have one, and only one, entry with a 1 in it. Therefore, summing over ALL the columns, we simply get the number of variables there are, J. The column margin is therefore a constant, the same for all individuals. And the sum of the values in the table is I times J, i.e., the number of individuals multiplied by the number of variables.

Now, let's calculate the row margin. As yik equals 1 if the i-th individual is in category k, the sum of the entries in the k-th column equals the number of individuals in category k. Another way of thinking of this number is the multiplication of I times pk, i.e., the number of individuals multiplied by the proportion of individuals, pk, in category k. Then, when we sum the Kj terms of the part of the row margin corresponding to the j-th variable, we get I. So, again, the row margin sums up to I times J.

Slide 5: Now, let's look at the goals we have when we run multiple correspondence analysis. Imagine that we have survey data. Our individuals have filled in the survey, and our variables are simply the survey questions. First, we're going to take a closer look at the individuals, and then follow this with a study of the variables. When we look at individuals in multiple correspondence analysis, as in PCA, it's a multidimensional approach. A given individual, representing a row in the table, is analyzed based on the set of categories they belong to.

This will be used when we want to compare individuals. We're going to say that individuals are similar to each other if they are in a similar set of categories. If we think about this in terms of social network profiles for instance, two individuals have similar profiles, and thus resemble each other, if they've chosen the same categories, and are different if they have not many categories in common. The set of these similarities and differences is what we call inter-individual variability.

In multiple correspondence analysis, it's exactly this, the variability between individuals, that we're looking at. If they've all answered the survey in the same way, there aren't any statistics to do! But of course, usually, all the individuals don't answer in exactly the same way, and multiple correspondence analysis can delve into this variability, using a multidimensional point of view.

As we are working in the principal component method framework, the way we're going to describe this variability is by extracting principal components. We're going to try to shine a light on dimensions which separate, for example, extremely different individuals from average individuals. These dimensions will be interpreted in terms of the categories. For instance, we're not going to say things like: one dimension separates individuals 1, 3 and 4 from 10, 11, and 12. Rather, we are going to say things like: this dimension separates men and women, workers from the north from workers in the south, manual workers from white-collar workers, etc., etc.

The second part of the analysis deals with the variables. Here, these are qualitative, so we are going to be interested in associations between categories, because we say that two qualitative variables are linked if the categories of one have some kind of connection with the categories of the other. We'd like to get an overall visual representation of the associations between categories, leading to an overall look at the links between variables. And lastly, we are going to build synthetic variables. We're going to look for a quantitative indicator based on the qualitative variables that best summarizes them.

This whole framework greatly resembles principal component analysis; both work with a table of individuals in the rows, versus variables in the columns. Though from a technical point of view, things are quite different, because in principal component analysis, we have quantitative variables, whereas in multiple correspondence analysis, we have qualitative ones.

Slide 6: We are going to illustrate multiple correspondence analysis using the results of a survey taken in 2003 (two thousand and three) on eight thousand-four hundred and three people aged eighteen and over. The INSEE institute ran a survey on how people construct their identity, which they called the "history of life" survey. Part of the survey looked into the leisure pursuits of the French.

The data set consists of eight thousand-four hundred and three rows and 22 columns. The variables can be divided into two categories. The first 18 variables deal with various leisure activities. Typical questions were like: have you been to the cinema in the last twelve months, without being obliged to? This is noted cinema (yes/no) in the survey.

Similarly, we have variables like: reading (yes/no), listening to music (yes/no), and so on. We also have the number of hours on average of television watching, with five categories: not at all, one hour, around two hours, around three hours, and four or more hours. The next four variables are sex (male/female), age, divided into intervals (eighteen to twenty, twenty-one to thirty, thirty-one to forty, etc.), marital status (single, married, widowed, divorced, remarried), and socio-professional status (farmer, manual labourer, professional, senior management, employee, or other).

The data table has the following structure: the eight thousand-four hundred and three individuals as rows, and the 18 activities and 4 other variables as columns. The first thing we look at in a survey like this, is the number of individuals in each category.

Slide 7: Here, we show the number of individuals doing each activity, sorted in decreasing order. Listening to music is the most common activity, with five thousand nine hundred and forty seven participants (out of eight thousand-four hundred and three people), followed by reading, walking, etc. At the bottom of the list, we find collecting and fishing as the least-performed activities. Following this, we show the number of people watching TV for each of the five lengths of time.

Now, if we look at the other variables, we see in particular that one thousand-four hundred and ninety-eight people did not reply to the socio-professional status question. What we can do is to treat "no-response" as just another category. This means that socio-professional status now has eight categories.

Slide 8: Before starting to perform multiple correspondence analysis, we have to ask ourselves what type of analysis we want to do exactly, considering that the variables naturally fall into two groups: the activity group, and the supplementary variable group.

One way to go about things is to consider the activity variables as "active" variables, and the other variables as "supplementary" ones. If we run an analysis like this, the individuals are only examined using the activity variables. In other words, an individual is an activity profile. These activity profiles exhibit a certain variability. And we will go and look for the principal dimensions of variability in these profiles. Once this has been done, we then bring back the supplementary variables for illustrative purposes, and the question becomes to find links between the dimensions of variability of the activity profiles, and the supplementary variables.

Another point of view, the reverse of what we've just described, is to make the supplementary variables active, and introduce the activities afterwards. From this point of view, individuals are characterized by descriptive questions. So, we are confronted with this set of variables, in which we look for main dimensions of variability. In the end, this means studying the survey design. Then, we would move on to study the links between these dimensions of variability, and the leisure activities.

There is even a third way to attack the problem. We could set ALL of the variables as active, meaning that individuals are represented by heterogeneous data. Without going into details, this type of analysis requires a certain equilibrium between the two types of data, and leads on to other methods like multiple factor analysis. We're not going to go into this anymore here, and in what follows next, we will just consider the first point of view. Thus, individuals will be represented in terms of their activity profiles.

Slide 9: Multiple correspondence analysis works from the complete disjunctive table, which has been previously constructed. A first thing to note is that in MCA, all individuals have the same weight. There is no reason to accord more importance to the responses of some individual over another. So, as the sum of the weights has to equal 1, the weight of each individual is 1 over I.

Remember also that in the CDT, yik equals 1 if the i-th individual is in category k of the j-th variable. This value of 1 doesn't depend on the category k, and in particular, doesn't depend on the number of individuals in this category. However, if an individual is in a rare category, this defines them much more than if they were in a common category. To give the most extreme case, if all of the individuals are in the same category of a certain variable, that variable is useless when it comes to characterizing the individuals.

This is where the idea of dividing yik by pk comes from. This gives us the value xik, which will be bigger whenever a category an individual is in, is rare.

By coding the complete disjunctive table in this way, the mean of the xik is equal to 1. In multiple correspondence analysis, we center the data, which in the end means that the entry xik will actually be yik over pk, minus 1.

The table with the xik is the one on which we do the projection. To make the connection with principal component analysis, xik is simply the centered and standardized data. So, we've now seen what the data looks like in MCA, and the kinds of question we want to answer. In the next videos, we'll see how to construct and interpret point clouds of individuals and categories, starting from the table of the xik-s.

Part 2. Visualizing the point cloud of individuals

(Slides 10 to 23)

Slide 10 (outline): We've seen what the data looks like in MCA, and the questions we want to answer. In MCA, like in all principal component methods, we are going to build point clouds: point clouds for the rows, and point clouds for the columns. We're going to begin now with the point cloud for the rows, i.e., the cloud of individuals.

Slide 11: In a complete disjunctive table, individuals are represented as rows. This means a set of K numbers, so a point in K-dimensional space. Each dimension corresponds to a category. In MCA, each category is associated with a weight proportional to the number of individuals in it. As the sum of weights has to equal 1, the weight for the k-th category is pk over J. The point Mi has the coordinate xik, and we've seen that each individual is given a weight, 1 over I.

When looking at the whole set of points, we call the cloud N I. This cloud N I has a center of gravity, G I, which is in fact the origin, as the variables have been centered.

Now, we bring in the point i prime, so that we can look at the distance between the individuals i and i prime.

The square of this distance is written: the sum of the squares of the differences of the coordinates, where each square of the differences of the coordinates is weighted, corresponding to the weight of that category, i.e., pk over J. If we express this distance as a function of the complete disjunctive table, we get the following formula, which, after simplification, looks like this. We write the distance in this way when we are performing a multiple correspondence analysis as a correspondence analysis on the complete disjunctive table.

Now a few remarks on the choice of this distance measure. If two individuals are in the same set of categories and thus have exactly the same profile, the distance between them is zero. If two individuals share many categories, the distance between them will be small. If two individuals share many categories, except one, a rare one, which one of them is in, then the distance between them will be relatively large, due to the pk which will be small for one of them. If two individuals share a rare category, the distance between them will be relatively small due to this rare, shared similarity.

Now let's calculate the distance of a point from the origin. The square of the distance between an individual and the origin is equal to the sum of the squares of the weighted coordinates. If we work from the complete disjunctive table, we have the following formula, which after simplification, looks like this. We can clearly see the sum of the yik over pk in it. This means that the sum gets bigger when the categories possessed by an individual are rarer, i.e., when that individual is associated with small pk-s. Basically, this means that the more an individual possesses rare categories, the further they are from the origin.

To finish, let's calculate the total inertia of the point cloud N I. First, we start by calculating just one point's inertia, i.e., its weight multiplied by its distance to the origin. Then, for the total inertia of the cloud, we sum over all the individuals. When all the calculations are done, we find that the total inertia is K over J, minus 1. This value doesn't actually depend of the content of the table itself, but only on some details of its format, that is, the number of categories and the number of variables.

This is different from the result in correspondence analysis, where the total inertia is phi squared, and therefore measures the deviation from independence. On the other hand, this result is analogous to the one from principal components analysis with normalized data, where the total inertia is equal to the number of variables, and doesn't depend on the table's content, only on its format.

Slide 12: So, we have defined what the point cloud of individuals is. Like in any principal component methods, we're now going to perform a projection of this point cloud into a smaller-dimensional space. Often, we will even simply project into a plane, i.e., two dimensions. In order to do this, we have to apply factor analysis to the cloud, i.e., project the cloud onto a sequence of orthogonal axes with maximal inertia. In terms of the first two dimensions obtained, we will end up with a factor plane for the individuals which is the "best" two-d representation of them.

Slide 13: Looking now at the data set on French leisure activities, the first 18 variables, each corresponding to one activity, are going to be considered the active ones. So, it's these 18 variables that will be used to calculate the distance between individuals. The supplementary variables are treated as such, and will be used later to help interpret the results.

Slide 14: Like in any principal component method, let's start by looking at the decreasing values of the inertia associated with each subsequent dimension. Here, the bar plot shows that the first dimension has a rather larger inertia than those that follow. This makes us want to try and interpret the first dimension on its own, and initially put the others aside. The percentage of inertia associated with this first dimension is 24 percent. Can we consider this to be a large number, or not? Well, recall that we are analyzing fairly complex activity profiles, as there are 18 leisure activities and high variability.

Among our sample population, there are young people, old people, men and women, and we expect a great diversity of activity profiles. And it would be foolish to expect that all this diversity could be captured with two dimensions. If we now look at the third and fourth dimensions, we see that they have comparable, and relatively large, inertia. In this video, we are not going to go on and analyze the plane defined by the third and fourth dimensions, but this doesn't mean that it can't be usefully interpreted. If we DO go up to four dimensions, we arrive at 35 percent of the total inertia. As for the rest of the inertia, it’s just a collection of individual variability.

Slide 15: In multiple correspondence analysis, as in principal component analysis, the first thing we look at is the general form of the point cloud of individuals. Here, there is nothing particularly unusual about it, it's quite normal. There's not much more to say, really. But you may ask: what would an unusual cloud look like?

Slide 16: Here is an example which shows three distinct classes of individuals. If something like this appears in the data visualization, clearly these three classes will be important in the interpretation step. This is why it's always important to first look at the general shape of the point cloud of individuals, even if we end up simply concluding that it's an entirely unremarkable point cloud, showing nothing out of the ordinary.

Here is another example of a point cloud of individuals with a quite particular shape, which comes up quite often in MCA, a horseshoe shape. This is known as the Guttman effect or the horseshoe effect. When the horseshoe effect is in action, what tends to happen is that on one dimension, here the horizontal one, individuals are placed according to categories with increasing values, for example, lowest value categories to the left, medium in the center, high to the right, and on the second dimension, here the vertical one, separate outlier individuals, which take very high or low values, from more average individuals.

Slide 17: To get a better understanding of the main dimensions of variability, we can use qualitative variables from the data set. One idea is to color the points, representing individuals, in terms of the categories of a chosen variable. For example, here for the gardening variable, those that garden are shown in red, those that don't in black. We see that the red points are mostly at the top of the plot, while the black ones are mostly at the bottom. It therefore seems that the second dimension separates individuals that garden from those that don't. Of course, we can do this kind of coloring exercise for each variable, but this quickly becomes tiring.

However, this leads us to the more general idea of using the variables' categories, or the variables themselves, to try and characterize and interpret the point cloud of individuals. First, let's think about how we can use the categories of a given variable.

A category can be seen as a group of individuals, so it makes sense to show this directly on the graph of the individuals.
To do so, we can put a point representing a category at the barycenter of the individuals with that category. For the gardening variable, we have therefore two points: one for "gardening yes" in red, and one for "gardening no" in black. Just remember that the origin of the cloud is at the barycenter of a given variable since each of its categories has a weight proportional to the number of individuals in it.

Here, the "gardening no" point is closer to the origin than "gardening yes". This means that there are fewer people that do gardening than don't. 3356 say they do, and 5047 say they don't, and these numbers are proportional to the distance between the categories and the origin. Hence, categories with fewer adherents will be further from the origin.

Slide 18: In this plot, we show all of the categories of all of the variables, along with the points for the individuals. As the categories' points represent means, because they are barycenters, they are all quite close to the center of gravity. The "yes" categories are in red, and the "no" categories in black.

Slide 19: If we forget about the points for the individuals, and zoom in, this is what the categories look like on the first two dimensions. The first thing we notice is how the "yes" activities are all clustered together, as are the "no" categories.

We should say straight away that this is no accident. In the indicator matrix, nothing sends a signal that the "yes" categories are any "different" to the "no" ones. This is therefore evidence of a real phenomenon in the data. This separation is not really with respect to the first dimension, nor to the second, but more like a diagonal across the two. Let's take some time and try to interpret what this means.

Slide 20: To help us, let's look at a few of the individuals that seem well-separated by this diagonal. We can take individuals 5938 and 2432 on one side, and 8325 and 203 on the other. The plot showing the categories shows us that individuals 5938 and 2432 are on the side of the "yesses", which suggests they participate in a lot of leisure activities. In contrast, individuals 8325 and 203 are far on the "no" side, suggesting that they don't participate in many leisure activities.

If we look back at the data table to check, this information jumps out at us. 5938 and 2432 participate in nearly all the activities, whereas 8325 and 203 don't do any, unless you count watching a lot of TV as an "activity". This contrast, defined by the diagonal, thus clearly represents how active people are, and separates the active ones from the less active.

Slide 21: Now that we've interpreted the first diagonal, let's look at the other diagonal, and try to understand what it means. The group of categories to the bottom right includes cinema yes, computer yes, shows yes, sport yes, playing a musical instrument yes, gardening no, cooking no, knitting no. Over on the top left, it's the opposite: cinema no, computer no, shows no, sport no, playing a musical instrument no, gardening yes, cooking yes, knitting yes.

Let's again take a few individuals to try and understand what is going on with this diagonal. Individuals 255 and 6766 do the activities that tend to be associated with younger active people, and not the more home-based activities like gardening and cooking. In contrast, individuals 1143 and 5676 participate in the home-based activities associated with older people, and none of the active outside activities. We should note that each of these four individuals do about the same number of activities in total, which explains why they have about the same coordinates with respect to the diagonal. We're in fact in the process of commenting on a diagonal which is orthogonal to the first, and thus answering the question: for those doing the same number of activities, what separates them into groups?

Overall, what we can say is that variability among the individuals can be understood using these two diagonals. One diagonal separates individuals doing many activities from those doing few, and the other diagonal separates the "types" of activities, seeming to separate younger-person activities from older-person activities. Later on, we're going to see how to project the categories of the supplementary variables onto the graph, in order to answer the following question: are the dimensions of variability that we have just seen, related to some of the supplementary variables, like age, for example?

Slide 22: For now, let's go back to the graph of the individuals colored according to the gardening variable. Let's see what happens when we project them all onto one of the dimensions.

Here, they've all been projected onto the second dimension. What we can try to do is calculate an indicator function of the link between the coordinates of the individuals on the second dimension, and the qualitative, gardening variable.

An indicator of the link between a quantitative and qualitative variable is the correlation ratio between them. This is the indicator used in one-way analysis of variance. Often, we work with the square of the correlation ratio, which is equal to the percentage of variability of the quantitative variable explained by the qualitative variable. This indicator varies between 0 and 1. It's equal to 0 if there's no link between the two variables, i.e., if the means of the quantitative variable for each category of the qualitative one are equal. And it's equal to 1 if the means are different, and the within-category variability is zero, i.e., if all of the individuals from the same category have exactly the same value of the quantitative variable.

Here, the square of the correlation ratio between the gardening variable and the second dimension is 0.453. This plot doesn't allow us to entirely "see" this, but it does give us an idea of the separation of the points from the different categories of the gardening variable. And it must be said that a squared correlation ratio of 0.453 really does represent a good separation of the points from the two categories.

The square of the correlation ratio between gardening and the first dimension is zero point 0 4 7. This is quite small, and, as we can see on the plot, the projections of the red and black points seem to be quite mixed; there is no real separation between the two.

Slide 23: So, what we do is calculate the squares of the correlation ratios of each variable with the first and second dimensions, which then helps us to plot a graph showing the variables. Here, all the points are located in the 1 by 1 square with non-negative values. Note that on this graph, a variable with three or more categories could well have a coordinate of 1 in both the first and second dimensions. For the gardening variable, we place its point at 0.047 on the first dimension, and 0.453 on the second.

In order to begin the interpretation, we can look at the variables with the largest correlation ratios for the first and second dimensions. But we still take into account the actual value of the squared correlation ratios, because even if a variable is the one the most connected to a certain dimension, it won't be seen in the same way if its value is 0.3 rather than 0.9, for instance.

In our example, the connections are not very strong, not very close to 1. If we zoom in to the graph, we see that the Exhibition, Show, Computer, and Cinema variables are linked to the first dimension. Their squared correlation ratios vary between 0.3 and 0.4. These are not very big, but given the large number of individuals, these squared correlation ratios are (very) significantly different to 0. It's therefore ok to take into account these variables when interpreting the first dimension, and the plane defined by the first two dimensions.

Be careful nevertheless not to interpret the p-values exactly like in a classical test. Indeed, the principal components have been constructed using the qualitative variables, so we shouldn't be surprised that connections can be found, and that some of the squared correlation ratios are large. We made this happen. In other words, these p-values can't be interpreted in the same way as for classical statistical tests, except for the supplementary variables. Still, the information we get from them can be quite useful.

We also make note of a very interesting property: MCA looks for principal components that are as related as possible to the variables in the data set, in terms of squared correlation ratio. Concretely, to find the s-th dimension, we look among the orthogonal directions to the previously-found dimensions, for the one that's most related to the variables of our data set, where "most related" means the one that maximizes the sum of the squared correlation ratios.

We'll see in the next video how to build a point cloud of categories, and how to get an optimal representation of it for a given number of dimensions. We'll also underline the link between this representation and the way we have plotted and represented the individuals. It's the link between these two representations that makes MCA so useful.

Part 3. Visualizing the point cloud of categories, and simultaneous representations

(Slides 24 to 29)

Slide 24 (outline): We have seen how to build a point cloud of individuals, and how to interpret it using the variable's categories. In this video, we're going to see how to build point clouds of categories, and then how to get an optimal representation of them. We will discover the link between the optimal representation of individuals and the optimal representation of categories, by way of some transition formulas.

Slide 25: Let's go back to the data table showing the individuals and the categories. In row i, and column k, we have xik, which is equal to yik over pk, minus 1. Here, yik equals 1 if the individual is in category k, and 0 otherwise. And pk is the proportion of individuals in category k, and we subtract 1 to center each column. Let's look at the point cloud of categories, i.e., for the columns.

Each column takes one value per individual, of which there are I. It's therefore a set of I numbers. From this point of views, it's a point in I-dimensional space, where each dimension corresponds to an individual. Now, here is the i-th individual, which is associated with a weight of 1 over I. The category is represented by the point Mk, with coordinates xik. Remember that the weight of a category is proportional to how common it is in the data.

When we move on to the whole set of categories, we have a cloud Nk of points. This point cloud has the following property: for a given individual i, the sum of the coordinates xik of the categories of a given variable is equal to 0. Therefore, the origin is the same thing as the centre of gravity of the categories of a given variable. As this property holds for all the variables, the centre of gravity of the point cloud Nk is itself the origin.

Let's now calculate the variance of category k. This is equal to the square of the distance between this category and the origin in the space R I. If we look at the square of this, it's indeed the sum of the squares of the coordinates, weighted by 1 over I. Now, what happens if we try to state this variance in terms of the complete disjunctive table? This is what we get: when all has been calculated, we get 1 over pk, minus 1. This means that the distance between a point Mk representing category k, and the origin, becomes larger as the category becomes less common. In multiple correspondence analysis, uncommon categories are located far from the origin.

Let's look at how this distance varies as a function of pk for a few values. We can see that when pk moves from one half to one fifth, the distance doubles. When pk moves from one fifth to one tenth, it increases another fifty percent. Clearly, this is something important to remember, that the distance increases greatly with rarity. But, don't forget, in principal component method, it's inertia, and not distance, that counts, when constructing the dimensions.

So, let's look at the inertia of the k-th category. The inertia is the square of the distance between k and the origin, multiplied by the category's weight. It is clear to see that when a category is rare, the distance increases, but the weight decreases. There is therefore a pull in both directions. When all calculations are done, we get: one minus pk, over J, which shows that rare categories have high inertia. They are therefore going to have an influence on the results coming out of the multiple correspondence analysis.

Let's look at how the inertia varies in terms of frequency, with J=10 variables. When pk goes from one half to one tenth, the inertia practically doubles, passing from 0.05 to 0.09. Clearly, rare categories have an important influence. On the other hand, if we look at what happens when we move from one tenth to one hundredth, i.e., from a rare category to an extremely rare one, the inertia barely increases. So, basically this means that MCA really does recognize rare categories, but doesn't over-exaggerate the influence of extremely rare ones.

Now, let's calculate the distance between pairs of categories, and express this distance as a function of the elements yik and yik-prime of the complete disjunctive table. This distance can be expressed uniquely as a function of pk time pk-prime, and of the proportion of individuals that are in both categories k and k-prime. We call this proportion p-k-k-prime. We see that the greater the number of individuals that are in only one of the categories, the greater the distance between the two categories.

Slide 26: Now let's calculate the inertia of the j-th variable. This just means summing the inertias of that variable's categories. So we get the inertia of the category k, and sum over the Kj categories of the j-th variable. The resulting sum is (Kj minus 1), over J. Therefore, we see that a variable's inertia is proportional to the number of categories it has, minus one.

For example, in a questionnaire, the variable "sex" has two categories, and therefore an inertia of 1 over J, i.e., one over the number of variables. In contrast, the "region" variable has 21 categories, so the total inertia of this variable is 20 over J. At a first glance, this possible disparity between inertias, varying here by a factor of 20, seems worrying. It might make us consider only working with variables with more or less the same number of categories.

However, no need to worry! Multiple correspondence analysis easily deals with these kinds of differences. Indeed, a variable j with Kj categories corresponds to Kj indicator functions in the complete disjunctive table, and therefore corresponds to a sub-space with Kj minus 1 dimensions. The minus 1 comes from the fact that all of these Kj indicator functions are linked, because their sum is 1.

And, we've seen that the inertia of the set of categories for a given variable j is proportional to Kj minus 1. Which is precisely the dimension of the subspace in which this inertia is found. In other words, variables with many categories have high inertia, but this inertia is, in a way, diluted into a correspondingly large-dimensional subspace. This explains why the variable "sex" can really only be strongly linked with one dimension. Only one dimension can strongly separate men and women. In contrast, the "region" variable could be linked with many dimensions, with, for example, one dimension separating the north from the south, another the west from the east, etc., etc.

And, to bring this to an end, we can calculate the total inertia by taking the sum of the inertias of all the variables. And, surprise, surprise, we get the same value as the inertia of the point cloud of individuals, that is, K over J, minus 1.

Slide 27: As has now become natural to us, we're going to get to work on this point cloud of categories using principal component method. That means sequentially constructing orthogonal dimensions that maximize the inertia.

This is how the visual representation of the categories comes out. It's the optimal representation. Notice how this graph strongly resembles the one we constructed earlier when we were using the categories to interpret the individual's plot, though it's not completely the same. Nevertheless, we can interpret this new graph in mostly the same way: a first dimension, or rather, a first diagonal, that separates individuals doing lots of activities from those doing few. And a second diagonal dividing generally young-person activities from older-person ones.

Slide 28: In a similar kind of way to what we did when studying the individuals, we can project individuals onto this plot of the categories. An individual can be considered like the set of categories they are in, so we can place an individual at the barycenter of its categories. We see that all the individuals end up in the center of the cloud, which makes sense, because they are essentially representing means here.

Slide 29: Now let's take a pause and a stock of all these graphs we've plotted, and try to see if we can find any connections between them. First, let's go back the optimal representation of the point cloud of individuals, that we constructed in the previous video.

To interpret it, we have placed each category at the barycenter of the individuals in it. If we note Gs(k) the coordinate of the k-th category in the s-th dimension, we have this formula. yik equals 1 if the i-th individual is in category k, and 0 otherwise. So here, we have the sum of the coordinates of the individuals in category k, divided by the number of individuals in category k. i.e., we're calculating the mean coordinates of the individuals in category k.

Now, we look again the optimal representation of the categories that we built earlier in the current video. We see that the category's coordinates are dilated, or more spread out, than in the graph on the left.

Now, let's project each individual onto the barycenter of the categories it's in. The coordinate value of the i-th individual on the s-th dimension is therefore the mean of the categories of that individual. In this formula, yik, which we multiply with Gs(k), corresponds to the sum of the coordinates of the categories of the i-th individual, and since an individual is in only one category per variable and there are J variables, we divide by J to have the mean of the coordinates of the categories of that individual. The two graphs, on the left and the right, are different. BUT...

One is just a dilated version of the other. In the graph on the left, if we fix the individual's positions but dilate by a factor of 1 over the square root of lambda-s, this is what we get. Lambda-s is the eigenvalue associated with the s-th dimension, and it's clearly a dilation because lambda-s is always less than 1. When writing this relationship, we just have to introduce the factor of 1 over the square root of lambda-s.

In the graph on the right, we fix the positions of the categories, and dilate now the point cloud of individuals, again by this coefficient of 1 over the square root of lambda-s. In the transition formula, we only introduce this term of 1 over the square root of lambda-s. And what do we see? Well, with these two dilations, the graph on the left is identical to the graph on the right.

This is the graph which we call the simultaneous representation. It is the graph which pops out of a software package when we run MCA. The representation of the individuals is optimal, and so is that of the categories. Both transition formulas show how to move from one representation to the other, and help our interpretation of the graph. The formula on the left says that a category is at the pseudo-barycenter of the individuals that are in it. The formula on the right says that an individual is at the pseudo-barycenter of the categories it is in.

Thus, an individual will be found on the same side as the categories it's in, and away from the categories it's not in. Similarly, a category will be near the individuals that are in it, and away from individuals that are not. This symmetry between the two properties means that we call it a double barycentric property. So, we've now seen how to build the point clouds of individuals and categories. In the next video, we'll have a look at some interpretation aids, and consider ways to use supplementary information, like the supplementary variables we saw earlier, when doing MCA.

Quatrième partie. Aides à l'interprétation

Part 4. Interpretation aids

(Slides 30 to 39)

Slide 30 (outline): So far, we've seen how to construct point clouds of individuals and categories in MCA. In this video, we're going to show you some interpretation aids, shared by all principal component methods, and also look at how to use supplementary information, including supplementary variables, in MCA.

Slide 31: The inertia of a dimension in MCA is special because it's equal to the mean of the squares of the correlation ratios between the dimension and the variables. This property validates our way of interpreting the synthetic variables defined by the MCA dimensions. So, these MCA factors are quantitative variables which summarize the qualitative one. Note also that, as the squares of the correlation ratios are between 0 and 1, the contribution of a given variable to a given dimension is bounded by 1. Furthermore, the mean of the correlation ratios, and thus the inertia, is also found between 0 and 1.

Now let's look at the percentages of inertia. These are generally lower than in PCA and CA because the individuals are located in a high, K-J dimensional space, which gets bigger and bigger as the number of categories per variable increases.

The following little calculation shows what the percentages of inertia associated with given dimension tend to be small. We calculate the inertia of a dimension, lambda-s, over the total inertia, and since the eigenvalue lambda-s is less than one, we find that the percentage of inertia has to be less than J over (K - J), multiplied by one hundred. So, with 10 variables and 10 categories per variable, even if the variables are strongly linked, the maximal percentage of inertia that can be in a given dimension is 11 percent.

Another quick calculation shows that the mean inertia equals 1 over the number of variables. This value can help us decide how many dimensions to interpret in MCA. Indeed, we can forget about interpreting dimensions with inertia less than 1 over J.

Slide 32: Interpretation aids like contribution, and quality of representation, can be calculated as for other principal component methods like PCA and CA. However, the projection quality tends to be weak on each dimension, which is normal because there are many dimensions. As for the categories, as a weight is associated with each of them, the most important contributors are not necessarily the most spread out points on the graph.

The contribution of a given variable to the construction of a given dimension can be calculated by summing the contribution of all of its categories. This contribution is equal to the square of the correlation ratio between the dimension and the variable, divided by the number of variables. To calculate the relative contribution of each variable, we divide by the inertia of the dimension, lambda-s.

Slide 33: We can now turn to looking at how to use supplementary information to interpret the simultaneous representation graph. To do this, we're going to use the transition formulas to calculate the coordinates of the supplementary elements. By "element", we mean either a row, or a column, of the table. So here, this means either an individual, or a category. In any principal component method, an element is labeled "supplementary" if it hasn't been used to build the dimensions. Note that, for categories, all the categories of a given variable must have the same status: they all have to be active, or all supplementary.

This is why we label variables "active", or "supplementary". Barycentric properties help us to calculate the position of a supplementary individual or category. Let's take for example a supplementary category. Once the positions of the individuals have been set in stone, we can calculate the barycenter of any subset of them. In practice, MCA almost always comes with supplementary variables, as these provide context to what we are doing, and are therefore extremely precious to us.

This plot shows the categories of the supplementary variables. In order not to put too much on the same plot, we don't show the active categories or the individuals. However, individuals as well as the active or supplementary categories can be put on the same plot. We see that the supplementary variables are linked to the dimensions.

For instance, the groups of ages appear in a rising fashion from bottom-right, up to the top and around to the left. Similarly, professions and marital status categories spread out in certain ways along the dimensions. Moving on to the interpretation of the MCA plot, we can say that towards the bottom-right, we have the younger individuals, mostly white-collar workers, living alone, with lots of activities like cinema, shows, and computers. At the top, the individuals are older, with activities like fishing, knitting and gardening. And lastly, to the left, we find manual labours and blue-collar workers, relatively older, with few leisure activities. These interpretations involved a two-step process. First, we used only active variables, and then we brought in the supplementary ones.

Slide 34: Also in our set of interpretation aids, we include the notion of supplementary quantitative variables. In MCA, active variables are by definition qualitative, so quantitative variables must be, by definition, supplementary. In our example, we were able to interpret the first diagonal in terms of the number of leisure activities undertaken. From which comes the idea to literally create a quantitative variable called "number of activities per individual", and see what it looks like.

To visually represent this variable, we're going to proceed like for PCA, and use the correlation circle representation. Each variable is represented by its correlation coefficient with the dimensions. This gives us the following graph, which shows that the number of activities performed is perfectly characterized by the plane defined by the first two dimensions. Essentially, it corresponds to the first diagonal, which comforts our initial interpretation.

Note that if we do want to make a quantitative variable active, we can discretize it into classes, making it qualitative, and it can then be used as an active variable.

Slide 35: Like in PCA, we can characterize the dimensions in MCA using active or supplementary variables, and quantitative or qualitative ones. For quantitative variables, we can calculate the correlation coefficient between each of them and a given dimension, and then sort them by importance. Here, we see that the variable "number of activities" is strongly related to the first dimension. For qualitative variables, we can construct an analysis of variance model for each variable, and then describe the coordinates in a given dimension in terms of the given variable. We can also do an F-test to have a look at the overall link between the variable and that dimension, and do Student t-tests to see which categories have specific coordinates in each dimension. Here, all of the p-values are extremely small, as there are a lot of individuals in this data set.

Slide 36: There is another way to present the data when we have a set of qualitative variables, which is called a Burt table. This table sets up the whole set of variables against themselves. In the table we show here, each block is a contingency table between two variables. The rows and columns of each contingency table are the categories. This means that a Burt table puts up the categories of all variables against themselves. The Burt table summarizes all of the pairwise relationships between variables. In a way, it is analogous to a correlation matrix which brings together all the pairwise correlations between variables.

So, what kind of analyses can we put into place to study a Burt table? One simple idea, coming by analogy from contingency tables, is to do the same kind of analyses, but for the Burt table. When we run CA on a Burt table, we get exactly the same dimensions as in MCA, but with different eigenvalues. The eigenvalues coming from a Burt table are the squares of the eigenvalues coming from the complete disjunctive table. So, which eigenvalues should we work with? Previously, we interpreted the eigenvalues coming from the complete disjunctive table, in terms of the mean of the squared correlation ratios. As this interpretation is quite useful, we're going to stick with it.

Lastly, this analogy teaches us something: that in the end, MCA really only depends on the pairwise links between variables. Therefore, it's again analogous with PCA, which only depends on the correlation matrix.

Slide 37: Well, it's time to wrap it up for multiple correspondence analysis, and make some conclusions. First, we have seen that MCA is the principal component method to use when we have tables of individuals versus qualitative variables. The general strategy in MCA is to look for the principal dimension explaining the variability between individuals, and to closely examine links between variables. From this general point of view, the situation is identical to PCA. Technically speaking, the main ways of interpreting the data are quite simple:

An individual is at the barycenter of its categories, and a category is at the barycenter of the individuals that are in it. This gives us a quite broadly useful methodology, because these types of data tables are quite common in reality, and can be easily interpreted, as we have seen. In particular, multiple correspondence analysis is a particularly useful method for dealing with survey data.

The possibility of interpreting the eigenvalues as the means of the squared correlation ratios is a fundamental property of MCA. This validates the interpretation of the MCA factors as synthetic variables. The MCA factors are quantitative variables, which summarize, in a way, the qualitative variables.

We have insisted in using the squares of these links. This representation is particularly precious to us when there are a large number of variables.

Like for other principal component methods, it's useful to validate our interpretations of the results by going back to look at the data. MCA will suggest to us links between qualitative variables, and if we want, we can then construct pairwise tables for such variable pairs, and analyze and visualize them using CA.

Also, the kind of convergence we see between the analysis of the complete disjunctive table and the analysis of a Burt table, is an argument in favor of the general method we've presented. It's always interesting to see when different points of view lead to the same result.

Lastly, MCA allows us to pass from a table of qualitative variables to a table of dimensions, i.e., quantitative variables. It's therefore possible to see the method as a preprocessing step for qualitative data, in order to pass it to a clustering algorithm, for example.

Slide 38: To finish, here are a few references related to what we've seen in the MCA videos. You've now completed all of the course videos for MCA, so, if you're ready to go on, you can now have a look at the video showing how to run MCA in practice, using FactoMineR. Good luck!