Audio Transcription of the Correspondence Analysis Course

Part 1. Data - Introduction

(Slides 1 - 14)

Slide 1: This week, we have for you 5 course videos on correspondence analysis.

Slide 2 (outline): In the videos, we will see the following: first, we start by describing the data, giving a little notation, and considering questions to ask when running correspondence analysis. We'll see that the main point of correspondence analysis is studying the links between pairs of qualitative variables. This really means looking at the difference between the given data, and what it would be like if the variables were independent. We're therefore going to see how the analysis captures deviation from independence. Our reasoning will mainly be geometrical, creating point clouds for the rows and point clouds for the columns. Projecting these clouds onto planes will give some useful representations.

We will also have a look at percentages of inertia. From this point of view, correspondence analysis is no different from other principal component methods, like principal component analysis. We will also give some technical results regarding inertia, and these indeed are quite specific to correspondence analysis. Similarly, the simultaneous representation of rows and columns is rather specific to correspondence analysis, and takes advantage of so-called barycentric properties. At the end, we'll give you some interpretation aids, including looking at contributions, and the quality of representation. In these, correspondence analysis has no real differences with, for example, principal component analysis.

Slide 3 (outline - continued): Let's start by describing the data.

Slide 4: The data we are working with here consists of n individuals for whom we have two qualitative variables. The data are put into a contingency table, in which rows are the possible categories of variable V1, and columns the possible categories of variable V2. The intersection of row i with column j has the entry xij, which is the number of individuals choosing or possessing category i of V1 and category j of V2. This table is also named a two-way cross tabulation. Let's look at a few examples of contingency tables.

As a first example, as well as a historic one - because it's without doubt one of the first applications of correspondence analysis - we turn to the study of the vocabulary of characters in the play Phèdre. Each row corresponds to a character, and each column to a word used in the play. So xij is the number of times the character i uses the word j.

As a second example, we can look at people's perception of luxury perfumes. We can ask people to associate words with each perfume. Each row is a perfume, each column a word, so xij is the number of times perfume i is described with the word j.

One more. In ecology, we may be interested in the ecological diversity of several locations. Each row is a location, and each column a plant or animal species. As before, xij is therefore the number found of plant or animal j in location i.

Data like this is everywhere, and for those of you who have heard of the chi2 test for independence, it's used for this type of data table.

Slide 5: The data we're going to use to illustrate this part of the course on correspondence analysis is a data table with the G8 countries as rows, and categories of Nobel prize as columns. In a given entry of the table, we have the number of people for one G8 country that have won the Nobel prize in a certain category. This table gives the allotment of the 570 Nobel prizes given out between 1901 and 2015 to these countries. For instance, 51 Americans

have won the Nobel prize in chemistry. This contingency table has marginals added to it. This means that the column added on the end contains the marginal sums of the values along each row. Similarly, the row added to the bottom of the table contains the marginal sums of each column. For instance, along the Germany row, the 80 in the last column is the sum along that row. Similarly, in the Chemistry column, 121 is the total number of Nobel prizes in chemistry given out over this period to G8 countries.

All of which leads us to the following question: Is there a relationship between country and type of Nobel prize? In other words, are some countries better at certain disciplines than others? Are certain prize categories more likely to go to one country than another?

Slide 6: Contingency tables are usually built using two variables. So, at the start, our data is n individuals, for which we have two qualitative variables. In the data table we see here, individual l is in category i for the variable V1, and category j for the variable V2.

We then build the contingency table with the categories of the first variable, V1, as rows for example, and the categories of the second variable, V2, as columns. At the intersection of row i and column j, we find xij, the number of individuals found in or choosing category i of V1 and category j of V2. If we add up all the values in the table, we get n. The table therefore shows the distribution of the n individuals over the I x J possible attributions.

Slide 7: Using this contingency table, we are going to calculate the corresponding probabilities. Then, correspondence analysis will go to work on this table of probabilities. To get a probability, for example fij, we simply divide the number xij by the total n. This gives us the probability of being both in category i of V1, and category j of V2. We fill the table with these probabilities. If we now sum them all up, we get 1, which is the sign of a true probability distribution.

We finish the table by first adding the marginal column, fi.. This is the sum of the i^th row, for each i.

Then we add the marginal row, f.j. This gives the sum of the j^th column, for each j.

We want to look at the interconnectedness of V1 and V2, i.e., how far away their relationship is from independence?

Slide 8 (outline): Let's take a closer look at what independence between pairs of qualitative variables means, and show how correspondence analysis captures deviation from independences.

Slide 9: Let's start by reminding ourselves what independence means for two events: we say that two events A and B are independent if the probability of A AND B is equal to the product of the two: the probability of A times the probability of B.

For two qualitative variables, we are going to see whether fij, the probability of i AND j, is equal to the probability of i: fi., multiplied by the probability of j: f.j. But this has to be true, of course, for all i and j. In terms of vocabulary, independence means that the JOINT probability is equal to the product of the MARGINAL probabilities.

There is another way to write this independence model, which we can see here. fij divided by fi. is equal to f.j. In this notation, we say that the CONDITIONAL probability is equal to the MARGINAL probability. This might seem a bit formal to you, but in the end, this way of writing is closer to the intuition of what independence is. Here, the conditional probability is the probability of being j, knowing that we are i. So, if there is independence, this probability is simply the probability of being j, even if we have no information about the other variable, i. In a nicely symmetric way, we also can divide fij by f.j, and under the hypothesis of independence, this new conditional probability, i.e, the probability to be i given that we are j, is indeed equal to the marginal probability of being i: fi..

Slide 10: The relationship between two qualitative variables is thus the deviation, or difference, between the observed data: fij, and the independence model: fi. times f.j.

When we study this relationship between two qualitative variables, usually, we first perform a test of significance of the relationship, using a chi2 test. Let's now remind ourselves of what a chi2 test looks like, and how we use it to compare observed values with theoretical values. The theoretical values are simply the probabilities under the independence model: fi. times f.j, multiplied by the total number of individuals: n

In this chi2 expression, we can factor out the n, and turn it into n phi2, where phi2 looks at the difference between the observed and theoretical probabilities.

So it's an indicator of the strength of the relationship - this difference between observed and theoretical probabilities. Why do we say: the strength of the relationship? Because this term no longer depends on the number of individuals, only on the probabilities.

There is something else to think about too: what is the nature of the relationship? i.e., how do the categories of the two variables relate to each other? Correspondence analysis works with the table of probabilities, but says nothing about significance. It really just aims to visualize the nature of the relationship between the two variables.

With this in mind, again focus on what is really happening when we study this kind of relationship. In particular, let's take an intuitive look at what it means to have a significant relationship. Imagine for example, that we did a little survey and discovered at the end of it that everyone who had blue eyes, wore glasses. This could be a stronger relationship. But how about now if I said to you: In fact, this survey involved 4 people. As you can well imagine, without even doing a chi2 test, you can already sense that this relationship is not statistically significant.

Slide 11: So, how does correspondence analysis get a fix on the near-ness or far-ness to independence? Let's first look at doing an analysis, row by row. In this analysis, we use the independence model that: the conditional probability given i equals the marginal probability. For this, in the table, we are going to divide each element in the row by it's marginal. This gives us what's called the row profile, which is nothing but a conditional distribution. In this row, we have the distribution of the variable V2 given that we are in category i of variable V1.

This row profile, we are going to compare it to the mean row profile, i.e., the marginal distribution of the variable V2.

So, correspondence analysis is going to compare each row profile to the mean profile.

This is basically a multidimensional way to look at the deviation from independence, as we will simultaneously consider ALL the profiles. So we are indeed comparing the conditional probability fij over fi., with the marginal probability f.j, over the whole set of j's.

Slide 12: As this all sounds a little bit complicated, let's illustrate what we mean with our Nobel example: what does comparing a profile to the mean profile signify? Here is the table giving the row profiles with each country, as percentages. The sum of each row therefore gives 100.

Now, color in the columns, i.e., the prize categories, with a block whose width is proportional to the percentage of that category in the mean profile, i.e., the marginal at the bottom of the table. So, the physics block is quite wide, whereas the peace block is quite narrow.

Now, for each row, i.e., each country, build a bar plot, with each bar joined the next, to visualize that country's profile. Look at Italy for example. The bar for Italy gives the distribution of the number of prizes won, per category, by Italians. More precisely, 31.6% of the prizes were in literature, for example. To get an understanding of this percentage, you have to compare it to the percentage of Nobel prizes given in literature out of all Nobel prizes over all the countries, which is 8.6%. Thus, the percentage of Nobel prize winners in literature from Italy is particularly

high. In this way, we can quantitatively compare the distribution of Nobel prize winners in Italy to that of all Nobel prize winners from the G8 countries.

This comparison of the Italian row profile with the mean profile helps us to answer the question: do Italians tend to win their Nobel prizes in some categories rather than others?

Slide 13: Now we're going to study the same table, but this time looking at the columns. Here we're dealing with the independence model: fij over f.j equals fi.. We build the table, including the column profiles. Look at column j. fij over f.j, it's the conditional probability to be in i, given that we are already in j. The sum of terms in this column equals 1, so it's a probability distribution.

Like for the rows, we can now build the mean column profile. This is made up of the terms fi., which is the probability to be in i.

The idea in correspondence analysis is to compare the column profiles with the mean profile. Again, it's a multidimensional approach for looking at the deviation from independence, multidimensional because we're going to consider ALL the terms in a column, and compare them with ALL the terms in the corresponding mean column profile.

Slide 14: Let's now look at our example, and see what "compare a column profile with the mean profile" actually means. Here are the column profiles, which each sum to 100.

We can color in each row with a block whose height corresponds to the value of that row in the mean column profile.

Then, like before, for each column profile, we build the bar plot with the bars piled one on top of the other. This helps us see better what's going on. We then compare each profile to the mean column profile. For example, look at the literature column. The distribution of Nobel prizes in literature by country seems to be quite different to the mean profile. For instance, the percentage of literature prizes given to Americans is very low in comparison to its overall performance: with 16.3% in literature, compared with 45.1% overall. In contrast, in literature, France and Italy are over-represented with respect to their overall piece of the Nobel pie.

So, when we compare this column profile to the mean profile, we're asking the following question: Is the distribution of Nobel prizes in literature per country the same as the distribution of all Nobel prizes?

We have now seen what types of data correspondence analysis works on. We've also seen how it involves comparing the data to an independence model. In the next video, we'll see how to visualize the divergence from independence by constructing a cloud for the rows, and a cloud for the columns.

Part 2. Visualizing clouds for the rows and columns

(Slides 15 - 21)

Slide 15 (outline): We have seen that correspondence analysis works with respect to an independence model. Now, to further describe correspondence analysis, we are going to work with essentially geometric ideas, and construct point clouds in space.

Slide 16: First, let's build the cloud of row profiles. A row profile is a set of J numbers, which can be seen as a point in J-dimensional space, R^J. Each dimension corresponds to a category j of the variable V2. Therefore, in the jth dimension, the row profile i has coordinate value: fij over fi.. If we then consider all of the row profiles i together, we obtain a cloud of row profiles, points in R^J, which we can call N_I.

To this point cloud, we can add what we called the mean profile as another point. Its jth coordinate is simply f.j. We call this G, for center of GRAVITY. In fact, G can be seen as the center of gravity of the cloud N_I as long as we associate each point i with a weight, proportional to its marginal value, fi.. In this space, what interests us the most, is to compare the positions of the row profiles i with the mean profile. For this, we're first going to take the mean profile and translate it to the origin, and the other points along with it.

In this space, we need to be able to calculate distances. The distance measure used in correspondence analysis is called the chi2 distance. It greatly resembles the usual Euclidian distance, indeed, it involves a sum of differences. We calculate the difference between profiles i and i-prime, that is, between fij over fi., and fi-prime-j over fi-prime.. We take the square, and then sum over all j. The difference between this and the Euclidean distance is that here, each dimension j is associated with a weight, 1 over f.j. The justification for this distance measure will become clear later in the course.

The centre of gravity of the cloud, corresponding to the mean profile, has coordinates f.j. Therefore, the distance between it and profile i is simply this distance.

Slide 17: Just like for the cloud of row profile points, we can build the cloud of column profile points. A column profile is a set of I values, so it becomes a point in I-dimensional space. In this space, each dimension corresponds to a category i of variable V1. For this category, the coordinates of the column profile are fij over f.j. There are J column profiles in all, which together make up the cloud of points: N_j.

We then add the mean profile G_j to the cloud, which has coordinates fi. in the i-th dimension. This mean profile can be considered the cloud's center of gravity, as long as we assign to each column profile j a weight corresponding to its marginal sum, or more precisely, its marginal probability, which here is f.j. In this space, we will be greatly interested in the distance between the column profiles and the mean profile. To simplify, we translate the centre of gravity to the origin, and the other points along with it.

To calculate distances in this space, we use, as for the cloud of row profiles, the chi2 distance. Here, this represents a sum of squares of differences in coordinates, where each difference is weighted by 1 over fi.

The cloud's center of gravity, corresponding to the mean profile, has coordinates fi.. Therefore, the distance between profile j and the mean profile is simply the square root of this quantity.

Slide 18: So, what happens if there is independence?

Well, if so, the conditional probability equals the marginal probability. Which means that all the profiles are the same as the mean profile. In other words, the cloud ends up just being a single point at the origin.

And this is true for both point clouds.

Slide 19: Therefore, the further the data is from independence, the further the profiles are from the origin.

Let's now calculate the inertia of the point cloud N_I with respect to its center of gravity. The inertia of a point cloud is the sum of the inertia of each point. The inertia of a point is the mass, times the square of the distance to the center of gravity. So the inertia of point i is fi. times the square of the chi2-distance between i and G_I. When we calculate this, we get an intermediate result, which is none other than phi2, or if you prefer, chi2 over n. The point cloud's inertia is therefore clearly an indicator of deviation from independence.

As what interests us is this deviation from independence, we end up studying the inertia of N_I. This is the correspondence analysis point of view.

The same reasoning holds for the cloud N_J, and we even have a surprising result: the inertia of N_I is equal to the inertia of N_J. This is an important thing to realise, wrapped up in what we call duality, a kind of paired equivalence. In fact, what this means is that it's the same thing to analyze the table in terms of rows OR in terms of columns. This is a fundamental result in correspondence analysis.

We see that in correspondence analysis, rows and columns have perfectly symmetric roles. This property distinguishes correspondence analysis from other methods, like principal component analysis, for example.

Slide 20: We have turned around our initial question on the link between qualitative variables, into a question of deviation from independence. This deviation can be interpreted geometrically as the inertia of a point cloud N_I with respect to the origin. And how does correspondence analysis proceed? As all principal component methods, it simply projects the point cloud onto a series of axes, of dimensions, of maximum inertia.

When we have a sequence of axes, we can take pairs of them and draw planes. Let's look at how we find the first plane. So, the best planar representation of the point cloud N_I is found, as we said, by projection, onto some plane P. Here we have the point Mi, which corresponds to the i-th profile. This point Mi is projected onto the plane as the point Hi.

In fact, what you see are the points Hi, and you hope that they are not too different from the points Mi, so that you can make some useful conclusions. So, what criteria should we use to find the plane P? Remember that we want to best represent the inertia of the point cloud N_I. So, we want to maximize the projected inertia.

Let's have a closer look at this. Recall that the inertia of the point Mi is its mass fi., times the square of its distance to the origin. What we therefore want is to find the plane P so that the sum of the inertia of the projected points Hi over all i, is maximal. To begin, we will find just one axis, u1, with maximal inertia. Then, a second one, u2, that maximizes inertia, with the constraint that it must be orthogonal to u1. The pairing u1, u2, gives the plane P.

As for notation, the inertia of the s-th axis will be called lambda s. We use the word lambda because this value is in fact an eigenvalue, and we say it's the eigenvalue of rank s because it turns out to be the s-th largest.

Slide 21: Here is the graphical output of correspondence analysis, applied to our data set. So, it's a plane, showing both the rows' and columns' projections. The rows, countries, are blue; the columns, prize categories, are red. So, how can we interpret this plot? Something we've gone on and on about is the distance to the center of gravity. So let's look at this for the UK. It's very close to the centre of gravity.

Look back at the row profile for the UK, which is approximately 25%, 6%, 8%, 28%, 12% and 22%. And compare this with the mean profile. We can see that it's pretty close: 25% versus 21%, 6% versus 11%, 8% versus 9%, 28% versus 25%, 12% versus 9% and 22% versus 26%. Clearly, the UK's profile is close to the mean profile. On our graphical

output, this means that the UK ends up close to the origin. How about Italy? Well, it's quite far from the origin, which means that it's profile is quite far from the mean profile.

Let's take a closer look at Italy. For example, 5% of the Italians' prizes are in chemistry, while the mean value is 21%. And it's reversed for literature: 32% of Italy's prizes are in this, compared with the average of 9% over the G8 countries.

Now, usually, when we have a plot, we want to label the dimensions. How can we do that here? Well, let's look at each group at a time, for example, let's start with the blue points, representing the countries.

We see a split between the North American countries to the left, the Latin countries (France and Italy) to the right, and Japan and Germany to the bottom. It's quite difficult to label the dimensions using the countries, unless we know that, for example, the North Americans are known for economics, and the French and Italians for literature.

Now, let's look at the red points, i.e., the prize categories. These are spread out so that the medicine and economics prizes are to the left, and literature and peace to the right. The second, vertical axis, separates the physics and chemistry prizes from the economics one.

We could hypothesize that the 1st dimension separates the scientific prizes from the others, while the 2nd separates physics and chemistry from economics. In fact, the interpretation of the dimensions is the same for both the rows and columns, due to the duality principal we talked about earlier. In reality, we're studying the same table in terms of its rows, or columns, but it's the SAME table, which we compare with the one we would have if there was independence.

So, in summary, we've seen how to visualize the point cloud of the rows, and the point cloud of the columns. In the next videos, we'll look more at two features of correspondence analysis: inertia, and simultaneous representations.

Part 3. Inertia and percentage of inertia

(Slides 22 - 26)

Slide 22 (outline): We've now seen how to plot point clouds for the rows, and point clouds for the columns, and how to project them onto graphical plots. In correspondence analysis, like all principal component methods, the first indices we want to look at are the percentages of inertia. We'll see in this section that in correspondence analysis, the inertia is a rather particular thing.

Slide 23: Let's start by talking about the percentage of inertia. Like in all principal component methods, percentages of inertia are the first indices we look at. The question we want to ask is the following: we have a point cloud representation. How good is it? Generally speaking, the quality of the representation is measured with respect to the ratio of the projected inertia over the total inertia. Usually, we multiply this by one hundred to get a percentage. In the precise case of a cloud N_I and the s-th dimension, the quality of representation of the cloud N_I onto the s-th dimension is the projected inertia of N_I onto this dimension, divided by the total inertia of N_I. We can write this in a way you've already seen, with lambda_s over the sum of the lambda_k. As we just said, this is then given as a percentage.

Let's look at the percentages of inertia in our example. We see that in the first dimension, the percentage of inertia is 54.75, which is quite high. We can therefore say that the 1st dimension represents 54.75 percent of the deviation from independence. In the second dimension, the percentage of inertia is 24.60. Thus, the first two dimensions together represent almost 79 percent of the deviation from independence. Essentially, this means that we can just stop, and interpret these two dimensions only.

Here's a mathematical property of what we have done: the projected inertia, i.e, the eigenvalues, can be added up from each dimension to the next. This comes from the fact that the dimensions are orthogonal. Therefore, the sum of all the eigenvalues, or the sum of the projected inertia if you like, equals the total inertia of the point cloud N_I. This is true for all principal component methods. In the particular case of correspondence analysis, this total inertia equals Phi squared.

Let's do a few calculations using correspondence analysis on our example: we can multiply the total inertia, 0.1522, by n, the sum of the table, which is 570, and we get a chi2 value of 86.75. Given the number of degrees of freedom here, the p-value is 2.77 times ten to the power of minus 6, which is tiny. There is clearly a highly significant connection between the countries and the prize categories in our example.

Here is something else we can do with the percentages of inertia. As these decrease as the rank of the dimension increases, we can use this decreasing sequence as a guide to find a cut-off, i.e., choose the number of dimensions to keep. As an example, here is the decreasing sequence of eigenvalues of a correspondence analysis on a contingency table facing off ten white wines from the Loire region with 30 descriptors. The wines are in the rows, the descriptors in the columns, and xij is the number of times the descriptor j was associated with wine i.

When we look at the decreasing sequence of eigenvalues using a bar plot, it's clear that the first two eigenvalues are much larger than the others. The first two dimensions therefore dominate in terms of inertia, suggesting that the best way to interpret the data is to just look at the plane defined by these two dimensions.

Slide 24: In correspondence analysis, it is useful to distinguish between the inertia, and the percentage of inertia, because the inertia are themselves interesting. They make up part of phi², which gives a global measure of the link between two variables. In correspondence analysis, the following theoretical result is very important: the eigenvalues are always between 0 and 1. Recall that in principal components analysis, where the variables are standardized, it's different, because the first eigenvalue is automatically greater than or equal to 1.

What does it mean to have an eigenvalue of 1 in correspondence analysis? This limit case is quite interesting, as we will see. And what does the data look like in this case? Well, it's like this:

We can separate the rows into two blocks, I1 and I2. The columns can also be separated into two blocks, J1 and J2. And suppose that this double division represents exclusivity between blocks, i.e., rows of block I1 are only linked with J1, and not at all with J2. And rows in block I2 are exclusively linked with J2, and not at all with J1. This represents a very strong association, because we have an exclusive link between categories of one variable and those of the other.

What does this look like graphically? We end up with this graph. The first dimension, corresponding to the eigenvalue of 1, perfectly separates the block I1 from I2. i.e., inside I1, no distinctions are made, and we perfectly separate J1 and J2. Inside J1, no distinctions are made.

Slide 25: These curious inertia, let's look at them again in another data set. This data is about recognizing three tastes: sweet, sour and bitter. The experimental design is: for each taste, we have asked ten people to try and recognize the taste of a sample they are given. Here is the little data table of results. Let's read it row by row. The sweet sample was identified as sweet 10 times out of 10, and never mistaken for sour or bitter.

The sour sample was never taken for sweet, but was once mistaken for bitter. Similarly, the bitter sample was never taken for sweet, but three times mistaken for sour. When we now do correspondence analysis on this table, we obtain the first eigenvalue of 1, which is the signal telling us we have a diagonal block structure in the data. What this means is: if we look back at the table, we see that all the non-zero data is found in blocks along the diagonal. For instance, here, sweetness is only ever perceived as sweetness, and no other taste is ever perceived as sweet.

We therefore have this eigenvalue of one. The corresponding plot's first dimension therefore perfectly separates sweet and perceived sweetness, with one point on top of the other, with bitter, perceived bitterness, sour, and perceived sourness all close together on the other side. Now, let's move on to look at the second dimension. This separates bitter and perceived bitterness on one side, from sour and perceived sourness on the other. It basically shows that bitter is usually perceived as bitter, and sour as sour.

And yes, this is exactly what happened. BUT, when we look back at the data table, it's a little confusing. Sour isn't always perceived as such, and nor is bitter. How can we tease this out of the math? Well, quite simply by looking at the eigenvalues. If there had been no perception errors between sour and bitter, the eigenvalue would have been 1. But here it's only 0.375, so, much less than 1. This is the indicator that there has been confusion between sour and bitter. i.e., from time to time, one is perceived as the other.

This is clearly visible in the graphical output. On the first dimension, we have a kind of perfect situation, in that we clearly see the separation between sweet on one side, and sour and bitter on the other, with a much larger inertia in this separation than between sour and bitter. These two are much closer to each other than with sweet. So, the plot clearly shows that there has been much more confusion between sour and bitter than between either of these two and sweet.

Let's now look at another data table in which the confusion between the categories is increased. i.e., this time, sour was only perceived as such 7 times instead of 9. And bitterness was only correctly detected 5 times. Still, sweetness was always correctly detected, and nothing else was perceived as sweetness. So, how does this affect the math? Well, we still find the first eigenvalue of 1, corresponding to the perfect separation between sweetness and the others. But now, for the second dimension, the eigenvalue has dropped from 0.375 to 0.04.

Now, look at the graphical output. It still looks similar to the previous one. The first dimension still perfectly separates sweet from the others. What's changed is that the separation between sour and bitter is much less than

before. This is the way the plot tells us that there has been a lot of confusion between sour and bitter. This is what leads to the small eigenvalue of 0.04 associated with this 2nd dimension.

How about in the most extreme case, where there is no confusion at all, and the data table is diagonal? Well, we'd end up with two eigenvalues equal to one. So what have we learned from the two examples we see here? We see that in both, the second dimension is always showing the same thing: it's separating bitter and perceived bitterness from sour and perceived sourness. So, basically, the second dimension means the same thing in both examples. It's showing that bitter is mostly perceived as bitter, and sour as sour.

But now, from one example to the other, overall it's not the same thing, because, in the first example, we can say that recognition is good, whereas in the second, it's pretty poor. This teaches us that the plot indeed shows the contrast between sour and bitter, but says nothing about the strength of this contrast. It's the eigenvalue that's going to tell us whether the separation is strong or not. If it's equal to 1, the separation is very strong, and in fact represents a total separation, an exclusive relationship.

If on the other hand, the eigenvalue is tiny, like we see in the second example with 0.04, we've struggled to get a majority of the predictions right. Bitterness is only correctly recognized half the time, and sourness 7 times in 10. Really, this means that in correspondence analysis, we should always start by looking at the eigenvalues, because it shows whether the links we find in the data are weak or strong.

The graphical output doesn't show us this, because it doesn't give information on the strength of links, only their type. Here, the link is simply that sour is mostly associated with perceived sourness, and bitter with perceived bitterness.

Slide 26: Let's now go back and have a look at the Nobel prize data. We've already talked about the percentages of inertia of 55 and 25, and how these were large compared to the next dimensions. We therefore decided to keep just these two when moving to the interpretation phase. However, the actual values for the inertia are 0.083 and 0.037, which are quite weak, especially compared with the value 1. We would have had a value of 1 if there had been at least one exclusive association between categories and countries. For example, all the prizes for one category went to only one country, and that country had no prizes in any other category. Clearly, therefore, we are far away from this situation. Indeed, the Nobel prizes for the various categories are well spread out across the countries. If we now look at the sum of the inertias, and knowing that in correspondence analysis, this sum is Phi squared, it can have at most a value of 5, if each of the inertias equals 1. We are very far from this case. Looking at the table again, we are clearly extremely far from having exclusive links between the categories of the two variables.

We've seen that inertias have a specific meaning in correspondence analysis and should be analyzed before interpreting. In the next video we'll see how correspondence analysis allows us to simultaneously look at the rows and the columns on the same plot.

Part 4. Simultaneous representation

(Slides 27 - 33)

Slide 27 (outline)

Now, let's go back to interpreting the graphical output of correspondence analysis. This is so we can talk about another important feature of correspondence analysis. So far, we've commented on the plot in terms of the rows, and in terms of the columns, separately. However, correspondence analysis lets us simultaneously look at the rows and the columns on the same plot. Intuitively, this simultaneous representation is possible because the rows and columns are the same types of object: categories of qualitative variables.

Slide 28: This simultaneous representation of the rows and columns, WORKS, in correspondence analysis, due to what we call transition formulas, or barycentric properties. Here we are looking at the transition formula on the s-th dimension. Let's be clear what the notation means: F_s(i) is the coordinate value for row i on the s-th dimension. G_s(j) is the coordinate value for column j on the s-th dimension. We can thus see how to link the row values with the column values. Let's look at the details. As there is a sum over the j, we know how to express the coordinate value for one of the rows with respect to the coordinate values of all the columns. fij over fi., this is the profile's j-th term. So, we are in the process of summing over the j, i.e., over all the columns, the coordinate values of the columns, weighted by the elements of the i-th profile.

So in a way, we do a kind of mean of the coordinate values of the columns, and in this mean, column j is more dominant the more it has a large conditional probability for the i-th profile.

Once we've calculated these centers of gravity, or barycenters if you like, we spread them out, because the coefficient, 1 over the square root of lambda_s is, by construction, larger than 1. We can write this transition formula in the following way: row i is at the barycenter of all the columns, as each column has been given the weight fij over fi., i.e., the j-th term of the i-th row profile. This barycentric property can be seen intuitively in the following way: a row will be close to the columns with which it is most associated. i.e., if fij over fi. is large, then G_s of j plays an important role in calculating the coordinate value of i.

In correspondence analysis, rows and columns play symmetric roles, so if we permute the roles played by each, we find the second barycentric property: basically, a column is next to the rows with which it associates the most. It is this double barycentric property that lets us use the simultaneous representation we have seen, and therefore help us extract interesting information.

Slide 29: Let's now apply these barycentric properties to a data set we've see before: recognizing the fundamental tastes: sweet, sour, bitter. We're going to look at two data sets, one where there was little confusion between sour and bitter, with only 3+1=4 errors out of the 20, and another where there was more confusion between sour and bitter, with 5+3=8 errors, i.e., a mistake 40% of the time.

So, 20% confusion to the left, 40% to the right. We've already commented on the results of the correspondence analysis of these two tables. The first eigenvalue is 1 in both cases, due to the exclusive link between perceived and actual sweetness. So, what consequences does this have on the barycenters?

Recall for a moment the barycentric property: if lambda_s is 1, this means that one column will be at the exact barycenter of the rows. Take for example perceived sweetness. This is completely mixed up with sweet, because it's only associated with sweet. On the other side of the plot, we have, along the first dimension, bitter, sour, perceived bitterness, and perceived sourness, which all have the same coordinates in the first dimension. Perceived bitterness

is therefore exactly at the barycenter of sweetness, bitterness and sourness, because it is never linked with sweetness. As it's never associated with sweetness, the weight for sweetness is zero in the calculation of the barycenter for the position of perceived bitterness.

Now let's look at the second dimension. This shows clear associations between bitterness and perceived bitterness on one side, and sour and perceived sourness on the other. This is true for both plots we see here. The relative placements of all these points are completely the same to the left and to the right. This shows that it's this relative placement of the points that gives us information on their connections, i.e., which categories are linked. But it says nothing about the strength of the links.

For instance: is the association strong? Very strong? Almost exclusive? If we want to get information on the strength of the link, we have to look at the relevant eigenvalue, which is relatively high in the first case, 0.375, and quite weak in the other, 0.042. The larger eigenvalue tells us that we have a much stronger association in that case. Though, to be sure, we're far from having an exclusive association here, which would have given an eigenvalue of 1.

Slide 30: Let's now take a precise look at how the barycentric property works, by zooming in to the second dimension, to take a closer look at the sourness and bitterness points. For the barycentric property, we must first calculate the barycenter, and afterwards, dilate or stretch the points by a factor of (1 over the square root of lambda_s).

Take for example constructing the perceived bitterness point. To begin with, it's the barycenter of the sour and bitter points with coefficients 1 and 7. So, if we calculate the barycenter of sour and bitter with coefficients 1/8 and 7/8, we land on the red circle, which is much closer to the bitter point than the sour one. Let's do the same calculation for the right hand side. This time we calculate the barycenter of sour and bitter whose coefficients are 3 and 5. We get the red circle, which this time next to bitter, but only slightly. Thus, we see that looking at the barycenter gives us a good idea of the intensity of the link between the two.

What does correspondence analysis do next? It multiplies these values by 1 over the square root of lambda_s. This value, 1 over the square root of lambda_s, is quite different between the two examples, because it equals 1.6 in the first case, and 4.9 in the second. In other words, the point cloud of barycenters will be much more spread out in the right hand side, i.e., when we have a weak link, than when we have a stronger one.

In this way, correspondence analysis neutralizes the effect of the strength of the link, and only keeps the information about the type of link. You might ask: why do we do this? The first reason is obvious: it's that we have exactly the same relationship between a row and the set of columns, and a column and the set of rows. We can clearly see the symmetric role played by the rows and columns. Another way to see this is to say: if we look at the barycenters, when there is a weak link, we're going to have difficulty detecting them.

Therefore, correspondence analysis spreads out the plot, a bit like a microscope, to make associations clearer. In conclusion, we can say that correspondence analysis says nothing about the significance of a link. Instead, the strength of a link is found in the eigenvalues, whereas the type of link is found in the relative positions of the points.

Slide 31: Let's now use the double barycentric property on the Nobel prize data. Recall that the center of gravity of the cloud is at the barycenter of the red points, weighted by the mean profile. We should thus note down each category, the weight of the category that maps the mean profile to the cloud's barycenter.

Now look at Japan for instance. Japan is the profile 26, 0, 9, 13, 4 and 48. It's therefore the barycenter of the red triangles that have been assigned these numbers.

If we use a font size proportional to the weight of a category, we see that Japan has very strong weights for physics and chemistry compared with the size of these weights in the mean profile. Japan will therefore be positioned at the

barycenter of the categories assigned these weights, which pulls Japan towards the bottom of the plot, on the physics and chemistry side. This is indeed the barycentric property, a row point goes towards the columns it is the most associated with, and away from columns it's the least associated with. Note that once the barycenter is calculated, the point is dilated with a coefficient 1 over the square root of lambda_s. This best explains why a barycenter can be outside of its points.

Now, if we look at Italy, we know that it's much more associated with literature than economics, and so clearly, it's point is close to literature. Of course, we know that Italy has large percentages for physics and medicine too, but in reality, these are similar to that of the mean profile. Thus, these two percentages don't really push Italy away from the cloud's barycenter.

Slide 32: Thanks to these barycentric properties, we can now simultaneously interpret the red and blue points, that is, the rows and columns. We've already said that the first dimension separates the scientific prizes on the left from the others to the right. Now, we can also say that the U.S. mostly gets scientific Nobel prizes, while the French and Italians are more likely to get the peace and literature prizes.

The fact that Italy is more towards the edge than France shows that the Italians are even more specialized in the literature category than the French. Quantitatively, they get less literature prizes than the French, but in Italy, the literature prize dominates with respect to the other categories.

As for the second dimension, it pushes physics and chemistry towards the bottom, and we see that it's the Germans and Japanese who obtain these prizes, more than they obtain other prizes.

We've seen how to interpret the simultaneous plot of correspondence analysis. In the next video we'll see some useful interpretation aids.

Part 5. Interpretation aids.

(Slides 33 - 43)

Slide 33 (outline): Let's now move on to classical interpretation aids, including: the quality of the representation, and, contributions. These aids are common to all of the various principal component methods we see in this course.

Slide 34: The quality of representation of a point can be measured using the same criteria we use for a cloud: i.e., the projected inertia of the point, divided by its total inertia. This criterion shows us how much the deviation of the profile from the mean profile is represented by each dimension. If we switch back to the geometric interpretation, we have a point, Mi, representing profile i, which is projected onto the axis Us, as the point His.

To calculate the quality of representation of profile i by the s-th dimension, we thus calculate the projected inertia of Mi on Us, divided by the total inertia of Mi, which gives us fi. times Ohis^2, divided by fi. times Omi^2 . The fi.'s cancel, which means that the quality of representation doesn't depend on the weights. In fact, this ratio is none other than the cosine of the angle between Omi and Us.

This way of looking at things, in terms of the cosine, brings up the following question: how well is the deviation between the profile and the mean profile well-represented by the dimension? Is the point Mi pointing in the same direction as the dimension?

Slide 35: Let's look now at the quality of representation in our toy data set on tastes, in which there was confusion between sour and bitter. Look at the sour point: it has a quality of representation of 0.89 on the first dimension, and 0.11 on the second. This means that it is much further away from the mean in the first dimension than in the second. And clearly, this makes sense: along the first dimension, sourness is never associated with perceived sweetness, while on the second dimension, sourness is occasionally associated with bitterness. Thus, the deviation from the mean profile is much larger on the first dimension than on the second.

In practice, the quality of representation is used in the following way: when we have lots of points, to start interpreting things, we select a few of them that are well-represented, and with fairly extreme coordinates, i.e., far from the origin on the dimension or dimensions we are interested in. This means that we are choosing points which should be helpful for interpretation, because it should be easy to find in these data the meaning of the dimensions, remembering that the deviation between the profile being looked at and the mean profile is essentially expressed with respect to these dimensions.

Slide 36: The second classical interpretation aid, found in all the principal component methods, is the idea of contribution. To calculate the contribution of a point i to the inertia of the s-th dimension, we first calculate a raw indicator, that of the projected inertia of the point, which equals fi. time OHis squared.

This raw indicator is then turned into a percentage by dividing by the total inertia of the s-th dimension, and multiplying by one hundred.

We can also add up the contributions of several elements to a given dimension, so as to measure their total contribution.

What is the point of these contributions?

Well, they indicate how much we can say that a given dimension exists because of one or several elements. If, for example, for a given dimension, we see that one element contributes ninety percent of the inertia, we can simply

limit the interpretation of this dimension to this one element. Which begs the question: what is our definition of a "big" contribution?

Well, we say it's big if both the distance OHis is quite large, AND the mass fi. is quite large too. In this sense, we can say that the idea of what a large contribution is, is a kind of compromise between distance to the origin, and weight.

In practice, contributions are usually used when dealing with big data tables, in order to select a subset of elements to start interpreting. So, we want to select those that contribute the most. Ideally, this means finding the set of elements that both contribute highly and are well-represented.

Slide 37: Let's illustrate this idea on an example. We're not going to use the examples we've seen before, because their row and column marginals are not so different. Instead, we'll use this little data set in which the row marginals are quite varied. For instance, the marginal sums of A and D are more than ten times smaller than those of B and C.

When we do correspondence analysis on this table, we get the following results. The first dimension dominates, with 83 percent of the total inertia, so we're going to stop right there. It separates columns X1 and X4, where we see that X1 was mostly characterized by rows A and B, and X4 by rows C and D. On this dimension, the coordinates of A and D are much more extreme, or further out, than those of B and C.

Now let's look at contributions. When we look at the contributions table, we see that the contributions of B and C are much larger than those of A and D, even though A and D have much more extreme coordinates. This comes from the marginal sums. B and C have, clearly, the smallest coordinate values, but by far the largest weights.

This example thus clearly shows the importance of looking not just at the graphical output of the correspondence analysis, but also at the contributions. But before we go overboard, don't forget that this is quite an extreme example, with huge differences in marginal sums, of a factor of ten or more.

As a broad conclusion therefore, if we have huge differences in marginal sums, it's very important to have a look at contributions. On the other hand, if there is little difference between the marginal sums, looking at contributions won't bring much more information to the table with respect to the plot, as the coordinate values in the plot also represent the contributions quite well.

Slide 38: Like in principal component analysis, it's easy here to bring supplementary elements into the analysis. We can simply apply the transition formula to calculate, for example, the coordinates of an extra column.

Here, in our Nobel example, we've added a column for the Fields medal, which is often considered the equivalent of a Nobel prize in mathematics.

We see in the plot that mathematics is located next to the Nobel prizes for peace and literature. Does this hint at the underlying connections between mathematics and philosophy, perhaps? We also see that mathematics is close to France and Russia, which are two countries with a particularly strong mathematical heritage.

Slide 39: Now, let's move on to talk about a remarkable property of correspondence analysis. It's called distributional equivalence. As we will see, it's extremely useful when analyzing textual data. If two rows (or two columns) have exactly the same profile, we may ask the question: is it useful to group them together as one? If we do this for two rows, for example, we then replace them with one row in which we have summed the two rows together. The next question would be: which table should we analyze? Should we do correspondence analysis on the first table or the modified one? Well, the distributional equivalence property of these tables is that both analyses lead to the same result!

This is extremely important in the analysis of lexical tables, and calms the debate about whether or not we should group together singular and plural versions of words, verb conjugations, synonyms, etc. Due to their equivalent distributions, we know that these words will all have the same profile. It doesn't matter if we group them together or not. And of course, if they DON'T have the same profile, they SHOULDN'T be grouped together, because they must be representing different notions.

Slide 40: We finish this presentation of correspondence analysis by thinking about how many dimensions with non-zero inertia there are. Let's consider this for a point cloud of rows, with I points, in a J-dimensional space. As there are J dimensions, we might guess that we can get J axes with non-zero inertia. In fact, not at all! Remember that we are analyzing a profile, which has a specific feature: its sum is 1. What this means is that the cloud exists, in reality, in a J minus 1 dimensional sub-space. Therefore, the maximum number of axes with non-zero inertia is less than or equal to J-1. Next, let's think a bit more about the number of points. There are I points. So, if I equals 2, we can represent them using just one dimension. In the general case, we can perfectly represent I points with at most I-1 dimensions.

Overall, where does this lead us? Well, the maximum number of dimensions with non-zero inertia in correspondence analysis has to be less than or equal to the minimum of these two numbers: i.e., the number of columns minus 1, and the number of rows minus 1. So, for the Nobel prize example, as we have a table with 8 rows and 6 columns, we can be absolutely sure that with five dimensions, we will have found 100 percent of the inertia.

So, what does this mean for Phi^2, which is equal to the sum of the eigenvalues, and remembering that these are all less than or equal to one? Well, it means that Phi^2 must be less than or equal to the number of dimensions with non-zero inertia. i.e., less than or equal to the smallest of I-1 and J-1.

So, this gives us the idea of taking the ratio of the value of Phi^2 over its maximum possible value. This gives us what is called Cramer's V, a measure that is between 0 and 1. It captures the intensity of the link between pairs of qualitative variables. We therefore have two measures of link intensity: the Phi^2, and Cramer's V. They are quite similar measures, with the advantage for Cramer's V that it's always found between 0 and 1.

Let's calculate Cramer's V for the Nobel prize data set. We get 0.03. This is quite small, which tells us that we are far from the situation of a strong exclusive link between pairs of categories: one country, one prize category. So, what about if we calculate Cramer's V for the tables with the three flavors? Here we get 0.69 and 0.52, which are much higher. This means that there's a fairly strong relationship between the two variables.

Slide 41: So, what is the overall conclusion we can make about the Nobel prize data? Well first, it shows that even on small data sets, correspondence analysis can give us a nice visual representation of the deviation from independence, which can help us decrypt the table better. This will be even more true for larger data tables. In the Nobel prize data, we have seen that essentially, the deviation from independence is structured as a separation between the scientific prizes and the rest, and, to a smaller degree, between physics and chemistry,....., and the economic sciences. The position of each country in the plot shows their relationships with the prizes. The United Kingdom, right in the middle, has not specialized in any particular prize, and wins them in the proportion to which they have been given out overall. France and Italy, on the other hand, specialize in peace and literature prizes, while Japan and Germany tend to get them in physics and chemistry, and the US and Canada in economics.

Slide 42: So, the overall conclusion is this: to study possible links between qualitative variables, we build a contingency table, the foundation. If there are links, they are to be found in the deviation between what the real data table looks like, and what it would look like if there were independence. Correspondence analysis means constructing a point cloud for the rows, and a point cloud for the columns. The total inertia of a point cloud measures the strength of the deviation from independence. This total inertia is then decomposed into a sequence of dimensions of decreasing importance, each representing some aspect of the link between the variables. Lastly,

correspondence analysis provides us with a simultaneous representation of the rows and columns, in which a point's position reflects its participation in the deviation from independence.

Slide 43: If you want to go further into the details of correspondence analysis, we invite you to delve into the book, which works though the subject in a similar way.

You've now arrived at the end of the course videos for correspondence analysis. Next, go and watch the video on how to use correspondence analysis in practice, using FactoMineR, and have a go at the suggested exercises.