Audio transcription of the
Principal Component Analysis course

Part I. Data - Practicalities

(Slides 1 to 8)

Slide 1:

This week, we have for you three videos which together present the main details of principal component analysis.

Principal component analysis is a set of tools which allow us to study and visualize large data sets. We will present the method from a theoretical as well as a practical point of view.

Slide 1 bis:

The outline of this week's work is as follows: We will first define the types of data we can use principal component analysis on; then we'll look at examples of situations in which principal component analysis can be performed. Then we'll define some useful notation.

Next, we'll focus on the individuals, then on the variables. At the end, we will spend some time looking at how to interpret the results, and provide interpretation aids.

Slide 2:

So, what kind of data are we looking at?

Principal component analysis, also known as PCA, applies to data tables where rows can be considered like individuals and columns like quantitative variables. Let x-i-k be the value taken by individual-i for variable k.

i varies from 1 to capital I, the number of individuals, and k from 1 to capital K, the number of variables. Overline x-k is the mean of variable k calculated over all individuals, and s-k the standard deviation of the sample for variable k. Here we use 1 over I, which means that the standard deviation is calculated on the data, and we don't try to estimate the standard deviation of the population.

Slide 3: Data tables, with individuals in rows and variables in columns, can be found in many different areas, which means that we can perform PCA on quite a diverse range of data sets.

Here, I have listed a few applications with examples. Here is a first example involving sensory analysis, where products are described by a set of variables often called sensory attributes, like acidity, bitterness, sweetness, and so on. Thus, we have a table with various products, for example different wines, and each is going to have a score for acidity, bitterness, sweetness, etc.

So, the aim of PCA is to study this data table.

Here is an example from ecology, where "individuals" are rivers, and variables are different pollutants; so we end up with a data table with rivers in the rows, and pollutants in the columns.

In Economics, we might have years in the rows, and economic indicators in the columns. We can then track the evolution of economic indicators through the years. Instead of years for rows, it could be countries, if we want to compare the economic situation of several countries.

Often in genetics, patients are represented in terms of their genes. This kind of data set can be large, since there are lots of genes.

In marketing, we could have a set of brands, and several measures of satisfaction.

And in sociology, we could have different social classes and different activities, with the table entries being the average time spent by individuals of each social class at a given activity.

Clearly, tables like these, with individuals in rows and variables in columns, are found across many fields.

Slide 4:

We are going to work with a running example throughout this course, involving wine. In the rows, we have ten wines, and in the columns, twenty-seven quantitative variables. These are sensory attributes, like sweetness, bitterness, fruity odor, and so on. The values in the data table correspond to the average score given by several judges for the same wine and descriptive variable. So, for example, the wine S Michaud has a mean score of 4.3 for the "fruity" odor. I.e., this is the average over all judges. And so on for all wines and all variables. In our data set, we also have two quantitative variables that correspond to preference: preference in terms of odor, and overall preference. We also have a qualitative variable representing different wine labels. There are 2 labels here: the first is Sauvignon and the second is Vouvray. We will see later how we can consider this information in the analysis. The aim of doing PCA here is to characterize the wines according to their sensory characteristics. So, first we will focus on the twenty-seven sensory characteristics for characterizing the wines.

How can we study this data table?

Slide 5:

This data table can be studied in different ways: we can see it as a set of rows, and try to look for differences between rows. We can also look at the data table as a set of columns, and investigate the similarities or links between columns. To study individuals, that is, the rows, we need to know when two individuals are "close" and when they are "different", from the point of view of all the variables. If there are many individuals, which are the most similar (and the most dissimilar)? Are there any groups of individuals which are homogeneous in terms of their similarity? In addition, we may want to look for axes of variability which can separate extreme individuals from more normal ones. In our example, two wines are considered similar if they are evaluated similarly for all sensory characteristics. In practice, this means that the two wines will consistently vary in the same direction with respect to the mean for many characteristics, and can thus be said to have the same sensory "profile". More generally, we may want to know whether or not there are groups of wines with similar profiles, that is, similar sensory profiles which might separate extreme wines from more average ones.

Slide 6:

Following the approach taken to study individuals, might it also be possible to interpret the data in terms of the variables? For instance, which variables provide similar information to each other?

Between variables, rather than similarity, we in fact talk about "relationships". And the most well-studied relationships between variables are linear. Indeed, PCA focuses on linear relationships between variables. More complex connections also exist, such as quadratic, logarithmic and exponential ones, but these are not studied in PCA. This may seem restrictive, but in practice many relationships can be considered linear, at least as a first approximation.

In exactly the same way as for individuals, creating groups of variables may provide useful information. When we have only a very small number of variables, it's possible to draw conclusions from the correlation matrix. This matrix holds all of the linear correlation coefficients between pairs of variables. However, when working with a large number of variables, the correlation matrix is huge, and it's therefore essential to have a tool capable of summarizing the most important relationships between variables, in a visual way. The aim of PCA is to draw conclusions from the linear relationships between variables by detecting the main directions of variability. As we will see, these conclusions can be enriched using the synthetic variables generated by PCA. It then becomes easier to characterize the data using a small number of synthetic variables, rather than all the original ones. In our example, the correlation matrix brings together the 351 correlation coefficients! So, trying to group variables using the correlation matrix would be rather tedious.

The average itself is a synthetic variable, but this is defined *a priori*. Here, we want to define indices from the data, so, *a posteriori*.

Slide 7:

Obviously, since we are working with the same data table, looked at from either the point of view of rows or columns, there exists a link between the two. Therefore, simultaneously studying individuals and variables will only improve their respective interpretations.

When studying individuals, we can build groups of individuals, but then we want to characterize these groups, and to do this, we would like to use the variables. For instance, we would like to be able to say that certain products are similar because they are acidic and bitter, whereas others are similar because they are sweet. So for this, we need an automated method, especially if we have a lot of variables.

Similarly, when there are groups of variables, it may not be easy to interpret the relationships between large numbers of variables in a group, so we could think about making use of specific individuals, that is, individuals who are "extreme" with respect to these relationships. For example, the connection between height and weight could be illustrated by the contrast between two extreme individuals: a dwarf and a giant.

In summary, PCA aims to delve into and decrypt data. In PCA, we can visualize the data with simple plots, which give nice summaries of the data. Essentially, we take the information contained in a data table and view it graphically.

Slide 8:

Let's now start to get into the details. We said earlier that we can look at our data table as a set of rows or a set of columns. If we see it as a set of rows, we aim studying individuals, if we see it as a set of columns, we aim studying variables. Studying individuals can be seen as considering a cloud of points, where each point corresponds to an individual. This point cloud lives in a space with many dimensions. If there are K variables, it lives in a K-dimensional space.

When studying variables, each is a point in an I-dimensional space, i.e., there are I coordinates for each variable. Therefore, I will see a cloud of variables in R^I.

We have seen what data tables look like in PCA applications, and what kinds of questions we can ask with PCA. In the next videos, we'll see how to put PCA into practice.

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

Part II. Studying individuals and variables

(Slides 9 to 26)

Slide 9 (outline):

Now that we've defined the types of data we can do PCA on, and what kinds of questions we can examine with PCA, let's now see how to put PCA into practice.

Slide 9 bis:

We said that an individual is a row of the data table, and therefore a point in K-dimensional space. If K is equal to 1, there is only 1 dimension, and it is easy to represent the individuals: I put them on a one-dimensional line, according to their value.

If there are two variables, I can put them on a standard scatter plot, like we can do before linear regression.

If there are three variables, it becomes a little more difficult, but there are softwares that can show data sets in 3 dimensions. The software can help visualize a point cloud by pivoting the axes in various directions, thus giving us a better feel for where the point cloud is.

But if there are 4 variables or more than 4, it is impossible to represent a point cloud in this many dimensions. However, the mathematical concept behind it is simple and stays the same: there are K coordinates for K dimensions.

So, how are we going to be able to "look at" the cloud of individuals if it exists in a very high-dimensional space? Well, we can start by defining a notion of similarity between pairs of individuals. When can we say that 2 individuals are similar? How about: two individuals are similar if they are "close"? If so, we can calculate the squared distance between individuals, thanks to Pythagoras, by the sum of the squared differences between each variable for the two individuals.

This is a simple definition of similarity, so finally, when I say I want to study the cloud of individuals, this just means studying the shape of this cloud, to see which individuals are close to each other, or far apart.

Slide 10:

The data table can be studied geometrically via the distances between individuals. And this turns out to mean studying the shape of the point cloud. We can see in the photograph a flock of starlings. The "cloud" of birds exists in 3 dimensions, but the photo allows us to visualize it in a lower-dimensional space, here, 2 dimensions. This 2-d representation nevertheless gives us a good idea of the cloud's shape in the original 3-d space, and thus a good idea of distances between birds. The moral of the story is: when our individuals exist in a high-dimensional space, we will try to study the shape of the cloud and visualize it in 2 dimensions.

Slide 11:

To study the cloud's shape, we first have to mention centering and standardization.

Centering data means translating the cloud of points so that its center of gravity ends up at the origin. The cloud's shape remains the same when translated. Centered data gives technical advantages, and is always done in PCA.

Slide 11 bis:

Next, we can consider whether or not to standardize the data. In the graph on the right, we have a representation of individuals in terms of size and weight. The size is in centimeters and the weight in quintals. Note that the data are centered here. We can see that the point cloud is very elongated and vertical.

In the middle graph, the size is in meters and the weight in kilograms. The shape of the cloud is quite different from the first one. We now have a very elongated horizontal point cloud.

Slide 11 ter:

And as we said, we want to study the shape of the point cloud. But depending on the units chosen (quintal and centimeters, or meters and kilograms), the cloud's shape can be very different. How can we handle this? One idea is to standardize the data. If we standardize our variables, they then have no units, and are therefore comparable. This type of data standardization is essential if measurement units are different from one variable to the next.

If the variables are already in the same units, we can choose to standardized or not. We will see later than standardizing leads to giving the same importance to each variable. Without it, the bigger the variance, the more important a variable is. It is not obvious in this case whether to standardize or not.

In our example in the course, we will standardize the dataset.

Slide 12:

The standardized data set already provides some useful information: for instance, we see that the Buisse Domaine wine has a very low score for bitterness and aroma intensity. Values below -2 are below the average for that variable, and indeed quite extreme. We know that when values are from a normal distribution, they are between -1.96 and 1.96, 95% of the time. Here it is not necessarily the case that the data follow a normal distribution, but a standardized value below -2 is still very extreme. So the Buisse Domaine wine is especially "not bitter", with very low aroma intensity. The wine V Aub Silex is rather sweet and non-acidic.

Our method, Principal Component Analysis, will then be used to analyze this standardized data table. This means that we want to "visualize" it. It is difficult to "visualize" this table, since there are twenty-seven variables. As it is not humanly possible to visualize 27 dimensions, we aim to get an "approximate" look at the cloud of wines, from a "good" viewpoint, in a low-dimensional space.

Slide 13:

For such a representation to be successful, it must select a "good" viewpoint. In fact, PCA turns out to mean searching for the best "smaller" space (lower-dimensional) that gives an optimal view of the K-dimensional cloud's shape. We often work in a plane, though this can prove inadequate when dealing with particularly complex data.

So, what is a "good" viewpoint? it's one that lets us get an idea of the distance between individuals, in a way that it distorts the higher-dimensional "truth" as little as possible. i.e., the viewpoint should maintain the point cloud's shape as closely as possible.

Here is a 3-d animation to help us think about the question. Here is a point cloud in 3 dimensions. I can move the point cloud to see it better. This gives us an idea of what the point cloud really looks like in 3 dimensions.

Then, it's possible to only have a 2-d representation of the 3-d point cloud. This is a first suggestion for a 2-d viewpoint. Here is a second. And a third.

Slide 13 bis:

And so the question is: among these three possibilities, which one should you choose to get the best understanding of the point cloud? We are only allowed to have a single photograph of the 3D cloud, which one should we choose? Well, I would intuitively choose the third one. Why? Because it spreads out the points, it seems to better "see" the distances between individuals. And this intuition is good. So viewpoint quality means the quality of reproduction of the cloud's shape: a viewpoint is good if we can see the diversity and variability in the data, and is especially good if it does not overly distort distances between individuals. The true distances are calculated between individuals in a high-dimensional space, and now we have put them into a 2-d space. This modifies the distances between individuals, and we want these distances to be as little modified as possible.

Slide 14:

So, how can we quantify a viewpoint's quality? The answer is: with the concept of dispersion. A point cloud can be best seen if its projection onto a 2-d plane is "spread out", if we can see a lot of variability. This variability exists across several dimensions, and is called inertia. Thus, we will talk about inertia from now on. Inertia is simply a generalization of variance to several dimensions.

Slide 15:

Here is another small illustration, again a visualization of a 3D cloud. Here I have a picture of an animal that "lives" in a 3-d space, but I have a photo of it that's only in 2d. So, which is the photograph that best reproduces the shape of my "cloud" in the original space? It would seem to be the one on the right. I can see the camel the best because I can see more information, and in particular, 2 humps and 4 legs.

Slide 16:

If we go deeper into the details, how can we find the best approximate viewpoint of the cloud? Well, we begin by finding the first component, also known as the "first axis", that distorts the cloud as little as possible. Let us denote Hi the projection of individual i onto an axis, and O the center of gravity of the cloud. Then, iHi^2 is the (squared) distance between the individual in the cloud and its projection onto the axis. And what we want is that the (squared) distance between an individual and its projection be as small as possible. Since the distance Oi between an individual and the origin is always the same, iHi is small if OHi is large (thanks to Pythagoras). So, since Oi^2 is constant, iHi^2 is small if OHi^2 is large. So we want OHi^2 to be as large as possible. Since we want each OHi^2 to be as large as possible, we really want the sum of the OHi^2 to be as large as possible. And the sum of the OHi^2 is none other than the dispersion of the projected points. i.e., we want these projected points to be as spread out as possible. This is how we find the first PCA axis, also named the first PCA dimension!

Let's now consider what it means to find the best viewpoint of the cloud in 2 dimensions. Now, the cloud is projected onto a plane, also chosen to minimize distortion of the point cloud. The plane is selected so that the distances between the projected points on the plane Hi (Hi is now the projection onto the plane) are as close as possible to the distances between the initial points. Since in projections, distances can only decrease, we can try to make the projected distances as large as possible, so once again, we want the sum of the OHi^2 to be as large as possible.

What is quite nice, is that the best plane (in 2d) contains the best component (in 1d). So to find the best plane (2d), you don't have to start from scratch. You can simply find the first PCA dimension, then look only at orthogonal directions to the first one, and find that which maximizes the sum of the OHi^2. In the same way, once we have found the second dimension, we can find the 3rd, and so on, sequentially. So, each new axis is orthogonal to all previous ones, and maximizes the inertia.

From a technical point of view, PCA axes are obtained by a singular value decomposition, or by diagonalizing the correlation matrix to extract the eigenvectors and eigenvalues. The eigenvectors are vectors u_s, associated with the eigenvalue of rank s. The eigenvalues are ranked in decreasing order. The eigenvalue lambda_s can be interpreted as the inertia of the cloud projected onto the s^th component. In other words, lambda_s corresponds to the "explained variance" for the component of rank s.

Slide 16 bis: We can visualize this with a small illustration.

Here we have a teapot in 3 dimensions, and we want to find the best direction in which to visualize it from just one viewpoint. Well, to start, we find the first axis. Here is the one that allows us to best see the teapot. So if we had to project the teapot's points into one dimension, we would choose the red dimension. And now we want a second axis, orthogonal to the first. We have defined the first, so now we find such an orthogonal axis to it. This gives us the second axis, in green.

Slide 17:

Let's now look at the wine dataset, with the 10 wines and 27 sensory attributes, 2 preference variables, and the qualitative variable corresponding to the wine labels. We want to characterize the wines from a sensorial point of view, so first we just consider the 27 quantitative sensory variables.

Slide 18:

Well, if we plot the individuals using PCA, we get this graph. It shows, for instance, that S Michaud and S Trotignon are very "close". What does it mean when I say "S Michaud and S Trotignon are very close "? It really means that the scores for S Michaud and S Trotignon are approximately the same, whatever the variable. The scores are almost the same, so the points are close. In the same way, Aub Marigny and Font Coteaux are wines with similar sensory scores for the 27 attributes.

On the other hand, Font Brulés and S Trotignon have very different sensory profiles, because the first principal component, representing the main axis of variability between wines, separates them strongly. Since these wines are far apart in the first dimension, they must be very different, because the first dimension is that which separates the points as much as possible. So then, why are these two wines so far apart in the first dimension? To answer this, it might help to be a French wine expert. I, personally, am not. So we have to use the variables we have, to interpret the dimensions.

Slide 19:

What are the differences between Font Brulés and S Trotignon? To answer this, we can look at the individual's coordinates in each dimension. For instance, Aub Silex has a value of 1.1 in the first dimension. Let's call this value Fi1 (1 for the 1st dimension). Similarly, we have Fi2 in the second dimension, which equals -6. So, for each individual i, we consider its values on the horizontal and vertical axes. Thus, we can create two vectors, one which gives the values for all individuals in the first dimension, the other that does the same in the second dimension. Therefore, these vectors have I entries, the same number as there are individuals.

Slide 20:

Then, in order to interpret the plot for the individuals, we will calculate the correlation between each variable, e.g. vanilla odor, and the first dimension. So we do this, and then, calculate the correlation between vanilla odor and the 2nd dimension. If this odor is related to the 1st dimension, it means that its values are related to the coordinates in this dimension. Thus, if the correlation is close to 1, it means that individuals who have small values for vanilla odor have small values in the 1st dimension. And thus, such wines will be to the left of the plot. And individuals who have large values for vanilla odor will have large values in the 1st dimension, and will be found to the right. If the correlation is negative, it means that individuals with small values in the 1st dimension have high values for vanilla odor. And that individuals with large values in the 1st dimension have small values for vanilla odor.

The same logic applies in the 2nd dimension. Using this information, we can build a graph that represents all the variables. They will all be found inside a circle, known as the correlation circle.

Slide 21:

In our example, the graph of the correlation circle shows that:

The variables astringency, visual intensity, mushroom odor and candied fruit odor, found to the right, have correlations close to 1 with the first dimension. Since the correlation with the 1st dimension is close to 1, the values of these variables move in the same direction as the coordinates in the 1st dimensions. Wines with a small value in the 1st dimension have low values for these variables, and wines with large values in the 1st dimension have high values for these variables. Thus, the wines that are to the right of the plot have high (and positive) values in the 1st dimension and thus have high values for these variables. With the same logic, wines that are to the left have a small value in the 1st dimension, and thus low values for these variables.

For the variables passionfruit odor, citrus odor and freshness, everything is the other way around. The correlation with the 1st dimension is close to -1, and thus the values move in the opposite direction. Wines with a low value in the 1st dimension have low coordinate values, and thus have high values for these variables, and wines with large values in the 1st dimension have small values for these variables.

Overall, we see that the first dimension splits apart wines that are considered fruity and flowery (on the left) from wines that are woody or with vegetal odors. And this is the main source of variability.

So then, how can we interpret the 2nd dimension, the vertical axis? At the top, wines have large values on the vertical axis. Since the correlation coefficients between the 2nd dimension and variables such acidity or bitterness are close to 1, it means that wines at the top take large values for these variables. And wines at the bottom have small values in the 2nd dimension, and thus small values for these variables. For sweetness, the correlation coefficient is close to -1, so wines that have a small value in the 2nd dimension are sweet, while wines that have large values are not.

Overall, the 2nd dimension separates the wines at the top, acidic and bitter, from sweet wines at the bottom.

Slide 21bis:

So, for our data, the main direction of variation involves odor, and the second separates sweet wines from acidic and bitter ones.

Slide 22:

Let's now take a closer look at the variables.

Slide 22bis:

Well, a variable is a point in an I-dimensional space, since each variable has I values (one per individual). So we can consider the variables' cloud that "lives" in an I-dimensional space. Variables are represented by arrows in principal component analysis. So, variable k is an arrow from the origin. The cosine of the angle theta(kl) between the arrows for variables l and k is equal to the scalar product of the variable k with variable l, divided by the norm of variable k times the norm of variable l. This is equal to the sum over all i of Xik times Xil, divided by the square root of the sum of the squares of the Xik times the sum of the squares of the Xil.

Slide 22ter:

As the data matrix X is centered, we can actually look at this formula as the sum of the Xik minus the mean of Xk times Xil minus the mean of Xl, divided by the standard deviation of the variable k times the standard deviation of the variable l. Which means that when the data are centered, we see that we are looking at the well-known correlation coefficient between two variables. So we have here a geometric representation of the correlation coefficient between two variables, k and l. The geometric representation of the correlation is none other than the cosine of the angle between variables k and l.

Slide 22 quater:

If the variables have been standardized, the lengths of arrows in this space equal 1, and all arrows will have their tip on the surface of an I-sphere of radius 1. So, for each variable, we have an arrow from the center of the I-sphere that reaches the I-sphere of radius 1.

Slide 23:

The question now is: how can we actually "see" the variables' cloud? It's in an l-dimensional space, and humans can't visualize that easily when l is 4 or more.

Well, what we can do is look for directions from which we can see the variables' cloud as well as possible. So it's the same trick as for individuals: we look for orthogonal axes that best represent the variables. The first dimension, the one that lets us see the variables the best, will maximize the sum of the correlations between that dimension's direction, and each variable. So the best axis, V1, is that which is the most related to all the variables, in terms of the squared correlation coefficient. So, essentially, V1 is like a synthetic variable that best summarizes all variables. Once we have the first axis, we can find a second one, orthogonal to the first, which is the next-most linked to the variables. This second dimension contains information not found in the first. Once we've found it, we can sequentially find the 3rd dimension, and so on.

So what happens when we fit the variables for our dataset in this way? We find:

Slide 24:

The same representation as before! Earlier, we had built a representation of the variables to help us interpret the individuals' point cloud. And then, when we build the best representation to see the variables' cloud, we find the same thing.

Slide 24 bis:

So that's a little magic! Amazing!

Slide 24 ter:

So, in summary, this representation helps us interpret the individuals' point cloud, to characterize individuals. In our example, this means dividing them into fruity and flowery wines, sweet wines, acidic and bitter wines. It is also, as we have seen here, an optimal representation of the variables' cloud. And we also get to visualize all the correlations between pairs of variables by the cosine of the angle between them. So this representation is also a way to visualize the correlation matrix. What is magical is that it's the same thing!

Slide 25:

The relationships between the individuals' and variables' point clouds are called "duality relationships". This term refers to the dual approach to the one single data table, by considering it either by row or column. This approach also involves ``transition formulas" that let us calculate the coordinates in one space from those in the other. Let us note Fis the coordinates of individual i and Gks the coordinates of variable k in the rank s component or dimension. The vectors F.s are called the scores, and the vectors G.s, divided by the square root of the s-th eigenvalue, are called the loadings.

The transition formulas are the two following relationships: give an individual's coordinates using the data and the variables' coordinates; and second: give a variable's coordinates using the data and the individuals' coordinates.

This result is very important for data interpretation, and makes PCA a rich and reliable experimental tool. What we find is: individuals are found on the same side as their variables with high values, and opposite their variables with low values. Remember also that the x_ik are centered and have both positive and negative values. This is one reason why individuals can be so far from variables for which they have small values.

Slide 25 bis:

For instance, the variable sweetness has a low value in the 2nd dimension, and the value of Aub Silex for sweetness is high. Thus, the multiplication gives a negative value, which contributes to the fact that the value of Aub Silex is low in the 2nd dimension. The value for acidity is high in the 2nd dimension, and Aub Silex has low acidity -- less than the mean -- so the positive value for acidity, multiplied by this negative value for Aub Silex, gives a negative value. This contributes to the fact that the value for Aub Silex is low in the 2nd dimension.

It is exactly the same reasoning for the variables.

Slide 26:

Now, let's go back to the graphical view of the variables. We've seen that the correlation coefficient between variable A and variable B is equal to the cosine of the angle between them. However, this is in the high-dimensional space. And we are visualizing it in the projected 2-d space.

So here is an illustration that allows us to see the scatter plot of the variables (on the left) and their projections (on the right). For example, variables D and E point very close to the projection plane. The projection of D, called H_D, is very close to D because the D is close to the projection plane. This is the same for H_E, which is close to E. So the angle between D and E is very close to the angle between H_D and H_E. And therefore, the angle in the projection plane can indeed be used to visualize the correlation between D and E. More precisely, the cosine of the angle H_D H_E is almost the same as cosine of the angle between E and D, because the variables are well projected.

On the contrary, when variables are not well-projected, which is the case for example for A and B, the angle in the projection plane is small while that in the original space is very large. The arrow A takes a large value in the 3rd dimension, while B takes a small one. So these two arrows are approximately projected to the same place, but poorly. Therefore, the cosine of the angle here is nowhere near the cosine of the angle in the original space, and therefore we can't really get the correlation coefficient between A and B. The cosine of the angle we see is nowhere near the correlation coefficient. Our conclusion is: you can get a feel for the correlation coefficients ONLY for well-projected variables. i.e., variables pointing close to the correlation circle. Indeed, what we have in general is an I-sphere of radius 1 cut by a plane. Thus, the I-sphere is projected onto a circle of radius 1. If two arrows are close to the edge of the circle, their variables are well represented, and we can visualize the angle between them in the global space. On the contrary, it is impossible, with only the projection, to have an idea of the correlation coefficient between two poorly represented variables, like A and B in our example. So, basically, we can only really interpret well-represented variables.

We have seen how to construct plots of individuals and variables, we will see in the next video some useful interpretation aids, which will be of great use for finishing up a PCA analysis.

Part 3. Aids for interpretation

(Slides 27 to 36)

Slide 27:

In the earlier video, we have seen how to construct plots of individuals and variables. Let's now take a look at some useful interpretation aids, which will be of great use for finishing up a PCA analysis.

Slide 27 bis:

We have seen that what helps interpretation is the quality of the projection. This can be measured by the percentage of inertia explained by the 1d or 2d mapping. This percentage can be seen as the percentage of the overall information contained in the 1d or 2d mapping. We can plot a bar chart with the percentage of variance, or percentage of inertia, explained by each dimension. So here we see that the percentage of explained variance in the first dimension is approximately 43%, and around 25% in the 2nd. Together, the two dimensions explain 68 % of the information contained in the data set.

Back to slide 25:

These percentages are displayed on the graphs of individuals and variables. This is what is written here: 43% of the information on the first dimension, 25% on the 2nd.

Slide 27:

So, initially there were 27 variables, and we summarized them in two dimensions, containing 68% of the information. So we get a fairly good summary of them. But then in the next dimensions, there is less and less information. This bar chart is useful for choosing the number of dimensions that should be used for interpretation. We look for a visible jump or gap in the bar chart. There are also tests for selecting the number of dimensions, or even cross-validation can be used. The function estim_ncp performs cross-validation.

We can also see PCA as a denoising method that separates the signal, found in the first dimensions, from noise in later dimensions. This is a bit beyond the scope of the course, but PCA can be used for preprocessing before clustering, for instance.

Slide 28:

Now, here is a table which gives us the 95% quantile of the percentage of inertia found in the first two PCA dimensions, when we have independent variables. So, for example, with 15 individuals and 12 variables, the value 47.8 tells us that having performed 10 000 PCA with 15 individuals and 12 independent variables, we have calculated the percentage of variance found in the first two dimensions, and 95% of these values are less than 47.8.

Slide 29:

Here is another table where the number of variables is greater. With 10 individuals and 27 variables, the 95% quantile of the percentage of inertia is approximately 43%. In our example, we had a value of 68%, which is higher than 43%. So, our plane explains more information than a PCA plane would have with independent variables. This means that our summary is very concise, and that we don't only have independent variables. There must be strong relationships in our dataset.

Slide 30:

Here are some more aids for interpreting results if we have access to supplementary variables or individuals. What we call "active elements" are individuals and variables used to build the principal components. So "supplementary elements" or "illustrative elements" are extra individuals or variables, that have not been used to build the principal components. They can be used instead to illustrate or further interpret the PCA. Illustrative elements can be quantitative variables, qualitative variables, or individuals.

By definition, supplementary quantitative variables play no role in calculating distances between individuals. They can be represented in the same way as active variables, to hopefully assist in interpreting the individuals' point cloud. The value of a supplementary variable in a given dimension corresponds to the correlation coefficient between the variable and the principal component. Supplementary variables can therefore be represented on the same plot as active variables.

In our example, we see that the overall appreciation variables are not strongly linked to the sensory attributes, since they are not well-projected onto the first two dimensions.

For supplementary qualitative variables, we can project the centre of gravity of individuals from each category. For the "label" variable, with categories Sauvignon and Vouvray, we put Sauvignon at the centre of gravity of the 5 Sauvignon wines, and Vouvray at the centre of gravity of the 5 Vouvray wines. Note that the qualitative variable is represented on the plot of the individuals, not on that of the variables. If we have several qualitative variables, we can calculate the centres of gravity for each category of each qualitative variable.

Slide 31:

Another indicator useful for interpretation is the quality of the representation. We can look at the quality of representation of a variable or individual. The quality of representation of a variable is measured by the squared cosine of the angle between it and its projection onto the plane or axis. Similarly, the quality of representation of an individual is the squared cosine of the angle between the vector from the cloud's center of gravity to the point i corresponding to the individual, and the vector from the cloud's center of gravity to the projected point Hi.

The cosine is calculated dimension by dimension, so we see for example that for Renaudie, the $\cos^2$ with the first dimension is 0.73, and 0.15 with the second. Thus, to define the quality of projection onto the plane, we sum 0.73 + 0.15 = 0.88. This means that the angle is close to 0, and therefore this individual is well-projected. We can thus allow ourselves to interpret the distance between this individual and other well-projected individual. However, if 2 individuals are close to each other on the plane, but poorly projected, we should not be quick to interpret their proximity, because they may be far apart in the other dimensions.

As for the variables, passionfruit odor is well-projected, whereas fruity odor is poorly projected onto the first two dimensions.

Slide 32:

Another thing we can look at is a variable's or individual's contribution to the construction of a given dimension. A variable's contribution is just the square of the correlation coefficient between it and that dimension, divided by the sum of squared correlations between all variables and that dimension. This will tell us whether there is some variable that contributes significantly to that dimension with respect to the others.

Next, individuals. An individual's contribution is the square of its coordinate value on the dimension of interest, divided by the sum of the squares of the coordinate values of all individuals. These contributions are expressed as percentages. Again, if some individual contributes significantly to a dimension, it means that this dimension is probably mainly due to that individual. And, without this individual, the dimension would be quite different. If this happens, we can state that this individual is important, and try to explain why. Then, it could be interesting to perform the analysis again without them, to see if and how much the dimensions change, to see if the first dimension of variation is still the same, or if it is a bit different without this individual.

Slide 33:

Next, here are some tools that can help interpretation, especially when there are lots of variables. The components provided by PCA can be automatically characterized using all of the variables, whether quantitative or categorical, supplementary or active. For quantitative variables, the principle is the same whether they are active or supplementary. First, the correlation coefficients between the coordinates of the individuals on component s and each variable are calculated. We then sort the variables in descending order, from highest to lowest coefficients, and keep variables with the highest correlation coefficients (in absolute value).

In our example, for the first dimension, we see that the most correlated variables are candied fruit odor, and Grade, with correlations close to 1. These therefore best describe the 1st dimension. There are other variables which are highly positively correlated with the first dimension, and others that are highly negatively correlated, with coefficients close to -1. For example, freshness, passionfruit odor, etc.

So clearly, there are many variables that characterize the first dimension. There are however fewer that characterize the second. This is expected, since the second dimension is less related to the variables than the first, by construction, since the first is the main direction of variation. We see that odor intensities are strongly positively related to the second dimension, and sweetness is fairly strongly negatively related.

Slide 34:

For categorical variables, we do one-way analysis of variance, i.e, one-way ANOVA, and seek to characterize the coordinates of individuals on component s in terms of the categorical variable. We use the constraint that the sum of the coefficient equals 0. Then, for each categorical variable, an F-test is performed to see if it is related to the given dimension. Second, a Student t-test is performed, to compare the average of the individuals who have that category, with the general average. We test whether the coefficient equals 0, supposing that the variances of the coordinates are equal for each category. The one-way ANOVA coefficients are sorted, according to p-values, in descending order for the positive coefficients, and ascending order for the negative ones.

For our variable "label", the squared correlation ratio is 0.87, which means that 87% of the variance of the coordinates of individuals in the first dimension is explained by the variable "label". And, also, this is significant. We see also that the Vouvray have significantly positive coordinates in the first dimension, whereas the Sauvignon have significantly negative ones.

These tips for interpreting data are particularly useful for understanding components when we have a lot of variables.

Slide 35:

To conclude, how do we implement PCA in practice?

First, we must choose which variables will be active, that is to say, which variables will be used to calculate the distances between individuals, and which variables will be kept aside as supplementary. These supplementary variables don't contribute to building the dimensions, but may be used for interpreting them.

Then, we must decide whether to standardize variables or not. Standardization is essential if the variables have different units. When they have the same units, both solutions are possible, and lead to separate analyses. This decision is therefore crucial. Standardization means that every variable has the same importance. Not standardizing the data therefore means that each variable is weighted, with a weight corresponding to its standard deviation. This choice is therefore all the more important if the variables have quite different variances.

The next step is to perform the PCA.

Then, we have to choose the number of dimensions we want to interpret: do we interpret the first 2 only, or also the third and fourth? If we do go to 3 or 4, we now must extend the PCA to this many dimensions.

Next, we interpret the PCA results, using both the plots of the individuals and variables. We're going to go back and forth between these two plots, trying to quantify and understand the differences and similarities between individuals, and relationships between variables.

We can use indicators to enrich our interpretation, for instance, calculating contributions using the squared cosine, to be sure that distances seen in the plane are representative of distances in the global space.

And lastly, when we conclude that two individuals are similar, it is important to go back to the raw data, to be sure that we have not misinterpreted things. Whenever the analysis suggests an interesting interpretation, we should return to the raw data or standardized data to validate what has been inferred.

Slide 36:

To conclude, here are some useful supplements to the course.

There are many books on exploratory multivariate data analysis, and we suggest in particular the one that follows the structure of this course: Exploratory Multivariate Analysis by Example using R.

There is also an R package, called FactoMineR, that lets you do PCA with supplementary quantitative and qualitative variables, with many indicators such as quality of interpretation, contributions, etc. And lastly, there are several videos on Youtube, and a Youtube channel playlist with videos in French and English. Some of these videos show how to perform PCA without missing values, others how to handle missing values in PCA.