From Tables to Visuals:
Principal Component Analysis, Part 1

D.L. Massart and Y. Vander Heyden, Vrije Universiteit Brussel, Belgium.

Chromatographic analysis[1] often leads to data tables. When many samples are analysed for more than a few components, a table results in which each sample is characterized by many variables (e.g., peak areas and heights). Such tables are not easy to interpret. Chemometrics often relies on visualization to help the chemist obtain the required information, and the most used method in this respect is principal component analysis (PCA). PCA extracts information from data tables by transforming them into plots, such as the one shown in Figure 1. In such a figure it is easy to see that the samples in this (fictitious) example consist of two groups, namely 1–4 (group I) and 5–8 (group II), separated along PC1. Along PC2, sample 1 is somewhat different from the other three samples of group I and, in group II sample 8 is different from the three others. In Part 1 of this column we will concentrate on the visualization of the information about the samples, while the variables will be discussed in Part 2. We will then discover how PCA can tell us which variables yield similar (or different) information, have characteristic values for a specific group of variables etc.

An Example and Some Terminology

The aim of this and the next column is to provide some background on the PCA method, without going into mathematical detail, and to introduce some of the terminology used by practitioners of the method.

An example from food analysis is given in Figure 2. The concentrations of eight fatty acids were determined for 572 olive oil samples originating in nine different regions of Italy.[2] The analyst wants to know if certain regions have different fatty acid profiles and if the origin of the olive oil samples can be ascertained by analysing the fatty acids. Figure 2 is called a *score plot* or more precisely *the plot of the scores on principal components 1 and 2 (PC1 and PC2)*. The score plot provides the information required. It is indeed easy to see that, for example, the olive oils from Liguria are different (have a different fatty acid pattern) from those of Sardinia, while those of Sicily and Apulia are rather similar. Also, one East Ligurian sample (to the left on the figure) is different from the others.

A data table with *m* variables has *m* dimensions. The data are said to be *m-dimensional* or to be present in an *m-dimensional space*. PCA reduces the dimensionality of the data from m dimensions to just a few so that graphical representation is possible. The original olive oil data are eight-dimensional and the PCA plot reduces this to two. The two new variables are the principal components. In this column we will assume that there are only two principal components, but we will see later that it is possible and often useful to consider more. PCA practitioners often refer to variables as *features* and the reduction in the number of variables is termed *feature reduction*.

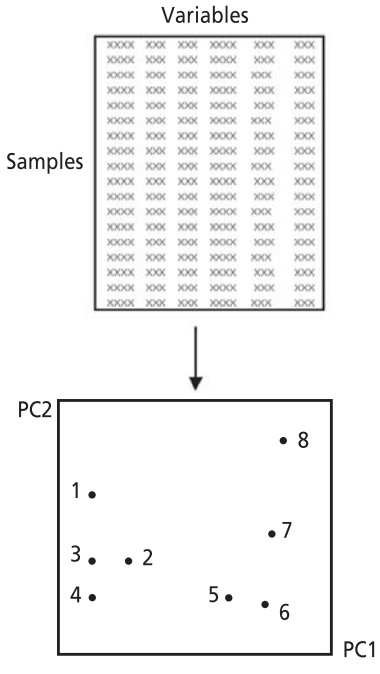
Variation = Information

Mathematically a table is a matrix, and the mathematics of PCA are therefore based on matrix algebra. It is not necessary to understand in detail this algebraic approach to be able to apply PCA, but to make the best possible use of PCA it is useful to understand the basic philosophy of the method. We will not explain the algebra, but apply a more intuitive and simpler approach, which is more useful in the interpretation of the PCA results.

By reducing the dimensionality, it can be expected that some of the information will be lost. The *feature reduction* from *m* dimensions to only two must therefore be performed in a logical manner to minimize this loss. An important definition to clarify first is that of the term 'information', which until now we have used very loosely.

Let us consider the very simple case represented in Figure 3. Two variables x_1 and x_2 were measured; the data table generated would consist of two columns and as many rows as there were samples. Because there are only two variables they can be plotted against each other as is done in Figure 3. We can then observe the

Figure 1: Data tables are transformed into (score) plots. The points and numbers represent samples.



data structure, the main feature of which is that there are two groups of objects. However, imagine that we have only one-dimensional sight: we can make observations only in a one-dimensional space; that is, along a line. The two variables constitute a two-dimensional space and we, therefore, have to reduce this two-dimensional space (the plane) to a one-dimensional space (the line), and do that in such a way that the essential characteristics of the data structure are preserved: along the line the data should still consist of two groups of objects

The operation by which the *feature reduction* is achieved is an orthogonal projection from the points in the plane onto a line. Many lines can be drawn and the question is which one is best. As shown in Figure 4, the choice of the direction of the line is important. In Figure 4(a) the projections of the two groups of objects coincide, and with an imaginary one-dimensional eye we would conclude that there is only one cluster of objects; that is, the information present in the two-dimensional space is lost. In Figure 4(b) a better choice is made. The projections on this line show the two clusters of points. This line is the *first principal component*.

Chemometrics often relies on visualization to help the chemist obtain the required information, and the most used method in this respect is principal component analysis (PCA).

By definition, the first principal component, PC1, is drawn in the direction of the largest variation (mathematically: variance) in the data. The PCA method assumes that the large variation in the data also carries most information. Other choices are possible and some methods that resemble PCA, in that they perform *feature reduction* by projection, apply other criteria. For instance, in a group of methods called collectively *projection pursuit* the direction of the line is chosen such that it maximizes an index of inhomogeneity and, depending on the exact choice of criterion, the method then focuses on finding outlying objects or on finding clusters. The results of PCA and projection pursuit are often similar but there are instances where different characteristics of the data structure are highlighted.

The projection of an object on PC1 is called the score of that object on PC1. The scores on PC1 for the objects of Figure 4(b) are shown in Figure 5. The number of features is reduced from 2 (x_1 and x_2) to 1 (PC1) and this was done successfully because the main feature of the data, namely the existence of two groups, is still perceived.

The Second Principal Component

Figure 4(b) shows that the objects are dispersed around PC1. This means that there is some variation which is not expressed by PC1. This unexplained variation can be shown on a second principal component, PC2. PC2 is by definition orthogonal to PC1. For 2 x-values, this means that its direction is

Figure 2: Score plot on PC1 and PC2 for 572 olive oil samples characterized by the composition of eight fatty acids.

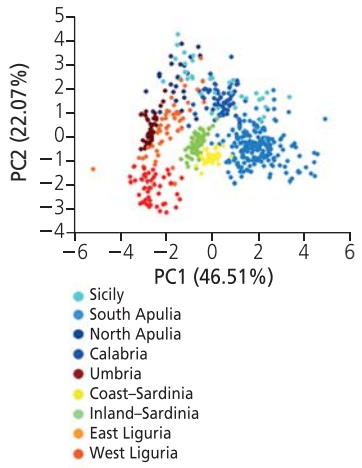


Figure 3: Samples characterized by two variables.

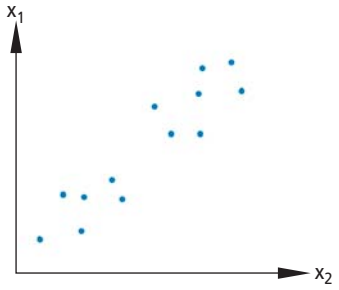


Figure 5: The projections on the line (principal component 1) of Figure 4(b). The data dimensions have been reduced from 2 (Figure 3) to 1.

Figure 4: (a) Orthogonal projection from the points in Figure 3 on the black line hides the data structure. (b) Orthogonal projection on the blue line preserves the original data structure.

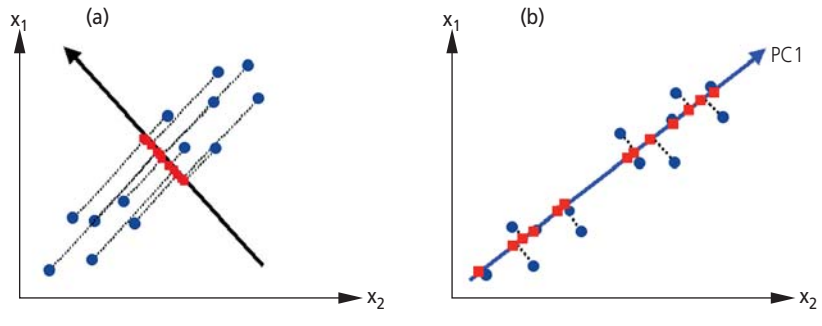
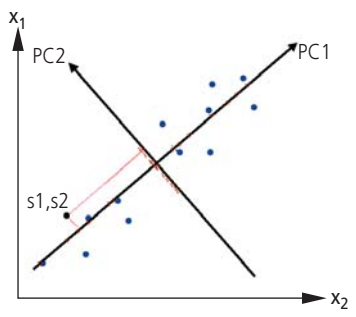


Figure 6: Principal components 1 and 2 represented in the space of the original two variables.



Practical Data Handling

Figure 7: Principal components 1 and 2 define a new two-dimensional space.

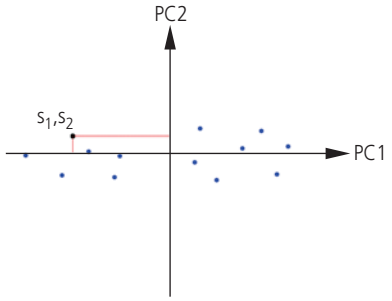


Figure 8: The variation in the data along PC1 is larger than along PC2; PC1 is more important than PC2.

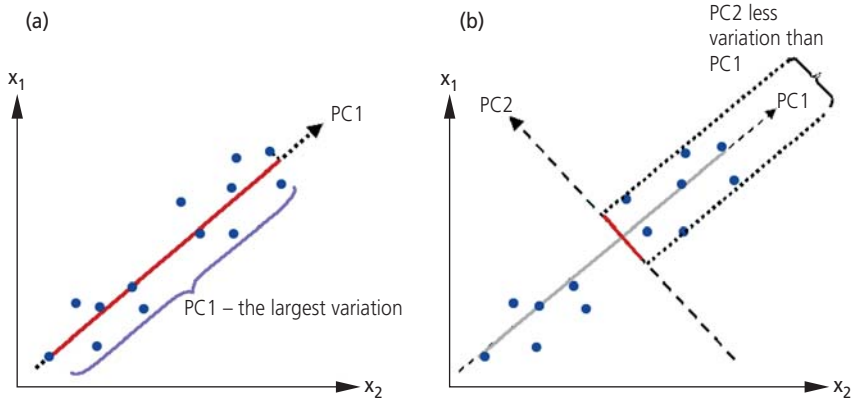
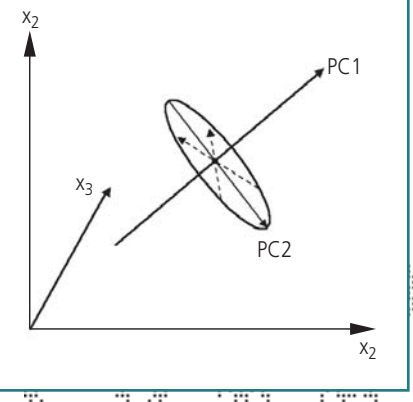


Figure 9: PC1 and PC2 in the space of the original three variables; PC2 is the direction of largest variation around PC1.



fixed by the direction of PC1. This is shown in Figure 6. The objects can be represented by their projections on to PC2 in the same way as on PC1. These projections are the scores on PC2 and each object is now characterized by two scores, one on PC1 (s_1) and one on PC2 (s_2). In PCA jargon, it is said that the objects are now present in the new two-dimensional space PC1-PC2 with as coordinates their scores (s_1,s_2) (Figure 7).

One could wonder what the advantage is of going from the x_1–x_2 plane to the PC1–PC2 plane. It should be remembered that we are viewing the data with a one-dimensional eye and that, therefore, we are unable to observe the two-dimensional PC-space. As variation is equated with information, and there is less variation in PC2 than in PC1, PC2 is always less important than PC1. This can be verified on Figure 8: the variation is larger along PC1 than along PC2. Because PC2 is less important, we can decide to ignore PC2 and to retain only PC1 and, in this way, *feature reduction* to one dimension is again achieved.

PCA Decorrelates Correlated Data

An aspect of PCA that merits attention is that it eliminates correlation between variables. In Figure 3, x_1 and x_2 are clearly correlated, but Figure 7 shows that in the PC1–PC2 plane the correlation has disappeared. This observation gives some insight into why PCA can achieve *feature reduction* with little loss of information. Correlated variables give similar information. In our simple example a plot of the data on x_1 would be very similar to the plot on x_2 and, therefore, the information provided by one of the two variables is to a large extent redundant. It can be compressed in a smaller number of variables and this is exactly what PCA does: in the example PC1 replaces x_1 and x_2. By eliminating redundancies PCA concentrates the information in less variables.

The decorrelation has another advantage, which we mention here only in passing as it is not of immediate use in the present application of PCA — it allows regression with many correlated variables. Multiple linear regression cannot be applied well when variables are strongly correlated. By using the scores on the (orthogonal, and therefore not correlated) PCs as variables instead of the original variables, this difficulty can be avoided. This technique is called *principal component regression*, and we hope to discuss it in somewhat more detail in a later column.

More than Two Original Variables

PC1 was obtained by finding the direction in the data along which the variation is largest. Then PC2 was obtained. It is orthogonal to PC1 and, when there are only two original variables, the direction of PC2 is therefore fixed by that of PC1.

However, for more than two original variables, this is no longer the case. PC2 must still be orthogonal to PC1, but it can take different directions. It is easiest to imagine this in a three-dimensional space (see Figure 9). All possible orthogonal directions form a plane orthogonal to PC1 and from them the direction is chosen along which the largest variation occurs: PC2 is drawn in the direction of largest variation in the data around PC1. It could also be said that the PC1–PC2 plane is the plane that contains the largest variation of all possible planes that can be drawn through the three-dimensional data. What is true for three dimensions is also true for *m* dimensions: the PC1–PC2 plane contains the most information of all planes that can be drawn through the data in the *m*-dimensional space.

Let us now go back to the example of Figure 2. The figure is the plane that contains the largest variation and, therefore, the most information of all possible planes that can be drawn through the eight-dimensional space of the original variables. We also know that PC1 is more important than PC2, because the amount of variation explained by PC1 is higher than that for PC2. How much variation can be computed, and is usually shown on the PC plot. In this instance (see Figure 2) PC1 (46.51% of the variation) is about twice as important as PC2 (22.07%). This also means that differences seen along PC1 are larger than those along PC2. Because of the way the plots are made, this cannot usually be seen. Indeed, a normalization is performed by most software when plotting the sample points, so that the resulting figure gives the impression that the differences along the two PCs are equivalent.

What More do we Need to Know about PCA?

There are a certain number of things we still need to learn to make full use of the capabilities of PCA. The following will be explained in the next column:

- It is possible and sometimes necessary to obtain more than PC1 and PC2, because some of the information is present in PC3 or even higher PCs.
- The information that can be obtained about the variables and the role they play resides in the so called loadings and loading plots. They allow us to decide which variables are most important for the differences observed between the samples.
- As for any method, there are some pitfalls in the use of PCA. For instance, when strong outliers are present, these are displayed but the structure in the remaining data may be hidden.
- In many instances it is necessary to apply a (simple) pretreatment to the data, for instance to avoid scale effects.

Acknowledgements

The authors thank Dr Michal Daszykowski for performing the computations leading to Figure 2.

References

1. D.L. Massart and Y. Vander Heyden, *LC•GC Eur.*, **17**(9), 467–470, 2004.
2. M. Forina and E. Tiscornia, *Annali di Chimica*, **72**, 143, 1982.

Column Editor, **Desire Luc Massart** is a part-time professor at the Vrije Universiteit Brussel, Belgium. He performs research on chemometrics in process analysis and its use in the detection of counterfeiting products or illegal manufacturing processes.

Yvan Vander Heyden is a professor of analytical chemistry at the university and heads a research group in chemometrics and separation science.