Introduction

I would like to welcome readers who are new to this column with a few words of explanation. Chemometrics is a subject which has generated (and continues to generate) much interest and excitement in analytical spectroscopy. While there is a rather small band of experts who are developing new techniques, it is not necessary to be an expert to utilise chemometrics given some basic understanding of the limitations and potential pitfalls for the unwary user of chemometric computer software. I am not a chemometrician, but I have friends who are. The aim of this column is to provide a bridge between chemometricians (who are expert mathematicians) and potential spectroscopic users who are probably not mathematicians. The intention is that users should develop an understanding which will provide a suitable balance between rejection of unknown methods and unqualified enthusiasm for "black-box" software. This should enable them to make successful use of these powerful enhancements to good spectroscopy. Photocopies of the previous articles in the Chemometrics Column series in Spectroscopy World *are available from the publishers at a very modest cost and after this issue it will be assumed that readers have the knowledge or the reprints.*

Former readers of Spectroscopy World *will recognise my article on Principal Component Analysis (PCA) which is repeated to enable new readers to fully comprehend the new article by Ian Cowe. While Ian may claim NOT to be a chemometrician, his paper on the utilisation of PCA is probably one of the most frequently referenced papers in near infrared spectroscopy. It is a great pleasure to welcome him to the first of these Columns in* Spectroscopy Europe.

The principles of principal component analysis*

by Tony Davies, Column Editor

Since the beginning of this column we have been taking a fairly relaxed tour of chemometric concepts while attempting to exclude mathematics as far as possible. I do not intend to change this approach, but in future columns we will have to be able to make assumptions of comprehension of some key topics. Principal Component Analysis (PCA) is one of the fundamental methods of multi-variate analysis and hence of chemometrics. It was introduced in an early column [*Spectroscopy World* 2(2), 32 (1990)] but it is so important that this and the next column will be devoted to it.

PCA is a method of data analysis which requires a matrix of samples and variables. It finds the maximum variations in the data and forms new variables [known as Principal Components (PCs)] such that:

each successive PC accounts for as much of the remaining variability as possible except that,

each new variable must be orthogonal (at right angles) to all other variables.

PCA is easily defined by matrix algebra but the intention of this column is to present ideas in diagrammatic forms. This makes life difficult, because we are visually restricted to three dimensions and thus we can only illustrate the working of PCA in terms of three variables. It is important to realise that the power of PCA is in being able to examine large numbers of variables and to compute many principal components which are mathematically orthogonal to each other. In some discussions of PCA this ability is not emphasised because of the difficulties of demonstrating it and the reader could be left with the impression that we only use two or three principal components. Except for very simple data, the number of principal components is more likely to be between 10 and 20.

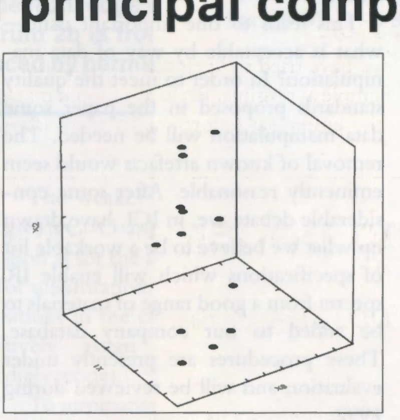

Figure 1. 3D data.

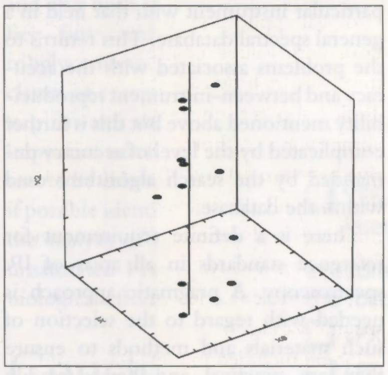

Figure 2. First PC.

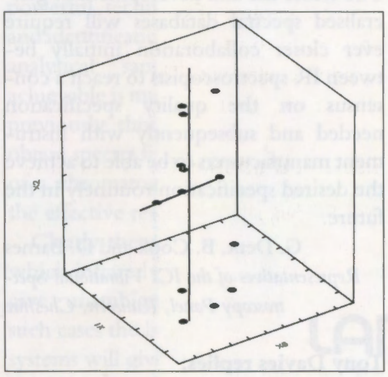

Figure 3. Second PC.

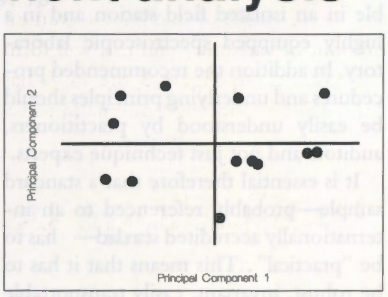

Figure 4. Scores plot.

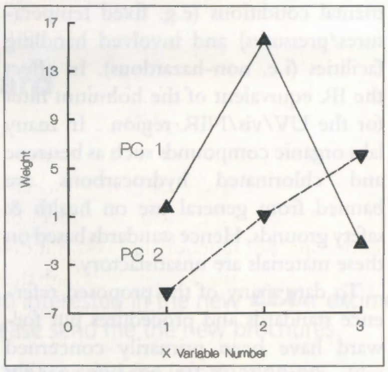

Figure 5. Weights plot.

The output from PCA is in the form of two tables and some statistical information. The first of these contains values for each sample on each Principal Component. These are known as scores. The other contains coefficients used to compute the components from the original variables which are known as weights (or sometimes coefficients). Both contain useful information. The scores are mainly concerned with the samples and can be used in place of the original variables, while the weights show how the components are formed and tells about the distribution of information in the data set. If you remember the article about cutting the data cake [*Spectroscopy World* 2(1), 35 (1990)] then the weights

*reprinted from *Spectroscopy World* 4(1), 23 (1992).

are represented by the shape of the cutter and the table of scores are new slices of the computed cake. One of the important statistics from a PCA is the total percentage of variance explained. This should be very close to 100%. The first few PCs will contain the majority of the variance but experience with PCA soon leads one to take notice of the later PCs which may explain only very small variances; sometimes this can be the crucial information in your data. Not retaining sufficient PCs can be like throwing out the baby with the bath-water.

Figure 1 shows a three-dimensional plot for three variables measured on a set of 13 samples. In Figure 2 PCA has found the vector which contains the maximum amount of variability and this will form the first PC. In Figure 3 the PCA has found the position and orientation of a vector which is at right angles to the first PC and contains the maximum amount of variability compared with all the vectors which conform to the specification; this is the second PC. Figure 4 shows the scores for the samples as a plot of the two PCs and Figure 5 shows plots for the PC weights. Figure 4 contains 98% of the variation present in the original three variables. The first component accounted for 73% and the second for 25% of the total variance. It can be seen from the weights plot that the first component is dominated by the second variable, while the second component is largely a product of the first and third variable.

Notes

have tried to keep this explanation as simple as possible. Perhaps I should make the point that before carrying out PCA it is usually necessary to transform the data. This involves correcting for the mean (i.e. subtracting the mean value of that variable) and sometimes standardising by making the variance of each variable equal to 1. Most software packages will do this for you so that my simple model is sufficient until you want to check that your program is giving correct answers!

The orthogonality of PCA is actually a dual orthogonality. Not only are the vectors orthogonal but also the scores are uncorrelated (i.e. orthogonal).

Acknowledgement

I am grateful to Tom Fearn for making sure that during my efforts to obtain simplicity I have not strayed from a valid description of PCA.

Applications using principal component analysis

Ian A. Cowe

10 Buddon Drive, Monifieth, Dundee DD2 5DA, Scotland.

In a previous article, Tony Davies explained how principal components are derived and defined some of their basic properties. In this article, I will look at one application of Principal Component Regression (PCR) to predict composition and also consider applications where components are used as an assessment of some aspect of functionality without direct use of constituent data. Although the applications discussed will relate to near infrared diffuse reflectance spectroscopy, the same general principles apply in other fields.

PCR is a chemometric technique which uses all the spectral data to predict composition. It provides two new variates, "weights", which represent the relative importance of each of the original data values to the components and which can be used for spectral interpretation and "scores" which condense the original data into a few uncorrelated values which can either be regressed against chemical values, or examined by other techniques such as discriminate analysis to reveal some underlying trend or relationship.

Scores are derived solely from the spectral data and we obtain a score for each sample on each principal component. Each orthogonal vector (or PC) represents in turn a decreasing amount of the spectral variation. By monitoring the cumulative variance as each component is derived we can determine easily how many components are needed to model the variation that relates to major physical and chemical effects.

A real application, in this case wheat flour with values for protein and moisture,[1] shows how easy it is to use PCR. Table 1 shows a summary for the first few components. Although we normally derive between 10 and 20 components, in this case only the first few correlated with moisture and protein. The remainder had uniformly low correlations and together represented less than 0.02% of the spectral variation.

To be included in a model, a component should have a significant correlation with the constituent of interest and express an amount of spectral variation in proportion to its concentration and absorption coefficient. This avoids the inclusion of later components which express practically zero variation but have statistically significant correlations due to random chance.

Table 1. Statistics for wheat flour.

PC No.	%Var.	%Cum.Var.	r_m	r_p
1	98.60	98.60	–0.16	–0.71
2	0.99	99.59	0.97	–0.08
3	0.22	99.81	0.09	0.21
4	0.11	99.92	–0.05	0.66
5	0.05	99.97	0.07	0.10
6	0.01	99.98	–0.02	–0.01

In Table 1, only two components (PC1 and PC2) correlate with oven dried moisture. Water is one of the strongest absorbers and should be present at about 12% in these samples. So we should expect that early components would be dominated by water. In fact, the second component ($r = 0.97$) alone would be enough to adequately predict moisture content. With protein, a weaker absorber, the first, fourth and, to a lesser extent, the third components showed some correlation.

One of the main advantages of PCR over conventional wavelength regression is that spectral interpretation of a model is much easier. When all the "x" data are spectral values then plots of the weights become analogous to spectra.

Table 2. Building regression models for moisture and protein in wheat.

For moisture		
PC2 + PC1	$= 0.97^2 + 0.16^2 + 0.09^2$	= 0.983
PC2 + PC1 + PC3	$= 0.97^2 + 0.16^2 + 0.09^2$	= 0.987
For protein		
PC1 + PC4	$= 0.71^2 + 0.66^2$	= 0.969
PC1 + PC4 + PC3	$= 0.71^2 + 0.66^2 + 0.21^2$	= 0.992
PC1 + PC4 + PC3 + PC5	$= 0.71^2 + 0.66^2 + 0.21^2 + 0.10^2$	= 0.997

Figure 1 shows the shapes of the first four components. These are plots of the weights against wavelength. Typical NIR spectra consist of approximately 700 data points covering the range 1100 to 2500 nm and so we have 700 weights. For each component we get a weight at each wavelength and the weights are scaled in such a way that as the sum of the squared weights across the spectrum always equals one. This means that, for any component, wavelengths with large weights are proportionally more important in determining the sample score on that component than wavelengths with weights close to zero.

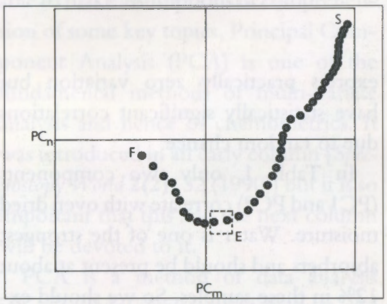

Figure 1. First four principal components for wheat flour.

If we look at Figure 1, we see that PC2 (which correlated highly with water) has a shape similar to a water spectrum. PC4, which correlated highly ($r = 0.66$) with protein, shows protein bands as high positive weights at 1980, 2050, 2180 and 2210 nm. But how do we interpret PC1? It expressed almost all the spectral variation, correlated highly with protein, yet showed no evidence of protein bands.

In fact, PC1 relates mainly to baseline shifts caused by variation in particle size between samples. The particle size of the ground flour is determined largely by grain hardness, and protein in one of the factors which affects hardness. Thus, PC1 relates to protein through a secondary correlation with a physical factor. Normally the use of an indirect correlation should be avoided but, as this relationship is likely to be stable, here we can exploit it in our protein model.

The orthogonality of principal components makes regression modelling a simple and predictable process. The multiple correlation for any combination of components relates directly to the individual component correlations. We simply sum the squared individual correlations and take the square root to obtain the multiple correlation (see Table 2). Adding PC3 only marginally improved the model for moisture, while for protein adding PC5 made little difference. Thus we would predict moisture using the first two components and protein using PCs 1,3 and 4. The form of the models is as follows: % Protein = 10.99 + 2.219 × PC1 + 13.72 × PC3 + 60.55 × C4; % Moisture = 13.47 + 0.24 × PC1 + 14.47 × PC2

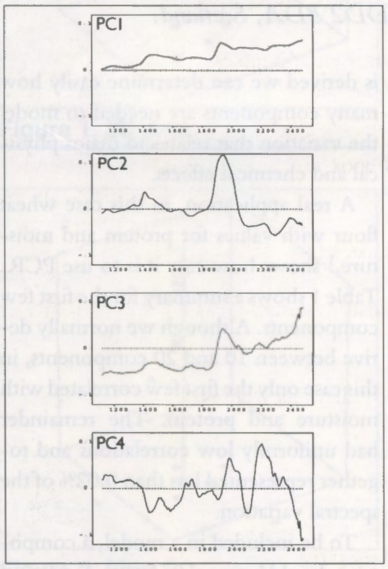

Figure 2. Scores/scores plot for process control. S = start or initial value, F = final value. The dashed line represents the normal end point for the reaction.

One strength of PCR is that, because of orthogonality, values of regression coefficients do not change when terms are added to or subtracted from the model. Thus for protein, the largest coefficient (PC4) always has a value of 60.55.

One strength of principal components is that they are derived solely on the spectral data. They can be used even where no suitable reference values are available. Take, for example, the problem of monitoring progress of a batch process. An example was presented recently by Griffin, Kohn and Cowie.[2] Using the sample scores we can represent each sample as a single point in a p dimensional space (where p is the number of components). The scores are Cartesian co-ordinates defining where each point lies within the space. As we cannot visualise more than three dimensions we normally select two components to provide a suitable two dimensional "window" on the p dimensional space.

If, for an imaginary example, we took samples every few minutes throughout the life of a process to a point beyond where it normally would be stopped, we might find that the scores form a "track" across a plane defined by two components (Figure 2). This is not surprising as the samples form a time series and adjacent samples are closely related. If the batch process were repeated several times, then we could measure the errors associated with "normal" operation throughout the process. Finally, we could identify a small area of the two dimensional space which represents an acceptable end point for the reaction.

When subsequent batches are run, the operating conditions can be modified to keep the reaction " on track" , and when scores within the end point space are encountered the process can be stopped. When linked with feedback control systems this forms a powerful system.

These examples show two contrasting ways in which principal components can provide a solution to basic chemometric problems. There are several statistical programs currently available for personal computers which provide PCA as an option.

References

1. I.A. Cowe and J.W. McNicol, "The use of principal components in the analysis of near-infrared spectra" , *Appl. Spectrosc.* **39**, 257 (1985).
2. J.A. Griffin, W. Kohn and J. Cowie, in *Making Light Work: Advances in Near Infrared Spectroscopy*, I.A. Cowe and I Murray (Eds), Proc. of the 4th Int. Conf. on NIR Spectrosc., 13–19 Aug, 1991, Aberdeen, Scotland. VCH, Weinheim, Germany (1992).