Back to basics: when you need more than Principal Component Analysis

A.M.C. Davies[a] and Tom Fearn[b]
[a]Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK
[b]Department of Statistical Science, University College London, Gower Street, London, UK

Introduction

We hope that after our previous two columns[1,2] you will have realised that we think it is vital that you understand the importance of principal component analysis (PCA). So we feel confident that in this column we can tell you that PCA is not the answer to every problem that can be solved by chemometrics, without you thinking that PCA is not useful. It is, but sometimes you need more than just PCA. You will remember that PCA gives us loadings and scores and we have shown that scores plots can be a very useful method for visually accessing a data set. If we have samples of different pure components then PCA will very often gives scores plots which separate the samples and this has lead people to use scores plots for determining identity and other sample attributes. **There is nothing in PCA to justify this procedure!** It may produce a good-looking result but if this is the main purpose of the study then you can do even better by using an appropriate chemometric tool. PCA is an analysis of the variation within a data set. If the main causes of this variation are related to the identity of the sample then PCA will produce the desired result but this is fortuitous and cannot be expected to arise in all situations. The best way to see this is to look at some examples and that is what we will show in this column. We will discuss the methods employed in more detail in later columns.

An example where PCA produced a good result

We are not trying to tell you that PCA is never an adequate procedure. An example from a recent paper by Schwanninger *et al.*[3] is shown in Figure 1. These are some of the results from an NIR study of the biodegradation of wood by three strains of fungi. The figure very clearly shows that NIR spectroscopy can distinguish between the same wood being attacked by different fungi.

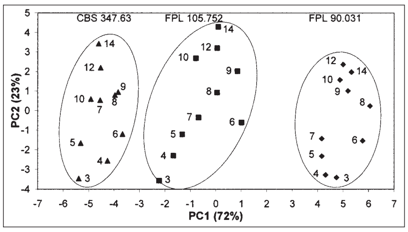

Figure 1. Separation of wood samples subjected to different fungi by PCA. Reproduced with permission of NIR Publications from Reference 3.

Examples where PCA alone was not adequate

Having shown you one example where PCA was very obviously all that was required to support the authors' conclusions, we will now show three examples where this was not the case. You will not be surprised that they are all based on NIR spectroscopy but the results can be generalised to any type of multivariate data.

The first is a study of the authentication of commercial wheat flours. The PCA scatter plot of scores shown in Figure 2 (note the use of 1st and 4th PCs) gave quite a good separation and some authors might have stopped at this point. However, Sirieix and Downey wanted a better result so they took the PC scores and put them into a factorial discriminant analysis program. The result is shown in Figure 3. Not only is this a better visual result, it also enabled them to compute the confidence in the discrimination.

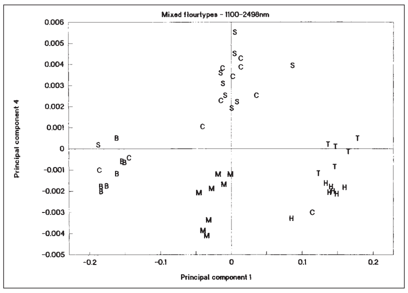

Figure 2. Scatter plot using the 1st and 4th PCs from the PCA of wheat flours. Reproduced with permission of NIR Publications from Reference 4.

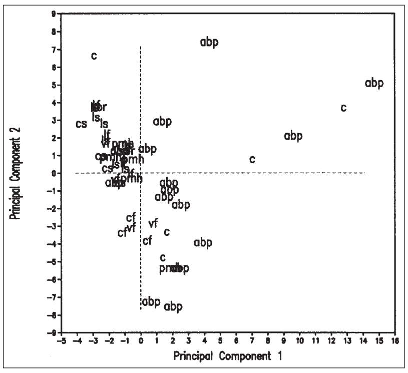

Figure 4. Scatter plot of a PCA of ruminant feeds. Reproduced with permission of NIR Publications from Reference 5.

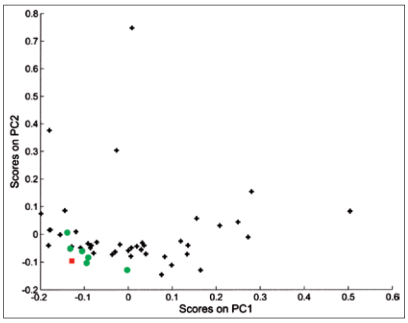

Figure 6. A scatter plot of the first two PCs indicating the position of acacia samples (green circles). The red square symbol indicates the position of an acacia sample which was excluded from the PCA. Reproduced with permission of NIR Publications from Reference 6.

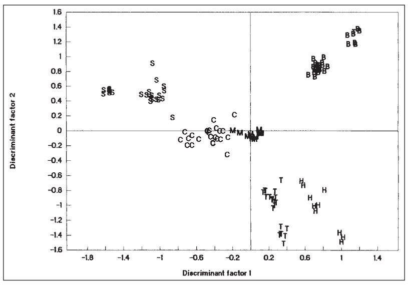

Figure 3. Factorial discriminant analysis plot of the data shown in Figure 2. Reproduced with permission of NIR Publications from Reference 4.

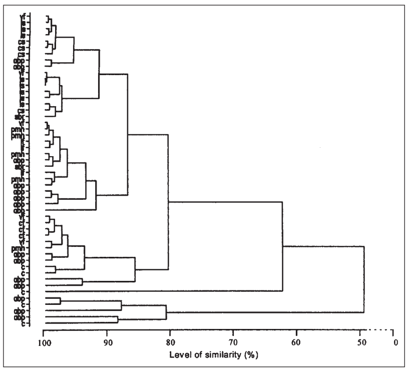

Figure 5. Clustering of the different ruminant feeds. Reproduced with permission of NIR Publications from Reference 5.

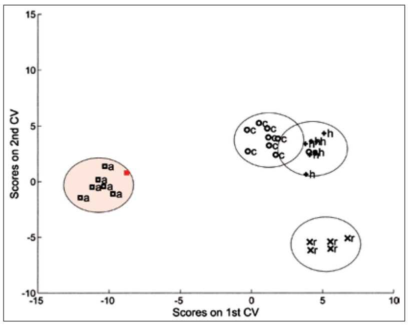

Figure 7. Scatter plot of the first two CVs showing that the test sample was correctly identified as acacia. The ellipses round the groups are probability ellipses indicating a 95% probability of group membership. Reproduced with permission of NIR Publications from Reference 6.

The second example comes from some work by Susan Lister and colleagues on ruminant feeds from the Mediterranean region. The PC scatter plot, Figure 4, showed quite good separation of the 47 samples by type but this was considerably improved by hierarchical clustering analysis as shown in Figure 5. This separation based on similarity allows us to see which samples are most different or most similar.

The final example comes from our own work on the characterisation of honey. There were 48 samples of honey; 28 of them were from four plant sources, acacia, chestnut, heather and rapeseed. The remaining 20 samples came from a range of different sources with too few members to be characterised as a group. The scatter plot of the first and second PC based on the group members indicates some grouping of samples but also considerable overlap. In order to achieve some separation, the PC scores from 10 or 15 PCs were entered into a canonical variate analysis (CVA) program which separated the groups by the computation of three canonical variates. Because of the small number of samples in this study the results had to be validated by "cross-validation". In cross-validation, one sample is left out of the PCA and CVA calculations and then projected into the CVA space. This was repeated for all 28 samples. Figure 6 shows the PC scatter plot with one of the acacia samples excluded from the PCA and Figure 7 is the CVA plot for this sample indicating that it was identified as acacia.

Discussion

It is important to note that while PCA was not the optimum technique to produce the desired result it was an essential part of the study because it was able to reduce the number of variables (more than 700 in most NIR spectra) to a small number of PCs which could be used as the input variables for the final stage of data processing.

There is another use of PCA, which we have not mentioned so far; that is the use of PCA in quantitative analysis in the form of PCA regression or PCR. While PCR is a very good technique it is very similar to the more widely used partial least squares regression, PLS. We think there is very little difference in the advantages of the two techniques so, because of its popularity we are going to concentrate your attention in later columns on PLS rather than on PCR.

References

1. A.M.C. Davies and T. Fearn, *Spectrosc. Europe* **16(6),** 20 (2004).
2. A.M.C. Davies, *Spectrosc. Europe* **17(2),** 39 (2005).
3. M. Schwanninger, B. Hinterstoisser, C. Gradinger, K. Messner and K. Fackler, *J. Near Infrared Spectrosc.* **12,** 397 (2004).
4. A. Sirieix and G. Downey, *J. Near Infrared Spectrosc.* **1,** 187 (1993).
5 S. Lister, M.S. Dhanoa, W. Ebenezer, S. Lopez and J. France, *J. Near Infrared Spectrosc.* **6,** A79 (1998).
6. A.M.C. Davies, B. Radovic, T. Fearn and E. Anklam, *J. Near Infrared Spectrosc.* **10,** 121 (2002).