TONY DAVIES COLUMN

The last furlong (3). Principal component analysis

A.M.C. Davies
Norwich Near Infrared Consultancy, 10 Aspen Way, Cringleford, Norwich NR4 6UA, UK. E-mail: td@nnirc.co.uk

Introduction

I am not sure when I first heard about principal component analysis (PCA); probably before I went to the Food Research Institute (FRI) in 1972. Before I went there I had been attempting to identify the country of origin of honeys from amino acid analysis data (I went to FRI as an amino acid analyst) and multivariable data was just beginning to be used. However, I do not think I had ever considered PCA for NIR spectroscopy until my friend Ian Cowe told me about the work he was doing (at the Scottish Crop Research Institute) which was published in *Applied Spectroscopy* in 1985.[1]

In 1989 Ian Michael invited me to write about chemometrics in the new publication *Spectroscopy World* (which became *Spectroscopy Europe* in 1992). The first column was a gentle introduction to chemometrics but the second was about the idea of computing new variables, as happens in PCA. It was called "The Data Cake"[2] and in it I tried to present a simple mental model of how raw data containing many variables can be transformed to new but fewer variables in which the useful data is concentrated. I was invited to present this idea at the third ICNIRS conference in Brussels in 1990. After the lecture I asked an American friend, Steve Buco, what he thought of it. His surprising answer was "I was going to ask you if I could borrow your slides for the next training course at Chambersburg". He was referring to the International Diffuse Reflectance Conference (IDRC) which had been held at Wilson College, Chambersburg, USA, since 1982 (and is still running) in every even numbered year. Steve had been presenting a very popular pre-conference course on statistics and chemometrics, so I said "If you like, I will come and show them". The conference organiser agreed to fund my attendance and I learnt enough to continue the course with Tom Fearn when Steve got too busy to run it. I am not sure that many people found the Data Cake very useful but it was very helpful to me!

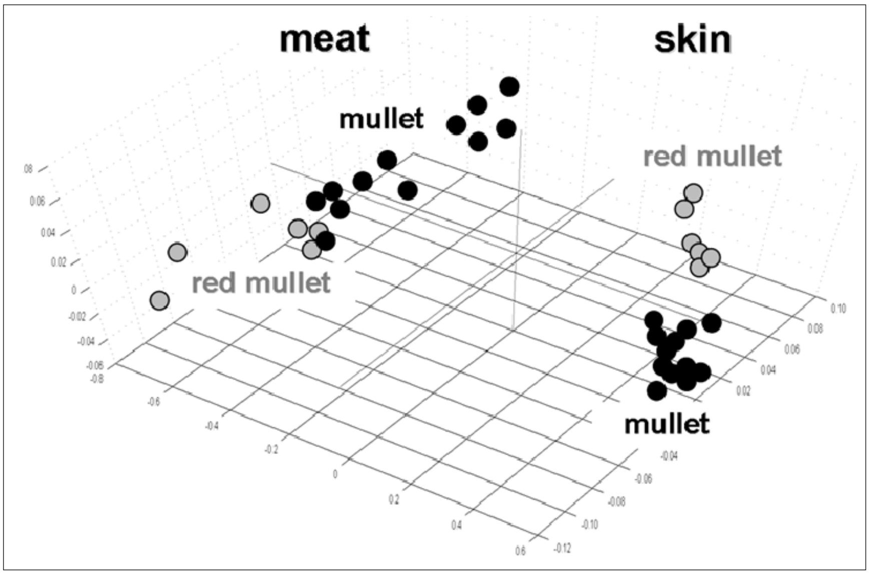


Figure 1. 3D score plot of the PCA (principal component analysis) model for the discrimination of the different mullets. © IM Publications 2013. Reproduced with permission from Reference 5.

PCA

My first article that was directly about PCA was the last in *Spectroscopy World* and the first TD column in *Spectroscopy Europe*.[3] In 2004 Tom Fearn and I wrote the first of the "Back to Basics" columns with the same title as that first column.[4] It was about PCA and I do not think I can improve on it (you can find the article on the web).

However, there are two additional comments to make. The first is that PCA is a very useful tool but it will not solve all our problems. PCA is good at showing the groups that some samples naturally form, but this is not the same as forming the groups you would like to see. Sometimes it will work and other times something more specific is required. A good example of this was published in a recent paper in the *Journal of Near Infrared Spectroscopy*. It concerns the possibility of testing the identity of expensive fish compared to similar but less expensive alternatives which might be sold as a substitute.[5] They tested the skin and the meat of three pairs of fish. PCA on one of the pairs (mullet and red mullet) was sufficient to distinguish them from skin spectra but not from the meat spectra, Figure 1. Neither of the other pairs could be separated by PCA of skin or meat spectra but all were successfully classified by soft independent modelling of class analogies (SIMCA),[6,7] Figure 2.

I was intending to ask Tom Fearn if there had been any recent developments in PCA when my copy of *NIR news* arrived and I found an article in Tom's "Chemometric Space" in which he reported on the addition of ANOVA to PCA—ASCA.[8] This is possibly useful in designed experiments but Tom thinks that there are problems with testing the significance of the results.

Conclusion

PCA is a very important method but it is not the only tool in the box!

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). doi: 10.1366/0003702854248944
2. Tony Davies, "Cutting the data cake—having your cake and sampling it", *Spectrosc. World* **2(1),** 35 (1990). http://bit.ly/1jsxyxG
3. A.M.C. Davies, "The principles of principal component analysis", *Spectrosc. Europe* **4(2),** 38 (1992). http://bit.ly/1aDgtA4
4. A.M.C. Davies and Tom Fearn, "Back to Basics: The principles of principal component analysis", *Spectrosc. Europe* **16(6),** 20 (2004). http://bit.ly/1bJ0CAd
5. N. O'Brien, C.A. Hulse, F. Pfeifer and H.W. Siesler, "Near Infrared spectroscopic authentication of seafood", *J. Near Infrared Spectrosc.* **21,** 299 (2013). doi: 10.1255/jnirs.1063
6. A.M.C. Davies and Tom Fearn, "Back to Basics: Multivariate qualitative analysis, SIMCA", *Spectrosc. Europe* **20(6),** 15 (2008). http://bit.ly/1bJ0oco
7. T. Næs, T. Isaksson, T. Fearn and T. Davies, *A User Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester, Chapter 10 (2002). http://bit.ly/1byjy14
8. T. Fearn, "ASCA", *NIR news* **24(7),** 20 (2013). doi: 10.1255/nirn.1400

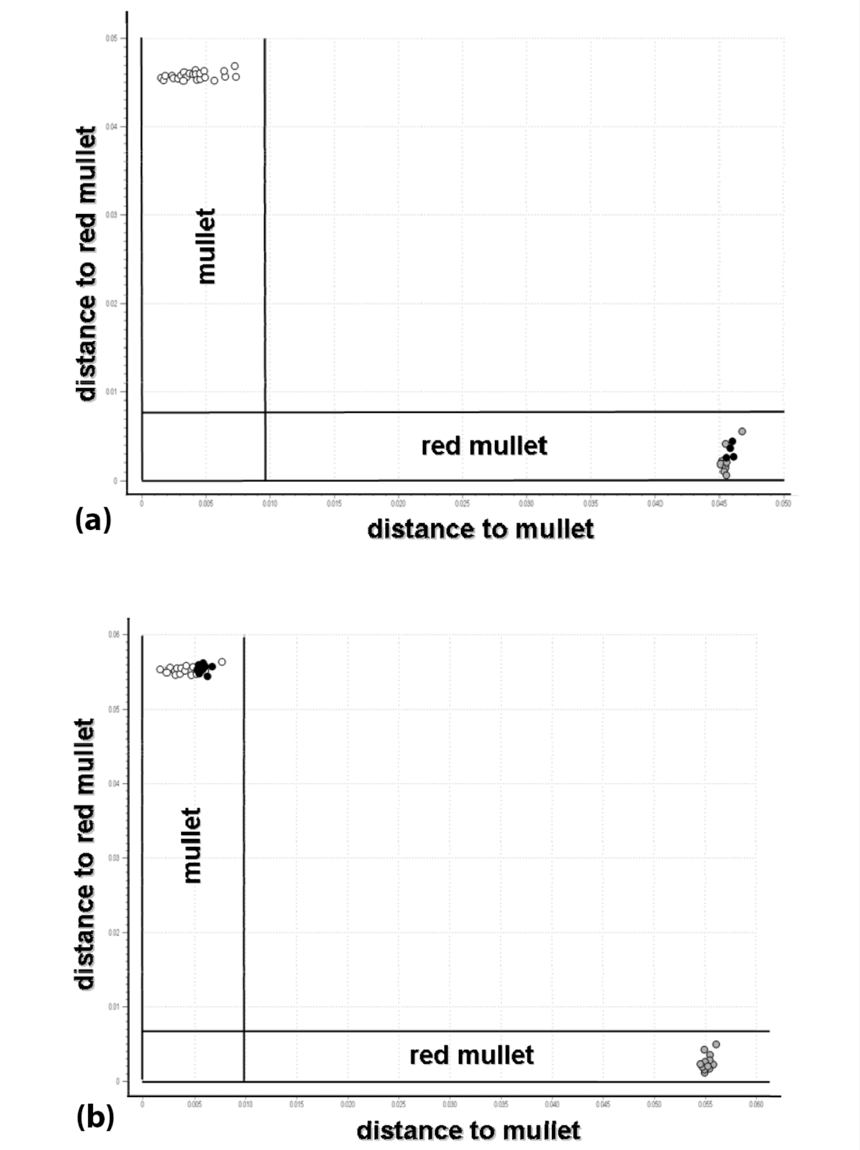
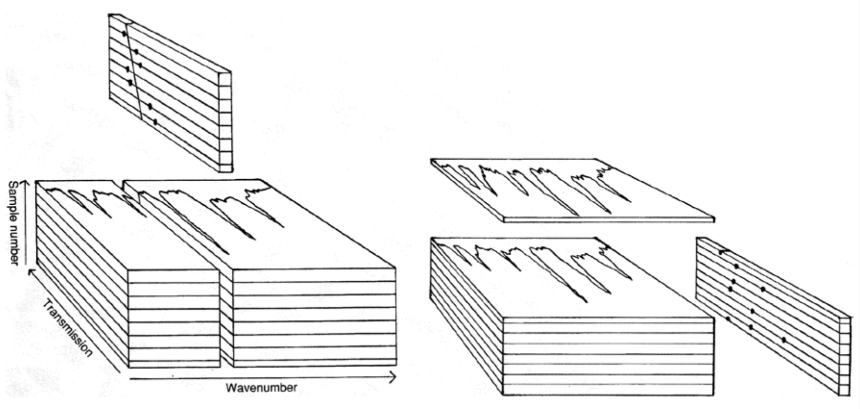


Figure 2. Coomans plots of SIMCA (soft independent modelling of class analogies) analyses (5% significance) demonstrating the correct identifications of (a) a red mullet test fish and (b) two mullet test fishes [the spectra of the test fishes are represented by the symbol (•)]. © IM Publications 2013. Reproduced with permission from Reference 5.



The data cake: read more in Reference 2.