A BIT ABOUT MULTIVARIATE ANALYSIS

Royston Goodacre
Department of Chemistry, UMIST, PO Box 88, Sackville St, Manchester M60 1QD, UK.
R.Goodacre@umist.ac.uk   T: +44 (0) 161 200 4480   F: +44 (0) 161 200 4519

Cluster analyses

The typical procedure for multivariate analysis is detailed in the figure below. The initial stage involves the reduction of the dimensionality of the hyperspectral data by principal components analysis (PCA) (Causton, 1987; Jolliffe, 1986). PCA is a well known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance, and the Matlab routine employs the NIPALS algorithm (Wold, 1966). Discriminant function analysis (DFA) is then used to discriminate between groups on the basis of the retained principal components (PCs) and the *a priori* knowledge of which spectra are replicates (MacFie *et al*., 1978; Windig *et al*., 1983). This process does not bias the analysis in any way. DFA was programmed by Dr Bjørn Alsberg according to Manly's principles (Manly, 1994).

DFA is not performed on the original feature space (spectra) because one can not feed co-linear variables or too many variables into DFA. The starting point for DFA is the inverse of the pooled variance-covariance matrix within *a priori* groups. This inverse can only exist when the matrix is non-singular, i.e., its determinant is other than zero, which implies that it is of full rank (Dixon, 1975; MacFie, *et al*., 1978); i.e.

Generally if:

$$(N_s - N_g - 1) > N_v \qquad \text{Equation 1}$$

where N_s = Number of samples
N_g = Number of groups
N_v = Number of inputs (variables); i.e., mass intensities, absorbances at particular wavenumbers, or photon counts at particular wavenumber shifts for MS, FT-IR and Raman respectively.

Singularity can be caused by collinearity, and PCA removes collinearities whilst also reducing the number of inputs (so as to obey the above) to the DFA algorithm.

Finally, the Euclidean distance between *a priori* group centres in DFA space is used to construct a similarity measure, with the Gower similarity coefficient S_G (Gower, 1966), and these distance measures were then processed by an agglomerative clustering algorithm (OCNT.EXE) to construct a dendrogram (Manly, 1994).

Some points to remember

The MVA process: Spectra → normalisation → PCA → DFA → HCA

There are however a few decisions to be made:

1. Spectra → normalisation - what pre-processing regime do I use?
2. PCA → DFA – how many PCs do I feed into DFA?
3. DFA → HCA – how many DFs do I feed into HFA?

These are not easy questions to answer and this is where the trial and error comes in.

1. As mentioned in the pre-processing part, all the methods have advantages and disadvantages so play with these. Using Savitzky & Golay derivatives is popular but as well as removing any baseline artifacts in the spectra they will also introduce noise.
2. Eventually you will be able to get this bit right based in part on what the clusters in DFA look like. Too many PCs into DFA and the clusters are too tight and artificial, too few PCs and no clusters relating to the *a priori* groups are evident. A good starting point to the uninitiated is to run all samples 6 times. These can be called two groups for DFA and if you are over doing the PCA extraction then you will see the two groups from nominally identical material separate.
3. Likewise for doing HCA, although this should be easier since the DFs are weighted according to their eigenvalues on input to the construction algorithm for the similarity matrix.

References

Causton, D. R. (1987). *A Biologist's Advanced Mathematics*. London: Allen and Unwin.

Dixon, W. J. (1975). *Biomedical Computer Programs*. Los Angeles: University of California Press.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis. *Biometrika* **53**, 325-338.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

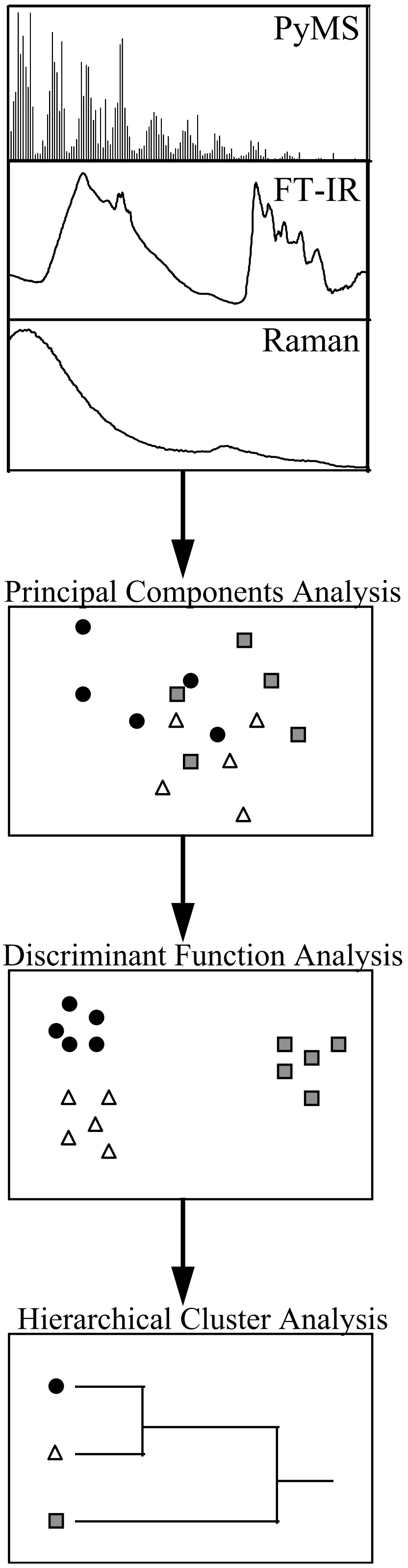
MacFie, H. J. H., Gutteridge, C. S. & Norris, J. R. (1978). Use of canonical variates in differentiation of bacteria by pyrolysis gas-liquid chromatography. *Journal of General Microbiology* **104**, 67-74.

Manly, B. F. J. (1994). *Multivariate Statistical Methods : A Primer*. London: Chapman & Hall.

Windig, W., Haverkamp, J. & Kistemaker, P. G. (1983). Interpretation of sets of pyrolysis mass spectra by discriminant analysis and graphical rotation. *Analytical Chemistry* **55**, 81-88.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, pp. 391-420. Edited by K. R. Krishnaiah. New York: Academic Press.

FLOWCHART OF UNSUPERVISED LEARNING MULTIVARIATE ANALYSIS USED TO CLUSTER THE HIGH DIMENSIONAL SPECTRA.


PyMS
FT-IR
Raman
Principal Components Analysis
Discriminant Function Analysis
Hierarchical Cluster Analysis

Spectra are high dimensional:

- 150 masses from PyMS
- 882 wavenumbers from FT-IR
- 2283 wavenumbers from Raman

PCA transforms the original set of variables to a new set of uncorrelated variables called PCs. PCA is a data reduction process and the first few PCs will typically account for >95% variance.

DFA has *a priori* information based on spectral replicates and uses this to minimise within group variance and maximise between group variance.

A similarity matrix can be constructed from the DFA space. HCA can then use this to produce a dendrogram, using average linkage clustering.