Sorting the wheat from the chaff

Tony M.C. Davies

Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK

Tom Fearn

Department of Statistical Science, University College London, Gower Street, London

Some of you will know that I am a fan of the artist M.C. Escher[a] (it has no connection with my middle initials; that is just a happy coincidence!). Authors of papers submitted to the *Journal of Near Infrared Spectroscopy* receive a postcard of one of his famous pictures if their paper is accepted for publication. I mention Escher because I am not able to draw pictures of impossible objects but I need you to imagine them from my figures.

Figure 1(a) is a plot of a near infrared spectrum. We know that this spectrum will contain information about our analyte of interest but it will also contain information about other constituents and physical effects (e.g. particle size). It will also contain noise from the instrument and the environment. In Figure 1(b) I have divided up the area of the plot and coloured the segments, red for noise, yellow for unrelated information and green for related information. This is very diagrammatic; you have to imagine three sets of coefficients which would relate the absorption to the different information at any point in the spectrum. If we had such a diagram then life would be much easier. We would set all the red and yellow coefficients to zero and be left with information which was only related to the analyte of interest!

Unfortunately we do not have Figure 1(b), what we have is indicated in Figure 1(c); all the information wanted and unwanted and the noise are distributed in an unknown manner. This column is about the progress of chemometrics in getting from Figure 1(c) to the stage after Figure 1(b), having only the related information.

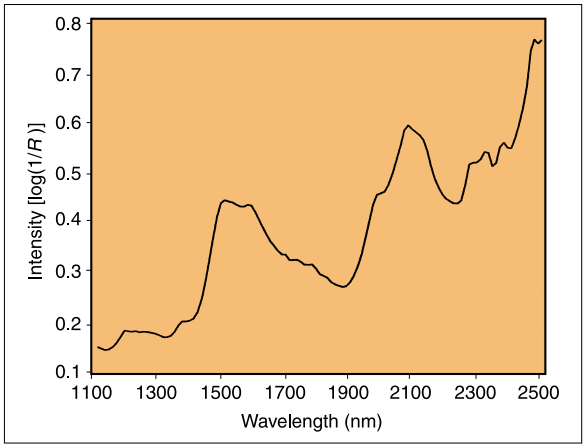

Figure 1(a). An NIR spectrum.

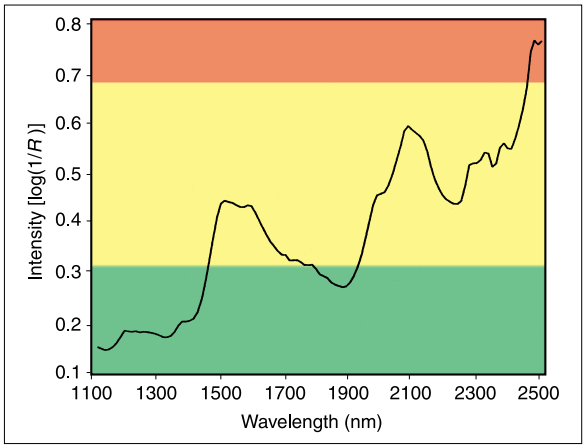

Figure 1(b). Information content (hypothetical).

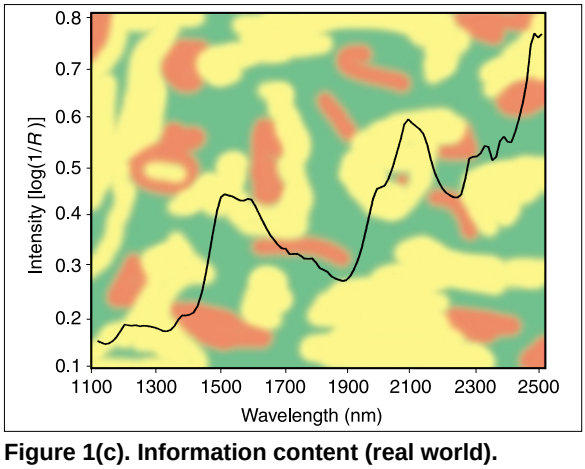

Figure 1(c). Information content (real world).

Noise reduction

Noise reduction can be achieved on a spectrum or a sample set basis. The spectral methods pre-date chemometrics; these include smoothing by averaging, Savitzky–Golay and Fourier. Savitzky–Golay fits a series of polynomials to the data and then uses the data computed from the curves. Fourier removes high frequency noise by computing a Fourier transformation and setting a large proportion of the higher frequency coefficients to zero and then retransforming. The simple moving average is by far the most popular. The sample set methods include the well-known principal component analysis regression (PCR) and partial least squares (PLS) methods. Both methods limit the regression to a few terms and this will help to lose some of the noise.

Removing unrelated information

Almost all the popular pre-processing methods have been devised to remove particle size (and other) effects. These

[a]Anyone who is not familiar with Escher should have a look at http://www.mcescher.com/.

include: derivatives, multiple scatter correction (MSC), standard normal variate (SNV), optimised scaling (OS) and orthogonal signal correction (OSC). These are all described in a recent book.[1] OSC is one of the most recent methods and is the one that prompted this column. It was original devised[2] as a pre-processing method which could have been followed by any regression method, but it is very well suited to PLS. In the latest development it has been incorporated with PLS and is known as O2-PLS.[3]

It should perhaps be emphasised that PCR and PLS methods cannot avoid including some unrelated data. When originally proposed by Ian Cowe *et al.*,[4] PCR was based on the inclusion of only those PCs which were correlated with the analytical (*y*) data, however, some quite low correlations were allowed and in most PCR programs this criterion has been forgotten. Although PLS attempts to form factors that are correlated with the *y*-data it is also influenced by very large reductions in the spectral variation so it too includes some unrelated data.

The idea of OSC is to find factors which explain variance in the spectral data but have very low correlation to the *y*-data. These factors are then subtracted from the data. This is usually restricted to one or two factors. The program computes corrected spectra and these corrected spectra are then used in the chosen regression method. In O2-PLS the two stages have been combined. I am not going to describe the matrix algebra (anyone interested should read Tom Fearn's Chemometric Space column in *NIR news*[5]) but the important advantage of the method is that it may provide more readily understandable plots of the OSC weights as well as (perhaps) more understandable (recognisable) plots of the PLS factors.

We can, of course, also plot the noise by calculating differences between the original and smoothed data. So now we can see the red, yellow and green data in our spectra. Chemometrics for the ordinary spectroscopist is about looking at pictures and understanding the data (I may have mentioned this before in previous columns!).

Anything that moves us away from the "Black-box" approach is to be welcomed!

Acknowledgement

I have included Tom Fearn as an author because without him I would not understand OSC or O2-PLS, but in fact this column has been written by TD, and TF may wish to publish a correction at a later date!

References

1. T. Næs, T. Isaksson, T. Fearn and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester, Chapter 10 (2002).
2. S. Wold, H. Antti, F. Lindgren and J. Öhman, *Chemometrics and Intelligent Laboratory Systems* **44,** 175-185 (1998).
3. J. Trygg, *Parsimonious Multivariate Models*, Doctoral Thesis, University of Umeå (2001).
4. I.A. Cowe and J.W. McNicol, *Appl. Spectrosc.* **39,** 257 (1985).
5. T. Fearn, *NIR news* **13(2),** 15 (2002).