TONY DAVIES COLUMN

Back to basics: running your first PLS calibration

A.M.C. Davies
Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK. E-mail: td@nnirc.co.uk

Introduction

You have worked really hard, doing the things I told you about in the last column[1] and now the day has arrived at last (no not Christmas!); the day of your first partial least squares (PLS) calibration.

You have collected samples, run the NIR spectroscopy and got the lab analysis. These are pharmaceutical tablets containing an unidentified (secret) active ingredient code-named "assay". You have a "Test set" (T1) containing 460 tablets and a "Calibration set" (C1) of 155 samples. Each of these tablets has been measured on an NIR spectrometer (Foss/NIRSystems Multitab spectrometer) and their spectra recorded over the range 600–1898 nm at 2 nm intervals. The tablets were weighed after which they were sent to the laboratory for the reference analysis of "assay" mg/tablet. In reality these data were used for the 2002 "Software Shootout" at the International Diffuse Reflection Conference, Chambersburg, USA and are still available on the IDRC website[2] and an article providing a detailed analysis of the problem (which is larger than suggested by this article) has been published by the competition winner, David Hopkins.[3] I am trying to avoid causing confusion by retaining the original names of the files but for this article we are going to use the T1 set for calibration and the C1 set for validation. I am working with Unscrambler 9.6 but other packages should produce similar if not identical results. If you have difficulty in finding out how to perform operations in Unscrambler you are welcome to e-mail me for additional directions. If you are using other packages you will have to ask for assistance from their source.

There is one other complication with these data, because it is the analysis of individual tablets. Patients take tablets, so the amount per tablet of the active ingredient is the important criterion. Spectroscopy measures concentrations, so we need to calibrate against a concentration parameter and then convert the results back to "assay". We have "assay" and weight variables so we can calculate a "Conc" variable as:

$$\text{Conc} = 100 \times \text{"Assay"} / \text{weight of tablet} \quad (1)$$

This can be done within Unscrambler so we have a new variable called "Conc".

When we have the results it is easy to rearrange Equation (1) as :

$$\text{"Assay"} = \text{Conc} \times \text{weight of tablet} / 100 \quad (2)$$

But we have to export the results and weights to Excel™ and do the calculation there.

Getting started

What to do first? Two things, both equally important; look at the spectra and use principal component analysis (PCA) to check that T1 and C1 occupy the same dimensional space. You have, of course, been checking each spectrum as it was recorded but now we need to see them together; Figure 1 is a plot of all 615 spectra. This plot shows several things. First, there are no gross outliers; second, the spectra are noisy above about 1800 nm; and third, there are a few spectra with low absorption between 600 and 1000 nm which may be un-typical. With experience you can also look at these plots and think that some form of data pre-treatment would be beneficial. More about that latter. The PCA plot is shown in Figure 2. The two sets are identified and this demonstrates that they occupy approximately the same space within the first two principal components, which represents 91% of the total variance. The next two PCs were even less interesting. There is some evidence that the data is bimodal but as both sets are similar this is not an immediate problem. If there was one really bad outlier it would have a very marked effect on the PCA and that is our first concern. Our next

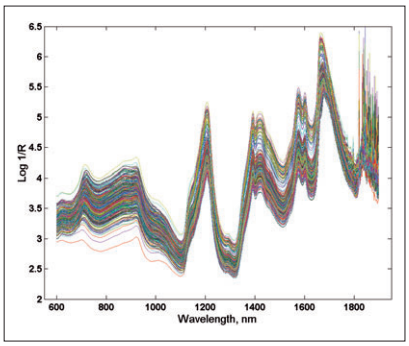

Figure 1. Plot of the 615 spectra in sets T1 and C1.

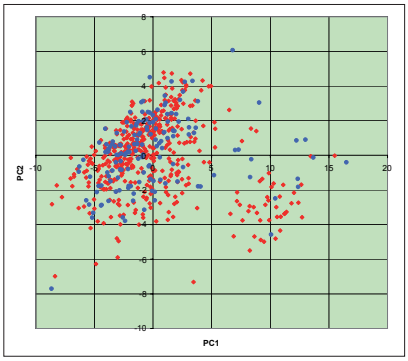

Figure 2. PCA on the calibration data. A plot of 615 samples on PC1 and PC2; red diamonds, set T1; blue ellipses, set C1.

task is to break the training set into two; a calibration set and a factor determination set. If T1 and C1 had appeared to be different then we would have had to combine them and then make three sets. The usual method of dividing a set into two is to sort the set on the analyte of interest and then put alternate samples into the two sets. This is what I did to produce sets T1O and T1E (odd and even). Having done this we are ready to run PLS.

The calibration phase

Having got this far, the next part is very easy; we just tell the computer to get on with it! You tell it where to find the data, what range of variables you want to use and which is the dependent analyte. You could take some time wondering what range of variables to use, you could be considering if you should first do a data pre-treatment. I prefer to see if there are any serious problems that will not be solved by these refinements. If your program allows you to select the number of factors to compute, I would set it at 10 for the first trial. For some applications you might require more for the optimum solution, but usually 10 is more than you need. The program will tell you what it regards as optimum. So tell the program where to find the data, use the full range of variables, select "Conc" as the dependent variable and the test set for "validating" the calibration. This is the terminology used by Unscrambler but it should be called "factor determination". You will need to determine how many factors should be used by the calibration then you can validate the calibration using the validation set (C1).

It doesn't take long for the computer to calculate the calibration. This was not always the case! You may remember that a few issues ago Tom Fearn and I estimated[4] that a modern PC was some 100,000 times faster than the computers we used in the early 1980s. At first we might start it running a calibration before we went home in the evening because it would take several hours, a few years later there was still time for a leisurely cup of coffee. I mention this because I want you to realise that today you can and should make use of the computer power and run a calibration several times. The only thing you need to be careful about is running the validation set too often. This set must be reserved until you think you have a done your best or the best for a particular treatment. If you use the validation set frequently you are using it to guide the calibration and it will no longer be an independent set.

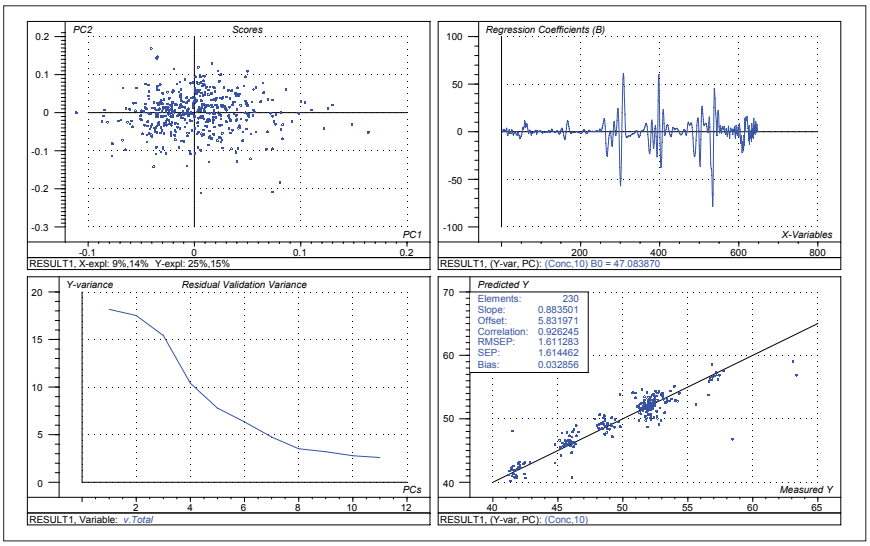

Figure 3. Overview of first trial calibration.

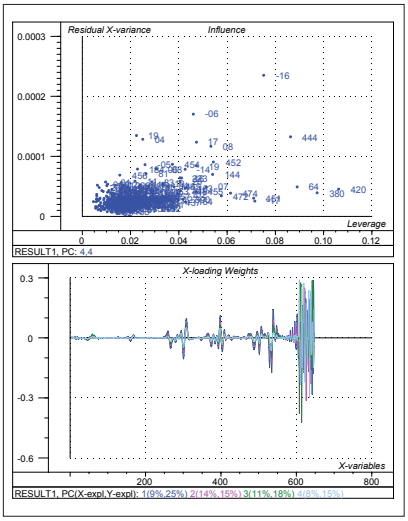

Figure 4. Influence plot and loadings plot.

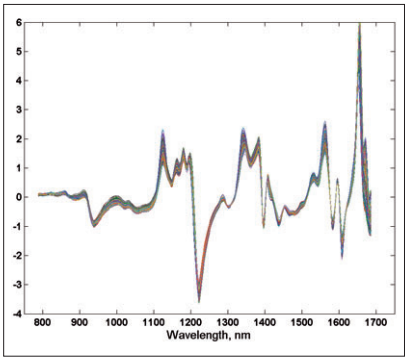

Figure 5. Plot of the T1 spectra after data pre-treatment.

Now the calibration has run and we have many things to consider, most of these come in the form of plots: Unscrambler provides an "overview" of four plots, Figure 3. These are not necessarily the most important views. I discussed these plots sometime ago but the articles are still available from the *SE* website.[5,6] From these plots we can see in the scores plot (top left) that it looks symmetrical with no gross outliers, the regression coefficients (top right) look noisy above variables greater than 600, the residual variance (bottom left) is still quite large and that there are some serious outliers in the predicted results of the validation set (bottom right). If we then look at two other important plots in Figure 4; the influence plot (top) shows that there are a few samples having a large influence on the calibration and we might need to look at them, and the X-loadings plot shows us that there is a serious noise problem. The noisy loadings plot tell us that we need to remove these noisy data by restricting the wavelength range of the input variables. From the inspection of this plot and the original spectra (Figure 1), I decided to use the

TONY DAVIES COLUMN

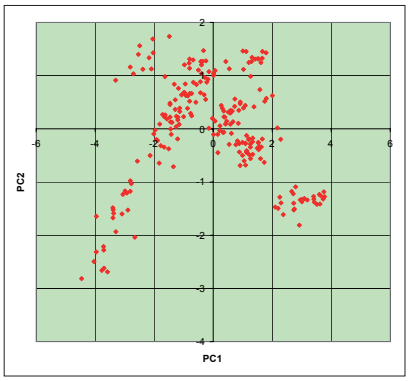

Figure 6. Plot of PC1 and PC2 for the T1 sample set after data pre-treatment.

range from 788 to 1686 nm. The residual variance plot shows a slow progress, we expect to see a much more rapid reduction in the variance and the most likely cause is that the PLS is finding it difficult to compensate for the variation in scatter. We should do some pre-treatment to try to reduce the scattering problem.

I will discuss data pre-treatment in detail in the next column but for the present we will use a fairly standard combination of second derivative (2d) followed by standard normal variates (SNV).[7] 2d will remove offsets and sloping baselines while SNV removes offsets and multiplicative effects from spectra. The effect of these transformations is seen in the plot of the spectra, Figure 5 and the PCA scatter plot, Figure 6. The spectra show only localised variation while the PCA now shows much more structure revealing the presence of several clusters. (These are the result of the way the set was constructed as a mixture of different batches of tablets.) Now we run the PLS calibration again but with the transformed data. Figures 7 and 8 show that we have improved the calibration. The scores plot shows the same sort of structure that we saw in the PCA, the regression coefficients and the x-loadings appear well defined. The influence plot still shows that there are some high influence spectra and also a few outliers but the plot of measured against predicted shows a large improvement, the *RMSEP* has decreased from 1.71 (Figure 3) using 10 factors to 0.71 (Figure 7) using only three factors, as recommended by the program.

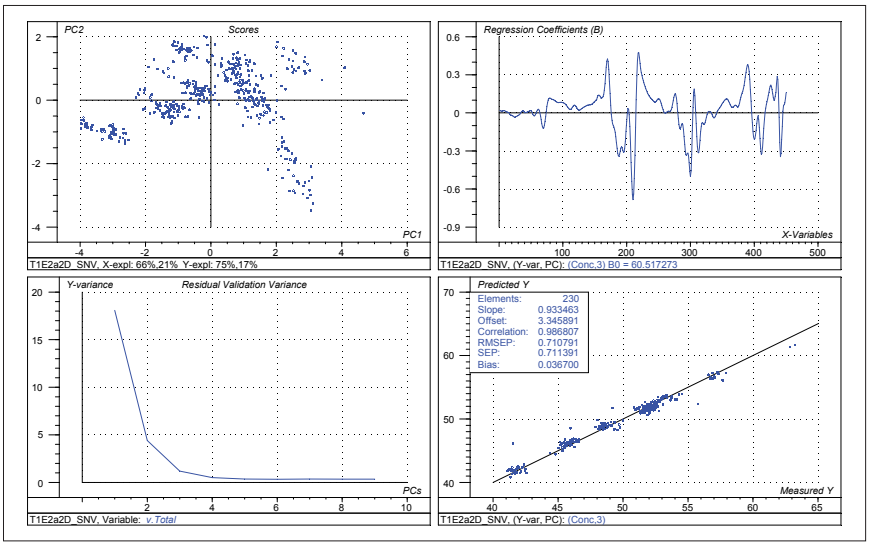

Figure 7. Calibration overview after spectral pre-treatment.

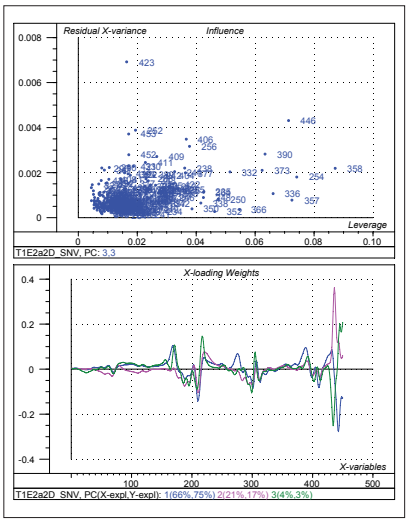

Figure 8. Influence plot and loadings plots after spectral pre-treatments.

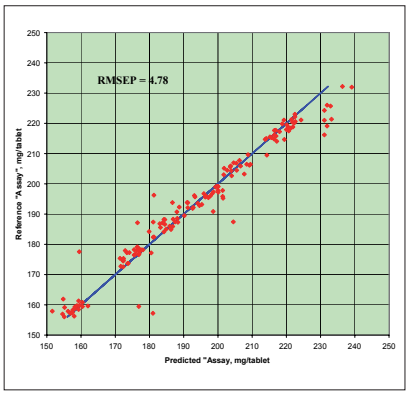

Figure 9. First (potential) result for "Assay" in the validation (C1) set of tablets.

Interim conclusion

It is time to calculate our final calibration using all the T1 data and cross-validation and selecting the three-factor solution. We save this calibration, use it to predict C1 and export the "Conc" results with the tablet weights to Excel™ to convert them to "Assay" using Equation (2). The result in Figure 9, a *RMSEP* of 4.78 mg/tablet, is our first potential result. In the next column we will see if other pre-treatments can produce a superior answer.

Acknowledgements

I would like to acknowledge that these data were created and put into the public domain by Gary Ritchie and colleagues.

I am, as usual, grateful to Tom Fearn for his guidance and also to Ian Cowe for his help with some of the complexities of Unscrambler that I had forgotten.

References

1. A.M.C. Davies, *Spectrosc. Europe* **18(4),** 23 (2006).
2. www.idrc-chambersburg.org/shootout_2002.htm
3. D.W. Hopkins, *NIR news* **14(5),** 10 (2003).
4. A.M.C. Davies and T. Fearn, *Spectrosc. Europe* **14(6),** 24 (2002).
5. A.M.C. Davies, *Spectrosc. Europe* **10(4),** 28 (1998).
6. A.M.C. Davies, *Spectrosc. Europe* **10(6),** 20 (1998).
7. R.J. Barnes, M.S. Dhanona and S.J. Lister, *Appl. Spectrosc.* **43,** 772 (1989).