Back to basics: preparing for PLS calibration

A.M.C. Davies
Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK

Introduction

It is not profitable to wake up one morning and say "Today I will run a partial least squares (PLS) calibration", **unless** a considerable amount of planning and work has already been done. This column is about the required preliminary work.

Programs

For many people their first PLS calibration involves the calibration of near infrared (NIR) spectroscopic data to make predictions of an analyte for which there is a (slow and expensive) reference method. There are many other uses of PLS but this discussion will be confined to this more common application. If this is your situation it is quite likely that you have an NIR spectrometer which has been successfully performing analyses using calibrations provided by the spectrometer manufacturer or some third party. Then your boss or friend(?) suggests that you should develop a calibration for some new analyte. By "new" I mean that a calibration is not available but there is good reason to believe that it should be possible to develop one.

The first thing you are going to worry about is a PLS program. If you have a modern NIR system it is very likely that it will have been provided with a suite of programs including PLS. If not you will have to buy one. I cannot suggest a "best" program; I can tell you that I have been using Unscrambler™ for many years but there are others which are as good if not better. The variations are in the ease of use, the plots that can be obtained etc. I hope that all programs are mathematically sound. In addition to PLS you should also have a principal component analysis (PCA) program.

Most programs will come with some test data; there are also a couple of nice examples that have been published[1,2] which derive from the "Software Shootout" at the IDRC at Chambersburg, USA (where I will be when this goes to the printers!). Use these to practice with your software and see that you can get similar results.

Spectroscopy

With a new type of sample you will have to determine the best way of presenting the sample to the spectrometer and make sure that it is reproducible. You should have a quality control procedure for the spectrometer. It will probably take some time to obtain and scan all the samples that you will need for the calibration; so it is very important that the spectrometer is correctly maintained over the calibration period. I learned the importance of this early in my NIR career. While I was stockpiling samples for a calibration the spectrometer was being used for other work. At some point the spectrometer window was contaminated and not cleaned. The problem was not discovered until an assistant decide the instrument needed a good clean and our spectra changed!

Samples

One of the most common questions is "How many samples do I need for a PLS calibration"? I tried to answer it some time ago;[3] we need to know the answer to several other questions before it can be answered. Some people think that they can use samples which they have in store. These are OK if you want the calibration to work on stored samples but if your calibration is going to be applied to "fresh" samples then that is what must be used to generate it. The samples should be scanned on the spectrometer and then sent for reference analysis. Much has been written and discussed about reference analysis. It is possible for PLS results to be better than the reference method but it can be difficult to prove. You need the best analysis that money can buy! Some ten years ago Tom Fearn and I[4] pointed out the danger of fooling yourself by the mis-use of duplicate analysis. It is a quite widely held belief that if a pair of results are outside a given tolerance then the analysis should be repeated until a pair of results are within the tolerance; this is not correct. A good way of monitoring analytical results is to add ten replicate (but differently coded) samples in the batch of samples sent for analysis. These ten results give you a much better idea of the quality of the analysis and the sort of error that you might approach with a PLS calibration. Sometimes you may have many more samples that you can afford to analyse and you would like to be able to make a selection that would contain a wide range of analyte. Your computer and the NIR spectra can help you but do use it in the correct manner; as detailed in that previous TD column.[3] These are known as sample selection programs and they can find the spectroscopically most diverse set of samples. The important thing is to ask it for the total number of samples and then randomly distribute them among the different sets.

You should also use a PCA program to check for outliers in your spectroscopic data. Figure 1 is a scatter plot of the first two principal components (PCs) of the data used in Susan Foulk's article. It is obvious that sample 9 is a long way from the rest of the samples and this might remind you that you should **always** look at the spectra (Chemometrics is about using your chemical and spectroscopic knowledge in conjunction with mathematics). Figure 2 shows the change in the scatter plot with sample 9 excluded. Sample 10 is now on the edge of the population but I would not judge it to be an outlier. You only need to remove samples that have an obvious problem.

Sample sets

For a PLS calibration you should have three sets of samples containing approximately the same number of samples and with similar variation in the analyte in each set. The three sets are called calibration, factor selection and validation sets. The specification for these sets is that they should be: Wide, Even, Precise and Typical (WEPT). The sets need to be wide to give the regression program a good chance of "seeing" the analyte of interest. Many years ago the NIR team at the Plant Breeding Institute, Cambridge (who were among the first people to employ NIR technology in the UK) suggested a range-error-ratio (RER) statistic[5] where range is the range of the analyte concentrations and "error" is the standard error of the reference analysis. They recommended that the RER should be greater than 10. Ideally samples should be evenly distributed in the sets but there are two cautions that must be given. Tom Fearn studied the problem[6] and recommended that you should not discard samples in order to improve the look of the set; it is generally more beneficial to have a larger number of samples in the regression than a flat distribution. Sometimes the natural distribution of a set of samples is not the typical bell shaped normal distribution but is bimodal—having two maxima. It has been known for people to "fill in" the missing samples by making mixtures of samples from the two sides of the distribution. The reason why you should NOT do this is that this does not conform to the final requirement of being typical. It also means that you should not spike samples with the analyte to make some additional high concentration samples.

When you run PLS you should use the calibration set to obtain a trial calibration and use the factor selection set to determine the optimum number of factors. Then you can combine the calibration and factor selection sets to make a final calibration using the selected number of factors. In this way the final calibration will contain about twice as many samples as you use for the validation, which is the distribution that many experts recommend.

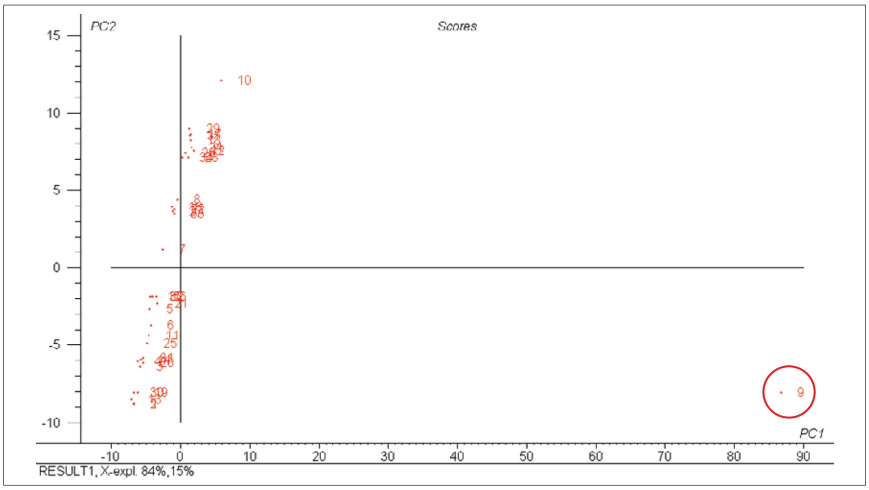

Figure 1. PCA scatter plot of scores for the first two PCs from SH-96 data.

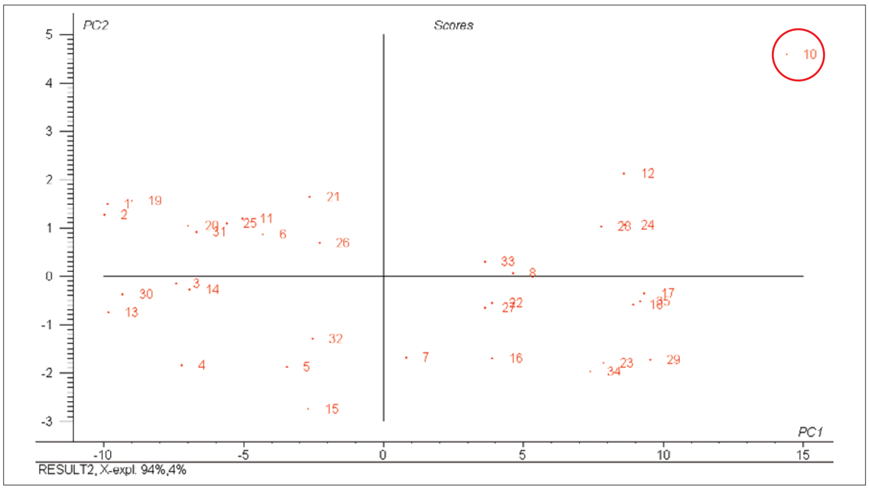

Figure 2. PCA scatter plot of scores for the first two PCs from SH-96 data after removal of sample 9 in Figure 1.

Conclusion

Next time we will really do the PLS calibration; by which time you should have sufficient samples. How many? I didn't tell you! At least one hundred, although you might be able to make a useful trial calibration with as few as thirty. Commercial calibrations may be based on several thousand samples.

References

1. Susan J. Foulk, *Spectroscopy Europe* **9(2),** 25 (1997).
2. D.W. Hopkins, *NIR news* **14(5),** 10 (2003).
3. A.M.C. Davies, *Spectroscopy Europe* **8(4),** 27 (1996).
4. A.M.C. Davies and T. Fearn, *Spectroscopy Europe* **8(2),** 36 (1996).
5. C. Starr, A.G. Morgan and D.B. Smith, *J. Agric. Sci., Cambs.* **97,** 107 (1981).
6. T. Fearn, in *Near Infra-Red Spectroscopy: Bridging the Gap between Data Analysis and NIR Applications*, Ed by K.I. Hildrum, T. Isaksson, T. Næs and A. Tandberg. Ellis Horwood Limited, Chichester, p. 61 (1992).