

# Cross-validation: do we love it too much?

A.M.C. Davies

Norwich NIR Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK.

I have come across several uses of cross-validation (CV) in recent weeks which has made me question why people continue to apply this technique to inappropriate data. Then I realised that although this column has often referred to CV we have never really described it, so I decided that I should rectify this omission as soon as possible.

## What is CV?

CV is a cunning technique to make up for shortage of data. It allows you to test a calibration without a set of validation samples. You take out one of

your calibration samples, do the calibration on the remaining samples then use that calibration to predict the sample you left out. This is an independent sample so you have a calibration, which has been validated by one sample. Not very useful? Wait, this is where you can have your cake and eat it (if you remember the Data Cake!). You can repeat the exercise leaving out another sample and getting another prediction and you (or rather your friendly PC) can go on repeating this until all the samples have been left out and predicted (see box to admire the simplicity of the idea). CV is often known as "Leave-One-Out Method".

## What is it used for?

CV is used whenever there are insufficient samples to form a proper validation set. I first met it in discriminant analysis where it is very popular. If you have many groups to discriminate it can be very hard work to obtain the samples for the calibration (training) set and CV saves having to do it all over again in order to obtain a validation set. The popularity of CV has become much greater in recent years because of PLS calibration which requires three sets of samples. CV is often used to save having a factor determination set.

## When to use it

In a PLS calibration the standard error of calibration (SEC) will continue to reduce the more factors are included, so we need a set of samples to determine the standard error of prediction (or performance) (SEP). After a given number of factors the SEP will begin to rise and this indicates how many factors should be retained in the calibration—usually one less than the number that gives the minimum SEP. This set of samples is now part of the calibration and cannot be correctly used to assess the performance of the calibration. So another independent validation set is required. CV is frequently used to save having to find this

### A Cross-Validation Program

```
Given n samples:
Let k = 0
For k = 1 to n
  k = k + 1
  Remove sample k
  Calibrate on n - 1
  samples
  Predict sample k
  Save result
  Replace sample k
Next k
Calculate statistics on
  saved results
```

extra set, by using cross-validation on the calibration set to determine the number of factors. If the calibration set has been well chosen, then this is a reasonable course of action.

As mentioned previously, CV is often used in qualitative analysis, such as canonical variates analysis (CVA) when obtaining sufficient numbers of authentic samples can be very difficult. However, it is not without dangers. If the CV results are much worse than the calibration results, then be warned. This is telling you that the model is unstable so that the absence of a single sample results in a significantly different model.

It is also reasonable to use CV in other regression methods (MLR, PCR) for preliminary tests but all regression methods need to be tested with truly independent samples before too much confidence is placed on that calibration.

## How to use it

Using an CV program is usually very simple, you tell the computer to do it and go for a cup of coffee. There is one complication, which leads to what I regard as an unsatisfactory application of CV. The complication is the existence of duplicate samples in the calibration set. If we were to do CV with a duplicate sample present it is obvious that we no longer have an independent sample. So all replicates of a sample must be removed at the same time and most CV programs have ways of han-

dling the problem. An extension of this procedure was introduced to save on computing time by taking out several different unrelated samples at the same time. This is done by dividing the calibration set into a number of blocks and then leaving out a complete block, while calibrating on the remainder. This might have been justified a few years ago when CV did take a long time to run but PCs are very much faster now. As far as I am aware there is no EC directive limiting the number of hours worked by a PC. If it needs to be left running over night, that's OK; we always used to run computers over night! [The authors of a paper in the *Journal of Near Infrared Spectroscopy*<sup>2</sup> estimated that the work they were reporting had taken more than 170 days (yes, days) of processing on 486PCs—that is what I call making PCs earn their keep!] If you want to leave out blocks then you have to be very careful about the structure of the data set; you still need to make sure that all duplicates are in the same block.

If CV really is taking a long time then you probably have sufficient number of samples to form calibration and validation sets. You will get a faster and safer calibration if you split the data into calibration and validation sets. When you report results from CV remember to say that, the results were obtained by CV and call the SEP an SECV.

## When not to use it

Judging from some of the papers I see, there is a belief that CV is superior to a normal validation with an independent set. ("All that work must make for a better estimate of the model".) CV is never better than a properly collected validation set. The problem with CV is that the samples are not truly independent. The very fact that they were collected to form a set of samples means that they are related in some way. Even worse; there may be unknown duplicates in the set, the same sample from different sources but if you do not know this you will not be able to remove all the duplicates at the same time.

The only time we should use CV is when we are short of samples. This applies to the number of factors step in PLS. As explained in a previous article<sup>3</sup> once the number of factors has been determined, all the samples used for calibration and factor determination can be combined into a "super" calibration set.

## Conclusions

CV is a useful technique but it should be seen as a stepping stone to justify further work rather than being used as the final justification for proposing the use of a chemometric method.

## References

1. A.M.C. Davies, *Spectroscopy World* **2(1)**, 35 (1990).
2. G. Sinnaeve, P. Dardenne and R. Agneessens, *J. Near Infrared Spectrosc.* **2**, 163–175 (1994).

3. A.M.C. Davies, *Spectroscopy Europe* **8(4)**, 27 (1996).

## Footnote to the footnote in *Spectroscopy Europe* 9/6

My apologies to those who were confused by my statement that the square root of  $10^{100}$  was  $10^{10!}$ . It is (of course)  $10^{50}$ . I think the error may have been due to mega-myriad-phobia but do not blame Tom Fearn. I confess to writing the footnote as an entertainment.