TONY DAVIES COLUMN

An evergreen problem in multivariate calibration

N.M. Faber[a,*] and R. Rajkó[b]
[a]Chemometry Consultancy, Rubensstraat 7, 6717 VD Ede, The Netherlands. E-mail: nmf@chemometry.com
[b]Department of Unit Operations and Food Engineering, Szeged College of Food Engineering, University of Szeged, H-6701 Szeged, POB 433, Hungary

Over recent months we have been in discussion with Klaas Faber about an article for the column looking into the issues around over-fitting multivariate calibration models. I hope their ideas expressed in this article will lead to some more discussion on this subject!—A.N.D.

Background

Multivariate calibration models are important in many fields. They apply in the chemical, petrochemical, pharmaceutical, cosmetic, colouring, plastics, paper, rubber and foodstuffs industries, as well as in forensic, environmental, medical, sensory and marketing research. Various methods have been developed for building multivariate calibration models of which partial least squares (PLS) regression is currently the *de facto* standard in applied work. The following discussions are equally relevant for other multivariate calibration methods.

The first step towards constructing a PLS model is to remove undesirable features from the X-data by pre-treatment techniques such as filtering[1] or differentiation.[2] The next critical step serves to select the optimum model rank, which is the number of PLS components that constitute the multivariate model. The state of the art concerning commercially available software has been criticised by A.N. Davies in this publication:[3]
"Back in 1998 more advanced chemometric tools were being made available as standard in spectrometer control packages. This had, however, raised fears that the *inherent dangers of over-fitting data were not being sufficiently addressed* in order to help inexperienced spectroscopists handle the additional computing power that was becoming available. I must admit that the work of my co-column Editor in pushing for "Good Chemometrics Practice" has hopefully raised awareness in the community of the potential pitfalls in using these packages without due consideration, but *I personally have not been aware of clear unambiguous automated warnings starting to appear when data was being over-fitted.*" (Our italics.)

Over-fitting causes harm because one not only incorporates predictive features of the data in the model, but also noise. The implication is degraded model performance in the prediction stage.

Many methods have been developed to tackle over-fitting, of which model validation is the most frequently applied one in practice. In the context of multivariate calibration, validation amounts to assessing the ability of a model to predict the property of interest for future samples. This can be performed in two essentially different modes, namely *externally* and *internally*. "External" refers to the requirement that the validation samples be independent of the samples used for constructing the model, i.e. the calibration set; otherwise one does not properly assess the ability to predict for truly unknown future samples. The predictive ability is estimated by applying the model to these (independent) validation samples and averaging the squared prediction errors, i.e. the differences between model prediction and the associated "known" reference value. The square root of this average squared error is known as the root mean squared error of prediction (*RMSEP*). In equation form, for increasing number of components (A),

$$RMSEP(A)=\sqrt{\frac{1}{N_{val}}\sum_{n}^{N_{val}}\left(\hat{Y}_{A,n}-Y_{ref,n}\right)^2} \quad (1)$$

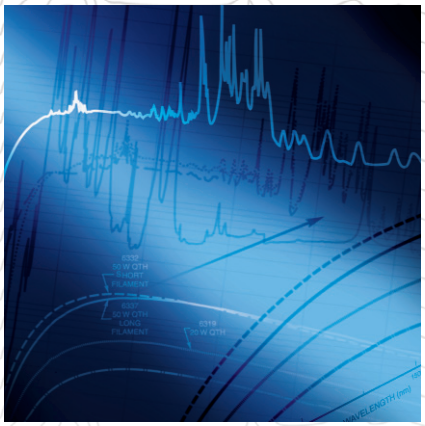
where N_{val} is the number of validation samples and $\hat{Y}_{A,n}$ and $Y_{ref,n}$ denote the model prediction with A components and "known" reference value for sample n ($n=1, \ldots, N_{val}$), respectively. Ideally, the results of this calculation lead to a clear (i.e. not too broad and shallow) minimum *RMSEP* for the optimum model rank.

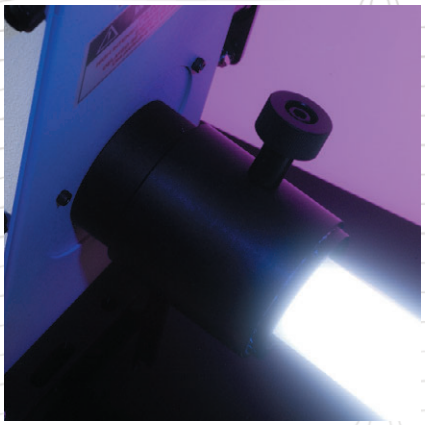
Internal validation differs in the sense that the validation samples are taken from the calibration set itself, i.e. the validation samples are not truly independent. To execute an internal validation, one has the choice between (1) cross-validation, (2) bootstrapping and (3) leverage correction. In cross-validation, one constructs models after judiciously leaving out segments of calibration samples. Then an estimate of *RMSEP* follows by averaging squared prediction errors for the left-out samples, as in external validation. To emphasise that this estimate is not based on independent validation samples, it will be denoted as root mean squared error of cross-validation (*RMSECV*) in the remainder of this paper. Bootstrapping performs similarly to cross-validation,[4,5] whereas leverage correction is only a "quick and dirty" alternative when applied to PLS.[6] In the remainder we will therefore focus on internal validation using cross-validation.

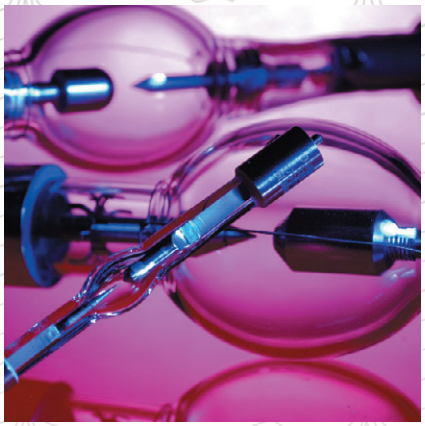
As we have discussed in the past, conventional validation-based component selection is problematic for various general and specific reasons which have been well documented.[7–10]

Want Oriel? Think Newport!

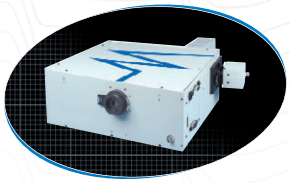
Direct, Rapid, Reliable.



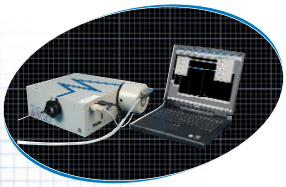




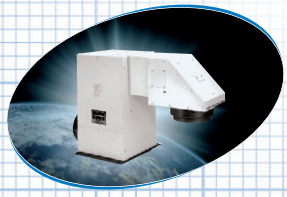
Light Sources



Monochromators & Spectrographs



Spectrometers



Solar Simulators

Newport is your one-stop source for Oriel products. We now supply Oriel products in Europe, directly.

With Newport's acquisition of Spectra-Physics, Oriel Instruments has joined the Newport family. You have come to know and rely upon Oriel's 40 years of industry experience and comprehensive offering of light sources and spectroscopic instruments. As part of Newport, Oriel is now stronger than ever. We look forward to supporting and servicing you directly.

Our global team is prepared to assist you at every level. No other company has the combined experience, service resources, expertise and breadth of technology. And no one else can meet all your needs from a single source.

For all your future Oriel needs, please call your local Newport sales office, and visit **www.newport.com/oriel** to see Newport's complete Oriel product offering.

FASTLINK / CIRCLE 014 FOR FURTHER INFORMATION

Newport®
Experience | Solutions

Spectra-Physics
A Division of Newport Corporation

Belgium
Tel: +32-(0)16 402927

France
Tel: +33 (0)1.60.91.68.68

Germany
Tel: +49 (0)61 51 / 708-0

Italy
Tel: +39 (0)2/92.90.921

Netherlands
Tel: +31-(0)30 6592111

United Kingdom
Tel: +44 (0)1635.521757

In 2004 Spectra-Physics including Oriel Instruments, Richardson Gratings, Corion Filters, Opticon Mirrors and Hilger Crystals was acquired by Newport.

MAKE LIGHT | MANAGE LIGHT | MEASURE LIGHT

AD-090607-EN

TONY DAVIES COLUMN

A novel approach

Wiklund *et al.*[11] suggested assessing the statistical significance of *each* individual component that enters the model. Theoretical approaches (using a *t*- or *F*-test) have been put forward but they are all based on unrealistic assumptions about the data, e.g. the absence of spectral noise, see Wiklund *et al.*[11] for examples. A pragmatic *data-driven* approach is therefore called for. A so-called randomisation test is a data-driven approach and therefore ideally suited for avoiding unrealistic assumptions. For an excellent description of this methodology, see van der Voet.[12] The rationale behind the randomisation test in regression modelling is illustrated in Figure 1. Randomisation amounts to permuting indices and the randomisation test is often referred to as a permutation test. In QSAR (quantitative structure–activity relationship) applications it is known as "Y-scrambling". Clearly, "scrambling" the elements of Y, while keeping the corresponding numbers in X fixed, destroys any relationship that might exist between the X- and Y-variables. Randomisation therefore yields PLS regression models that should reflect the absence of a real association between the X- and Y-variables. For each of these random models, a test statistic is calculated. The value for a test statistic obtained *after randomisation* should be indistinguishable from a chance fluctuation. For this reason, it will be referred to as a "noise value". Repeating this calculation a number of times generates a histogram for the null-distribution, i.e., the distribution that holds when the component under scrutiny is due to chance—the null-hypothesis (H_0). Next, a critical value is derived from the null-distribution as the value exceeded by a certain percentage of "noise values" (say 5% or 10%). Finally, the statistic obtained for the original data—the value under test—is compared with the critical value. The (only) difference with a conventional statistical test is that the critical value follows as a percentage point of a *data-driven histogram* of "noise values" instead of a *theoretical distribution* that is tabulated, e.g., *t* or *F*.

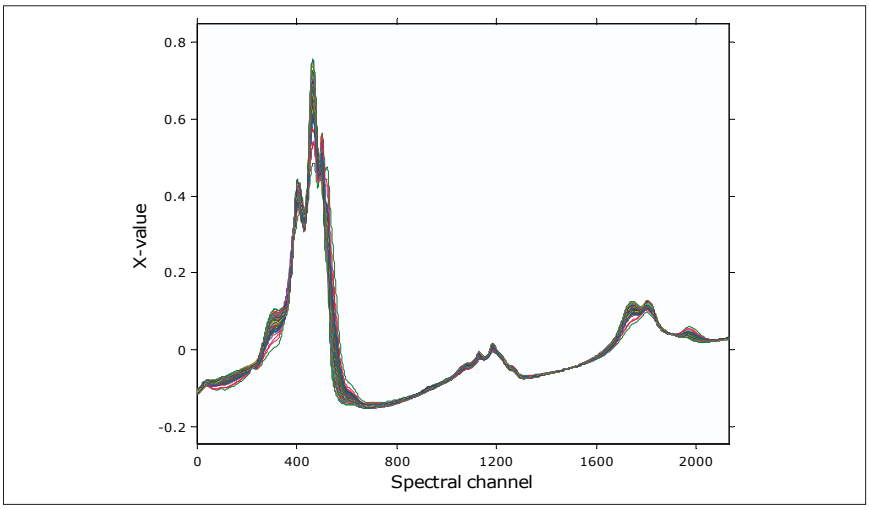


Figure 2. NIR spectra of the example data set.

Experimental

The example data set

A near infrared (NIR) spectral data set serves to illustrate the problems with the conventional validation approach to avoid over-fitting. This type of spectral data provides critical test cases for PLS component selection procedures because tiny substructures may have predictive value. The example data set contains NIR spectra (X) for 239 gas oil samples measured between 4900 and 9000 cm^{-1} (Figure 2). The property of interest (Y) is the hydrogen content. The reference values were determined by nuclear magnetic resonance, which has an estimated measurement error standard deviation σ_{ref}=0.025 g / 100 g. Eighty-four samples were used for calibration and 155 samples for validation. It is noted that the majority of the available samples was selected for (external) validation, which is quite unusual in practice. However, Fernández Pierna *et al.* had chosen this particular data split to test expressions for multivariate sample-specific prediction uncertainty.[13] In other words: focus was more on assessing the predictive ability of a model than on obtaining the best model. For the current study it is useful to have a relatively large validation set because external validation is generally considered to be the "gold standard".

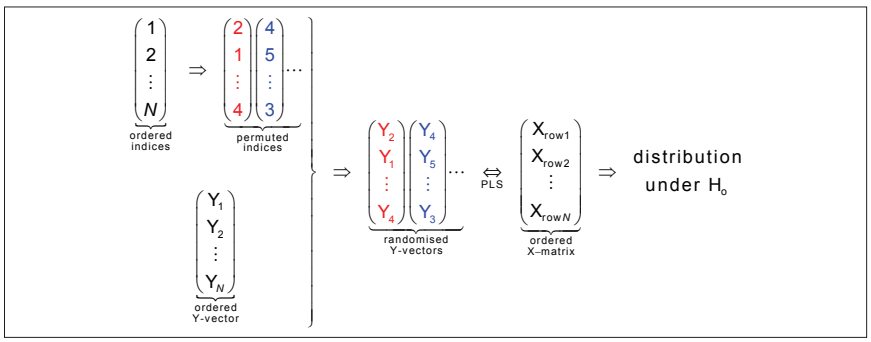


Figure 1. Generating the distribution under the null-hypothesis (H_0) by building a series of PLS models after pairing up the observations for predictor (X) and response (Y) variables at random. Any result obtained by PLS modelling after randomisation must be due to chance. The statistical significance of the test statistic obtained for the original data follows from a comparison with the corresponding randomisation results.

Calculations

The randomisation test was implemented in Matlab 7.0 (The Mathworks, Natick, MA, USA) and the program is available from the first author. Histograms of "noise values" were generated using

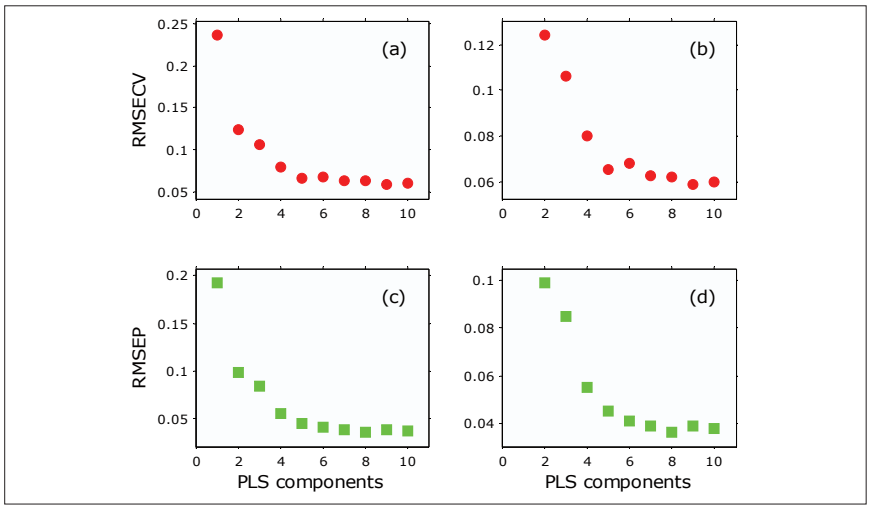


Figure 3. Validation results for the example data set: (top panels) internal *RMSECV* (●) for the 84 calibration samples and (bottom panels) external *RMSEP* (■) for the 155 independent validation samples. To better exploit the vertical scale, the first point is omitted in panels (b) and (d).

1000 permutations. Although as few as 100 permutations can be used,[12] this relatively large number ensures that the resulting histograms are fairly smooth. For the current example data set (84 samples × 2128 wavelengths), the computations were completed within seven CPU seconds on a 3.4 GHz personal computer. To calculate the risk of over-fitting when, in fact, none of the "noise values" exceeds the value under test, the so-called inverse Gaussian function is fit to the "noise values". This function is often suited for modelling positive and/or positively skewed data.[14]

Results and discussion

The conventional validation approach

Both internal validation using cross-validation and external validation—the "gold standard"—lead to a rather subjective decision process (Figure 3). The global minimum in *RMSECV* is achieved for nine components (see top panels). However, the five-dimensional model achieves the first local minimum in *RMSECV* and the "improvement" in *RMSECV* obtained by adding components 6–9 is hard to appreciate. Likewise, the external *RMSEP*

TONY DAVIES COLUMN

estimates continue to decrease until eight components have been fitted, but the rate of "improvement" is difficult to assess (see bottom panels). The analyst faces major difficulties to decide objectively whether a further decrease of *RMSEP* is worthwhile or merely results from "statistical fluctuations". We suspect that to obtain a clear minimum, many more samples are required since the law of diminishing returns is in force—Equation (2). However, the currently available total number of samples (239) is already quite favourable.

The proposed alternative

Histograms of "noise values" generated for components 1–8 are presented in Figure 4. It is observed that the probability that the value under test is due to chance (α) is extremely small for components 1 (0.0009%), 2 (0.02%), 4 (0.0006%) and 5 (0.002%). Interestingly, the significance of component 3 is only 3.3%. We speculate this to be due to component 3 taking care, with some difficulty, of subtle non-linearities in the spectra, after which the remaining linear contributions are conveniently handled by components 4 and 5. The high α-values for components 6–8 constitute a *clear unambiguous warning* that over-fitting starts after the 5th component.

Concluding remarks

The conventional validation approach to component selection is problematic in practice because, often, the *RMSEP* estimates do not yield a clear global minimum. In such a case, the analyst has to resort to "visual inspection" and its associated "soft" decision rules.

This all leads to a rather subjective decision process, which makes the proposed statistical alternative rather attractive. So if software were to make use of this approach the requirement for automated warnings could well be met!

Well done Klaas!—A.N.D.

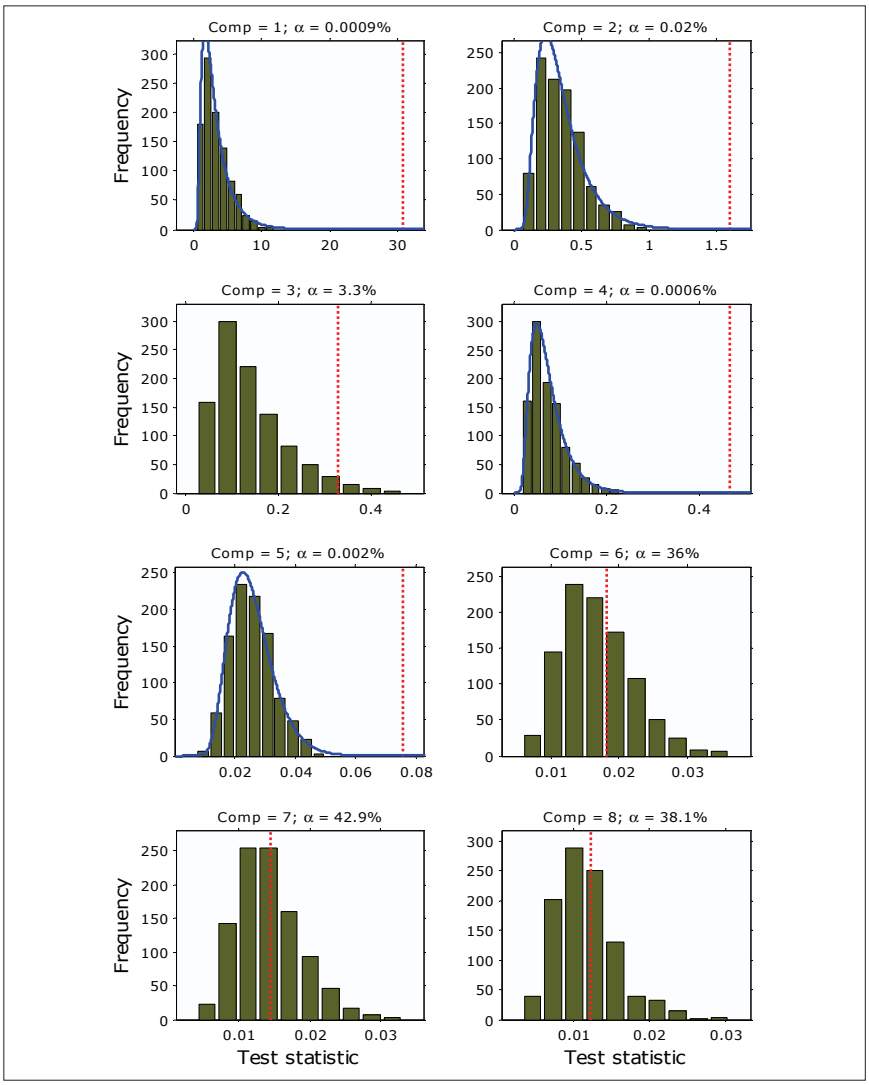


Figure 4. Randomisation results for the example data set: histogram of 1000 "noise values", fit using the inverse Gaussian function (——) and value under test (- - -). The symbol α stands for the significance level.

References

1. S. Wold, H. Antti, F. Lindgren and J. Öhman, *Chemometr. Intell. Lab. Syst.* **44,** 175 (1998).
2. A. Savitzky and M.J.E. Golay, *Anal. Chem*. **36,** 1627 (1964).
3. A.N. Davies, *Spectrosc. Europe* **16(3),** 26 (2004).
4. M.C. Denham, *J. Chemometr.* **14,** 351 (2000).
5. R. Wehrens, H. Putter and L.M.C. Buydens, *Chemometr. Intell. Lab. Syst.* **54,** 35 (2000).
6. A. Lorber and B.R. Kowalski, *Appl. Spectrosc*. **44,** 1464 (1990).
7. R. DiFoggio, *Appl. Spectrosc*. **49,** 67 (1995).
8. H.A. Martens and P. Dardenne, *Chemometr. Intell. Lab. Syst*. **44,** 99 (1998).
9. L. Xu and I. Schechter, *Anal. Chem*. **68,** 2392 (1996).
10. E.T.S. Skibsted, H.F.M. Boelens, J.A. Westerhuis, D.T. Witte and A.K. Smilde, *Anal. Chem*. **58,** 264 (2004).
11. S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold and K. Faber, *J. Chemometr.*, submitted.
12. H. van der Voet, *Chemometr. Intell. Lab. Syst*. **25,** 313 (1994).
13. J.A. Fernández Pierna, L. Jin, F. Wahl, N.M. Faber and D.L. Massart, *Chemometr. Intell. Lab. Syst*. **65,** 281 (2003).
14. R.S. Chhikara and J.L. Folks, *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Marcel Dekker, New York (1989).