A statistics-based expert calibration system

Paul R. Mobley and Bruce R. Kowalski

Center for Process Analytical Chemistry, Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, WA 98195, USA.

Introduction

The value of employing the power of chemometric methods is significant, as can be seen by the large number of publications and the proven success of industrially applied chemometric calibration methods in the past twenty years.[1] This success has been spurred on by several significant factors. The development of faster, more powerful computers has allowed for rapid retrieval and interpretation of data which has, in turn, lent itself to monitoring chemical processes. The ability to monitor a chemical process in real time creates the potential to reduce waste and optimize for energy efficiency and product quality—all factors which weigh heavily in the commercial viability of an industrial product. A second factor key to the proliferation of chemometric techniques is the growing awareness of the mathematical advantages inherent in data produced by certain types of analytical techniques. Analytical instrumentation may be categorized according to the type of mathematical tensor created by the technique and notable advantages are achieved by working with data of higher order.[2] Zero order instrumentation, e.g. a pH meter, produces data consisting of a scalar and allows for calibration in cases where the sample contains only the analyte or where the chemical sensor responds only to the analyte of interest. Detection of outliers, for example unanticipated additional chemical components, is not possible with zero order data. Near infrared (NIR) spectroscopy well exemplifies the advantages attained in progressing to first order calibration techniques. First order data consists of a vector of numbers and allows for calibration of an analyte in the presence of interferents which overlap spectrally with the analyte *as long as the interferents are included in the calibration set*. Unlike zero order calibration, outliers can be detected; however, it is not possible to correct them. Second order instrumentation, which produces a matrix of data per sample, e.g. gas chromatrography-mass spectrometry, provides the holy grail of the analytical chemist: prediction of an analyte in the presence of unknown interferents. Although still early in development, applications using second order calibration are beginning to find their way into the literature.[3–5]

Although first order chemometric techniques have proven their worth and second order techniques have shown potential for even more powerful applications, several significant stumbling blocks prevent the field from gaining wider acceptance. The complexity and relative youth of chemometrics as an area of research has resulted in a dearth of qualified, competent practitioners capable of applying chemometrics techniques. The difficulties raised by the scarcity of chemometricians is further amplified in process applications where many process analysis techniques are not automated and constant supervision is required. Another hindrance to more frequent application of chemometric techniques is the lack of even a single ASTM approved application in the chemical industry. With these problems in mind, development of an expert calibration system is underway at the Center for Process Analytical Chemisty (CPAC).

From Near Infrared Spectroscopy: The Future Waves

Overview of expert calibration system (ECS)

The initial step taken in building the ECS was to gather a group of expert chemometricians and collectively formulate a flow chart describing the steps required to perform optimal calibration. Among others, these steps include sample selection, data preprocessing, outlier detection, model selection, model validation and subsequent diagnostics to ensure the model remains valid over time. The flowchart is shown in Figure 1. It is also anticipated that the ECS will at some point be enabled to interact with the user before any data is collected to optimize experimental design. To further facilitate development of the ECS, the project was divided into three phases; each phase being directed to a user with a specific level of chemometrics expertise. Phase I, recently completed, consisted of the assembly of the MATLAB M-files an expert user would require in carrying out calibration. Basically, this phase gathered the building blocks for the second phase which begins to add expert capabilities. Phase II of the project is geared for a spectroscopist who has the background of a research and development scientist, is knowledgeable of the data being collected and can make some informed decisions. The goal of this phase is for the ECS to make the decisions requiring a chemometrics background yet still take advantage of the chemical knowledge of the user. The final phase of the project is aimed at the basic operator and is essentially a black box with the ECS making all decisions involving calibration.

In developing the expert calibration system, two design goals have been identified as being priorities for enhancing the usefulness and effectiveness of the system. The first of these is to strive to make the system functional and effective in producing the optimal calibration model given any data set for which it might be applied (e.g. near infrared, ultraviolet/visible or Raman data.) The goal here is to minimize the number of assumptions required for the program to operate with the intent to widen the utility of the expert system. The second design goal of the expert calibration system is to require and make use of error estimates of both the spectral and reference values. This requirement seems reasonable as analytical chemists should be aware of the errors in their measurements and the increased mathematical and statistical rigor will likely be required for the ECS to make decisions. In addition, the added statistical rigor should expedite the ASTM approval process. Further, a new method of model selection, designed to overcome the overfitting problem in current state-of-the art selection methods, requires the exploitation of this additional information to indicate which model will predict optimally. This new method of model selection, based on the Parsimony Principle, presents a major departure from current selection methods. Since model selection plays a pivotal role in the structure and function of the ECS, the Parsimony Principle is presented here in some detail.

The Parsimony Principle: model selection

In performing calibration, selection of the complexity of the model plays a crucial role in the model's overall predictive ability when faced with future unknown samples. In the early development of calibration methodology, factor selection methods were based either on principal components analysis (PCA) selection rules[6], which are concerned with modeling the data matrix, or basic fit statistics[7] which test the model's ability to do self-prediction of some physical or chemical property. The problem inherent in both of these approaches is that the quality most desired in the model, i.e. good predictive ability of unknown samples, is not the criterion.

In response to this problem, leave-one-out cross-validation was developed and has since become the standard method of factor selection.[8–12] Leave-one-out cross-validation entails removing one sample from the data set, building the calibration model and then using the model to predict the left out sample. By continuing this process until each sample has been left out, an estimate of the predictive ability is attained. Although this method is widely acknowledged as the best of those currently available and is the most commonly used stopping rule, cross-validation tends to overfit

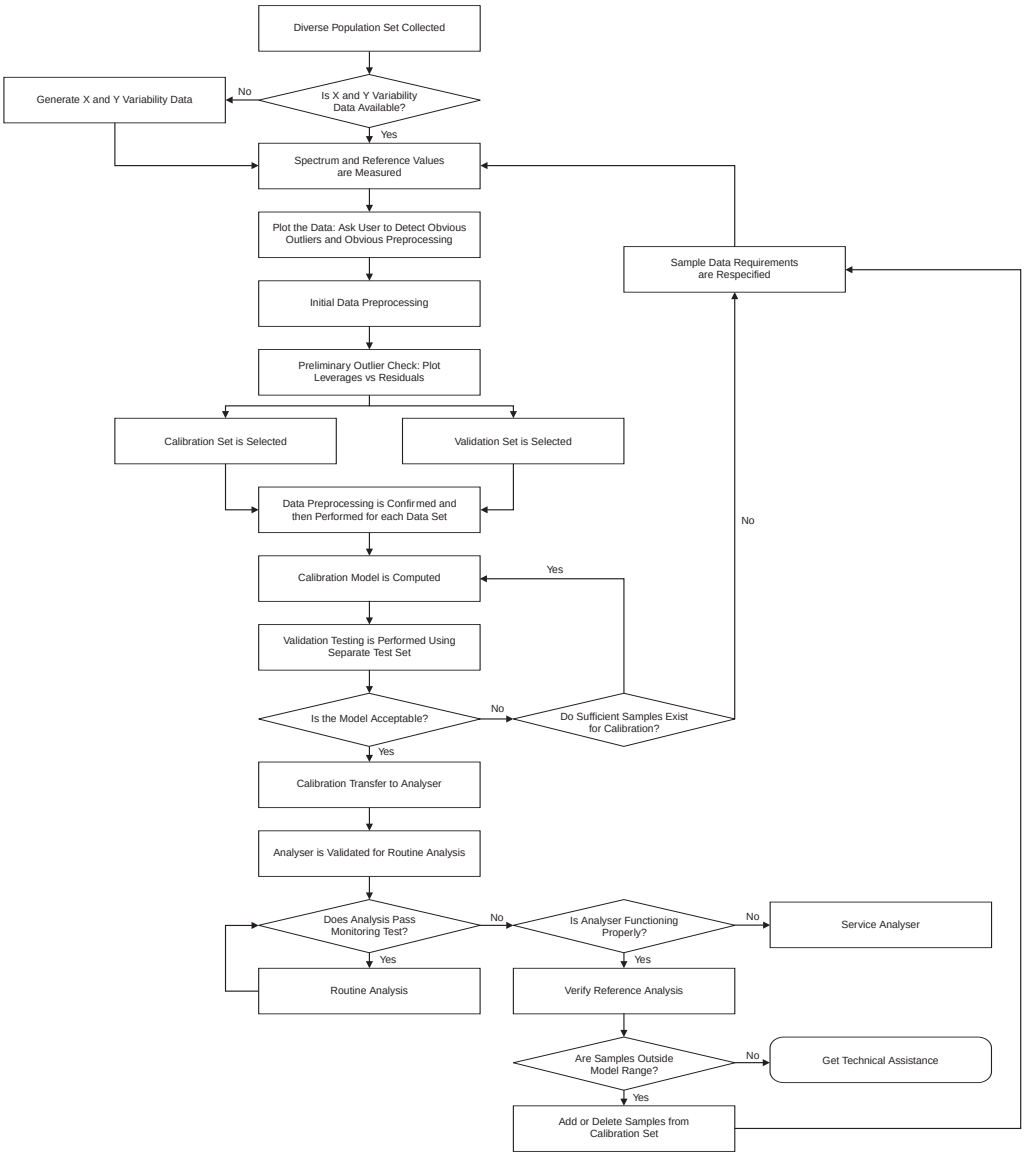

Figure 1. Expert calibration system flowchart.

many data sets which means that noise is incorporated into the calibration model and predictive ability is compromised. In an effort to overcome the problem of overfitting, Seasholtz and Kowalski revisited the mathematical and statistical literature to investigate whether a solution

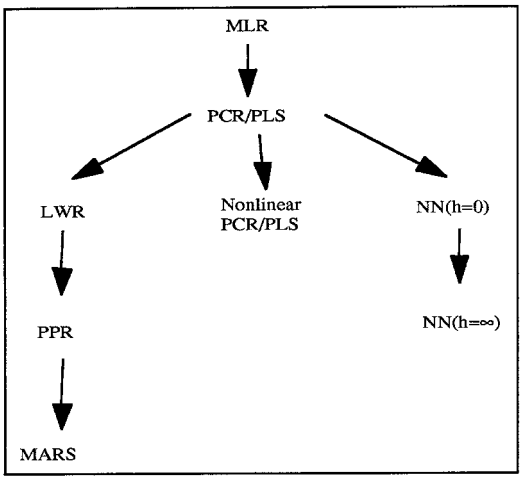

Figure 2. Parsimony flowchart: modelling methods include multiple linear regression (MLR), principal components regression (PCR), partial least squares (PLS), locally weighted regression (LWR), projection pursuit regression (PPR), multiplicative adaptive regression splines (MARS) and neural networks (NN) with zero to infinite number of hidden nodes.

might already be available.[13] This search led to the Parsimony Principle which says that, on average, the simplest effective model will yield the best predictive ability.

The Parsimony Principle has two important results. From the wide perspective of calibration, the principle provides direction as to which of the many modeling methods available should be constructed first. That is the simplest model, a stepwise method such as multiple linear regression (MLR), should be tried first and more complex modeling methods should then be implemented only as required by the data set and as allowed by the level of measurement error. The flow chart depicted in Figure 2 shows the order in which several of the more commonly used modeling methods should be applied according to the dictates of the Parsimony Principle. Note that the Parsimony Principle relies on the various modeling types being nested (simpler models being a subset of the more complex models) in order to determine relative complexity. Therefore, the flow chart branches in the lower section where models are not nested.

A second result of the Parsimony Principle is to stress the need for simple, parsimonious calibration models. The difficulty in implementing this concept defines the primary problem this paper confronts, which is to determine when a simple and effective model has been attained. In order to determine when enough factors have been included to allow effective calibration, while at the same time limiting the number of factors to prevent incorporation of noise into the model, demands a departure from current factor selection rules. In an effort to accommodate the requirements of the Parsimony Principle, a selection criterion was designed with the intent that it indicate more factors should be added to the model only as long as the fit error is greater than the error due to noise in both the measured and reference values. The extent to which the model fits the data is ascertained by a simple fit statistic, the sum of squared residuals (*SSR*):

$$SSR = \sum_{i=1}^{I} (y_i - \hat{y}_i)^2 \qquad (1)$$

• **Noise-added sets are created:**

Measured variables: (X is a data matrix)
X = first randomization of noise ⇒ X_{n1}
X + second randomization of noise ⇒ X_{n2}
Reference values (y is a vector)
y + first randomization of noise ⇒ y_{n1}
y + second randomization of noise ⇒ y_{n2}

•**Regression is performed:**

b_1 = PLS (X_{n1}, y_{n1}) or PCR (X_{n1}, y_{n1})
b_2 = PLS (X_{n2}, y_{n2}) or PCR (X_{n2}, y_{n2})

•**Self-predict both sets:**

$$\hat{y}_{n1} = X_{n1}\, b_1$$
$$\hat{y}_{n2} = X_{n2}\, b_2$$

Figure 3.

where I is the number of samples, y_i is a vector containing the reference values (i.e. analyte concentration), and $\hat{y}_i$ is the vector of model predicted values. In order to determine the effect measurement error has on the model, the stopping rule then takes advantage of the merits of adding known amounts of noise to a data set; a concept that is not new, but one that has perhaps been under-utilized,[9,10]

$$\text{Error due to noise} = \sum_{i=1}^{I} (y_{n1,i} - \hat{y}_{n2,i})^2 \qquad (2)$$

where $\hat{y}_{n1}$ and $\hat{y}_{n2}$ are the self-predicted values of data sets which have had noise added at similar, pre-determined levels but at two different randomizations, $n1$ and $n2$. To further clarify, a brief outline of the algorithm is given in Figure 3.

The purpose behind perturbation of the data set is to determine the impact of measurement errors on the calibration model. As is shown in the following theory section, subtracting one estimation from another (i.e. $\hat{y}_{n1,i,r} - \hat{y}_{n2,i,r}$) removes error due to bias which occurs when too few factors are included in the model. The presence of additional noise then yields an estimate of the error due exclusively to measurement noise. Several mathematical and statistical considerations must also be realized to ensure optimal results. Since Equation 2 is the result of artificially added noise, many replicates using different randomizations of noise can be applied in order to better characterize the results. Also, because Equation 2 includes subtraction of one normally distributed population from another population exhibiting the same noise level, a scaling factor of 1/2 is needed as a correction to ensure that a fair comparison can be made of the error due to noise and the error due to bias. Including both of the above considerations into Equation 2 yields,

$$\textit{Corrected}\ \text{error due to noise} = \frac{1}{2R} \sum_{r=1}^{R} \sum_{i=1}^{I} (\hat{y}_{n1,r,i} - \hat{y}_{n2,r,i})^2 \qquad (3)$$

where R reflects the number of replicate randomizations.

Equations 1 and 3 both describe variances which must be compared in order to discern when the fit error is no longer larger than the error due exclusively to noise. In fact, this also describes the purpose of statistical F-tests: comparing two variances. By placing Equations 1 and 3 into the standard format of an F-test, the final stopping rule statistic is achieved,

$$F = \frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{\frac{1}{2R} \sum_{r=1}^{R} \sum_{i=1}^{I} (\hat{y}_{n1,r,i} - \hat{y}_{n2,r,i})^2} \tag{4}$$

In order to discern when the two variances compared in the F-test are no longer significantly different, a method involving bootstrapping has been used; however, research is currently underway to determine the correct number of degrees of freedom which would allow for use of established F tables.

Expert calibration system attributes

A preliminary list of data preprocessing measures includes log transformations, Kubelka–Munk, mean centering, autoscaling, derivatives, smoothing, multiplicative scatter correction, background subtraction and baseline correction. Calibration methods include stepwise multiple linear regression (MLR), principal components regression (PCR), partial least squares (PLS), nonlinear versions of PCR and PLS, locally weighted regression II (LWR2) and neural networks. A distinct advantage of the stopping rule described above is the flexibility of implementation regardless of the noise distribution, preprocessing measure or calibration method.

References

1. W.W. Blaser, R.A. Bredeweg, R.S. Harner, M.A. LaPack, A. Leugers, D.P. Martin, R.J. Pell, J.J. Workman and L.G. Wright, *Anal. Chem.* **67,** 60R (1995).
2. K.S. Booksh and B.R. Kowalski, *Anal. Chem.* **66,** 782A (1994).
3. J.M. Henshaw, L.W. Burgess, K.S. Booksh and B.R. Kowalski, *Anal. Chem.* **66,** 3328 (1994).
4. R. Tauler, A.K. Smilde, J.M. Henshaw, L.W. Burgess and B.R. Kowalski, *Anal. Chem.* **66,** 3337 (1994).
5. A.K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess and B.R. Kowalski, *Anal. Chem.* **66,** 3345 (1994).
6. E. Malinowski, *Factor Analysis in Chemistry,* Second Edition. John Wiley & Sons, New York (1991).
7. N.R. Draper and H. Smith, *Applied Regression Analysis,* Second Edition. John Wiley & Sons, New York (1981).
8. M.J. Stone, *Roy. Statist. Soc. B,* 36 (1973).
9. B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia (1982).
10. D.M. Haaland and E.V. Thomas, *Anal. Chem.* **60,** 1193 (1988).
11. D.W. Osten, *J. Chemom.* **2,** 39 (1988).
12. I.N. Wakeling and J.J. Morris, *J. Chemom.* **7,** 291 (1993).
13. M.B. Seasholtz and B.R. Kowalski, *Anal. Chim. Acta* **277,** 156 (1993).

From Near Infrared Spectroscopy: The Future Waves