Chemometrics

Technical Note

Comparison of Factor-Based PLS and PCR to Traditional Calibration Methods

Abstract

When attempting to model spectroscopic data for the determination of bulk property or concentration, multivariate regression methods are particularly useful due to the increased precision attained from including multiple channels of data. The analyst is reminded, however, that spectral data is composed not only of relevant information, but of irrelevant information and noise as well. Thus, care should be taken in the creation of regression models to include only the relevant portion of the data.

Data	=	Relevant Information	+	Irrelevant Information	+	Noise

The two traditional multivariate regression techniques used in spectroscopy are Classical Least Squares and Inverse Least Squares. Two factor-based methods, Principal Component Regression (PCR) and Partial Least Squares (PLS), have been shown to offer advantages which often produce more robust modeling of analytical data.

This note will summarize the relevant differences among the above methods, with information gleaned from a few key literature sources (1-3). The factor methods described are as implemented in the Infometrix® Pirouette® software.

Traditional Regression Methods

CLS and ILS are Traditional Data-based Methods

The need for robust regression methods which function in the presence of highly correlated data has driven spectroscopists to multivariate approaches. The techniques first used for this work were Classical Least Squares and Inverse Least Squares (Multiple Linear Regression).

Classical Least Square (CLS)

CLS (the K Matrix method in infrared spectroscopy) assumes Beer's Law and that absorbance at each frequency is proportional to component concentration. Model error is assumed to derive from the measurement of spectral absorbance. Thus, a model generated using CLS in its simplest form, requires that all interfering chemical components be known and included in the calibration data set.

This combination of constraints conspires against using the classical approach in many analytical systems. With limited exceptions (*e.g.*, gas phase spectroscopy, some process monitoring), it is common for unknowns or chemical variations to creep into a routine analysis. However, CLS does have the advantage of improved precision when using many frequencies, due to signal averaging.

Inverse Least Squares (ILS)

ILS (the P Matrix method in infrared spectroscopy) applies the inverse of Beer's Law and assumes that component concentration is a function of absorbance. Model error is assumed to derive from error in the measurement of component concentration, whereas no error is assumed to be inherent in the absorbance values. An ILS model has a significant advantage in that it does not need to know and include all components in the calibration set. However, ILS does not have the signal averaging gain as with CLS.

ILS does assume that the intensities for each measured variable in the analysis all behave perfectly independently. Further, you are restricted from using all of the spectral channels in making the model: the number of channels of spectral information used cannot exceed the number of calibration standards. Precision will be degraded if more channels are included than the number of independent sources of variation in the data.

Factor Analysis Methods

PCR and PLS are Factor-based Methods

To get around some of the disadvantages of the CLS and ILS methods, two factor-based regression methods, PCR and PLS, have been developed. Both employ factor analysis, then use a subset of the resulting factors to complete the regression modeling. Both PLS and PCR have the signal averaging advantages of a full-spectral technique such as CLS while retaining the ILS advantage of being able to perform a calibration without needing to know the characteristics of all interferences. With both PLS and PCR, you must optimize the number of factors to use.

Principal Components Regression (PCR)

PCR is a factor analysis followed by a regression step. In essence, it uses the same inverse approach as ILS except that it uses a subset of the principal components instead of absorbances. It, therefore, combines the advantage of using all of the spectral channels, avoids noise (which is relegated to the unused factors) and retains the ILS independence of uncalibrated components. The factor analysis reduces errors in absorbance, while the following regression minimizes errors in concentration.

Partial Least Squares (PLS)

PLS combines both CLS and ILS approaches using factor data. As with PCR, PLS assumes that the error can stem from both absorbance readings and from the measurement of component concentration. As a factor-based

method, PLS uses all of the channels and avoids noise. Like ILS and PCR, PLS does not require the user to prepare a calibration standard where all of the interfering species are known. PLS tries to find those factors which have the greatest relevance for prediction, whereas PCR finds its factors independent of correlations to concentration.

Conclusions

The salient differences can be summarized as presented in the table below. Classical Least Squares is sensitive to the unknown components in the calibration mixture. This sensitivity makes CLS less useful in many technical applications, although improvements have been recently suggested (1). Inverse Least Squares, on the other hand, takes care of the unknown component problem but can be sensitive to noise in typical spectroscopy applications. The fact that a portion of the data must be discarded in order to apply ILS can add to the error.

There is a significant advantage in using a factor-based method in that it models data in a more compact form and, because it organizes the data based on the similarity of information content, it can also aid in chemical interpretation of the system. Unlike ILS, there is no variable selection necessary, and there is less of a tendency to overfit the data due to noise. Unlike CLS, unknowns in the calibration set do not present a problem.

Note that no modeling technique will correct for large errors in assessing the concentration in the training set. If the data is bad, CLS, ILS, PCR and PLS will all yield unacceptable models.

(1)Thomas, E.V. and D.M. Haaland, Anal. Chem. (1990) 62: 1091-1099.

(2)Haaland, D.M. and E.V. Thomas, Anal. Chem. (1988) 60: 1193-1202.

(3)Beebe, K.R. and B.R. Kowalski, Anal. Chem. (1987) 59: 1007A-1017A.

Table 1. Summary of the basis and source of error for four multivariate calibration techniques

Technique	Spectral Basis	Data Used	Assumed Error	Primary Problems
CLS	Full Spectrum	Raw	Absorbance	Unknown Components
ILS	Partial Spectrum	Raw	Concentration	Noise, Variable Choice
PCR	Full Spectrum	Factors	Absorbance & Concentration	Spectral Errors
PLS	Full Spectrum	Factors	Absorbance & Concentration	Large Concentration Errors