SAND96-2243C
CONF-9610166--1

CHEMOMETRICS APPLIED TO VIBRATIONAL SPECTROSCOPY: OVERVIEW, CHALLENGES AND PITFALLS

RECEIVED
SEP 18 1996
OSTI

David M. Haaland

Sandia National Laboratories, Albuquerque, New Mexico, USA 87185-0342

Introduction

Chemometric multivariate calibration methods are rapidly impacting quantitative infrared spectroscopy in many positive ways. The combination of vibrational spectroscopy and chemometrics has been used by industry for quality control and process monitoring. The growth of these methods has been phenomenal in the past decade. Yet, as with any new technology, there are growing pains. The methods are so powerful at finding correlations in the data, that when used without great care they can readily yield results that are not valid for the analysis of future unknown samples. In this paper, I discuss the power of the multivariate calibration methods while pointing out common pitfalls and some remaining challenges that may slow the implementation of chemometrics in research and industry.

MASTER

Multivariate Calibration Methods

Because multivariate calibration methods are the most widely used of the chemometric methods, they are generally thought by many to be synonymous with chemometrics. Although chemometrics encompasses many methods in addition to multivariate calibration, I will limit

this discussion to multivariate calibration. Multivariate simply means that multiple variables are used in the analyses of the data. In the case of spectroscopy, the intensities of multiple samples at many frequencies are involved in the multivariate calibration. As in univariate analyses that involve the analysis of a single spectral intensity, multivariate calibration seeks to build an empirical model that relates concentration or sample property information to the spectra of a series of known calibration samples. This empirical model is then used in the multivariate prediction step that applies the calibration model to the spectrum of the unknown sample to estimate the component concentrations and/or the properties of the unknown sample.

Three factor based multivariate calibration methods are generally used in multivariate calibration. These are classical least squares (CLS), principal component regression (PCR), and partial least squares (PLS) methods [1]. Each method factor analyzes the spectral calibration intensity data for all calibration samples into the product of two smaller matrices known as score and loading vector matrices. The loading vectors are a set of full-spectrum basis vectors that contain all the information in the spectra required to model the component concentration of interest. The scores represent the amounts or "intensities" of these basis vectors in the reduced factor analysis space of the loading vectors.

CLS uses the Beer's law model in factor analyzing the spectral data. Thus, the CLS loading vectors are representations of the infrared active pure components that vary in the calibration samples, and the scores represent the concentrations of these infrared active components. It is clear that the CLS representation of the spectral data has useful qualitative information for the spectroscopist since the loading vectors represent the chemically relevant space of the pure-component spectra. The PCR model factor analyzes the spectral data into loading vectors that represent the maximal sources of spectral variance without regard for the component concentrations of the calibration samples. PLS, on the other hand, finds those loading vectors that maximize the covariance of the spectral and concentration information. Thus, the PLS loading vectors are those that maximally correlate with the component being analyzed. PCR and PLS methods are more flexible and more generally applicable because they are not constrained by the Beer's law model. Each method has its own set of advantages that have been reviewed [2].

In general, these multivariate methods are capable of being more precise, accurate, and reliable than the univariate methods. They are more precise because the factor analysis methods preferentially reject those loading vectors that contain higher levels of noise. They also are able to signal average the redundant component information across the spectra to achieve a further improvement in analysis precision. Unlike the univariate methods that are blind to outlier or problem samples, multivariate methods are able to detect these outlier samples. If outliers are appropriately removed form the calibration set, greater calibration accuracy is achieved. During prediction, outlier samples whose spectra are different from the set of calibration spectra can be identified with statistical tests, and their analyses should be considered suspect. Thus, the reliability of the analyses can be improved with the multivariate methods. Finally, because intensities are used at multiple frequencies, many sources of overlapping spectral variation can be included in the model, and the effects of spectral interferences and spectrometer drift can be accommodated. Thus, the model can be made robust against these other sources of spectral variation.

Experimental Design

Because multivariate calibration methods are so efficient at finding correlations present in the calibration data, it is desirable that good statistically designed calibration sets be used in the calibration phase of the analysis. This generally means that when a calibration set can be designed, orthogonal statistical designs should be used. Factorial designs are one example of experimental designs that can be employed [3]. Each component or experimental parameter that can be controlled can represent a factor in the design. Each factor is then varied at two or more levels. The design allows all factors to be varied simultaneously and, therefore, allows interactions between components and other variables to be modeled. If more than two levels are included in the designs, then quadratic and higher order effects can also be modeled. A feature of these designs is that all of the controlled variables are made to be orthogonal, i.e., there is no correlation between components in the calibration. This is important since we want to avoid using the spectral variation of one component to inappropriately aid in the calibration of other components that are not guaranteed to be correlated in the unknown samples. Another very important part of experimental design is the randomization of the sample run order. This will allow the model to guard against incorporating spectrometer drift or other sources of non-sample variation into the calibration model.

The inclusion of repeat samples in the analysis can be a useful tool for determining the repeatability and reproducibility of the chemometric analyses. By using statistical experimental designs coupled with the process of cross validation in the calibration, realistic estimates of analysis precision can be obtained during the calibration. Cross validation is a procedure that can minimize overfitting of the calibration data and is used to determine the optimal number of loading vectors to use in the model [1]. Cross validation involves removing single samples or sets of samples from the calibration model determination. These removed samples are then used as test samples for obtaining true predictions. Once predicted these samples are placed back in the calibration and another sample or set of samples is removed. The process is repeated until each sample has been removed from the calibration and predicted. It is important that all repeat samples are removed together from the calibration or overfitting of the calibration data will occur and an optimistic estimate of the prediction precision will be obtained. We recommend that a separate validation set of known samples be run on a different date to assure that spectrometer drift or some unknown lurking variable is not present to degrade the quality of the analysis.

If designed calibrations are not possible, then calibrations based upon randomly selected samples can be performed. However, the orthogonality of components cannot be assured, and validation test samples should definitely be run. Sometimes, calibration samples can be selected from a larger number of samples based upon their spectra. Methods are available to select sets of calibration spectra that adequately span the spectral space of the samples available for calibration [4].

Outlier Detection

The ability of multivariate calibration methods to detect outlier samples greatly improves the reliability of the methods. A variety of outlier detection methods are available with multivariate calibration and prediction. One can examine the spectral F ratios to detect spectra that are not modeled well [1]. The spectral F ratio is a statistical metric that compares the sum of squared spectral residuals for a given sample to the average of the sum of squared spectral residuals for the calibration. The spectral residuals are the difference between the measured spectrum and the model for the sample spectrum based upon the multivariate calibration model. This outlier detection method can be used during calibration to detect outlier calibration samples. It can also be used during prediction of an unknown sample. However, during calibration, it is important that this spectral F ratio be calculated for samples left out of the calibration during the cross-validation procedure. If all samples are included in the calibration, outlier samples tend to require a separate factor to describe their variation, and their spectral residuals can be quite small. During the cross-validation procedure, spectra of the outlier samples left out of the calibration will be poorly modeled, and the spectral F ratios will tend to be large. In addition, by examining the spectral residuals of the outlier samples, it is often possible to identify the reason for a sample being an outlier. For example, unexpected components in the predicted sample that are not present in the calibration tend to have the unexpected component visible in the spectral residual. Thus, the spectral residuals can be used for both identifying and possibly understanding the reasons for a sample being an outlier.

Another outlier detection method is the concentration F ratio [1]. It is simply the squared concentration error of the prediction sample divided by the average of the squared concentration errors of the calibration samples. If this ratio is larger than expected based upon normally distributed errors, then the sample is flagged as an outlier. This metric can only be used for calibration or validation samples since the concentration error of a true unknown sample is not known. Again, this metric is only valid for samples left out of the calibration during the cross-validation procedure or for separate validation samples.

A third outlier detection method is the Mahalanobis distance metric which is used to identify extreme samples among the calibration or the prediction samples [5]. The Mahalanobis distance uses the scores to define the multidimensional space of the samples. It is a distance metric that is scaled by the size of each dimension of the score space and, therefore, is not dominated by the scores of the first few loading vectors as is the case with the Euclidean metric. This metric identifies samples that are far from the center of the spectral space spanned by the calibration samples. Since the multivariate calibration methods are excellent at interpolation but problematic in extrapolation, the Mahalanobis distance is a useful metric for identifying extreme samples. If accurate predictions are desired for these extreme samples, then more samples must be added to the calibration set to assure that this extreme space is adequately modeled by the calibration.

In combination, these statistical outlier detection methods greatly improve the reliability of the multivariate methods. Outlier detection is not possible with univariate methods that are based upon a single intensity or ratio of intensities for each sample.

Instrument Standardization

In industry, it is important that calibrations be maintained over time and that they be able to be transferred between spectrometers and locations. Since spectrometers drift with time or are changed as components fail, it is useful to have methods to update multivariate calibrations as simply and as fast as possible. In addition, since calibration is a slow and expensive proposition, it is desirable that calibrations on one instrument be transferred between spectrometers at multiple locations rather than having to build individual calibrations separately on each spectrometer. The process of maintaining calibrations on a single instrument or transferring calibrations between instruments is called instrument standardization. Kowalski and coworkers [6] have developed preliminary methods to perform this instrument standardization using a small subset of calibration samples. When these methods are finally perfected, the calibration of a drifting spectrometer will be more easily maintained and calibrations will be transferred between spectrometers with a minimum of additional work. The need for instrument standardization is dependent on the data set.

Examples of Chemometrics Applied to Quantitative Spectral Analyses

IR Analysis of Glasses. An example of the use of statistical experimental designs and the use of outlier detection methods to improve the accuracy of the chemometric data analysis is given in our work on the analysis of phosphosilicate based glasses used in making glass to aluminum seals. It was desired to make a rapid spectral analysis of the composition and properties of these glasses that consist of P, Al, Na, Ba, and OH groups. Originally, these glasses required a full day of analyses by conventional methods for quality control. By using a fractional factorial experimental design, we built a PLS calibration model with just 17 samples that could measure the composition and physical properties of each new glass from its IR spectrum in a few minutes to the precision required for quality control. However, during the model building phase, a univariate model had no prediction ability while the PLS calibration yielded predictions but the calibration had a precision that was inadequate for quality control. No outliers were detected from the full PLS calibration model. It was not until we implemented outlier detection during cross validation that a single outlier sample was detected. Examination of this sample under the microscope revealed that it was the only calibration sample that contained microcrystals of a separate crystalline phase. When the sample was removed from the calibration, the analysis of all components improved. For example, the standard error of prediction (SEP) for P improved from 0.9 wt. % to 0.3 wt. % (i.e., the same precision of the reference ICP emission method used in building the calibration).

Noninvasive NIR Analyses of Glucose in Diabetic Patients. In using near infrared spectroscopy to monitor the blood glucose levels of diabetic patients, we built our original calibrations on spectra that were taken during multiple glucose meal tolerance tests [7]. Multiple meal tests were run over the course of the day, and the glucose level was ramped up and down multiple times in order to minimize the possibility of incorporating instrument drift into the calibration models. However, the fact that the glucose values were not fully randomized may have contributed to an optimistic glucose precision. Since the instrument drifts over the same time, any components of instrument drift that by chance correlate with the time dependent glucose values will be incorporated into the PLS model. This problem was demonstrated by Arnold and coworkers [8] by using absorbance spectra of a single blood sample taken as a function of time. They then added to the calibration data dummy values of glucose variation that mimicked standard glucose meal test variations with time. The resulting PLS model was able to obtain glucose calibrations with these data since the glucose levels varied systematically as did the spectrometer drift. Arnold's result calls into question the role non-random sampling might play in the noninvasive glucose precision results. In order to eliminate these problems, calibration over multiple days at random times and glucose levels or the calibration of multiple patients over long calibration times should be performed. Validation data should also be obtained to independently validate the PLS models. We have completed experiments on multiple diabetic patients randomly brought into the hospital. In this case, PLS calibrations yielded cross-validated PLS calibrations that were a factor of two poorer when compared to our earlier single-patient calibrations that were subject to overfitting.

NIR Analysis of Water Temperature and of Dilute Aqueous Samples. NIR spectra of pure water (7500 to 11000 cm^{-1}) were twice obtained as a function of temperature. In the first set, spectra were collected every 5° C as the temperature of the water was systematically raised from 10° to 70° C. In the second run, the spectra were obtained at the same temperatures, but the run order was randomized with respect to time. In each case, spectra were obtained in triplicate. PLS calibrations for temperature showed that the cross-validated SEP for temperature was 0.03° C and 0.02° C with single-sample cross validation for the systematic and random run-order experiments, respectively, it was 0.08° C for the systematic run when all three repeat spectra were removed during each cross-validated rotation, and 0.09° C for the random run-order experiment with triplicates removed. However, when the calibration for the systematic model with triplicates removed was applied to predicting the random data, the SEP was 0.25° while when the random run-order model (with triplicates removed) was applied to the systematic run order temperature spectra, the SEP was 0.09° C. The single-sample cross-validated random run-order calibration predicted the systematic run-order data with an SEP of 0.11° C. Thus, there is a clear indication of overfitting of both the single-sample cross-validated model and the systematic-run-order model since they both yield optimistic precisions and result in poorer predictions when applied to true prediction samples. These models have incorporated spectral noise and systematic spectrometer drift into the model causing true predictions on independent samples to be degraded. However, in the case of the randomized run order model with the triplicate samples removed during cross validation, overfitting of spectral noise and incorporating spectrometer drift into the model are guarded against, and the predictions on the independent sample data are not degraded with this more robust model.

In another experiment, 32 samples of 0 to 3 % urea, creatinine, and NaCl were prepared in an orthogonal Latin Hypercube experimental design and sealed in 1 cm path length cuvettes [9]. Their spectra from 7500 to 11000 cm^{-1} were obtained in random order on three different spectrometers with each sample temperature controlled at 23° C. Cross-validated SEP's on each instrument were in the range of 30 mg/dl. However, when the spectra from one spectrometer were predicted using a model developed on another spectrometer, the SEP values increased to as high as 3000 mg/dl. Clearly, these models do not transfer between spectrometers. However, by using spectra from five samples obtained on both spectrometers to adjust the spectral data from the secondary spectrometer to look like it had been obtained on the primary spectrometer, the SEP's improved to 50 mg/dl. Thus, in this case, instrument standardization was degraded but improved greatly with the use of only five common samples run on each spectrometer.

Conclusions

It is clear the multivariate methods have large advantages over the older univariate methods of quantitative spectral analyses. They are more precise, accurate, and reliable and can be applied to many problems that could not be addressed with the old methods. However, they must be used with great care to avoid a series of potential pitfalls. Proper orthogonal experimental designs with run order randomization is important for obtaining reliable calibrations. Sensitive outlier detection allows improved reliability. Separate validation data sets are recommended to assure the robustness of any models developed. Methods are available for instrument standardization, but more work needs to be completed in this area. Finally, as commercial software packages become more sophisticated and complete, chemometrics will make a greater impact in industrial process monitoring and quality control. However, these methods should not be used as black box approaches since they require careful implementation in order to avoid a variety of common problems.

Acknowledgments

H. D. T. Jones is acknowledged for making the aqueous samples and collecting their spectra. Ed Thomas aided in all experimental designs. David Melgaard programmed the data analyses software. This work was performed at Sandia National Laboratories and was supported by the U. S. Department of Energy under Contract Number DE-AC04-94AL85000.

1. D. M. Haaland and E. V. Thomas, Anal. Chem. 60, 1193 and 1202 (1988).
2. E. V. Thomas and D. M. Haaland, Anal. Chem. 62, 1091 (1990).
3. G. E. P. Box, W. G. Hunter, and J. S. Hunter, Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, Wiley, New York, (1978).
4. W. R. Hruschka and K. H. Norris, Appl. Spectrosc., 36, 261, (1982).
5. P. M. Fredericks, J. B. Lee, P. R. Osborn, and D. A. J. Swinkels, Appl. Spectrosc., 39 , 303 and 311 (1985).
6. Y. Wang, D. Veltkamp, and B. R. Kowalski, Anal. Chem. 63, 2750 (1991).
7. M. R. Robinson, R. P. Eaton, D. M. Haaland, G. W. Koepp, E. V. Thomas, B. R. Stallard, and P. L. Robinson, Clin. Chem., 38, 1618 (1992).
8. S Pan, H. Chung, M. A. Arnold, and G. W. Small, Anal. Chem. 68, 1124 (1996).
9. D. M. Haaland and H. D. T. Jones, 9th International Conference on Fourier Transform Spectroscopy, J. E. Bertie and H. Wisser, Editors, Proc. SPIE Vol. 2089, 448 (1993).