REVIEW

MULTIVARIATE CALIBRATIONS IN UV SPECTROPHOTOMETRIC ANALYSIS

M. SAEED ARAYNE, NAJMA SULTANA* AND SAIMA SHER BAHADUR

Department of Chemistry, University of Karachi, Karachi-75270, Pakistan
**Research Institute of Pharmaceutical Sciences, Faculty of Pharmacy, University of Karachi, Karachi-75270, Pakistan*

ABSTRACT

Calibration allows the user to relate instrumental measurements to the sample of interest. Multivariate calibration allows for the analysis of several measurements from several samples or specimens. The method contributes to the two steps procedure where step one involves the calibration of data and second step involves the prediction that are made or based on the calibration. In calibration, indirect measurements are made from samples where the amount of the analyte has been predetermined, usually by an independent assay or technique. These measurements, along with the predetermined analyte levels, comprise a group known as the calibration set. This set is used to develop a model that relates the amount of sample to the measurements by the instrument. In some cases, the construction of the model is simple due to a certain relationship, such as Beer's Law in the application of UV spectroscopy. Unlike spectroscopy, other cases can be much more complex, and it is in these cases where construction of the model is time-consuming step. Once the model is constructed, it can predict analyte levels based on measurements of new samples. It can be used to separate samples from interferences without the need of highly selective measurements for the analyte. Calibration techniques (used in the calibration step) differ in determining coefficient values for the preceding or similar equations.

Keywords: Chemometrics, multivariate calibration, UV spectrophotometry.

INTRODUCTION

Multivariate calibration can be used to separate samples from interferences without the need of highly selective measurements for the analyte (Thomas, 1994). In the case of HPLC, certain overlapping or anomalous peaks can be systematically separated or deleted from the data set based on certain linear combinations of measurements derived from one of several multivariate calibration techniques.

Theory of multivariate calibration

The term multivariate calibration refers to the process of constructing a mathematical model that relates a property such as content or identity to the absorbances of a set of known reference samples at more than one wavelength. A calibration model, *i.e.*, a mathematical expression, is determined from a set of samples of known content and identity, the calibration set. This can be done by means of partial least squares (PLS), and the resulting model is used to predict the content and identity of new unknown samples from their digitized UV–Vis spectra. PLS is a regression technique that has been used in many different fields of chemistry and the theories behind it have been thoroughly described elsewhere (Geladi and Kowalski, 1986; Sjöström *et al*., 1983 and Brereton, 2000), together with the theories of multivariate calibration (Sjöström *et al*., 1983 and Brereton, 2000). The best calibration is obtained when the calibration model contains all the sources of variation that can occur in the actual measurement.

Preview

Multivariate calibration has historically been a major cornerstone of chemometrics as applied to analytical chemistry. However, there are a large number of diverse schools of thought. To some, most of chemometrics involves multivariate calibration. Certain Scandinavian and North American groups have based much of their development over the past two decades primarily on applications of PLS algorithm. At the same time, the classic text by Massart and co-workers (Massart *et al*., 1988) does not mention PLS, and multivariate calibration is viewed by some only as one of a large battery of approaches to the interpretation of analytical data. In Scandinavia, many use PLS for almost all regression problems (whether appropriate or otherwise) whereas related methods such as multiple linear regressions (MLR) are more widely used by mainstream statisticians. There has developed a mystique surrounding PLS, a technique with its own terminology, conferences and establishment. Although originally developed within the area of economics, most of its prominent proponents are chemists. There are a number of commercial packages on the market-place that perform PLS calibration and result in a variety of diagnostic statistics. It is important to understand that a major historic (and economic) driving force was near infrared spectroscopy (NIR), primarily in

Corresponding author: Email: lab9@gawab.com

the food industry and in process analytical chemistry. Each type of spectroscopy and chromatography has its own features and problems, so much software was developed to tackle specific situations which may not necessarily be very applicable to other techniques such as chromatography or NMR or MS. In many statistical circles NIR and chemometrics are almost inseparably intertwined. However, other more modern techniques are emerging even in process analysis, so it is not at all certain that the heavy investment on the use of PLS in NIR will be so beneficial in the future. Despite this, chemometric approaches to calibration have very wide potential applicability throughout all areas of quantitative analytical chemistry. There are many circumstances in which multivariate calibration methods are appropriate. The difficulty is that to develop a very robust set of data analytical techniques for a particular situation takes a large investment in resources and time, so the applications of multivariate calibration in some areas of science are much less well established than in others. It is important to distinguish the methodology that has built up around a small number of spectroscopic methods from the general principles applicable throughout analytical chemistry. There are a whole series of problems in analytical chemistry for which multivariate calibration is appropriate, but each is very different in nature.

Types of multivariate calibrations

Many methods for multivariate calibrations have been proposed. It turns out that many of the methods perform similarly. To avoid confusion due to use of many different methods, it is suggested that only the following should be considered.

1 Multiple linear regressions (MLR)
2 Principal component regression (PCR)
3 Partial least squares (PLS)
4 Neural networks (NN)
5 Locally weighted regression (LWR)
6 Radical basis function combined with PLS (RBF-PLS)

They have been included on the basis of theoretical considerations, confirmed by an intercomparison of their performances.

The simplest is calibration of the concentration of a single compound using a spectroscopic or chromatographic method, an example being determining the concentration of chlorophyll by electronic absorption spectroscopy (EAS) (Araujo *et al*., 1996). Instead of using one wavelength (as is conventional for the determination of molar absorptivity or extinction coefficients), multivariate calibration involves using all or several of the wavelengths.

A more complex situation is a multi-component mixture where all pure standards are available, such as a mixture of four pharmaceuticals (Demir and Brereton, 1998). It is possible to control the concentration of the reference compounds, so that a number of carefully designed mixtures can be produced in the laboratory. Sometimes the aim is to see whether a spectrum of a mixture can be employed to determine individual concentrations, and, if so, how reliably. The aim may be to replace a slow and expensive chromatographic method by a rapid spectroscopic approach. Another rather different aim might be impurity monitoring (Zissis *et al*., 1999), how well the concentration of a small impurity may be determined, for example, buried within a large chromatographic peak.

A different approach is required if only the concentration of a portion of the components is known in a mixture, for example, the polyaromatic hydrocarbons within coal tar pitch volatiles (Cirovic *et al*., 1996). In the natural samples there may be tens or hundreds of unknowns, but only a few can be quantified and calibrated. The unknown interferents cannot necessarily be determined and it is not possible to design a set of samples in the laboratory containing all the potential components in real samples. Multivariate calibration is effective provided that the range of samples used to develop the model is sufficiently representative of all future samples in the field. If it is not, the predictions from multivariate calibration could be dangerously inaccurate. In order to protect against samples not belonging to the original dataset, a number of approaches for determination of outliers have been developed.

A final case is where the aim of calibration is not so much to determine the concentration of a particular compound but a group of compounds, for example protein in wheat (Osborne and Fearn, 1986). The criteria here become fairly statistical and the methods will only work if a sufficiently large and adequate set of samples are available. However, in food chemistry if the supplier of a product comes from a known source that is unlikely to change, it is often adequate to set up a calibration model on this training set.

Disadvantage

There are many pitfalls in the use of calibration models, perhaps the most serious being variability in instrument performance over time. Each instrument has different characteristics and on each day and even hour the response can vary. How serious this is for the stability of the calibration model needs to be assessed before investing a large effort. Sometimes it is necessary to reform the calibration model on a regular basis, by running a standard set of samples, possibly on a daily or weekly basis. In other cases multivariate calibration gives

only a rough prediction, but if the quality of a product or the concentration of a pollutant appears to exceed a certain limit, then other more detailed approaches can be used to investigate the sample. For example, on-line calibration in NIR can be used for screening a manufactured sample, and any dubious batches investigated in more detail using chromatography. There are many excellent articles and books on multivariate calibration which provide greater details about the algorithms (Geladi and Kowalski, 1986; Martens and Naes, 1989; Wold *et al*., 1983; Höskuldsson, 1988; Manne, 1987; Brown, 1982; Beebe *et al*., 1998 and Wold *et al*., 1983).

Developing a relationship

There are n samples which were spectroscopically measured at p wavelengths. The data are present in an nxp data matrix 'X'. Moreover each sample is characterized by the values of the concentration or other characteristic, 'y'. These form together the **y** vector. The object of the multivariate calibration is to develop a model y = f(X), which can be used to predict the 'y' values of a new sample.

When the relationship between 'y' and some or all of the 'X' variables is non linear, then the data are said to show non linearity. PLS and PCR are able to some extent to model non linear behavior. NN are called universal approximators, which means that they are able to model also strongly non linear behavior. Non linear relationships can be locally approximated by linear models, so that local methods such as LWR and RBF-PLS can be applied also for non linear data.

At the selection stage, when the first preliminary investigation of the data is carried out, sometimes samples with very extreme characteristics may be detected. Such data points constitute gross outliers and they are often due to errors. These could influence the conclusions to such an extent that one needs to decide whether such a sample should be retained. It is possible for instance that all data except the extreme one show linear behavior, but that on the basis of the gross outlier one would conclude that there is non linearity. The decision to retain the sample must be made on the basis of practicality and of chemical reasoning. It should be clear however, that the region in data space described by the gross outlier is described by a single data point. It may then be preferable to limit the prediction range so that the gross outlier is not relevant anymore.

In principle one should not make predictions for data that are not inside the calibration domain, i.e. concentrations predicted should fall within the concentration range (the 'y' calibration domain) defined by the calibration samples. If the model is correct this also means that the sample to be predicted will show a signal, e.g. absorbance, within the signal range (the 'X' calibration domain) defined by the calibration samples. In multivariate calibration, this is not always easy to ascertain. It is for instance possible that a sample within the 'y' calibration domain includes a source of spectral variation, not present in the calibration samples, so that it is outside the 'X' calibration domain. In such cases, one may be required to extrapolate the model outside the calibration domain. This should be avoided if at all possible, but this is not always possible. Some methods are less robust towards extrapolation than others.

Preliminary investigation of data structure

When contemplating the development of a multivariate calibration model, one should first ask whether one should expect one of the data structures described above. For instance, when the data concern a few grades of the same product and the difference between grades are larger than within grades, clusters should be expected. Additionally one should decide whether extrapolation is allowed or expected and inquire about the number of samples that will be available to construct the model.

The following plots can help in reaching conclusions about the data structure.

i A histogram of the y-values. This will make apparent clusters and gross outliers in 'y'.
ii An overlay plot of the spectra. This will make apparent gross outliers and clear clusters in 'X'.
iii Score plots on the first principal components, e.g. PC1-PC2, PC2-PC3. This will provide additional insight in clustering or the presence of extreme data in 'X'.
iv A plot of the first few PC, e.g. PC1-PC5 or PLS components against 'y'. Strong non linearities will probably show up on one of these plots. Some PC may not be relevant to the calibration. Therefore, the fact that no relation is seen for some PC should not surprise.

It is recommended that, even if one has already decided to use a specific method (for instance, because it is the only one included in the available software), these plots should be obtained.

It should be understood that at this stage no exhaustive search into the presence of non-linearities, clusters and outliers is carried out. Minor non-linearities, clustering and less gross outliers do not need to be detected at this stage: they can be handled once the method has been selected in ways that are appropriate for that method.

Selection of method

MLR is the method which is best known from a statistical point of view. The method is based on selection of variables. It should certainly be preferred when the

selection of variables is simple, i.e. when some variables are rather selective for the compounds or characteristics being determined. When MLR is used with strongly collinear variables, as is the case with spectroscopic variables, there is what might seem a statistical paradox.

PCR is a well known method that is closest to MLR. It is a two step procedure. In the first step, one determines principal components that are linear combinations of the original variables. They can be considered as new variables that summarize in an optimal way the variation present in the spectra. In the second step, MLR is applied to the newly obtained latent variables. When collinearity between original variables occurs, principal component plots often allow better interpretation of the variations observed in the data set than plots of original variables selected by MLR. As a modeling method, it is somewhat less performant than MLR when performing prediction within the calibration domain and when the model is indeed linear, but it is more reliable if extrapolation may be required. It is very comparable in performance to PLS, but it is less rapid (Estienne et al., 2004).

PLS is the method that is often used in multivariate calibration. It resembles PCR, but works in one step. It also determines latent variables, that are linear combinations of the original variables, but the criterion applied is maximal covariance between the y - values and the spectral variables. Because of this criterion the algorithm yields models by an iterative procedure which is perceived by the user as a single regression step. As explained above, it generally performs equally well as PCR and has the same properties towards difficult data structures (Estienne *et al*., 2004). It follows that the method performs better than the linear methods described in the preceding sections when there is a pronounced non linearity in the relation between 'y' and 'X'. Their main disadvantage is that extrapolation outside the calibration domain can lead to very bad results. Additionally, it is more difficult to use NN for interpretation purposes.

Methods that are not considered

In the intercomparison, many other methods were tried out, namely non-linear variants of PLS, PCR and MLR, nearest-neighbor methods, methods in which the regression is carried out on Fourier coefficients or wavelets. Some of these do not yield good results, several others perform equally well as (some of) the methods selected, but are less easy to understand or are less generally used. For this reason their use is not recommended for users that are not experienced in chemometrics.

SIMCA classification

Soft independent modeling of class analogy (SIMCA) is a technique that uses principal component analysis (PCA) or PLS for classification (Dunn and Wold, 1995; Maesschalck *et al*., 1999; Mertens *et al*., 1994 and Kvalheim *et al*., 1983). Each class is modeled by a separate PCA analysis. From the residuals of the samples in the calibration set, a confidence region for the class is constructed around the principal component. New objects are regarded as a member of the class if their distance from the principal component space does not exceed a critical limit defined by the confidence region. Based on the residual variance of the objects in the calibration set, the residual standard deviation (s_0) is calculated (pooled residual standard deviation):

$$S_0 = \frac{\sqrt{\sum_{k=1}^{m}\sum^{p} e_{ik^2}}}{(m-1-r)(m-r-1)} \qquad (1)$$

Where e_{ik}^2 is the residual of object k in the calibration set at variable i, m is the number of observations in the calibration set, p is the number of variables and r is the number of principal components. The number of degrees of freedom given in equation 1 is used when the number of observations is less than the number of variables, $m \leq$ p (Kvalheim *et al*., 1983 and Maesschalck *et al*., 1999).

The absolute residual standard deviation (s_i) for an object in the calibration set when $m \leq p$ is given

$$S_i = \frac{\sqrt{\sum_{i=1}^{p} e_{ik^2}}}{(m-1-r)} \qquad (2)$$

The ratio s_i/s_0 in the software used in this study, Simca P 7.01, is called the normalised distance to model in X space and in the following text is denoted DmodX.

The normalised distance to model in X space values for new objects are calculated from the ratio s_i/s_0 and are called DmodXPS in the software. However, if the predicted score value of a new object is outside the model score range, the DmodXPS value is calculated as a combined measure of the object's residual standard deviation (s_i) and the distance of its score from the normal score range of the model (UGS, 2000; Wold and *Sjöström*, 1977 and Albano *et al*., 1981). In this study, no critical limit was used for classification; instead, simple comparisons between the calculated distances to model values for the samples in the external test sets compared with the calibration set were used. Higher values of DmodXPS of the samples in the test sets compared with the calibration set give a negative identity or outlier. The distance to model values calculated from the PLS regression was in this way used for the SIMCA classification.

One advantage of using this type of outlier detection is that the DmodXPS value gives a measurement of how the spectra of the sample fit into the multivariate calibration

model. This is a major advantage of the multivariate approach over univariate calibration.

Data pre-treatment

Several types of pre-treatments of spectroscopic data in multivariate calibration are described in the literature. The following pre-treatments are commonly used in UV/ Vis spectroscopy: multiplicative scatter correction (MSC) (Martens and Naes, 1989), standard normal variate (SNV) (Barnes *et al*., 1989) and orthogonal signal correction (OSC) (Wold *et al*., 1998). OSC removes variations from X that do not contain any information about Y, *i.e.*, removing data in X that is orthogonal against Y. As a result of this OSC also give a simpler calibration model with fewer components.

Multiple linear regressions (MLR)

Models constructed from spectroscopy are relatively simple due to linear combinations of the instrumental measurements which make the model a correlation-based model. Models for a broader range of conditions have been constructed in order to separate overlapping peaks elicited from the analyte plus other unknown components or conditions. These methods for separating "outliers" are based upon the following equation:

$$x_i = b_0 + b_1 * y_{i1} + b_2 * y_{i2} + ... + b_q * y_{iq} + e_i$$

where x_i = analyte level of the *i* th specimen, y_{ij} = *j* th instrumental measurement with the *i* th specimen, b = model parameters, and e_i = error associated with y_i.

Next, using Equation, $x(estimated)=a_o+a_{1*}\ y_1+a_2*\ y_2+\ldots\ldots a_q*y_q$ the analyte levels of new specimens can be predicted when estimated b_j is substituted for a_j.

MLR does not require knowledge of the amount of interference or other samples in the calibration set. But the maximum amount of measurements (q) used in this method is restricted to approximately 2 to 10 in most cases; therefore, selecting an appropriate set of instrumental measurements is paramount (Thomas, 1994). For example, in critical care environments, use of Vis-Near-IR spectroscopy noninvasively monitors oxygen saturation in arterial blood (Mendelson, 1983). One wavelength in the Visible spectrum monitors blood pulsatile volume and oxygen saturation; the other in the Near-IR measures pulsatile blood volume only. Since the interference in this case is the blood volume, the information from the two wavelengths can be used to filter out the interference and provide a measurement of the oxygen content.

Principle components regression (PCR) and partial least squares (PLS)

These methods belong to a class of statistical methods known as continuum regression. Models for contiuum regression techniques are described by:

$$x_i = b_0 + b_1 * t_{i1} + b_2 * t_{i2} + ... + b_h * t_{iq} + e_i$$

Where t_{ik} is the *k* th score affiliated with the *i* th specimen. Every score has a linear combination of the instrumental measurements, such that

$$t_{ik} = \gamma_{k1} * y_{i1} + \gamma_{k2} * y_{i2} + \ldots + \gamma_{kq} * y_{iq}$$

Where $gamma_{kj}$ are coefficients. The calibration set data determines these, along with the estimated model parameters (b) and the metaparameter, h, which is the model size and is usually much smaller than number of measurements, q (Thomas, 1994). The prediction step follows where **estimated** x and **estimated** b are correlated in the following equation:

$$x = b_0 + b_1 * t_1 + b_2 * t_2 + ... + b_h * t_h$$

where

$$t_k = \gamma_{k1} * y_1 + \gamma_{k2} * y_2 + \ldots + \gamma_{kq} * y_q$$

The prediction step equation is similar to Equation1, where estimated 'x' is simply a linear combination of the measurements affiliated with the new specimen.

The difference between PCR and PLS is the coefficients and how they are obtained: in PCR, only the calibration set measurements are used to acquire the coefficients, while in PLS, the measurements and the analyte values are used (Thomas, 1994). For information on how to calculate and obtain these coefficients, detailed references are provided (Martens and Naes, 1989; Höskuldsson, 1988).

Briefly, it is important to mention that before methods such as PCR or PLS are utilized, data pretreatment is usually necessary. The most prevalent and easiest type of transformation is centering the data. This method is executed for both the analyte values and the instrumental measurements. It tends to make subsequent calculations less prone to round-off and overflow problems, and it can reduce the size of the model by one factor (Thomas, 1994).

Calibration methods

A. Univariate calibration

1. Classical calibration.

There is a huge literature on univariate calibration (Höskuldsson, 1996; 1992; Thompson *et al*., 1994; Miller, 1991; Miller and Miller, 1988; Miller and Miller, 1988; Eisenhart, 1939 and Draper and Smith, 1981). One of the simplest problems is to determine the concentration of a single compound using the response of a single detector, for example a single spectroscopic wavelength or a chromatographic peak area. Mathematically a series of experiments can be performed to give $\boldsymbol{x \approx c.s}$

Where, in the simplest case, $\boldsymbol{x}$ is a vector consisting of

absorbance at one wavelength for a number of samples (or the response), and ***c*** is of the corresponding concentrations. Both vectors have length *I*, equal to the number of samples. The scalar *s* relates these parameters and is determined by experiments. A simple method for solving this equation is as follows:

$\boldsymbol{c}'.x \approx \boldsymbol{c}'.c\ .s$ so $(\boldsymbol{c}'.c)^{-1}.\boldsymbol{c}'.\ x \approx (\boldsymbol{c}'.c)^{-1}.\ (\boldsymbol{c}'.c).\ s$

$$S \approx (c'c)^{-1}\, c'.x = \frac{\sum_{i=1}^{I} x_i c_i}{\sum_{i-1}^{I} c_i^2} \qquad (2)$$

Where the $'$ is the transpose.

Many conventional texts use summations rather than matrices for determination of regression equations, but both approaches are equivalent.

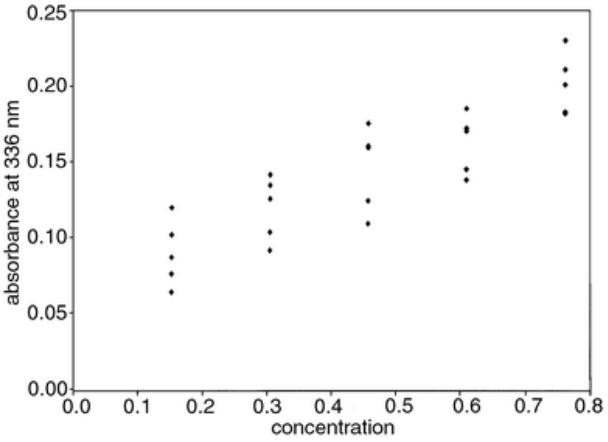

Fig. 1

In fig. 1 (Draper and Smith, 1981), the absorbance of the spectra of case study 1A at 336 nm is plotted against the concentration of pyrene. The graph is approximately linear, and provides a best fit slope calculated by

$$\sum_{i=1}^{I} x_i c_i = 1.849$$

and

$$\sum_{i=1}^{I} x_1^2 = 6.354$$

so that $\hat{x} = 0.291\ \hat{c}$. Note the hat (^) symbol indicates a prediction.

The quality of prediction can be determined by the residuals (or errors) *i.e.* the difference between the observed and predicted, *i.e.* $x - \hat{x}$; the less this is the better. Generally the root mean error is calculated,

$$E = \sqrt{\sum_{i=1}^{I} (x_i - \hat{x}_i)^2 / d}$$

where *d* is called the degrees of freedom. In the case of univariate calibration this equals the number of observations (*N*) minus the number of parameters in the model (*P*) or in this case, 25 − 1 = 24, so that

$$\sqrt{0.0289/24} = 0.0347$$

This error can be represented as a percentage of the mean $E_{\%} = 100\ (E/\bar{x}) = 24.1\%$ in this case. It is always useful to check the original graph (Fig. 1) just to be sure, which appears a reasonable answer. Note that classical calibration is slightly illogical in analytical chemistry. The aim of calibration is to determine concentrations from spectral intensities, and not *vice versa* yet the calibration equation in this section involves fitting a model to determine a peak height from a known concentration.

For a new or unknown sample, the concentration can be estimated (approximately) by using the inverse of the slope or $\hat{c} = 3.44\ x$

The spectrum of pure pyrene is given in fig. 2 (Brereton, 2000), superimposed over the spectra of the other compounds in the mixture. It can be seen that the wavelength chosen largely represents pyrene, so a reasonable model can be obtained by univariate methods. For most of the other compounds in the mixtures this is not possible, so a much poorer fit to the data would be obtained.

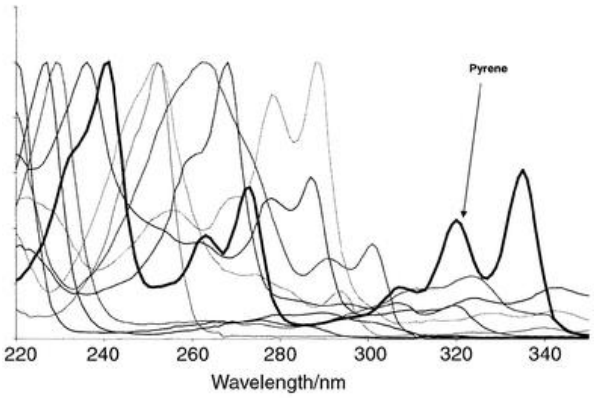

Fig. 2: Spectrum of pyrene superimposed over the spectra of the other pure PAHs.

2. ***Inverse calibration***

Although classical calibration is widely used, it is not always the most appropriate approach in analytical

chemistry, for two main reasons. First, the ultimate aim is usually to predict the concentration (or factor) from the spectrum or chromatogram (response) rather than *vice versa*. There is a great deal of technical discussion of the philosophy behind different calibration methods, but in other areas of chemistry the reverse may be true, for example, can a response (*e.g.* a synthetic yield) be predicted from the values of the independent factors (*e.g.* temperature and pH). The second relates to error distributions. The errors in the response are often due to instrumental performance. Over the years, instruments have become more reliable. The independent variable (often concentration) is usually determined by weighing, dilutions and so on, and is often the largest source of error. The quality of volumetric flasks, syringes and so on has not improved dramatically over the years. Classical calibration fits a model so that all errors are in the response [fig. 3(a)] (Brereton, 2000), whereas with improved instrumental performance, a more appropriate assumption is that errors are primarily in the measurement of concentration fig. 3b) (Brereton, 2000).

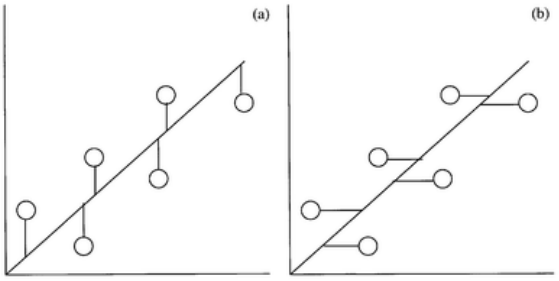

Fig. 3: Errors in (a) Classical and (b) Inverse calibration.

Calibration can be performed by the inverse method where $\boldsymbol{c} \approx \boldsymbol{x}.b$ or

$$b = (x'.x)^{-1}.x'.c = \frac{\sum_{i=1}^{l} x_i c_i}{\sum_{i-1}^{l} x_i^2}$$

Giving for this example, $\hat{c} = 3.262\ x$. Note that b is only approximately the inverse of s, because each model makes different assumptions about error distributions. However, for good data, both models should provide fairly similar predictions, if not there could be some other factor that influences the data, such as an intercept, non-linearities, outliers or unexpected noise distributions. For heteroscedastic noise distributions (Massart *et al*., 1997), there are a variety of enhancements to linear calibration. However, these are rarely taken into consideration when extending the principles to the multivariate calibration.

Including the intercept

In many situations it is appropriate to include extra terms in the calibration model. Most commonly an intercept (or baseline) term is included to give an inverse model of the form $c \approx b_0 + b_1 x$ which can be expressed in matrix/vector notation by $c \approx X\ .\ b$ for inverse calibration where $\boldsymbol{c}$ is a column vector of concentrations and $\boldsymbol{b}$ is a column vector consisting of two numbers, the first equal to b_0 (the intercept) and the second to b_1 (the slope). $\boldsymbol{X}$ is now a matrix of two columns, the first of which is a column of 1's, the second the spectroscopic readings.

Applications

An initial approach to the simultaneous determination of several amino acids by using multivariate calibration methods is proposed by Saurina (1995). The method is based on the development of the reaction between amino acids and 1,2-naphthoquinone-4-sulfonate. Spectra of the corresponding derivatives obtained from a continuous-flow system were used to provide multivariate data for the multivariate procedure. Two multivariate calibration methods were applied, principal component regression (PCR) and partial least-squares regression (PLS).

Multivariate chromatographic calibration technique is a powerful mathematical tool for optimizing chromatographic multivariate calibration and elimination of fluctuations coming from instrumental and experimental conditions. Multivariate chromatographic calibration technique was subjected to HPLC data for simultaneous analysis of binary mixture of clopidogrel and aspirin (Sultana *et al*., 2007).

HPLC data based on the analyte peak areas were obtained at five wavelengths. The mathematical algorithm of multivariate chromatographic calibration technique is based on the use of the linear regression equations. In this study, classical HPLC provided good results than multivariate HPLC as mentioned in the table.

Another multivariate HPLC-UV technique (Arayne *et al*., 2007) anticipated that under optimized conditions considerable resolving power, sensitivity, rapidity, quality control and routine analysis of subject compounds is achieved. This is a reversed-phase liquid chromatographic method for the determination of levofloxacin contents in pharmaceutical formulations and in human serum has been developed and validated which has ability to determine levofloxacin concentration up to 40 ng/ml. The developed method is very simple and results obtained confirm suitable accuracy, specificity and precision. Therefore, the developed RP HPLC method was proved to be suitable for the levofloxacin determination in routine Quality control analysis as well as R&D.

In this case it is evident from table that the concentration range effects the results obtained through multivariate HPLC.

Table: Multivariate precision and accuracy of clopidogrel and aspirin

Added conc. Found concentration (μg mL-1)						Multi-variate calibra-tion	% accuracy					Multi-variate calibra-tion
Clopidogrel	225	230	235	240	245		225	230	235	240	245	
0.25	0.242	0.246	0.249	0.249	0.247	0	96.98	98.44	99.69	99.67	99.14	0
1.0	0.985	0.985	0.999	1.00	0.994	0	98.55	98.58	99.94	100.10	99.43	0
5.0	4.95	5.01	5.00	4.96	4.99	6.57	99.11	100.32	100.01	99.32	99.97	131.43
12.5	12.12	12.82	12.64	12.90	12.63	12.57	97.01	102.60	101.15	103.27	101.11	100.56
25.0	25.10	25.00	25.35	24.40	25.17	26.26	100.40	100.02	101.40	97.61	100.70	105.04
50.0	50.05	50.03	50.02	49.96	50.35	49.23	100.11	100.07	100.04	99.92	100.71	98.46
						Mean	98.69	100.00	100.37	99.98	100.17	108.87
						RSD	1.47	1.50	0.71	1.84	0.78	3.50
						RE	1.30	-0.005	-0.37	0.01	-0.17	-8.87
Aspirin												
0.25	0.250	0.252	0.252	0.257	0.259	0.196	100.09	101.00	100.85	103.02	103.95	78.46
1.0	1.00	1.01	0.994	1.00	1.03	0.023	100.07	101.07	99.42	100.02	103.30	1.21
5.0	4.99	4.99	5.00	4.98	5.01	3.482	99.80	99.96	99.70	99.70	100.03	69.64
12.5	12.06	12.34	13.06	12.86	12.97	13.28	96.48	98.73	104.49	102.88	103.78	106.29
25.0	25.10	24.97	25.33	24.81	25.08	20.24	100.41	99.91	101.32	99.24	100.34	80.99
50.0	49.76	49.77	49.77	50.01	48.47	50.46	99.92	99.92	100.96	100.03	96.95	100.92
						Mean	99.46	100.09	100.96	100.81	100.39	72.91
						RSD	1.48	0.86	1.86	1.66	2.73	51.84
						RE	0.53	-0.09	-0.81	-1.39	-1.39	27.08

A simple multivariate spectral calibration method for the quantitative analysis of gliquidone in methanol was developed by Arayne *et al.* (2007) and has been compared with the classical ultraviolet spectrophotometric method for estimation of the drug. In this developed method, analytical parameters such as selectivity, accuracy and precision have been established and evaluated statistically to assess the application of the method. The method was found to have advantage for simplicity, stability, sensitivity (0.1ppm), reproducibility (±0.1, n=21) and accuracy for using as an alternate to the existing spectrophotometric and non-spectrophotometric methods for the routine analysis of the drug and its pharmaceutical formulations. This method is a powerful tool with very simple mathematical content, is more reliable than other spectrophotometric methods.

Andrew and Worsfold (1994) found that relative precisions of five multivariate calibration methods [direct multicomponent analysis (DMA), stepwise multiple linear regression (SMLR), principal components regression (PCR), and PLS1 and PLS2 (PLS = partial least-squares)] are evaluated for the determination of transition metal ions in model multicomponent systems. These systems represent simulated industrial process streams containing mixtures of two, three and five metal ions. Physical and chemical interferences have been incorporated to provide a rigorous test of the calibration techniques. Diode-array spectrophotometry has been used to obtain spectra of the inherent absorbances of the metal ions in the visible region. Multivariate calibration models have been constructed from these data and used to predict concentrations of metal ions in test solutions. Calibrations based on first-derivative data afforded greater precision than those based on absorbance data.

An empiric procedure of successive approximations for the selection of the most convenient wavelengths and instrumental parameters to execute an UV multivariate analysis is described by Vetuschi and Ragno (2002). The selection criterion was according to the minimum PRESS value. For the identification of the best wavelength intervals a tree-fraction scheme has been used, starting from the full spectrum and going down to 3 nm intervals. General considerations for applying this procedure on any multicomponent mixture are finally reported. The study has been developed on a mixture of propyphenazone, paracetamol and caffeine, present in several pharmaceuticals, whose spectra so totally overlap to make inconsiderable the classical spectrophotometric methods. The obtained results were enthusiastic demonstrating a very high improvement of accuracy and precision. The mean error on propyphenazone, paracetamol, caffeine passed respectively from 12.8, 5.49, 41.88 %, reading the full spectrum at the default conditions, to 1.06, 0.52, 1.18 %, employing the selected conditions (Vetuschi and Ragno, 2002). In-line Raman, near infrared and UV-visible spectometries, and at-line low-field NMR spectrometry have been used to monitor the acid-catalysed esterification of crotonic acid and butan-2-ol by Colin A. McGill and others (McGill *et al*., 2002).

In forensic science, soil is an important form of trace evidence and as such it can serve to link a potential source of it with an item on which it is found. Soil samples are often found as deposits on shoes and surfaces or smears on clothing and skin. Since soil can easily get trapped on the undersole of footwear and on vehicle tires, it can possibly be transferred between the whereabouts of a potential suspect to the scene of the crime and vice versa. Provided that an appropriate method for the characterization of the samples is available, a useful link could be established between the suspect and the crime scene and corroborative evidence could emerge (Lombardi, 1999).

Thanasoulias and others (2002) analyzed 44 soil samples from five different areas on the basis of the UV-Vis spectrum of the acid fraction of humus with a view to achieve good discrimination between them. Principal component analysis (PCA) for the removal of outliers was applied on the data. An overall 85% correct classification of the training dataset was observed (Wilks' λ=0.0420) and the method was judged satisfactory for supporting exclusionary forensic purposes.

A new method for the rapid determination of pharmaceutical solutions is proposed by Wiberg *et al* (2000). A conventional HPLC system with a Diode Array Detector (DAD) was used with no chromatographic column connected. The method was tested on the local anaesthetic compound lidocaine. The UV spectrum (245-290 nm) from the samples analysed in the detector was used for multivariate calibration for the determination of lidocaine solutions. The content was determined with PLS regression. The results show that in respect of accuracy, precision and repeatability the new method is comparable to the reference method. The main advantages compared with liquid chromatography are the much shorter time of analysis (<30 s) as well as the automatic and simple analytical procedure and the low consumption of organic solvents (Wiberg *et al*., 2003).

The extraction of amino acids at 298 K through hydrophobic liquid membranes is studied using the bulk liquid membrane (BLM) method. The transport of different mixtures of phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp) in a buffer solution (pH 1.5) was monitored with a diode array spectrometer by recording their UV-vis spectra at the receiving phase in small time intervals for 30 min. The ordinary least squares (OLS) method was applied separately to each mixture, using the experimental spectra of pure compounds. The transport rates, which are given by the slope of the kinetic curves, were found to be very small, of the order of 10^{-5}, 10^{-6} and 10^{-7} mol l^{-1} min^{-1} for Phe, Trp and Try respectively. The trilinear decomposition (TLD) method was used to estimate the kinetic profiles for each compound in the ternary mixtures and also their transport rates. They are in fair agreement with those obtained by OLS. The transport rates are of the same order of magnitude as those obtained for the ternary mixtures, and in this case the agreement between TLD and OLS results is excellent (Alexandre *et al*., 2002).

In another study Dinç and coworkers (2003) developed a new method of quantitative two-component analysis of a mixture of hydrochlorothiazide (HCT) and spironolactone (SP) in the presence of strongly overlapping signals by using one-dimensional continuous wavelet analysis. This new method was built on the simultaneous use of both continuous wavelet transform and the zero crossing technique for the quantitative resolution of this binary mixture. The absorption spectra of the standard solutions were recorded in the wavelength range of 215–330 nm. They selected 400 points from the absorption spectra and subjected the corresponding signal to Daubechies2 (DAUB2) and Biorthogonal1.5 (BIOR1.5) one-dimensional wavelet transform. In the transformed signals, the amplitude of HCT was measured at the position of a zero crossing point of SP and vice versa. Thus simple linear regression analysis can be applied to establish calibrations for both components. These calibrations were validated with synthetic mixtures of HCT and SP. *MATLAB 6.5* software was used for one-dimensional wavelet analysis, and the basic concepts about wavelet method are briefly explained. A second-order reaction between benzophenone and phenylhydrazine to give benzophenone phenylhydrazone was followed using UV/vis and mid-infrared spectroscopic probes. Established kinetic and partial least squares modeling chemometrics methods were applied to both datasets in order to compare the information acquired with each probe. It is shown that several different spectroscopic techniques can be used in reaction monitoring, with increasing benefits in terms of information and interpretation of the results. The profiles obtained agreed well which was also demonstrated when comparing the different rate constants obtained (Carvalho *et al*., 2006).

The influence of metal ions (Na^+, Mg^{2+} and Cd^{2+}) on the thermal unfolding of phenylalanine transfer ribonucleic acid (tRNA(Phe)) was studied by UV spectroscopy-monitored melting experiments. Absorbance data were obtained during the unfolding process in the range 220-340 nm and later analyzed by a multivariate curve resolution approach (MCR-ALS) based on factor analysis. This procedure determines the number of spectroscopically distinct conformations present during the unfolding process and reveals their concentration profiles and pure spectra, without any initial assumption having to be made about the number of steps in the unfolding pathway. From the concentration profiles and pure spectra, information such as T(m) values can be recovered. The results were compared with those obtained

previously in spectroscopic and calorimetric unfolding experiments, showing that the multivariate approach recovers information that complements that obtained in traditional spectroscopic melting experiments (Vives *et al*., 2002). Anti-bacterial drugs such as sulfonamides and trimethoprim, are commonly used in medical and veterinary applications. The tablets often contain mixtures of drugs such as sulfadiazine, sulfadimidine, sulfamethoxazole, sulfanilamide and trimethoprim. For the simultaneous determination of these drugs this method was researched and developed with the aid of chemometric methods comparing the classical least squares (CLS), principal component regression (PCR) and partial least squares (PLS) models. The % Rating of Perceived Exertion (RPE) values for analysis of the individual drugs in mixtures were in the range of 2-10 with % recoveries being close to 100 for the PCR and PLS models while the CLS method performed relatively poorly. The four PCR and the PLS models constructed could not be distinguished on prediction performance for practical purposes, and the PCR calibration based on the raw data matrix was utilised to illustrate successfully the application of the proposed method for analysis of the drugs in commonly available pharmaceutical tablets (Yongnian *et al*., 2006).

The formation of Pd nanoclusters in solution is studied. The system has two types of light-absorbing species: Pd ions which absorb light *via* electronic transitions and Pd clusters and aggregates which absorb light *via* valence–conduction transitions and also scatter light due to their nanometric dimensions. Here we monitor these dynamic changes using UV-visible spectroscopy. The reduction and clustering concentration profiles are extracted from the raw data using a combination of net analyte signal (NAS) and principal component analysis (PCA) methods. The *in situ* approach enables the quantification of both the reduction of the Pd^{2+} ions and the growth of the Pd clusters. Kinetic models that account for ion reduction, cluster growth and aggregation are presented and the influence of the counteranions and the reducing agents on these processes is discussed (Gaikwad and Rothenberg, 2006).

CONCLUSIONS

This paper has illustrated how spectroscopic techniques can be used in concurrence with chemometric tools in order to achieve rapid and simple analytical methods for the analysis of solutions. The strength of multivariate calibration and SIMCA classification on UV-Vis data for determination of the content and identity of pharmaceutical solutions is unmatched. The paper illustrated how the same calibration model could be used for both quantitative and qualitative determinations. The results obtained were comparable or even slightly better than the reference method, HPLC. It was further shown that UV/ Vis absorption spectra contain a great deal of information that can be utilised for classification and identity determination. The main advantages of the method proposed are rapid analysis, automation and the absence of organic solvents. UV-Vis absorption spectroscopy can be used for rapid and automated quantitative and qualitative determination of solutions with the help of chemometric techniques. Through investigation of multivariate residuals and use of control samples, the calibration model could also be used over time. Moreover it is shown that how chemometric tools such as experimental design, PCA, PLS, SIMCA and PARAFAC can be used to obtain efficient multivariate spectroscopic methods for the analysis of solutions.

REFERENCES

Albano C, Blomqvist G, Coomans D Dunn WJ III, Edlund U, Eliasson B, Hellberg S, Johansson E, Nordén B, Johnels D, Sjöström M, Söderström B, Wold H and Wold S (1981). *Proceedings of the Symposium on Applied Statistics, Copenhagen.*

Alexandre M, Antunes MM, Ferreira C and Pedro LO (2002). Chemometric study of amino acid transport through liquid membranes using UV-Vis spectroscopy. *Journal of Chemometrics*, **16**: 111-116.

Andrew KN and Worsfold P (1994). J. Comparison of multivariate calibration techniques for the quantification of model process streams using diode-array spectrophotometry. *Analyst*, **119**: 1541-1546.

Araujo PW, Cirovic DA and Brereton RG (1996). Multivariate Calibration of Chlorophyll a using PLS and Electronic Absorption Spectroscopy, *Analyst*, **121**: 581-584

Arayne MS, Sultana N and Siddiqui FA. Optimization of levofloxacin analysis by RP-HPLC using multivariate calibration technique. *Pak. J. Pharm. Sci.*, **20**(2): 100-106.

Arayne MS, Sultana N, Mirza AZ and Siddiqui FA. Optimization of quantitative analysis of gliquidone using multivariate spectral calibrations in bulk drug and dosage formulations (submitted to Journal of Chinese Chemical Society).

Barnes RJ, M Dhanoa S and Lister SJ (1989). Standard normal variate transformation and De-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.*, **43**: 772-777.

Beebe KR, Pell RJ and Seasholtz MB (1998). *Chemometrics: A Practical Guide*, Wiley, New York.

Brereton RG (2000). Introduction to Multivariate Calibration in Analytical Chemistry. *Analyst*, **125**: 2125-2154.

Brereton RG (2000). Introduction to multivariate calibration in analytical chemistry. *Analyst*, **125**: 2125-2154.

Brown PJ (1982), Multivariate calibration. *J. R. Stat. Soc., Ser. B.*, **44**: 287-321.

Carvalho D, Antonio R, Nogal D, Miguel S, Jirut W and Bretereton RG (2006). Comparison of PLS and kinetic models for a second-order reaction as monitored using ultraviolet visible and mid-infrared spectroscopy. *Talanta*, **68**: 1190-1200.

Cirovic DA, Brereton RG, Walsh PT, Ellwood J and Scobbie E (1996). Application of PLS calibration to Measurements of Polycyclic Aromatic Hydrocarbons in Coal Tar Pitch Volatiles, *Analyst*, **121**: 575-580.

Demir C and Brereton RG (1998). Multivariate calibration on designed mixtures of four pharmaceuticals. *Analyst*, **123**: 181-189.

Dinç E, Baleanu D and Üstündağ Ö (2003). An Approach to Quantitative Two Component Analysis of a Mixture Containing Hydrochlorothiazide and Spironolactone in Tablets by OneDimensional Continuous Daubechies and Biorthogonal Wavelet Analysis of UV Spectra. *Spectroscopy Letters*, **36**: 341-355.

Draper RN and Smith H (1981). *Applied Regression Analysis*, Wiley, New York, 2nd Edition.

Dunn WJ and Wold S (1995). *Methods Princ. Med. Chem.*, **2**: 179.

Eisenhart C (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Stat.*, **10**: 162.

Estienne F, Despagne F, Walczak B, Denoord OE and Massart DL (2004). A comparison of multivariate calibration techniques applied to experimental NIR data sets. Part III: Robustness against instrumental perturbation conditions. (*Chemometr. intell. lab. syst*.), **73**: 207-218.

Gaikwad AV and Rothenberg G (2006). *In-situ* UV-visible study of Pd nanocluster formation in solution. *Phys. Chem. Chem. Phys.*, **8**: 3669-3675.

Geladi P and Kowalski BR (1986). Partial least squares regression: A tutorial. *Anal. Chim. Acta.*, **185**: 1-17.

Höskuldsson A (1988). PLS regression methods. *J. Chemometr.,* **2**: 211-228.

Höskuldsson A (1988). PLS regression methods. *J. Chemometr*., **2**: 211-228.

Höskuldsson A (1992). The H-Principle in Modeling With Applications to Chemometrics. *Chemometrics Intell. Lab. Syst.*, **14**: 139-153.

Höskuldsson A (1996). Prediction methods in science and technology. Vol.1 - basic theory, Thor Publishing: Warsaw, Poland.

Kvalheim OM, Oygard K and Grahl-Nielsen O (1983). SIMCA multivariate data analysis of blue mussel. *Anal. Chim. Acta*, **150**: 145-157.

Lombardi G (1999). The contribution of forensic geology and other trace evidence analysis to the investigation of the killing of Italian Prime Minister Aldo Moro. *J. Forensic Sci.*, **44**: 634-642.

Maesschalck R De, Candolfi A, Candolfi DL, Massart DL and Heuerding S (1999). Decision criteria for SIMCA applied to near infrared data. *Chemom. Intell. Syst*., **47**: 65-77.

Maesschalck R De, Candolfi A, Massart DL and Heuerding S (1999). Decision criteria for SIMCA applied to near infrared data. *Chem. Intell. Lab. Syst*., **47**: 65-77.

Manne R (1987). Analysis of two partial-least squares algorithms for multivariate calibration. Chemometrics *Intell. Lab. Syst.,* **2**: 187-197.

Martens H and Naes T (1989). *Multivariate Calibration*, Wiley, New York.

Martens H Naes T (1989). *Multivariate Calibration*; Wiley: Chichester, England.

Massart DL, Vandeginste BGM, Buydens LMC, Jong S de, Lewi PJ and Smeyers-Verbeke JB (1997). *Handbook of Chemometrics and Qualimetrics, Part A*, Elsevier, Amsterdam.

Massart DL, Vandeginste BGM, Deming SN, Michotte Y and Kaufman L (1988). *Chemometrics: A Textbook*, Elsevier, Amsterdam.

McGill CA, Nordon A and Littlejohn D (2002). Comparison of in-line NIR, Raman and UV-visible spectrometries and at-line NMR spectrometry for the monitoring of an esterification reaction. *Analyst*, **127**: 287-292.

Mendelson Y (1983). Ph. D. dissertation, Case Western University, Cleveland, OH.

Mertens B, Thompson M and Fearn T (1994). Principal component outlier detection and SIMCA: a synthesis. *Analyst*, **119**: 2777-2784.

Miller JC and Miller JN (1988). Basic statistical methods for analytical chemistry. Part I. Statistics of repeated measurements: A review. *Analyst*, **113**: 1351-1356.

Miller JN (1991). Basic statistical methods for analytical chemistry. Part 2. Calibration and regression methods. A review. *Analyst*,**116**: 3.

Miller JN and Miller JC (1988). *Statistics for Analytical Chemistry*, 2nd Edition, Ellis. Horwood, Chichester.

Osborne BG and Fearn T (1986). *Near Infrared Spectroscopy in Food Analysis*, Longman, London.

Saurina J (1995). Hernandez-Cassou S. Simultaneous determination of several amino acids with multivariate calibration methods by using a continuous-flow system, *Analyst*, **120**: 305-312.

Sjöström M, Wold S, Lindberg W, Persson J-Å, HM (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Anal. Chim. Acta.*, **150**: 61-70.

Sultana N, Arayne MS, Kiran AA and Nawaz M. Simultaneous determination of clopidogrel and aspirin by RP- HPLC from bulk material and dosage formulations using multivariate calibration technique (submitted in the *Acta. Chromatographica*, Poland).

Thanasoulias NC, Piliouris ET, Melina-Spyridoula EK and Nicholas P (2002). Evmiridis. Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus. *Forensic science International*, **130**: 73-83.

Thomas E (1994). A primer on multivariate calibration. *Analytical Chemistry*, **66**: 797.

Thompson M, Brown DW, Fearn T, Greenhow MJ, Howarth R, Miller JN, Newman EJ, Riple BD. Y, Swan KJ, Williams A, Wood R and Wilson JJ (1994). Is my calibration linear? *Analyst,* **119**: 2363.

Users Guide to Simca-P 7.01, Umetrics, Umeå, Sweden (2000).

Vetuschi C and Ragno G (2002). Improvement of multivariate UV analysis by selecting wavelengths and instrumental parameters. *Analytical Letters*, **35**: 559-57.

Vives M, Tauler R and Gargallo R (2002). Study of the influence of metal ions on tRNA(Phe) thermal unfolding equilibria by UV spectroscopy and multivariate curve resolution. *J. Inorg. Biochem.,* **10**: 89(1-2): 115-22.

Wiberg K, Hagman A and Jacobsson SP (2003). Rapid determination of lidocaine solutions with non-column chromatographic diode array UV spectroscopy and multivariate calibration. *J. Pharm. Biomed. Anal.,* **30**: 1575-86.

Wold S and Sjöström M and Kowalski B (1977). Chemomefrics: Theory and Practire, *ACS Symp.,* **52**: 243-282.

Wold S, Albano C, Dunn III WJ, Esbensen K, Hellberg S, Johansson E and Sjøstrøm M (1983). *Food Research and Data Analysis*, H. Martens and H. Russworm, Applied Science Publishers, London, p.147.

Wold S, Antti H, Lindgren F and Öhman J (1998). Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.*, **44**: 175-185.

Wold S, Martens H and Wold H (1983). *Proc. Conf. Matrix Pencils, Lecture Notes in Mathematics*, Springer-Verlag, Heidelberg, p.286.

Yongnian NI, Zhengbao QI and Serge K (2006). Simultaneous ultraviolet-spectrophotometric determination of sulfonamides by multivariate calibration approaches. *Chemometrics and intelligent laboratory systems*, **82**: 241-247.

Zissis KD, Brereton RG, Dunkerley S and Escott REA (1999). Two-way, unfolded three-way and three-mode partial least squares calibration of diode array HPLC chromatograms for the quantitation of low-level pharmaceutical impurities. *Analytica. Chimica. Acta.*, **384**: 71-81.

Received: 22-03-2007 – Accepted: 03-04-2007