*Analytica Chimica Acta*, 223 (1989) 75–93
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands


# CHEMOMETRICS IN NEAR-INFRARED SPECTROSCOPY

HOWARD MARK
*Bran & Luebbe, 103 Fairview Industrial Park, Elmsford, NY 10523 (U.S.A.)*
(Received 18th May 1988)

SUMMARY

Spectroscopy methods of chemical analysis are excellent for the application of chemometric methods, because the measurements at many different wavelengths provide inherently multivariate data. The chemist generally requires three categories of information from specimens under investigation: quantitative data, qualitative data, and fundamental information on the properties of the material. Spectroscopy has long been used for all three purposes; the recent application of chemometric algorithms has assisted greatly in these endeavors. Although there is some overlap, three chemometric methods correspond to the three types of information: multiple regression, discriminant analysis, and principal components analysis. The basis of these chemometric methods and some of their strengths and limitations in application to near-infrared spectroscopy are discussed.

The science of what is now called chemometrics, the application of multivariate mathematical/statistical techniques to chemical problems, developed slowly as an evolutionary process from a background of statistical investigation, separating itself from other branches of statistics in the early 1970s [1]. The application of multivariate statistical techniques to near-infrared (NIR) spectroscopy, however, was developed independently from research at the U.S. Department of Agriculture [2]. This application of chemometrics is certainly one of the earliest to be put to practical use, and is among the most successful applications of chemometric algorithms. The best current estimate of the number of published scientific papers dealing with various aspects of NIR usage is over 1200 [3], and the number of NIR instruments in routine operation is over 10 000 worldwide [4].

In order to be of value to chemists, chemometrics must address relevant problems. These problems tend to fall into three classes: qualitative analysis, quantitative analysis, and understanding the principles underlying the observable phenomena. Numerous chemometric techniques have been developed over the years, and recently they have been compiled in monographs (see, e.g. [1]) These techniques have been applied in many areas of chemistry, including

0003-2670/89/$03.50 

various spectroscopies. Spectroscopic methods are eminently suitable for application of chemometric algorithms, because they are inherently multivariate, with many wavelengths (or frequencies) available at which to measure responses.

Several chemometric techniques have been applied to each of the three classes of chemical problems mentioned above. When applied to NIR spectroscopy, some overlap has been recognized, but in general, particular techniques have been applied to particular types of problems. As will be seen, there is considerable overlap between several of the chemometric techniques that are in common use, but in NIR studies they tend to be considered distinct. Discriminant algorithms have been applied to problems in qualitative analysis, multiple regression algorithms have been applied to quantitative problems, and principal component analysis has been the method of choice in studies of underlying principles. While some papers dealing with application of principal components have appeared [5–7] and Martens and Martens [8] have generated interest in applying partial least squares to quantitative NIR spectroscopy, these methods are not yet widely used.

Parts A and B of Fig. 1 show the NIR spectra of water, methanol and acetic acid in the 1100–2500-nm and 600–1400-nm regions. The three materials are completely miscible and exhibit marked differences in their NIR spectra, so that their mixtures are eminently suitable for illustrating the similarities and differences between various chemometric algorithms.

## EXPERIMENTAL

Because of the large differences in absorption coefficients of the three liquids used between the 1100–2500-nm and 600–1400-nm regions, the spectra in these two regions had to be obtained separately, with sample cups of suitable pathlength for each region. For the longer-wavelength region, a standard liquid drawer accessory for the NIR instrument was used; the sample cell in this drawer had a spacing of ca. 0.075 mm. For the short-wavelength region, the special sample cell used was made of Teflon and had a spacing of 10 mm.

The chemicals were reagent-grade (Aldrich); the water was treated with deionizing resin. Mixtures were prepared on a % (w/w) basis.

All measurements were made with a Technicon InfraAlyzer model 500. Data were collected at 4-nm intervals.

## RESULTS

Each sample was read twice in the instrument, to generate an estimate of the noise contribution to the data. As will be seen in the figures, the two readings from each sample coincide in virtually all cases, indicating that the noise is very small compared to other effects. The discussion will therefore ignore the random noise.

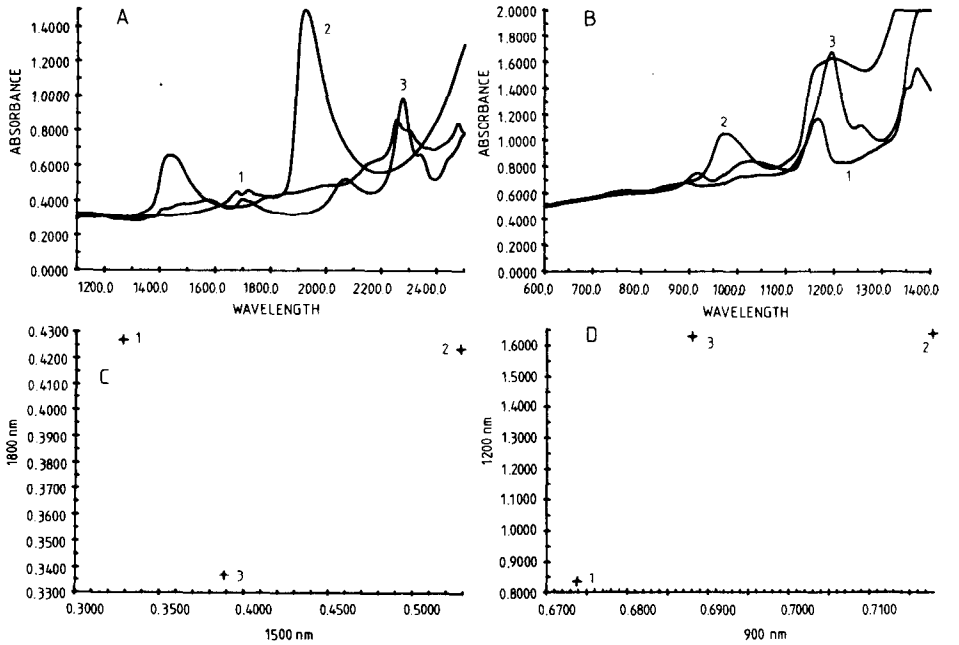


Fig. 1. Data from the pure materials used. (A) The spectra in the 1100–1500 nm region; (B) the spectra in the 600–1400 nm region; (1) acetic acid; (2) water; (3) methanol. (C) Absorbance/absorbance plot of the data at 1500 nm and 1800 nm; (D) absorbance/absorbance plot of the data at 900 nm and 1200 nm.

*Disciminant analysis*

The key to achieving qualitative analysis by NIR spectroscopy is the application of the multivariate algorithms that are classified as "supervised learning" [1]; the supervised training method used in NIR spectroscopy is discriminant analysis [8], through the use of multidimensional distance measures [10].

The approach needed is to examine the spectrum from the point of view of the mathematician rather than that of the chemist. In this vein, the spectra of Fig. 1, for example, can be examined in the following way. The data at certain wavelengths will allow a computer to distinguish among the various materials of interest, and thus classify the samples into their respective categories. The spectral data for the three materials shown in Fig. 1 constitute the training set, and the materials can be distinguished by using data from only two wavelengths. The three materials have suitable characteristics at several pairs of wavelengths. Suitable pairs include 950 nm and 1200 nm in the short-wavelength region, and 1500 nm and 1800 nm in the long-wavelength region. The spectra presented in Fig. 1 show that water and acetic acid have similar values

of absorbance at 1800 nm while methanol has a smaller absorbance. At 1500 nm, the absorbances increase in the order acetic acid, methanol, water. In the short-wavelength region, methanol and water have the same absorbance at 1200 nm while acetic acid has a smaller absorbance. At 900 nm, the absorbances increase in the order acetic acid, methanol, water.

It would be possible to examine the spectrum of an unknown sample at either wavelength pair and decide, by comparing the spectral information at only two wavelengths with the training set samples, which of the three materials the new sample was. In fact, the definition of the term "spectrum" could be extended, so that the data at the two wavelengths could be said to constitute the "spectrum" of the sample for the purpose of this identification. This is the chemist's view of the situation.

From the viewpoint of the mathematician, because two wavelengths suffice to allow the desired discrimination, the rest of the spectrum is superfluous. When only the data at those two wavelengths are retained, the absorbances of the three materials constitute a list of paired data. This is a common form of data presentation, and at least one obvious operation can always be done to paired data: they can be plotted as an ordinary graph. Parts C and D of Fig. 1 present such plots, by using the two wavelengths pairs under consideration, one in each spectral region. Examination of these diagrams shows that these absorbance/absorbance plots reflect the same characteristics as were noted in the original spectral presentation. At 1800 nm, water and acetic acid show the same absorbance, i.e., they have the same ordinate value, the ordinate in Fig. 1C representing the absorbance at 1800 nm. Similarly, water and methanol exhibit the same absorbance at 1200 nm (Fig. 1D). In both frames, the projection of the three materials onto the abscissa places them in the order acetic acid, methanol, water; the abcissa represents 1500 nm in Fig. 1C and 900 nm in Fig. 1D. Thus, in this presentation, the materials are described by their locations in the two-dimensional spaces represented by Fig. 1C and D. Each dimension represents the data at a different wavelength. Clearly, other materials could be contained in the same space and could be distinguished from these three, the only requirement being that the absorbance of the new materials must differ from all the materials currently in the training set.

An interesting aspect of this is that a new material need not differ from all the materials at all wavelengths. Indeed, it is easy to show that the spectrum of the new material could match that of each material in the current set at all wavelengths except one, and still be distinguishable from all the materials in the training set. For example, in Fig. 1D, the three materials fall at what might correspond to three corners of a quadrilateral. A new material, having the same absorbance as water at 900 nm, and the same absorbance as acetic acid at 1200 nm, would be placed on the plot at the fourth corner of the hypothetical quadrilateral, and be easily distinguished from all the other materials, although its absorbances at both wavelengths equal the absorbances of a material already in the training set.

There remains the development of a method of deciding whether the data for an unknown material match those for any of the materials in the training set. The criterion for this is that the data from the unknown sample appear in a location that is sufficiently close to those of any of the known materials. In order to make this decision, it is necessary to have some measure of the amount of variability to be expected from the data. In the case of solid samples, the Mahalanobis distance [11] has been used to take into account the variability of natural products. These multidimensional distance measures, which are well-described by Gnanadesikan [10], are computed from the matrix equation:

$$D_{ij}^2 = (\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)'\mathbf{M}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j) \quad (1)$$

where $D_j$ is the Mahalanobis distance from the $i$th sample to the location of the $j$th material in the multidimensional space, $\mathbf{X}_i$ is the vector of absorbance data from the unknown sample, $\mathbf{M}$ is inverse pooled covariance matrix from all the materials in the training set, and $\bar{\mathbf{X}}$ is the location in multidimensional space of the $j$th material. Conceptually, the nature of this equation is to surround the data from each material with an ellipse, the ellipse being matched to the data in such a way that it measures the multidimensional equivalent of a standard deviation. Thus, when distances are measured, they are measured in terms of this ellipse, so that the distance from the centroid of the data to the ellipse in any given direction becomes one "unit distance". These unit distances are then used as the measuring rod for data in that given direction. The distance measures used are called "Mahalanobis distances" in recognition of the pioneering work of Mahalanobis [12].

The basic assumption in this use of Mahalanobis distances is that the data from each material in the training set can be described by ellipses with the same size, shape and orientation. In cases where this assumption is not met, then the ellipse to match the data from each material must be modified. The nature of NIR data is such that a straightforward modification is suitable for the purpose of classification. For NIR data measured on powdered solids, the largest sources of variation are particle size and repack phenomena; these have the effect of causing data at all wavelengths to vary in proportion to each other. Accordingly, a simple normalization procedure is satisfactory; the Mahalanobis distance of the data from an unknown sample to a known material is divided by the *RMS* group size of the multidimensional cloud of data from the known material in the training set [13]. The *RMS* group size used to normalize the Mahalanobis distances is calculated as

$$RMS_j = [\sum_i D_{ij}^2/(n_j - 1)]^{\frac{1}{2}} \quad (2)$$

where $RMS_j$ is the size of the data from the $j$th material, $D_{ij}$ is as defined above, and $n_j$ is the number of training samples of the $j$th material.

For liquids, the situation is simultaneously simplified and more difficult. It is simplified because there are no particle-size effects or other effects that de-

pend to any great extent on the physical properties of the sample. For natural products, then, only changes in the spectrum related to different sample compositions would be observed. For pure materials such as those in the present training set, even these differences are non-existent. This raises the problem that, under these conditions, the ellipse defining the Mahalanobis distance collapses virtually to a point, which results effectively in a "divide-by-nearly-zero" problem for calculating the Mahalanobis distances. To circumvent this problem, Euclidean distances should be used instead. Little work in this area has been done to establish the optimum size of the space surrounding each material, and such situations should be dealt with on a case-by-case basis.

*Multiple regression analysis*

The relationship between discriminant analysis, used for qualitative purposes, and multiple regression analysis, used for quantitative purposes, can be examined via a synthetic sample set. A suitable experimental design consists of a set of samples made in accordance with a three-component mixture diagram, such as that shown in Fig. 2A. The corners represent the pure materials used to make the samples (water, methanol and acetic acid). The edges represent all possible two-component mixtures; the edge between the water corner and the methanol corner, for example, represents all possible mixtures of water and methanol, and the position of a point along that edge indicates the actual composition of a given mixture. Similarly, the other two edges represent mixtures of acetic acid with water and methanol, respectively. Each point in the interior of the diagram represents the composition of a three component mixture. For example, a line joining the midpoints of the water-methanol edge and the water–acetic acid edge represents all mixtures containing 50% water. Thus, the midpoint of that line represents a mixture containing 50% water, 25%

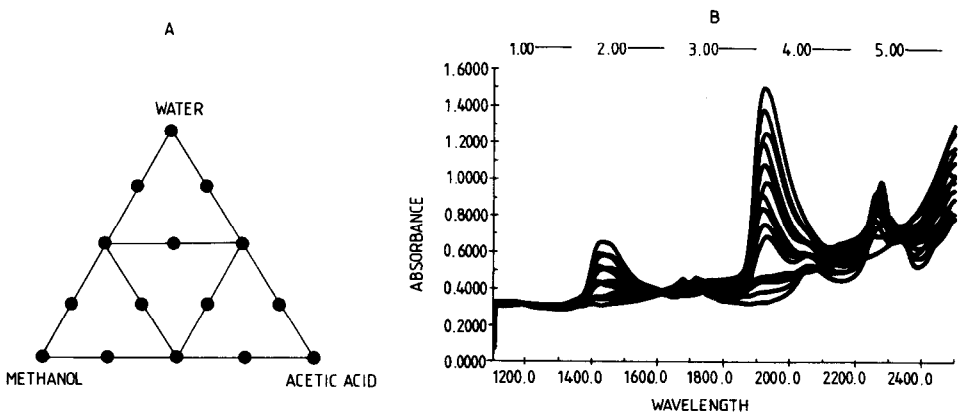


Fig. 2. (A) Mixture diagram for three-component mixtures of water, methanol and acetic acid. (B) Spectra of all the samples comprising the set described by the mixture diagram.

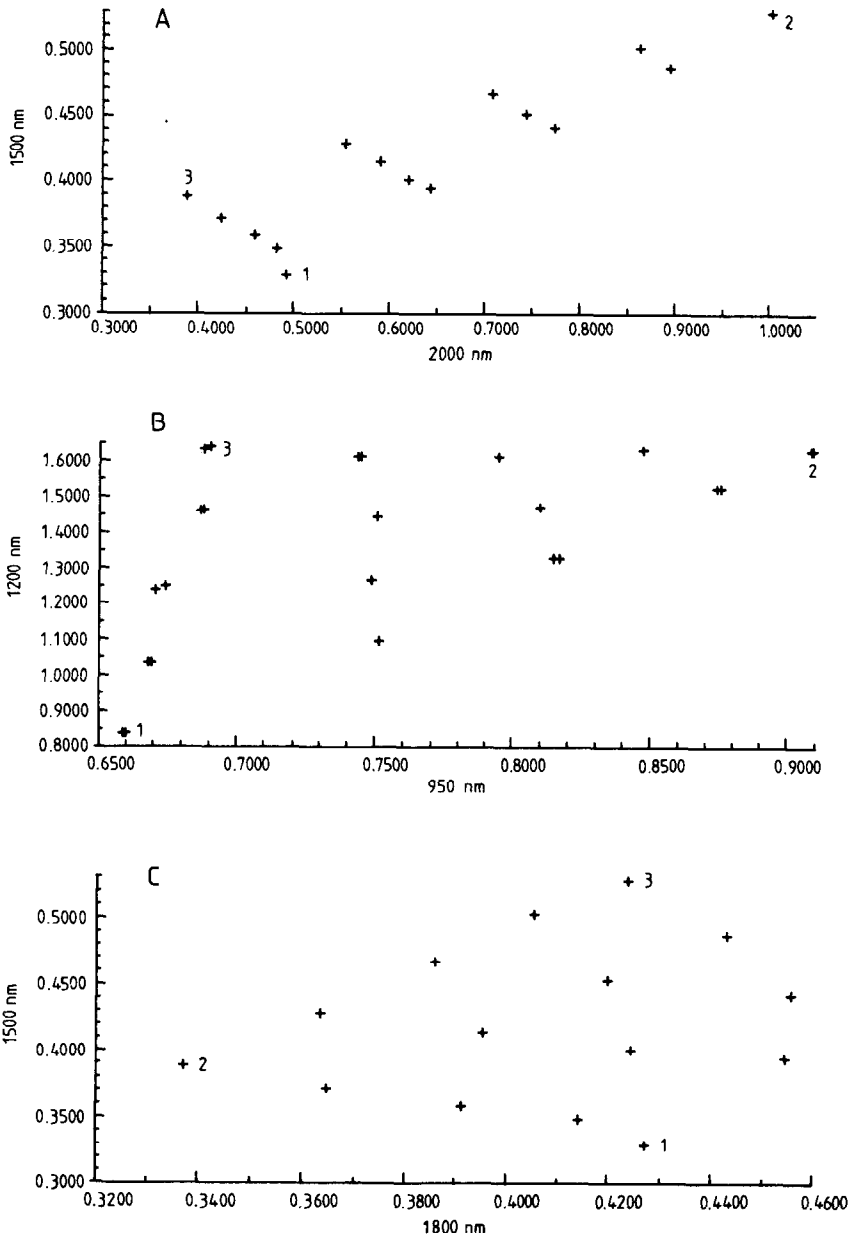


Fig. 3. Absorbance/absorbance plots of the data from the full set of three-component mixtures: (1) Acetic acid; (2) water; (3) acetic acid. (A) Plots in the long-wavelength region with 1500 and 2000 nm; (B) plots in the short-wavelength region with 1200 and 950 nm; (C) plots in the long-wavelength region with 1500 nm and 1800 nm.

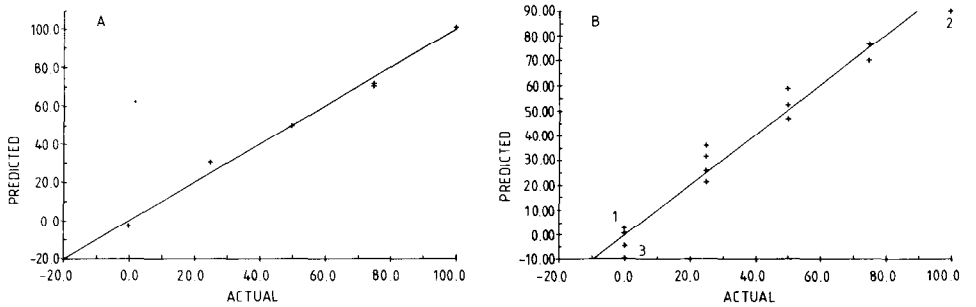


Fig. 4. Calibrations based on only one wavelength. (A) Calibration for methanol in methanol/acetic acid binary mixtures, from data at 1500 nm. (B) Calibration for water in the full set of ternary mixtures; (1) acetic acid; (2) water; (3) methanol.

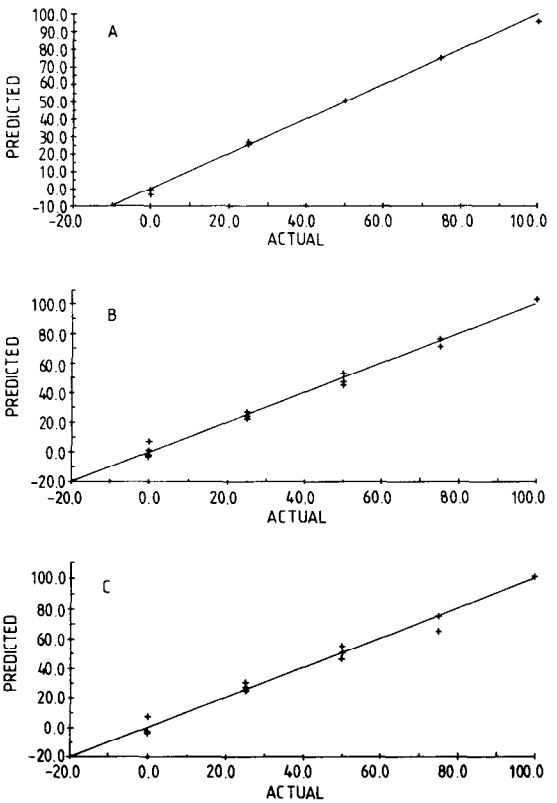


Fig. 5. Calibrations based on two wavelengths, for the full set of ternary mixtures. (A) water; (B) methanol; (C) acetic acid. In each calibration, the same two wavelengths were used (1500 and 2000 nm).

TABLE 1

Calibration characteristics for each component in ternary mixtures based on one independent variable

| Constituent | Wavelength data (1940 nm) | | Principal component | |
|---|---|---|---|---|
| | SEE[a] | $r^b$ | SEE | $r$ |
| Water | 6.11 | 0.9819 | 7.49 | 0.9727 |
| Methanol | 25.42 | 0.6162 | 24.36 | 0.6558 |
| Acetic acid | 30.03 | 0.3656 | 30.61 | 0.3168 |

[a]Standard error of estimate. [b]Correlation coefficient.

TABLE 2

Calibration characteristics for each component in ternary mixtures based on two wavelengths (1500 and 2000 nm)

| Constituent | Coefficients | | SEE | $r$ |
|---|---|---|---|---|
| | 1500 | 2000 | | |
| Water | 218.5 | 108.7 | 1.83 | 0.9984 |
| Methanol | 997.5 | −399.5 | 3.22 | 0.9952 |
| Acetic acid | −1215.9 | 290.8 | 4.15 | 0.9919 |

methanol and 25% acetic acid. The filled circles in Fig. 2A mark the compositions of the mixtures used in the current study.

Figure 2B presents the spectra corresponding to the fifteen mixtures used. Except for small spectral regions, Fig. 2B is very cluttered and very little information could be gained from it directly. A possible exception is the 1400–1500 nm region, where there is some structure to the set of spectra; the spectra can be seen to group together at distinct absorbance levels, with some variation between the spectra at each level. The different levels are due to the fact that in that wavelength region water is by far the strongest absorber, so that all the samples containing the same amount of water have almost the same absorbance. Within each level, the spectral differences are due to the differences between methanol and acetic acid.

An absorbance/absorbance plot of the spectra of the mixtures shows an interesting aspect to the data: if such plots are made at suitable wavelengths, they reproduce the underlying structure of the experimental design by recreating the mixture diagram in the data. Examples are shown in Fig. 3. Because they reflect the experimental design, these plots reveal the relationships between the mixtures and the data obtained from them, and also indicate the requirements for multivariate methods of dealing with such data. As can be

seen in Fig. 3, the spectra of the pure materials are at the corners of the triangles containing the data. These corner points alone are the same as those presented in Fig. 1, that formed the basis of the algorithm for qualitative analysis. Figure 3 shows the relation between those data and the data from the full mixture set; the qualitative data form a subset of the full data. Also, the sets of data points representing the two-component mixtures fall on straight lines, just as they did in the mixture diagram. This fact, however, depends on choosing the proper wavelengths for the display. In Fig. 1C, the long-wavelength pair 1500 nm/1800 nm was used to illustrate the qualitative algorithm, because those wavelengths showed the best separation. Figure 3C displays the full data at those wavelengths; this plot clearly shows how non-linear phenomena can affect the data at different wavelengths. When the data are linear, as in Fig. 3A, it is equally clear that, because two-component mixtures are represented by the data along straight lines between the points representing the pure materials, the composition of a two-component mixture can be determined by the position of its spectrum along that line. Indeed, because there is only one degree of freedom for motion along the line representing a two-component mixture, it is possible to obtain a result by using only one wavelength. An example of this is given in Fig. 4A, which shows the calibration line for methanol in methanol/acetic binary mixtures. Except for a slight non-linearity, this line is an eminently suitable calibration; it represents data corresponding to one edge of the mixture diagram. Not shown are the calibration lines corresponding to the other two edges (the water/methanol and water/acetic acid edges) are separated from the full data set, equally good calibration lines are obtained even when the same wavelength is used.

When the data are non-linear, as in Fig. 3C, a single wavelength would not be satisfactory. For example, an attempt to use data at 1800 nm to analyze mixtures of methanol and acetic acid would fail; a reading of 0.44 absorbance at 1800 nm could correspond to a mixture containing either 80% methanol or 10% methanol. In this case, it would be necessary also to use the absorbance at 1500 nm, to decide between the two possibilities. The presence of non-linearities adds the equivalent of a degree of freedom to the data, thus requiring the inclusion of an additional variable to the model in order to achieve accurate results. In some cases, the non-linearity can be separated from the data and its unique contribution determined [14]. From Fig. 3C, one might argue that data at 1500 nm would suffice for the analysis of the binary mixtures. However, it is easy to imagine data that are double-valued along both wavelength axes, so that neither value alone would suffice.

To analyze ternary mixtures, the operation corresponding to locating a binary mixture along the line representing all possible binary mixtures is locating the point corresponding to the ternary mixture within one of the triangles described by the data in Fig. 3. Indeed, if a non-linear situation such as that shown in Fig. 3C is encountered, the non-linearity need not prevent accurate

analysis of the sample, as long as the space is well mapped. This would correspond to using discriminant analysis algorithms for quantitative purposes. The disadvantage of this use of discriminant analysis is that each point within the triangle must be defined separately; this algorithm does not use external chemical knowledge of the behavior of absorbance of the mixture when the composition changes.

Regardless of the algorithms used, more than one wavelength is also needed when mixtures containing all three components are present in the training set. Figure 4 and Table 1 present the results of attempting to use only one wavelength to model the data. The results are predictably poor, because three-component mixtures actually contain two degrees of freedom. Even when the wavelength containing the most information is used, this remains true. The NIR wavelength that has by far the strongest absorbance is the 1940-nm band of water. An attempt to calibrate the set of mixtures for water, based only on that wavelength, is shown in Fig. 4B. The structure of the experimental design can be seen in this plot; the effect of the uncompensated variation arising from the degree of freedom represented by the varying concentrations of methanol and acetic acid is clear in the residuals.

Figure 5 presents the calibrations obtained from water, methanol and acetic acid, respectively, based on data at two wavelengths (1500 and 2000 nm). The absorbance data at those wavelengths are presented in Fig. 3A. It is clear in Fig. 5 that some residual non-linearity is not accounted for by the model and produces curvature of the calibration line. The two degrees of freedom accounted for by the two wavelengths minimize the effects of composition variation on the data. As was noted above, non-linearity represents another degree of freedom and would require another wavelength to model. The spread of the data around the calibration lines is also due to uncompensated non-linearity. Figure 5B best illustrates this effect; close inspection shows that the topmost point of the three at ACTUAL = 0 is itself composed of two almost overlapping points. These two points correspond to the pure water sample and the pure acetic acid sample; one is effectively looking along the edge of a surface that has the shape of a shallow dish.

If the residual non-linearities are ignored, and only first-order linear relationships apparent in Fig. 3B are considered, then the data at these two wavelengths are clearly sufficient to describe all points within the triangle representing the data space. There remains only the development of suitable relationships (calibration equations) to relate any position in the triangle to the composition. The relations for the data from the ternary mixtures of Fig. 3B are presented in Table 2. It is interesting to note that at each wavelength, the coefficients for the three components sum to zero, indicating the dependence of the calibrations created by the compositions.

In developing practical calibrations, one of the important considerations is the choice of wavelengths. Spectral information can be helpful but normally

does not provide the complete answer because the spectrum shows only which wavelengths correspond to absorbance bands of known constituents in the specimens of interest, and so account only for those degrees of freedom attributable to composition variations. Information which is equally important to accurate calibrations includes the degrees of freedom related to non-linearity, or to physical phenomena such as repack effects [15]. For this reason, it is necessary to select at least some of the wavelengths for the calibration via automatic computerized search algorithms. Recent studies have demonstrated that all such algorithms have the inherent characteristic that the selected wavelengths are subject to random variation because of the electronic noise superimposed on the data [16].

*Principal component analysis*

Figures 6–8 present the spectra of all the binary mixtures, and their first principal components. In each case, it can be seen that the principal component is the difference between the spectra of the two materials in the mixture.

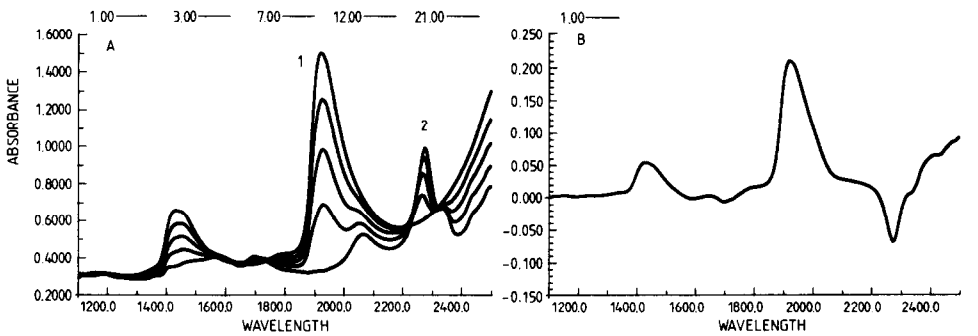


Fig. 6. (A) Spectra of the water/methanol binary mixtures. (B) First principal component of the water/methanol binary mixtures. (1) Water; (2) methanol.

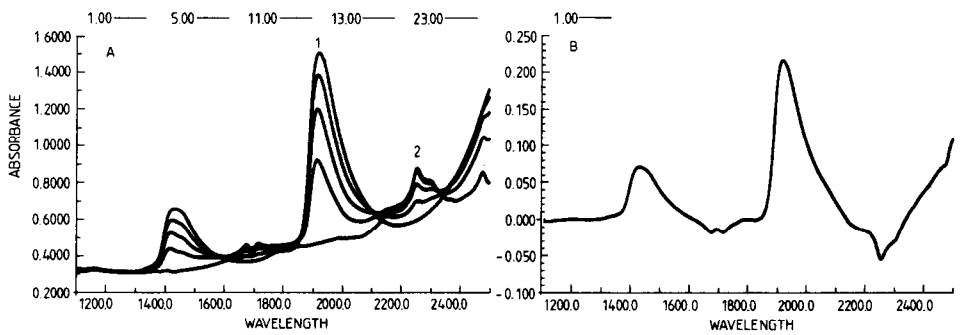


Fig. 7. (A) Spectra of the water/acetic acid binary mixtures. (B) First principal component of the water/acetic acid binary mixtures. (1) Water; (2) acetic acid.

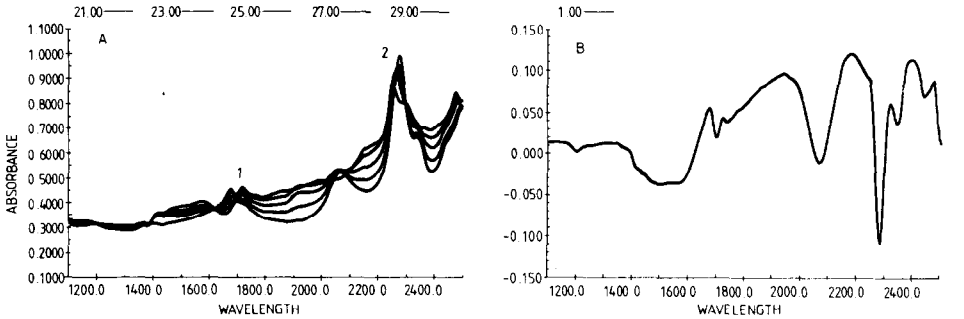


Fig. 8. (A) Spectra of the methanol/acetic acid binary mixtures. (B) First principal component of the methanol/acetic acid binary mixtures. (1) Acetic acid; (2) methanol.

TABLE 3

Comparison between calibration with a single wavelength and a single principal component (PC) for binary mixtures

| Mixture | Constituent | Standard error of estimate | |
|---|---|---|---|
| | | 1500 nm | 1 PC |
| Water/methanol | Water | 3.15 | 1.38 |
| Water/acetic acid | Acetic acid | 3.66 | 6.80 |
| Methanol/acetic acid | Methanol | 3.59 | 5.58 |

For example, the first principal component of the water/acetic acid mixtures (Fig. 7B) shows maxima around 1445 nm and 1940 nm, corresponding to the absorbance bands of water, and minima around 1700 nm (a doublet) and 2250 nm, corresponding to the absorbance bands of acetic acid. The first principal component of an NIR data set is often attributed to repack or particle size effects, which are usually the major source of variation of the data for solids. As these data show, however, the first principal component of NIR data is not inherently due to such phenomena as is sometimes erroneously claimed.

A calibration for any of the constituents based on one principal component is essentially equivalent to a calibration for that constituent with one wavelength, as can be seen from Table 3 where one-variable calibrations for one constituent in a binary mixture are compared on the basis of their standard errors. The main source of error in all these cases is non-linearity; clearly the benefit of using principal components versus wavelength data depends on the nature of the data.

Some of the principal components from the ternary mixtures are presented in Fig. 9. Interpretation of these principal components is not as clear as those for the binary mixtures, although some information can be obtained about the

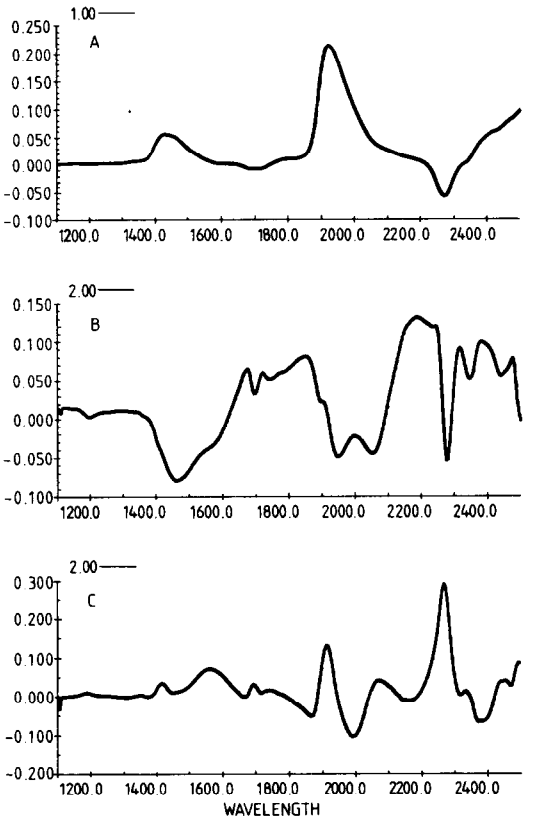


Fig. 9. Some principal components of the full set of ternary mixtures: (A) first principal component; (B) second principal component; (C) third principal component.

bands seen. Because water has by far the strongest absorbance bands, the first principal component consists of positive-going water bands (at 1940 nm and 1445 nm) with the combined bands of methanol and acetic acid going negative. Most of the remaining variance is due to the methanol/acetic acid differences, and this shows up in the second principal component, which is very similar (although not identical) to Fig. 8B, the principal component of methanol/acetic acid binary mixtures.

The result of plotting the first two principal component scores against each other, as is usually recommended, is shown in Fig. 10. Again, the structure of the experimental design is clear in the data. Thus, as in the case of the absorbance/absorbance plots shown (Fig. 3), the space occupied by the principal component scores can be used for qualitative or quantitative analysis, by noting where the data from an unknown sample lie, in exact parallelism to the methods applied to absorbance data. If only the corners of the somewhat cur-

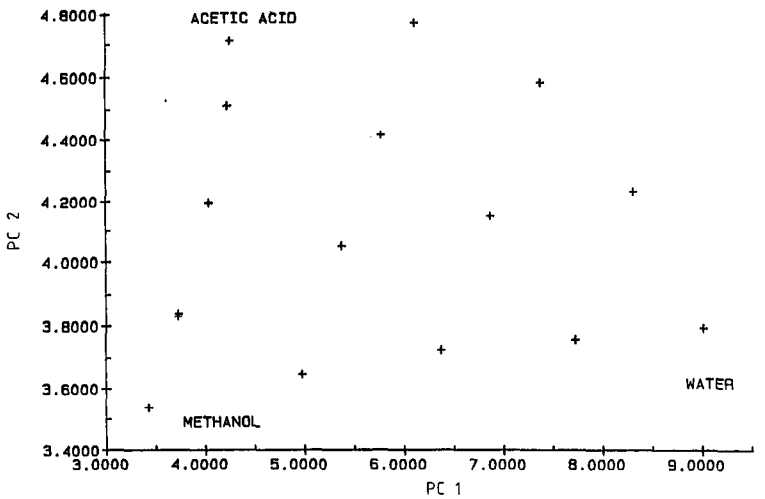


Fig. 10. Plot of the first two principal component scores from the full set of ternary mixtures, showing the non-linearity.

TABLE 4

Calibration statistics for each component of the ternary mixtures based on two principal components

| Constituent | Coefficients | | SEE | $r$ |
|---|---|---|---|---|
| | PC1 | PC2 | | |
| Water | 17.89 | −17.56 | 2.37 | 0.9973 |
| Methanol | −12.06 | −56.44 | 8.53 | 0.9657 |
| Acetic acid | −5.83 | 74.00 | 6.19 | 0.9821 |

vilinear triangle shown in Fig. 10 are utilized, then only qualitative analysis is achieved, by distinguishing the pure materials from each other. This is closely related to the SIMCA algorithms (see p. 242 [1]). The entire interior of that triangle is utilized for quantitative analysis of the ternary mixtures.

The principal components can also be used to obtain calibrations. When calibration computations are done for each of the constituents, the first principal component is found to be the most important predictor in each case, but as with the calibrations based on wavelength data, it is necessary to use the first two principal components to achieve accurate results. The results of using the first principal component for the calibration of each constituent of ternary mixtures are presented in Table 1 along with the corresponding results for calibrations based on data from a single wavelength. The results are poor, but are virtually identical between the two types of calibration. Similarly, compar-

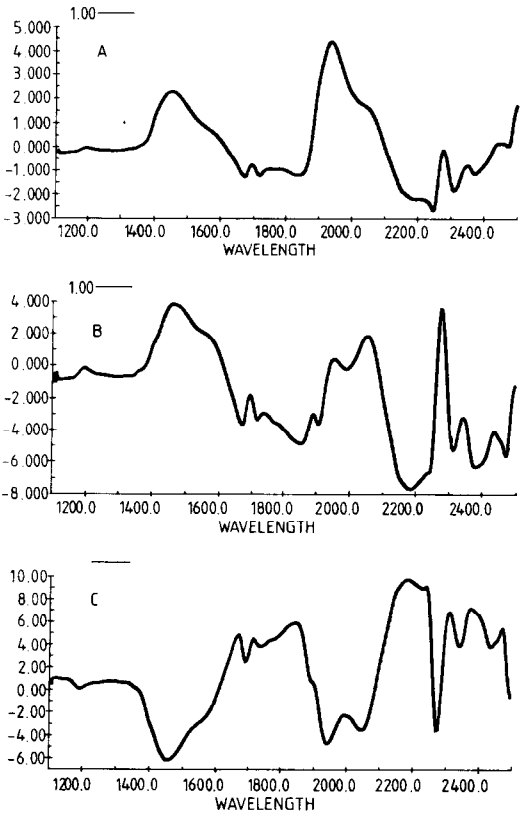


Fig. 11. "Spectra" of the principal component equations for the three constituents: (A) water; (B) methanol; (C) acetic acid.

ing the results in Table 4 (for which the first two principal components were used) with those in Table 2 shows that the two approaches are again effectively equivalent. If anything, the data at the individual wavelengths used appear to compensate slightly better for the non-linearity.

One aspect of the use of principal components that has been largely ignored is the way in which calibration results can be applied to gaining understanding of the data, in addition to their quantitative application. The route to this is the conversion of the principal component calibration to a set of coefficients of the original spectral data. This can be done via the following relationship:

$$k_i = \Sigma b_j P_{ij},$$

where $k_i$ is the coefficient for the $i$th wavelength, $b_j$ is the calibration coefficient for the scores from the $j$th principal component and $P_{ij}$ is the principal component loading for the $j$th principal component at the $i$th wavelength [7].

When coefficients are calculated from this equation for all the wavelengths

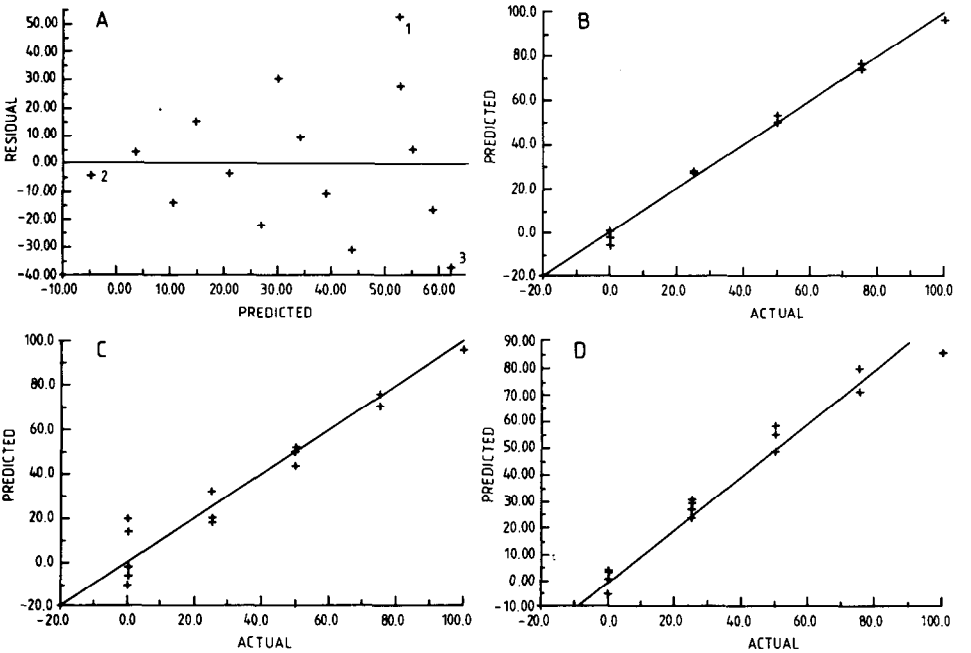


Fig. 12. Plots obtained from principal component calibrations. (A) Residuals of methanol calibration based on one principal component. (B-D) Principal component calibrations based on the first two principal components; (B) for water; (C) for methanol; (D) for acetic acid.

in the original data set, the set of coefficients themselves forms a "spectrum" (if a "spectrum" is defined as any set of numbers that are associated one-to-one with a corresponding set of wavelengths). For the data in the current set, the spectra corresponding to the calibrations obtained by using the first two principal components for the three constituents are shown in Fig. 11. Comparison of these spectra to the spectra of the constituents gives insight into which wavelengths are useful for predicting the constituent of interest in each case, and which are correcting for the absorbance of interfering compounds. Also, as in the case of calibration coefficients obtained directly from the individual wavelength data, the sum of the coefficients at each wavelength for the spectra in Fig. 11 is zero, i.e., those three spectra sum to zero.

The plots of the principal component calibrations themselves contain useful information. It is often noted that when principal components are calculated, there is some ambiguity, requiring that the eigenvector must be normalized to unit vector length. It is not usually appreciated, however, that even after this has been done, there is still an ambiguity of $\pm 1$, leaving the resulting principal component in one of two orientations, depending on how the computation happens to converge. When the principal component calibration is calculated, this

remaining ambiguity disappears. Bands in the plotted principal component calibration that are positive correspond to those wavelengths that aid in the prediction of the constituent, while those bands that are negative correspond to correction for the various interferences and physical phenomena.

Associated with the poor performance of the calibration based on one principal component are residual plots such as that shown in Fig. 12A. Here again, the structure of the mixture design is clear in the residuals; residuals from the calibrations for the other constituents showed the same effect.

The calibration plots obtained from the use of the first two principal components are shown in Fig. 12B–D. As in the case of the calibrations based on wavelength data, the errors of the calibration are due mainly to non-linearity. The plotted points in Fig. 12B–D also lie on a cup-shape surface. In the case of principal components, this error appears much worse than when the calibrations were obtained by using data from individual wavelengths, particularly for methanol and acetic acid. This can be seen by comparing the calibrations in Fig. 12 to those in Fig. 5. Given that the wavelengths used for the calibrations shown in Fig. 5 were not optimized for linearity, it is clear that the principal component approach is less accurate when such effects are present in the data. Of course, the non-linearities could be corrected by including a principal component representing the non-linear terms, but three principal components would then be needed.

*Conclusions*

As has been shown, the same data can be analyzed by several nominally different algorithms. To a large extent, any of the three algorithms considered can be used to achieve any of the specified analyses. Discriminant analysis can be used for either qualitative or quantitative analysis of the mixtures in the dataset, simply by focussing attention on part or all of the space occupied by the data, respectively. Multiple (linear) regression (MLR) can be used for quantitative analysis, as it is normally used, and for qualitative analysis by noting which material is present at 100% concentration. Principal component analysis (PCA), of course, is well-known for both its qualitative and quantitative aspects.

Proponents of various sophisticated multivariate algorithms such as PCA make claims such as "PCA is better than MLR". Detractors make counterclaims such as "PCA is not as good as PLS". The current work shows that neither of these claims is correct: each algorithm has its own set of characteristics that make it different than the others, but inherently neither better nor worse. As has been seen, any algorithm can perform any function, yet there is a certain artificiality to this, clearly certain algorithms are better suited for certain types of analysis than others. In comparing different algorithms, it is necessary to consider both the strengths and weaknesses of each.

For example, principal component analysis, when used for quantitative pur-

poses, has the advantages of robustness and orthogonality, avoiding the need to execute complicated and time-consuming variable selection routines; it also allows all wavelengths in a calibration data set to be used without overfitting the data. But PCA has the disadvantage of being sensitive to non-linearities that a proper choice of wavelengths can avoid. This sensitivity can be seen in both Fig. 10 and Fig. 12. Because of this sensitivity to non-linearity, as well as the fact that the limiting error of a calibration for most substances of practical interest is usually due to the error in the dependent variable (as regression theory says should be the case [17], there is little justification for the widespread belief that quantitative calibrations based on principal components, partial least squares, or other sophisticated multivariate algorithms are inherently more accurate than calibrations based on individual wavelength data, although of course they may be in particular cases.

It behooves the chemometric community to investigate each algorithm on its own merits and to establish which types of problems each one is best suited for, because none of them is optimal for all possible situations.

REFERENCES


1 M.A. Sharaf, D.L. Illman and B.R. Kowalski, Chemometrics, Wiley, New York, 1986, pp. v, 228, 242.
2 D.R. Massie and K.H. Norris, Trans. ASAE, 8 (1965) 598.
3 F.E. Barton, USDA Russell Research Center, Athens, GA, personal communication, 1987.
4 C. Maddix, Bran & Luebbe, Elsmford, NY, personal communication, 1988.
5 I.A. Cowe and J.W. NcNichol. Appl. Spectrosc., 39 (1985) 257.
6 H. Mark, Anal. Chem., 58 (1986) 2814.
7 H. Mark, Chim. Oggi, Sept. (1987) 57.
8 M. Martens and H. Martens, Appl. Spectrosc., 40 (1986) 303.
9 W.W. Cooley and P.R. Lohnes, Multivariate Data Analysis, Wiley, New York, 1971, p. 243.
10 R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York, 1977, pp. 90–103.
11 H. Mark and D. Tunnell, Anal. Chem., 57 (1985) 1449.
12 P.C. Mahalanobis, Proc. Nat. Inst. Sci. India, 2 (1936) 49.
13 H. Mark, Anal. Chem., 58 (1986) 379.
14 H. Mark, Appl. Spectrosc., 42 (1988) 832.
15 H. Mark, Anal. Chem., 58 (1986) 1454.
16 H. Mark, Appl.. Spectrosc., 42 (1988) 1427.
17 N. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 1981, p. 29.