Principles of Multivariate Analysis with Emphasis on Factor Analysis

David G. Watson

Physical Electronics, Inc., 6509 Flying Cloud Dr., Eden Prairie, MN USA 55344

(Received September 30 1998; accepted January 22 1998)

Multivariate analysis (MVA) has slowly gained wide acceptance in the field of surface analysis since the first examples were published some 20 years ago. In the past few years, the greater availability and lower cost of scientific computing have resulted in a steady increase in multivariate applications. Most of the successful multivariate methods in use today – Linear Least Squares (LLS), Target Factor Analysis (TFA), and Partial Least Squares (PLS) – are quite closely related. This talk will focus on these multivariate data analysis methods which are often referred to under the broader heading of *chemometrics*. In particular, factor analysis will be discussed in detail with emphasis on the mechanics of the technique and on the practical implications.

1. Introduction

As a data reduction approach, multivariate analysis[1] has slowly gained acceptance in surface analysis during the last two decades and many analysts now employ these techniques on a routine basis. Most commercial software produces analytical results which are, unavoidably, based on information – *e.g.* examples of pure component spectra – provided by the user for the samples or system under consideration. Multivariate manipulations of the data are also more complex than those traditionally used for intensity measurement in surface analysis. It is very easy to misinterpret MVA results if the user is unaware of the underlying assumptions and intricacies of the techniques. It is, therefore, of great benefit to the user to understand the workings of MVA and this can be done without having to fully understand the complexities of the mathematics.

This aim of this article is to present the fundamental concepts behind MVA without relying too heavily on notation and mathematical proficiency. It should be made clear that there are many more thorough treatments of MVA available (*e.g.*) [1] and that this article is meant to describe the workings of MVA to surface analysts in relation to surface analysis. The assumptions, data processing steps, and danger areas in MVA are presented in a way that the author hopes is easy to understand, thereby, increasing the confidence in the results of this type of data analysis.

2. Objectives and Assumptions

The basic assumption underlying the three multivariate techniques discussed here is *linearity*. It is assumed that the data (spectra) under analysis can be described by *linear combinations* of the spectra of the chemical components in the sample, as can be demonstrated by the construction of the simulated "measured data" shown in Fig. 1. The spectra in the lower portion of Fig. 1 represent those that might be recorded during a

[1] The term *chemometrics* – a discipline that includes MVA – is often used in place of the term MVA.

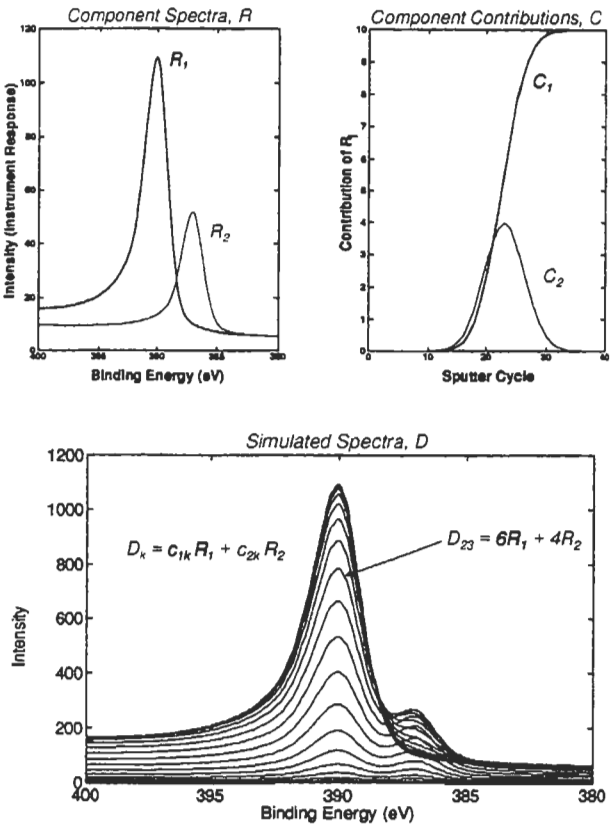

Figure 1 The simulated pure component spectra (top left) and their 'contributions' (top right) to each linear combination spectrum (bottom) in the simulated "measured dataset".

sputter depth profile of a sample composed of the pure chemical species R_1 and R_2 in the amounts C_1 and C_2 as specified by the 'contribution' profile in Fig. 1. Each spectrum, D_k, has a composition dictated by the C_{jk} and the construction of the entire dataset can be expressed in matrix notation,

$$\mathbf{D} = \mathbf{RC} \qquad (1)$$

where, **D** contains the spectra recorded during the experiment, **R** contains the pure component spectra, and **C** contains the contribution, C_{jk}, of each R_j to each spectrum, D_k, in the data matrix. Note the assumption of linearity is explicitly stated in Eq. 1. Given this model of the dataset, we can now specify the objectives of the data analysis:

1. Determine the number of chemical species or *factors* – *i.e.* find the number of spectral components that vary in contribution independently of one another;

2. Determine the identity of each component by finding its spectrum, R_j

3. Determine the contribution, C_{jk} of each spectral component to each recorded spectrum; this is equivalent to calculating the *profile* for each component by solving Eq. 1.

Note that the word 'contribution' and not 'concentration' is used here; any calculated intensities are subject to the same matrix effects and required sensitivity factor corrections to obtain the final chemical concentrations.

One final assumption is that we have collected enough data to determine the sample composition; this will be addressed later.

3. Named Multivariate Techniques

Three multivariate techniques are discussed here. Each technique has its advantages and disadvantages and, therefore, each has a place in the analyst's toolbox.

1. *Classical Least Squares (CLS)*, also called *Linear Least Squares (LLS)*, allows the analyst to quickly extract component intensities from complex spectra with S/N and dynamic range that are usually improved over traditional peak height or area calculations [1]. It also requires that we specify both the number of components and their spectra completely before any calculation of contributions can be made;

2. *Target Factor Analysis (TFA)*, essentially a more rigorous version of CLS, allows the

analyst to determine with confidence the number of components responsible for the spectral variation in the dataset and to test each proposed component spectrum individually [2]. It is, therefore, both more rigorous and more flexible than CLS. TFA also allows greater improvements in S/N and dynamic range because of the noise rejection inherent in the first steps of the procedure (*vide infra*). As with CLS, however, TFA requires a complete solution before any component contributions can be determined;

3. *Partial Least Squares (PLS)*, a calibration/prediction scheme, is the most powerful MVA technique for data modeling in that it allows the analyst to construct a calibrated model for a particular component hidden in a multivariate dataset without providing the complete solution [3]. Also, due to the calibration step, it can produce concentrations as the analytical result. In contrast with CLS and TFA, however, PLS requires standard samples for the calibration step.

This paper will address the first two techniques, CLS and TFA, in detail since they are by far the more commonly used in surface analysis at present; PLS has been used mainly in TOF-SIMS analyses as briefly discussed in a later section.

The benefits of using MVA instead of traditional surface analytical methods of spectral intensity calculation can be seen in Fig. 2 where the Auger depth profile of a Au/Ta/SiC sample is shown as calculated by traditional (derivative peak-to-peak) method and by CLS.

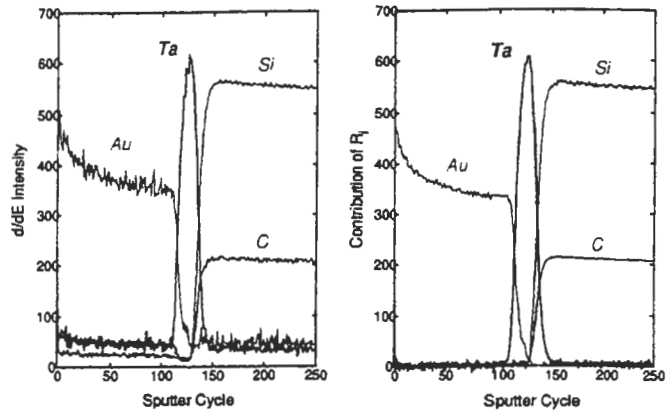

Figure 2. Au/Ta/SiC sample depth profiles by traditional derivative p-p method (left) and CLS (right).

In Fig. 2, the p-p profiles of the three sample components are lower in dynamic range because of S/N and spectral overlap whereas, the application of a simple CLS fit of pure spectra to the data produces the profiles on the right in Fig. 2.

4. Classical Least Squares (CLS)

CLS is the simplest of the multivariate techniques and is the most productive in terms of results versus effort required. The contributions to the recorded spectra of the various sample components are given by the following:

If $\mathbf{D} = \mathbf{RC}$ (Eq. 1) *then* (given $\mathbf{R}$)

$$\mathbf{C} = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{D} \quad (2)$$

by inversion where the first part of the right hand side of Eq. 2 – $(\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T$ – is called the pseudoinverse of $\mathbf{R}$ when $\mathbf{R}$ is not square (almost always). Eq. 2 is the *prediction* step of CLS and can be used to estimate the intensity of each spectral component, R_j, in each measured spectrum, D_k. A graphical depiction of the procedure given by Eq. 2 is shown in Fig. 3 where the Auger *basis* spectrum, R_1 of Ta to D_{146}, a suspected spectrum of Ta; the fit

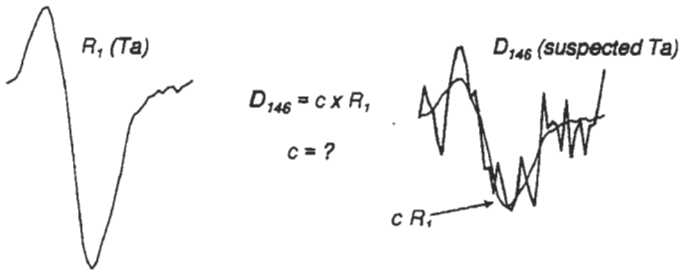

Figure 3. The CLS model.

coefficient, c, is a scaling factor for R_1 that fits it to D_{146} in a least squares sense. Fig. 4 shows the extension of this method to the Ta spectral region of sputter depth profile data from the Au/Ta/SiC sample mentioned above. In this case, the sample consists of relatively pure layers of materials, therefore, the *basis* or *pure component* spectrum is taken from the dataset itself at Sputter Cycle 128 (in TFA R_1 would be called a 'typical' factor [2]).

The differences between the results of the two methods of intensity calculation are obvious in Fig. 4: first, the p-p method suggests a background level, which is due solely to measurement of p-p noise, of Ta of 42±6 in both the Au and SiC layers – CLS shows Ta at 0±4 in the same regions; second, the CLS profile is overall smoother due to the use of the full spectrum instead of two data channels.

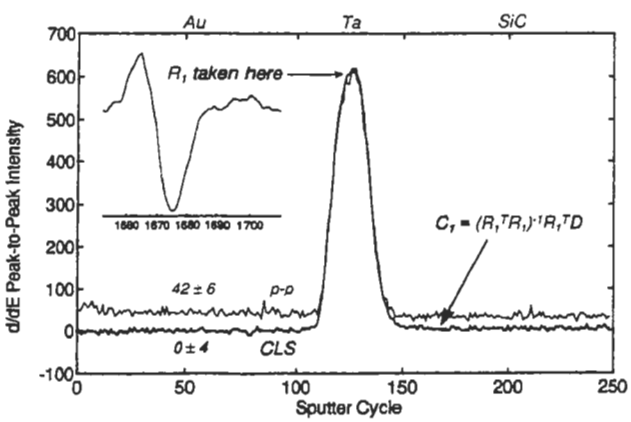

Figure 4. The Auger depth profile of Ta calculated both by p-p and CLS methods. The inserted spectrum, R_1 (D_{128}) was used as the *basis spectrum* for the least squares fit calculated by Eq. 2.

Common Sources of Error

The following are the most common sources of error in CLS (and TFA) analyses:

1. Non-linearity in the data including charge induced spectral shifts, detector saturation (lack of a dead time correction), and sample decomposition;

2. Co-linearity in the component spectra, meaning that one *basis spectrum* is contaminated with another used in the fit – this leads to instability in the calculation and can result in large, negative values in the calculated contributions.

Summary and Evaluation of CLS

CLS is shown by example to reduce spurious backgrounds, extract chemically significant results from the data, lead to a better understanding of the sample, and all this is done with a very simple implementation and application of the technique.

On the other hand, we have guessed at the number of components, have little confirmation that the *basis spectra* are the correct ones, and we have retained more noise in the data than is necessary. These disadvantages are addressed by TFA.

5. Target Factor Analysis[2] (TFA)

TFA is a more sophisticated version of CLS and it begins by considering the same problem with the same assumptions. TFA is performed in two steps: *Principal Component Analysis* (PCA) produces a mathematical or *abstract* solution to Eq. 1 that allows the analyst to

[2] For a complete mathematical treatment of TFA applied to the physical sciences, see Ref. 2.

determine the number of independent *components* or *factors* underlying the measured dataset and; *Target Transformation* (TT) rotates the PCA solution expressed in the abstract space to a solution in a physically and chemically meaningful space. At first glance it appears that the abstract PCA solution, **D** = **RC**, is simply arrived at by magic, however, this solution too comes from the rotation of another: the trivial solution,

$$\mathbf{D} = \mathbf{I}\,\mathbf{D} \qquad (3)$$

where, **I** is the identity matrix and contains the basis vectors that define a co-ordinate space and the elements of **D** (the spectral data points) are co-ordinates on the axes defined by **I**. In the abstract solution, $\mathbf{I} \rightarrow \underline{\mathbf{R}}$ and $\mathbf{D} \rightarrow \underline{\mathbf{C}}$ where $\underline{\mathbf{R}}$ is a new set of basis vectors that define the same data space with a new, rotated co-ordinate system and $\underline{\mathbf{C}}$ contains the co-ordinates of the data points on the new axes in $\underline{\mathbf{R}}$. The axes are rotated by a technique called *eigenvector* rotation [4], which rotates the co-ordinate system of the data space into a very convenient orientation, as will be seen.

Principal Components Analysis (PCA)

To understand the rest of this description we must now think of spectra in an unconventional way. A spectrum, like those displayed in Fig. 5, can be thought of as a highly compressed representation of a point in n-dimensional space where, n = 200 in Fig. 5. For purposes of illustration, the spectra in Fig. 5 are reduced to two data points and plotted against the two energy axes, e_1 and e_2 in Fig. 6. This plot shows that each measured spectrum is really a point, or the end of a vector, in the data space, as are the basis vectors, I_j,, also shown in Fig. 6.

In Fig. 7, the same plot is made for a series of spectra from a single component system. The points in this case fall on a straight line, as might be expected since each spectrum is the same shape (*i.e.* each data vector has the same

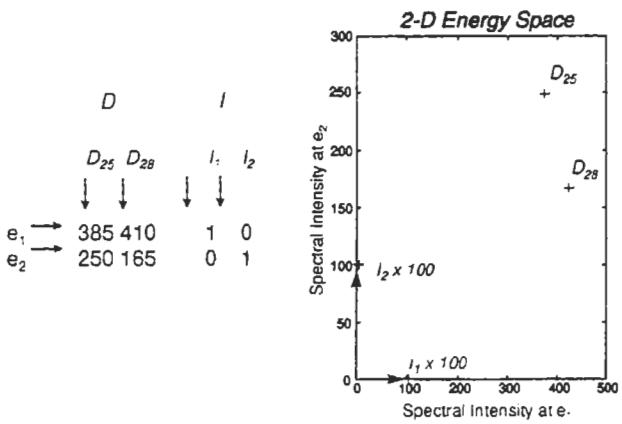

Figure 6. Two spectra plotted against two of 200 showing the spectra to be points in the data space.

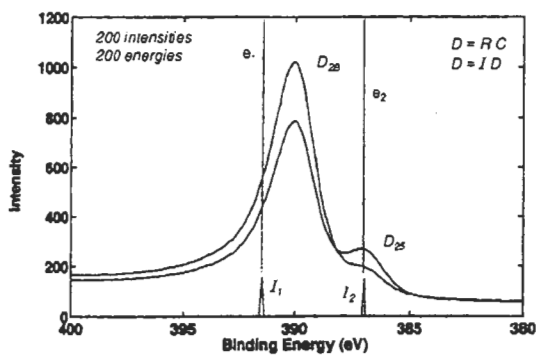

Figure 5. Synthetic spectra at 200 energies. Two of the energy axis (e_1 and e_2) *basis vectors*, **I**, of the data space are also plotted as 'spectra'.

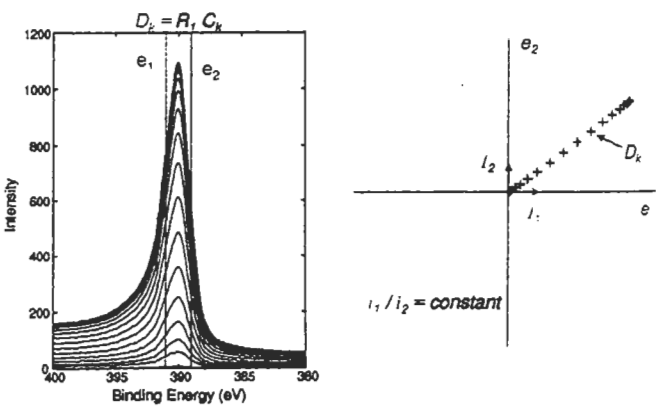

Figure 7 Plot of single component data set in a 2-D data space; the ratio of intensities, i_1/i_2, is constant.

direction) as all the others, therefore, the only differences are in the relative magnitudes of the vectors (*i.e.* intensities of the spectra).

This model of the data space can now be used to determine the number of components for any dataset. To do so, the axes of the space are rotated so the first axis lies in a direction that maximizes the projections of the data vectors on it. Second and subsequent axes are found by rotation about the first, maintaining orthogonality at each step. Fig. 8 shows the first two such eigenvector[3] axes in relation to the data points and the original nominal axes. Once the preferentially oriented eigenvector axes, $\underline{\mathbf{R}}$, are calculated, it is desirable to use

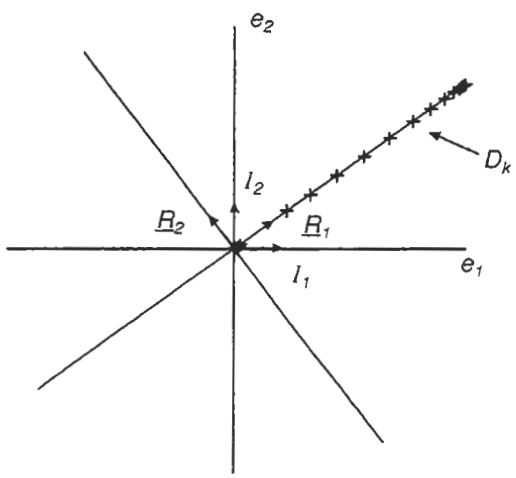

Figure 8. The data space showing the newly calculated eigenvector axes and their orientation with respect to the spectra.

them to define the data space, therefore, we need to calculate the co-ordinates, $\underline{\mathbf{C}}$, of the spectra against these new axes. To do this, a simple inversion of Eq. 1 can be used because $\underline{\mathbf{R}}$ is, at the moment, square, and orthonormal ($\underline{\mathbf{R}}\underline{\mathbf{R}}^{-1} = \underline{\mathbf{R}}\underline{\mathbf{R}}^{T} = \mathbf{I}$),

$$\underline{\mathbf{C}} = \underline{\mathbf{R}}^{T}\,\mathbf{D} \qquad (4)$$

Note that in Fig. 8, our single component system requires only a single eigenvector axis to locate all the data spectra in the data space. This fact would be reflected in the eigenvalues associated with the eigenvector axes; the eigenvalues indicate the total projections of the data spectra on each eigenvector axis.

In Fig. 9, a two component dataset is displayed in a similar manner along with the two most significant eigenvector defined axes. In this case, both eigenvectors have significant rôle in defining the space occupied by the data.

At this point we suspect that the number of eigenvector axes required to account for the true dimension of the data space corresponds to the number of components or *factors* underlying the dataset. It is now useful to look at the eigenvectors, or *eigenspectra*,

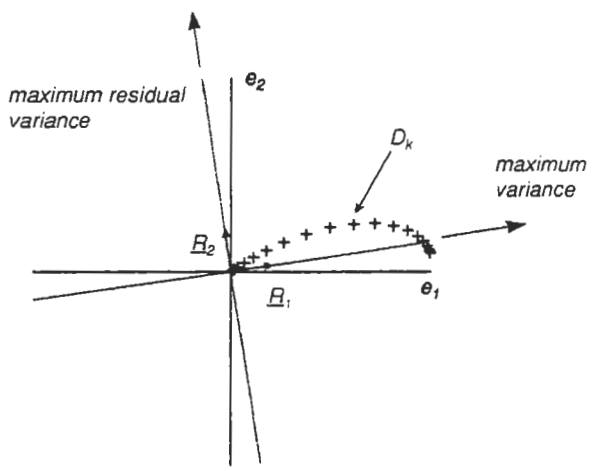

Figure 9. First two eigenvector axes calculated for a two component dataset. The data vectors now line in a plane and show projections against both eigenaxes.

since they are linear combinations of the spectra, in $\underline{\mathbf{R}}$ and the corresponding *eigencontributions* in $\underline{\mathbf{C}}$ as shown in Fig. 10. The first three eigenvectors are plotted in the left hand pane of Fig. 10 and the contributions – the co-ordinates of the spectra on the eigenvectors are plotted in the right hand pane.

[3] 'Eigen' is a German word meaning 'own' or 'peculiar to', therefore, the eigenvector axes can be thought of as those axes that are peculiar to or that best represent the data set.

The two most significant eigenvector axes, $\underline{R}_1$ and $\underline{R}_2$, show spectral intensity while $\underline{R}_3$, which is out of plane, does not. The contributions $\underline{C}_1$ and $\underline{C}_2$ show that all data intensity is accounted for by two eigenvector axes – the third is due to significance error in the double precision calculations, as indicated by the scale factor on the plot. Since the data lie in a plane, *the factor space* is 2-dimensional – there are two factors or components in the system under study.

The above discussion also indicates the need for at least $n + 1$ data points in each spectrum, where n is the number of significant eigenvectors.

We can now see that the MVA is redefining the data co-ordinate system: single-energy axes become multiple-energy or full spectrum axes. In the example above a data space previously described by 200 single-energy axes is now described by 2 multivariate eigenvector axes; in the process we have determined that there are two independently varying (in contribution as a function of depth) components in the system.

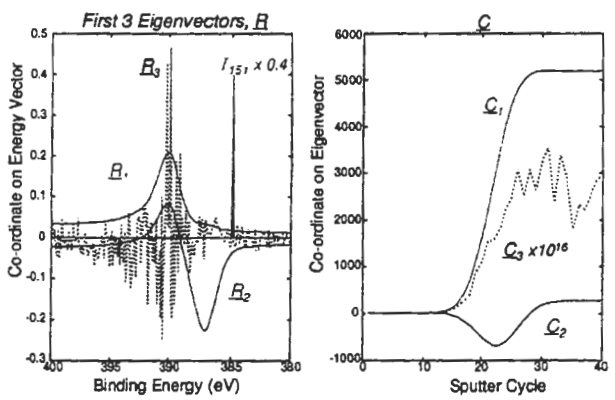

Figure 10. The eigenvector axes (left) and the corresponding contributions or co-ordinates of each data spectrum on each axis (right). A nominal energy space basis vector is also shown. The contribution vector, C3, is due to significance error in the calculations.

The Significant Factors

The question remains: How do we rigorously determine the number of significant factors required to explain the measured data? The answer is that the model of the data, **RC**, must be able to reproduce all the data to within the measurement uncertainty [2]. An n-factor reconstruction, $^{n}\underline{\mathbf{D}}$, of the data matrix is calculated as shown in Eq. 5.

$$^{n}\underline{\mathbf{D}} = \underline{R}_1 \underline{C}_1 + \underline{R}_2 \underline{C}_2 + \ldots + \underline{R}_n \underline{C}_n \quad (5)$$

There are several tests available [2] that allow the analyst to determine when this criterion has been met. However, these tests all assume absolute linearity in the data, something that is rarely realized in surface analysis. Therefore, most analysts rely on visual inspection of the residual matrix, $\mathbf{D} - {}^{n}\underline{\mathbf{D}}$, to estimate the number of factors necessary. Fig. 11 shows the 1- and 2-factor reconstructions of the data matrix, **D**, shown in Fig. 1.

In Fig.11, the first factor (left pane) models the major spectral component at 390 eV reasonably well, however, the minor component at 387 eV is not well modeled – it requires a second factor to accurately reproduce the data. It is important

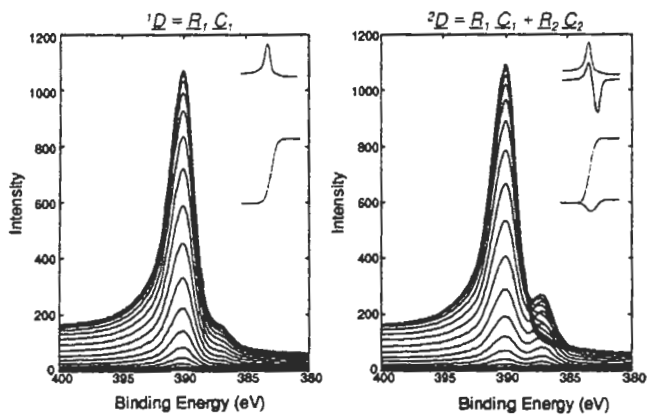

Figure 11. The 1-factor (left) and 2-factor (right) reconstructions of the data matrix, **D**, in Fig. 1. Note that the minor spectral component is not very well modeled by the 1-factor reproduction.

to note that both factors are required to model both spectral components – *i.e.* the factors do not represent pure chemical components and the abstract solution that currently exists must be transformed to reflect the physical reality of the sample.

Separation of Signal and Noise

One of the major benefits of the PCA step is that the orientation of the eigenvector axes naturally allows the elimination of some of the measurement uncertainty in the data as argued in the following:

If the eigenvectors are oriented in the direction of maximum variance and if that variance is due to spectral intensity *then* the minor variance is due to noise (and unique behavior of some spectra), therefore, the most significant (primary) eigenvectors model *mainly* signal and the minor (secondary) eigenvectors model *mainly* noise or uncertainty.

Fig. 12 shows an example of a single-component system to which some noise has been added. The eigenvectors are oriented roughly as before, in Fig. 8, however, we can

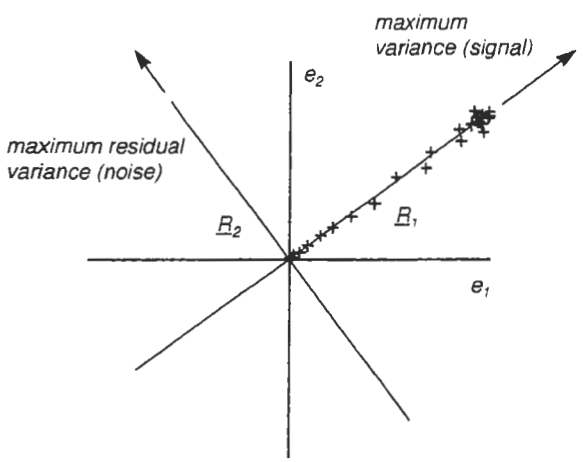

Figure 12. The eigenvectors for a single-component dataset + noise. In the case, the noise accounts for most of the projections of the data onto the minor axis, $\underline{R}_2$.

see that the projections of the data points (*i.e.* the spectra) on $\underline{R}_2$ are now non-zero. It can also be seen in Fig. 12 that the eigenvectors are averaging the spectra and this is the origin of some of the S/N improvement: the spectra are being signal-averaged across the entire dataset. The 1-factor reconstruction of one of the Ta region spectra from the Au/Ta/SiC dataset is shown in Fig. 13 where it can be seen that the

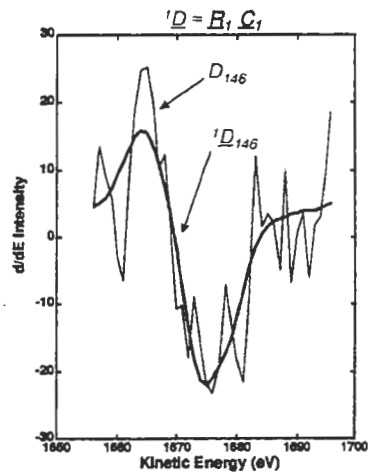

Figure 13. Reconstruction ($^1\underline{D}_{146}$) of Ta spectrum D_{146}.

reconstructed spectrum, $^1\underline{D}_{146}$, is a much higher S/N estimate of the measured spectrum D_{146} and
will lead to a more accurate calculation of the spectral intensity whether by the traditional derivative p-p method (for a single-component system) or by CLS fitting.

Defining the Real Spectral Axes by Target Transformation

We now have a PCA model of the dataset under examination, however, we do not have the physical model that allows us to measure the contributions of the individual sample components. To do this we must transform the linear combination eigenvector axes, $\underline{\mathbf{R}}$, to the real axes, $\mathbf{R}$, that represent the spectra of the components. By doing this, we can also transform the co-ordinates, $\underline{\mathbf{C}}$, on the

eigenvectors to co-ordinates on the real axes, thereby, determining the contribution of each

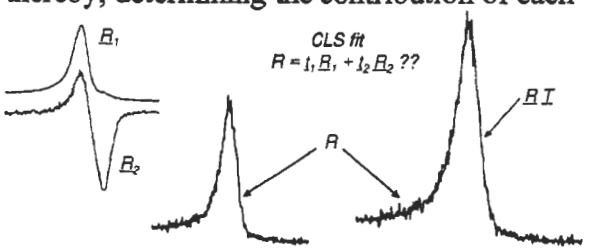

Figure 14. Target transformation test of a suspected pure component spectrum which attempts to find a linear combination of **R** that models the test spectrum, *R*.

component to each measured spectrum. The procedure used to do this is called Target Transformation [2]. Target transformation finds linear combinations of the eigenvectors (*i.e.* the eigenspectra) that approximate the known spectra of the pure sample components. This is done by a procedure called *target testing* where the idea is to find a transformation matrix, **T**, that rotates the significant eigenvectors into the real spectrum orientations. Once found,

$$\mathbf{R} = \underline{\mathbf{R}}\,\mathbf{T} \text{ and } \mathbf{C} = \mathbf{T}^{-1}\,\underline{\mathbf{C}} \qquad (6)$$

Here a connection to CLS is made – the target tests are in fact CLS fits of the eigenvectors, $\underline{\mathbf{R}}$, to the suspected pure component spectrum, R_j, and the fit is done as in Eq. 1 and shown in Fig. 14. The resulting real spectrum axes, shown in Fig. 15, are normally not orthogonal which means simply that the spectra have intensity in the same energy or mass region.

TFA Summary and Evaluation

The advantages of TFA over CLS appear to lie in the rigor that is applied to the determination of the number of factors (guessed at in CLS), the independent testing of the target spectra (assumed to be appropriate in CLS if the full solution looks acceptable), and the reduction of noise giving improved detection limits. The undesirable characteristic is that the eigenvector modeling does not allow discrimination against

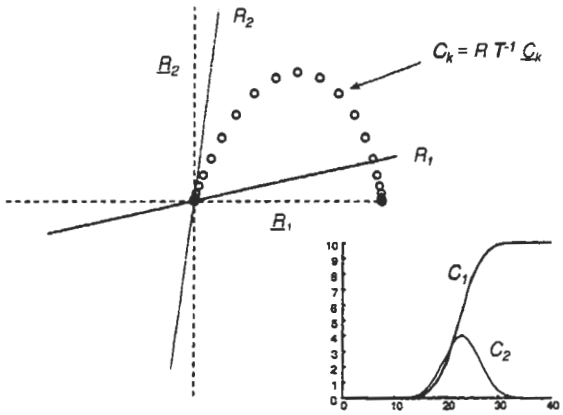

Figure 15. The target transformed eigenvector axes and a plot of the co-ordinates of the data spectra on each of the spectral axes (insert).

sample components that are of no interest – a complete solution must always be found even if we are interested in only a single component. This disadvantage can be overcome to some extent by the use of Partial Least Squares (PLS).

6. Partial Least Squares (PLS)

PLS is somewhat different from CLS and TFA in that there is a rigorous calibration step required before the PLS model can be used to estimate concentrations in unknown spectra. This requires spectra from standard samples with known surface concentrations of the component of interest. The major advantage of PLS over CLS and TFA is that it is not necessary to find the complete solution – *i.e.* the composition of the entire sample – rather, it is possible to build a PLS model for a single component in the presence of many.

PLS is performed in the following steps [3]:

1. Calibration – the eigenvectors are calculated but only for spectral intensity that is correlated to the concentration of the component of interest; build the PLS model for the component

2. Prediction in which the unknown spectrum is projected against the PLS model

An excellent example of a PLS analysis using TOF-SIMS data for the PLS model and XPS data to arrive at calibration sample concentration information is given in Ref. 5.

7. Summary and Conclusions

MVA techniques provide certain advantages over more traditional methods of calculating intensities from spectral data including better S/N, dynamic range, and the separation of contributions to spectra from different chemical species. CLS is conceptually simple, fast to use, and relatively easy to implement. TFA can be used instead if the complexity of the sample is not well known or if the data are especially noisy. PLS can be used where the analyst is interested in or capable of analyzing only some of the sample components.

All three MVA techniques discussed here assume that the dataset is linear, therefore, the analyst must be aware of the effects of non-linearity on the results.

8. References

[1] D. M. Haaland and E. V. Thomas, *Anal. Chem.*, **60**, (1990) 1193.
[2] E. R. Malinowski and D. G. Howery in *Factor Analysis in Chemistry*, John Wiley and Sons, New York, 1980.
[3] H. Wold in *Multivariate Analysis*, P. R. Krishnaiah (Ed.), Academic Press, New York (1966).
[4] M. A. Sharaf, D. L. Ilman, and B. R. Kowalski, *Chemometrics*, Vol. 82 in *Chemical Analysis*, Wiley-Interscience, ISBN 0 471-83106-9.
[5] A. Chilkoti, B. D. Ratner, and D. Briggs, *Anal. Chem.*, **65**, (1993) 1736.