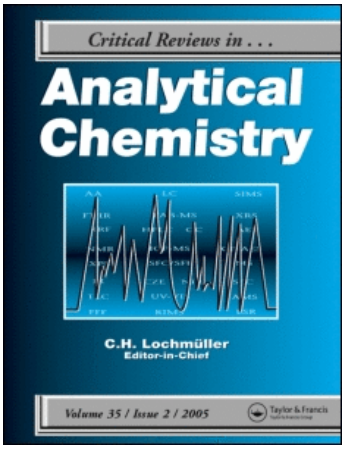

Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Introduction to Factor Analysis

C. H. Lochmüller; Charles E. Reese

Online publication date: 03 June 2010

To cite this Article Lochmüller, C. H. and Reese, Charles E.(1998) 'Introduction to Factor Analysis', Critical Reviews in Analytical Chemistry, 28: 1, 21 — 49
To link to this Article: DOI: 10.1080/10408349891194162
URL: http://dx.doi.org/10.1080/10408349891194162

Critical Reviews in Analytical Chemistry, 28(1):21–49 (1998)

Introduction to Factor Analysis

C. H. Lochmüller[1] and Charles E. Reese
Department of Chemistry, Duke University, Durham NC 27708

[1] Author to whom correspondence should be addressed.

I. INTRODUCTION

Factor analysis is a mathematical tool that can be used to examine a wide range of data sets. It has been used in disciplines as diverse as chemistry, sociology, economics, psychology and the analysis of the performance of racehorses. This tutorial is designed to provide a basic understanding of the principles underlying factor analysis. The focus of the tutorial is the analysis of a 'factor space' or 'data space'. It was written to introduce the undergraduate chemistry major to the basic concept of a 'data space' and to demonstrate how factor analysis can be used to study a 'data space'. As an aid to conceptualization, a geometric approach is used wherever possible and the actual linear algebra involved is illustrated.

II. FACTOR ANALYZABILITY

Factor analysis requires a set of data points in matrix form; following the terminology used by Malinowski and Howry, the terms 'row designee' and 'column designee' will be used to refer to the row and column identifiers of the matrix. This terminology is used because of the very wide range of data matrix types that may be analyzed by factor analysis. To be factor analyzable the data must be bi-linear; this means that the row entities and the column entities must be independent of each other.

Example 1

For the factor analysis of UV absorption data the row designees might be 'wavelength' (wl) and the column designees might be 'Solution' as shown in Table 1. In this case, data common to each solution is contained in the columns and data for each of the wavelengths is contained in the rows. If the sample concentrations are such that Beer's Law holds, then the data elements in the Table are true absorption values. The data elements are the linear sums of the absorptions of each of the components in a solution (column) at the wavelength indicated by the row. This is shown in Equation 1.

$$A_{jk} = \sum_{i=1}^{n} \varepsilon_{ji} b C_{ik} \qquad (1)$$

where A_{jk} is the absorbance measured in the j'th row and i'th column, n is the number of components, e_{ji} the molar absorbance for the i'th component at the j'th wavelength, b is the cell path length (in cm.), and C_{ik} is the concentration of the i'th component in the

k'th solution (in moles/liter). Table 2 shows the concentrations of FNB and DMP in these solutions and Table 3 shows the molar absorptivity of FNB and DMP at the wavelengths in Table 1. The cell length (b) used to collect this data was 1 cm; thus, the value in row 2 column 3 ($A_{2,3}$) in Table 1 (0.4252) is the linear sum of the absorbance of 0.00013 molar FNB at 220 nm.(e 2406) and 0.00004 molar DMP at 220 nm (e 2810). The "linear sum of terms property" is a requirement a problem must meet to be factor analyzable. It is expressed in more general terms in Equation 2.

$$d_{jk} = \sum_{i=1}^{n} f_{ji}^{r} f_{ik}^{c} \qquad (2)$$

where d_{jk} is the data element in the jth row and kth column of the data matrix, n is the number of contributing terms to each data element, f_i^r is the ith function of the row designees and f_{ji}^r is the value of this function for the jth row of the matrix. f_i^c is the ith function of the column designee and f_{ik}^c is the value of this function for the kth column of the matrix. Comparing Equation 2 and Equation 1 it is clear that for UV absorbance data, f_i^r is simply the molar absorptivity (ε) for i'th component and f_i^c is the molar concentration (C) for the i'th component. Since the path length is a scalar quantity it can be partitioned into either function or taken outside the summation altogether. For the UV absorbance data the absorbing species are the factors and the functions are the same (i.e., ε and C); for other types of data the factors may be different functions of the same or different variables.

Example 2

For an analysis of chromatographic retention data the row designees might be the sol-

TABLE 1
Absorption's of 6 Solutions of 1-Fluoro-3-Nitrobenzene (FNB) and Dimethy Phthalate (DMP) at 3 Wavelengths*

wl	Solution 1	Solution 2	Solution 3	Solution 4	Solution 5	Solution 6
215 nm	0.0997	0.0191	0.1189	0.1955	0.3183	0.25666
220 nm	0.3128	0.1124	0.4252	0.8749	1.0508	0.96276
225 nm	0.2892	0.1096	0.3988	0.8371	0.9772	0.9073

TABLE 2
Concentrations of FNB and DMP in the Six Solutions

	Solution 1	Solution 2	Solution 3	Solution 4	Solution 5	Solution 6
FNB	0.00013	0.0	0.00013	0.00013	0.00039	0.00026
DMP	0.0	0.00004	0.00004	0.0002	0.00004	0.00012

TABLE 3
Molar Absorbances of FNB and DMP at 3 Wavelengths

	215 nm	220 nm	225 nm
FNB	767	2406	2225
DMP	477	2810	2740

vent systems (eg. %methanol/%acetonitrile/%water) and the column designees might be the compound names as shown in Table 4.

In Table 4 the value of element (2,3) is 2.2533 ,the column designee for column 3 is Benzaldehyde and the row designee for row 2 is 30/0/70. Each factor is a product of a row function and a column function who's natures have yet to be determined. The sum of the products of the values of the column and row functions is the value of the data point. One published theory of chromatographic retention states that Equation 3 can be used to describe chromatographic retention in ternary mobile phase systems.

$$\ln(k) = A_1\varphi_1^2 + A_2\varphi_1^2 + B_1\varphi_1 + B_2\varphi_2 + C + D_1\varphi_2 \qquad (3)$$

Where A1, A2, B1, B2, and D are functions of the solute molecules, φ_1 and φ_2 are the concentrations of the organic solvents and C is a constant. This equation predicts that ternary retention would require six factors. Examining equation 3 it can be seen that there are four types of solvent (column) functions, namely, the zero, first and second power of the concentration and the product of the first power of the two concentrations.

As an example of a solute (row) function:

$$A_1 = (v_1 + RT)(\delta_1 - \delta_3)^2 \qquad (4)$$

where v_1 is the molar volume of the solute, R is the universal gas constant, T is the absolute temperature and δ_1 and δ_3 are the solubility parameters for solvents 1 and 3 (water) respectively. Using the terminology of equation 2, A1 would be the first row function, B2 would be the fourth row function, and $\varphi_1\varphi_2$ would be the sixth column function.

Note that some of the functions are not linear functions of the underlying variables (δ_1, φ_1, etc.); however, the data is still a *linear sum of the products* of the functions.

Example 1 (continued)

Table 5 is an expansion of Table 1, it contains the UV absorbance measurements made at 35 wavelengths for the six solutions of Table 3. The row designees are the 35 wavelengths and the column designees are the solution names.

A plot of the absorption intensities of the data in the first two columns verses wavelength is shown in Figure 1; this is a normal absorption spectra plot since the rows are in order of increasing wavelength.

There are three significant areas in the two spectra. Area A is a region where both signals increase and decrease together. In mathematical terms, we can say that in this region the two spectra have a *positive correlation coefficient* and it is probably close to one. In region B component 1 shows a peak and component 2 does not; in this region the correlation is near zero. In region C both spectra are near zero and the correlation is subject to random noise.

TABLE 4
HPLC ln(k′) Values for 5 Compounds in 3 Mobile Phases

	Acetophenone	Anisole	Benzaldehyde	Benzene	Benzonitrile
20/0/80	3.443	3.8287	2.9151	3.4445	3.0852
30/0/70	2.682	3.1653	2.2533	2.9108	2.4146
40/0/60	1.9212	2.4787	1.611	2.3743	1.7254

TABLE 5
UV Absorbance of the 6 Solutions at 35 Wavelengths

Wavelength	Cmp1	Cmp2	Mix 1	Mix 2	Mix 3	Mix 4
215	0.0997	0.0191	0.1189	0.1955	0.3183	0.2569
220	0.3128	0.1124	0.4252	0.8749	1.0508	0.9628
225	0.2892	0.1096	0.3988	0.8371	0.9772	0.9071
230	0.1344	0.0489	0.1833	0.3787	0.4521	0.4154
235	0.0555	0.0145	0.0701	0.1283	0.1812	0.1547
240	0.0498	0.0050	0.0548	0.0749	0.1545	0.1147
245	0.0742	0.0024	0.0766	0.0864	0.2250	0.1557
250	0.1128	0.0013	0.1141	0.1193	0.3397	0.2295
255	0.1532	0.0007	0.1539	0.1565	0.4603	0.3084
260	0.1763	0.0003	0.1766	0.1780	0.5291	0.3535
265	0.1680	0.0003	0.1683	0.1695	0.5043	0.3369
270	0.1275	0.0004	0.1280	0.1298	0.3830	0.2564
275	0.0726	0.0006	0.0732	0.0755	0.2183	0.1469
280	0.0325	0.0004	0.0328	0.0343	0.0978	0.0660
285	0.0168	0.0000	0.0168	0.0169	0.0503	0.0336
290	0.0148	0.0000	0.0148	0.0148	0.0445	0.0297
295	0.0158	0.0000	0.0158	0.0159	0.0474	0.0317
300	0.0163	0.0001	0.0164	0.0167	0.0490	0.0329
305	0.0147	0.0000	0.0147	0.0148	0.0441	0.0294
310	0.0113	0.0001	0.0114	0.0117	0.0341	0.0229
315	0.0066	0.0000	0.0066	0.0068	0.0197	0.0133
320	0.0037	0.0000	0.0037	0.0039	0.0110	0.0074
325	0.0028	0.0002	0.0029	0.0036	0.0085	0.0061
330	0.0027	0.0003	0.0030	0.0042	0.0084	0.0063
335	0.0027	0.0004	0.0031	0.0048	0.0085	0.0067
340	0.0026	0.0005	0.0031	0.0051	0.0082	0.0066
345	0.0023	0.0005	0.0028	0.0046	0.0073	0.0060
350	0.0019	0.0003	0.0022	0.0036	0.0060	0.0048
355	0.0014	0.0001	0.0014	0.0017	0.0041	0.0029
360	0.0008	–0.0002	0.0006	–0.0001	0.0023	0.0011
365	0.0005	–0.0003	0.0001	–0.0011	0.0010	–0.0000
370	0.0002	–0.0003	–0.0001	–0.0014	0.0003	–0.0005
375	0.0001	–0.0004	–0.0003	–0.0017	–0.0001	–0.0009
380	0.0001	–0.0004	–0.0003	–0.0019	–0.0002	–0.0010
385	0.0001	–0.0003	–0.0002	–0.0015	–0.0000	–0.0008

III. SPECTRA SPACE

An alternate way to display the spectral data of the pure components in Table 5 is to use columns 1 and 2 as the X and Y axis and the values in each row as the coordinates of a spectral point. A Cartesian plot of the first twenty rows of the data in Table 5 is shown in Figure 2. Column 1 of Table 5 was used as the X coordinate and Column 2 was used as the Y coordinate. Each point in this plot represents two absorbance measurements, one for each solution, made at one spectral wavelength. There are a couple of details to be noted about this plot. First, points 2, 3 and 4 lie on a straight line. These points are in region A (see Figure 1) of the spectra and, in general, highly correlated data will lie on a line. Second points 9, 10 and 11, found in spectral region B, also fall on a straight line. This

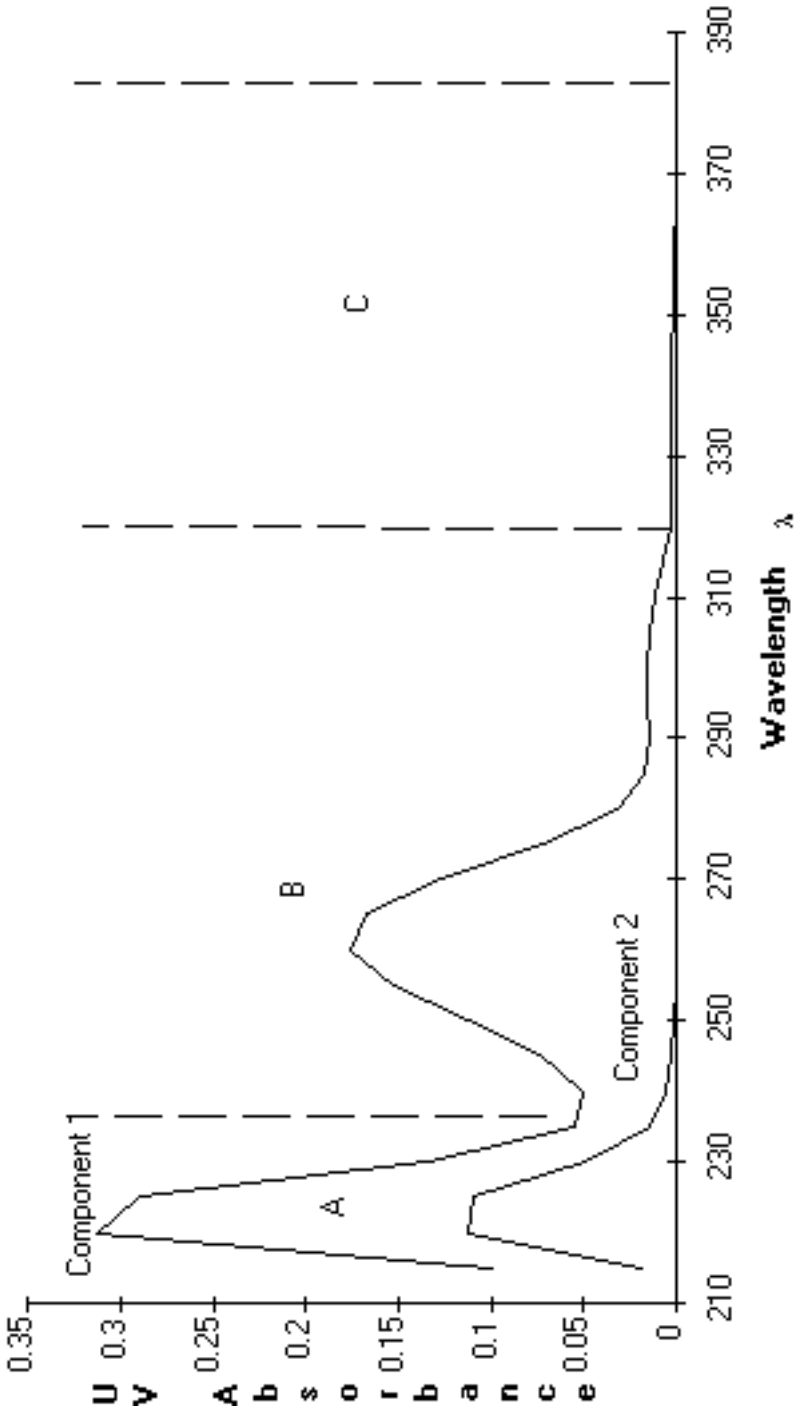

FIGURE 1. UV Absorption Spectra of Pure Components

line is, in fact, the X-axis of the plot; this is because only component 1 absorbs in this region. If rows 21-36 had been plotted then there would have been a number of points at the origin indicating the region where neither of the components adsorbs (region C).

There is very little advantage in using this type of plot for the absorbance data in Table 5, however, suppose for a moment that you had the same data as in Table 5, but the rows were not in any particular order. A normal spectra plot would not make any sense at all in this case but the Cartesian plot in Figure 2 would not change. There are many types of data where there is no obvious order in the rows and/or columns. The HPLC retention data is an example where the order of the rows (compounds) is arbitrary. These type of plots are often used to identify rows or columns of a matrix which are clustered together. For certain data sets, points that lie close together in these type of plots may imply a similarity or association between the row variables.

A disadvantage of the Cartesian plot is that the relationship of the plotted points is partially determined by the scaling and common variance (correlation) of the data vectors used to plot the points. This is illustrated in Figure 3 where the solutions 3 and 4 are used to plot the spectral data. There are some similarities between Figure 2 and Figure 3 but there are also a number of differences.

Again we see that points in the two regions of the spectra fall on straight lines but in this case none of the lines are on the axis. We can also see that the distances between the same pairs of data points in Figure 2 and 3 are different. Extracting useful information common to these two plots would be quite difficult.

IV. VECTOR PLOTS

We can get a much more useful data plot if we use the columns of our data as vectors and plot the spectral points according to their values (loadings) on these vectors. In order to make vector plots, the data in the matrix is normalized by dividing each column of the matrix by the square root of the sum of the

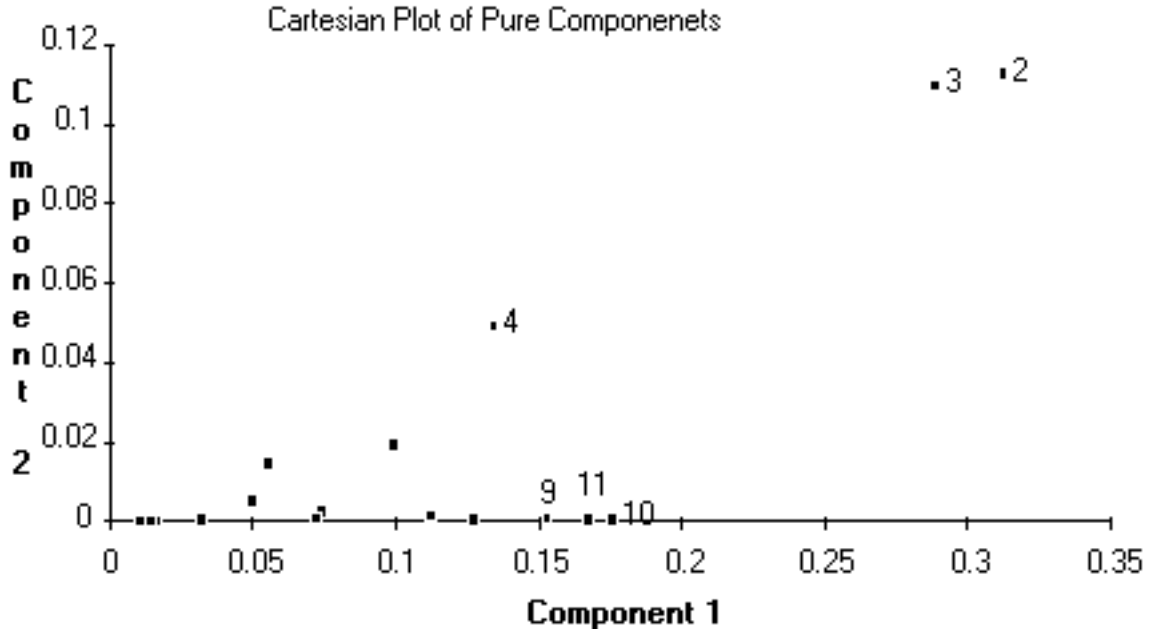

FIGURE 2. Cartesian plot of pure components.

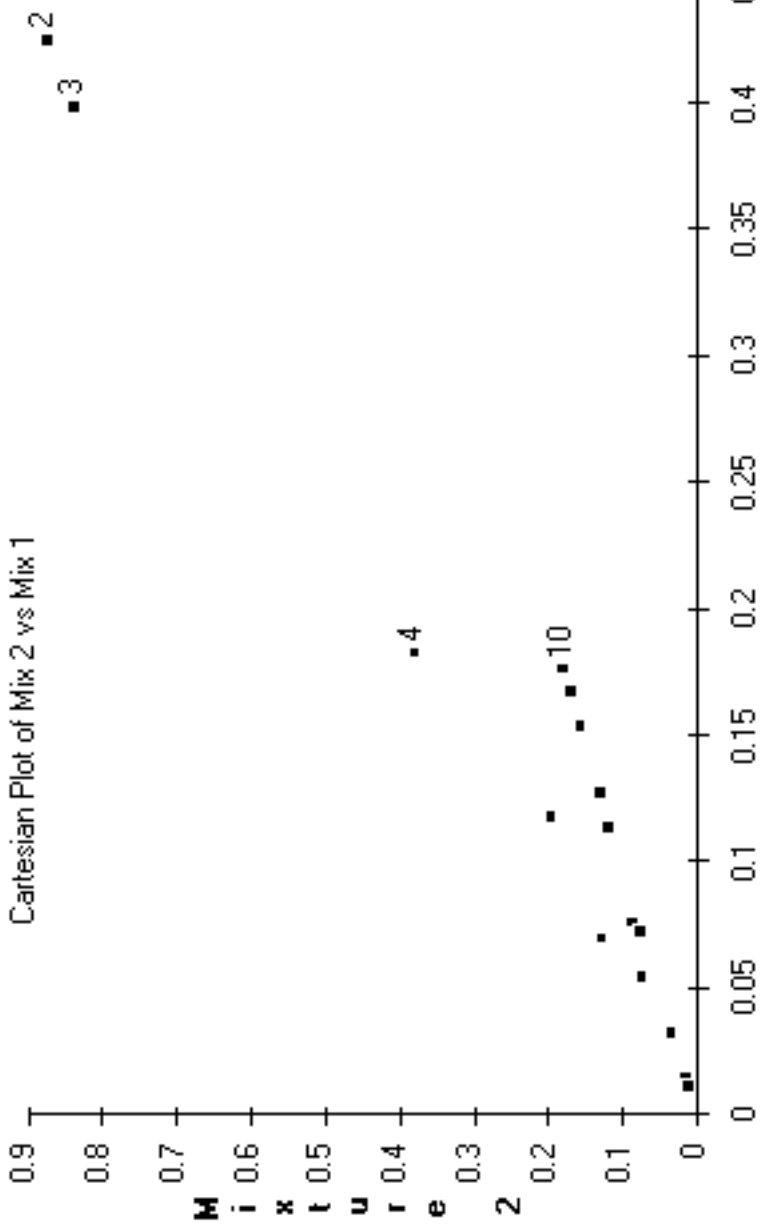

FIGURE 3. Cartesian plot of Mixture 2 vs. Mixture 3.

squares of each column element. This makes each of the vectors have the same length or scale. In order to find the angles between the vectors we form the cross product of our normalized matrix. This gives a correlation matrix which has as its elements the cosines of the angles between vectors. Table 6 shows the normalized matrix, Table 7 shows the complete correlation matrix and Table 8 shows the angles between vector 1 and each of the other vectors.

Note that in Table 8 only the angle each vector makes with column 1 (i.e., component 1) is shown and the rows have been sorted according to increasing angle from column 1. One observation can be made immediately from Table 8. In the sorted column, the largest vector angle measured from Column 1 is 37.380 for Column 2. The *vectors* for all of the *mixtures* lie *between* the *vectors* for the *pure components.*

A second observation is that the angles are additive, that is, Col 1 to Col 4 (22.085 deg.) + Col 4 to Col 2 (15.2925 deg.) is equal to Col 1 to Col 2 (37.378 deg.). The later observation is indicative of there being only the same two components in each of the solutions. If one of the mixtures had an additional UV absorbing substance then the angles would not

TABLE 6
Normalized UV Absorbance for 6 Solutions at 35 Wavelengths

Wavelength	Cmp 1	Cmp2	Mix 1	Mix 2	Mix 3	Mix 4
215	0.0854	0.0302	0.0759	0.0558	0.0820	0.0721
220	0.1272	0.0147	0.1061	0.0644	0.1194	0.0978
225	0.1934	0.0079	0.1579	0.0889	0.1803	0.1442
230	0.2627	0.0040	0.2130	0.1166	0.2443	0.1938
235	0.3022	0.0020	0.2445	0.1326	0.2808	0.2221
240	0.2881	0.0019	0.2330	0.1263	0.2676	0.2117
245	0.2187	0.0027	0.1772	0.0967	0.2033	0.1611
250	0.1244	0.0036	0.1013	0.0563	0.1159	0.0923
255	0.0557	0.0022	0.0455	0.0256	0.0519	0.0415
260	0.0288	0.0001	0.0232	0.0126	0.0267	0.0211
265	0.0254	0.0000	0.0205	0.0111	0.0236	0.0186
270	0.0271	0.0002	0.0219	0.0119	0.0252	0.0199
275	0.0280	0.0004	0.0227	0.0124	0.0260	0.0206
280	0.0252	0.0001	0.0204	0.0110	0.0234	0.0185
285	0.0194	0.0005	0.0158	0.0087	0.0181	0.0144
290	0.0113	0.0003	0.0091	0.0050	0.0105	0.0083
295	0.0063	0.0002	0.0051	0.0029	0.0058	0.0047
300	0.0048	0.0010	0.0041	0.0027	0.0045	0.0038
305	0.0047	0.0018	0.0042	0.0031	0.0045	0.0040
310	0.0046	0.0026	0.0043	0.0036	0.0045	0.0042
315	0.0044	0.0030	0.0042	0.0038	0.0044	0.0042
320	0.0039	0.0027	0.0038	0.0034	0.0039	0.0037
325	0.0032	0.0020	0.0031	0.0027	0.0032	0.0030
330	0.0023	0.0004	0.0020	0.0012	0.0022	0.0018
335	0.0014	–0.0011	0.0009	–0.0001	0.0012	0.0007
340	0.0008	–0.0019	0.0002	–0.0008	0.0006	–0.0000
345	0.0004	–0.0020	–0.0001	–0.0011	0.0002	–0.0003
350	0.0001	–0.0022	–0.0004	–0.0013	–0.0001	–0.0006
355	0.0001	–0.0023	–0.0004	–0.0014	–0.0001	–0.0006
360	0.0002	–0.0019	–0.0003	–0.0011	–0.0000	–0.0005

TABLE 7
Correlation Matrix of Data in Table 6

	Cmp 1	Cmp2	Mix 1	Mix 2	Mix 3	Mix 4
Cmp 1	1.0000	0.7946	0.9902	0.9266	0.9986	0.9817
Cmp 2	0.7946	1.0000	0.8717	0.9646	0.8260	0.8956
Mix 1	0.9902	0.8717	1.0000	0.9701	0.9963	0.9987
Mix 2	0.9266	0.9646	0.9701	1.0000	0.9454	0.9812
Mix 3	0.9986	0.8260	0.9963	0.9454	1.0000	0.9905
Mix 4	0.9817	0.8956	0.9987	0.9812	0.9905	1.0000

TABLE 8
Angles between the Vectors in Degrees

Col ID	Name	Angle Col 1	Angle Col 2
Col 1	Cmp 1	0.0000	37.3783
Col 5	Mix 3	3.0707	34.3076
Col 3	Mix 1	8.0323	29.3460
Col 6	Mix 4	10.966	26.4116
Col 4	Mix 2	22.085	15.2925
Col 2	Cmp 2	37.378	0.0000

all be additive and more then two dimensions would be required to plot the data.

These relationships along with others can best be understood by looking at a vector plot of the data in Table 6. Figure 4 is a vector plot of columns 1 and 2 of the normalized data in Table 6.

In order to simplify plotting the vector C1 has been plotted as the Y-axis. Vector C2 is at an angle of 37.378 deg. from the Y-axis. The loadings of point 2 are illustrated. Dashed lines are drawn from point 2 perpendicular to the vectors. The vector loading is the distance from the origin to the interception of this perpendicular with the vector. This is the distance C1-L2 for vector 1 and C2-L2 for vector 2. The actual loadings as given in Table 2 are 0.536 and 0.676. The loading is *NOT* the length of the perpendicular from the point to the vector (e.g., C1-P2). Again we observe that points 9,10 and 11 lie on a straight line and points 2,3 and 4 also lie on a straight line.

Since the loadings in this example are (normalized) UV absorption's, we can also observe is that no vector can be at a greater angle from the Y-axis then the vector C2 in the plot. The reason we can state this is that the loading of point 10 and several other points would have negative values on any vector drawn at a larger angle than 37.4 deg. from vector C1.

The great advantage that vector plots have over Cartesian plots for this type of data analysis is that the *data pattern is INVARIANT with the choice of vector axes*. This is illustrated in Figure 5 where points have been plotted for all pairs of C1 with the other vectors in Table 3.

All of the plotted points for a given wavelength lie at the same location no matter which set of vectors is used for axes in this two dimensional plot. The data is therefore invariant with the coordinate axes and any pair of coordinate axes can be used to analyze the relationships between the spectral points. If we had

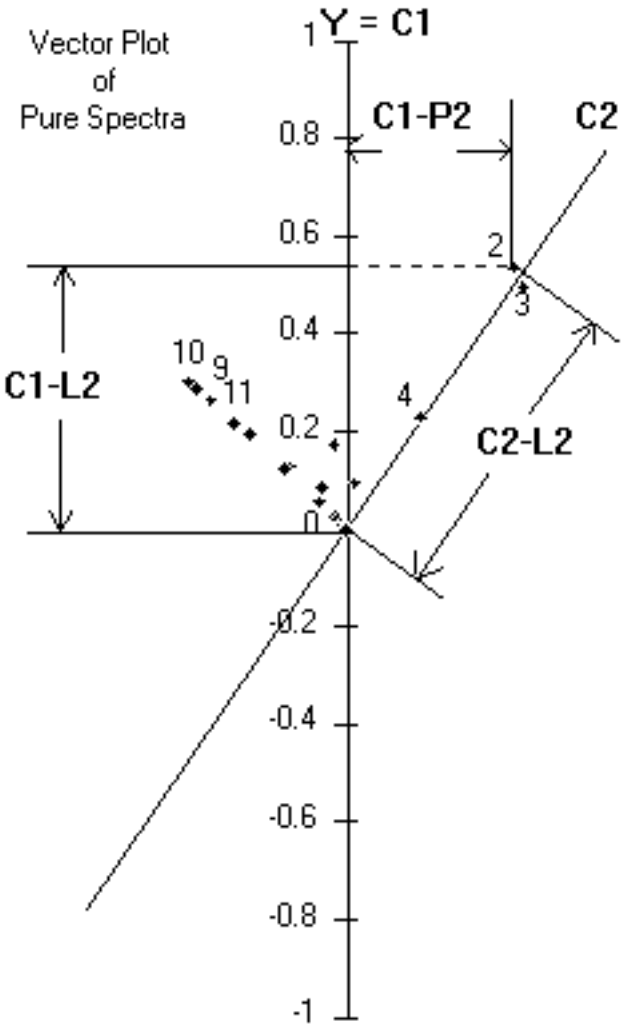

FIGURE 4. Vector plot of pure components.

plotted this data matrix without knowing anything about it we could conclude that there are only two components or 'factors' that contribute to the variance in the data. Note the last point carefully we do NOT know how many components there are in the mixture only that there are two factors contributing to the variance.

There are several reasons that there may be more components in the mixture than there are factors contributing to the variance:

1. There may be non-absorbing components.
2. There may be components that have the same spectra.
3. There may be components that are at *fixed, constant ratios.*

By 3 we mean that if you take a mixture of several components and dilute that mixture serially, you will still have only a single factor. The exception to this is if the spectra is dependent upon the absolute concentration in the range examined. Some examples of this are spectra where the initial concentration is to high so that Beer's Law is not applicable. In this case the number of factors is larger because a single function (e.g., e) is not sufficient to describe the absorption of the component at a given wavelength. Fluorescence spectra

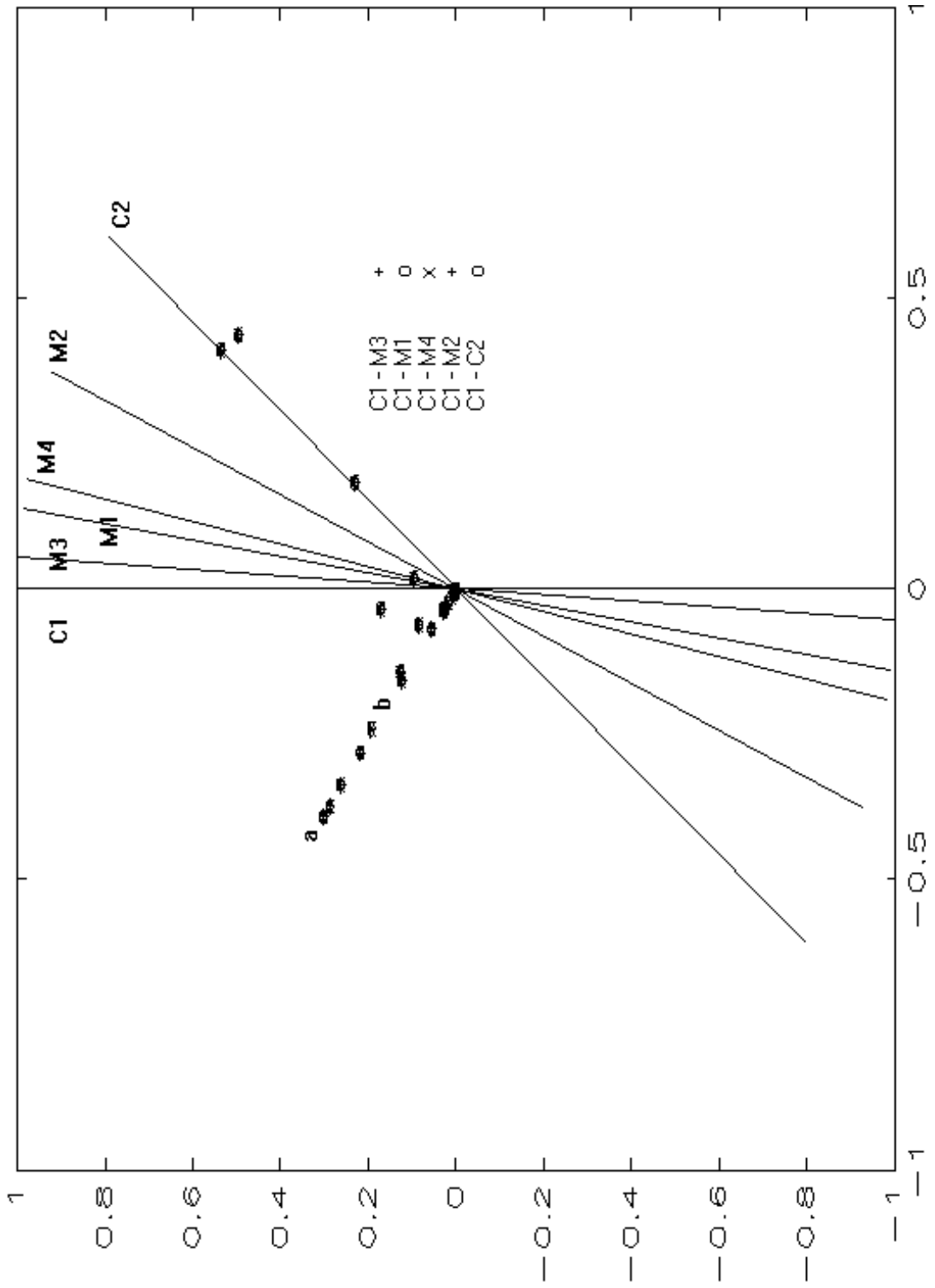

FIGURE 5. Plot of UV absorbances in Table 6 using Col 1 as the Y axis and each of the other columns as the second axis.

of substances that undergo direct energy transfer are also non-linear with concentration.

We also note that all of the (mathematical) solutions to this data matrix can be represented by a vector in this plot. That is, the spectrum of ANY mixture of the two components in this mixture is a set of loadings on a vector in this plot, as long as Beer's Law holds. If one only had the mixtures to plot these data and the spectrum of a compound suspected to be one of the components, one could test to see if there was a vector we could place in this plot which had the same loadings as the suspected spectrum. If it did not, then we would know that this compound was NOT one of our components. Likewise we could test unknown mixtures to see if they contained only these two compounds; however, remember the three points listed above.

If the spectrum of one of the components contains a region where it does not absorb and the other components do, we can recover the spectra of the pure component. In our example, component 2 does not absorb in region B. Referring to Figure 5, we can draw a line through points (of region B) from a to b. We can then draw a vector through the origin perpendicular to this line. In the case of Figure 5, we can see that C2 is such a perpendicular vector and the loadings of the points on C2 are by definition the spectra of component 2. We cannot extract the spectra of component 1 this way as there are no regions where it alone does not absorb.

One of the most important features of the vector plot analysis for many data types is the invariant structuring of the data. The vector plot completely details the STRUCTURE of the data, however, the graphic vector approach to data analysis has some severe limitations.

It is difficult to analyze 3-dimensional (i.e., 3-component) data with the graphic approach and it is impossible to extend it to even higher order (n-dimensional) data. Also, it can be difficult to determine the number of components and the placement of real data points is affected by noise in the vectors chosen for plotting. A better method of obtaining an invariant data structure is the use of Principal Component Analysis (PCA) or singular value decomposition (SVD).

V. FACTOR ANALYSIS

PCA is an eigenanalysis technique that extracts a set of eigenvectors and their associated eigenvalues by a step-wise procedure. Singular Value Decomposition is a computationally more robust procedure than the PCA which gives very similar results. SVD will be used exclusively for computation in this text, however, the terminology will be that which is usually associated with PCA. The first eigenvector is extracted in a manner that causes it to account for a maximum amount of variance in the data. After each eigenvector is extracted a residual data matrix is calculated and the procedure is repeated until there are no significant eigenvectors left. The variance accounted for by each eigenvector is measured by its eigenvalue. The variance is equal to the square of the eigenvalue. Examination of the eigenvalues and their relative magnitudes allow an estimation of the number of significant 'factors' or components in the matrix. The singular values, obtained from the singular value decomposition, for the data matrix of Table 5 are the diagonal elements in Table 9; these singular values are the square roots of the eigenvalues. Note that there are only two non-zero singular values. This is what is expected from a matrix of two component mixtures. With real data the other eigenvalues ((SVs)2) would not be exactly zero because of noise which is always found in real data. A significant part of the effort in many data analyses is the determination of the number of real (non-noise) factors; a number of tests have been developed to help decide the number of real factors. The number of real factors is called the rank of the data matrix.

TABLE 9
Eigenvalues for the PCA Analysis of the Data in Table 5

2.9374	0	0	0	0	0
0	0.3839	0	0	0	0
0	0	0.0000	0	0	0
0	0	0	0.0000	0	0
0	0	0	0	0.0000	0
0	0	0	0	0	0.0000

The eigenvectors are orthonormal. This means that they are orthogonal (at right angles) to each other and they are normalized to unit length. The first two SVD eigenvectors for the data in Table 5 are shown in Table 10. The PCA eigenvectors are the same as the SVD eigenvectors except for an occasional change in the direction of an eigenvector and the numerical errors associated with the computational procedures. It can be confirmed that they are orthogonal by forming there cross product and that they are normalized can be demonstrated by summing the squares of their elements.

A scatter plot of these eigenvector loadings is shown in Figure 6. Note that this is a vector plot; the axes are at right angles because the cross product of the eigenvectors is zero which is the cosine of 90 deg.

This is of course the same data structure that we saw in Figures 3 and 4. PCA produces a set of abstract factors or coordinates delineating our data structure. Looking at the plot, we can see that the first eigenvector seems to be very similar to the real component 1 vector in Figures 3 and 4. We can quantitate how similar these vectors are by forming their cross prod-

TABLE 10
Eigenvectors (Row) for the SVD Analysis of the Data in Table 5

Wavelength	EV-1	EV-2	Wavelength	EV-1	EV-2
215	0.1631	–0.0618	300	0.0217	–0.0401
220	0.5968	0.3178	305	0.0195	–0.0365
225	0.5610	0.3489	310	0.0151	–0.0275
230	0.2573	0.1419	315	0.0087	–0.0160
235	0.0971	0.0035	320	0.0049	–0.0087
240	0.0741	–0.0753	325	0.0039	–0.0053
245	0.1021	–0.1613	330	0.0041	–0.0039
250	0.1513	–0.2687	335	0.0043	–0.0026
255	0.2036	–0.3759	340	0.0042	–0.0016
260	0.2336	–0.4366	345	0.0038	–0.0013
265	0.2226	–0.4163	350	0.0030	–0.0014
270	0.1693	–0.3140	355	0.0019	–0.0028
275	0.0969	–0.1754	360	0.0008	–0.0039
280	0.0435	–0.0774	365	0.0001	–0.0042
285	0.0222	–0.0417	370	–0.0002	–0.0038
290	0.0196	–0.0370	375	–0.0005	–0.0037
295	0.0209	–0.0392	380	–0.0005	–0.0040
			385	–0.0004	–0.0034

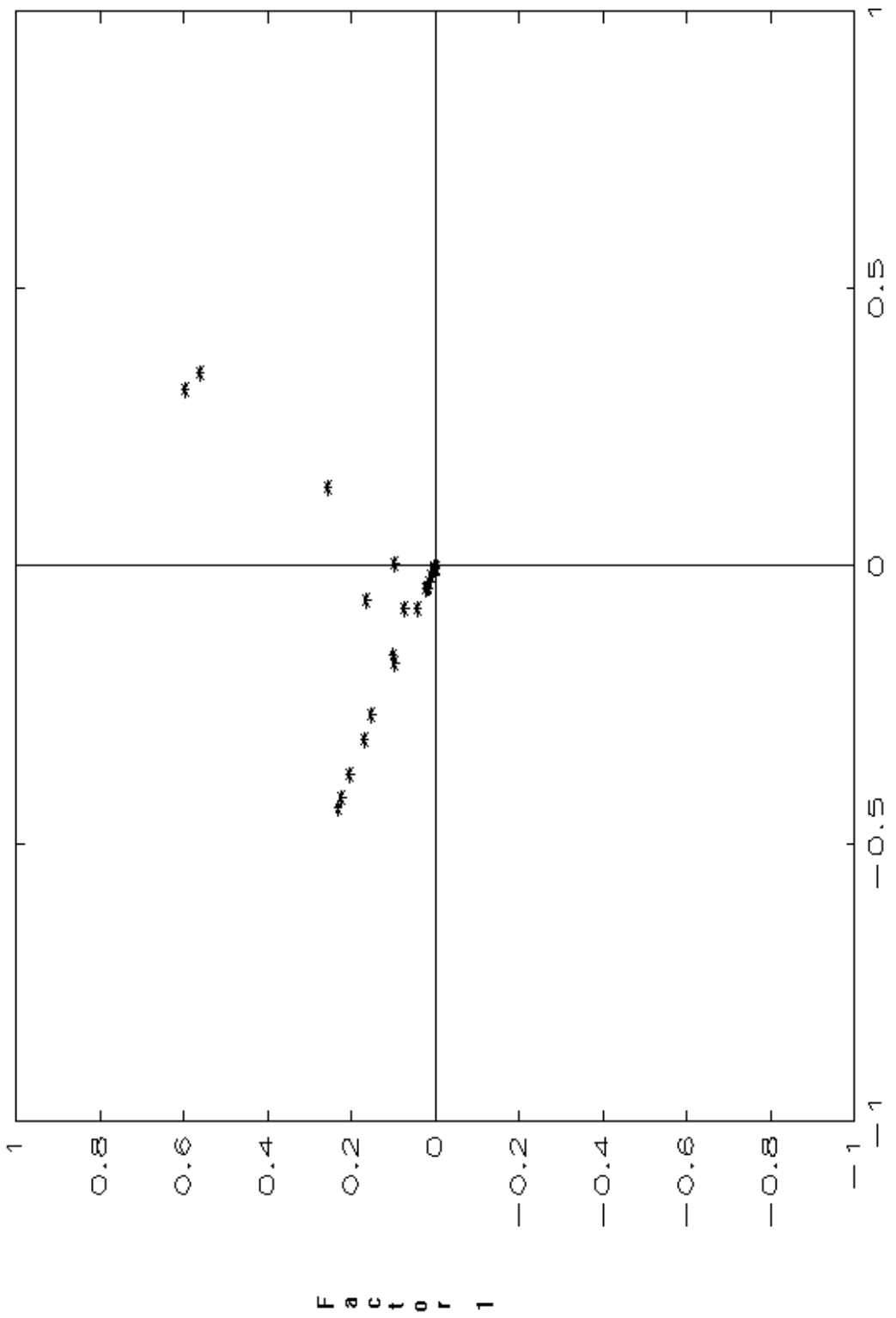

FIGURE 6. Plot of eigenvector loading.

TABLE 11

Fac	EV	Var	IE	RE	IND	REV	F	P
1	1.96e–01	98.32	1.18	2.90	0.12	4.11e–02	1.37e+02	0.000
2	–2.50e–01	1.68	0.00	0.00	0.00	8.67e–04	9.99e+14	0.000
3	9.21e–01	0.00	0.00	0.00	0.00	1.09e–33	7.41e–16	1.000
4	–1.38e–01	0.00	0.00	0.00	0.00	8.43e–35	2.80e–17	1.000
5	1.63e–01	0.00	0.00	0.00	0.00	6.10e–35	6.59e–18	1.000
6	–7.32e–02	0.00	NaN	NaN	NaN	2.72e–35	0.00e+00	1.000

IE, RE, IND are multiplied by 100.

Fac = Factor number, EV = eigenvalue, Var = Percent variance accounted for IE = imbedded error, RE = Real error in factor, IND = Malinowski indicator function, REV = reduced eigenvalue, F = Malinowski F test for factor, P = Probability factor is random (From F and degrees of freedom). NaN means not a number as there is no valid value.

ucts (using the normalized vector in Table 2). Remember that this cross product is equal to the cosine of the angle between the vectors. The cross product gives a value of .9864 for the cosine and 9.46 deg. for the angle; Table 4 shows that the first eigenvector is nearly coincident with the vector of Mix 4 (10.96 deg.). Mixture 4 actually has a ratio of C1 to C2 of 2 to 3. This illustrates an important point, the PCA solution is an ABSTRACT solution. One should not try to equate these ABSTRACT factors with REAL factors. Note that in this case we can see immediately that factor 2 *CANNOT be a REAL factor* as many of the points have NEGATIVE loadings on it and UV spectra do not have negative absorbencies.

VI. RANK

One of the important pieces of information that can be obtained from the abstract factor analysis is the number of components contributing to the variance. This is called the rank of the matrix and a number of tests have been developed to determine this number. In general, the greater the level of noise in the data relative to the variance caused by the *smallest* real factor, the pore difficult it becomes to determine the rank. Table 11 shows a rank analysis of the PCA analysis of the data in Table 5.

The individual tests work better for some data sets than for others but the most generally reliable are the IND, REV and F (P) tests. All of the tests give the correct number of factors for data which has no noise in it. Table 12 is the result of an analysis of the data of Table 5 to which one percent 'normal' noise has been added. Normal means that the noise has a mean of zero and has a gaussian distribution.

These tests are interpreted as follows:

1. Var is the percentage of the variance accounted for by the factor; for all factors greater than the rank of the matrix the variance is it is due to noise in the data. Factor 2 accounts for 1.79% of the variance and factor 3 accounts for only .02% of the variance and we conclude that the rank of this matrix is two. Observe that for this data set most of the noise is also accounted for by the first two factors. Although we added 1% (RMS) noise to the pure data, the noise factors (3-6) account for only .05% of the variance. This

TABLE 12

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.62e+00	98.16	1.24	3.04	0.12	4.11e–02	1.24e+02	0.000
2	1.57e–01	1.79	0.33	0.58	0.04	9.25e–04	6.38e+01	0.001
3	1.95e–03	0.02	0.36	0.51	0.06	1.47e–05	1.03e+00	0.385
4	9.97e–04	0.01	0.40	0.49	0.12	1.04e–05	5.63e–01	0.531
5	9.20e–04	0.01	0.43	0.47	0.47	1.48e–05	5.72e–01	0.588
6	7.78e–04	0.01	NaN	NaN	NaN	2.59e–05	0.00e+00	1.000

Total variance in pure data matrix:	8.7754
RMS error added to data matrix:	0.0879
RMS percent error	1.0014%
RMS relative to component 1	1.6151%
RMS relative to component 2	8.5912%
RMS error in reproduced data (2F)	0.6329%

IE, RE, and IND are multiplied by 100.

will be discussed in the section on Error analysis.

2. The imbedded error (IE) is the error in the reproduced data matrix; it should decrease for real factors and then increase for the noise factors. It should show a minimum at the number of factors and in Table 8 it shows a minimum for factor 2 as expected.
3. The real error (RE) is an estimate of the error in the raw data matrix; it should be less than the experimental error estimated from the conditions of the data acquisition. Since this is a synthesized data matrix the estimated error is just the error added or 1%. The real error for 1 factor is 3.04% so we conclude that at least two factors are needed. The real error for 2 factors is .58%, which is less than 1%, so we conclude that 2 factors are sufficient.
4. The indicator function (IND), developed by Malinowski, should have a minimum at the rank of the matrix and in Table 8 there is a distinct minimum at 2 factors as expected.
5. The reduced eigenvalue (REV) should show a sharp decrease for factors greater than the rank and the REV's for the noise factors should be statistically equal. This is again found to be the case in Table 8.
6. The statistical F-test with its associated probability (F and P) gives the probability that a factor is random (i.e., noise); it should have a probability of zero for real factors and 1 for noise factors. This test is one of the most reliable test for finding real factors. As with most statistical F-tests a score of <.05 or <.10 is considered to indicate a real factor and larger values indicate a noise or random factor. The probability used as a cutoff for true factors depends upon the expected noise level in the data. For noisier data use a higher cutoff value. The P data in Table 8 indicates that there are two real factors.

The factors from the PCA analysis can be used to reconstruct the data matrix. The data analyzed in Table 8 was reconstructed using two factors. Plots of the first two column of this data (components 1 and 2) are shown in Figure 7a and 7b respectively.

The black dots are the original data, the triangles are the data with added noise and

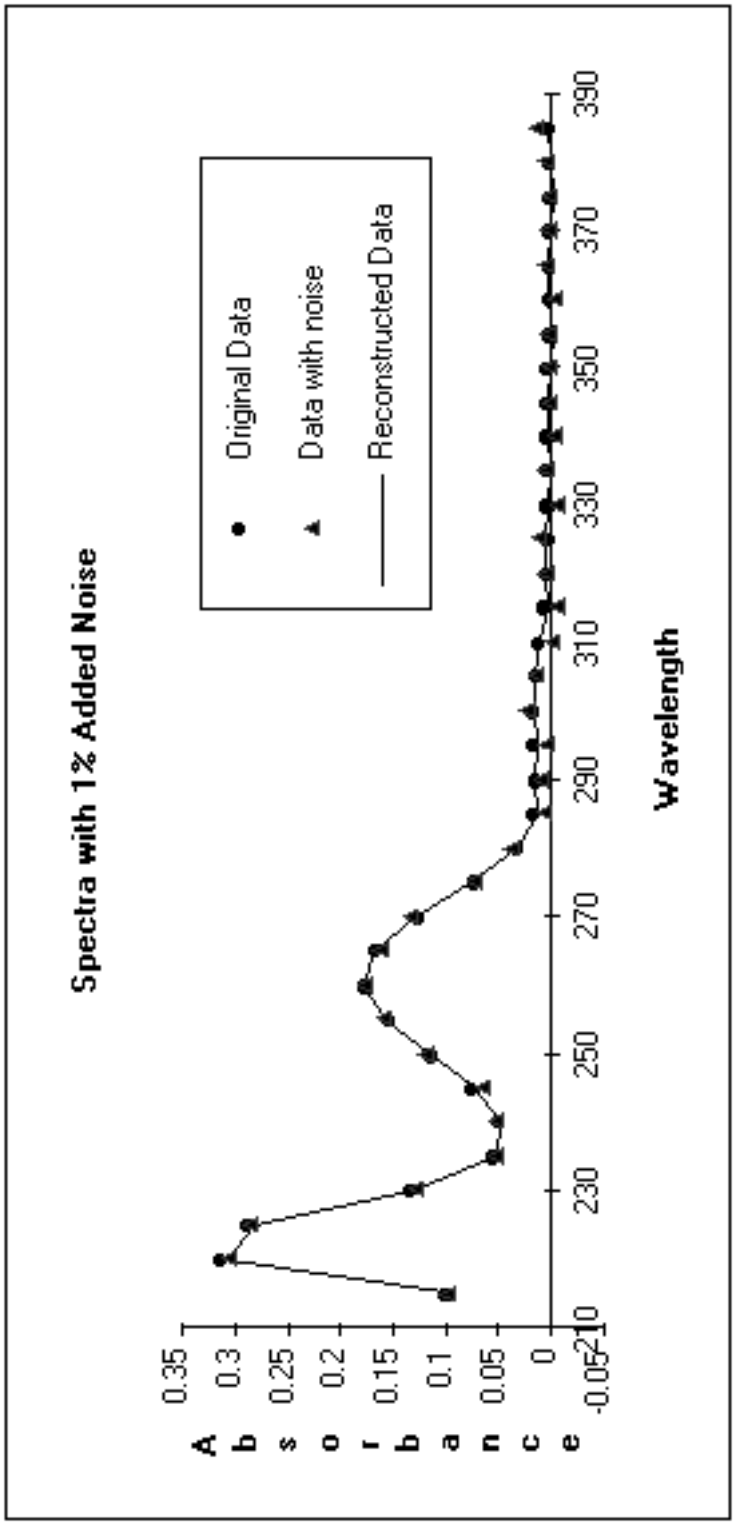

FIGURE 7a. Reconstructed data for pure component 1 from a data set with 1% added noise.

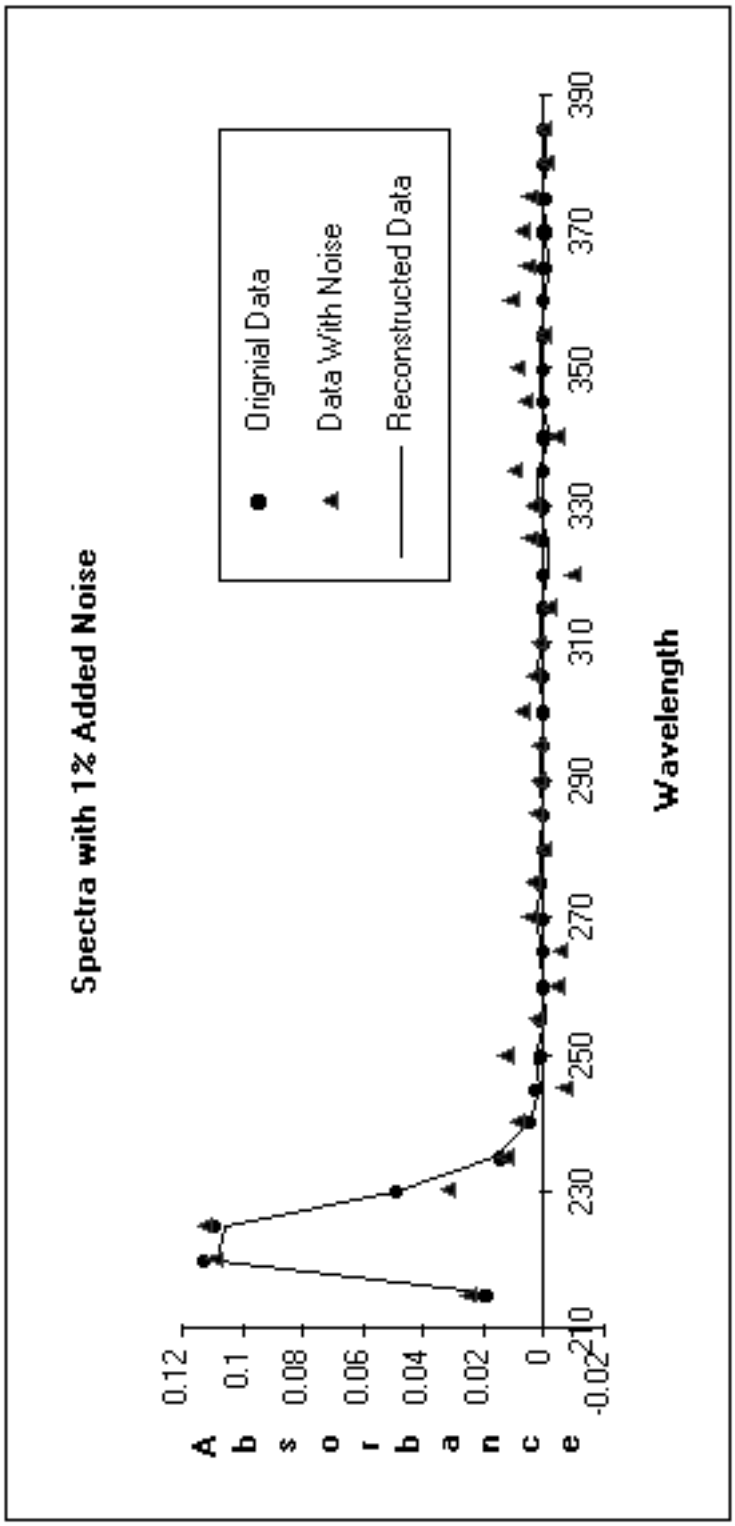

FIGURE 7b. Reconstructed data for pure component 2 from a data set with 1% added noise.

the line is the reconstructed data. The reconstructed data is much closer to the original data then it is to the noisy raw data. This is because of the noise extracted from the data matrix by the factor analysis.

VII. ERROR/RANK ANALYSIS

A comprehensive discussion of error analysis is beyond the scope of this tutorial; readers are referred to 'Factor Analysis in Chemistry' for a detailed treatment of this topic. A number of analyses of data sets will be presented here, however, to act as a guide as to what error levels are expected for different levels and types of noise which might be expected in chemical data. Tables 13–19 are for addition of 'normal' noise to the data of Table 1 in various percentages as given (RMS percent error).

All of the rank tests give correct results for the 1% and 2% noise cases. Note that the noise is given as a percentage of the total variance in the data. For 2% noise this means that the noise is 17.6% of the variance of the second component. This would be regarded as rather noisy data. When 5% noise is added only IE and REV indicate two factors. The P test might be considered as giving a 'lukewarm' indication as a second factor is more than twice as likely as a third factor.

At 10%, all of the tests fail and we would conclude that there is only one factor accounting for all of the variance. This is not surprising given that the noise is almost equal to component 2 (86%). Quite often, a signal 3 times the (RMS) noise level is required to qualify as a real response.

In addition to the magnitude of noise that is added, the distribution of the noise also affects the error (rank) analysis. Table 18 shows an error analysis of data containing 2 percent noise, in which, the noise is uniformly distributed with a mean of zero.

As in Table 11 all of the error tests indicate 2 factors. Table 19 shows an error analysis of data with 2 percent added noise; in this case the noise is not only uniformly distributed but it is also positive valued. In this case, the IE value indicates 3 factors and the P value for the third factor is a little lower.

Table 20 shows the error analysis results from an analysis of random, i.e., all noise, data. The data had a normal distribution and the data matrix was 35×6 just as in the ex-

TABLE 13
1% Normal Noise

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.62e+00	98.25	1.21	2.96	0.12	4.10e–02	1.31e+02	0.000
2	1.48e–01	1.69	0.36	0.61	0.04	8.71e–04	5.27e+01	0.002
3	1.94e–03	0.02	0.40	0.57	0.06	1.47e–05	8.21e–01	0.432
4	1.30e–03	0.01	0.44	0.54	0.14	1.36e–05	6.09e–01	0.517
5	1.15e-03	0.01	0.46	0.51	0.51	1.85e–05	6.12e–01	0.577
6	9.07e–04	0.01	NaN	NaN	NaN	3.02e–05	0.00e+00	1.000

Total variance in pure data matrix:	8.7754
RMS error added to data matrix:	0.0879
RMS percent error	1.0014%
RMS relative to component 1	1.6151%
RMS relative to component 2	8.5912%
RMS error in reproduced data (2F)	0.5624%

IE, RE, and IND are multiplied by 100.

TABLE 14
1% Normal Noise

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.60e+00	98.22	1.22	2.99	0.12	4.10e–02	1.28e+02	0.000
2	1.51e–01	1.73	0.35	0.61	0.04	8.89e–04	5.55e+01	0.002
3	1.87e–03	0.02	0.39	0.56	0.06	1.41e–05	8.13e–01	0.434
4	1.44e–03	0.02	0.42	0.51	0.13	1.50e–05	7.54e–01	0.477
5	1.14e–03	0.01	0.40	0.44	0.44	1.84e–05	8.08e–01	0.534
6	6.85e–04	0.01	NaN	NaN	NaN	2.28e–05	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.0879
RMS percent error 1.0014%
RMS relative to component 1 1.6151%
RMS relative to component 2 8.5912%
RMS error in reproduced data (2F) 0.5803%
IE, RE, and IND are multiplied by 100.

TABLE 15
2% Normal Noise

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.54e+00	97.85	1.34	3.28	0.13	4.07e–02	1.06e+02	0.000
2	1.68e–01	1.92	0.69	1.20	0.07	9.89e–04	1.58e+01	0.016
3	7.70e–03	0.09	0.77	1.08	0.12	5.83e–05	8.91e–01	0.415
4	5.06e–03	0.06	0.83	1.02	0.25	5.27e–05	6.70e–01	0.499
5	4.61e–03	0.05	0.79	0.87	0.87	7.43e–05	8.47e–01	0.526
6	2.63e–03	0.03	NaN	NaN	NaN	8.77e–05	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.1802
RMS percent error 2.0534%
RMS relative to component 1 3.3117%
RMS relative to component 2 17.6163%
RMS error in reproduced data (2F) 1.2733%
IE, RE, and IND are multiplied by 100.

ample data. This data clearly shows no clear number of factors for IE, IND and P. Many analysts use the Var value as guide to the number of factors. This is not recommended, as the Var value is very susceptible to misinterpretation especially for different sizes of matrices. Table 20 shows that for this 35×6 matrix the first factor accounts for more than six times the variance as does the 6th factor. Since the data is all noise and has a uniform distribution, how can this be? The answer is that the PCA procedure selects the eigenvectors in a very non- random way. The first eigenvector is chosen to account for the most variance, and as a corollary, there is less variance left for the remaining eigenvectors. Although the data is distributed randomly, there is still a distribution of loadings on the vectors existing in the data space. The eigenanalysis will *always* use the vector with the largest load-

TABLE 16
5% Normal Noise

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.71e+00	96.47	1.74	4.27	0.17	4.15e–02	6.37e+01	0.000
2	1.99e–01	2.21	1.69	2.92	0.18	1.17e–03	3.14e+00	0.151
3	5.02e–02	0.56	1.82	2.57	0.29	3.80e–04	1.03e+00	0.385
4	3.14e–02	0.35	1.90	2.33	0.58	3.27e–04	7.92e–01	0.467
5	2.34e–02	0.26	1.87	2.04	2.04	3.77e–04	7.75e–01	0.541
6	1.46e–02	0.16	NaN	NaN	NaN	4.87e–04	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.4394
RMS percent error 5.0070%
RMS relative to component 1 8.0753%
RMS relative to component 2 42.9561%
RMS error in reproduced data (2F) 3.1370%
IE, RE, and IND are multiplied by 100.

TABLE 17
10% Normal Noise

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.84e+00	91.71	2.76	6.76	0.27	4.21e–02	2.58e+01	0.004
2	2.99e–01	3.11	3.45	5.97	0.37	1.76e–03	1.13e+00	0.348
3	1.82e–01	1.88	3.89	5.50	0.61	1.38e–03	8.13e–01	0.434
4	1.28e–01	1.33	4.25	5.21	1.30	1.33e–03	6.45e–01	0.506
5	1.23e–01	1.28	4.00	4.38	4.38	1.98e–03	8.86e–01	0.519
6	6.72e–02	0.70	NaN	NaN	NaN	2.24e–03	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.8783
RMS percent error 10.0083%
RMS relative to component 1 16.1414%
RMS relative to component 2 85.8635%
RMS error in reproduced data (2F) 6.4788%
IE, RE, and IND are multiplied by 100.

ing as the first eigenvector and so it is NOT random *by definition*. The fewer the number of data points and the greater the number of eigenvectors extracted the more pronounced this distribution of loadings would be. Table 21 shows the results of analyzing various size matrices of random numbers. The variance distribution is very misleading for the 10×6 case and, in the case of real data, it is the number of data points containing appreciable amounts of variance that need to counted in the 'effective' matrix size. For example in our 35×6 example data for component 1 most of the pixels (rows) greater than 21 have zero absorbance so the 'effective' size of its matrix is 21×6 and for component 2 the 'effective' size is only 7×35. Unfortunately, in most cases larger data matrices mean more work, therefore, careful attention must be paid to error analysis.

TABLE 18
2% Uniform Noise with a Mean of 0

Fac	EV	Var	IE	RE	IND	REV	F	P
1	8.55e+00	97.84	1.34	3.28	0.13	4.07e–02	1.06e+02	0.000
2	1.68e–01	1.92	0.70	1.21	0.08	9.87e–04	1.54e+01	0.017
3	9.08e–03	0.10	0.74	1.04	0.12	6.88e–05	1.14e+00	0.364
4	5.66e–03	0.06	0.74	0.90	0.23	5.90e–05	9.51e–01	0.432
5	3.48e–03	0.04	0.73	0.80	0.80	5.62e–05	7.59e–01	0.544
6	2.22e–03	0.03	NaN	NaN	NaN	7.41e–05	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.0316
RMS percent error 2.0254%
RMS relative to component 1 3.2666%
RMS relative to component 2 17.3766%
RMS error in reproduced data (2F) 1.2021%
IE, RE, and ND are multiplied by 100.

Table 19
2% Uniform Noise with All Positive Values

Fac	EV	Var	IE	RE	IND	REV	F	P
1	9.08e+00	98.31	1.22	2.99	0.12	4.32e–02	1.36e+02	0.000
2	1.49e–01	1.62	0.40	0.70	0.04	8.78e–04	4.12e+01	0.003
3	3.98e–03	0.04	0.37	0.52	0.06	3.02e–05	2.00e+00	0.252
4	1.13e–03	0.01	0.40	0.49	0.12	1.17e–05	6.30e–01	0.510
5	1.05e–03	0.01	0.40	0.43	0.43	1.70e–05	7.76e–01	0.540
6	6.57e–04	0.01	NaN	NaN	NaN	2.19e–05	0.00e+00	1.000

Total variance in pure data matrix: 8.7754
RMS error added to data matrix: 0.0316
RMS percent error 2.0254%
RMS relative to component 1 3.2666%
RMS relative to component 2 17.3766%
RMS error in reproduced data (2F) 1.8002%
IE, RE, and IND are multiplied by 100.

The second major source of problems in error analysis is systematic error. This is the problem of error introduced as a response to instrument drift. It may be caused by changes in temperature, line voltage, air pressure, aging of lamps or any other variable that cannot be held constant. Use of the term 'error' for these types of changes is dependent on the variables being studied. For example changes in lamp characteristics in a spectrophotometer are noise to the spectroscopist interested in analyzing some compound, but they may be the factors of interest to a manufacturer of lamps trying to determine why the lamps age.

In this example, five percent noise is about as much as can be tolerated if information about component 2 is desired. Figure 8 shows a plot of Components 1 and 2 along with the noisy data and the data reconstructed using two factors.

Figure 8a clearly shows that the spectrum of component 1 is recovered quite well. Even the

TABLE 20
Random Data Analysis for 36×6 Matrix

Fac	EV	Var	IE	RE	IND	REV	F	P
1	9.75e–03	30.86	0.46	1.12	0.04	4.64e–05	1.04e+00	0.354
2	7.14e–03	22.59	0.59	1.02	0.06	4.20e–05	9.13e–01	0.393
3	6.33e–03	20.03	0.63	0.89	0.10	4.79e–05	1.08e+00	0.376
4	4.93e–03	15.61	0.57	0.70	0.18	5.14e–05	1.37e+00	0.362
5	1.99e–03	6.31	0.59	0.64	0.64	3.21e–05	6.62e–01	0.565
6	1.46e–03	4.61	NaN	NaN	NaN	4.85e–05	0.00e+00	1.000

IE, RE, IND are multiplied by 100.

TABLE 21
Variance Distribution for Set of Random Data

Fac	10×2	100×2	1000×2	10×3	100×3	1000×3	10×6	100×6	1000×6
1	77.68	53.22	0.52	61.46	39.42	35.73	45.00	22.97	19.80
2	22.32	46.78	49.48	22.67	32.81	32.89	29.88	21.02	17.19
3				15.87	27.77	31.37	12.13	16.73	16.36
4							5.92	15.49	16.33
5							4.86	12.99	15.98
6							2.21	10.78	14.34

large noise value at pixel 11 has been smoothed out. On the other hand if one only had the noisy data to plot (triangles) it would be quite difficult to identify the spectrum as being that of pure component 1. Thus, even with 5% noise (8% relative to component 1's variance) satisfactory spectra are obtained. Figure 8b is quite noisy and it would be difficult to identify component 2 by its recovered spectra using the data reproduction approach. There is however another way that we can test to see if a particular spectrum is consistent with a given data set. This approach is called target testing.

VIII. TARGET TESTING

Remember that in the section on vector plotting we said that the data STRUCTURE is invariant. It is the same for the example data set regardless of which data columns are used to define it. As a corollary of this we can say that all of the spectra consistent with this data set can be represented by a vector in the plot of Figure 6. Even if the data set consists of mixtures only and does not include the pure spectra, the data structure will be the same. Figure 9 shows a vector plot of the loadings on the first 2 eigenvectors of the data in Table 5 with 5% added error. Note that this is the same as an ordinary Cartesian (XY-scatter) plot because the eigenvectors (factors) are orthogonal. This data structure is very similar to that shown in Figure 6 for the case with no added error. This is because of the *averaging* and *error extraction* accomplished with the factor analysis.

Using the target testing procedure we can determine the vector with the best fit (i.e., loadings) to any spectra we might suspect is contributing to the variance in the data. Figure 10 shows the results of target testing the pure spec-

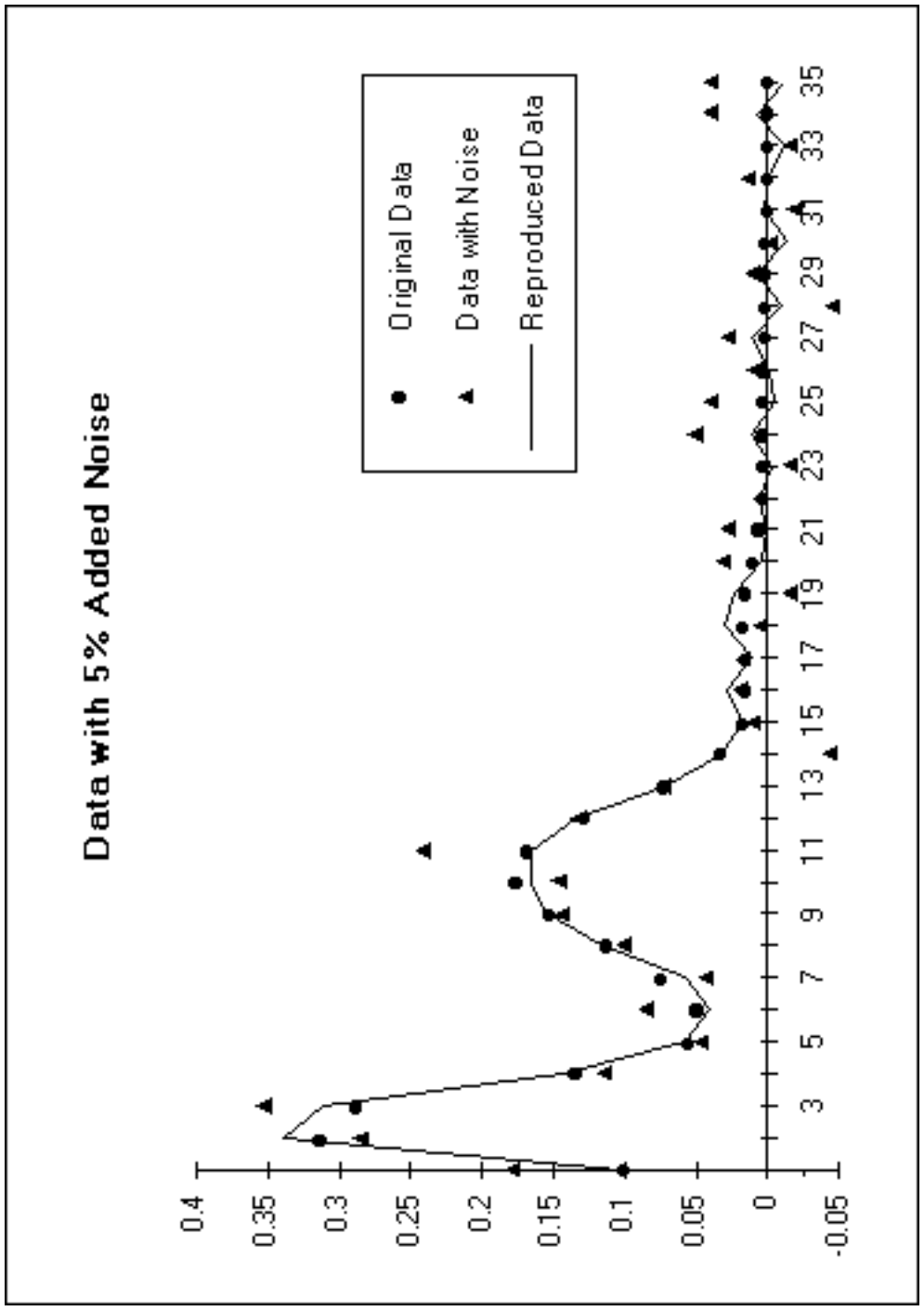

FIGURE 8a. Plot of Component 1 data with 5% added noise.

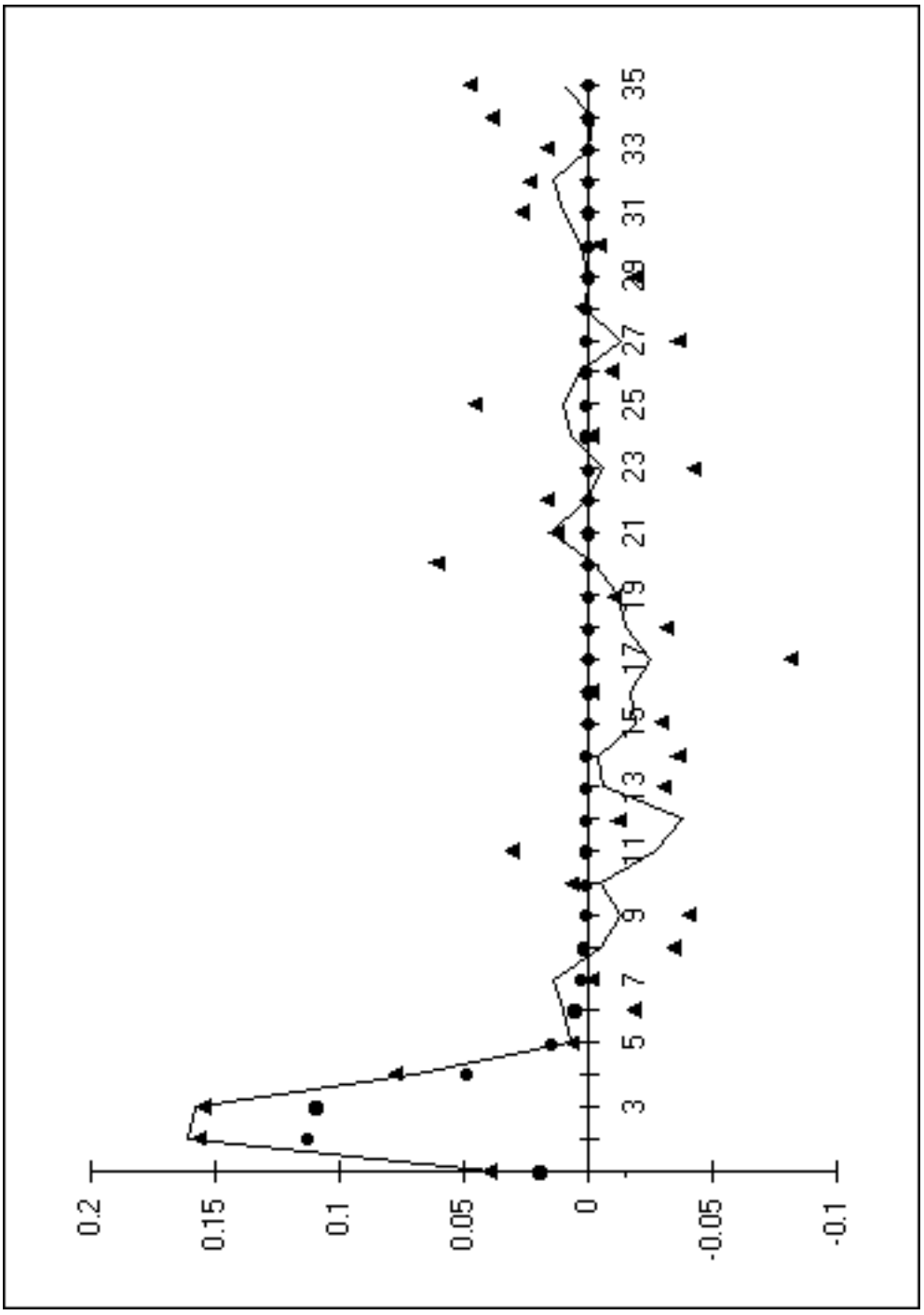

FIGURE 8b. Plot of Component 2 data with 5% added noise.

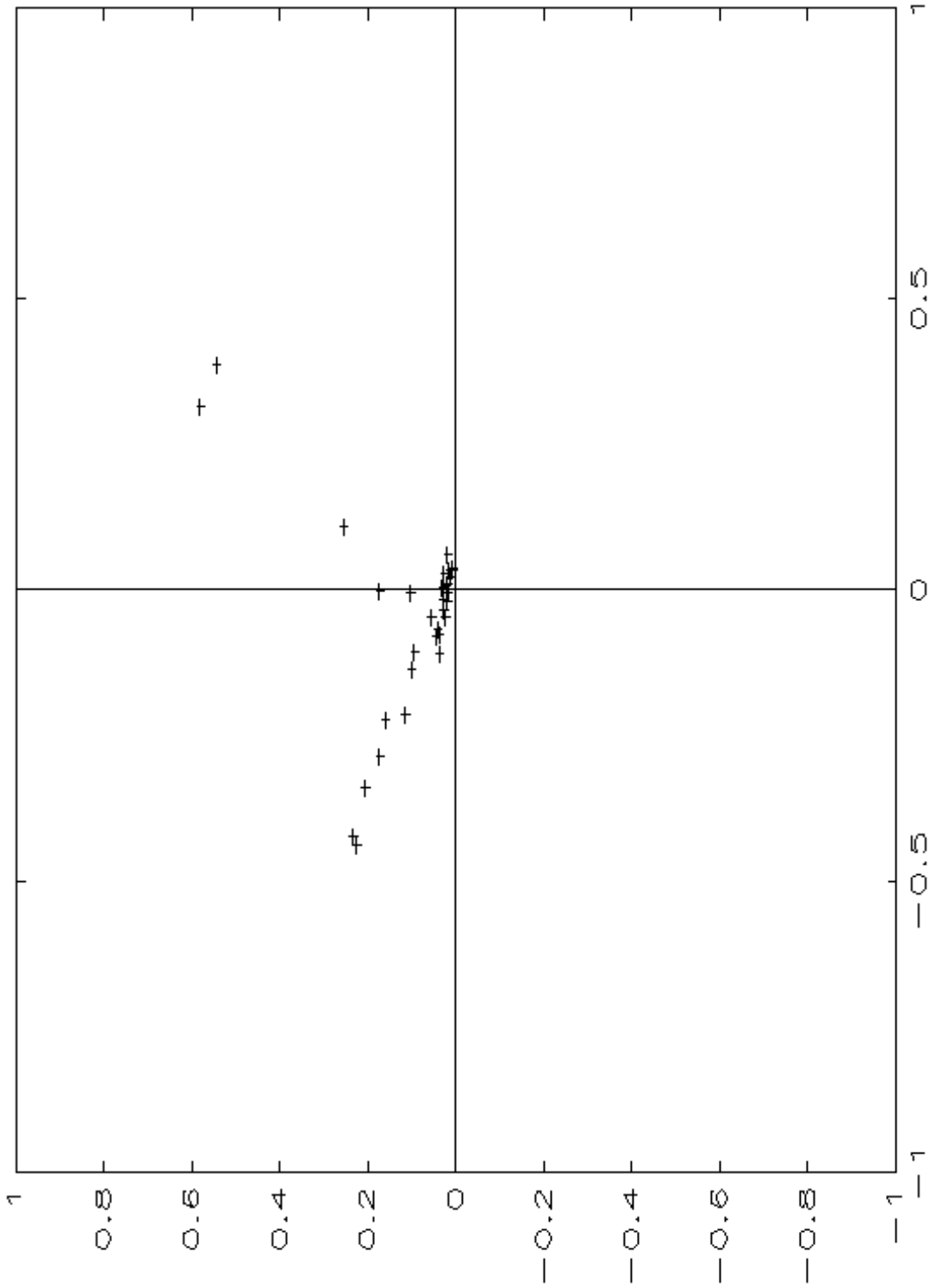

FIGURE 9. Eigenvector plot of data with 5% added noise.

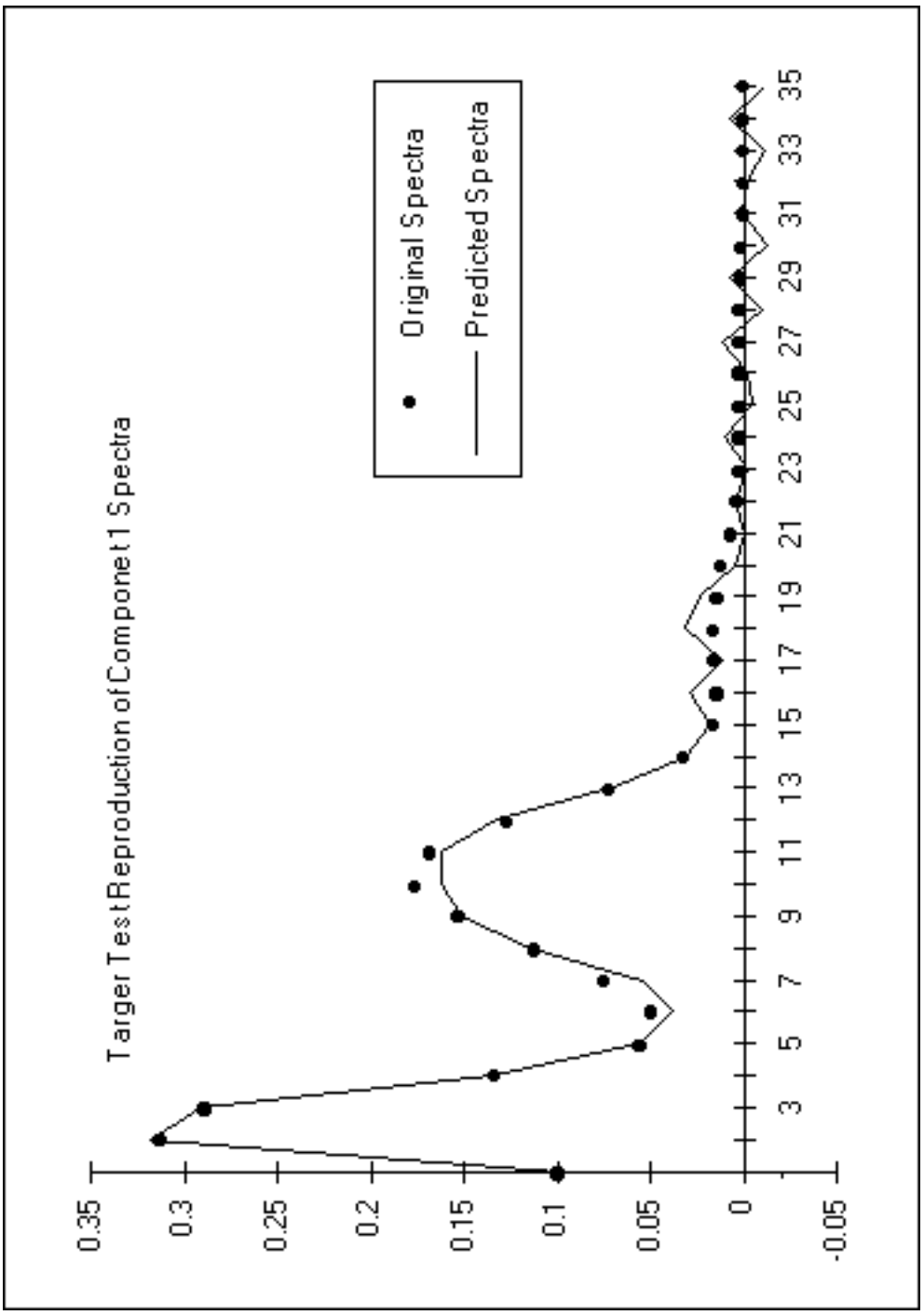

FIGURE 10a. Target test of Component 1 spectra.

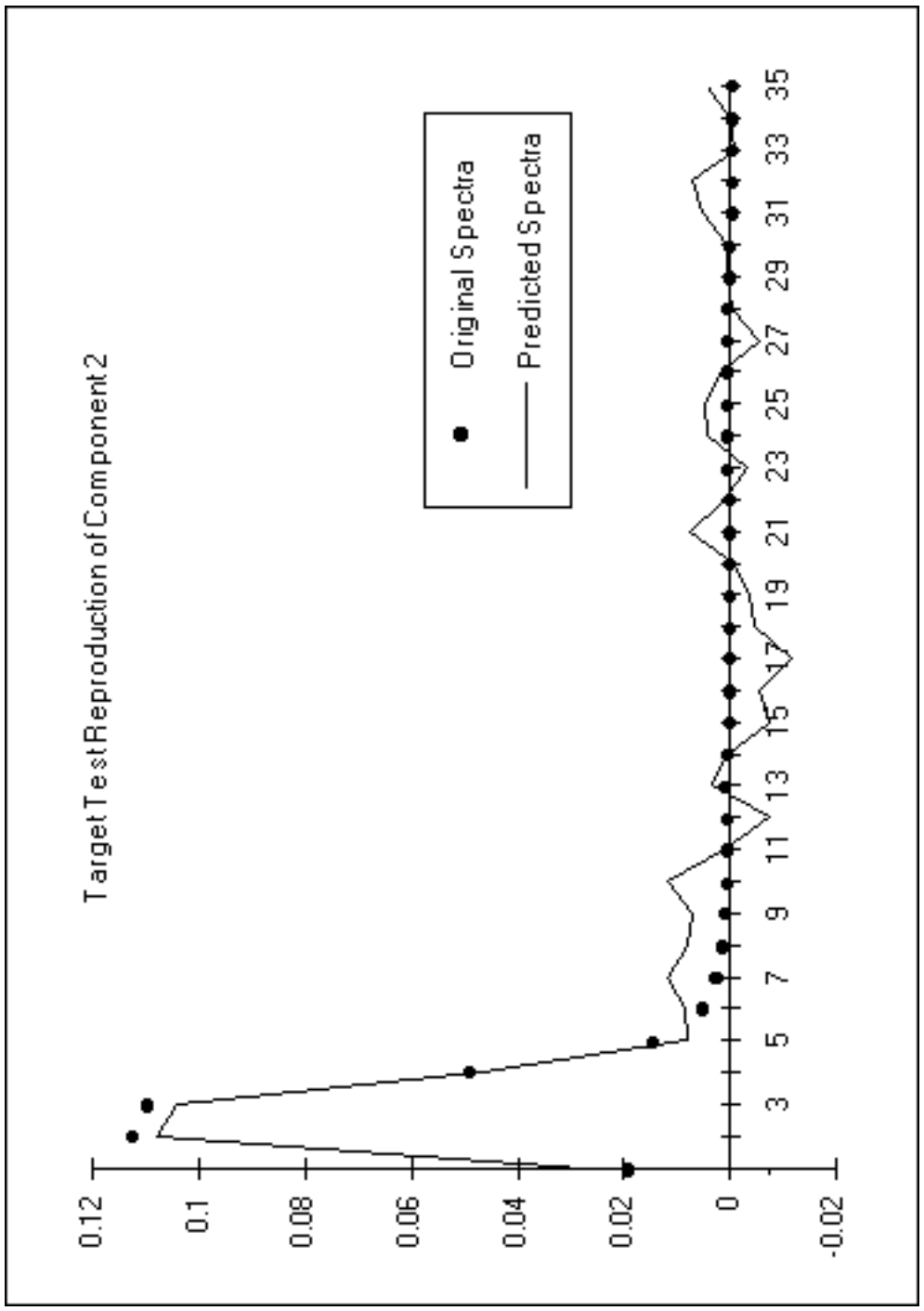

FIGURE 10b. Target test of Component 2 spectra.

tra into the vector space of Figure 9. The solid dots are the input test vector (the known spectra) and the lines are the loadings on the vector which has the closest loadings to the known spectra and which exists in the data space defined by the first two eigenvectors.

The target testing procedure indicates that both of the suspected spectra are *compatible* with the data set. By compatible we mean that the spectra *may* be contributing to the data space but it is not proven. The positive result my be due to the presence of some other compound with similar spectra or even several compounds who's spectra sum up to give spectra similar to one of our compounds.

Target testing is also valuable for indicating that a compound may be incompatible with a given data set. In this case the best fit vector does not resemble the target vector and we can conclude that the given spectra is incompatible with the data set. Again, we have not *proven* that the compound does not exist in the data space. A false negative test may result from the presence of an absorbing compound who's concentration is at the same ratio in each mixture. In this case the factor analysis would behave as if the composite spectra were due to a component in the mixtures and the pure spectra of the real components were absent.

ACKNOWLEDGMENT

Developed under support in part by a grant from the National Science Foundation (NSF #DUE-92-50282). Leadership in Laboratory Development Program awarded to Duke University, Principal Investigator: Professor Charles H. Lochmüller.