The observed Pb/Ba intensity ratio for the muffler sample was off-scale for the range of concentrations used for the calibration plot in Figure 4. However, the calibration curve should be linear and the Pb/Ba atomic ratio could be estimated by substituting the observed intensity ratio into the equation for the least squares calibration curve. The results of the ESCA analysis using this procedure agree to within 10% of the lead concentration measured by atomic absorption and exactly with the results from spark source mass spectrometry. The concentration of lead in gasoline determined by ESCA again using the calibrated plot, Figure 4, was 0.09% which agreed to within 10% of the value determined by spark source mass spectrometry.

CONCLUSION

Use of an inert matrix improves quantitative measurements by ESCA in cases where a homogeneous blend can be achieved. This improvement is reflected by a decrease in relative standard deviation and a more realistic x–y intercept.

The same cation showed different atomic sensitivities for different salts. Because the sensitivity differences are derived from quantitative plots for separate salts, they are real differences and are not the result of some artifact. The crystal structures of the various salts probably cause sensitivity differences by affecting the average escape depth of the photoejected electrons. If this is true, one should view with caution any tables which give listings of relative atomic sensitivities in ESCA, since the use of different compounds in these studies might have resulted in vastly different relative sensitivities.

Sensitivity differences in ESCA caused by crystal structures of different compounds complicate the use of ESCA for total metal if the exact nature of the salt is unknown. Quantitative determination of total metal concentration can be achieved if all the metal atoms can be chemically converted into the same form and thus have the same crystal structure. The use of ESCA for real samples is limited only by the availability of the chemical means necessary to achieve this conversion.

ACKNOWLEDGMENT

The authors acknowledge help of the following: C. Blount, for the soil sample; R. Smith for the atomic absorption analyses; and C. Taylor, for the spark source mass spectrometry work.

LITERATURE CITED

(1) K. Siegbahn, C. Nordling, A. Fahlman, R. Nordberg, K. Hamin, J. Hedman, G. Jokansson, T. Bergmark, S. E. Karlsson, I. Lindgren, and B. Lindberg, "ESCA-Atomic, Molecular and Solid State Structure Studied by Means of Electron Spectroscopy", Almquist and Wiksells, Uppsala, 1967.
(2) P. E. Larson, *Anal. Chem.*, **44**, 1678 (1972).
(3) W. E. Swartz, Jr., and D. M. Hercules, *Anal. Chem.*, **43**, 1774 (1971).
(4) G. D. Nichols, D. M. Hercules, R. C. Peek, and D. J. Vaughan, *Appl. Spectrosc.*, **28**, 219 (1974).
(5) A. C. Shead and G. F. Smith, *J. Am. Chem. Soc.*, **53**, 483 (1931).
(6) ASTM, D 526-70, 211 (1970).
(7) M. G. Natrella, "Handbook 91", U.S. Department of Commerce, National Bureau of Standards, August 1, 1970, p 5–16.
(8) P. W. Carr, Department of Chemistry, University of Georgia, Athens, GA, personal communications, 1974.
(9) P. J. Durrant and B. Durrant, "Introduction to Advanced Inorganic Chemistry", 2nd ed., John Wiley and Sons, Inc., New York, 1970, p 390.
(10) C. D. Wagner, *Anal. Chem.*, **44**, 1050 (1972).
(11) B. L. Henke, *J. Phys. (Paris), Colloq.*, **4**, 115 (1971).
(12) C. J. Power, "Attenuation Lengths of Low Energy Electrons in Solids," (unpublished) National Bureau of Standards, Washington, DC, 20234.
(13) R. C. Weast, Ed., "Handbook of Chemistry and Physics", 52nd ed., The Chemical Rubber Co., Publishers, Cleveland, OH, 1971–72.
(14) L. Bragg, C. F. Claringbull, and W. H. Taylor, "Crystal Structures of Minerals", Cornell University Press, Ithaca, NY, 1965.
(15) F. A. Cotton and G. Wilkinson, "Advanced Inorganic Chemistry", 2nd ed., Interscience Publishers, New York, 1966.
(16) H. Krebs, "Fundamentals of Inorganic Crystal Chemistry", McGraw-Hill, London, 1968.
(17) L. B. Leder, H. Mendlowitz, and L. Marton, *Phys. Rev.*, **101**, 1460 (1956).
(18) L. E. Cox and D. M. Hercules, *J. Electron Spectrosc. Relat. Phenom.*, **1**, 193 (1973).
(19) M. Klasson, J. Hedman, A. Berndtsson, R. Nilsson, and C. Nordling, *Phys. Scr.*, **5**, 93 (1972).

RECEIVED for review September 30, 1974. Accepted March 17, 1975. This work was supported by the National Science Foundation under Grant GP-32484.

Methods of Factor Analysis of Mass Spectra

Richard W. Rozett and E. McLaughlin Petersen

Chemistry Department, Fordham University, Bronx, NY 10458

The multivariate statistical technique of factor analysis is applied to the interpretation of mass spectra. We systematically investigate the effect of data selection (masses/intensities), and data transformation (correlation/covariance matrix, about the mean/about the origin, direct data/transpose data) upon the results of the factor analysis. Criteria for data compression (eigenvalue methods/data recalculation techniques) are studied. Analytical methods of factor orientation (principal component/varimax/quartimax) are compared. Empirical methods of factor transformation, useful for hypothesis-testing, prediction, and interpretation, are illustrated. Throughout, the mass spectra of the 22 isomers of the alkyl benzenes with the formula $C_{10}H_{14}$, are used as examples. These methodological studies determine the best methods for the factor analysis of mass spectra.

One impressive difference between organic mass spectrometry and other spectral methods such as IR and NMR, is the absence of a convenient theoretical method capable of relating the experimental measurements to the fundamental properties of the sample. The fragmentation patterns of complex compounds are obscure (*1–4*). To become meaningful, they must be supplemented by the study of isotopically-labeled variants, metastable transitions, and energy distributions (*5*). Without such studies, the information contained in the spectrum must be extracted by chemical intuition supported by qualitative theories from carbonium ion chemistry. As a consequence, the hundreds of measurements made during the recording of a mass spectrum are consistently under-utilized. On the other hand, the ordinary mass spectrum is overly detailed. It contains redundant measurements which complicate the stor-

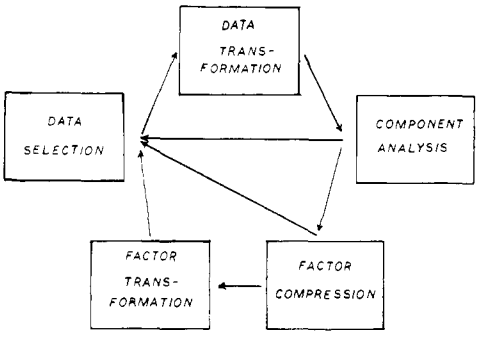

Figure 1. Steps in factor analysis

age of mass spectra and aggravate the procedure of library search and sample identification. Mass spectrometry needs a method of data reduction which will compress spectral data, and extract structural and mechanistic information in a rapid, objective and quantitative fashion. We propose that factor analysis is such a technique.

Factor analysis (FA) is a multivariate statistical technique; it employs the simultaneous analysis of multiple measurements on many compounds. It is a well-developed, computerized method capable of handling the large amount of detail found in organic mass spectra. Although it was first applied in the psychological and social sciences (*6–9*), it has been applied to chemical and physical problems (*10, 11*), including the analysis of IR spectra (*12*), NMR chemical shifts (*13–15*), GC retention times (*16–18*) and peak deconvolution (*19*), and determination of the number of components in a mixture (*20–22*).

One function of FA is the compression of data: the segregation of significant information from random or redundant measurements. A second function is the classification of data by establishing the optimal number of categories and their class definitions. FA is a natural prelude to pattern recognition because of its ability to define the minimum number of patterns needed for the pattern discrimination procedure. FA can interrelate the mass spectra (MS) of different compounds, and relate MS to other chemically-significant properties, such as structural, thermodynamic, and kinetic properties. It can test hypotheses about the origin of MS quantitatively and predict the value of related properties. Finally, FA can identify the hidden variable or factors behind the mass spectra and thus help us interpret mass spectral information (*9*).

An outline of FA includes the topics of data selection and transformation, component analysis, factor compression and transformation, and data regeneration (Figure 1). FA operates on a data matrix with n rows and m columns. Each row represents the mass spectrum of a single compound; each column, the intensities of many compounds at a certain mass (data selection). From the data matrix, one calculates an m by m covariance or correlation matrix (data transformation). Eigenanalysis of this symmetric matrix produces m eigenvectors and eigenvalues (component analysis, cf. section on factor compression). The eigenvalues arranged in decreasing order can be used to determine the dimensionality of the original data (factor compression). The procedure up to this point is equivalent to the construction of an orthogonal set of reference axes in a subspace of the original measurements which contains almost all of the original information. This reference system may be inappropriate for many reasons. Rotation of the reference axes allows one to relate the axes more closely to chemically-significant properties (factor transformation). As a final step, and to provide one more quantitative measure of success, one can regenerate the data.

Our study of the factor analysis of mass spectra has taken two directions (*23*). First, we investigated the many variants of FA to discover procedures appropriate for mass spectral data. Second, we applied the methods to a number of cases. We report only the methodological studies here.

DATA TRANSFORMATION

The data matrix, X, is not analyzed directly. Generally, it is first transformed into a covariance or correlation matrix, C, and this matrix is analyzed. One may formulate the calculation of C from X as a two-step process. First, a linear transformation of X is carried out to generate Y, a matrix of transformed data, Equation 1.

$$Y = XA + B \tag{1}$$

Then one calculates the matrix of product moment coefficients, C, Equation 2.

$$C = Y'Y/n \tag{2}$$

Y' is the transpose of the Y matrix. The unsubscripted C is used here and subsequently in equations which refer to *both* the correlation matrix and to the covariance matrix.

Four variants of the linear transformation (Equation 1) occur commonly, and they produce four different C matrices, namely, correlation about the mean, R_m, correlation about the origin, R_o, covariance about the mean, C_m, and covariance about the origin, C_o. They differ in their definition of A and B in Equation 1. Since A is a diagonal matrix, only the diagonal elements, a_{jj}, need be defined. All the elements in any one column of B are identical.

Correlation about the mean, R_m:

$$a_{jj} = \left(\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_j)^2\right)^{-1/2}$$
$$b_{ij} = \bar{x}_j a_{jj} \tag{3}$$

Correlation about the origin, R_o:

$$a_{jj} = \left(\frac{1}{n}\sum_{j=1}^{n} x_{ij}^2\right)^{-1/2}$$
$$b_{ij} = 0 \tag{4}$$

Covariance about the mean, C_m:

$$a_{jj} = 1$$
$$b_{ij} = -\bar{x}_j \tag{5}$$

Covariance about the origin, C_o:

$$a_{jj} = 1$$
$$b_{ij} = 0 \tag{6}$$

$\bar{x}_j$ is the average of the jth column of the data matrix. The n in the denominator of Equations 2 and 3 becomes $n - 1$ when one calculates R_m and C_m using an unbiased estimator.

The matrix A is a scaling matrix. In R_o and R_m, it normalizes the length of the variable (column) vector to the square root of the number of cases (rows). In these two instances, each transformed variable is assigned equal weight in the subsequent calculations. The magnitude of the original data is destroyed, only the variable pattern is preserved. The B matrix is a centering matrix which adjusts the origin of the space to the variable means in R_m and C_m. In these two instances, no account is taken of variation about the original scale origin, but only of deviations from the variable mean.

The importance of these four different data transformations becomes clear when we consider the peculiarities of

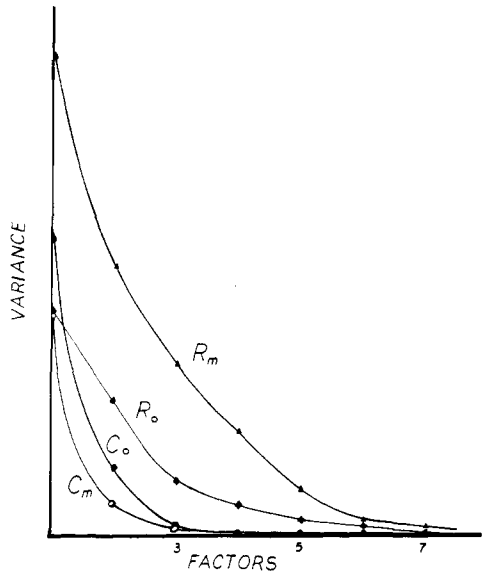

Figure 2. Cumulative percentage of the total variance accounted for by a given number of factors. Principal component analysis (*R*) of absolute intensities at twenty masses for the twenty-two isomers of $C_{10}H_{14}$

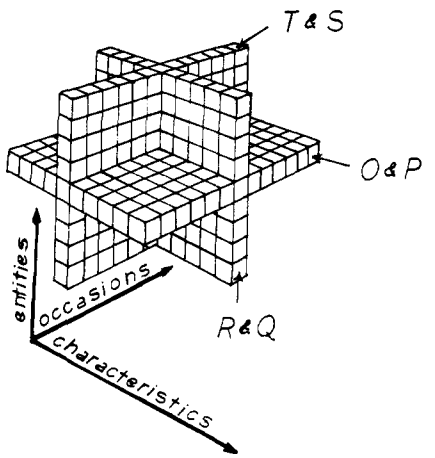

Figure 3. The six different kinds of two-dimensional slices through the three-dimensional data cube

mass spectral data. The intensities reported in a mass spectrum have a real zero, like the Kelvin and unlike the Fahrenheit temperature scale. The intensities at all masses for the same compound and, in principle, the intensities at all masses for all compounds, are reported in the same units. The intensities of mass spectra can be recorded on the *same ratio scale*. These characteristics of mass spectral data are highly unusual in FA and they drastically affect the results of the analysis. One can see this in Figure 2. The experimental data for benzene derivatives with the formula $C_{10}H_{14}$ used in this example, will be discussed later. The cumulative percentage of the total variance of the data accounted for, is plotted vs. the number of factors. A necessary (but not sufficient) condition that two analyses produce the same result, is the coincidence of the curves for the different types of analysis. They differ markedly. This is not surprising, since R_m and C_m have lost the information about the zero point of the experimental scale and R_m and R_o have lost the information about the comparable size of intensities at different masses for the same compound. Only C_o, covariance about the origin, retains the zero and scale information characteristic of mass spectral measurements (*9*).

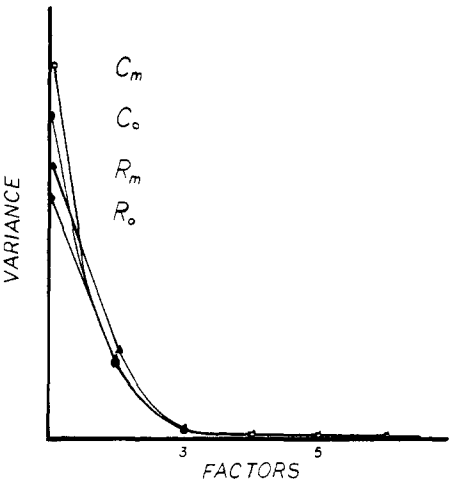

Figure 4. Cumulative percentage of the total variance accounted for by a given number of factors. Principal component analysis (*Q*) of absolute intensities at twenty masses for the twenty-two isomers of $C_{10}H_{14}$

Table I. The Six Different Kinds of Two-dimensional Slices through the Three-dimensional Data Cube

Type	Case or row	Variable or column	Constant
R	entities	characteristics	occasion
Q	characteristics	entities	occasion
O	characteristics	occasions	entity
P	occasions	characteristics	entity
T	entities	occasions	characteristic
S	occasions	entities	characteristic

DATA TRANSPOSITION

R_m, R_o, C_m, and C_o are not the only options available for data transformation. In general, measurements occur in a three-dimensional array, the data cube. The axes of the cube represent entities (e.g., compounds), characteristics (e.g., intensities at a mass), and occasions (e.g., times) (*9*). Mass spectra are not reported as time-dependent, so only two of the six different kinds of slices through the data cube interest us, *R* and *Q* analysis (Table I, Figure 3). Figure 2 reported an *R* analysis; Figure 4 presents the corresponding *Q* analysis (analysis of the transpose of the original data matrix). The *R* and *Q* analysis results are quite different. The results for the four different data transformations of *R* analysis (Equations 3–6) differ considerably from each other (the first four columns of Table II). The results for the four different data transformation methods of the *Q* analysis are quite similar to each other (the final four columns of Table II). The evidence from the cumulative percentage variance accounted for presented in Table II is corroborated by the angles between the various resulting eigenvector solutions. As we mentioned previously, the eigenanalysis is equivalent to the construction of an orthogonal set of reference axes in a subspace of the original measurements. The different kinds of data transformations produce coordinate systems oriented differently in space. One can describe the similarity of two solutions by calculating the angles through which it is necessary to rotate the one coordinate system into the other by a rigid rotation. The first four columns of Table III present the angles in degrees needed to rotate the solutions for the four methods of data transformation into the C_o solution for each of the factors of a three-factor solution. The final four columns of the Table list the same angles for the *Q* analysis. The angles of the *Q* analysis are much smaller and more homoge-

Table II. Cumulative Percent of the Total Data Variance Accounted for by p Factors. Principal Component Analysis of the Absolute Intensities at Twenty Masses for the Twenty-two Isomers of $C_{10}H_{14}$

	R analysis				Q analysis			
P	C_o	C_m	R_o	R_m	C_o	C_m	R_o	R_m
1	60.1	72.8	68.7	36.8	60.1	54.3	70.1	66.1
2	90.9	96.3	82.1	64.4	90.9	90.2	90.5	89.5
3	99.3	98.6	93.2	76.8	99.3	99.1	99.1	98.9
4	99.7	99.4	96.2	86.4	99.7	99.7	99.7	99.6
5	99.9	99.8	98.1	93.8	99.9	99.9	99.9	99.9

Table III. Angles in Degrees through Which the Three Factors of the Principal Component Solution of the C_m, R_o, and R_m Data Transformations Must Be Rotated to Coincide with the Factors of the C_o Data Transformation Solution

	R analysis				Q analysis			
Factor	C_o	C_m	R_o	R_m	C_o	C_m	R_o	R_m
1	0.0	76.0	55.5	86.3	0.0	20.5	12.6	19.4
2	0.0	82.3	79.3	80.5	0.0	20.1	13.1	20.4
3	0.0	64.5	80.1	79.8	0.0	8.5	3.6	9.3

Table IV. Definitions of Six New Types of Factor Analysis

Type	Case or row	Variable or column	Constant
U	entities (a)	entities (b)	characteristic, occasion
V	entities (b)	entities (a)	characteristic, occasion
W	characteristics (a)	characteristics (b)	entity, occasion
X	characteristics (b)	characteristics (a)	entity, occasion
Y	occasions (a)	occasions (b)	entity, characteristic
Z	occasions (b)	occasions (a)	entity, characteristic

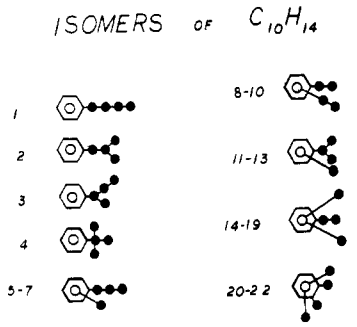

Figure 5. Structural formulas of the twenty-two alkyl benzenes with the formula $C_{10}H_{14}$

neous for the four different data transformation methods. A mass spectroscopist should not be surprised by the similarity of the results from the different methods of data transformation for Q analysis. In R analysis, R_o and R_m, one renormalizes the data along a mass (a data column); this destroys the information about the relative peak size within a compound. In Q analysis, one renormalizes the data along a compound (data row), preserving the fragmentation pattern within each mass spectrum. It is this pattern which represents the most significant information present in mass spectra. From the Figures and Tables, it is also evident that only C_o analysis is the same for both R and Q analysis. This is one more reason to prefer C_o analysis, covariance about the origin, for mass spectral data. Whether this preference will extend to other kinds of data depends on the presence of a meaningful zero and unit in the measurement scale, and on weighting considerations.

To avoid confusion, one should note that the six slices through the data matrix, R, Q, O, P, S, and T, do not represent all possible types of factor analysis. This seems to have escaped the notice of the standard works on FA (*9*). Table IV presents a terminology for representing the other variants which occur in the literature. U and V analysis have been applied to NMR chemical shifts (*13–15*), and GC retention times (*16–18*). When the mass spectra of many compounds produced by many different ionizing particles become available (chemiionization and ion cyclotron resonance), these types of analysis will be valuable to the mass spectroscopist. Y and Z analysis of meteorological data has been performed (*24*).

DATA SELECTION

The mass spectra of the benzene derivatives with the formula $C_{10}H_{14}$ (*25*) were chosen as the test case for these methodological studies. Deliberately this is a simple case. Only two elements are present, and all the compounds have the same molecular weight. The MS of benzene derivatives are more regular than those of the aliphatic hydrocarbons (*4*). There is a relatively small number of isomers, 22 (Figure 5). On the other hand, these compounds are sufficiently complex to be interesting. About 125 intensities are available for each. The ring isomers and side chain isomers provide variety. Interesting processes such as ring opening, ring expansion, and ion-neutral complexing are known to occur. The isomers have been investigated extensively by isotope labeling and by the study of metastable transitions and appearance potentials (*5*), so one can hope to calibrate the new method by known results.

Studies were carried out to determine how many of the measurements at different masses should be included in the analysis. The complete spectrum includes intensities from 12 to 136 amu. Q analysis of the complete spectrum (compounds as variables or columns) is feasible. The covariance matrix is 22 by 22; an eigenanalysis of this magnitude can be carried out even on relatively small computers. R

Table V. Rotation Angles (1-3) and Cumulative Percent Total Variance Accounted for by p Factors (1-5). Principal Component Solution, C_o Transformation for Twenty Masses (20 × 22), for All Masses Greater than 71 amu (65 × 22) and for Summed Spectra (9 × 22)

Factor/p	20 × 22	65 × 22	9 × 22
1	0.0	1.1	7.9
2	0.0	1.0	7.6
3	0.0	0.9	5.8
1	60.1	59.6	67.7
2	90.9	91.0	88.2
3	99.2	99.3	99.1
4	99.7	99.8	99.7
5	99.9	99.9	99.9

analysis, on the other hand, must work with a 125 by 125 covariance matrix. The C matrix has over 15 thousand elements. Even with a large computer, round-off errors may be considerable. Since one major purpose was to study the relationship of R and Q analysis, we attempted to generate a fairly square data matrix by selecting only some of the 125 masses. The ten largest intensities for all 22 isomers occur at one or other of the following 20 masses: 134, 133, 120, 119, 117, 106, 105, 103, 93, 92, 91, 79, 78, 77, 65, 51, 43, 41, 39, and 27. Unless explicitly contraindicated, the intensities of the 22 isomers at these 20 masses comprise the data analyzed. Another method to reduce the number of columns in the data matrix was also used; the intensities of ions having the same number of carbon atoms, but different numbers of hydrogen atoms were summed. This resulted in a 22 by 9 matrix (22 isomers and 9 different carbon groupings) which could be analyzed easily. A final method of selection was the Q analysis of all intensities with masses equal to or heavier than the C_6 ring, mass 72. The results of these different analyses are shown in Table V. The first column of the Table lists the results for C_o analysis of intensities at the twenty peaks described earlier. The second column is concerned with the analysis of all intensities at mass 72 or higher. The final column lists the results for the nine intensities produced by summation over all the ions with the same number of carbon atoms for the twenty-two isomers. The upper half of the Table is composed of the angles in degrees necessary to rotate the various solutions into the 20 by 22 analysis. The lower half of the Table contains the cumulative percentage variance accounted for by 1 to 5 factors. The analysis of the twenty large peaks and the analysis of all those peaks at mass 72 and heavier, produce very similar solutions. One might say naively that ring fragmentation has little influence on the factors or, more accurately, that the low weight ions introduce no new element into the dissociation pattern. Analysis of the summed spectra produces a solution quite similar to the other two.

We also studied the effect on the solution if one used relative intensities as data, or if one used absolute intensities. As we mentioned previously, in principle, all mass spectral measurements can be made in the same units. Such spectra can be called "absolute" mass spectra. Often the spectral intensities for each compound are normalized so that the measurements range from 0 to 100. Such measurements are called "relative" mass spectra. A calibration factor which standardized the sample pressure, electron current, and detector sensitivity was reported with our spectra, so we were able to analyze both the absolute and relative mass spectra of the 22 isomers (25). The results are shown in Tables II and VI. The factors are similar but not identical. Unless otherwise indicated, absolute measurements are used in all examples.

A summary of the conclusions of our studies of the effect of different kinds of data selection and transformation might be helpful. In general, the R and Q analysis of the four kinds of transformed data produces different results. The four solutions for Q analysis are much more similar to each other than the four solutions for R analysis with our data. R and Q analysis of untransformed data, C_o, produces the identical solution, if one ignores the different normalization conventions. That is, the eigenvectors (F of Equations 7, 8, 9, 10) of the R analysis become the scores (S of Equations 8 and 10) of the Q analysis, and vice versa. R analysis of the ordinary correlation matrix, R_m, the most usual form of FA, produces the most idiosyncratic solution because of the peculiarities of mass spectral data. With similar isomers, relative intensities produce factors similar to, but not identical with, those from absolute measurements. With sets of dissimilar compounds, R analysis of correlation coefficients calculated from relative intensities will produce meaningless results. The (row) normalization used to produce the relative intensities combined with the (column) normalization of the method of analysis multiplies each measurement by a different arbitrary number. The best technique uses R or Q analysis of untransformed (C_o) absolute intensities. This preserves the information present, and accounts for the major processes at work in the mass spectra.

FACTOR COMPRESSION

Data selection and transformation is followed by (principal) component analysis, PCA (Figure 1). PCA constructs a unique orthogonal coordinate system for the data such that the first axis accounts for most of the variance of the data—i.e., it is oriented along the direction on which the sum of the projections of the variable vectors is a maximum. The second reference axis is orthogonal to the first, and in the direction which maximizes the sum of the projections of the residual variance, and so forth. Mathematically, this procedure is equivalent to an eigenanalysis which generates m eigenvectors and m associated eigenvalues. Each eigenvector, a column of the factor loading matrix, F, is normalized to the square root of its eigenvalue, so that

Table VI. Cumulative Percent of the Total Variance of the Data Accounted for by p Factors. Principal Component Analysis of the Relative Intensities at Twenty Masses for the Twenty-Two Isomers of $C_{10}H_{14}$

	R analysis				Q analysis			
P	C_o	C_m	R_o	R_m	C_o	C_m	R_o	R_m
1	71.4	58.6	71.4	33.6	71.4	65.8	70.1	66.1
2	89.8	90.4	84.9	59.3	89.8	88.6	90.5	89.5
3	99.1	97.5	93.5	81.2	99.1	98.8	99.1	98.8
4	99.7	99.2	96.2	88.7	99.7	99.6	99.7	99.6
5	99.9	99.7	98.1	95.1	99.9	99.9	99.9	99.9

$$\Lambda = F'F \quad (7)$$

where Λ is a diagonal matrix composed of eigenvalues as the diagonal elements. Each element of the F matrix is called a "loading". In the R_m case, each element of the factor loading matrix, f_{ij}, is the correlation coefficient between variable i and factor j, and the cosine of the angle between data variable vector **i** and the factor vector **j**. Each eigenvalue, the diagonal elements of the diagonal matrix, Λ, is a measure of the variance accounted for by the corresponding eigenvector. The factor score matrix, S, is calculated from the transformed data, Y, from the loading matrix, F, and from the eigenvalue matrix, Λ (Equation 8).

$$S = YF\Lambda^{-1} \quad (8)$$

Each column of S is normalized to the square root of the number of cases, n. Each element of the factor score matrix, s_{ij}, is called a "score". It is a measure of the importance of the jth eigenvector in the variability of the ith case (row). Two important properties of the factor loadings and factor scores are represented by recalculation of the covariance matrix, Equation 9, and recalculation of the data matrix, Equation 10.

$$C = FF' \quad (9)$$

$$Y = SF' \quad (10)$$

For those familiar with factor analysis, we might note that we are not performing *common* factor analysis (*9*). The diagonal elements of the C matrix are unchanged. The importance of component analysis, just described, becomes clear when we discuss factor compression.

Mass spectra are highly redundant. One can effectively reduce the number of measurements in a group of spectra if one can determine the number of independent variables at work in the data. Mathematically equivalent ways of expressing this goal are many: how many independent rows or columns are present in the data, what is the rank of the data matrix, what is the dimension of the subspace needed to express practically all of the significant information present in the data? Two sets of techniques which allow one to reach these goals will be discussed.

The first set of techniques is based on the size of the eigenvalue of the eigenvector produced by the principal component analysis (*9*). The first method accepts all those dimensions with eigenvalues greater than the average eigenvalue. In both R_m and R_o analysis, the average eigenvalue is 1. A second method accepts all those eigenvectors with eigenvalues greater than the variance expected from the random error present in the measurements. For example, a 1% increase in the mass spectral intensities increases the total variance of the data by 101 (R analysis, C_o data transformation). Eigenvalues are a measure of the variance accounted for by the corresponding eigenvector. Eigenvalues smaller than 101 can therefore be neglected. A third technique defines some reasonable highest cumulative percent of the total variance of the data which must be accounted for by the factors. When this cumulative percent variance is accounted for by a certain number of factors, the subsequent factors are neglected. A graphical procedure related to this is the scree plot (Figure 2) (*9*). The cumulative percentage variance accounted for is plotted vertically (origin is 100%) against the (horizontal) number of factors used. A discontinuity in the scree plot (Figure 2) can be taken as indication of a threshold eigenvalue ($\lambda p \gg \lambda p + 1$). Finally, if one extrapolates back the tail of a scree plot (presumably due only to random variation) to the sharply increasing portion of the plot, one obtains an estimate of the eigenvalue which excludes random variations, λ_∞. With the

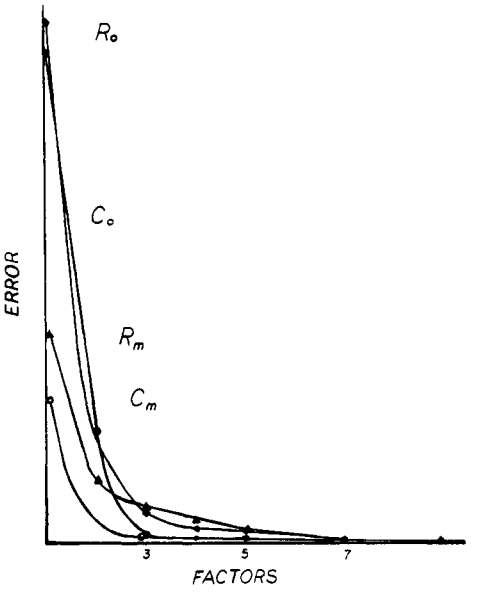

Figure 6. Mean square deviation of the recalculated data from the original measurements as a function of the number of factors. R analysis, principal component orientation

Table VII. Number of Factors Accepted According to the Criterion of the Average Eigenvalue (λ), and the Eigenvalue Which Reaches a Cumulative Percent Variance Accounted for of 99%

	R analysis				Q analysis			
	C_o	C_m	R_o	R_m	C_o	C_m	R_o	R_m
$\bar{\lambda}$	3	3	3	5	3	3	3	3
$\lambda(99\%)$	3	4	6	8	3	3	3	3

$C_{10}H_{14}$ test data, C_o transformation, all five criteria agree that three factors are present. Other transformations lead to a larger number of factors (Table VII). These results are coherent with what we said previously about different results from the different forms of data transformation.

The second important set of techniques which are useful for choosing the number of independent variables at work in a set of mass spectra, are based on the deviation of the recalculated data from the original data matrix (Equation 10) (*14*). First of all, one can use a single number, the mean square deviation, of the two sets of data. Second, one can use the error (mean square) for each column (compound), or for each row (mass). The n plus m criteria can pinpoint compounds or masses not adequately represented by the factors. Finally, the deviation of the recalculated data (Equation 10) from the original measurement provides n times m measures of fit which can pinpoint troublesome data points. These complementary measures of fit can be used in several ways. For example, the global measure of fit may be plotted as a function of the number of factors, and the graph interpreted like a scree plot. The error of an analysis varies with the number of factors used and with the type of analysis (Figure 6). R and Q analysis, C_o transformation, however, have the same error measures. If one plots mean square deviations, all the data deviation methods agree with the eigenvalue methods.

FACTOR TRANSFORMATION

Principal component analysis helps one define the minimum dimensionality of the data, but the unique orientation of the eigenvectors is based upon mathematical, not chemical properties (*9, 14*). The eigenvectors may be linear

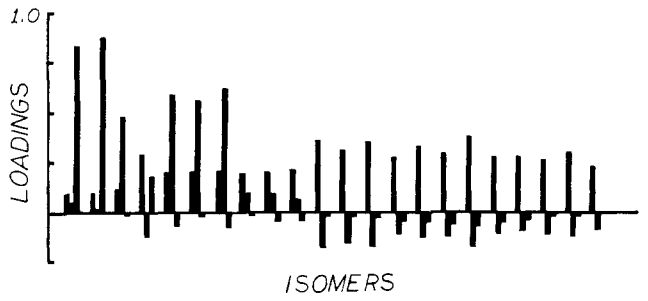

Figure 7. Three values for each of the twenty-two isomers of $C_{10}H_{14}$ in turn (cf. Figure 5). Loadings from *Q* analysis, scores from *R* analysis, principal component orientation, covariance about the origin. Each factor is normalized to 1

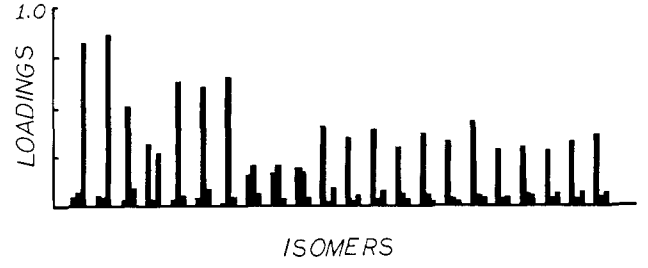

Figure 8. Three values for each of the twenty-two isomers of $C_{10}H_{14}$ in turn (cf. Figure 5). Loadings from *Q* analysis, varimax rotation, covariance about the origin. Each factor is normalized to 1

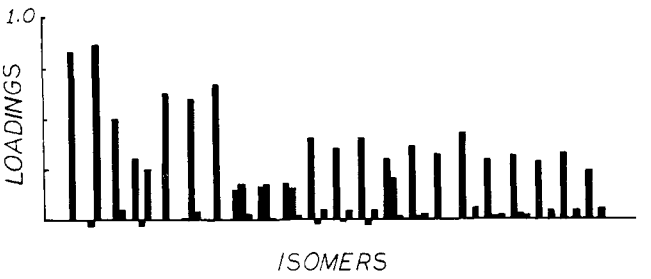

Figure 9. Three values for each of the twenty-two isomers of $C_{10}H_{14}$ (Figure 5) in turn. Loadings from *Q* analysis, target rotation, covariance about the origin. Suggested scores were the intensities at twenty masses for isomers 16, 5, and 1 in that order

Table VIII. Angle in Degrees through Which Each of the Three Factors Must be Rotated to Coincide for Principal Component Analysis (PC), Varimax (VM) and Quartimax (QM) Rotations

	R analysis			*Q* analysis		
Factor	PC–VM	PC–QM	VM–QM	PC–VM	PC–QM	VM–QM
1	18.9	2.0	20.6	17.5	12.9	4.6
2	32.5	3.4	34.6	30.7	23.7	6.9
3	25.5	14.0	36.2	21.1	13.1	8.9

combinations of important chemical properties, rather than the important properties themselves. The scores or coordinates of an isomer in the coordinate system of the eigenvectors may be both positive and negative (Figure 7), while the original mass spectra were all positive. Finally, one may wish to determine whether a known property of the compounds is responsible for the mass spectra. Positive scores, correlation with known properties, and hypothesis-testing demand rotation of the coordinate system. Rotation or transformation of the reference axes may be oblique or orthogonal, analytical or empirical (*9*). Orthogonal rotations based on some analytical criterion (varimax, quartimax), or according to some empirically-defined criterion (target rotation) will be discussed in turn.

Analytical rotation simplifies the results of the analysis according to some definable rule. This may be motivated by the hope that the simplest solution will be the best, or perhaps only a best first step. One may have the more concrete goal of finding uniformly positive loadings and scores. Two common analytical techniques of factor transformation are varimax and quartimax rotation (*9*). The varimax method, VM, defines the equations necessary to simplify the relations of the ith factor with the m original variables (masses). This makes interpretation of the factor simpler, since the significance of the original variables is presumably clear. VM rotates the factor so that it will depend strongly on some variables and weakly or not at all on others. It tries to eliminate the partial similarity of a factor with many variables in favor of strong similarity with some, and strong dissimilarity with the others. In terms of angles, VM rotates the factor to make it as parallel or perpendicular to as many of the variable vectors as possible. A comparison of the positive and negative loadings of the PC orientation (Figure 7) and the positive loadings of the VM solution (Figure 8) is instructive. Quartimax rotation sets up the criterion necessary to simplify the relationship of the ith measured variable with the p factors. The factors are rotated so that the ith variable will depend strongly on some factors, and weakly or not at all on others. The factors are rotated to make them as parallel or perpendicular to the variable in question as possible. They are not strikingly different from the varimax results with our data. Table VIII lists the angles between the three orientations.

Target rotation is a more empirical method of reference axis rotation (*9, 14*). If one specifies the coordinates or scores of the cases, then the reference axes can be rotated so that a least squares approximation to the suggested scores is generated. If the scores are one of the original variables, the procedure is equivalent to a regression analysis. If the scores are other known properties of the compounds, one can test the hypothesis that the property in question is correlated with the mass spectrum. Since only p values of the scores are needed to define the vector, some of the $n - p$ values of the property may be predicted by the procedure (*14*). A quantitative measure of the fit of the suggested set of scores to the scores on the rotated factors is produced and a second measure of fit results from the regeneration of the data. Figure 9 shows the results of using the measurements of isomers 1, 5, and 16 of Figure 5 as hypothetical scores in a *Q* analysis. These isomers provided the best fit when tested together as factors of a three-dimensional solution. Figure 10 shows the three experimental intensities at masses 119, 105, and 91 for each of the twenty-two isomers in turn. Other more chemically significant examples of target rotation will be left for a subsequent article. It is clear from the examples that varimax and target rotation are both valuable aids to the interpretation of mass spectra by factor analytic methods.

We end our discussion of data selection and transformation, factor compression and transformation, with a few remarks on the computer programs which were used. Of greatest importance is BMDX72 from the Berkeley Biomedical software collection (*26*). We adapted it to recalculate the data and calculate the error (mean square deviation) for each case and each variable, as well as the total error. Professor Malinowski's program was helpful for target rotation and hypothesis testing (*14*). We adapted it to perform all four types of data transformation from the original R_o, and to calculate the mean square deviation for each case, variable and the over-all measure of fit. It is impor-

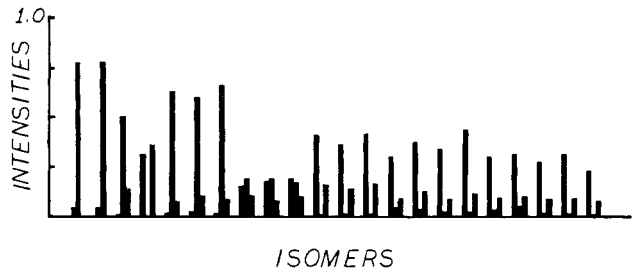

Figure 10. Three experimental intensities for each of the twenty-two isomers of $C_{10}H_{14}$ in turn (Figure 5). The masses are 119, 105, and 91 amu in that order. The intensities at each mass are normalized to 1

tant to realize that normalization conventions differ in different programs. We have consistently used those of BMDX72. The terminology loadings and scores, rather than property factor matrix and molecular factor matrix also corresponds to BMDX72. More details are available concerning terminology, conventions, and computational procedures (*27*).

ACKNOWLEDGMENT

We thank William Lawlor for his help, and recognize the preliminary work of Vincent Waldron (Honors Thesis, Fordham University, 1969).

LITERATURE CITED

(1) K. Biemann, "Mass Spectrometry—Organic Chemical Applications", McGraw-Hill, New York, 1962.
(2) J. H. Benyon, "Mass Spectrometry and Its Applications to Organic Chemistry", Elsevier, Amsterdam, 1960.
(3) F. W. McLafferty, ed., "Mass Spectrometry of Organic Ions," Academic Press, New York, 1963.
(4) H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Mass Spectrometry of Organic Compounds", Holden-Day, San Francisco, 1967.
(5) H. M. Grubb and S. Meyerson, "Mass Spectrometry of Organic Ions", F. W. McLafferty, Ed., Academic Press, New York, 1963.
(6) L. L. Thurstone, "Multiple Factor Analysis", University of Chicago Press, Chicago, 1947.
(7) H. H. Harman, "Modern Factor Analysis", University of Chicago Press, Chicago, 1967.
(8) P. Horst, "Factor Analysis of Data Matrices", Holt, Reinhart and Winston, New York, 1965.
(9) R. J. Rummel, "Applied Factor Analysis", Northwestern University Press, Evanston, 1970.
(10) E. R. Malinowski, Doctoral Dissertation, Stevens Institute of Technology, Hoboken, NJ, 1961.
(11) P. H. Weiner, Doctoral Dissertation, Stevens Institute of Technology, Hoboken, NJ, 1971.
(12) J. T. Bulmer and H. F. Shurvell, *J. Phys. Chem.*, **77**, 256 (1973).
(13) E. R. Malinowski and P. H. Weiner, *J. Am. Chem. Soc.*, **92**, 4193 (1970).
(14) P. H. Weiner, E. R. Malinowski, and A. Levinstone, *J. Phys. Chem.*, **74**, 4537 (1970).
(15) E. R. Malinowski and P. H. Weiner, *J. Phys. Chem.*, **75**, 1207 (1971).
(16) P. T. Funke, E. R. Malinowski, E. E. Martire, and L. Z. Pollara, *Sep. Sci.*, **1**, 661 (1966).
(17) P. H. Weiner and D. G. Howery, *Anal. Chem.*, **44**, 1189 (1972).
(18) P. H. Weiner and J. F. Parcher, *Anal. Chem.*, **45**, 302 (1973).
(19) D. Macnaughton, Jr., L. B. Rogers, and G. Wernimont, *Anal. Chem.*, **44**, 1421 (1972).
(20) J. J. Kankare, *Anal. Chem.*, **42**, 1322 (1970).
(21) N. Ohta, *Anal. Chem.*, **45**, 553 (1973).
(22) J. C. Stover, Doctoral Dissertation, Fordham University, New York, 1974.
(23) R. W. Rozett and E. McLaughlin, Twenty-Second Annual Conference on Mass Spectrometry, Philadelphia, PA, May 19–24, 1974, p 441.
(24) John Jalickee, private communication.
(25) Mass Spectral Data, API Project 44, Nos. 210, 212, 214, 439–441, 459–463, 486, 494, 863, 934, 1184, 1429, 1431, 1432, 1570, 1655, 1957.
(26) W. J. Dixon, Ed., BMD, Biomedical Computer Programs, X Series Supplement, University of California Press, Berkeley, 1970, p 90.
(27) E. McLaughlin Petersen, Doctoral Dissertation, Fordham University, New York, 1975.

RECEIVED for review November 5, 1974. Accepted March 12, 1975.

Atmospheric Pressure Ionization (API) Mass Spectrometry: Formation of Phenoxide Ions from Chlorinated Aromatic Compounds

Ismet Dzidic, D. I. Carroll, R. N. Stillwell, and E. C. Horning

Institute for Lipid Research, Baylor College of Medicine, Houston, TX 77025

Phenoxide ions, $(M - Cl + O)^-$, are formed by ion-molecule reactions: $M^- + O_2 \rightarrow (M - Cl + O)^- + OCl$, and $O_2^- + M \rightarrow (M - Cl + O)^- + OCl$, when certain chlorinated aromatic compounds are ionized in an API source in the presence of nitrogen containing approximately 0.5 ppm of oxygen and also in air. While *o*- and *p*-chloronitrobenzenes form mainly chloride and nitrophenoxide ions, *m*-chloronitrobenzene ionizes under the same conditions to form a negative molecular ion. Chlorobenzene and *o*-dichlorobenzene yield only chloride ions, while more highly substituted polychlorobenzenes form phenoxide ions. Subpicogram detection of 2,3,4,5,6-pentachlorobiphenyl is demonstrated by selective monitoring of the corresponding phenoxide ion.

The usual method of analysis for residues of many insecticides, herbicides, fungicides, and polychlorobiphenyls is gas chromatography with electron capture (EC) detection. A wide variation in detector response with structural changes is a feature of EC detection of organic compounds.

The direct experimental observation of negative ion formation, under conditions approximating those of EC detection (atmospheric pressure, carrier gas environment), is not possible with conventional mass spectrometers. The development of the "Plasma Chromatograph" (*1*) provided an opportunity for the measurement of mobilities of negative ions formed in a carrier gas under conditions of temperature and pressure similar to those used in EC detection. When mobilities of negative ions were measured for several polychlorobiphenyls (*2*), and for isomeric chloronitrobenzenes (*3*), it was found that some compounds formed several negative ions. These ions were assumed to correspond to M^-, Cl^-, and $(M - Cl)^-$ (*2, 3*). The nature of these ions was not, in fact, firmly established, since mobilities cannot be used for ion identification (*4*).