1

Every problem has a solution.

INTRODUCTION

The use of computers to solve chemical problems has increased tremendously during the past 20 years. Indeed, the discipline of chemistry called chemometrics is flourishing as a result of this revolutionary marriage of chemistry and computer science. Chemometrics is the use of mathematical and statistical methods for handling, interpreting, and predicting chemical data. Powerful methodologies have opened new vistas for chemists and have provided useful solutions for many complex chemical problems.[1–29] Factor analysis has proved to be one of the most potent techniques in the chemometric arsenal.

Chapters 1 and 2 provide a simplified, qualitative introduction to factor analysis. These two chapters furnish an overview of factor analysis, written especially for readers who do not wish to wrestle with mathematical details. Chapter 1 explores the purposes and advantages of factor analysis, while avoiding detailed methodology and mathematical derivations. For the mathematically inclined reader and for all who wish to acquire an in-depth understanding of factor analysis, a rigorous development is presented in Chapters 3 and 4.

1.1 DEFINITION OF FACTOR ANALYSIS

Factor analysis (FA) was founded by the behavioral scientists. Its early history has been recorded by Harman,[30] who ascribes its origin to a paper on orthogonal regression lines published by Pearson[31] in 1901. The first real development was

accomplished by Hotelling[32] in 1933. Although an ideal tool for solving chemical problems, the method went unnoticed by the chemical profession until the birth of chemometrics in the 1970s.

In the early years factor analysis suffered a tarnished reputation because many variations of the method did not yield the same results. This was due to the different simplifying assumptions and auxiliary conditions needed during the precomputer era. Many of these ingenious methods have fallen by the wayside. During the years, the computer greatly influenced the development of factor analysis. Today, chemists, and scientists in general, are familiar with computers, mathematics, and statistics, the prerequisites for factor analysis. The field has grown so large that it is impossible for us to examine all the methodologies. In this text, we concentrate on those techniques that have become popular in chemistry.

If you ask for a definition of *factor analysis*, you will obtain a different answer from each practitioner. The definition has changed over the years, encompassing a much wider selection of techniques than originally intended. As an attempt to provide a global definition, we offer the following:

Factor analysis is a multivariate technique for reducing matrices of data to their lowest dimensionality by the use of orthogonal factor space and transformations that yield predictions and/or recognizable factors.

1.2 PHILOSOPHICAL BASIS

The nature and objectives of factor analysis can be illustrated with an example from academic life. As we know, the same laboratory report will receive different grades from different professors because of variations in marking criteria. The assigned grade represents a composite of a variety of factors (i.e., subjects), such as chemistry, physics, mathematics, grammar, and organization. Each factor is weighted in importance according to the personal judgment of each professor. The various grades a given report receives are due to the differences in importance conferred upon each factor by each professor.

In the technique of factor analysis, the grade is viewed as a linear sum of factors, each factor being weighted differently. Grade d_{ik} received by student i from professor k is assumed to have the form

$$d_{ik} = s_{i1}l_{1k} + s_{i2}l_{2k} + \cdots + s_{in}l_{nk} = \sum_{j=1}^{n} s_{ij}l_{jk} \tag{1.1}$$

where s_{ij} is the true *score* of student i in factor j, l_{jk} is the relative *loading* (importance) given by professor k to factor j, and the sum is taken over the n *factors* or subjects.

Factor analysis deals with a battery of grades involving a number of students and professors. The grades are arranged in a *matrix* such that each row concerns a particular student and each column concerns a particular professor. Such a data

matrix, involving four students and three professors, is shown in (1.2). If only two factors, such as chemistry and English, were considered important in the grading, each data point could be broken down into a sum of two factors, as shown in the first equality:

$$
\begin{array}{c} \\ \text{Students} \end{array}
\begin{array}{c} \text{Professors} \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \end{array}
\begin{array}{c} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \end{array}
\begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \end{bmatrix}
= \begin{bmatrix} s_{11}l_{11} + s_{12}l_{21} & s_{11}l_{12} + s_{12}l_{22} & s_{11}l_{13} + s_{12}l_{23} \\ s_{21}l_{11} + s_{22}l_{21} & s_{21}l_{12} + s_{22}l_{22} & s_{21}l_{13} + s_{22}l_{23} \\ s_{31}l_{11} + s_{32}l_{21} & s_{31}l_{12} + s_{32}l_{22} & s_{31}l_{13} + s_{32}l_{23} \\ s_{41}l_{11} + s_{42}l_{21} & s_{41}l_{12} + s_{42}l_{22} & s_{41}l_{13} + s_{42}l_{23} \end{bmatrix}
$$

$$
= \begin{array}{c} \\ \text{Students} \end{array}
\begin{array}{c} \text{Factors} \\ \begin{array}{cc} 1 & 2 \end{array} \end{array}
\begin{array}{c} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \end{array}
\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ s_{31} & s_{32} \\ s_{41} & s_{42} \end{bmatrix}
\begin{array}{c} \text{Professors} \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \end{array}
\begin{array}{c} \\ \begin{array}{c} 1 \\ 2 \end{array} \end{array}
\begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \end{bmatrix}
\begin{array}{c} \\ \text{Factors} \end{array}
\tag{1.2}
$$

The second equality in (1.2) is the result of applying standard rules of matrix multiplication. In matrix notation (1.2) becomes

$$\underset{\text{data}}{\mathbf{D}} = \underset{\text{scores}}{\mathbf{S}} \underset{\text{loadings}}{\mathbf{L}} \tag{1.3}$$

Here **D** is the *data matrix* that consists of the grades; **S** is the matrix of the students' true scores in each subject, called the *score* matrix; and **L** is the matrix of importance conferred upon each subject by each professor, called the *loading matrix*.

The purpose of factor analysis, as visualized by psychologists, is to extract the student score matrix from the data matrix in order to determine the students' true abilities in each subject, in effect removing the professors' prejudices (the loadings) from the grades. Such reasoning forms the philosophical basis of factor analysis.

The form of the data matrix discussed above is analogous to many types of data matrices encountered in chemistry, where, for example, molecules emulate the students and chemical measurements emulate the professors. Physical observations, such as boiling points, melting points, spectral intensities, and retention values, are data analogous to grades. In a chemical problem, a row of data may concern a particular molecule and a column may concern a particular measurement. Factor analysis yields a molecule score matrix, which depends solely on the characteristics of the molecules, and a measurement loading matrix, which depends solely on the nature of the measurements. Such a separation of the features of the molecules from the features of the measurements provides the chemist with a better insight into the true nature of the chemical phenomenon involved.

1.3 GENERALIZATIONS

This section concerns notation and terminology, which can be used in a general manner. Data matrix **D**, consisting of r rows and c columns, is written as

$$\mathbf{D} = \begin{array}{c} \\ \text{Row designee} \end{array} \overset{\text{Column designee}}{\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1c} \\ d_{21} & d_{22} & \cdots & d_{2c} \\ \vdots & \vdots & & \vdots \\ d_{r1} & d_{r2} & \cdots & d_{rc} \end{bmatrix}} \tag{1.4}$$

The row and column headings of the matrix are called *designees*. Each measured data point in **D** is specified by a subscript denoting its row and column position in the matrix. The symbol d_{ik} represents the data point associated with the ith row and kth column of the matrix.

Abstract Model. The first objective of factor analysis is to obtain a mathematical, "abstract" solution wherein each point in the data matrix is expressed as a linear sum of product terms. The number of terms in the sum, n, is called the *number of factors*. Specifically, we seek solutions of the form

$$d_{ik} = \sum_{j=1}^{n} r_{ij} c_{jk} \tag{1.5}$$

In this equation r_{ij} and c_{jk} are called *factors*. For the jth factor in the sum, *row factor* r_{ij} is associated with the ith row of the data matrix, and the corresponding *column factor* c_{jk} is associated with the kth column of the matrix. In classical abstract factor analysis, the row factors are called *scores* and the column factors are called *loadings*.

For data modeled by (1.5), the data matrix can be decomposed into two matrices:

$$\underset{\substack{\text{data} \\ \text{matrix}}}{\mathbf{D}} = \underset{\substack{\text{row} \\ \text{matrix}}}{\mathbf{R}_{\text{abstract}}} \underset{\substack{\text{column} \\ \text{matrix}}}{\mathbf{C}_{\text{abstract}}} \tag{1.6}$$

where

$$\mathbf{R}_{\text{abstract}} = \begin{matrix} \\ \text{Row designee} \end{matrix} \overset{\text{Factor}}{\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{r1} & r_{r2} & \cdots & r_{rn} \end{bmatrix}} \tag{1.7}$$

$$\mathbf{C}_{\text{abstract}} = \begin{matrix} \\ \text{Factor} \end{matrix} \overset{\text{Column designee}}{\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1c} \\ c_{21} & c_{22} & \cdots & c_{2c} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nc} \end{bmatrix}} \tag{1.8}$$

Since this solution is purely mathematical and is devoid of physical meaning, these matrices are called *abstract matrices*. The columns of $\mathbf{R}_{\text{abstract}}$ are called *abstract factors*. *Row matrix* $\mathbf{R}_{\text{abstract}}$ contains a row for each of the r row designees and a column for each of the n factors, while *column matrix* $\mathbf{C}_{\text{abstract}}$ has a column for each of the c column designees and a row for each factor. The factor analytical solution isolates the row-designee factors from the column-designee factors.

Methodologies for determining the number of factors and for calculating the abstract row and column matrices are discussed in Chapter 3. Since the abstract solution should involve a physically meaningful number of factors, determination of n, the correct factor "size," is a particularly important step. As a result of this step, an estimate of the complexity of the data space, information normally lacking even for the simplest chemical problems, is obtained.

Interpreting Factors. The ultimate objective of factor analysis is to develop a complete, physically meaningful model for the data. Hence the second objective of factor analysis is to convert the abstract solution into a real solution. To do this, the abstract factors are mathematically "transformed" into physically significant, "real" factors. Transforming the abstract solution into a real solution is a difficult, but realizable, goal of factor analysis.

To carry out the transformations, an appropriate information matrix, $\mathbf{T}$, is required. Postmultiplying $\mathbf{R}_{\text{abstract}}$ by $\mathbf{T}$ and premultiplying $\mathbf{C}_{\text{abstract}}$ by the inverse of the transformation $\mathbf{T}^{-1}$, the data matrix in (1.6) can be expressed as

$$\begin{aligned} \mathbf{D} &= \{\mathbf{R}_{\text{abstract}}\mathbf{T}\}\{\mathbf{T}^{-1}\mathbf{C}_{\text{abstract}}\} \\ &= \mathbf{R}_{\text{transformed}}\mathbf{C}_{\text{transformed}} \end{aligned} \tag{1.9}$$

If the transformed solution can be shown to have physical significance, a real solution to the problem will have been found so that

$$\mathbf{D} = \mathbf{X}_{\text{real}}\mathbf{Y}_{\text{real}} \tag{1.10}$$

where $\mathbf{X}_{\text{real}} = \mathbf{R}_{\text{transformed}}$ and $\mathbf{Y}_{\text{real}} = \mathbf{C}_{\text{transformed}}$. This equation summarizes the ultimate objective of factor analysis.

How such magical transformations can be carried out is one of the main topics of Chapters 2 through 6. Using factor analysis, we may be able to ascribe meanings to chemical data that initially appear to be exceedingly complicated because of the myriad of uncontrollable factors at play. The potential for modeling data with real factors is the most exciting feature of factor analysis.

A technique called *target testing* is especially valuable for achieving meaningful transformations. Suspected parameters (such as physical properties or structural features of molecules) can be tested individually as possible factors, and complete models of real factors can be systematically pieced together. This individual testing ability is one of the most valuable features of the target factor analysis method.

1.4 CHEMICAL EXAMPLE

To illustrate how factor analysis can be applied to chemical problems, let us consider the following data matrix, **A**, involving the ultraviolet absorbances of five different mixtures of the same absorbing components measured at six wavelengths:

$$\mathbf{A} = \begin{array}{c} \\ \text{Wavelength} \\ 278\text{ nm} \\ 274\text{ nm} \\ 270\text{ nm} \\ 266\text{ nm} \\ 262\text{ nm} \\ 258\text{ nm} \end{array} \begin{array}{c} \text{Mixture} \\ \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.005 & 0.031 & 0.063 & 0.091 & 0.046 \\ 0.040 & 0.172 & 0.356 & 0.444 & 0.218 \\ 0.103 & 0.283 & 0.484 & 0.471 & 0.208 \\ 0.116 & 0.323 & 0.562 & 0.548 & 0.241 \\ 0.125 & 0.318 & 0.516 & 0.450 & 0.185 \\ 0.104 & 0.267 & 0.430 & 0.376 & 0.154 \end{bmatrix} \end{array} \tag{1.11}$$

Such special information can be obtained from a liquid chromatograph where samples are collected at five different elution time intervals. Similar data can be collected from a chemical kinetics study if samples of the reaction mixture are collected at different times during the experiment. The problem here is to determine the number of components, to identify the chemical constituents, and to ascertain their concentrations.

According to (1.5), factor analysis will automatically furnish an abstract solution for each absorbance datum, A_{ik}, in the form

$$A_{ik} = \sum_{j=1}^{n} w_{ij} m_{jk} \tag{1.12}$$

Here w_{ij} and m_{jk} are the jth abstract row and column factors associated with the ith wavenumber and the kth mixture, respectively. To account for the absorbances within experimental error, n factors are included in the sum. According to (1.12), the absorbance data matrix has an abstract factor analytical solution expressed by

$$\mathbf{A} = \mathbf{W}_{\text{abstract}} \mathbf{M}_{\text{abstract}} \tag{1.13}$$

where $\mathbf{W}_{\text{abstract}}$ and $\mathbf{M}_{\text{abstract}}$ are wavenumber-factor and mixture-factor matrices, respectively.

The most important feature of the abstract solution is that it reveals the number of factors responsible for the absorbance data. Ultimately, we search for an appropriate transformation matrix that will convert the abstract solution into a physically significant real solution:

$$\mathbf{A} = \mathbf{W}_{\text{real}} \mathbf{M}_{\text{real}} \tag{1.14}$$

Going from (1.13) to (1.14) is not automatic. On the contrary, this step presents the most difficult challenge to the chemist, requiring a great deal of effort, knowledge, and intuition. If theoretical speculations can be invoked, the transformation has a better chance of being successful.

If the absorbance data obey Beer's law, the factors can be interpreted chemically. For a mixture containing n absorbing components, Beer's law models each absorbance datum by the equation

$$A_{ik} = \sum_{j=1}^{n} \varepsilon_{ij} c_{jk} \tag{1.15}$$

Here ε_{ij} is the molar absorptivity per unit path length of component j at wavelength i, and c_{jk} is the molar concentration of component j in the kth mixture. Equation (1.15) involves a linear sum of products analogous to (1.12); therefore, data that obey Beer's law should have meaningful factor analytical solutions. To solve the problem completely, we must find a transformation matrix that will convert the abstract solution into the real solution. When this is done correctly, (1.14) will take the form

$$\mathbf{A} = \mathbf{E}_{\text{real}} \mathbf{C}_{\text{real}} \tag{1.16}$$

Each column of the molar absorptivity matrix, $\mathbf{E}_{\text{real}}$, corresponds to the absorbances of one of the pure components at the five wavelengths, essentially tracing out the spectrum of the pure component. Each row of the molar concentration matrix, $\mathbf{C}_{\text{real}}$,

corresponds to the concentrations of one of the n components in each of the four mixtures.

Let us now summarize the kinds of information that might be furnished by factor analyses of absorbance data. First and quite important, determining the number of factors is tantamount to finding the total number of absorbing components in the mixtures. Second, astute transformations of the abstract factor analytical solution furnishes a good factor analytical representation of the real situation. The molar absorptivity matrix and the molar concentration matrix are the desired, physically significant transformations of $\mathbf{W}_{\text{abstract}}$ and $\mathbf{M}_{\text{abstract}}$. Successful transformation to $\mathbf{E}_{\text{real}}$ identifies each component chemically via its spectrum. The concomitant transformation to $\mathbf{C}_{\text{real}}$ furnishes the concentrations of the components in each sample mixture.

In summary, the ultimate payoff from factor analysis in this type of problem might be to determine:

1. The number of absorbing components
2. The concentration of each component in each mixture
3. The spectrum of each component

The factor analytical approach is far more useful than the popular determinant method for finding the concentrations of components in multicomponent mixtures, since the spectra of all components must be specified initially in the latter approach. By contrast, factor analysis can furnish the number of components, the concentrations, and the spectral information via a purely mathematical route. Describing exactly how factor analysis can accomplish these and even other, more difficult tasks is the primary objective of this book.

1.5 ATTRIBUTES

Factor analysis often allows us to answer the most fundamental questions in a chemical problem: How many factors influence the observable? What are the natures of these factors in terms of physically significant parameters? Factor analysis enables chemists to tackle problems that in the past had to be avoided because too many uncontrollable variables influenced the data. Factor analysis not only enables us to correlate and explain data, but also to fill in gaps in our data store.

In this section we list some of the virtues of the factor analytical approach. In particular, the following five general attributes illustrate why a chemist might want to use factor analysis:

1. *Data of great complexity can be investigated*. Factor analysis, being a method of "multivariate" analysis, can deal with many factors simultaneously. This feature is of special importance in chemistry, since interpretations of most chemical data require multivariate approaches.

2. *Large quantities of data can be analyzed*. Factor analyses can be carried out efficiently using standard factor analytical computer programs. Methods such as factor analysis are needed to properly use the voluminous data sets of chemistry.
3. *Many types of problems can be studied*. Factors analysis can be applied regardless of the initial lack of insight into the data. Although, ideally, factor analysis is used in conjunction with theoretical constructs, the approach can yield valuable predictions based on empirical applications.
4. *Data can be simplified*. Matrices can be modeled concisely with a minimum of factors, and generalizations that bring out the underlying order in the data can be obtained. Huge volumes of data can be compressed into small packages without loss of accuracy.
5. *Factors can be interpreted in useful ways*. The nature of the factors can be clarified and deciphered, and data can be classified into specific categories. Complete physically significant models can be developed systematically, and these models can be employed to predict new data.

In general, factor analysis provides a means to attack those problems that appear to be too difficult to solve. Such problems are bountiful in chemistry, making factor analysis an ideal probe for exploration. Finding the controlling factors is akin to an engineer drilling for oil. To increase the chances for success, the engineer must use every scientific resource available; blind drilling can be extremely expensive, time-consuming, and fruitless. A great deal of scientific input and intuition are required in the factor analytical approach.

Thousands of publications bear witness to the power and utility of factor analysis in chemistry. Howery,[14,15] Weiner,[16] and Llinas and Ruiz[17] have reviewed the role of factor analysis in chemistry during the early years. Its importance in mixture analysis has been reviewed by Gemperline[18] and by Hamilton and Gemperline.[19] A tutorial on target transformation factor analysis has been written by Hopke.[20] Bro[21] has published an exposition of multiway analysis, including parallel factor analysis (PARAFAC). Multiway analysis has been reviewed by Bro et al.[22] Geladi[23] has compiled the history of partial least squares (PLS). Classical PLS methodology is explained in the studies of Geladi and Kowalski.[24] Various chemical applications are described in Chapters 6 through 12.

REFERENCES

1. B. R. Kowalski (Ed.), *Chemometrics: Theory and Applications*, ACS Symp. Ser., 52, American Chemical Society, Washington, DC, 1977.
2. R. F. Hirsch (Ed.), *Statistics*, Franklin Institute Press, Philadelphia, 1978.
3. B. R. Kowalski (Ed.), *Chemometrics and Statistics in Chemistry*, Reidel, Dordrecht, Holland, 1983.
4. M. A. Sharaf, D. L. Illman, and B. R. Kowalski, *Chemometrics*, Wiley-Interscience, New York, 1986.

5. G. L. McClure (Ed.), *Computerized Quantitative Infrared Analysis*, American Society for Testing and Materials, Philadelphia, 1987.
6. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman, *Chemometrics: A Textbook*, Elsevier Science, Amsterdam, 1988.
7. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, New York, 1989.
8. J. E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.
9. R. Q. Yu, *Introduction to Chemometrics*, Hunan Educational Press House, Changsha, 1991.
10. Z. X. Pang, S. Z. Si, S. Z. Nie, and M. Z. Zhang, *Chemical Factor Analysis*, Publishing House of Science and Technology University of China, Hehui, 1992.
11. J. H. Kalivas and P. M. Lang, *Mathematical Analysis of Spectral Orthogonality*, Dekker, New York, 1994.
12. E. J. Karjalainen and U. P. Karjalainen, *Data Analysis for Hyphenated Techniques*, Elsevier Science, Amsterdam, 1996.
13. R. Kramer, *Chemometric Techniques for Quantitative Analysis*, Marcel Dekker, New York, 1998.
14. D. G. Howery, *Am. Lab.*, **8**(2), 14 (1976).
15. D. G. Howery, in R. F. Hirsch (Ed.), *Statistics*, Franklin Institute Press, Philadelphia, 1978, p. 185.
16. P. H. Weiner, *Chem Tech.*, **1977**, 321.
17. J. R. Llinas and J. M. Ruiz, in G. Vernin and Chanon (Eds.), *Computer Aids to Chemistry*, Wiley, New York, 1986, Chap. V.
18. P. J. Gemperline, *J. Chemometrics*, **3**, 549 (1989).
19. J. C. Hamilton and P. J. Gemperline, *J. Chemometrics*, **4**, 1 (1990).
20. P. K. Hopke, *Chemometrics Intell. Lab. Syst.*, **6**, 7 (1989).
21. R. Bro, Multi-Way Analysis in the Food Industry: Models, Algorithms and Applications, Doctoral Thesis, Royal Veterinary and Agricultural University, Denmark, 1998.
22. R. Bro, J. J. Workman, Jr., P. R. Mobley, and B. R. Kowalski, *Appl. Spectrosc. Rev.*, **32**, 237–261 (1997).
23. P. Geladi, *J. Chemometrics*, **2**, 231 (1988).
24. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1 (1986); **185**, 18 (1986).
25. O. Matthias, *Chemometrics*, Wiley-VCH, New York, 1999.
26. D. Livingstone, *Data Analysis for Chemists: Applications to QSAR and Chemical Products Design*, Oxford University Press, Oxford, 1995.
27. K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide*, Wiley-Interscience, New York, 1998.
28. E. K. Kemsley, *Discriminant Analysis and Class Modeling of Spectroscopic DataI*, Wiley, Chichester, 1998.
29. A. Höskuldsson, *Prediction Methods in Science and Technology*, Thor Publishing, Arnegaards Alle 7, Copenhagen, Denmark, 1998.
30. H. H. Harmon, *Modern Factor Analysis*, 2nd ed., University of Chicago Press, Chicago, 1967.
31. K. Pearson, *Philos. Mag.*, Series 6, **2**, 559 (1901).
32. H. Hotelling, *J. Educ. Psych.*, **24**, 417 (1933).