Volume 93, Number 3, May-June 1988
Journal of Research of the National Bureau of Standards

Evolutionary Factor Analysis

K. Schostack, P. Parekh, S. Patel, and E. R. Malinowski

Department of Chemistry and Chemical Engineering
Stevens Institute of Technology
Castle Point, Hoboken, NJ 07030

Because of chemical interconversion, many chemical systems cannot be physically separated, making chemical identification and quantification difficult. The spectra (IR, UV, Visible, Raman, CD, etc.) of such systems exhibit overlapping contributions of uncataloged components, confounding the identification as well as the quantification. Strategies based on factor analysis [1], a chemometric technique for handling complex multi-dimensional problems, are ideally suited to such problems. Abstract factor analysis (AFA) reveals the number of spectroscopically visible components. Evolutionary factor analysis (EFA) [2–4] takes advantage of experimental variables that control the evolution of components, revealing not only the concentration profiles of the components but also their spectra even when there are no unique concentrations or spectral regions.

Evolutionary factor analysis makes use of the fact that each species has a single, unique maximum in its evolutionary concentration distribution curve. We have recently applied this self-modeling method to the infrared spectra of stearyl alcohol in carbon tetrachloride solution. The evolutionary process of this system was achieved by increasing the concentration of stearyl alcohol from 0.0090 to 0.0800 g/L in 15 stages, each time recording the IR spectra from 3206 to 3826 cm^{-1}. The spectra were corrected for baseline shift, solvent absorption and reflectance losses. The 15 spectra were then digitized every 3 cm^{-1} and assembled into a 35×15 absorbance matrix $[A]$ appropriate for factor analysis.

The factor indicator function [1], the reduced eigenvalue [5] and cross validation [6] indicated that three species contribute to the observed spectra. Thus AFA expresses the data matrix as a product of a 35×3 absorptivity matrix $[E]_{abst}$ and 3×15 abstract concentration matrix $[C]_{abst}$.

$$[A]=[E]_{abst}\,[C]_{abst}$$

Because the abstract matrices are mathematical solutions devoid of chemical meaning, they must be transformed into physically meaningful absorptivities and concentrations. This is accomplished by target transformation factor analysis (TFA) [1], a powerful technique which allows one to test factors *individually* without requiring any *a priori* information concerning the other factors. A test vector C_{test} emulating an evolutionary profile can be converted into a predicted vector C_{pred} that lies completely inside the factor space by finding a transformation vector T that minimizes the sum of squares of the difference between C_{test} and C_{pred}. Accordingly, C_{pred} is given by:

$$C_{pred}=T[C]_{abst}$$

in which $T=C_{test}\,[C]'_{abst}\,\{[C]_{abst}\,[C]'_{abst}\}^{-1}$

The prime indicates matrix transposition.

These equations were used to target test 15 ideal (Dirac delta function) concentration profiles, represented by uniqueness tests for each column of the absorbance matrix. A uniqueness vector consists of zeros for all elements except the one in question, which contains unity:

$$\begin{aligned}C_1&=(1,0,0,\ldots,0,0,0)\\C_2&=(0,1,0,\ldots,0,0,0)\\&\cdots\\C_{15}&=(0,0,0,\ldots,0,0,1)\end{aligned}$$

The three predicted vectors with maxima corresponding to the unique point of the respective test vector were retained as likely candidates. These crude profiles were refined, individually, by applying simplex optimization to the respective transformation vector, using a response function designed to minimize negative concentration points and double maxima in the profile.

Further refinement was achieved by the following iteration. Because negative regions are meaningless, all data beyond the boundaries marked by the first negative regions encountered on the left and on the right of the peak maximum were truncated. These profiles were normalized so the sum of squares equals unity and then assembled into a concentration matrix $[C]$. The pseudoinverse equation

$$[E]=[A]\,[C]'\,\{[C]\,[C]'\}^{-1}$$

was used to calculate the spectral matrix, followed by another pseudoinverse

$$[C]=\{[E]'\,[E]\}^{-1}\,[E]\,[A]$$

to recalculate the concentration profiles. This process (truncation, normalization and pseudoinverse followed by pseudoinverse) was repeated until no further refinement occurred.

The concentration profiles and spectra of the three unknown components of stearyl alcohol in carbon tetrachloride obtained in this manner were found to make chemical sense.

This EFA procedure, unlike others, was successful in extracting concentration profiles from situations where one component profile was completely encompassed underneath another component profile.

References

[1] Malinowski, E. R., and Howery, D. G., Factor Analysis in Chemistry, Wiley Interscience, New York (1980).
[2] Gemperline, P. G., J. Chem. Inf. Comput. Sci. **24**, 206 (1984).
[3] Vandeginste, B. G. M., Derks, W., and Kateman, G., Anal. Chim. Acta **173**, 253 (1985).
[4] Gampp, H., Maeder, M., Meyer, C. J., and Zuberbuhler, A. D., Talanta **32**, 1133 (1985).
[5] Malinowski, E. R., J. Chemometrics **1**, 33 (1987).
[6] Wold, S., Technometrics **20**, 397 (1978).

Chemometrics in Europe: Selected Results

Wolfhard Wegscheider

Institute for Analytical Chemistry
Mikro- and Radiochemistry
Technical University Graz, A-8010 Graz, Austria

Chemometrics is a very international branch of science, perhaps more so than chemistry at large, and it is therefore appropriate to question the suitability of the topic to be presented. It is, however, the author's opinion that the profile of European chemometric research has a couple of distinct features that may originate more in the structure of the educational system than in the actual research topics. The profile as it will be presented is the one perceived by the author, and therefore comprises a very subjective selection of individual contributions to the field. Obviously, this is not the place to offer a review on chemometrics, let alone one that is restricted to a continent.

The definition of chemometrics [1] comprises three distinct areas characterized by the key words "optimal measurements," "maximum chemical information" and, for analytical chemistry something that sounds like the synopsis of the other two: "optimal way [to obtain] relevant information."

Information Theory

Eckschlager and Stepanek [2–5] pioneered the adaption and application of information theory in analytical chemistry. One of their important results gives the information gain of a quantitative determination [5]

$$\hat{I}\,(q\|p)=\ln\frac{(x_2-x_1)\sqrt{n_A}}{s\sqrt{2\pi e}} \qquad (1)$$

where q and p are the prior and posterior distribution of the analyte concentration for the specific cases of a rectangular prior distribution in (x_1,x_2) and a Gaussian posterior with a standard deviation s determined from n_A independent results. The penalty for an inaccurate analysis is considerable and can be expressed as

$$\hat{I}\,(r;q,p)=\hat{I}\,(q\|p)-\frac{n_A}{2}\left(\frac{d}{s}\right)^2 \qquad (2)$$

with d the difference between obtained value and the true value of x. The concept has also been extended to multicomponent analysis and multimethod characterization. In the latter case, correlations between the information provided by the different methods need to be accounted for. Given the cost of and time needed for an analysis, *information efficiency* can be deduced in a straightforward manner [2]. Recently, work was published [5] suggesting the incorporation of various relevance coefficients; this, indeed, is a very important step since it provides a way to single out the information that is judged to be relevant for a given problem. It also opens up the possibility to draw on information theory for defining objective functions in computer-aided optimization of laboratory procedures and instruments.