Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 500 (2003) 365–377
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Review

The evolution of chemometrics

Philip K. Hopke*

Departments of Chemical Engineering and Chemistry, Clarkson University, Box 5708, Potsdam, NY 13699-5708, USA

Received 24 April 2003; received in revised form 10 July 2003; accepted 14 July 2003

Abstract

Chemometrics is the application of statistical and mathematical methods to chemical problems to permit maximal collection and extraction of useful information. The development of advanced chemical instruments and processes has led to a need for advanced methods to design experiments, calibrate instruments, and analyze the resulting data. For many years, there was the prevailing view that if one needed fancy data analyses, then the experiment was not planned correctly, but now it is recognized that most systems are multivariate in nature and univariate approaches are unlikely to result in optimum solutions. At the same time, instruments have evolved in complexity, computational capability has similarly advanced so that it has been possible to develop and employ increasing complex and computationally intensive methods. In this paper, the development of chemometrics as a subfield of chemistry and particularly analytical chemistry will be presented with a view of the current state-of-the-art and the prospects for the future will be presented.

Keywords: Chemometrics; Multivariate calibration; Pattern recognition; Mixture resolution

1. Introduction

Chemometrics has been evolving as a subdiscipline in chemistry for over 30 years as the need for advanced statistical and mathematical methods has increased with the increasing sophistication of chemical instrumentation and processes. As defined by Massart et al. [1], "chemometrics is a chemical discipline that uses mathematics, statistics, and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems."

In the 1972 review of Statistical and Mathematical Methods in Analytical Chemistry [2], there are only two areas of active study; "Curve Fitting" and "Statistical Control" where curve fitting is attributed to be the analytical chemists' area of interest whereas chemical engineers were reported to be primarily concerned with quality control. In 1971, Svante Wold of the University of Umeå in Sweden coined the term "chemometrics" in a grant proposal [3] and shortly after, his collaboration with Bruce Kowalski of the University of Washington, brought the name to the United States. As part of the development of chemometrics as a separate subdiscipline, the International Chemometrics Society was formed in 1974.

The first paper with chemometrics in the title appeared in 1975 [4]. In which Kowalski suggested that chemometrics had developed to the point that it was now a functioning research area in the chemical sciences. He points to the value of pattern recognition and indicates that there were vehicles for publication of research results in journals such as the *Journal of Chemical Information and Computer Sciences*, but

* Tel.: +1-315-268-3861; fax: +1-315-4410.
E-mail address: hopkept@clarkson.edu (P.K. Hopke).

0003-2670/$ – see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/S0003-2670(03)00944-9

that it would be appropriate for *Analytical Chemistry* to publish such work. Although there were a limited number of papers published in these journals, there were significant impediments to publishing chemometrics articles as there remained considerable skepticism in the analytical community as to the need for complex data analysis tools. In the view of many established chemists, the need for complicated data analysis tools was a sign that the proper experiments were not performed, rather than understanding that advanced data analysis was integral to the maximal use of evolving new technologies. The introduction of the section "Computer Techniques and Optimization" in *Analytica Chimica Acta* in 1977 [5] was the first journal publication that was clearly dedicated to this developing area. In 1980, *Analytical Chemistry* changed the name of the review section on Statistical and Mathematical Methods in Analytical Chemistry to Chemometrics, and entered the mainstream of the field. Subsequently, in 1982, the separate section in *Analytica Chimica Acta* was terminated because chemometrics had become sufficiently well accepted to eliminate the need for this special attention [6].

Subsequently, two journals dedicated to chemometrics were launched: *Chemometrics and Intelligent Laboratory Systems* and *Journal of Chemometrics*, but now to provide a vehicle for discussion of more of the details of the methods while applications would generally be published in the broader analytical journals. Thus, the critical aspect of chemometric methods is that they had become commonly accepted into the practice of chemistry.

There has continued to be the simultaneous development of new analytical instruments that have produced data that demanded new and more effective data analysis methods while the increasing capability of personal computers has permitted more computationally intensive calculations to be performed without the need for access to "super" computers. This combination of developments has opened many new options for data analytical method improvement. Thus, over the intervening years, chemometrics has emerged to have a significant role within analytical chemistry including the incorporation into the operating systems of a number of commercial analytical instruments.

This paper will review the major areas of chemometrics related to analytical chemistry including multivariate calibration, pattern recognition, and mathematical mixture resolution and then highlight some of the new directions that chemometric methods are taking.

2. Multivariate calibration

2.1. Introduction

In many chemical studies, the concentration of one or more species are to be estimated based on measured properties of the system. For example, the absorption of electromagnetic radiation at a specific wavelength can be related to concentration through the Beer–Lambert law.

$$\frac{I(\lambda)}{I_0(\lambda)} = \mathrm{e}^{-\varepsilon_\lambda cl} \tag{1}$$

where $I(\lambda)$ is the intensity of light at wavelength λ passing through a sample of length l, $I_0(\lambda)$ the light intensity incident on the sample, ε_λ the molar extinction coefficient for the species, and c is the concentration. Typically, the molar extinction coefficients are not well characterized and corrections need to be made for other light absorbing species in the sample. Thus, a calibration of the system is made by measuring the light absorbance of a series of samples for which the concentrations of the species of interest are known. The problem is then to identify the relationship between the measured property and the concentration.

$$\ln\left[\frac{I_0(\lambda)}{I(\lambda)}\right] = \alpha_\lambda = \varepsilon cl \tag{2}$$

or

$$\log\left[\frac{I_0(\lambda)}{I(\lambda)}\right] = 2.303\varepsilon cl = \alpha cl \tag{3}$$

where α is the absorptivity of the sample at the specific wavelength.

For samples that contain multiple species, the problem becomes more complicated because there can be several components in the system that absorb at a given wavelength. The properties of the system can be considered to be a linear sum of the term representing the fundamental effects in that system times appropriate weighing factors. For example, the absorbance at a particular wavelength of a mixture of compounds for

a fixed path length, l, is considered to be a sum of the absorbencies of the individual components

$$a(\lambda) = \varepsilon_1(\lambda)c_1 l + \varepsilon_2(\lambda)c_2 l + \cdots + \varepsilon_p(\lambda)c_p l \quad (4)$$

where $\varepsilon_k(\lambda)$ is the molar extinction coefficient for the kth compound at wavelength λ and c_k is the corresponding concentration. Thus, if the absorbencies of a mixture of several absorbing species are measured at m various wavelengths, a series of equations can be obtained.

$$a_j(\lambda_i) = \varepsilon_1(\lambda_i)c_{1j} l + \varepsilon_2(\lambda_i)c_{2j} l + \cdots + \varepsilon_p(\lambda_i)c_{pj} l = \sum_{k=1}^{p} \varepsilon_k(\lambda_i)c_{kj} l \quad (5)$$

where i is the index for wavelengths *1–m*, j the index for samples *1–j*, and the number of components to be determined is p. For the calibration samples, the concentrations are known and the absorbances at a series of wavelengths is measured. A similar set of equations can be written for emission of electromagnetic radiation as would occur in X-ray fluorescence or atomic emission spectroscopy as can equivalent equations for other means of detecting a signal that is linearly related to the concentrations present in the samples. Expanding this equation to a general matrix form yields

$$A = BC \quad (6)$$

Ordinary least-squares regression has been widely applied to solve this problem by estimating the $\varepsilon_k(\lambda_i)l$ products in Eq. (5). It is very successful when a single component in present in the sample. However, for multiple analyte species, there are a variety of problems including collinearity, correlation of adjacent wavelengths, and measurement errors such that other approaches such as biased regression methods can provide better prediction of the components in the mixture. These methods are commonly called multivariate calibration methods.

Multivariate calibration methods has been widely applied because many common analytical methods provide analyses of multiple species. In these methods, it is assumed that there are a series of mixtures for which the amounts of each component are known and for which a series of properties has been measured. The methods include principal components regression (PCR), partial least-squares (PLS), simulated annealing (SA), the genetic algorithm (GA), and artificial neural networks (ANN).

2.2. Principal components regression

The measurement of the spectral characteristics of a series of calibration samples for which the concentrations are known provides a set of data in which there are significant correlations and collinearities. To reduce these problems, a principal components analysis (PCA) is performed on the matrix of spectral characteristics, **B**. The PCA produces a set of orthogonal vectors that are fully uncorrelated and by choosing a subset of the vectors can help to enhance the signal to noise ratio. This approach is called principal components regression. A principal components analysis maximizes the included variance in the analysis, but does not necessarily maximize the quality of the resulting predictions of concentrations from measurements on an unknown sample. In order to improve the quality of prediction, the approach to extracting components form the **B** matrix can be modified as described in the next section.

2.3. Partial least-squares

Partial least-squares has been the most widely applied multivariate calibration method (e.g. Martens and NFs [7]). In this case, the components of the **B** matrix are extract so as to maximize the covariance with the measured absorbances in the set of calibration samples. The relationships are developed from the training set and then applied to the set of unknowns. For the unknown sample data, the concentrations of the various constituents in the samples can often be predicted with better accuracy.

2.4. Parallel calibration

K-matrix calibration [8] or parallel calibration [9] use the spectral data as predictors in a least-squares fit of the calibration spectra to the spectra measured on the unknown concentration sample. The estimated least-squares coefficients are then applied to the calibration concentrations to estimate the unknown concentration. This method of estimation can be shown to be equivalent to a procedure known to statisticians

as generalized inverse prediction or generalized inverse regression. It is a relatively simple method and competes favorably with PLS in situations where the data is not too noisy. It has problems when the spectra are too co-linear and averaging replicate spectra is frequently a good idea before implementing this method.

2.5. Simulated annealing

Simulated annealing [10,11] is a category of stochastic optimization algorithms first used by Kirkpatrick et al. [12] for the solution of combinatorial optimization. It has been generalized by Bohachevsky et al. [13] for searching the global optimum on a continuous *m*-dimensional response surface. Thus, the search is based on an objective function defined for the problem The search space for the SA algorithm can be constrained. The search space for the problem defines a high-dimensional surface. For a realistic number of sources and samples, the surface has numerous local minima in addition to the global minimum. SA is very attractive because of its mechanism for walking out of local optima.

A typical simulated annealing step starting from a current point (C_c) generates a trial point (C_t) in its neighborhood. If the objective function value corresponding to this trial point, $f(C_t)$, is less than that corresponding to the current point, $f(C_c)$, then the trial point is accepted unconditionally and the SA step repeated. Otherwise, the trial point is accepted only with a given probability. Which is dependent on a parameter, T, called the temperature. It controls the acceptance probability. The higher the value of T, the more likely that a trial solution with $f(C_t) > f(C_c)$ is accepted. The merit of the algorithm comes from the acceptance of the detrimental state with a non-zero probability. It is this biased random walk that provides a chance to escape from any possible local minimum.

The process of generating a trial point and making it the current point if it is accepted is repeated until there is no change in the best objective function value obtained for a given number of successive steps, N_s. At this time, the process is considered to be at a steady state. When this steady state is reached, the temperature T is lowered by a factor T_f, $0 < T_f < 1$. Again the SA step is applied at the new temperature. The algorithm is stopped when the temperature is reduced to the size necessary for obtaining sufficient precision of the source contribution estimates in terms of the mixing fractions. The corresponding stopping criterion can be expressed by a predefined maximum number of steps for lowering the temperature, N_m. The detailed description of the algorithm is given by Song and Hopke [14]. The performance of SA is influenced by the parameters T, T_f, N_s and N_m. For different data sets, they may have different values in order to give satisfactory performance.

SA has been applied to parameter estimation in linear and non-linear models [15] and for feature selection [16].

2.6. Genetic algorithm

Genetic algorithms [17] were developed by Holland [18] as part of the study of adaptive processes. It is different from deterministic approaches in that the genetic algorithm tries to mimic the probabilistic evolutionary process in nature. The procedure builds a random population of strings, each of which represents one possible solution to the problem being investigated. In general, such a string can be a simple bit sequence, a vector of sequence numbers, or any other suitable linear representation of the parameters of the problem to be optimized. For each member of the population, a fitness value is calculated by transferring it to the function to be optimized. This fitness value determines the string's relative position within all members of the population regarding their closeness to their optimum value.

A new generation of values is then generated from copies of the strings of the old population, considering their relative fitness. Strings with a fitness higher than the average are copied more times than strings with lower values. As an iterative application of this reproduction rule, it would only select the best string from the initial population. Two innovative features are therefore introduced that changes the values of arbitrarily chosen strings. A mutation event can randomly change one or more information unit of a string with a specific probability. A second change is made to some randomly chosen members of the population in which two strings are mixed by cutting them and recombining them crosswise providing that a considerable amount of information of each string remains intact.

This new population then enters the next cycle and is exposed to the selective pressure and within each cycle, the string showing the highest fitness value is memorized. The process is repeated until an predefined end condition is met. It thus represents a versatile method for intelligent searching of the complex surface of solutions that may exist in the calibration problem [19–21]. It has also been widely used for feature or variable selection in a variety of problems [22,23]. There have been several comparisons of the genetic algorithms with other feature selection methods [24–26]. They can also be used in pattern recognition studies [27,28]. They tend to be computationally intensive, but with the increasing computing power that is readily available make the Genetic Algorithm more practical for many applications.

2.7. Signal to noise ratios

Signal to noise ratios were used to build models for NIR data for detergent in Brown et al. [29], and glucose measurements in McShane et al. [30]. The signal to noise methods build models based upon R^2 and quantities related to partial R^2 using a forward selection method. The procedure in McShane et al. was shown to be competitive to the Genetic Algorithm and runs much faster. Genetic algorithms can take several hours to several days to run depending upon the data size and model complexity. The signal to noise algorithms delivers results in a fraction of that time.

2.8. Artificial neural networks

The calibration problem is normally thought to be a linear one. However, Long et al. [31] found that the calibration becomes non-linear when high levels of noise are present. In this case, an artificial neural network approach can provide better prediction results.

A schematic diagram of an artificial neural network is shown in Fig. 1, where the large circles are artificial neurons and the squares represent weights that describe the importance of the signal being transmitted along a given path.

The conceptual basis for artificial neural networks has been in the literature for a long time [32–34]. However, it was not until Hopfield [35] introduced the novel concept of non-linearity between the total input received by a neuron from the other neurons and the output produced and transferred onward to other neurons. The non-linear output and feedback coupling of outputs with inputs gave new flexibility to this old architecture and sparked a major new field of interest across a number of disciplines. A detailed description of artificial neural networks in chemistry has been provided by Zupan and Gasteiger [36]. A number of applications have been made to a variety of problems including multivariate calibration [37].

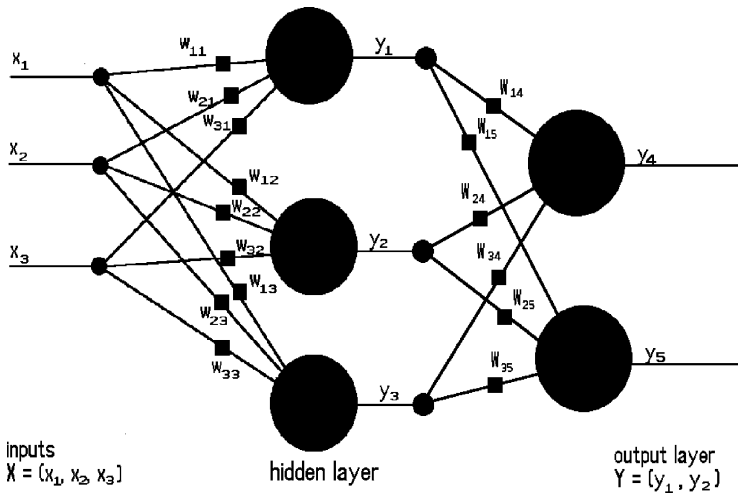

Fig. 1. Schematic diagram of a neural network with one hidden layer.

2.9. Summary

These methods are now quite mature and because of the computational capabilities available in current generation personal computers, it is possible to effectively use these methods. For example, the genetic algorithm and artificial neural networks are getting much more use and it is now possible to develop sophisticated real-time applications for process control.

3. Pattern recognition

One of the important problems in chemometrics is the identification of the relationships among chemically characterized objects. Classification and pattern recognition are widely used in chemical problem solving. For the chemist, the objects to be classified can be samples for which chemical analysis of their constituents are obtained or the spectral data measured for a compound. In pattern recognition, set of objects for which chemical data are available and the objects are to be sorted into groups of "similar" samples. It is assumed that there should be groups of objects that have similar characteristics. For example, a manufacturer is producing a product and the quality control laboratory analyzes random samples from the production line. Those samples are also tested to determine their acceptability for performing the function for which they were created. Thus, there will be two groups; acceptable and unacceptable. Suppose, it is easier to do the chemical measurements such as collecting a spectrum than the direct acceptance testing. Then, it is necessary to use the chemical data to determine whether batches of product are acceptable or not.

There are basically two approaches to the problems of sorting the objects depending on what is known a priori. If there is a series of objects for which the group assignments are known, then it is necessary to quantitatively establish the basis on which those objects were classified so that other objects of unknown class can be accurately sorted. This process is termed supervised pattern recognition and is what would be needed for the manufacturing problem above. Alternatively, if there is no a priori information of the classification of any of the objects, then the problem is to use an unsupervised pattern recognition method to find the group structure in the data as would be needed to find the groups of similar pottery material.

3.1. Unsupervised pattern recognition

3.1.1. Cluster analysis

The most commonly used unsupervised pattern recognition methods are various forms of cluster analysis. These methods are dependent on quantitative definition of the dissimilarity between pairs of objects and mathematical definitions of what is meant by groups. Cluster analysis has been applied to a number of chemical problems using a variety of clustering methods [38].

3.1.2. Eigenvector methods

The objective of these methods is to compress the *m*-dimensional information content of the data into two or three dimensions so that each sample can be plotted and visually evaluated. These methods take advantage of the correlation structure within the data to help separate signal from noise and to compress the system variance into as few components as possible. The analysis begins with the calculation of a dispersion matrix that characterizes the relationships among the variables measured for each object. In most cases, the matrix consists of the product moment correlation coefficients. The eigenvector analysis can take one of several forms including principal components analysis, correspondence analysis, or projection pursuit [39].

3.1.3. Neural networks

Artificial neural networks are methods designed to organize knowledge in ways that mimic human reasoning. They can be used for solving both unsupervised and supervised pattern recognition problems. For unsupervised problems, several types of networks can be used [40].

3.1.3.1. Kohonen networks. In the Kohonen neural network [41], an approach has been developed that tries to find groups of similar objects. The essence of Kohonen's algorithm is a repeated comparison of the vectors representing the sample data with weight vectors using a distance metric such as the Euclidian distance or the correlation coefficient. During each single comparison, a winner among the weight vectors

is found that has the highest degree of similarity (smallest distance) the sample. The winning weight vector is then adapted to be closer to the actual input sample vector.

Together with the winner, further weight vectors are modified around the winner within a limited topological neighborhood. To start, the region is large but decreases slowly with the training time. After the adaptation, the subset of weight vectors within the domain became slightly more similar to the input vectors in terms of the distance metric employed. A comparison of all of the sample vectors with all of the weight vectors and their modifications is called one epoch. This process is repeated over many epochs yielding a self-organized behavior of the samples in the low-dimensional neural array. The samples form a visual structure related to the original *m*-dimensional variables space.

The final result of this training process is the collection of subsets of sample vectors aggregated to some of the weight vectors. The number of these "loaded" neurons is *v* with *v* less than or equal to *u*. The trained weight vector for such an aggregated cluster of input vectors comes numerically very close to their mean vector. The remaining *u*–*v* weight vectors do not have any associated input vectors. Their corresponding neurons are defined as "unloaded" neurons.

At this point, only an unsupervised pattern recognition result similar to those obtainable from hierarchical cluster analysis, non-linear mapping, or principal component scores plots has been obtained. Although the topology of the multi-space will be preserved by this special type of projection, the quantitative interpretation of such a Kohonen map remains difficult especially when information about the correct neighborhood disappears during such a projection into a two-dimensional plane. The interpretability of the results is improved by adding a visualization of the global relationships among the loaded neurons in the map. A Minimal Spanning Tree [42] is added to provide this visualization. Wienke and Hopke [43] provide a complete description of the combined algorithm.

3.1.3.2. Adaptive resonance theory networks. Adaptive resonance theory (ART) was introduced by Grossberg [44,45] as a mathematical model for the description of fundamental behavioral functions of the biological brain such as learning, neglecting, parallel and distributed information storage, short- and long-term memory and pattern recognition. The aim of the ART-based models is to understand the seemingly paradoxal situation that a biological brain is able to identify an unexpected event as what it is: as belonging to or not belonging to the existing knowledge base. Moreover, ART-based models try to describe the ability of the brain to expand its knowledge by learning deviating unknowns without disturbing or destroying stored knowledge. The ART-2a [46] has been chosen for this study because its algorithm is mathematically simple and computationally inexpensive as compared with other types of ART based neural networks. The essence of the ART-2a algorithm is the dynamic formation of a weight matrix. Thus, each column of this weight matrix describes the nature of the future centroid of a class. The class size is controlled by a vigilance parameter which lies between 0 and 1. The algorithm is described elsewhere [47] and has been used with a number of problems related to the characterization of single airborne particles [48,49].

3.1.3.3. Support vector machines. Support vector machines are a relatively new class of classifiers that can incorporate a variety of kernel methods such as radial basis sets and Gaussian kernel or neural networks [50,51]. They are a method for creating functions from a set of labeled training data. The function can be a classification function (the output is binary: is the input in a category) or the function can be a general regression function. There has not yet been direct applications to chemical problems, but they are likely to become more common in the future.

3.2. Supervised pattern recognition

The other approach to pattern recognition is supervised pattern recognition in which a series of objects of known classes (training set) are used to develop mathematical rules or procedures for sorting new objects into their appropriate classes. Often additional samples of known class (test set) are sorted to test the ability of the system to perform the classification task accurately. There are a variety of methods available for performing such a task. Nearest neighbor methods and linear discriminant analysis are commonly used approaches. However, a number of other methods are available.

3.2.1. Linear learning machine

The linear learning machine [52] also finds linear boundaries among the classes. Such boundaries can be found by trial and error in which the errors from the prior trial permits the next trial to be developed more intelligently. For an *m*-dimensional space, the decision surface must be of lower dimension that *m*. The linear learning machine can be extended to multiple class separation through a sequence of surfaces that produce binary separations.

3.2.2. Potential or density methods

In potential methods [53], a potential field is assumed to exist around each object. Such a filed can be triangular or Gaussian and is characterized by its width. The cumulative potential function for the objects is determined by adding the heights of the individual object functions. This cumulative function is divided by the number of objects to obtain the mean function and assumes a probabilistic character. The classification of objects into one of the established classes can be determined by determining the class assignment that yields the maximum potential.

3.2.3. Class modeling methods

In these methods, a mathematical model is developed for each class in the training set. Such a model defines a domain in the space and procedures are developed for testing whether new objects fall within that domain. If so, these objects are assigned to the appropriate class.

3.2.3.1. SIMCA. Soft independent modeling of class analogy (SIMCA) [54] uses a principal component analysis to develop a model of each group within the training set. Thus, classes are modeled by one of a series of linear structures (a point, a line, a plane, etc.) depending on the number of components required to reproduce the class data. It is possible to define bounding surfaces around these linear structures based on the residuals of the data after fitting the components. New objects can be quantitative tested for their membership in the defined classes.

3.2.3.2. UNEQ. In this method, each group is modeled with a multivariate normal distribution [55]. The Mahalanobis distance for each object in a given class. The Mahalanobis distances follow a chi-squared distribution and a 95% confidence interval (CI) can therefore be defined. This confidence interval represents the class boundary for the class and new objects can be tested to determine if they lie within the defined class boundary.

3.2.4. Rule-building expert systems

These methods have been developed as one of the techniques of machine intelligence. The idea is that a training set is presented to an induction engine based on some classification procedure and a set of rules are derived from these examples that can be used for new objects [56]. One of the widely used methods, the ID3 algorithm [57], uses information theory as the basis for class separation and has been incorporated in several commercial programs. If there is a collection of objects representing different classes, the information entropy will be high. Separation of the objects into homogeneous groups will produce a low information entropy. This method sequentially separates the objects using a univariate classifier so as to produce the maximal decrease in entropy. The process continues until each object is in a homogeneous group. As a result of this classification, a series of rules is provided that can be used to classify objects from the same domain. Thus, instead of building rules by querying an expert, the system develops the rules from classifying the training set objects.

3.3. Summary

Again this area of study is relatively mature and there has not been many reports of new method developments and results are typically presented in application journals.

4. Mixture resolution

4.1. Introduction

In many chemical studies, the measured properties of the system can be considered to be the linear sum of the term representing the fundamental effects in that system times appropriate weighing factors. For example, rearranging Eq. (1) and expanding it to refer

to sample j in a series of mixtures yields

$$\frac{\alpha(\lambda_i)_j}{z} = \varepsilon(\lambda_i)_1 C_{1j} + \varepsilon(\lambda_i)_2 C_{2j} + \cdots + \varepsilon(\lambda_i)_p C_{pj} \quad (7)$$

where $\alpha(\lambda_i)_j$ is the absorbance of the jth mixture at wavelength λ_i, $\varepsilon(\lambda_i)_k$ the molar extinction coefficient for the kth compound at wavelength λ_i and C_{kj} is the corresponding concentration of the kth component in the jth sample. Thus, if the absorbencies of a mixture of several absorbing species are measured at m various wavelengths, a series of equations can be obtained.

$$\frac{\alpha(\lambda_i)_j}{z} = \sum_{k=1}^{p} \varepsilon(\lambda_i)_k C_{kj} \quad (8)$$

If we know what components are present and what the molar extinction coefficients are for each compound at each wavelength, the concentrations of each compound can be determined using a variety of techniques to fit to these data including the multivariate calibration methods described previously in this report.

Similarly, the assumption of mass conservation and the use of a mass balance analysis can be used to identify an apportion sources of airborne particulate matter in the atmosphere. This methodology has generally been referred to within the air pollution research community as receptor modeling [58,59]. The normal approach to obtaining a data set for receptor modeling is to determine a large number of chemical constituents such as elemental concentrations in a number of samples. Similar problems are found throughout analytical and environmental chemistry.

However, in many cases neither the number of compounds nor their absorbance spectra may be known. For example, several compounds may elute from an HPLC column at about the same retention time so that a broad elution peak or several poorly resolved peaks containing these compounds may be observed. Thus, at any point in the elution curve, there would be a mixture of the same components but in differing proportions.

4.2. Factor analysis

Principal components analysis leads to the model outlined in Eq. (6) but applied to matrices after the variables have been standardized. A quantitative mixture resolution can be obtained by uncentering the resulting component scores [60]. This approach has been termed absolute principal components analysis and has been used extensively in receptor modeling. Other forms of eigenvector analysis have been performed in which the data are not centered making the quantitative resolution of the mixture components simpler [61]. There is a problem of rotational freedom that make the mixture resolution problem ill-posed [62]. However, there are natural constraints on the results of the factor analysis that can be used to limit the space in which the solution. The elements of the component profiles and the contributions of each component to sample must be strictly non-negative. The model must reproduce the original data for each of the mixtures and the sum of the contributions should be less than or equal to the measured total mass of sample. These constraints can be included in the analysis [49].

4.2.1. Self-modeling curve resolution

One of the earliest methods to solve the indeterminacy of the factor analysis of mixtures is to utilize points in the spectra for which the value of one or more components is known to be zero [63]. There have been extensions to a variety of chemical spectroscopic measurement systems [64].

4.2.2. Target transformation

Another approach to the rotation of the factor axes to more physically realistic values is target transformation factor analysis (TTFA) originally developed by Malinou. The concept is that some idea as to nature of the properties of the mixture components (test vectors) will be known a priori. These factor component axes can then be rotated toward these test vectors. The quality of the fit to the test vectors can be examined to see if they are a reasonable representation of mixture component properties and the amounts of the components present in the mixture can then be estimated.

4.2.3. Iterative TTFA

In the absence of any information about the components, Roscoe and Hopke [65] found that unit vectors could serve as initial target vectors with the subsequently developed profiles forced to be non-negative in accordance with the natural physical constraints. Subsequently, iterative TTFA was applied to chemical

systems by Vandeginste et al. [66] and has come to be used in many mixture resolution and receptor modeling studies.

4.2.4. Evolving factor analysis

In many chemical analysis systems, there is a changing number of components in the measured mixture. For example, the effluent from a liquid chromatograph contains resolved and partially resolved components, thus producing a continuous changing mixture containing different number of components whose multiple wavelength spectra are sequentially determined. Evolving factor analysis makes repeated eigenvector analysis of a series of data matrices beginning with some base set of spectra and to which additional spectra are spectra are added [61,67]. When a new absorbing species appears, an additional significant eigenvalue should evolve. In this way, the presence of components may be more easily identified and quantitatively apportioned. This approach is applicable to any evolutionary chemical process in which a series of chemical data points are sequentially generated and which may contain a varying number of measurable components.

4.2.5. Window factor analysis

A similar approach to analyzing a series of measurements of changing components is window factor analysis. Instead of adding additional spectra to continuously make the matrix being analyzed larger, a fixed size window is sequentially analyzed [68] as the window slides along the wavelength or time axis. In this case the number of significant eigenvalues may decrease as well as increase since a subset of data are being examined that can have larger or smaller numbers of components that contribute to the observed spectral data.

4.2.6. SAFER/UNMIX and positive matrix factorization (PMF)

Other approaches to providing quantitative source resolutions have been developed and used in receptor modeling. These methods include source apportionment by factors with explicit constraints (SAFER) and positive matrix factorization (PMF). In these approaches, the space in which the factor solution lies is truncated by imposing one or more of the natural physical constraints discussed above. SAFER [69–71] can incorporate other constraints based on additional knowledge of the system. If the range in which the concentration of an element must lie can be estimated, then these limits can be added to the analysis. SAFER has been applied to interpretation of airborne organic compound data from Atlanta, GA [72] and airborne particulate matter in Los Angeles, CA [73] and Phoenix, AZ [74].

Another approach to factor analysis is positive matrix factorization. It recognizes that an eigenvector analysis is actually a process that minimizes the sum of squared residuals so that it is an implicit least-squares process. In PMF, an explicit least-squares approach is taken in which individual datum points can be individually weighted by its estimated uncertainty. Paatero and Tapper [75] have shown that a meaningful eigenvector analysis cannot be performed if the data matrix is scaled element-by-element. Paatero has developed a global optimization scheme in which the joint solution is directly determined [76]. This approach has been widely applied to the receptor modeling problem (e.g. [77–84]). It has also been applied to spectroscopic and calibration problems [82–84].

4.3. Mulimodal factor analysis

With development of combined analytical methods or the repeated sampling of complex environmental system, data sets can become multi-dimensional. For example, fluorescence spectroscopy of a series of mixtures also yields three-way data because there are excitation spectra and emission spectra for each of the mixtures. Hyphenated methods (e.g. ICP-AES) also produce three dimensional data. The multi-elemental analysis of a series of samples of airborne particles over a network of sampling sites yields a three-way data set of time, location, and composition. Thus, factor analysis has been expanded to deal with multi-way data [85–88].

The availability of the additional dimensionality to the data permits the resolution of components that are high similar in two modes, but have different behavior in the higher modes. Such components would be unresolvable by conventional two-way methods because of their collinearity, but are resolving in these data sets [89,90]. Some efforts have analyzed such data by unfolding them into a two-way matrix so the conventional methods described above can be used, but such

methods lose the unique advantage of the three-way structure. Many details of these methods still need to be resolved, but the analysis of multi-way data is a rapidly growing area of study in the solving the mixture resolution problem. There have been a number of minor alternative methods to solving such problems. Faber et al. [91] review a number of these methods.

4.4. Image analysis

One of the main areas to which mixture resolution methods are being applied is in the area of chemical image analysis. By using microscopic spectrometry, it is possible to obtain large data sets that represent the spectrum of the mixture that is present at each location across the sample. This image can then be analyzed by mixture resolution methods. The field was reviewed by Geladi and Grahn [92] and is highlighted in the recent review of chemometrics by Lavine and Workman [93]. The major challenge has been the size of the data matrix that can make it difficult to obtain rapid results. New methods such as modifications to alternating least-squares [94] and more efficient factor analysis algorithms [95] have been proposed specifically for application to multivariate images. This is an area of ongoing exploration.

5. Conclusions

Chemometrics has advanced in parallel with advancing analytical instrumentation and computational capability. It has expanded widely from its beginnings in curve fitting and quality assurance into a variety of other areas including multivariate calibration, pattern recognition, and mixture resolution. However, it can be seen that the field continues to develop and adapt to new challenges. The most recent review of the field [93] highlights the application of chemometrics to image analysis, sensor arrays, and chemoinformatics. As discussed in the Section 4.4, image analysis represents a mixture resolution problem that is an area of active research.

To make full use of sensor arrays, multivariate calibration methods are being applied. In many cases, the conceptual framework of the sensor array is to duplicate human capabilities such as an electronic nose or an electronic tongue. The idea is that although sensors are generally not highly selective, the combination of a series of sensors coupling with appropriate processing of the resulting signals can provide the necessary selectivity while retaining the low cost and power requirements that are typical of current generation devices [93].

It can be seen that the field continues to evolve and new tools will be developed and deployed as new data analysis challenges are posed to the chemometrics community.

References

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier Science, Amsterdam, 1997.
[2] L.A. Currie, J.J. Filliben, J.R. DeVoe, Statistical and mathematical methods in analytical chemistry, Anal. Chem. 44 (1972) 497R–512R.
[3] R.G. Brereton, Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems, Ellis Horwood, New York, 1990.
[4] B.R. Kowalski, Chemometrics: views and propositions, J. Chem. Inf. Comp. Sci. 15 (1975) 201–203.
[5] J.T. Clerc, E. Ziegler, Editorial, Anal. Chim. Acta 95 (1977) 1.
[6] A.M.G. Macdonald, H.L. Pardue, A. Townshend, J.T. Clerc, Editorial, Anal. Chim. Acta 134 (1982) 1.
[7] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chicester, 1989.
[8] G.L. McClure, P.B. Roush, J.F. Williams, C. Lehmann, Application of computerized infrared spectroscopy to the analysis of the principal lipids found in blood serum, in: G.L. McClure (Ed.), Computerized Quantitative Infrared Analysis, ASTM Special Technical Publication 934, American Society of Testing Materials, Philadelphia, PA, 1987, pp. 131–179.
[9] C.H. Spiegelman, J. Bennett, M. Vannucci, M.J. McShane, G.L. Coté, A transparent tool for seemingly difficult calibrations: the parallel calibration method, Anal. Chem. 72 (2000) 135–140.
[10] P.J.M. Van Laarhoven, E.H.L. Aarts, Simulated Annealing: Theory and Applications. Reidel, Dordrecht, 1987.
[11] J.H. Kalivas, Optimization using variations of simulated annealing, Chemom. Intell. Lab. Syst. 15 (1992) 1–12.
[12] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[13] I.O. Bohachevsky, M.E. Johnson, M.L. Stein, Generalized simulated annealing for function optimization, Technometrics 28 (1986) 209–217.
[14] X.H. Song, P.K. Hopke, Solving the chemical mass balance problem using an artificial neural network, Environ. Sci. Technol. 30 (1996) 531–535.
[15] J.M. Sutter, J.H. Kalivas, Anal. Chem. 63 (1991) 2383–2386.

[16] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024–2030.
[17] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[18] J.H. Holland, Adaptation in Natural and Artificial Systems, The University of Michigan Press, Ann Arbor, MI, 1975.
[19] C.B. Lucasius, G. Kateman, Applications of Genetic Algorithms in Chemometrics, in: M. Kaufmann (Ed.), Proceeding of the Third International Conference on Genetic Algorithms, Fairfax, VA, 1989.
[20] A. Bos, M. Bos, W.E. van der Linden, Anal. Chim. Acta 277 (1993) 289–295.
[21] T.-H. Li, B. Lucasius, G. Kateman, Anal. Chim. Acta 268 (1992) 123–134.
[22] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135–153.
[23] Z. Ramadan, X.-H. Song, P.K. Hopke, M.J. Johnson, K.M. Scow, Anal. Chim. Acta 446 (2001) 231–242.
[24] L. Xu, W.-J. Zhang, Anal. Chim. Acta 446 (2001) 475–481.
[25] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135–153.
[26] U. Hörchner, J.H. Kalivas, Anal. Chim. Acta 311 (1995) 1–13.
[27] B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, H.T. Mayfield, Anal. Chim. Acta 437 (2001) 233–246.
[28] B.M. Smith, P.J. Gemperline, Anal. Chim. Acta 423 (2000) 167–177.
[29] P.J. Brown, C.H. Spiegelman, M.C. Denham, Phil. Trans. R. Soc. A 337 (1991) 311–322.
[30] M.J. McShane, B.D. Cameron, G.L. Coté, M. Motamedi, C.H. Spiegelman, Anal. Chim. Acta 388 (1999) 251–264.
[31] J.R. Long, V.G. Gregoriou, P.J. Gemperline, Spectroscopic calibration and quantitation using artificial neural networks, Anal. Chem. 62 (1990) 1791–1797.
[32] W.S. McCulloch, W. Pitts, Bull. Math. Biophys. 5 (1943) 115.
[33] W. Pitts, W.S. McCulloch, Bull. Math. Biophys. 9 (1947) 127.
[34] D.O. Hebb, The Organization of Behavior, Wiley, New York, 1949.
[35] J.J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. 79 (1982) 2554.
[36] J. Zupan, J. Gasteiger, Anal. Chim. Acta 248 (1991) 1–30.
[37] G. Kateman, Neural networks in analytical chemistry? In: Proceedings of the Presentation at Fifth International Meeting on Computer Applications in Analytical Chemistry, Jena, Germany, 1982.
[38] D.L. Massart, L. Kaufman, Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[39] D.M. Glover, P.K. Hopke, Exploration of multivariate chemical data by projection pursuit, Chemom. Intell. Lab. Syst. 16 (1992) 45–59.
[40] Y.-H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley,, Reading, MA, 1989.
[41] T. Kohonen, Self-Organization and Associative Memory, third ed., Springer-Verlag, New York, 1989.
[42] R.C. Prim, Shortest connection networks and some generalizations, Bull. Syst. Tech. J. 36 (1958) 1389–1401.
[43] D. Wienke, P.K. Hopke, Projection of Prim's minimal spanning tree into Kohonen's neural network for identification of airborne particle sources by their multielement trace patterns, Anal Chim. Acta 291 (1994) 1–18.
[44] S. Grossberg, Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors, Biol. Cybern. 23 (1976) 121–134.
[45] S. Grossberg, Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, and illusions, Biol. Cybern. 23 (1976) 187–203.
[46] G.A. Carpenter, S. Grossberg, D.B. Rosen, ART-2a—an adaptive resonance algorithm for rapid category learning and recognition, Neural Netw. 4 (1991) 493–504.
[47] D. Wienke, L. Buydens, Adaptive resonance theory based neural networks—the 'ART' of real-time pattern recognition in chemical process monitoring, Trends Anal. Chem. 14 (1995) 398–406.
[48] Y. Xie, P.K. Hopke, D. Wienke, Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network, Environ. Sci. Technol. 28 (1994) 1921–1928.
[49] X.-H. Song, P.K. Hopke, D.P. Fergenson, K.A. Prather, Classification of single particles analyzed by ATOFMS using an artificial neural network, ART-2a, Anal. Chem. 71 (1999) 860–865.
[50] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.
[51] V. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.
[52] N.J. Nilsson, Linear Learning Machines, McGraw-Hill, New York, 1965.
[53] D. Coomans, I. Brookaert, Potential Pattern Recognition in Chemical and Medical Decision Making, Wiley, New York, 1986.
[54] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recog. 8 (1976) 127–139.
[55] M.P. Derde, D.L. Massart, UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution, Anal. Chim. Acta 184 (1986) 33–51.
[56] M.-P. Derde, L. Buydens, C. Guns, D.L. Massart, P.K. Hopke, Comparison of rule-building expert systems with pattern recognition for the classification of analytical data, Anal. Chem. 59 (1987) 1868–1871.
[57] J.R. Quinlin, Learning efficient classification procedures and their application to chess end games, in: R.S. Micahlski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach, Tioga, Palo Alto, CA, 1983, pp. 463–482.
[58] P.K. Hopke, Receptor Modeling in Environmental Chemistry, Wiley, New York, 1985.
[59] P.K. Hopke (Ed.), Receptor Modeling for Air Quality Management, Elsevier Science Publishers, Amsterdam, 1991.
[60] G.D. Thurston, J.D. Spengler, A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston, Atmos. Environ. 19 (1985) 9–25.
[61] E.R. Malinowski, Factor Analysis in Factor Analysis, third ed., Wiley, NY, 2002.

[62] P. Paatero, P.K. Hopke, X.H. Song, Z. Ramadan, Understanding and controlling rotations in factor analytic models, Chemom. Intell. Lab. Syst. 60 (2002) 253–264.
[63] W.H. Lawton, E.A. Sylvestre, Self-modeling curve resolution, Technometrics 13 (1971) 617–633.
[64] W. Windig, Self-modeling mixture analysis of spectra data with continuous concentration profiles, Chemom. Intell. Lab. Syst. 16 (1992) 1–16.
[65] B.A. Roscoe, P.K. Hopke, Comparison of weighted and unweighted target transformation rotations in factor analysis, Comput. Chem. 5 (1981) 1–7.
[66] B.G.M. Vandeginste, W. Derks, G. Kateman, Multicomponent self-modeling curve resolution in high-performance liquid chromatography by iterative target transformation factor analysis, Anal. Chim. Acta 173 (1985) 253–264.
[67] H.R. Keller, D.L. Massart, Evolving factor analysis, Chemom. Intell. Lab. Syst. 12 (1992) 209–224.
[68] H.R. Keller, D.L. Massart, J.O. De Beer, Window evolving factor analysis for assessment of peak homogeneity in liquid chromatography, Anal. Chem. 65 (1993) 471–475.
[69] R.C. Henry, B.M. Kim, Extension of self-modeling curve resolution to mixtures of more than three components, Chemom. Intell. Lab. Syst. 8 (1990) 205–216.
[70] B.M. Kim, R.C. Henry, Extension of self-modeling curve resolution to mixtures of more than three components. Part 2. Finding the complete solution, Chemom. Intell. Lab. Syst. 49 (1999) 67–77.
[71] B.M. Kim, R.C. Henry, Extension of self-modeling curve resolution to mixtures of more than three components. Part 3. Atmospheric aerosol data simulation studies, Chemom. Intell. Lab. Syst. 52 (2000) 145–154.
[72] R.C. Henry, C.B. Lewis, J.F. Collins, Vehicle-related hydrocarbon source compositions from ambient data: the GRACE/SAFER method, Environ. Sci. Technol. 28 (1994) 823–832.
[73] B.M. Kim, R.C. Henry, Application of SAFER model to the Los Angeles PM10 data, Atmos. Environ. 34 (2000) 1747–1759.
[74] C.W. Lewis, G.A. Norris, T.L. Conner, R.C. Henry, Source apportionment of Phoenix PM2.5 aerosol with the Unmix receptor model, J. Air Waste Manage. Assoc. 53 (2003) 325–338.
[75] P. Paatero, U. Tapper, Analysis of different modes of factor analysis as least-squares fit problems, Chemom. Intell. Lab. Syst. 18 (1993) 183–194.
[76] P. Paatero, Least-squares formulation of robust, non-negative factor analysis, Chemom. Intell. Lab. Syst. 37 (1997) 23–35.
[77] P. Anttila, P. Paatero, U. Tapper, O. Järvinen, Application of positive matrix factorization to source apportionment: results of a study of bulk deposition chemistry in Finland, Atmos. Environ. 29 (1995) 1705–1718.
[78] S. Juntto, P. Paatero, Analysis of daily precipitation data by positive matrix factorization, Environmetrics 5 (1994) 127–144.
[79] E. Lee, C.K. Chan, P. Paatero, Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, Atmos. Environ. 33 (1999) 3201–3212.
[80] S. Huang, K.A. Rahn, R. Arimoto, Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island, Atmos. Environ. 33 (1999) 2169–2185.
[81] A.V. Polissar, P.K. Hopke, R.L. Poirot, Atmospheric aerosol over vermont: chemical composition and sources, Environ. Sci. Technol. 35 (2001) 4604–4621.
[82] Y.-L. Xie, P.K. Hopke, P. Paatero, Positive matrix factorization applied to curve resolution problem, J. Chemom. 12 (1998) 357–364.
[83] A. Garrido Frenich, M. Martínez Galera, J.L. Martínez Vidal, D.L. Massart, J.R. Torres-Lapasió, K. De Braekeleer, J.H. Wang, P.K. Hopke, Resolution of multicomponent peaks by OPA, PMF and ALS, Anal. Chim. Acta 411 (2000) 145–155.
[84] Y.-L. Xie, P.K. Hopke, P. Paatero, Calibration transfer as a data reconstruction problem, Anal. Chim. Acta 384 (1999) 193–205.
[85] P. Geladi, Analysis of multi-way (multi-mode) data, Chemom. Intell. Lab. Syst. 7 (1989) 11–30.
[86] R. Henrion, N-way principal components analysis: theory, algorithms, and applications, Chemom. Intell. Lab. Syst. 25 (1994) 1–23.
[87] P. Paatero, A weighted non-negative least-squares algorithm for three-way 'PARAFAC' factor analysis, Chemom. Intell. Lab. Syst. 38 (1997) 223–242.
[88] A. Marcos, M. Foulkes, S.J. Hill, Application of a multi-way method to study long-term stability in ICP-AES, J. Anal. Atmos. Spectrom. 16 (2001) 105–114.
[89] Y. Zeng, P.K. Hopke, A new receptor model: direct trilinear decomposition followed by a matrix reconstruction, J. Chemom. 6 (1992) 65–83.
[90] E. Yakovleva, P.K. Hopke, L. Wallace, Receptor modeling assessment of PTEAM data, Environ. Sci. Technol. 33 (1999) 3645–3652.
[91] N.M. Faber, R. Bro, P.K. Hopke, Recent developments in CANDECOMP/PARAFAC algorithms: a critical review, Chemom. Intell. Lab. Syst. 65 (2003) 119–137.
[92] P. Geladi, H. Grahn, Multivariate Image Analysis, Wiley, New York, 1997.
[93] B.K. Lavine, J. Workman Jr., Chemometrics, Anal. Chem. 74 (2002) 2763–2770.
[94] J.-H. Wang, P.K. Hopke, T.M. Hancewicz, S.L. Zhang, Application of modified alternating least-squares regression to spectroscopic image analysis, Anal. Chim. Acta 476 (2003) 93–109.
[95] J.-H. Wang, P.K. Hopke, Equation-oriented system (EOS): an efficient programming approach to solve multilinear and polynomial equations by the conjugate gradient algorithm, Chemom. Intell. Lab. Syst. 55 (2001) 13–22.