Vol. 122 No. 3 September 15, 1980 ISSN 0378-4304
(Computer Techniques and Optimization, Vol. 4 No. 3)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry
Revue internationale consacrée à tous les domaines de la chimie analytique
Internationale Zeitschrift für alle Gebiete der analytischen Chemie

PUBLICATION SCHEDULE FOR 1980 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	113/1 113/2	114	115	116/1	116/2	117	118/1	118/2	119/1	119/2	120	121
Section on Computer Techniques and Optimization			122/1			122/2			122/3			122/4

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with every aspect of modern chemical analysis, both fundamental and applied. The section on *Computer Techniques and Optimization* is devoted to new developments in chemical analysis by the application of computer techniques and by interdisciplinary approaches, including statistics, systems theory and operation research. The section deals with the following topics: Computerized acquisition, processing and evaluation of data. Computerized methods for the interpretation of analytical data including chemometrics, cluster analysis, and pattern recognition. Storage and retrieval systems. Optimization procedures and their application. Automated analysis for industrial processes and quality control. Organizational problems.

Submission of Papers. Manuscripts (three copies) should be submitted as designated below for rapid and efficient handling:

Papers from the Americas to: Professor Harry L. Pardue, Department of Chemistry, Purdue University, West Lafayette, IN 47090, U.S.A.

Papers from all other countries to: Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box 363, Birmingham B15 2TT, England.

For the section on *Computer Techniques and Optimization:* Dr. J. T. Clerc, Universität Bern, Pharmazeutisches Institut, Sahlistrasse 10, CH-3012 Bern, Switzerland.

American authors are recommended to send manuscripts and proofs by INTERNATIONAL AIRMAIL.

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 111, p. 343 for detailed information. Reprints of this information are available from the Editors or from: Elsevier Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DE (Great Britain).

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An order form containing price quotations will be sent to the authors together with the proofs of their article.

Advertisements. Advertisement rates are available from the publisher.

Subscriptions. Subscriptions should be sent to: Elsevier Scientific Publishing Company, P.O. Box 211, 1000 AE Amsterdam, The Netherlands. The section on *Computer Techniques and Optimization* can be subscribed to separately.

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 10 volumes in 1980. The subscription for 1980 (Vols. 113–122) is Dfl. 1390.00 plus Dfl. 160.00 (postage) (total approx. U.S. $795.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 122) is Dfl. 139.00 plus Dfl. 16.00 (postage) (total approx. U.S. $79.50). Journals are sent automatically by airmail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–112) except Vols. 23 and 28 are available at Dfl. 153.00 (U.S. $78.50), plus Dfl. 11.00 (U.S. $5.50) postage and handling, per volume.

Claims for issues not received should be made within three months of publication of the issue, otherwise they cannot be honoured free of charge.

Customers in the U.S.A. and Canada who wish to obtain additional bibliographic information on this and other Elsevier journals should contact Elsevier/North Holland Inc., Journal Information Center, 52 Vanderbilt Avenue, New York, NY 10017. Tel: (212) 867-9040.

Analytica Chimica Acta, 122 (1980) 227—240
Computer Techniques and Optimization

Review

PATTERN RECOGNITION IN ANALYTICAL CHEMISTRY*

K. VARMUZA

Institut für Allgemeine Chemie, Technische Universität, Lehargasse 4, A-1060 Vienna (Austria)

(Received 16th November 1979)

SUMMARY

Pattern recognition methods are applied in order to classify unknown objects into categories, or to separate objects into categories. Some basic principles and simple methods of pattern recognition are described; typical applications in analytical chemistry are discussed; warnings about improper usage of pattern recognition methods are also emphasized.

During the last ten years some hundred papers have been published by more than 300 authors about applications of pattern recognition methods in chemistry. The titles of these papers include very promising ideas such as learning machines or the automatic prediction of biological activities, but also rather obscure concepts such as 200-dimensional vectors, planes and spaces. Table 1 lists those authors who have published the largest numbers of papers on these topics; this list is drawn from a representative, but certainly incomplete, reference collection [1] and may not be wholly correct. The majority of papers has, however, been published by 10—20 authors, most of whom are from the United States of America. About 250 authors have published only one or two papers on this topic. No attempt will be made here to interpret all their data.

Any scientific method can be, and may too frequently be, used inadequately. This is especially true for statistical methods, and pattern recognition is no exception. As a warning, but certainly not with the intention of discrediting the method, the ranking of authors shown in Table 1 will be used to construct a pattern recognition problem which demonstrates dramatically and also typically some of the risks attached to improper applications of pattern recognition. Suppose that the topic of interest is whether an author comes from the U.S.A. or from Europe. A glance at Table 1 shows that the number of leading names might be useful for this separation; two leading names is often accompanied by the classification U.S.A. Only one error is incurred by using this classification method, and the success rate is 94%.

*This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

TABLE 1

List of the most prolific authors working on pattern recognition in chemistry [1]

No.	Name	No. of papers	Origin	No.	Name	No. of papers	Origin
1	Isenhour T. L.	40	U.S.A.	10	Ritter G. L.	13	U.S.A.
2	Jurs P. C.	40	U.S.A.	11	Massart D. L.	12	Belgium
3	Kowalski B. R.	36	U.S.A.	12	Woodruff H. B.	12	U.S.A.
4	Wilkins C. L.	21	U.S.A.	13	Brunner T. R.	10	U.S.A.
5	Lowry S. R.	18	U.S.A.	14	Rotter H.	8	Austria
6	Wold S.	17	Sweden	15	Duewer D. L.	7	U.S.A.
7	Varmuza K.	16	Austria	16	Soltzberg L. J.	7	U.S.A.
8	Perone S. P.	14	U.S.A.	17	Stuper A. J.	7	U.S.A.
9	Bender C. F.	13	U.S.A.				

However, the facts to be noted in this example are that the data set is too small, the numbers of the members in the two classes differ significantly, and the separation was done with features which are probably irrelevant to the problem; nevertheless the success rate is high. The oddity of this example may seem slightly exaggerated, but at least two other similar examples have been presented in the literature [2, 3] to "prove" the absurdity of certain pattern recognition applications in chemistry.

This review contains a short introduction to some basic principles of pattern recognition; a few methods are described, with emphasis on mathematical simplicity, and some aspects of an objective evaluation of classification methods are discussed. Finally, some applications of pattern recognition methods in analytical chemistry are briefly reviewed. This paper is, of course, not intended to replace more systematic and exhaustive treatments [4—7].

BASIC PRINCIPLES

In analytical chemistry, a frequent problem is to determine or predict a property of an object or an event which cannot be measured directly but which has to be deduced from indirect measurements (Fig. 1). Examples are, the determination of molecular structures or biological activities of a compound. Here, such a property is called an "obscure" property. If a theoretical relationship between the indirect measurements and the obscure property is inadequately established, then pattern recognition methods may provide an approach for solving the problem.

If several measurements are made (e.g. several peak heights in a spectrum) from an object (e.g. a chemical compound), the resulting set of measurements belonging to the same object is considered as a pattern or in mathematical form as a "pattern vector"; each of the measurements corresponds to one vector component. The number of measurements defines the number of dimensions, and is an essential parameter for pattern recognition methods.

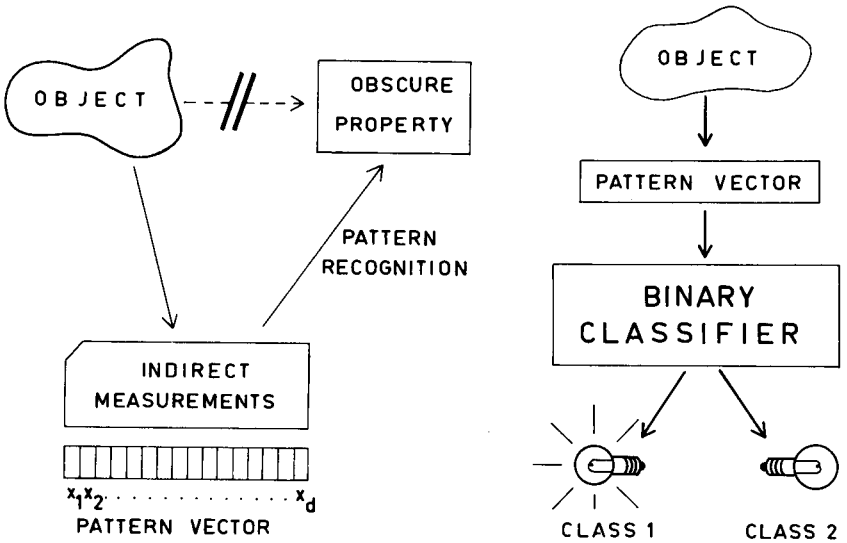


Fig. 1. Deduction of an obscure property from indirect measurements by the application of pattern recognition methods (left), and the use of a binary classifier indicating one of two alternative classes (right).

A pattern vector characterizes an object and is used to determine the unknown obscure property. Determination of an obscure property is often equivalent to the recognition of a certain class or category to which the pattern (and the object) belongs. The most widely used approach in chemical applications of pattern recognition methods is the concept of the "binary classifier" (Fig. 1). A binary classifier is an algorithm that uses the pattern components as input and produces an output which indicates one of two alternative classes to which the pattern is assigned. In an analytical problem, the pattern vector may be the mass spectrum of a compound and the two alternative classes may be defined by the presence of a certain molecular structure in class 1 and absence of this structure in class 2.

The central problem is, of course, the development of an appropriate classifier. A typical procedure used in pattern recognition is shown in Fig. 2. Firstly, a set of patterns is needed, all from objects with known class membership. This original data set is split into two parts (by the use of random numbers). One part is called the "training set", which is used for the computation of a classifier; the other is called the "prediction set", which is used to test the classifier. Computation of a classifier is often called the training of a classifier and some simple examples of training methods are explained below. Training is done so that a maximum number of patterns of the training set is assigned to the correct class. The patterns of the prediction set are used after the training to test the quality of the classifier. The percentage of correctly classified patterns from the prediction set (or preferably some other criteria) is used for objective evaluation of the classifiers. The com-

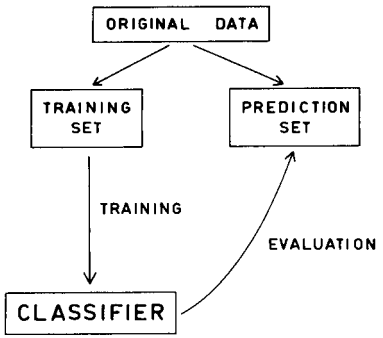


Fig. 2. Training and evaluation of classifiers.

putation of a classifier is usually laborious, requiring the use of a large computer, but the application of a given classifier to an unknown pattern is simple and fast.

A typical data set for a chemical application of pattern recognition may contain 20—1000 or more objects, each object being represented by 5—200 independent measurements. The first difficulty in all pattern recognition problems is the selection of appropriate measurements that are directly related to the classification problem. Because of the use of automatic instruments available today, large numbers of measurements are often possible. It is extremely important that the number of objects in the data set is at least three times the number of independent measurements, otherwise a random and meaningless separation of the patterns into classes may be possible, but will remain chemically meaningless.

A reduction of the number of measurements is therefore often necessary. This dimensional reduction is done by feature selection. More or less simple statistical methods are used to select those features which are most relevant to a given classification problem. At this point the chemical background of the classification problem should be recalled. Selection of appropriate features by the use of chemical knowledge may facilitate any classification problem significantly.

In the next stage, the numerical values of the selected features are mathematically transformed or scaled or weighted. The aim of this preprocessing is to facilitate the training process and to improve the classification results.

PATTERN RECOGNITION METHODS

Binary classifier

Numerous mathematical pattern recognition methods have been developed by statisticians and have been used in various fields of science, medicine and technology [8—10]. Most of the successful pattern recognition methods that have been proposed for chemical problems are conceptually simple; the understanding, and application, of these methods does not require an extensive knowledge of mathematics or statistics.

Consider a simple example where each object is characterized by only two measurements x_1 and x_2. Each object can then be represented by a point in a two-dimensional coordinate system (the "pattern space"). An equivalent representation is a vector (the pattern vector) from the origin to this point. The assumption for all pattern recognition methods is that similar objects will appear close together in the pattern space, albeit that the similarity is not chemically measurable. A very obvious case is shown in Fig. 3. The objects form two distinct clusters and each cluster contains objects of only one class. Classification of the unknown object (o) requires the determination of the cluster to which this point belongs. In real chemical applications a pattern space with much more than two or three dimensions is necessary. Clustering in a multidimensional space is of course not directly visible so that special methods are necessary.

Mere mention of multidimensional space is often enough to foreclose interest in pattern recognition, thus it should be emphatically stated that there is no qualitative difference between the geometry of, say, 100-dimensional space and that of two or three dimensions. The difference is only quantitative, which will pose no problems when computers are available.

Most simple classification methods can be explained by two-dimensional examples. Extension to more dimensions is largely formal, and does not require the imagination of more than three dimensions.

In a binary classification problem two mutually exclusive classes have to be distinguished. If the classes form well-separated clusters, they can be separated completely by a "decision plane" (a straight line in the two-dimensional example); in this case, the data set is said to be linearly separable. The decision plane is usually defined by a decision vector (weight vector) which is orthogonal to the plane (Fig. 3b). This weight vector decides whether a point belongs to class 1 or 2. In order to classify a pattern vector, it is necessary to compute the scalar product (dot product) s between the weight vector w and the pattern vector x.

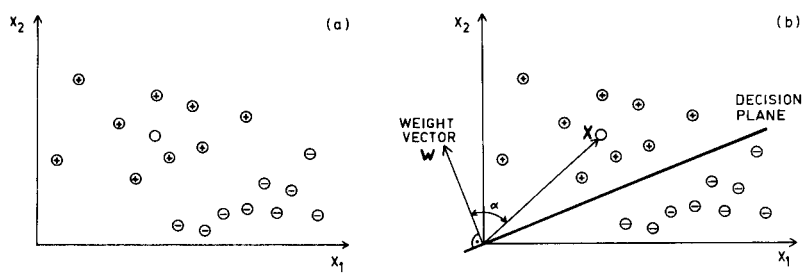


Fig. 3(a). Clustering of objects in the pattern space where x_1 and x_2 are the indirect measurements and the objects are represented by points; classification of the unknown object (o) requires determination of the cluster to which this point belongs and (b) class membership indicated by the sign of the scalar product between weight vector w and a pattern vector x.

$$s = \mathbf{w}.\mathbf{x} = |w|.|x|.\cos\alpha \tag{1}$$

The sign of the scalar product is positive for class 1, because the angle between the two vectors is less than 90° and the cosine is positive. The scalar product is negative for class 2 because the cosine is negative. A simple trick makes it possible for the decision plane always to go through the origin: all pattern vectors are augmented by an additional component (x_3 in the two-dimensional example) with the same constant value in all patterns.

Computation of the scalar product of vectors which have more than two dimensions is more easily done by another formula

$$s = \mathbf{w}.\mathbf{x} = w_1x_1 + w_2x_2 + \cdots w_dx_d \tag{2}$$

where (s) is the scalar product ($s < 0$ for class 1; $s > 0$ for class 2), (w_i) the components of the weight vector, (x_i) the components of the pattern vector, and (d) the number of dimensions (including the additional component). Thus, classification of an unknown requires only the execution of some multiplications and addition of the products. A pocket calculator is suitable for these computations.

Classification by distance to centres of gravity

The centre of gravity of a cluster (a class) corresponds to an averaged pattern and represents a prototype (a template) of that class

$$c_i = 1/n_1 \sum_{j=1}^{n_1} x_{i,j} \tag{3}$$

where (c_i) is coordinate i of the centre of gravity, ($x_{i,j}$) is the vector component i of the j^{th} pattern, and (n_1) is the number of patterns in that class. As a first approximation, the symmetry plane between the two centres of gravity can be used as a decision plane. An unknown pattern is assigned to that class which is associated with the nearest centre of gravity (see below for the computation of the Euclidean distance). Alternatively, a weight vector and computation of the scalar product may be used. This is a simple and clear classification method, even for a large number of objects and a large number of dimensions; it serves as a standard for comparisons with sophisticated methods [8, 11].

Linear regression

A better position of the decision plane can be found by a linear regression (least squares) analysis [4, 12]. Scope values z_1 and z_2 are defined for the scalar products of the patterns. For example, the scalar products should be $z_1 = -1$ for all patterns of class 1 and $z_2 = +1$ for all patterns of class 2. This cannot be achieved exactly but the decision plane is positioned in such a way that the sum of the squared errors is a minimum: $\Sigma(z-s)^2 \rightarrow \min$. The sum is taken over all patterns of the training set. Computation of a classifier by this method requires a lot of computational effort, but the performance of such a classifier is often better than for other methods. An even larger computational

effort is necessary if the optimal decision plane is sought by the simplex optimization technique [13].

Learning machine

The most popular pattern recognition method in chemistry is the learning machine [4, 14]. As used in chemistry, this is an iterative procedure which starts with an arbitrary plane. If the data are linearly separable, then the procedure will converge after a finite number of iterations by finding a decision plane that separates the classes completely. Figure 4 shows the training scheme of a learning machine. This iterative method has little similarity with human learning. Its promising, but misleading, name has probably initiated much of the enthusiasm on the one hand and opposition on the other, often connected with applications of pattern recognition in chemistry. The learning machine approach is computationally simple but has some severe disadvantages. Thus, the final position of the decision plane may be determined by atypical outlying pattern points. Furthermore, if the data set is not linearly separable, the decision plane oscillates and the training does not converge. Finally, the number of iterations necessary for a given linearly separable data set cannot be predicted. Some of these disadvantages have been eliminated

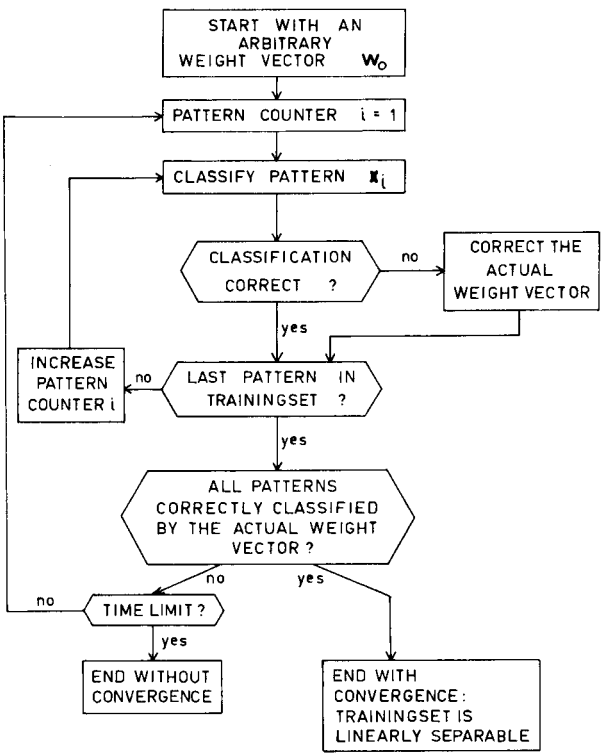


Fig. 4. Training scheme of the learning machine. The weight vector is corrected in such a way that the previously misclassified pattern is correctly classified after the adaptation.

by the use of a decision plane with a "dead zone" on both sides of the plane [15, 16].

K-nearest-neighbour method

In the clustering shown in Fig. 5 linear separation of the two classes is impossible. For such data the *K*-nearest-neighbour method (KNN method) is appropriate. Each point in this example can be considered as unknown and can be classified correctly by simply looking at the nearest-neighbour point. The nearest neighbour of a "plus" is always a "plus", and the nearest neighbour of a "minus" is always a "minus". In order to find the nearest neighbour for an arbitrary unknown, it is necessary to compute the distances from the unknown to all other pattern points. The large number of computations needed for each classification is the main disadvantage of this method. The classifier contains all patterns of the training set. Therefore this method is similar to a library search.

Instead of only one neighbour, a group of neighbours may be used, and classification of the unknown is done by a voting procedure. The number of neighbours used is normally denoted by the character *K*. The distance between two points in the multidimensional pattern space is usually defined by the Euclidean distance, which is computed in the same way as a distance in two or three dimensions:

$$D = [\sum_{i=1}^{d} (x_{i,A} - x_{i,B})^2]^{1/2} \tag{4}$$

where (D) is the Euclidean distance between points A and B, $(x_{i,A})$ and $(x_{i,B})$ are the coordinates of points A and B, and (d) is the number of dimensions.

The *K*-nearest-neighbour classification has remarkable advantages: the method is very simple, yet it can be used as a multiclass method; and, neither training of classifiers nor linearly separable clusters are necessary. New patterns may be added to the data set without difficulties. Because of the extensive storage requirements, the *K*-nearest-neighbour method is especially suited for small data sets with few dimensions. It also serves as a standard method in pattern recognition for comparisons with more sophisticated classification procedures.

Modelling clusters

In the example shown in Fig. 6 one class forms a compact cluster while the patterns of the other class are scattered throughout the pattern space. The compact class (+) may, for example, correspond to "good" samples and the other class to "bad" samples. For this asymmetric type of data it is often useful to construct a geometrical model of that class which forms a compact cluster [17]. A cluster may, for example, be modelled by a hypersphere around the centre of gravity (the hypersphere corresponds to a circle in this two-dimensional example). A more sophisticated method was developed by Wold [18]. In his SIMCA method (Soft Independent Modelling of Class Analogy), a cluster is approximated by a rectangular box (a so-called hyperbox, for more than three dimensions). The box is constructed from the plane that best fits the patterns of a class.

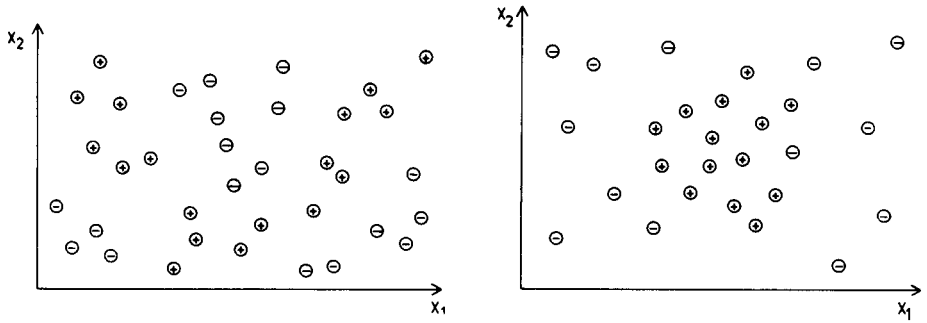


Fig. 5. Linear separation of the two classes is impossible. The K-nearest-neighbour method is appropriate for such types of data.

Fig. 6. Compact clusters can be approximated by a geometrical model (hypersphere, rectangular hyperbox).

Cluster analysis

It is difficult to construct computer programs that recognize and separate clusters of different shapes and sizes but it is often easy for the human to do this. It is evident that man is the best pattern recognizer known today. No computer program is able to compete with a human in the recognition of visual or acoustic patterns. Therefore, it may be helpful for pattern recognition problems to make a projection of the multidimensional space onto a two-dimensional plane. The scientist, with his excellent pattern recognition capabilities, can look at this projection in order to find clusters [19]. Of course, much valuable information may be lost by a simple projection and erroneous conclusions about the clusters in the multidimensional space are possible.

The mathematical problem for a good projection is to find the position of an optimum projection plane. This problem is solved by an eigenvector analysis of the data matrix. Another approach is the method of nonlinear mapping, which is not a projection onto a plane; the coordinates in the two-dimensional representation are not linear combinations of the coordinates in the hyperspace, and are found by an optimization technique.

Up to now, so-called supervised pattern recognition methods have been explained. It has been assumed that the classes of objects that are to be separated are known exactly. A more difficult question arises if one has, for example, analyzed a set of objects and wants to know if there are some natural groups of objects. An appropriate method for unsupervised pattern recognition problems is cluster analysis. Several algorithms have been described in the literature to find clusters in a multidimensional space of chemical data [20]. The main problem of all these methods is the difficulty of defining which configuration of points is a cluster and which is not. Heuristic and subjective parameters are always necessary to control the size, shape and number of clusters. The computational effort necessary for a cluster analysis is large. A cluster analysis is successful if chemical similarities can be found within the groups.

EVALUATION OF PATTERN CLASSIFIERS

The actual merit of a classification method can only be judged in connection with actual problems. This depends not only on the classification problem but also on subjective criteria such as the demands and knowledge of the user. However, initial implementation of a classifier must be based on a mathematically well-defined criterion that characterizes the quality objectively. The quality of a classifier is usually estimated by application of the classifier to a set of known patterns which have not been used for the training, the so-called prediction set.

First, an example of how an evaluation should not be done is given. Suppose that a binary classifier has been applied to a prediction set and it is reported that 96% of the patterns have been correctly classified. This high overall predictive ability may initially be greatly appreciated, but if the prediction set is examined more closely, it may be seen that, e.g. 95% of the patterns actually belong to class 1 and only 5% to class 2. A primitive classifier that always predicts class 1 would give an overall predictive ability of 95%. Accordingly, the classifier is essentially useless.

Unfortunately, many chemical applications of pattern recognition methods have been evaluated by this overall predictive ability, without regard to the composition of the prediction set. This has led to several incorrect conclusions, and over-estimates of pattern recognition methods.

A correct evaluation method is the characterization of a classifier by the predictive abilities for both classes separately. This means that the percentage of correctly classified patterns is given for each class separately. These two numbers are of course independent of the composition of the prediction set. Other objective quality criteria can be deduced from these two fundamental numbers. A single number that objectively characterizes a classifier is, for example, the mean value of both predictive abilities.

Another approach is the application of information theory [21, 22]. Before a classification has taken place, there is uncertainty about the class membership of the patterns. This uncertainty can be expressed by the entropy. For a two-class problem the entropy has a maximum of 1 bit if both classes are equally probable. After the classification, the uncertainty is lower. In terms of information theory, a classifier can be considered as a communication channel. The difference between the entropies before and after classification is called "transinformation".

In this application, the transinformation is only a function of the predictive abilities for both classes. It is an objective criterion for classifiers and has wide applicability. An ideal binary classifier always works correctly and has a transinformation of 1 bit; transinformation is 0 bit if a classifier works randomly. The primitive classifier which always predicts only one class has also a transinformation of 0 bit!

A classifier can also be interpreted as an instrument that changes the probabilities of class memberships [23]. This concept is very useful for the

user of classifiers, and an example may demonstrate this. Before classificatio usually (but not necessarily) equal probabilities for both classes are assumed. The so-called a priori probabilities for both classes are thus 50%. Now the mass spectrum of the compound is recorded and an appropriate classifier is applied to this mass spectrum (the quality of the classifier was previously tested and is defined by the predictive abilities for both classes). After the classification, the a posteriori probabilities are given for both classes. The change in the probabilities depends on the predictive abilities of the classifier and on the classification answer. The a posteriori probabilities can be used as a priori probabilities for a subsequent classifier based on the infrared spectrum of the compound, for example. The final result will be independent of the sequence in which the classifiers are used.

APPLICATIONS IN ANALYTICAL CHEMISTRY

The general concept of pattern recognition is applicable to various classification problems in science and technology [10]. Important applications are: recognition of printed or handwritten characters; analysis of pictures (recognition of chromosomes, interpretation of aerial photographs); speech recognition; medical diagnosis (interpretation of data in clinical chemistry); plant taxonomy; interpretation of seismic signals. Some applications and proposed applications for pattern recognition in analytical chemistry will be described below.

Numerous papers deal with the automatic prediction of molecular structures from spectral patterns. The generation of pattern vectors from spectra is usually obvious. A low-resolution mass spectrum contains peaks at integral mass numbers and can therefore be used directly as a pattern vector. Each mass number corresponds to a vector component [24]. Other types of spectra (i.r., Raman, n.m.r.) are digitized at small intervals and the absorptions are used as vector components [25]. More sophisticated methods for the generation of features have also been tested but often without a significant improvement of the classification results. Predictive abilities between 75 and 95% are reported for the recognition of molecular structures from spectra.

Although much effort has gone into the development of pattern recognition methods for spectral interpretations, applications in practical problems are rather poor. One reason seems to be the chronic use of data sets which are too small. Library search methods compete successfully with pattern recognition methods. While a classifier gives only a yes/no answer to one particular question, a library search may identify the unknown compound. Pattern recognition methods may have some future as a help for a preliminary interpretation of series of spectra, as recorded in a g.c.—m.s. analysis, for example. It is not normally possible to compare 100 or more spectra with a large spectral library, but a classifier may extract those spectra which stem from a certain interesting class of compounds. Subsequently, these spectra are compared with a library to identify the compounds. Spectral classifiers

may also be helpful for the identification of new compounds which are not in the library.

Applications of pattern recognition methods in polarography have been reported by Perone and co-workers [26, 27]. These papers deal mainly with the qualitative analysis for ions and recognition of the multiplicity of peaks. Special features have been generated from polarograms that describe the peak shape.

Pattern recognition methods have been applied very successfully for classification of materials, such as technological materials, food, archaeological artefacts, or environmental samples. The origin of a material can often be characterized by the contents of trace elements. A set of trace element concentrations can be used directly as a pattern vector for a sample.

The first impressive application of pattern recognition methods for a classification of materials was reported in 1972 by Kowalski et al. [28]. A total of 45 obsidian samples from different sources and 27 archaeological obsidian artefacts were analyzed by x-ray fluorescence spectroscopy. For each sample the concentrations of 10 trace elements were determined. Each obsidian sample therefore corresponded to a point in a 10-dimensional space. A cluster analysis was done by non-linear mapping from the 10-dimensional space to two dimensions. The authors identified clusters which corresponded to the different obsidian sources. The origin of almost all 27 archaeological samples could be classified. It is interesting that an independent human interpretation of the multivariate data gave the same result as the pattern recognition approach, but required several hours of tedious work.

Recently, the origins of the materials used for the Colossi of Memnon in Egypt were classified by pattern recognition methods [6]. Similar studies have been reported about the classification of paper samples [29, 30] and other industrial materials [25].

Another application of pattern recognition methods is classification of the origin of petroleum samples in environmental chemistry. Oil spills can be characterized by gas chromatograms, or infrared spectra, or trace element concentrations. Good results have been achieved even for severely weathered petroleum samples [31]. The elemental composition of atmospheric particulates has been used for classification of their origin [32].

Gas chromatograms of foods can often be correlated with quality, origin, and flavour. Characteristic peaks can be used to construct pattern vectors that characterize these properties. The origin of milk samples has been determined by using patterns of fatty oil concentrations; the method was able to distinguish between cow, sheep and goat milk [33]. Whisky samples have been classified by their gas chromatographic patterns, in order to distinguish between malt and other whiskies [34].

Interesting applications of pattern recognition principles have been reported in the field of chromatography [20]. A stationary phase can be characterized by the retention indices for some selected standard compounds. Each stationary phase therefore corresponds to a point in a multidimensional

space (the coordinates are given by the retention indices). A cluster analysis gives groups of similar stationary phases. A large distance between two points indicates a great difference in the separation properties of the two corresponding phases. Pattern recognition methods may therefore provide objective data for the selection of appropriate phases.

Promising results have been reported in applications of pattern recognition methods in investigations of relationships between molecular structure and biological activity [35]. Compounds are represented by a set of numerical descriptors that are derived from the molecular structure. A set of descriptors is used as a pattern vector. Some attempts have also been made to correlate mass spectra with biological activity, but the results have been violently criticized [2, 3].

An analytical method can be represented by a point — or better by a region — in a multidimensional "space of procedures". The coordinates are given by the parameters of the analytical method (e.g., accuracy, sensitivity, cost, etc.). Some preliminary attempts have been made to describe the problems of evaluation or selection of analytical methods in terms of pattern recognition [36, 37].

CONCLUSIONS

The actual value of pattern recognition methods in analytical chemistry is not completely clear at the moment. Many methods have been developed and proposed for chemical applications but only a few practical examples show a really impressive success. However, the philosophy of pattern recognition can be useful for many chemometric problems in analytical chemistry. There are three points which will require more attention in the future. First, more cooperation is needed between chemists (who understand the problems) and experts in pattern recognition (who know the methods). Secondly, packages [38] with classification methods which are prepared for non-expert users should be more widely distributed. Thirdly, more chemists should adopt more realistic views about the possibilities and limitations of pattern recognition methods.

Pattern recognition is an approach for data reduction which is essential today. Already there is far more chemical information than anyone knows how to manage. Pattern recognition can be a valuable part of a computer-assisted interpretation of chemical data. However, complete automatic processing of complex interpretations seems to be uneconomical and unrealistic in the near future.

REFERENCES

1 K. Varmuza, Pattern Recognition in Chemistry, in preparation.
2 J. T. Clerc, P. Nägeli and J. Seibl, Chimia, 27 (1973) 639.
3 C. L. Perrin, Science, 183 (1974) 551.
4 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.

5 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
6 J. R. McGill and B. R. Kowalski, Appl. Spectrosc., 31 (1977) 87.
7 C. L. Wilkins and P. C. Jurs, in P. R. Griffiths (Ed.), Transform Techniques in Chemistry, Plenum Press, New York, 1978.
8 R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
9 H. E. Steinhagen and S. Fuchs, Objekterkennung — Einführung in die mathematischen Methoden der Zeichenerkennung, VEB-Verlag Technik, Berlin, 1976.
10 B. G. Batchelor, Practical Approach to Pattern Classification, Plenum Press, New York, 1974.
11 H. Rotter and K. Varmuza, Anal. Chim. Acta, 95 (1977) 25.
12 H. Rotter and K. Varmuza, Anal. Chim. Acta, 103 (1978) 61.
13 S. L. Kaberline and C. L. Wilkins, Anal. Chim. Acta, 103 (1978) 417.
14 N. J. Nilsson, Learning Machines, McGraw Hill, New York, 1965.
15 L. E. Wangen, N. M. Frew and T. L. Isenhour, Anal. Chem., 43 (1971) 845.
16 D. R. Preuss and P. C. Jurs, Anal. Chem., 46 (1974) 520.
17 C. Albano, W. J. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.
18 S. Wold, Pattern Recognition, 8 (1976) 127.
19 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.
20 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
21 K. Varmuza and H. Rotter, Monatsh. Chem., 107 (1976) 547.
22 K. Varmuza and H. Rotter, Advances in Mass Spectrometry, 7 (1978) 1099.
23 K. Varmuza and H. Rotter, Advances in Mass Spectrometry, 8 (1980) in print.
24 J. R. Chapman, Computers in Mass Spectrometry, Academic Press, London, 1978.
25 B. R. Kowalski, in C. E. Klopfenstein and C. L. Wilkins (Eds.), Computers in Chemical and Biochemical Research, Vol. 2, Academic Press, New York, 1974.
26 Q. V. Thomas, R. A. de Palma and S. P. Perone, Anal. Chem., 49 (1977) 1376.
27 R. A. de Palma and S. P. Perone, Anal. Chem., 51 (1979) 825.
28 B. R. Kowalski, T. F. Schatzki and F. H. Stross, Anal. Chem., 44 (1972) 2176.
29 D. L. Duewer and B. R. Kowalski, Anal. Chem., 47 (1975) 526.
30 P. J. Simon, B. C. Giessen and T. R. Copeland, Anal. Chem., 49 (1977) 2285.
31 H. A. Clark and P. C. Jurs, Anal. Chem., 51 (1979) 616.
32 P. D. Gaarenstroom, S. P. Perone and J. L. Moyers, Environ. Sci. Technol., 11 (1977) 795.
33 J. Smeyers-Verbeke, D. L. Massart and D. Coomans, J. Ass. Offic. Anal. Chem., 60 (1977) 1382.
34 B. E. H. Saxberg, D. L. Duewer, J. L. Booker and B. R. Kowalski, Anal. Chim. Acta, 103 (1978) 201.
35 A. J. Stuper, W. E. Brugger and P. C. Jurs, Computer Assisted Studies of Chemical Structure and Biological Function, Wiley, New York, 1979.
36 H. Kaiser, Anal. Chem., 42 (1970) 24A.
37 B. G. M. Vandeginste, Anal. Lett., 10 (1977) 661.
38 A. M. Harper, D. L. Duewer, B. R. Kowalski and J. L. Fasching, in Chemometrics: Theory and Application, ed.: B. R. Kowalski, ACS Symp. Ser., 52, American Chemical Society, Washington, D.C., 1977.

Analytica Chimica Acta, 122 (1980) 241—247
Computer Techniques and Optimization

A STRUCTURE—BIOLOGICAL ACTIVITY STUDY BASED ON CLUSTER ANALYSIS AND THE NONLINEAR MAPPING METHOD OF PATTERN RECOGNITION

YOSHIMASA TAKAHASHI, YOSHIKATSU MIYASHITA, HIDETSUGU ABE and SHIN-ICHI SASAKI*

School of Materials Science, Toyohashi University of Technology, 1 Tempaku-cho, Toyohashi 440 (Japan)

YASUHIKO YOTSUI and MITSUJI SANO

Research Institute, Daiichi Seiyaku Co. Ltd., Edogawa-ku, Tokyo 132 (Japan)

(Received 18th December 1979)

SUMMARY

Cluster analysis is used in a study of structure—activity relationships of biologically active compounds. A hierarchal clustering technique was applied to 29 typical antibiotics using 27 antibacterial activities. These antibiotics were of various types; penicillins, cephalosporins, aminoglycosides, macrolides, tetracyclines, and peptides. The result was obtained as a branching tree diagram. The technique allowed the antibiotics to be distributed into 6 clusters, each cluster mostly consisting of compounds with a similar structure. Nonlinear mapping was used to display the 27-dimensional data structure of the antibiotics. The nonlinear map was compared with the clusters obtained by cluster analysis.

A great deal of effort has been devoted to studies of structure—activity relationships of drugs. Various methods have been developed for correlating potencies or types of activities of drugs with the chemical structures. One of these methods was initiated by Hansch and his co-workers [1]. The Hansch approach is based on the assumption that changes in the potencies of biological activities for a series of compounds can be related to changes in their various physicochemical properties. This technique has been applied to a number of studies of structure—activity relationships [2, 3]. Another approach was described by Free and Wilson [4]. This is a mathematical model based on the assumption that the effect of a substituent at a certain position on the potencies of a series of compounds is constant and additive.

Recently, pattern recognition techniques have been applied to investigations of structure—activity relationships [5, 6] where types of activities were classified by pattern recognition [7, 8]. In addition, an application of cluster analysis in structure—activity relationships was discussed by Hansch et al. [9]. They used a hierarchal clustering procedure in order to achieve a good initial choice of substituent in their study of quantitative structure—

activity relationships by the Hansch approach. Several studies exploring the utility of various analytical and elucidative techniques have been reported [10].

The present study is concerned with the application of pattern recognition to a study of structure—activity relationships between various antibiotics and antibacterial activities. Cluster analysis and a nonlinear mapping method were employed in analyzing the antibacterial spectral data. A clustering technique was applied to 29 antibiotics of diverse structural types. Further, the nonlinear mapping method was used to visualize the data structure and clustering results.

DATA SET

The samples used in this study are listed in Table 1. These compounds are 29 typical antibiotics which show antibacterial spectra against 27 species of bacteria (21 gram-negative and 6 gram-positive) measured by the two-fold agar dilution method. The 27 species of bacteria are listed in Table 2. The data set contains six different types of parent structures, including penicillins, cephalosporins, aminoglycosides, macrolides, tetracyclines, and peptide antibiotics. To examine the reproducibility of antibacterial activities, antibacterial spectra of tetracycline measured on different days were contained in the data set. Therefore, the data set consists of 30 samples. The antibacterial activity is represented as the logarithm of the reciprocal of minimum inhibitory concentration (MIC).

CLUSTER ANALYSIS

Cluster analysis is a powerful technique for finding "homogeneous" groups (clusters) in a given data set. The term "homogeneous" means here that all close points in the same cluster are similar to each other in some defined property and are different from points in the other clusters in this property. These algorithms are nonparametric in nature, i.e., no parametric model of an underlying probability density function for clustering multivariate data is used in the algorithms.

Patterns with m features can be represented as points in m-dimensional space. Each point (a pattern vector) represents an antibacterial spectrum of an antibiotic. In this case, the antibacterial spectrum becomes a 27-dimensional pattern. The squared Euclidean distance D_{ij} between pattern vectors, $\mathbf{X}_i = (x_{i1}, x_{i2}. \ldots, x_{im})$ and $\mathbf{X}_j = (x_{j1}, x_{j2} \ldots, x_{jm})$, is defined as

$$D_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \tag{1}$$

Hierarchal clustering was employed in this work. Initially, there are n distinct clusters for n patterns in the clustering set and each cluster consists of only one pattern. In the first step, the nearest pair of distinct clusters

C_i and C_j is sought, then these are merged, which results in a new cluster C_i by deleting C_j. Thus, the number of clusters decreases one by one. This procedure is repeated until one large cluster is formed. The distance between the clusters is measured by the Euclidean distance between the center points of the clusters. $D^{(i)}$ is a squared Euclidean distance between clusters in the i^{th} tep. If cluster C_g is generated by fusing clusters C_h and C_l in the next step, the squared distance between clusters C_f and C_g is given by $D_{fg}^{(i+l)}$.

$$D_{fg}^{(i+l)} = \frac{n_h}{n_g} D_{fh}^{(i)} + \frac{n_l}{n_g} D_{fl}^{(i)} - \frac{n_h n_l}{n_g^2} D_{hl}^{(i)} \tag{2}$$

where n_g, n_h and n_l are the numbers of patterns in clusters C_g, C_h and C_l, respectively.

Results

The hierarchal clustering algorithm was applied to the antibacterial spectra of 30 samples. The result is given by a branching-tree diagram (Fig. 1) which

TABLE 1

List of antibiotics, penicillins (1—7), cephalosporins (8—13), aminoglycosides (14—19), macrolides (20—23), tetracyclines (24—28) and peptides (29, 30)

1 Penicillin-G	9 Cephaloridine	17 Paromomycin	25 Oxytetracycline
2 Methicillin	10 Cephaloglycin	18 Gentamicin	26 Tetracycline (1)[a]
3 Oxacillin	11 Cephoxitin	19 Vistamycin	27 Tetracycline (2)[a]
4 Ampicillin	12 Cephazolin	20 Erythromycin	28 Pyrrolidinomethyl-tetracycline
5 Cloxacillin	13 Cephalexin	21 Spiramycin	
6 Sulbenicillin	14 Streptomycin	22 Kitasamycin	29 Colistin
7 Carbenicillin	15 Neomycin	23 Oleandomycin	30 Polymyxin B
8 Cephalothin	16 Kanamycin	24 Chlortetracycline	

[a]Antibacterial spectra were measured on different days.

TABLE 2

List of bacteria, gram-negative (1—21) and gram-positive (22—27)

1 *Escherichia coli*	15 *Pseudomonas cepacia*
2 *Shigella flexneri*	16 *Pseudomonas multophilia*
3 *Salmonella typhimurium*	17 *Pseudomonas putida*
4 *Salmonella enteritidis*	18 *Achremobacter xylosoxidans*
5 *Proteus vulgaris*	19 *Acinetobacter anitratus*
6 *Proteus mirabilis*	20 *Agrobacterium faecalis*
7 *Citrobacter freundii*	21 *Flavobacterium meningosepticum*
8 *Haemophilus alvei*	22 *Staphylococcus aureus*
9 *Yersina enterocolitica*	23 *Staphylococcus epidermidis*
10 *Klebsiella pneumoniae*	24 *Streptococcus pyogenes*
11 *Enterobacterium cloacae*	25 *Streptococcus pneumoniae*
12 *Enterobacterium aerogenes*	26 *Streptococcus faecalis*
13 *Serratia marcescens*	27 *Bacillus subtilis*
14 *Pseudomonas aeruginosa*	

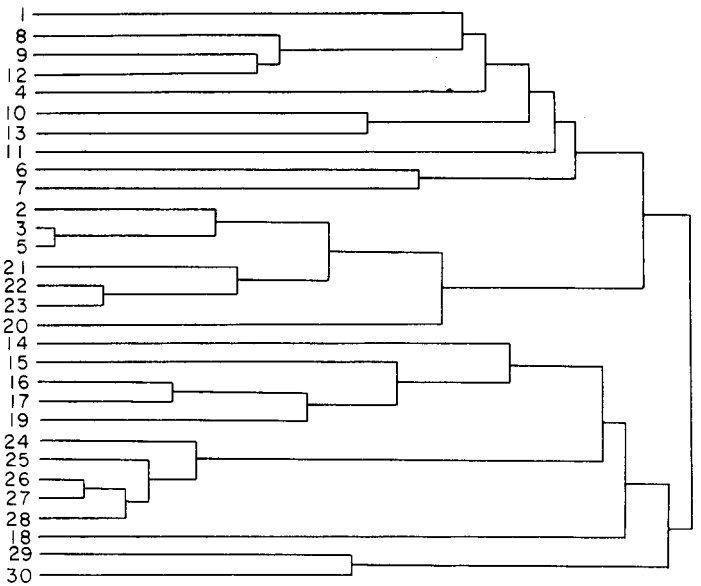


Fig. 1. Branching tree diagram for 30 samples.

indicates the successive clustering of the samples. The samples that show more distinct antibacterial spectra are merged later. That the antibacterial spectra of tetracycline are clustered in the second step shows the high reproducibility of the spectra.

Since there are six diverse types of compounds with respect to the fundamental structures, the whole series of samples is divided into six clusters on the basis of the diagram. The result is shown in Table 3. The clusters are assigned in order of the mean activity value of compounds in each cluster ($C_1, C_2, \ldots, C_6$).

The mean activity value of cluster C_p is defined as

$$A_{\text{mean}}^{(p)} = n_p^{-1} \sum_{i=1}^{np} \sum_{k=1}^{m} x_{ik}^{(p)} \tag{3}$$

TABLE 3

Members in each cluster

Cluster	A_{mean}	Sample no.	Group
C_1	196.0	18	Aminoglycoside
C_2	139.6	24 25 26 27 28	Tetracyclines
C_3	138.6	14 15 16 17 19	Aminoglycosides
C_4	131.0	29 30	Peptides
C_5	103.3	1 4 6 7	Penicillins
		8 9 10 11 12 13	Cephalosporins
C_6	60.0	2 3 5	Penicillins
		20 21 22 23	Macrolides

where the summations are over the measurements and the number of samples in each cluster, $x_{ik}^{(p)}$ is the activity of the i^{th} sample in cluster C_p for the k^{th} bacterium, and n_p is the number of patterns in cluster C_p. The clusters C_1, C_2, C_3 and C_4 are simple clusters, which consist of aminoglycoside, tetracyclines, aminoglycosides, and peptide antibiotics, respectively. As both penicillins and cephalosporins are β-lactam antibiotics, C_5 also can be considered as a simple cluster.

Thus, it was found that each of the clusters C_1 through C_5 consists of compounds with a similar parent structure. Only the cluster C_6 is a mixed cluster of penicillins and macrolides. The penicillins are dispersed between two clusters (C_5 or C_6), and all the penicillins with a benzyl group are concentrated in cluster C_5.

Therefore, it is confirmed that characteristic antibacterial activities are related to the basic chemical structures.

VISUAL DISPLAY OF THE DATA SET BY NONLINEAR MAPPING

The human eye is the best pattern recognizer, but features in n-dimensional space ($n > 3$) cannot be seen. However, a computer can easily reduce such multidimensional data to recognizable two- or three-dimensional space. One of the best ways of doing this is nonlinear mapping (NLM) [11], which is particularly useful when measurements are non-correlated. If measurements are highly correlated, the Karhunen–Loève method involving an eigenvector projection is preferable. An original Euclidean distance d_{ij}^* between sample points $\mathbf{X}_i$ and $\mathbf{X}_j$ in the data space is defined as the square root of D_{ij} (eqn. 1)

$$d_{ij}^* = D_{ij}^{1/2} = \left[\sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \right]^{1/2} \tag{4}$$

Here a m-dimensional pattern $\mathbf{X}_i$ is mapped to an r-dimensional pattern $\mathbf{U}_i = (u_{i1}, u_{i2}, \ldots, u_{ir})$ in a nonlinear manner. The distance d_{ij} between $\mathbf{U}_i$ and $\mathbf{U}_j$ in the reduced r-dimensional space is defined as

$$d_{ij} = \left[\sum_{k=1}^{r} (u_{ik} - u_{jk})^2 \right] \tag{5}$$

Then, new points are created by minimizing an error function E, which is a criterion for deciding whether or not one configuration is better than another. The error function is defined as

$$E(\rho) = \sum_{i>j} [(d_{ij} - d_{ij}^*)^2 / d_{ij}^{*\rho}] \tag{6}$$

where ρ is a parameter. For the present purpose, $\rho = 2$; this corresponds to an equal weighting of small and large distances. The minimization of the error function E was implemented by using the Powell function minimization method without derivative.

NLM was used here to display the data structure of 27-dimensional antibacterial activities in 2-dimensional space. Figure 2 shows the nonlinear map

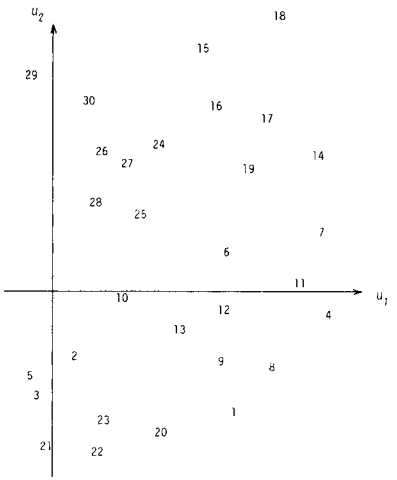


Fig. 2. NLM of antibacterial spectral data from 27-dimensional space to 2-dimensiona space.

of the antibacterial activity data. The set of antibiotics was divided into six clusters according to the branching diagram obtained by cluster analysis The six clusters are displayed as shown in Fig. 3. The clustering method proves to be realistic because most of the members in a cluster have the same structural feature, as shown in Fig. 3. Thus, the use of cluster analysis and NLM methods together is very valuable in visualizing multidimensional information.

DISCUSSION

As illustrated above, correlations between the structures of antibiotics and their antibacterial activities can be elucidated by the use of cluster analysis and NLM. A hierarchal clustering of a set of 29 antibiotics with their antibacterial activities has been implemented. As the set was separated into six clusters, the members in each cluster mostly consisted of compounds with a similar structure. The results confirm that antibiotics exhibit characteristic activities according to their basic chemical structures.

The data structure for antibacterial activities can be displayed by using NLM. This method is useful in gaining a comprehensive understanding of multidimensional data structure. In conclusion, from a practical point of view, if cluster analysis is used together with the NLM method, a more powerful and effective approach becomes available for the study of structure-biological activity relationships.

The authors thank Dr. Y. Osada and his co-workers for measuring the antibacterial spectra.

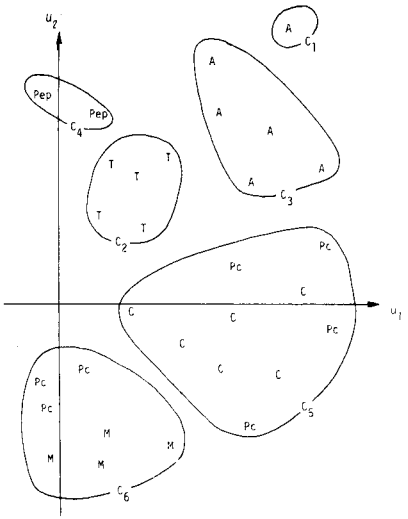


Fig. 3. Six clusters with NLM as display. (A) Aminoglycoside, (C) cephalosporin, (M) macrolide, (Pc) penicillin, (Pep) peptide, (T) tetracycline.

REFERENCES

1 C. Hansch and T. Fujita, J. Am. Chem. Soc., 86 (1964) 1616.
2 C. Silipo and C. Hansch, J. Am. Chem. Soc., 97 (1975) 6849.
3 J. A. Keverling Buisman (Ed.), Biological Activity and Chemical Structure, Elsevier, Amsterdam, 1977.
4 S. M. Free and J. M. Wilson, J. Med. Chem., 7 (1964) 395.
5 A. J. Stuper, W. E. Brugger and P. C. Jurs, Computer Assisted Studies of Chemical Structure and Biological Function, Wiley, New York, 1979.
6 A. Cammarata and G. K. Menson, J. Med. Chem., 19 (1976) 739.
7 K. C. Chu, Anal. Chem., 46 (1974) 1181.
8 H. Abe, S. Kumazawa, T. Taji and S. Sasaki, Biomed. Mass Spectrom., 3 (1976) 151.
9 C. Hansch, S. H. Unger and A. B. Forsythe, J. Med. Chem., 16 (1973) 1217.
10 B. R. Kowalski (Ed.), Chemometrics, Theory and Application, ACS Symposium Series, No. 52, Am. Chem. Soc., Washington, 1977.
11 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.

Analytica Chimica Acta, 122 (1980) 249—256
Computer Techniques and Optimization

APPLICATION OF ARTIFICIAL INTELLIGENCE SYSTEMS IN MOLECULAR SPECTROSCOPY†

L. A. GRIBOV

V. I. Vernadsky Institute of Geochemistry and Analytical Chemistry, U.S.S.R. Academy of Sciences, Vorobievskoye shosse, Moscow V-334 (U.S.S.R.)

(Received 24th September 1979)

SUMMARY

The merits and demerits of data retrieval and artificial intelligence systems for identifying polyatomic molecules from their molecular spectra are considered. It is concluded that the creation of artificial intelligence systems provides the most promising developments for the future of analytical molecular spectroscopy. In these systems, experimental spectra are compared with computer-generated data in the course of solution of the analytical problem and not with data already stored in the data bank.

Identification of polyatomic molecules is one of the challenging problems in modern chemistry. The synthesis of new compounds, studies of biologically active substances and technologies for environmental control all place heavy demands on facilities for identification of molecular products. Delays in providing adequate means for solving this problem are of course detrimental to progress in many fields.

The most satisfactory methods for identification of components in mixtures are the various types of molecular spectroscopy, e.g. mass, magnetic resonance and microwave spectroscopy, which have numerous advantages over purely chemical procedures. These spectroscopic methods can be remarkably sensitive and can be used over wide ranges of conditions (low and high temperatures, pressures, aggregate states, etc.). The instrumentation can be incorporated into systems directly linked to computers, thus allowing on-line processing of the measurement data, which is an important advantage. There is, therefore, great interest in the creation of analytical research centres, which utilize systems combining spectral sensors and computers.

At present, there are well developed techniques for recording spectra in the optical and radio ranges and constantly improving instrumentation is available. Linkage with computers is a technical and organizational problem rather than a chemical one. A scientific problem that still requires efficient solution is the processing of the measurement data to gain the maximum

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

amount of unambiguous information on the molecule studied. General concepts of the processing of spectral data, and the appropriate algorithms and software have been widely studied during the last decade. A new science has appeared at the interface of physics, chemistry, mathematics and cybernetics which might be called "spectro-cybernetics". The remarkable progress made in this area has been surveyed in two recent reviews [1, 2].

Despite the number of successful developments, it would be utterly wrong to imagine that further progress will consist simply of improving the approaches already developed or their technical and mathematical formulations. Instead, it can be confidently asserted that we are still in the initial stages and can speak only in terms of various models of research systems in molecular spectroscopy. In this paper, the experience gained so far is analyzed, and general modes of investigation which may lead to successful developments are outlined.

At present, trends in the development of computerized spectroscopy can be divided into three groups: mathematical pattern recognition procedures, data retrieval systems and artificial intelligence systems. Jurs and Isenhour [3] have published a comprehensive description of pattern recognition methods. These methods have provided some successful solutions but on the whole they do possess broad scope and cannot be applied to the identification of individual molecules. It is advisable to use them only for specific purposes, such as classification of the material studied to a certain chemical group.

The data retrieval systems are considerably more promising. There are systems for molecular spectroscopy in the U.S.S.R. [4], U.S.A. [5], Switzerland [6] and Yugoslavia [7] which have been in operation for a number of years. These systems have helped to solve various important practical problems, which has provided an impetus to create special data banks. An advantage of data retrieval systems lies in their relatively simple software and their ability to accumulate a large amount of information and to identify compounds from one or two types of spectral data without additional "non-spectral" information (e.g., without the empirical formula). However, basic shortcomings of such systems are also becoming apparent (see below).

Thirdly, there are the artificial intelligence systems for solving various wide-ranging problems in molecular spectroscopy. Systems described by Gray [8], Sasaki et al. [9], Carhart et al. [10], Beech et al. [11] and Gribov et al. [12, 13] are examples. These systems are still not very elaborate and can handle only relatively simple molecules. However, they seem to have practically unlimited resources for improvement.

Shortcomings of data retrieval systems

Identification of a polyatomic molecule from its experimental spectra by means of a data retrieval system is done by comparing sets of spectral characteristics with those available in the respective data bank. Essentially,

this is search by the "fingerprint" technique. Generally speaking, the software of data retrieval systems can also be used for solving other problems, e.g. determination of the presence of certain structural components in a molecule, classification of the system, etc.

With the data retrieval systems, capacity is directly proportional to the number of compounds whose spectral characteristics are stored in the bank. The largest systems store spectral data for about 10^5 molecules. This may seem an enormous number, but it has to be remembered that this is about the number of new organic compounds synthesized in the world annually. It becomes totally impossible to accumulate the spectral information in a bank of any size as soon as it appears. Moreover, the data on new compounds presenting the most interest for identification are often not introduced into the memory of the computer. Accordingly, although data retrieval systems pretend to be multipurpose, they are doomed to lag behind progress in synthetic chemistry. Because of the complexity of data accumulation in the data bank and necessity of standardizing them, these data do not actually meet modern spectral experimental requirements. Some spectral data banks are based on infrared spectra obtained using devices with a range up to 500—700 cm^{-1}. Meanwhile, the range of modern infrared spectrometers has been widened considerably. The absorption bands most advantageous analytically lie in the low-frequency range.

The application of laser Raman spectroscopy, low-temperature techniques and high-resolution methods has considerably widened. All this can hinder the application of spectral data obtained by the new technique to the identification of polyatomic molecules by data retrieval. This is a major disadvantage which it would be very difficult to eliminate by improvements in the data bank.

As a rule, any data retrieval system contains not only the absolutely necessary information but also much extraneous data. It is well known, for example, that, given information on spectra and structure of some molecules, the spectra of other molecules can be established. Thus, mathematically, only a limited number of spectra is independent. The availability of the dependent information makes the system irrationally complicated. For an accurate spectral identification of a molecule by the "fingerprint" method, experimental and standard spectra should be compared, strictly speaking, not at the level of single spectral features but at the level of the total sum of their features (e.g., at a level of precise spectral curves). However, a detailed description of spectra stored in the system makes it still more complicated.

There are other shortcomings but those outlined above seem adequate to support the conclusion that specialized data retrieval systems offer little potential for improvement and have various inherent disadvantages. Thus their improvement cannot be regarded as a guideline for completely automated identification of polyatomic molecules by means of spectral data. This certainly does not imply that any further development of data retrieval

systems is futile. Data banks in such systems are useful as a source of primary information for building up more sophisticated systems of artificial intelligence. More immediately, specialized data retrieval systems can be very helpful when designed for analyzing particular industrial products, or identifying the pollution status of natural and waste waters, atmospheres, etc.

PRINCIPLES IN DESIGNING ARTIFICIAL INTELLIGENCE SYSTEMS FOR MOLECULAR SPECTROSCOPY

Identification of compounds in initial data-processing systems following the artificial intelligence pattern is done by comparing experimental molecular spectra with computed spectra. However, unlike data retrieval systems, such comparative spectra are not stored in the computer memory originally but are generated by following certain rules in the process of solution. In this process, automatically formulated hypotheses on the structure of the molecule studied are involved. These hypotheses are continuously updated, and there is no need to store a large amount of data in the bank. It is sufficient to "memorize" relatively few basic axioms. Great complexity of the algorithms for a high-class computer is not crucial in designing such systems. The artificial intelligence systems by their very nature are free of the shortcomings inherent in the data retrieval systems. They can be multipurpose and flexible, and they can be readily updated by using any new knowledge from the theory of molecular structure and spectra, and quantum chemistry. This is what makes artificial intelligence systems capable of continuous and theoretically unlimited improvement.

As mentioned above, fairly sophisticated artificial intelligence systems are already available for molecular spectroscopy, though they are of course far from perfect. However, the experience gained to date allows some conclusions and predictions.

With artificial intelligence systems, sufficiently reliable identification even of isomeric structures can be obtained even when a computed spectrum is not identical but only relatively close to the experimental spectrum because of restrictions in different theories. Examples of this type have been cited [14].

In the first stage, the computer must isolate a set of "suspected" structures based on a number of features. In the second stage, the sequential generation of more accurate spectra and their comparison with the experimental spectra gradually exclude false structures of the initial set until an unambiguous result is achieved. Hence, any system of artificial intelligence must have two large blocks: one for building hypothetical structures and the other for generating spectra at different levels of accuracy. Moreover, there must be a feedback for comparing the theoretical and experimental spectra. Mathematically, the first block formulates a set of initial hypotheses, whereas the second block and the feedback test the validity of these hypotheses. The

algorithms of the first block must be capable of structural group analysis, computation of molecular structures at the adjacency matrix level and computation of spatial models for the "suspected" structures at varying levels of accuracy.

The algorithms of the second block, using the data on molecular structures, should be able to generate spectral features including complete spectral curves for these structures. For sequential estimation of the validity of the initial hypotheses, it is advisable to perform the primary filtering of hypotheses by comparing only very rough theoretical spectra with the experimental spectra. It is possible, for example, to build a reasonably representative vibrational spectrum of a structure by analyzing the closed additive groupings comprising it and the corresponding spectrostructural correlations. As experience has shown, this simple procedure can sometimes reduce the number of possible structures by half or even more. Further testing of the "suspected" structures requires more accurate spectral calculations that are possible only when detailed spatial characteristics are available, which implies availability of the atom coordinates or the bond distances for valence and torsion angles. Thus a certain hierarchy of algorithm complexity is obtained corresponding to the sequence of steps undertaken by the system to yield the final result and prove its unambiguity. This proof becomes possible because of the great differences in molecular spectra revealed by detailed analysis. Figure 1 illustrates the operating sequence of an artificial intelligence system for solving analytical problems from recorded molecular spectra.

This brief analysis of the artificial intelligence ideology shows that such systems are feasible only when high-level algorithms are available. For this purpose, it is necessary to develop a number of sections in the theory of molecular structure and spectra up to a quantitative forecasting level. Fortunately, considerable progress in this direction has been made. There is now a well developed theory of molecular structures and of molecular spectra, as well as quantum chemistry. Particularly favourable is the status of the theory of infrared spectroscopy which can be applied to detailed calculations of total spectral curves over the entire spectral range down to the lowest zone [15]. Other parts of the general theory of molecular spectra of different types are being constantly developed. This is why analytical systems built to the above pattern are already quite feasible.

A model of the artificial intelligence system for molecular spectroscopy

An operating model of an artificial intelligence system based on the above ideology has been described [12, 13]. The system uses data on vibrational, ultraviolet, nuclear magnetic resonance, and mass spectra separately and in their totality. Additionally, the empirical formula of the substance studied is introduced into the computer. It is assumed that the substance studied is a pure compound. The system is supplied with a data bank containing spectrostructural correlations for atom groupings, geometrical characteristics

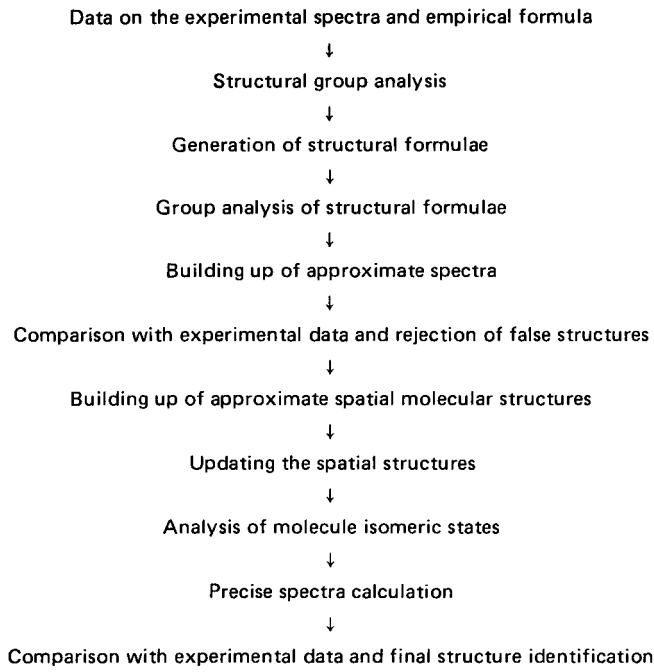


Fig. 1. Operating sequence of an artificial intelligence system for the interpretation of spectra.

of molecular fragments, and the parameters from quantum chemistry and spectroscopy required for calculating the geometry and spectra of the structures studied.

In the first stage, possible structural components are sought by using the spectrostructural correlation data stored in the bank. From these components, chemical rules and the empirical formula, the possible structural formulae are generated. These structural formulae form the initial set of hypotheses to be verified. Usually the initial set comprises dozens to several hundreds of molecules of average complexity. In some cases this initial set may also be obtained when the structural group analysis block has not isolated a single "suspected" structural component. In this case, the initial set is obtained by generation of structures based only on the empirical formula and chemical rules. Here, as in many other cases, the computer can be advised by data obtained in other experiments or the operator's intuition.

This step is followed by verification of the hypotheses. Structural components of hypothetical molecules are analyzed for their correspondence to the appropriate bank of characteristic structures. If any are found, additive rules are used to build up the approximate vibrational, n.m.r., and mass spectra of the corresponding molecular models. Comparison of these spectra with the experimental spectra, i.e. filtering, significantly reduces

the number of hypotheses to be verified. Essentially, even at this stage, as experimental runs of the system show, the result is almost unambiguous; at least the number of "suspected" molecular structures seldom exceeds 10. To verify so few hypotheses, more precise calculation techniques can be used. Many examples of the corresponding blocks of an overall program operating at this stage (STREC systems) have been given [12, 13].

In the subsequent stage, the remaining hypotheses are verified and complete validity of the structure identification is proved by more accurate calculation of spectra and their comparison with the experimental spectra. Here, spatial models of molecules are also built up and their vibrational and ultraviolet spectra are calculated. The calculation of the vibrational spectrum and comparison with the experimental spectrum are vital. The techniques of the structural chemistry, quantum chemistry and spectral theory are used. For the operation of this part of the program, stored data are required on geometric, quantum chemical and molecular spectral parameters of atoms and structural components of molecules. This allows not only the complete identification of the structure of the substance studied but also a solution to the rather intricate problems of isomeric origin (cis—trans isomers, rotational isomers, etc.) [13, 14].

This model of the artificial intelligence system takes care of the essential ideas. The experimental operation of this model for some years has proved the validity of these ideas and of the system as well as its reliability.

The artificial intelligence systems discussed here refer to a strictly defined category. Their operation is based on a number of basic axioms, including spectrostructural correlations, chemical rules of building structural formulae (rules of valency and forbidden combinations of atoms), vector models of atoms and structural components of molecules, and empirical parameters, required for calculating the geometry of molecules and their spectra. This list shows that the amount of formal data that must be introduced into the computer memory is far less than the number of spectra necessary for data retrieval systems. There may be up to several dozen quantum chemical parameters and several hundred other parameters. These parameters are of a more general and fundamental character than spectral curves. Such fundamental characteristics can include, for example, characteristics of potential surfaces of molecules used for calculating vibrational spectra, etc. Thus, the sets of initial axioms comprising the data banks of artificial intelligence systems will remain basically the same regardless of progress in either experimental spectral techniques or theories forming the basis of the algorithms. Moreover, the greater the progress, the less the number of initial independent parameters and the more fundamental they become. This could be very gratifying.

Since the spectra of molecular models are formed by the computation system itself, there should be no problem in generating not only spectra of components but also the spectra of their mixtures. This would open up the possibility of applying the systems discussed to the analysis of mixtures.

One more important aspect should be stressed. By one cycle in the artificial intelligence system, it is possible not only to identify a molecule but also to gain detailed information on its geometrical and electronic structure. The results can be applied to the solution of different problems of a wider scope than those formulated initially. Components of the artificial intelligence system can be applied efficiently to the solution not only of analytical problems but also of problems arising in various branches of the chemistry and physics of polyatomic molecules.

Conclusions

Artificial intelligence systems appear to be the most advantageous available for the future development of structure identification from spectroscopic data. These systems will not be able to surpass data retrieval systems, especially purpose-oriented ones, in the solution of specific problems in the immediate future. The success of artificial intelligence systems depends on the extent to which they can be made to imitate the human brain in drawing conclusions through complicated logic operations. Further concepts in the design of the analytical systems await development, and depend on experimental and theoretical developments in establishing molecular parameters. Self-learning systems will pose yet another challenge.

REFERENCES

1 L. A. Gribov and M. E. Elyashberg, Crit. Rev. Anal. Chem., 8 (1979) 111.
2 J. Zupan, Anal. Chim. Acta, 103 (1978) 273.
3 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
4 V. A. Koptjug, Z. Chem., 15 (1975) 41.
5 S. R. Heller, G. W. Milne and R. J. Feldmann, Science, 195 (1977) 253.
6 J. T. Clerc and J. Zupan, Pure Appl. Chem., 49 (1977) 1827.
7 J. Zupan, M. Penca, D. Hadzi and J. Marsel, Anal. Chem., 49 (1977) 2141.
8 N. A. Gray, Anal. Chem., 47 (1975) 2426.
9 S. Sasaki and H. Abe, Sci. Rep. Tohoku Univ., Ser. 1, 50 (1978) 157.
10 R. E. Carhart and D. H. Smith, J. Am. Chem. Soc., 97 (1975) 714.
11 G. Beech, R. T. Jones and K. Miller, Anal. Chem., 47 (1975) 714.
12 L. A. Gribov, M. E. Elyashberg and V. V. Serov, Anal. Chim. Acta, 95 (1977) 75.
13 L. A. Gribov, M. E. Elyashberg and M. M. Raikhshtat, J. Mol. Struct., 53 (1979) 81.
14 L. A. Gribov, M. E. Elyashberg and V. V. Serov, J. Mol. Struct., 50 (1978) 371.
15 L. A. Gribov, V. A. Dementiev and A. T. Todorovsky, J. Mol. Struct., 50 (1978) 389.

Analytica Chimica Acta, 122 (1980) 257—266
Computer Techniques and Optimization

GENERAL PRINCIPLES OF ALGEBRAIC MODELLING OF STRUCTURAL ORGANIC ANALYSIS†

G. G. SZÉKELY* and P. SZEPESVÁRY

Institute of Isotopes, Hungarian Academy of Sciences, P.O. Box 77, H-1525 Budapest (Hungary)

(Received 13th November 1979)

SUMMARY

A general theory is developed for modelling structural organic analysis, based on Boolean algebra and the theory of semi-lattices with certain simplifications. A linear system of Boolean equations is suitable for expressing the exact relationship of compositional and experimental information. The concepts introduced, and their unambiguous limits of applicability are discussed in terms of abstract algebra.

Inferences in qualitative chemical analysis always follow logical rules in a formal or informal way. If qualitative analysis is done by more or less automated systems, these rules must be formal. During the construction of an analytical system by Farkas and co-workers [1], there arose many problems whose solution demanded precise logical considerations.

The significance of formal logic in the formulation of the results of spectroscopic analyses has been recognized by Gribov and co-workers, who used predicate calculus in evaluating infrared spectra [2—4]. In order to handle and solve the logical problems of more or less automated qualitative analytical systems, an attempt is made here to build up a precise coherent mathematical apparatus, which also provides a good framework for studying basic problems in this field. During this work, it became clear that some very rigorous assumptions were needed to obtain a simple and manageable mathematical system. Accordingly, as a first step, the basic foundation of a general theory is described in this and later papers. In most cases, this theory will not be able to handle actual practical problems, but it should give a precise basis for later developments in this direction.

In the present paper, the basic definitions and lemmas (propositions) of the theory are described. In further papers, the necessary theorems and definitions for calculating the results of structural organic analyses using Boolean functions and equations will be given, and some theorems for the reverse problem, i.e. how to establish defined elementary entities of the analysis (e.g. functional groups) from an available collection of analytical

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

observations (e.g. spectral libraries), will be described. Finally, efficiency considerations regarding the possibilities of obtaining a unique solution for one measurement under certain conditions will be discussed.

In the present paper, two modes of presentation will alternate regularly. Firstly, the particular aspect is discussed informally and illustrated by examples; then it is formulated by means of mathematical definitions, lemmas and theorems. Most of the theorems shown in the mathematical parts are generally well known or special cases of well known theorems. Their proof is given in most cases in order to clarify the construction. The alternating general and mathematical parts can be read separately; the mathematical formulations are not essential for readers who are interested only in the applicability of the theories.

BASIC DEFINITIONS

General treatment

Very generally, any chemical analysis is a procedure linking the measured entity of the analysis with its properties [5]. In developing a mathematical model for the analysis, the entities of the analysis and the measured properties must be unambiguously described. There are many different ways of describing the entities. The way chosen here originates from the qualitative analysis of mixtures: in this case, there is a set of pure materials, and the entities of the analysis are qualitatively different mixtures of the pure species. In the general treatment, the term 'elementary entities' will be used; in this special case, it is synonymous with the pure species. In the general concept, the entities of the analysis will be characterized by the elementary entities contained in the entity.

Two basic restrictions follow from this concept of entity: (1) the quantity of the elementary entity contained in the entity of the analysis cannot be described, so that we can refer only to its presence or absence; (2) this mixture has no 'structure', i.e. the entity is simply a set of elementary entities, and there is no means of describing the connection between the elementary entities (i.e., elementary entities do not 'interact').

These requirements are trivial in the case of qualitative analysis of mixtures, but in the general case they are often restrictive. Despite this, the above type of representation is widely used in structural analysis systems, when the concept of functional groups is used.

For the description of properties, special binary properties called 'elementary properties' are used. They are binary, in the sense that they can have only two possible outcomes, true or false. A property will be characterized by the list of all elementary properties which proved to be true.

Given these concepts of description, the analysis can be described in the following way:

(i) fixing the set of all elementary entities;

(ii) fixing the set of all elementary properties;

(iii) making a list of all possible entities, which are all combinations of elementary entities;

(iv) giving the property of all entities (i.e. giving a list of all elementary properties proved true, for each entity).

Example 1. Define the qualitative analysis of mixtures containing alkanes up to 3 carbon atoms by binary mass spectra. (The spectra are taken from API project 44 tables [6]; the intensities are normalized by base peak intensity, and the threshold value for the presence/absence decision is 5% of the base peak intensity.)

(i) Elementary entities: methane, ethane, propane.

(ii) Elementary properties: presence of peaks at m/z 13, 14, 15, 16, 26, 27, 28, 29, 30, 38, 39, 41, 42, 43 and 44.

(iii) Entities of the analysis: M_1, methane; M_2, ethane; M_3, propane; M_4, methane + methane; M_5, methane + propane; M_6, ethane + propane; M_7, methane + ethane + propane.

(iv) Properties of entities:

M_1: 13, 14, 15, 16
M_2: 26, 27, 28, 29, 30
M_3: 15 26, 27, 28, 29, 38, 39, 41, 42, 43, 44
M_4: 13, 14, 15, 16, 26, 27, 28, 29, 30
M_5: 13, 14, 15, 16, 26, 27, 28, 29, 38, 39, 41, 42, 43, 44
M_6: 15 26, 27, 28, 29, 30, 38, 39, 41, 42, 43, 44
M_7: 13, 14, 15, 16, 26, 27, 28, 29, 30, 38, 39, 41, 42, 43, 44

Example 2. Define the entities and properties for the mass spectrometric detection of an alkane with not more than 3 carbon atoms.

(i) Elementary entities: CH_4 group, CH_3 group, CH_2 group

(ii) Elementary properties: see Example 1

(iii) Entities of the analysis: M_1, CH_4, methane; M_2, CH_3, ethane, CH_3-CH_3; M_3, CH_2, chemically meaningless; M_4, $CH_4 + CH_3$, chemically meaningless; M_5, $CH_4 + CH_2$, chemically meaningless; M_6, $CH_3 + CH_2$, propane, CH_3-CH_2-CH_3; M_7, $CH_4 + CH_3 + CH_2$, chemically meaningless.

(iv) Properties of entities:

M_1: 13, 14, 15, 16
M_2: 26, 27, 28, 29, 30
M_3: —
M_4: —
M_5: —
M_6: 15 26, 27, 28, 29, 38, 39, 41, 42, 43, 44

Chemically meaningless entities have no properties.

The procedure demonstrated in these examples is now generalized, in order to introduce the general notation used later.

(i) Suppose that there are n different elementary entities $a_1, a_2, \ldots, a_n$. Their set can be denoted by A: $A = \{a_1, a_2, \ldots, a_n\}$.

(ii) Suppose that there are m different elementary properties; their set can be denoted by T: $T = \{t_1, t_2, \ldots, t_m\}$.

(iii) The possible entities of the analysis are all possible combinations of $a_1, \ldots, a_n$. Here, 'possible' refers to logical rather than chemical possibilities. The set of these combinations is denoted by $\mathscr{P}(A)$, and is simply the power set of A. This power set has 2^n elements, namely: $\{a_1\}, \ldots, \{a_n\}$, $\{a_1, a_2\}$, $\{a_1, a_3\}, \ldots \{a_1, \ldots, a_n\}$.

(iv) For all elements of $\mathscr{P}(A)$, a list of elementary properties proved true is given; this is a subset of T, or, using the above notation, a member of set $\mathscr{P}(T)$.

Mathematical treatment

Definition 1. Let A be a finite set; its elements are the elementary entities. Let T' be a set which is not necessarily finite; its elements are the elementary properties. The function $R, R: \mathscr{P}(A) \rightarrow \mathscr{P}(T')$, is an analysis, if $dom\ R = \mathscr{P}(A)$. (For convenience, the notation used is summarized in Table 1).

Remarks. In the above definition, 'function' denotes any mapping between the two sets, for which any subset of A has a unique image in $\mathscr{P}(T')$.

The elements of $\mathscr{P}(A)$ are called as the entities of the analysis, so these are all possible subsets of A. One element of A, i.e. an elementary entity, is counted only once in a subset. This means that one entity of the analysis can be characterized exactly by stating whether or not it contains an elementary entity; but description of the multiple occurrence of an elementary entity is impossible. Similarly, for the subsets of T', the results of the analysis, represented exactly by the elements of $\mathscr{P}(T')$, can be characterized by stating which elementary properties prove to be true and which do not. It should be emphasized that $\mathscr{P}(T')$ denotes the set of finite subsets of T', regardless of the finiteness of T'.

The condition $dom\ R = \mathscr{P}(A)$ in the definition expresses the requirement that all entities must have properties. Empty property sets are allowed, of course.

The T' set introduced in Definition 1 differs from the set T shown in the general introduction of the notation. The difference can be illustrated as follows. In collecting the elementary properties in Examples 1 and 2, all existing m/z values could have been taken into account. If this had been done, set T' would have been obtained. However, it would be nonsense to use m/z values with no peaks present. This reduction is introduced by the following definition.

Definition 2. The set

$$T = \{t_i | t_i \in T' \text{ and there is a } X \subset A \text{ such that } t_i \in R(X)\} = \bigcup_{X \subset A} R(X)$$

is called as the reduced property set of the analysis R.

In consequence T is trivially a finite set, because $\mathscr{P}(A)$ and $R(X)$ for every X are finite (see Remarks after definition 1). Further, $im\ R \subset \mathscr{P}(T)$, and if for any $T^* \subset T'$, $im\ R \subset \mathscr{P}(T^*)$, then $T \subset T^*$. This follows immediately from the definition of T.

Remarks. In the following paragraphs, the reduced property set of the analysis R will always be used, whether or not this is emphasized. For the consequence above T is the closest subset of T', the power set of which contains $im\ R$.

TABLE 1

Notation used

$\{x_1 \ldots x_n\}$	description of a set by its elements
$\|X\|$	the number of elements of the set X
$\{x\|P(x)\}$	definition of a set by the property P depending on x
$x \in X$	x is contained in the set X
$A \subset B$	set inclusion (equality is also allowed)
$A \cup B$	union of set A and B
$\bigcup_{i=1}^{n} A_i$	union of the sets $A_1, \ldots, A_n$
$\emptyset$	empty set
X^n	the Descartes power of the set X
$\mathscr{P}(X)$	the power set of the set X
$\phi : X \to Y$	a mapping from the set X to the set Y
$\phi : x \mapsto y$	orders the element y to the element x
$dom\ \phi$	the set where the mapping ϕ is defined
$im\ \phi$	the image of the mapping ϕ
$<X, \cup, \emptyset>$	a semilattice over set X with operator $\cup$ and unit element $\emptyset$
$\phi\|_X$	the restriction of function ϕ to the set X
χ_X	the characteristic function of the set X
ϕ^{-1}	the inverse of the function ϕ
$\phi \circ \Psi$	the composition of the functions ϕ and Ψ (Ψ is applied first)
B_2	the two-element Boolean algebra $B_2 = \{0, 1\}$ with the usual Boolean operations

REGULAR ANALYSIS

General treatment

The concepts introduced above are very tedious to use in practice unless further assumptions are introduced. The number of entities can be too large to be managed even for relatively small numbers of elementary entities. However, in considerations of analysis of mixtures, the additivity of properties is very frequently assumed. In the binary case, additivity means that a mixture has an elementary property if, and only if, one of its components has this elementary property. If an analysis is additive in this sense, it will be called 'semiregular'. In the examples given above, it was assumed that for the 'empty mixture', i.e. for an analyzed entity which contained no elementary entity, all elementary properties were false, i.e., the property list is also empty. In reality, this is not necessarily so, but, as can be easily seen, if the analysis is semiregular (additive), the 'background' property list can be regarded as part of the properties of all entities of the analysis, and the background itself can be regarded as an empty list. For a semiregular analysis, if the empty mixture has no property proved to be true, it can be called a 'regular' analysis. In the following paragraphs, only regular analysis will be discussed.

The regularity in Example 1 is easily verified. In Example 2, the additivity is not valid; e.g. case M_4 can be regarded as the mixture of M_1 and M_2, but its property list is empty.

It is clear that any regular analysis can be exactly characterized by the property list of the special one-component mixtures, which contain only one elementary entity. Because of additivity, the property of any other mixture can be built up additively from the properties of the one-component mixtures which will subsequently be called basic mixtures. In Example 1, the three basic mixtures and their property lists are:

M_1: 13, 14, 15, 16
M_2: 26, 27, 28, 29, 30
M_3: 15 26, 27, 28, 29, 30, 41, 42, 43, 44

For example, mixture M_7 has all the peaks of methane, ethane and propane, and nothing else, showing additivity.

Mathematical treatment

Definition 3. The analysis $R:\mathscr{P}(A) \rightarrow \mathscr{P}(T)$ is said to be regular, if
(a) for every $X_1, X_2 \in \mathscr{P}(A)$, $R(X_1 \cup X_2) = R(X_1) \cup R(X_2)$ follows
(b) $R(\emptyset) = \emptyset$.

When (b) is not valid, the analysis is said to be semiregular.

In consequence, analysis R is regular only if $R: <A, \cup, \emptyset> \rightarrow <T, \cup, \emptyset>$ mapping between semilattices and unit element is homomorphism. Further, *im* R is a similattice for the operation $\cup$, and with unit element $\emptyset$, if R is regular, because it is the homomorphic image of a semilattice with unit element.

If R is regular, then

$$R(\bigcup_{i=1}^{k} X_i) = \bigcup_{i=1}^{k} R(X_i)$$

for every $X_i \in \mathscr{P}(A)$ $(i = 1, \ldots, k)$. It can be seen by induction from Definition 3.

A further consequence is that if R is regular, $A = \{a_1, \ldots, a_n\}$ and the notation $A_i = \{a_i\}$ is used (as is done below), then for every $X \in \mathscr{P}(A)$, where $X = \{a_{i_1}, \ldots, a_{in}\}$, the following is true

$$R(X) = R(\bigcup_{j=1}^{k} A_{i_j}) = \bigcup_{j=1}^{k} R(A_{i_j}).$$

This means that R can be exactly characterized by its values on the subsets $A_1, \ldots, A_n$; in other words, R is the homomorphic extrapolation to the set $\mathscr{P}(A)$ of the function $\bar{R} = R|_{\{A_1, \ldots, A_n\}}$.

Definition 4. The analysis $R:\mathscr{P}(A) \rightarrow \mathscr{P}(T)$ is called Boolean-homomorphic if R is homomorphism between the Boolean algebras $\mathscr{P}(A)$ and $\mathscr{P}(T)$.

Obviously, if analysis R is Boolean-homomorphic, then it is regular. Further, if analysis R is Boolean-homomorphic, then *im* R is a Boolean algebra.

DESCRIPTION OF A REGULAR ANALYSIS

General treatment

A regular analysis can be described more formally. Suppose that a number is assigned to each of the elementary entities from 1 to n, and to each of the

elementary properties from 1 to m. Then, instead of a list of all elementary properties proving true, an m-element Boolean vector can be constructed for each basic mixture. The i^{th} element of this vector is 1 if the i^{th} elementary property proves true, otherwise it is zero. When this procedure is followed for all basic mixtures, a Boolean matrix is obtained, which will completely describe the regular analysis.

For the case discussed in Example 1, this matrix is shown in Table 2. Analogously, any mixture can be described with the help of an n-element Boolean vector. The i^{th} element of this vector is 1 if the i^{th} elementary entity is present in the mixture, otherwise it is zero. For the case discussed in Example 1, these vectors are shown in Table 3. The logical vectors describing the possible properties can be constructed similarly.

For the example of a regular analysis with n elementary entities and m elementary properties, **R** will denote the Boolean matrix of the analysis, the structure of which is shown in Table 4. The vector describing the property of the entity **x** described by the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ can be obtained in the following way. If a + sign is used for the logical operation OR, and a · sign for the logical operation AND, the following equations will be true:

$$y_j = r_{1j} \cdot x_1 + r_{2j} \cdot x_2 + \ldots + r_{nj} \cdot x_n \quad (j = 1, \ldots, m)$$

because the j^{th} elementary property will be true only if it is true for at least one of the elementary entities present in the mixture described by vector **x**.

If the matrix notations of linear algebra are used, then the following general equation is obtained: **x R = y** (where **y** denotes the vector describing the properties), which gives the general connection between the entities of the analysis and its properties.

The following numerical example demonstrates the procedure described. Let n be 3 and m be 4. Suppose that the matrix of the analysis is

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

TABLE 2

Property matrix of a hydrocarbon analysis

Basic mixtures	Number and m/z values of the elementary property														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	13	14	15	16	26	27	28	29	30	38	39	41	42	43	44
M_1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
M_2	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0
M_3	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1

TABLE 3

Vectors of entities of a hydrocarbon analysis

Mixtures	Number and name of the elementary entity		
	1 Methane	2 Ethane	3 Propane
M_1	1	0	0
M_2	0	1	0
M_3	0	0	1
M_4	1	1	0
M_5	1	0	1
M_6	0	1	1
M_7	1	1	1

The properties of the particular entity x = (1 0 1) will then be given by the vector

$$(1 \quad 0 \quad 1) \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} = (1 \quad 1 \quad 1 \quad 1)$$

Mathematical treatment

Notation. The following notation is used:

$\mu(n,m)$ denotes the $n \times m$ matrices over B_2;

$\hat{\rho}(A,T)$ denotes the $\mathscr{P}(A) \to \mathscr{P}(T)$ analyses;

$\underset{\wedge}{\rho}(A,T)$ denotes the $\mathscr{P}(A) \to \mathscr{P}(T)$ regular analyses;

$\hat{\vartheta}(n,m)$ denotes the mappings $B_2^n \to B_2^m$;

$\vartheta(n,m)$ denotes the semilattice homomorphisms $B_2^n \to B_2^m$

Lemma 1. If $|A| = n$ and $|T| = m$, then there is a mapping $\phi : \hat{\rho}(A,T) \to \hat{\vartheta}(n,m)$ which is a bijection.

The proof follows. It is assumed that $R \in \hat{\rho}(A,T)$; $\phi(R)$ is defined as

$$\phi(R) = \chi_{\mathscr{P}(T)} \circ R \circ \chi^{-1}_{(A)} = \bar{R}$$

TABLE 4

The R matrix of the analysis

Elementary entities	Elementary properties			
	t_1	t_2	...	t_m
a_1	r_{11}	r_{12}	...	r_{1m}
a_2	r_{21}	r_{22}	...	r_{2m}
.	.	.		.
.	.	.		.
.	.	.		.
a_n	r_{n1}	r_{n2}		r_{nm}

Trivially, for all $R \in \hat{\rho}(A,T)$, $\tilde{R} \in \hat{\vartheta}(n,m)$. The mapping is superjective, because $\chi_{\mathscr{P}(A)}$ and $\chi_{\mathscr{P}(T)}$ are isomorphisms between Boolean algebras, so if $R \neq R'$, then

$$\phi(R) = \chi_{\mathscr{P}(T)} \circ R \circ \chi^{-1}_{\mathscr{P}(A)} \neq \chi_{\mathscr{P}(T)} \circ R' \circ \chi^{-1}_{\mathscr{P}(A)} = \phi(R')$$

and it is injective, because if $\tilde{R} \in \vartheta(n,m)$ then for $R = \chi^{-1}_{\mathscr{P}(T)} \circ \tilde{R} \circ \chi_{\mathscr{P}(A)}$, it is easily seen that $R \in \rho(A,T)$ and $\phi(R) = R$.

Lemma 2. The analysis R is regular only if $\phi(R) \in \vartheta(n,m)$; i.e. if ϕ is a bijection between $\rho(A,T)$ and $\vartheta(n,m)$ as well.

The proof of this proposition is readily induced from the definition of ϕ, because $\chi_{\mathscr{P}(A)}$ and $\chi_{\mathscr{P}(T)}$ are isomorphisms.

In the following paragraphs, the adjective 'regular' is also used for the elements of $\vartheta(n,m)$.

Theorem 1. There is a mapping $\psi : \rho(A,T) \to \mu(n,m)$, which is a bijection.

The proof follows. The elements of $\vartheta(n,m)$ are homomorphisms, and thus can be unambiguously characterized by their values on the unit vectors of B_2^n. In this case, if $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are the n unit vectors of B_2^n, and the mapping η: $\vartheta(n,m) \to \mu(n,m)$ is as follows

$$\eta(\tilde{R}) = \begin{pmatrix} \tilde{R}(\mathbf{e}_1) \\ \vdots \\ \tilde{R}(\mathbf{e}_n) \end{pmatrix} \quad \tilde{R} \in \vartheta(n,m)$$

then $\psi = \eta \circ \phi$ satisfies the conditions of the theorem.

It should be noted that the mapping $\tilde{R} \in \vartheta(n,m)$ can be handled as a function vector of m and n-variable Boolean functions, the variables of which are from B_2 Boolean algebra, i.e.

$$\tilde{R}(\mathbf{x}) = \begin{pmatrix} \tilde{P}_1(x_1, \ldots, x_n) \\ \vdots \\ \tilde{R}_m(x_1, \ldots, x_n) \end{pmatrix}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$.

In the following treatment, if $\alpha, x \in B_2$, αx will denote the usual conjunction operation of Boolean algebra. If $\mathbf{x} = (x_1, \ldots, x_n) \in B_2^n$ and $\alpha \in B_2$, then $\alpha\mathbf{x} = (\alpha x_1, \ldots, \alpha x_n)$.

Lemma 3. If $\tilde{R} \in \vartheta(n,m)$, then for every $\alpha \in B_2$ and $\mathbf{x} \in B_2^n$, $\tilde{R}(\alpha\mathbf{x}) = \alpha\tilde{R}(\mathbf{x})$.

The proof involves two possibilities for α. If $\alpha = 1$, then $\alpha\mathbf{x} = \mathbf{x}$ and $\alpha\tilde{R}(\mathbf{x}) = \tilde{R}(\mathbf{x})$, and so the statement is trivial. For $\alpha = 0$, $\tilde{R}(\alpha\mathbf{x}) = \tilde{R}(0\mathbf{x}) = 0 = 0\,\tilde{R}(\mathbf{x})$, because $\tilde{R}$ is regular.

The consequence of this proposition is as follows: if $\tilde{R} \in \vartheta(n,m)$, $\alpha, \beta \in B_2$ and $\mathbf{x}, \mathbf{y} \in B_2^n$, then $\tilde{R}$ is linear in the sense that $\tilde{R}(\alpha\mathbf{x}+\beta\mathbf{y}) = \alpha\tilde{R}(\mathbf{x})+\beta\tilde{R}(\mathbf{y})$. This is trivial from lemma 3 and the regularity of $\tilde{R}$.

Theorem 2. If $R \in \rho(A,T)$ and $\mathbf{R} = \psi(R)$ and $\mathbf{x} = \chi_{\mathscr{P}(A)}(X)$ and $\mathbf{y} = \chi_{\mathscr{P}(T)}(R(X))$ for any appropriate $X \in \mathscr{P}(A)$, then $\mathbf{y} = \mathbf{x}\,\mathbf{R}$. (Here, conjunction is used instead of multiplication and disjunction instead of addition in the usual matrix-vector production.)

The proof follows. As a consequence of lemma 2, there is a $\tilde{R} \in \vartheta(n,m)$ for which $\tilde{R} = \phi(R)$. Because of the definition of ϕ, R: $\mathbf{x} \mapsto \mathbf{y}$. The definition of R is $\mathbf{R} = \psi(R) = \eta(\tilde{R})$. However, for every $\mathbf{x} = (x_1,\ldots,x_n) \in B_2^n$, $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$, where $\mathbf{e}_i$ represents the unit vectors of B_2^n. In this case, application of lemma 3 gives

$$\mathbf{y} = \tilde{R}(\mathbf{x}) = \tilde{R}\left(\sum_{i=1}^{n} x_i \mathbf{e}_i\right) = \sum_{i=1}^{n} x_i \tilde{R}(\mathbf{e}_i) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \tilde{R}(\mathbf{e}_1) \\ \vdots \\ \tilde{R}(\mathbf{e}_n) \end{pmatrix} = \mathbf{x}\,\eta(\tilde{R}) = \mathbf{x}\,\mathbf{R}.$$

The consequence of theorem 2 is that, for every j, the mapping $\tilde{R}_j : B_2^n \to B_2$ (see discussion of Theorem 1) is a Boolean function, because $y_j = \tilde{R}_j(x_1,\ldots,x_n) = \sum_{i=1}^{n} r_{ij} x_i$ (for $j = 1,\ldots,m$), where $\mathbf{R} = (r_{ij})$, as stated in the previous theorem, and because this is a special canonical disjunctive form [7].

Conclusions

Many analyses intended to establish the qualitative characteristics of an investigated object can be described mathematically by mapping between properly defined entity and property sets. With some restrictive assumptions, a linear system of Boolean equations can express the relation between compositional and experimental information. Procedures for practical computing of the results of analyses of this type will be discussed in subsequent papers.

REFERENCES

1 G. Szalontai, Z. Simon, Z. Csapo, M. Farkas and Gy. Pfeifer, Anal. Chim. Acta., 133(1) (1981) in press.
2 M. V. Wolkenstein, L. A. Gribov, M. A. Elyashevitch and B. I. Stepanov, Vibrations of Molecules, 2nd edn., Nauka, Moscow, 1972. Appendix (in Russian).
3 L. A. Gribov and M. E. Elyashberg, J. Mol. Struct., 5 (1970) 179.
4 L. A. Gribov, M. E. Elyashberg and L. A. Moskovina, J. Mol. Struct., 9 (1971) 357.
5 H. Malissa, Fresenius Z. Anal. Chem., 271 (1974) 97.
6 Mass Spectral Data, API Research Project 44, College Station, Texas.
7 S. Rudeanu, Boolean Functions and Equations, North-Holland Publishing Co., Amsterdam, 1974.

Analytica Chimica Acta, 122 (1980) 267—277
Computer Techniques and Optimization

APPLICATION OF CORRELATION HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY TO THE REVERSE-PHASE SEPARATION OF TRACES OF CHLORINATED PHENOLS

H. C. SMIT*, T. T. LUB and W. J. VLOON

Laboratory for Analytical Chemistry, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam (The Netherlands)

(Received 4th December 1979)

SUMMARY

A correlation high-performance liquid chromatography instrument operating under on-line microprocessor control is described. Pentachlorophenol (PCP) and 11 other polychlorinated phenols are separated on a reverse-phase system. With an ultraviolet detector, the detection limit is about 2.5×10^{-7} absorption unit (about 6 μg PCP l^{-1} at 254 nm).

Correlation chromatography, essentially statistical by nature, is a powerful method in decreasing the detection limit in chromatography. The only physical modification needed to make a conventional chromatograph suitable for correlation chromatography is that of the injection system. The input flow of the chromatograph is switched between the sample and the background concentrations under control of a pseudo-random pattern. The chromatograph responds with a pseudo-random output as well. A cross-correlation function between the input pattern and the resulting output is evaluated; it has been proved [1, 2] that the function equals the response of the chromatograph to an impulse-shaped excitation at the input caused by the difference between the input concentrations, if certain conditions are met. Hence, the cross-correlation function is identical with a differential chromatogram of the input concentrations.

The most important advantage of correlation chromatography over straightforward chromatography is a rapid decrease in the detection limit in a relatively short time, because the noise of the chromatographic system is not cross-correlated with the input. Its contribution to the overall cross-correlation function converges to zero with increasing correlation time.

Several papers dealing with the theory of correlation chromatography have been published [1—5]. Applications to the analysis of gaseous mixtures by gas chromatography have been reported [2, 6]. In an earlier paper an application to h.p.l.c. was presented [3]; that experiment was intended to prove that correlation high-performance liquid chromatography (c.h.p.l.c.) was indeed possible and to demonstrate its potential. The input of the column was switched between background and sample by two electromagnetic

valves. In two hours, an improvement of the detection limit by a factor of 100 was achieved. However, the system was not very stable as a result of dilution of the sample in its holder during the correlation, and corrosion of the electromagnetic valves by the sample, which caused uncontrollable background absorption and clogging of the column. To overcome these difficulties, a c.h.p.l.c. instrument was constructed with total separation between the two analytical solutions and with a non-corrosive driving liquid in contact with the h.p.l.c. pump and electromagnetic valves.

Because the solute concentrations in both channels may be chosen at will, it is possible to perform c.h.p.l.c. in a differential mode. The chromatograph was used for the detection and separation of very low concentrations (down to 40 $\mu g\ l^{-1}$) of chlorinated phenols. The operation of the c.h.p.l.c. configuration was much facilitated by the use of a microprocessor-based control and data-processing unit [6b], which generates the pseudo-random binary sequence (PRBS) [8] input pattern and evaluates the cross-correlogram simultaneously.

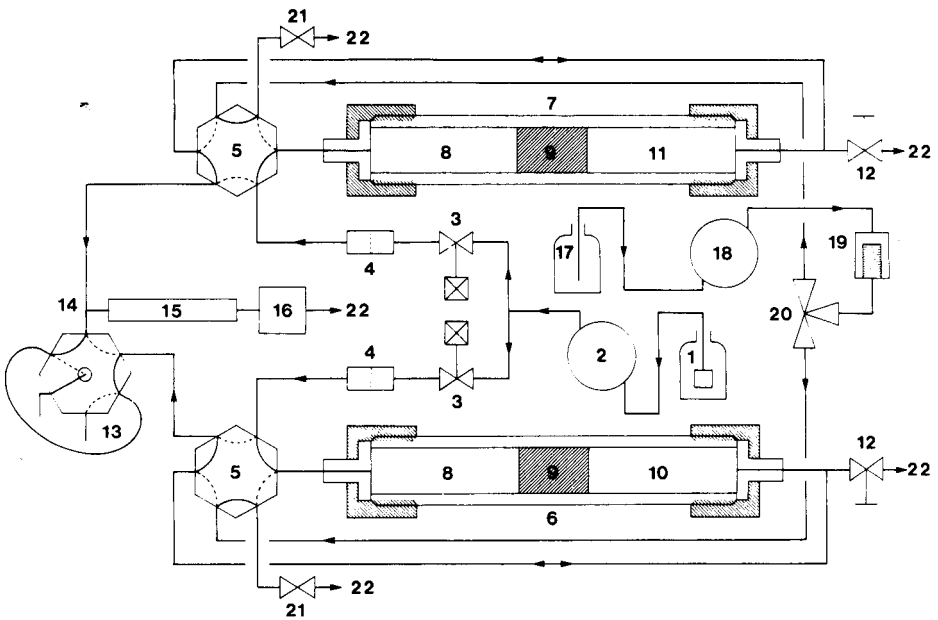


Fig. 1. Diagram of the chromatographic part of the c.h.p.l.c. instrument. (1) Reservoir of the driving liquid, 2 μm inlet filter immersed; (2) h.p.l.c. pump; (3) electromagnetic valves, controlled by the microprocessor; (4) in-line filters to retain particles released by the valves; (5) 6-way rotary valves; full lines show operating mode and dotted lines show filling mode; (6) background or eluent holder; (7) sample holder; (8) compartment containing driving liquid; (9) plungers; (10) compartment containing background solution or eluent; (11) compartment containing sample; (12) high-pressure open/close (column bypass) valves; (13) sample loop valve with needle port; full lines show inject position and dotted lines show load position; (14) low-dead-volume union-Tee; (15) h.p.l.c. column; (16) u.v. detector; (17) reservoir containing background (eluent) or sample; (18) low-pressure membrane pump; (19) in-line filter (2 μm); (20) three-way ball valve; (21) low-pressure open/close valve; (22) waste outlet.

DESCRIPTION OF THE C.H.P.L.C. CONFIGURATION

Figure 1 shows a diagram of the chromatographic part. The heart of the system is formed by two stainless steel sample holders (6, 7) of about 0.6 l each. One contains the sample, and the other the eluent or background solution. In both holders the analytical solutions are separated from and displaced by a driving liquid by means of an O-ring-sealed plunger (9). The driving liquid is displaced by an h.p.l.c. pump (2); its flow can be switched to either of the holders by two electromagnetic valves (3), which are controlled by the microprocessor. The outputs of the holders are alternately fed to an h.p.l.c. column (15) through a low-dead-volume union-T (14), controlled by the PRBS pattern generated by the microprocessor. The output of the column is monitored by a u.v. detector (16) (254 or 280 nm). The detector signal is sampled and digitized by an A/D converter under control of the microprocessor. The digitized signal is used for continuous evaluation of the cross-correlation function.

Several options have been added to this system to make it more manageable. A sample loop valve with needle port (13) is added to make conventional h.p.l.c. possible. Both sample holders can be switched from operating mode to filling mode by two 6-way rotary valves (5), and can be filled by an auxiliary membrane pump (18). Both can be emptied whilst bypassing the column through a high-pressure open/close valve (12).

The construction of the plungers is shown in Fig. 2. The body is made from Kel-F, a chlorofluorocarbon polymer. The permanent magnet (a magnetic stirrer was used) placed in the plunger body allows the plunger to be located outside the holder by means of a Hall element. This element is mounted in a hand probe together with two LED's; if a magnetic field is detected one LED is activated, depending on the direction of the field. A scheme of the electronics is given in Fig. 3. Of course an ordinary compass is a simple alternative.

The most marked property of the system is its operation by means of a driving liquid that does not enter the analytical part. Hence, the system can be easily used in a differential mode without the need to flush the h.p.l.c. pump whenever the background concentration is changed. Another very important consideration in this respect is the difficulty of obtaining high-pressure electromagnetic valves which are resistant to chemical corrosion. By choosing a non-corrosive driving liquid, corrosion of the valves is prevented and the lifetime of the seals and plungers of the pump is also prolonged.

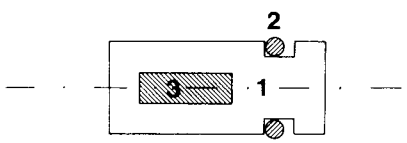


Fig. 2. Cross-section of a plunger used in the sample holders. (1) Kel-F body; (2) Kal-Rez O-ring; (3) permanent magnet.

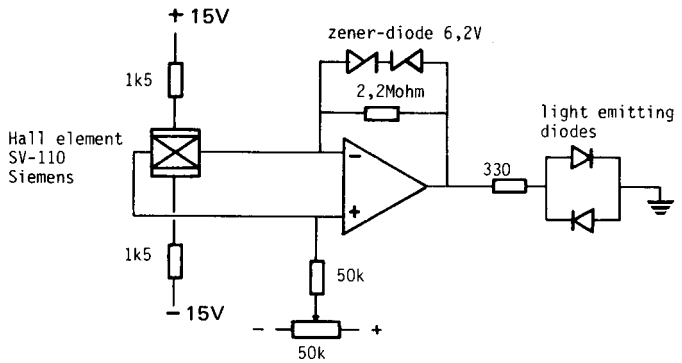


Fig. 3. The electronics used for the location of the plungers by means of their magnetic field.

The plungers in the sample holders can be sealed only with O-rings because the holders expand under pressure. The material from which the O-rings are made must be elastic as well as chemically inert. Kal-Rez (Dupont), a fluorocarbon elastomer, meets both conditions excellently.

The h.p.l.c. pump used should preferably be of a constant-flow type. Constant flow is very important in c.h.p.l.c. because variations of flow introduce non-stationarity which has a disastrous effect on the baseline noise of the correlogram [4]. For the same reason, the sample holders and column were kept at a constant temperature in an open water bath. Water was chosen as the driving liquid.

The microprocessor correlator is only briefly described; a detailed description will be given in a separate paper. It was designed for use by operators who have no particular skill in computer techniques and no detailed knowledge of the theoretical background of correlation chromatography. The parameters to be introduced by the operator are: expected length of the chromatogram up to and including the most retarded peak, standard deviation (width) of the first peak of analytical interest, and the number of PRBS sequences during which the correlation will run. The microprocessor responds with the best approximation of the input data (not every combination is allowed because of the mathematics of the PRBS). It generates the desired number of sequences and simultaneously correlates the output of the chromatograph with the input; intermediate results are displayed on a video screen. As soon as the desired number of sequences has elapsed, the correlation ends; the final result is displayed. Various simple operations can be made on these final data: enlargement of a part of the chromatogram, a simple baseline correction procedure, integration of peaks, and estimation of the standard error of the integral [9], based on a simplified model of the ACVF of the baseline noise. All data can also be output to a paper-tape punch and plotted on an x-t recorder.

SEPARATION AND DETECTION OF POLYCHLOROPHENOLS

The instrument was used to obtain chromatograms of mixtures of polychlorophenols (pentachlorophenol (PCP) and isomers of tetra-, tri-. and di-chlorophenol) at very low concentrations. PCP is used in wood preservation; the less chlorinated compounds are present as impurities or are formed in the environment. PCP pollution has been reported as an environmental problem [10]. It may contaminate soils as well as surface water.

An h.p.l.c. reverse-phase separation of several polychlorinated aromatic compounds has been reported [11]. In that work, 30% acetonitrile—water was used as eluent. The retention times of the more strongly retarded compounds (such as PCP) could be of the order of hours, even though the column temperature was raised to 60°C. Therefore a stronger modifier than acetonitrile had to be chosen; the sample and column were kept near ambient temperature to prevent chemical degradation of the sample. The separating system chosen was tetrahydrofuran—water (about 2 + 3 by volume) on Hypersil ODS (5 μm). The pH of the mobile phase was adjusted to a value about two units lower than the lowest single pK_a appropriate to the mixture to be separated. It is not possible to use tetrahydrofuran (THF) that has been "stabilized" against oxidation, because the stabilizer would greatly increase the background absorption and act as a modifier. Because the contraction of THF—water mixtures renders it difficult to prepare reproducible volume/volume compositions, the eluent was prepared by weighing the constituents.

Before mixing, the acidified water was filtered through a disc filter with a pore size of 0.2 μm (Sartorius SM 163 18 disc filter holder with filter type 11107, pressurized with helium). Both constituents were purged with helium before weighing and mixing, in order to minimize oxidation of the THF. Sample solutions were prepared by dilution of a stock solution with eluent. Eluent and sample were pumped into the system immediately after preparation; during this operation and all mixing operations they were kept under nitrogen. The stock solution was stored for longer periods under nitrogen in a freezer at −20°C.

Equipment and materials

A Spectraphysics 740 B or Orlita MK 00 h.p.l.c. pump was used with a Pro Minent B 2505 S pump as auxiliary for filling. The electromagnetic valves were Lucifer 121 A 54, with orifice 1.5 mm. The sample holders were 800 mm long (outer diameter 50 mm, inner diameter 33 mm) made of Sandvik 5R60 stainless steel; the increase in the inner diameter at 300 bar was 0.09 mm. The seals of the sample holders and the plungers in the holders were made of Kel-F, and the O-rings of Kal-Rez (Dupont), size ARP 317 or 215. The 6-way rotary valves were Rheodyne 70-10, and the sample loop valve with needle port was a Rheodyne 71-20. The column by-pass valves were Tescom 30-1101-204, the low-dead-volume union-Tree was a Swagelok SSLHT, and the three-way valve was a Whitey SS41XS2. The filter at the

output of the filling pump was a Nupro in-line filter, type 2F (2 μm). The low-pressure open/close valves were Whitey SS-OGS 2.

The column was a Chrompack 27668 model (length 25 cm, inner diameter 4.6 mm) packed with Hypersil ODS (Shandon). A Spectraphysics 8300 u.v. monitor was used.

The THF—water eluent was prepared by mixing doubly distilled water adjusted to pH 3 with perchloric acid, with THF (Merck LiChroSolv) in a ratio of 1:0.5701 (weight/weight).

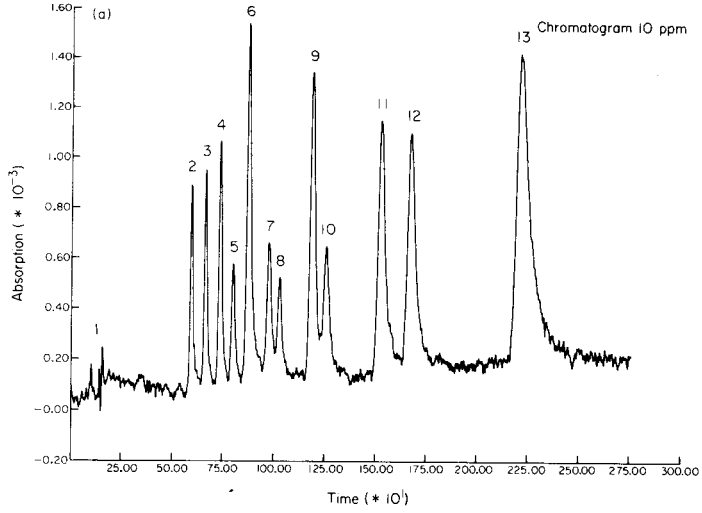
The polychlorophenols were "chemically pure".

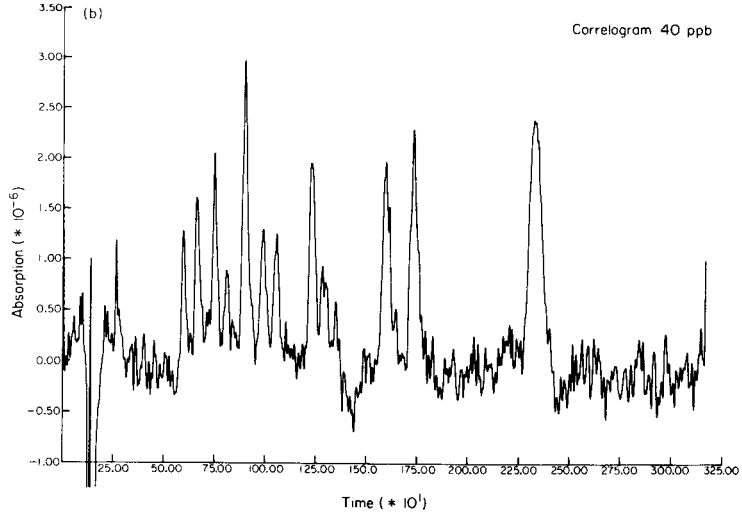
RESULTS

With the column and eluent used, the retention time of PCP could be kept well within 1 h at a flow rate of about 1.2 ml min^{-1}. Figure 4(a) shows a normal chromatogram of the stock solution diluted by a factor of 20 (all concentrations about 10 ppm), obtained with a 20-μl sample loop. Column temperature was not controlled; ambient temperature was about 22°C. The peaks present in this chromatogram are listed in Table 1. Figure 4(b) shows a correlogram of the stock solution diluted by a factor of 5000 (all concentrations about 40 ppb). Column and sample holders were kept at 21.9°C. The retention times are slightly different from those of Fig. 4(a), as a result of the different temperature and flow. Figure 4(c) shows the filtered correlogram of Fig. 4(b). The digital filter has a Gaussian shape; its bandwidth increases throughout the chromatogram, and at any point the width is equal to the expected peak width at that point. It is intended for cosmetic purposes only; the possible merits for improvement of the detection limit are yet to be investigated. Figure 5 shows a correlogram of a 1:1000 dilution (about 200 ppb) recorded at 254 and 280 nm.

All correlograms have large differential peaks near t = 150 s; these are probably caused by THF, its oxidation products and/or degradation products of the Kel-F plungers. The concentrations of these compounds in both sample holders are slightly different as a result of different treatment before filling. The detection limit reached in the 6-h correlation (Fig. 4(b)) seems to be of the order of 2.5×10^{-7} absorption unit (related to peak amplitude, not integration).

Fig. 4(a) Normal chromatogram of polychlorophenols. All concentrations are about 10 ppm (10^{-2} g l^{-1}); injected volume, 20 μl; column temperature, 22.3°C; pump, Spectraphysics 740 B; pressure, about 276 bar; flow rate, about 1.16 ml min^{-1}; elution time of dead volume (determined with $KCrO_4$), 92.0 s; full scale of u.v. absorption axis, 0.0018; wavelength, 254 nm. (b) Correlogram of polychlorophenols. All concentrations are about 40 ppb (4×10^{-5} g l^{-1}); virtual injection volume, about 124 μl; column temperature, 21.9°C; pump, Orlita MK00; pressure, 325 bar; elution time of dead volume, 87.6 s; sequence length, 3168.2 s; correlation time, seven sequences (6.16 h); clock period, 6.2 s; four digital samples taken per clock period, 2044 per sequence; full scale of u.v. absorption axis, 4.5×10^{-6}; wavelength, 254 nm. (c) As (b) after digital filtering by a Gaussian-shaped filter with optimum bandwidth for each data point, and first-order baseline correction.


(a)
Chromatogram 10 ppm
Absorption (* 10⁻³)
Time (* 10¹)
1
2
3
4
5
6
7
8
9
10
11
12
13


(b)
Correlogram 40 ppb
Absorption (* 10⁻⁶)
Time (* 10¹)

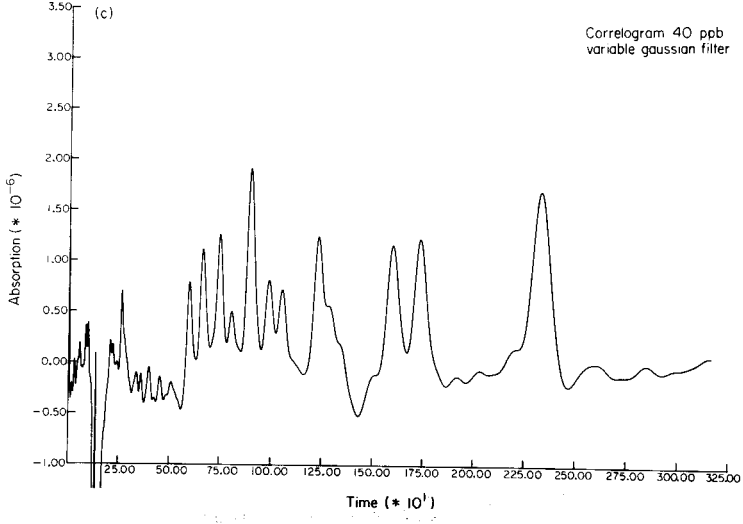

(c)
Correlogram 40 ppb
variable gaussian filter
Absorption (* 10⁻⁶)
Time (* 10¹)

TABLE 1

Listing of solutes present in the chromatogram of Fig. 4(a)

Peak no.	Solute	Capacity ratio	Concentration (ppm)
1	Impurities, THF	—	—
2	2,3-Dichlorophenol	5.52	10.5
3	2,6-DCP	6.34	12.6
4	3,4-DCP	7.07	10.9
5	2,5-DCP	7.78	10.0
6	2,3,4-Trichlorophenol	8.53	10.1
7	2,3,6-TCP	9.70	11.8
8	3,5-DCP	10.52	9.9
9	3,4,5-TCP	11.94	10.4
10	2,4,6-TCP	12.76	10.0
11	2,3,4,5-Tetrachlorophenol	15.52	10.1
12	2,3,5,6-TCP	17.21	10.3
13	Pentachlorophenol	24.96	10.4

EVALUATION

The concentrations of the chromatograms of Fig. 4(a) and (b) differ by a factor of 250. In theory [4], the detection limit (related to concentration of solutes) under the conditions leading to Fig. 4(b) is 525 times lower than it is in the case of Fig. 4(a). In this computation, the difference between the injected volume in Fig. 4(a) and the virtual injection volume [3] in Fig. 4(b) has been taken into account. If these volumes are the same, the expected factor of improvement is 85. In the computation of the expected factor of improvement between Fig. 4(a) and 4(b), it is assumed that the detector noise spectrum is band-limited white, which is not entirely true.

It is not easy to estimate the actual difference between the detection limits for Figs. 4(a) and 4(b). This is caused by the contribution of non-stationarity to the noise of Fig. 4(b) [4]. It is impossible to make a correct estimation of this contribution because it originates from uncontrollable variations of the flow, one or more of the input concentrations, or other system parameters. A reliable estimate of these detection limits could be made through the ACVF's of their baselines, but in this case these do not contain enough data to make a reliable ACVF.

Most chromatograms were recorded at 254 nm, which at first seems doubtful given an eluent that absorbs considerably at this wavelength and that yields oxidation products which have extremely high absorptivities at 254 nm. The next higher wavelength at which the u.v. monitor can operate is 280 nm; at this wavelength, THF and its oxidation products hardly interfere (Fig. 5(b)). However, PCP and the tetrachlorophenols have considerably lower absorptivities at 280 than at 254 nm, and the emission of the 280-nm fluorescent lamp is much lower than the emission of the uncoated mercury lamp at 254 nm. This causes the photodetector noise at 280 nm to be much

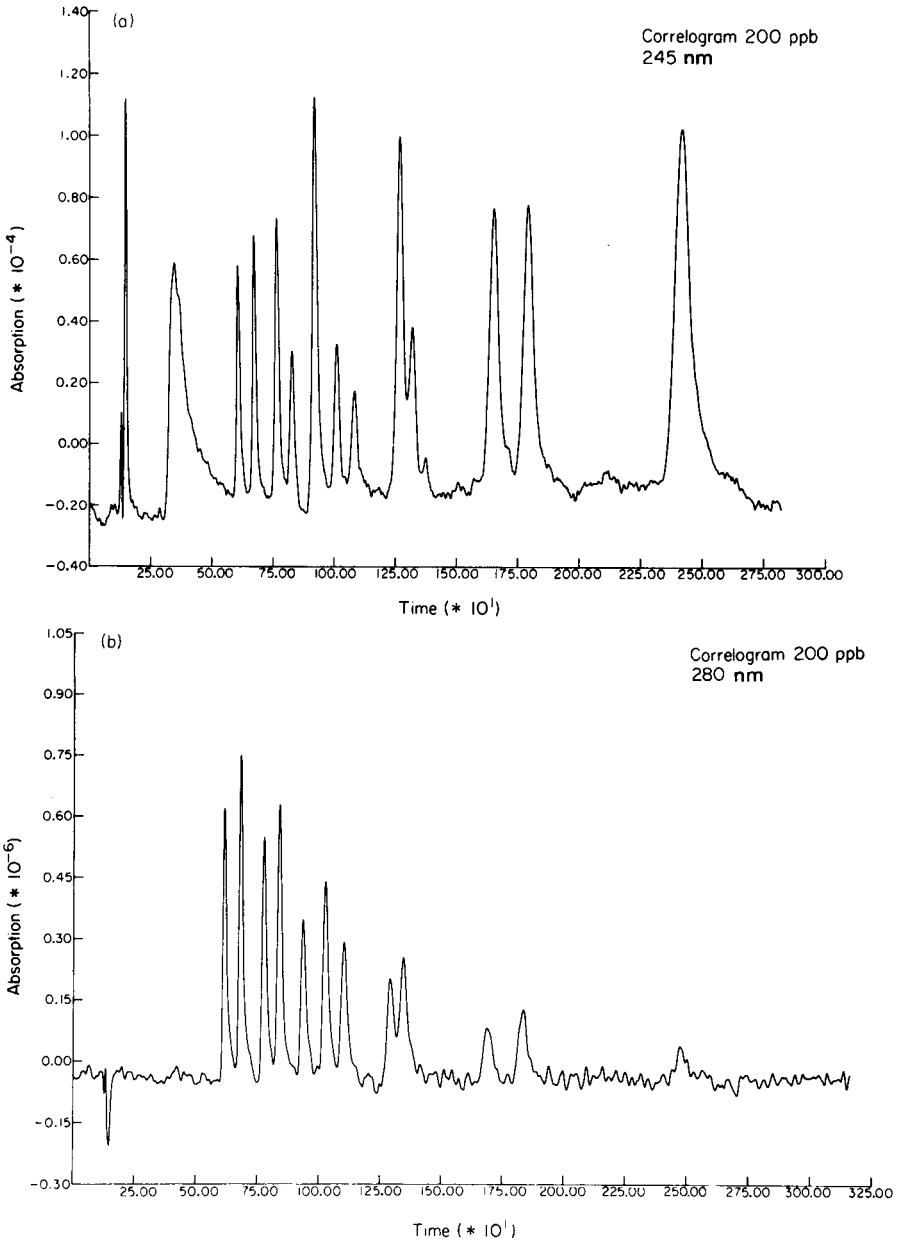


Fig. 5(a) Correlogram of polychlorophenols. All concentrations are about 200 ppb (2×10^{-4} g l^{-1}); nominal injection volume, about 110 μl; column temperature, 21.9°C; pump, Orlita MK00; pressure, 323 bar; elution time of dead volume, 87.2 s; sequence length, 2820.71 s; correlation time, three sequences (2.35 h); clock period 5.52 s; four digital samples per clock period, 2044 per sequence; full scale of u.v. absorption axis, 0.00018; wavelength, 254 nm. (b) As (a) except: wavelength, 280 nm; nominal injection volume about 124 μl; column temperature, 20.4°C; elution time of dead volume, 88.9 s; sequence length, 3168.2 s. Correlation time, four sequences (3.52 h); full scale of u.v. absorption axis, 0.00135.

higher than at 254 nm. Overall, PCP and the tetrachlorophenols are better detected at 254 nm, and the other compounds (whose peaks have much higher signal-to-noise ratios at 280 than at 254 nm) at 280 nm.

Much trouble caused by oxidation of the eluent could be eliminated by choosing a modifier that is less easily oxidized. This was not done because eluents prepared with such modifiers and giving the same eluting power would have a higher modifier content than a THF—water eluent. This is unimportant if the samples are artificially prepared. However, if an analytical sample drawn from an aqueous medium were to be prepared for c.h.p.l.c., the solute concentrations would be diluted by the amount of modifier added. For this reason, THF was preferred over weaker modifiers in spite of its unfavorable u.v. absorption properties.

The c.h.p.l.c. instrument can be used not only for the detection of very low solute concentrations, but also for the detection of very small differences between the solute concentrations in the background and the sample. These differences could be positive or negative, causing positive or negative peaks in the correlogram. This differential mode has not been used explicitly in the present work; the background concentrations of all polychlorophenols were zero. Nevertheless, the possibility of working differentially with only one column and one detector is indicated by Fig. 4(b): at lower retention times, several peaks appear which are the result of small differences in large concentrations present in both channels (such as THF, its oxidation products and, possibly, contamination from the corroding plungers).

Some general remarks concerning sample preparation for c.h.p.l.c. seem in order. In one important respect the procedures will differ from those used in normal h.p.l.c.: pre-concentration of the solutes of interest is not essential. This avoids the introduction of errors caused by the recoveries in the concentration procedures being less than 100%. A clean-up procedure may be necessary in c.h.p.l.c. as well as in h.p.l.c., and for the same reasons. Besides, it is advisable to keep the ratios of the peak surfaces in c.h.p.l.c. as close to unity as possible, because small peaks disappear in the contribution of large peaks to the baseline noise, if the system is not stationary. Solutes causing large peaks could be masked by making their concentrations in the background and the sample equal to each other.

For the future development of c.h.p.l.c., several improvements of the configuration described in this paper are proposed. The search for and/or development of a flow-feedback-controlled h.p.l.c. pump should be continued. All Kel-F parts should be replaced by PTFE replicas, because Kel-F appears to be slightly corroded by the eluent used. All stainless steel surfaces in contact with eluent or sample should be replaced by glass, quartz or PTFE. The dead volume in front of the plungers when they are at the farthest right position (Fig. 1) should be minimized; an option to flush this dead volume should be added.

It is hoped that this work will contribute to the development of c.h.p.l.c. as a fully operational analytical technique.

The authors thank Mr. H. J. Harmsen, Prof. Dr. O. Hutzinger, Mr. R. Logchies and Mr. H. Steigstra for their valuable contributions.

REFERENCES

1 K. Izawa, K. Furuta, T. Fujiwara and N. Suyama, Ind. Chim. Belge, 32 (1967) 223.
2 H. C. Smit, Chromatographia, 3 (1970) 515.
3 T. T. Lub, H. C. Smit and H. Poppe, J. Chromatogr., 149 (1978) 721.
4 T. T. Lub and H. C. Smit, Anal. Chim. Acta, 112 (1979) 341.
5 M. Kaljurand and E. Küllik, Chromatographia, 11 (1978) 328.
6 (a) M. Kaljurand and E. Küllik, in A. Zlatkis (Ed.), Proc. 14th Int. Symp. on Advances in Chromatography, Lausanne, 1979, p. 173; Chrom. Symp., Dept. Chem., Univ. Houston, TX, U.S.A. (1979).
(b) H. C. Smit, R. P. J. Duursma and H. Steigstra, to be published.
7 M. Kaljurand and E. Küllik, J. Chromatogr., 171 (1979) 243.
8 A. C. Davies, IEEE Trans. Comput., C-20 (1971) 270.
9 H. C. Smit and H. L. Walg, Chromatographia, 8 (1975) 311.
10 R. H. Pierce Jr., EPA Rep. No. 600/3-78-063 (July 1978), Nat. Tech. Info. Serv., Springfield, VA, U.S.A.
11 C. L. Bramlett, HPLC of chlorophenoxyacetic acids, polychlorinated phenols, dinitrobutylphenol, dinitrocresol, and pentachloroanisole; Report on SARAP project No. 52-74 (1975), U.S. Govt. Printing Office 1976-622-657/416.

Analytica Chimica Acta, 122 (1980) 279—290
Computer Techniques and Optimization

ALGORITHMS FOR HIGH-LEVEL DATA PROCESSING IN GAS CHROMATOGRAPHY†

Z. HIPPE*, A. BIEROWSKA and T. PIETRYGA
The I. Łukasiewicz's Technical University, 35-959 Rzeszów (Poland)
(Received 15th October 1979)

SUMMARY

Computer processing of gas chromatographic data is discussed, with special emphasis on the numerical operations needed. The CHADIC program is described. This program enables optimal separation conditions to be determined and isothermal gas chromatograms to be processed for qualitative and quantitative analysis.

Gas chromatography (g.c.) is probably the most extensively used technique in analytical chemistry. Characteristically, a huge number of chromatograms is supplied in a working day, and the bottleneck created by handling all these records requires computational aids. Various devices for automatic g.c. data handling, e.g., electromechanical or electronic analog integrators [1—5] and digital integrators [6—10], are available, and automation by application of digital computers with hardware configurations such as satellite [11], time-sharing [12] or multi-access systems [13] has been in use for some years. Many dedicated hardware/software g.c. systems are now commercially available. Depending on the techniques used for coupling, the computer with the gas chromatograph, the tasks of the algorithms and the programs may differ somewhat in different systems in terms of subordinate functions, but the basic functions remain practically the same. The normal operations include numerical filtration (smoothing), baseline correction, recognition of isolated and overlapping peaks, resolution of partly overlapping peaks, estimation of retention parameters and standard calculations of sample content, and optimization of the separation process itself. These functions are performed with various degrees of correctness and efficiency.

Processing of g.c. data is complicated and requires the resolution of instrumental problems in analog-digital (A/D) conversion, as well as advanced algorithms and computer programs. This is particularly true when the data have to be interpreted in considerable depth, matching the level of precision and accuracy offered by modern chromatographs. The present status of algorithms for high-level processing of g.c. data is surveyed in the following paragraphs.

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

Numerical filtration (smoothing)

The raw g.c. data are always contaminated by various distortions: comparatively large, spurious signals may originate from the chromatograph itself (e.g. flip-flop operation) and A/D conversion unit (peripheral errors), and noise is inevitable. Rejection of outlying results may be achieved routinely by intensity tests or area tests, but probably the most promising procedure is the approach of Nalimov [14].

Some studies of noise in g.c. data [15] suggest the possibility of treating noise as an ergodic random process with an expected mean value of zero and finite variance. Thus linear numerical filters can be applied

$$\tilde{y}_i = (a_m y_{i-m} + \ldots + a_0 y_i + \ldots + b_1 y_{i+1} + \ldots + b_m y_{i+m})/(a_m + \ldots + a_0 + b_1 + \ldots + b_m) \quad (1)$$

where $\tilde{y}_i$ and y_i denote the filtered and raw signals, respectively; a_0, a_j, $b_{j\,|j=1,\ldots,m}$ are coordinates of the characteristic vector of the particular filter; and i is the sampling counter.

Table 1 summarizes some results of the evaluation of various linear filters used widely for smoothing g.c. data. The second filter displays advantageous properties: its systematic error is comparatively low, whereas its smoothing efficiency is satisfactory [16]. It must be remembered that effective smoothing imposes some conditions on the number of measured points per peak. There are differing views in this respect: Baumann et al. [17] suggest at least 8 points per peak; other workers [18—20] found 13—15 points to be useful, whereas still others [21—24] show that 20—25 points or 40—75 points [25, 26], are necessary for full and precise peak description. A total of 30 points per peak seems to be limiting for full characterization of a peak [16].

Baseline correction

Baseline correction is necessary for estimation of the start and finish of g.c. peaks. This in turn permits calculation of true peak area and normalization of the signals. Various types of baseline correction can be differentiated [27—30]. Most frequently, the correction involves two distinct steps; first, the initial approximation of parameters for the drift component in the elution curve is established; secondly, the data are refined by the least-squares method. In a recent approach [31], local minima are sought in the chromatogram $\{y_i\}_{i=1,\ldots,n}$, where n is the number of data points in the chromatogram (see Fig. 1); those minima should be situated between two clearly separated peaks in the band. (Two neighbouring peaks having half-height widths w_1 and w_2, respectively, are resolved when $\Delta t_r/(w_1 + w_2) > 1.28$, where Δt_r is the difference between retention times.) Thus

$$<mih + p - \Delta y, mih + p + \Delta y> \quad (2)$$

where $m = (y_n - y_1)/[(n-1)h]$ and $p = y_n - mh$. Here, Δy denotes the mean standard deviation of the noise, and h is the sampling step. Thus, the array $\{N_{k,j}\}_{\substack{k=1,\ldots,K \\ j=1,2}}$ (see Fig. 1.) is calculated. $N_{k,1}$ and $N_{k,2}$ represent the

TABLE 1

Evaluation of principal linear filters[a]

No.	Filter type	Filter systematic error	Smoothing efficiency
1	$\tilde{y}_i = [-3(y_{i-2} + y_{i+2}) + 12(y_{i-1} + y_{i+1}) + 17y_i]/35$	$\frac{3}{35}h^4\phi_i^{(IV)}$	$\frac{6h^4}{35\bar{N}}\sum\omega_i^{(II)}$
2	$\tilde{y}_i = [-(y_{i-2} + y_{i+2}) + 4(y_{i-1} + y_{i+1}) + 6y_i]/12$	$h^4\phi_i^{(IV)}$	$\frac{24h^4}{\bar{N}}\sum\omega_i^{(II)}$
3	$\tilde{y}_i = [-2(y_{i-3} + y_{i+3}) + 3(y_{i-2} + y_{i+2}) + 6(y_{i-1} + y_{i+1}) + 7y_i]/21$	$0.4h^4\phi_i^{(IV)}$	$\frac{0.8h^4}{\bar{N}}\sum\omega_i^{(II)}$
4	$\tilde{y}_i = [-21(y_{i-4} + y_{i+4}) + 14(y_{i-3} + y_{i+3}) + 39(y_{i-2} + y_{i+2}) + 54(y_{i-1} + y_{i+1}) + 59y_i]/231$	$1.25h^4\phi_i^{(IV)}$	$\frac{2.5h^4}{\bar{N}}\sum\omega_i^{(II)}$

[a] ϕ_i denotes the exact value of the g.c. curve at point i; ω_i is the noise value at point i, and h is the sampling step.

start and finish of baseline fragments; there are k fragments altogether. Further, the equation for each fragment is calculated from the general rule:

$$m_k = (y_{Nk,2} - y_{Nk,1})/[(N_{k,2} - N_{k,1})h],$$
$$p_k = y_{Nk,1} - m_k N_{k,1} h \text{ (for } k = 1,...,K) \tag{3}$$

(on the assumption that the baseline consists of linear segments). This approach is characterized by small errors in peak-area and peak-height calculations. The algorithm described is unusually general, and can be merged with any data system for any experimental conditions.

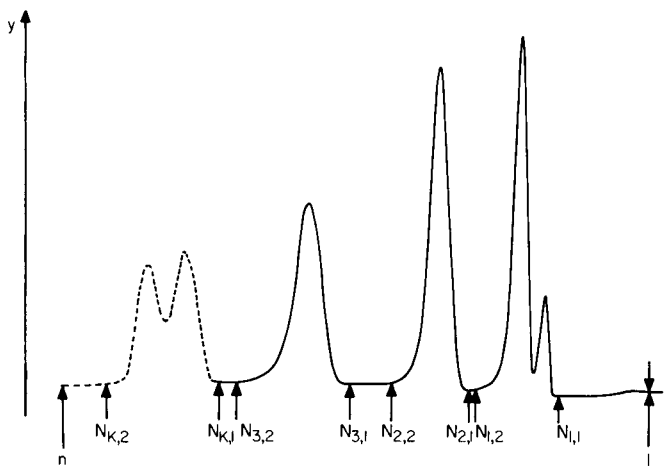


Fig. 1. Baseline correction.

RECOGNITION OF ISOLATED AND OVERLAPPING PEAKS

This recognition process is of great importance, because in chromatographic practice it is often impossible to achieve separation conditions which yield completely isolated peaks. Thus, multifunctional computer systems for processing g.c. data must contain subroutines for recognition of isolated and overlapping peaks and for separation of unresolved complexes into the correct number of component peaks. Critical reviews of this problem indicate that, especially in commercial g.c. data systems, expensive though they are, the algorithms used are very unsatisfactory, e.g. the drop-line method [32, 33]. In this situation, the results of splitting are usually marred by serious errors, unless the proper initial information (so-called zero approximation) can be settled.

The identification of overlapping peaks having distinct maximum signals is not very difficult. In such cases, the number of maxima is taken as the number of component peaks. Far more awkward and difficult is the frequently encountered case where component peaks must be recognized in a complex contour which has only some shoulders.

With regard to machine identification of the type of peaks in g.c. curves (this is really a sophisticated case of artificial intelligence), there are two recent general approaches. The first approach, which involves numerical integration, is based on the observation that the asymmetry and excess coefficients of an isolated peak are similar to the corresponding coefficients in the mathematical model of the peak. The asymmetry (S) and the excess coefficient (E) are described by: $S = m_3/(m_2^3)^{1/2}$, and $E = (m_4/m_2^2) - 3$ where m_p denotes for $p > 1$ the moment of the order p

$$m_p = \sum_{i=N_1}^{N_2} (ih - m)^p y_i / \sum_{i=N_1}^{N_2} y_i \qquad (4)$$

where $m = h \cdot \sum_{i=N_1}^{N_2} i \cdot y_i / \sum_{i=N_1}^{N_2} y_i$

In these formulae, N_1 and N_2 denote the serial numbers of the starting and closing points of a peak, respectively. This method can be used only when the values of the asymmetry and excess coefficients of the mathematical model are constant, a condition which is fulfilled when a peak can be represented by a Gaussian function. However, as this function usually approximates real peaks unsatisfactorily, this approach shows low efficiency.

In the second approach, the convexity of the chromatographic peak is established from the sign of the second derivative of the function describing the peak. This method is efficient only when, under defined experimental conditions, the differences in the retention times of sequential overlapping peaks have a distinct influence on the convexity of the g.c. curve. This approach enables any g.c. signals to be assigned as isolated or overlapped. This is based on the preliminary mathematical proposition (or lemma): if the

array $\{y_i\}_{(i=N_{k,1},\ldots,N_{k,2})}$ describes exclusively an isolated peak, then the natural constants n_1 and n_2, where $N_{k,1} < n_1 < n_2 < N_{k,2}$, are such that for the set $(N_{k,1}h,\ n_1h) \cup (n_2h,\ N_{k,2}h)$, $\{y_i\}$ is a convex function, whereas, for the set $(n_1h,\ n_2h)$, $\{y_i\}$ is a concave function. This means that the detection of two deflection points for the array $\{y_i\}$ proves the isolated character of the peak tested, whereas a larger number of deflection points detected suggests an unresolved complex of peaks.

Identification of deflection points solely on the basis of analysis of derivatives of g.c. curves is very difficult, because the signals $\{y_i\}$ are distorted by noise from the detector itself and by errors of A/C due to conversion. Hence, these methods require multiple filtration of the second derivative of the chromatogram, which in turn extends the identification of unresolved peaks. Results obtained in this laboratory indicated, for instance, that M-fold filtration of a curve represented by n discrete points by means of a numerical filter described by k-point convolution rule, requires $M \cdot n$ entire $(k/2)$ multiplications [34]. It should be emphasized that the value of the M should be at least 20 for satisfactory results.

These drawbacks of both approaches to peak character recognition, as well as the random distribution of g.c. noise led to the development of a novel method of fast recognition of peak type [34]. In this method, the relationship between the convexity of an isolated peak and the monotonic nature of its first derivative is used to find the most probable deflection points. The subroutine based on this idea is used both to recognize unresolved peaks and to determine the number of component peaks in the contour.

Identification of isolated and/or unresolved peaks is done sequentially in intervals $<N_{k,1}h,\ N_{k,2}h>$ of the chromatographic signals $\{y_i\}_{(i=1,\ldots,n)}$, where k represents the number (position) of the processed segment of the chromatogram (see Fig. 1). Whether the tested segment $\{y_i^{(k)}\}_{(i=N_{k,1},\ldots,N_{k,2})}$ is a collection of signals representing an isolated peak or comprises n unresolved peaks, the first stage of identification is done by location of maxima and minima in the segment. Further identification is accomplished by counting the number of deflection points in the intervals over which the function is monotonic (decreasing or increasing). In sub-intervals, the deflected line is passed through points i and $(i + t)$ of the real chromatogram and the increments at the extremes are calculated from $Z_j = |y_{i+t} - y_i|$, where $t = 3, 4, 5$. Because the values of these increments can be falsified by distortions of chromatographic signals, the true extremes are those which are determined by deflection points, satisfying the condition

$$|Z_{\max\, i} - Z_{\min\, i}| > a_1 \wedge |Z_{\max\,(i+1)} - Z_{\min\, i}| > a_1 \qquad (5)$$

In this notation, $a_1 = i_1 dy_1$, but $i_1 = 5, 6, 7, 8, 10, 12, 15, 20$, and $dy_1 = 0.001$ are connected with measurement errors in the chromatographic signals. From these calculations is obtained the matrix $\{M_i\}_{(i=1,\ldots,ei)}$ of intrinsic maxima $\{Z_i\}_{(i=1,\ldots,i)}$.

In the next stage of identification of unresolved peaks, the number of

deflection points for the given sub-interval is obtained by statistical estimation of the number of maxima, *ei*, determined for appropriate variants $t = 3, 4, 5$ and $a_1 = 0.001\ i_1$ (for $i_1 = 5, 6, 7, 8, 10, 12, 15, 20$), by creation of the matrix $\{DEC_l\}_{(l=1,\ldots,p)}$. The matrix determines the size of variants t and a_1 which have $ei = l$ deflection points, whereas the l^{th} element of the matrix DEC which fulfills the condition $DEC_l = \max_l \{DEC_l\}$, determines the number of deflection points in the estimated sub-interval of the segment $\{y_i^{(k)}\}_{(i=N_{k,1},\ldots,N_{k,2})}$. The number of shoulders on the contour follows from the equation, $l = l - 1$. In this way, the number of shadows and the number of maxima found previously in the $\{y_i^{(k)}\}$ segment analyzed forms the basis for a decision if these signals contain an unresolved peak complex or an isolated peak. Thus, one maximum and no shoulders suggests an isolated peak, whereas more than one maximum and shoulders indicates an unresolved peak complex. The total number of component peaks in an unresolved complex equals the number of maxima and shoulders detected.

The method described was tested on a large data set comprised of complex contours with different numbers of component peaks and various degrees of overlapping. It was found that the subroutine yields accurate results and may be regarded as versatile. In comparison with methods based only on differentiation of the g.c. curve, computer time was considerably reduced because fewer iterations were required for identification.

In these problems of separation of unresolved peaks, it should be stressed that, apart from estimation of peak character, other questions remain to be solved, mainly with respect to the mathematical model describing the shape of an isolated peak and a suitable algorithm for separation.

Table 2 summarises peak models discussed in the literature. Formally, function (5) of Table 2 approximates an isolated peak most accurately. Unfortunately, this function cannot be recommended as the resolution step requires lengthy computing time. In this respect, function (8) of Table 2 is satisfactory, giving reasonable execution time and minor changes in peak parameters (area ca. 0.7—3.6%, height ca. 0.0—1.2%, half-height width ca. 0.3—5.4% depending on the size and asymmetry of the peak). Other advantageous features of function (8) are the limited number of parameters needed to describe the function and the possibility of calculating the first and second moments of the peak area directly from equations rather than numerically.

With regard to the separation algorithm, the procedure normally used here involves a least-squares method to minimize the function

$$\phi(a_1^{(j)},\ldots,a_k^{(j)}) = \sum_{i=1}^{n} [y_i - \sum_{j=1}^{r} F_j(t_i,a_1^{(j)},\ldots,a_k^{(j)})]^2 \qquad (6)$$

in which $\{y_i\}_{(i=1,\ldots,n)}$ represents the (discrete) shape of a peak having r components, and $\sum_{j=1}^{r} F_j(t_i,a_1^{(j)},\ldots,a_k^{(j)})$ is its mathematical model. However, two essential conditions must always be met: the mathematical model applied for an isolated peak has to be suitable for calculation of derivatives, and the

initial values of parameters $[a_1^{(j)}\ldots,a_k^{(j)}]$ should fall within rather a narrow region of convergency to shorten the iteration process of eqn. (6). The difficulties in fulfilling such conditions explain why rather primitive unsatisfactory algorithms for peak resolutions such as the drop-line method, are normally used in g.c.

Results obtained here suggest usage of the function

$$Q(A^{(j)}b_1^{(j)}b_2^{(j)}dr^{(j)}) = \sum_{i=1}^{m}[y_i - \sum_{j=1}^{r} F_j(i\cdot h, A^{(j)}b_1^{(j)}b_2^{(j)}dr^{(j)})]^2 + (P-\hat{P})^2 + (M_1-\hat{M}_2)^2 + (M_2-\hat{M}_2)^2 \quad (7)$$

where h is the sampling step, function F is the mathematical model of the j^{th} unresolved complex with r components defined as

$$\begin{Bmatrix} A^{(j)}\exp[-b_1^{(j)}\cdot(i\cdot h - dr^{(j)})^2], & i\cdot h \leqslant dr^{(j)} \\ A^{(j)}\exp[-b_2^{(j)}\cdot(i\cdot h - dr^{(j)})^2], & i\cdot h > dr^{(j)} \end{Bmatrix}$$

$$\hat{P} \approx \text{const}\sum_{j=1}^{r} A^{(j)}[(b_1^{(j)})^{1/2} + (b_2^{(j)})^{1/2}]$$

$$\hat{M}_1 \approx \text{const}\sum_{j=1}^{r} A^{(j)}\{[dr^{(j)} - \text{const}'(b_1^{(j)})^{1/2}](b_1^{(j)})^{1/2} + [dr^{(j)} + \text{const}'(b_2^{(j)})^{1/2}]/(b_2^{(j)})^{1/2}$$

$$\hat{M}_2 \approx \text{const}\sum_{j=1}^{r} A^{(j)}\{[dr^{(j)} - \text{const}'(b_1^{(j)})^{1/2}]^2/(b_1^{(j)})^{1/2} + [dr^{(j)} +$$
$$+ \text{const}\,(b_2^{(j)})^{1/2}]^2/(b_2^{(j)})^{1/2} + \text{const}''[b_1^{(j)})^{1/2}/b_1^{(j)} + (b_2^{(j)})^{1/2}/b_2^{(j)}]\}$$

$\hat{P}$, $\hat{M}_1$ and $\hat{M}_2$ represent the area and the first and second moments of the mathematical model, respectively, whereas P, M_1 and M_2 denote the corresponding parameters for the approximated peak complex. To minimize eqn. (7), the algorithm of Hooke and Jeeves [35] was used with special modifications [36] which make it possible to estimate the initial values of the parameters sought, so that even the first step of the iteration is effective; details of such programs are available from the authors. These algorithms display some progress in the evaluation of overlapping peaks: the initial values of peak parameters need not be estimated very precisely, which is particularly important when overlapping is fairly large; and the parameters of component peaks can be estimated within reasonable error. The computing time depends mainly on the number of iterations, L, which is in turn connected with the number of component peaks, r, in the peak complex

$$L = 32r\max(\vec{X}_b[K]/d[k])$$

where $\mathbf{X}_b$, $\mathbf{d}$ are the vectors of initial approximation and corrections. The actual computing time is no greater than that needed for any other separation method described in the literature.

OTHER FEATURES

Estimation of retention parameters and standard calculations of sample content are too trivial to be discussed here. A recent option in algorithms for

TABLE 2

Characteristics of various mathematical models for an isolated g.c. peak.

No.	Model	Approximation error (%)		Computing needed for separation of r unresolved peaks
		Stated in literature	Present results	
1	2	3	4	5
1	$y(t,t_0,w,A) = A \cdot \exp\,[-(t-t_0)^2/2w^2]$	≈ 6	$\gg 6$	Solution of $3r$ transcendental equations
2	$y(t,t_0,w,A) = A(t_0/t)^{1/2} \exp[-2t_0(t^{1/2}-t_0)^2/w^2]$	Not given	—[a]	Solution of $3r$ transcendental equations
3	$y(t,t_0,w,A) = A \exp\,[-t_0(t-t_0)^2/(2w^2 t)]$	Not given	—[a]	Solution of $3r$ transcendental equations
4	$y(t,t_0,w,A,p_{c_1},pc_2,p_{c_3},p_{c_4}) = A\,\{\exp[-(t-t_0)^2/2w^2] + (1-y_1)y_2\}$ where: $y_1 = 0.5\{1 - tgh[p_{c_1}(t-p_{c_2})]\}$ $y_2 = p_{c_3} \exp\{-0.5p_{c_4}[(t-p_{c_5}{}^2)^{1/2} + (t-p_{c_5})]\}$	Not given	—[a]	Solution of transcendental equations with $7r$ unknowns[b]
5	$y(t,t_0,w,A,\tau) = Aw\tau^{-1}(2\pi)^{1/2}$ $\exp\{(w^2/2\tau^2) - [(t-t_0)/\tau]\}\mathrm{erf}\{[(t-t_0)/w\sqrt{2} - (w/\tau\sqrt{2})\}$ where: $\mathrm{erf}[x] = \pi^{-1/2}\int_{-\infty}^{x}\exp[-u^2]\,du$	≈ 1	≈ 1	Solution of transcendental equations with $4r$ unknowns
6	$y(t,t_0,w,C_i) = \frac{1}{w\sqrt{2\pi}} \exp\left[\frac{-(t-t_0)^2}{2w^2}\right]\left[1 + \sum_{i=3}^{K} \frac{C_i}{i!} H_i\left(\frac{t}{w}\right)\right]$	Not given	≈ 15	Solution of $(K+2)r$ transcendental equations with $(K+2)r$ unknowns[b]
7	$y(t,t_0,w) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(t-t_0)^2}{2w^2}\right]\left[1 + \frac{S}{6}\left\{\left(\frac{t-t_0}{w}\right)^3 - 3\left(\frac{t-t_0}{w}\right)\right\} - \frac{E}{24}\left\{\left(\frac{t-t_0}{w}\right)^4 - 6\left(\frac{t-t_0}{w}\right)^2 + 3\right\}\right]$	Not given	≈ 15	Solution of $4r$ transcendental equations with $4r$ unknowns[b]

8	$y(t,t_0,w_1,w_2,A) = A \exp[-(t-t_0)^2/2w_1^2]$ for $t \leq t_0$ $= A \exp[-(t-t_0)^2/2w_2^2]$ for $t > t_0$	≈ 3	≈ 2	Solution of $r(r+1)$ transcendental equations with $3r$ unknowns
9	$y(t,t_0,w,A,p_{m_1}) = A \exp[-(t-t_0)^2/2w^2]$ for $t < t_0$ $= A \exp\{-(t-t_0)^2/[2(w + p_{m_1}(t-t_0)]^2\}$	Not given	—[c]	Solution of $r(r+1)$ transcendental equations with $4r$ unknowns[b]

[a]Not tested; function is unsuitable for separation.
[b]Estimation of additional parameters is necessary.
[c]Not tested; many additional parameters must be estimated.

g.c. processing which is found only in sophisticated software, is the automatic selection of the main setting of the chromatograph (carrier-gas flow rate, and column, detection and injection port temperatures) to obtain well-resolved chromatograms in a reasonable time. This involves computer analysis of the dependence of peak resolution on the settings of the instrument. In the optimization procedure of Morgan and Deming [37], for every mixture separated on a given column, the following equations are defined as

$$CRF = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2 + \beta_2 x_2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (8)$$

$$t = \beta'_0 + \beta'_1 x_1 + \beta'_{11} x_1^2 + \beta'_2 x_2 + \beta'_{22} x_2^2 + \beta'_{12} x_1 x_2 \quad (9)$$

Here, $CRF = \sum_{i=1}^{j} \ln P_i$ and represents the degree of peak separation [38]; t is the duration of the analysis, x_1 the column temperature, and x_2 the carrier-gas flow rate; β,β' are empirical constants connected with the physical mechanism of the process.

In another approach [39], the optimal flow rate of carrier gas is established from

$$Z = (a/u^2) + (b/u) + c + d + eu + fu^2 \quad (10)$$

in which the dependence of peak asymmetry (coefficient Z) on flow rate (u) is described by means of six empirical constants a, b, c, d, e and f. The real optimization criteria are the relative retention of two neighbouring peaks, α, and coefficient D

$$D = [(t_R - t_M)/t_R] - 18/(N)^{1/2}$$

where t_R and t_M are the retention time and zero retention time respectively, and N is the number of theoretical plates. In the case $D > 0$ and $\alpha > 1.5$, a new column temperature, T, is selected according to the rule: $T = T + (15 + \alpha)$.

The principal drawback of these methods is the need for an extensive series of analyses for each column system, which is very time-consuming despite the use of a computer for preliminary data processing. In the computer program CHADIC developed here for optimization of analysis, the kinetic theory of gas chromatography [40] is basically used. The well-known van Deemter equation $H = A + (B/u) + Cu$ may be rearranged to $H - A = h = B/u + Cu$. For optimal flow rate, $h_0 = (B/u_0) + Cu_0$. Since $u_0 = (B/C)^{1/2}$ then

$$h = (h_0/2)[(u_0/u) + (u/u_0)] \quad (11)$$

$$u_0 = (u/h_0)[h \pm (h^2 - h_0^2)^{1/2}] \quad (12)$$

$$C = hu/(u_0^2 + u^2) \quad (13)$$

The consequence of this approach is clearly seen: only two chromatograms are needed to establish the optimal flow rate. The temperature of the column (and then those of the detector and injection port) are found from the classical equations

$\ln r_{i,j} = a + b/T$, and $\ln k_i = A + B/T$

where $r_{i,j}$ is the relative retention and k_i is the separation coefficient.

In the actual procedure, the first chromatogram is run at temperature T_1 with carrier-gas flow rate u_1, whereas the second is run at the same flow rate, but for a different temperature T_2. An example of the optimization is shown in Fig. 2 for multicomponent analysis. Of course, in the optimization algorithm, all the chromatographic peaks are taken into consideration. The algorithm is fast and suitable for real-time processing. It can be used to control the performance of the gas chromatograph in routine analysis of a given mixture, once the temperature dependence of the relative retention for two neighbouring peaks has been established. As the CHADIC system is used in the batch mode, the computer prints out some advice on how to settle the instrumental parameters to obtain the best resolved chromatogram for a given analysis time.

The future development of algorithms for high-level processing of g.c. data may follow two general directions. First, the central processor unit of a large computer could be used for advanced calculations exploiting accurate algorithms. This is an extension of the concept of post-run processing of chromatographic data [41] in which peak information is produced in real time, but baseline correction is done later. Such processing (done cheaply at night) could be suitable for modernizing numerous ordinary gas chromatographs because the instrument would only require fitting with a suitable numerical output on magnetic tape. Of course, the application of the best algorithms, processed by powerful computers, is necessary to gain adequate experience for other developments. The second direction of development is the application of processor-based chromatographs. Such instruments are capable of producing a complete chromatographic analysis, calculations and printed report without operator intervention. However, the number and

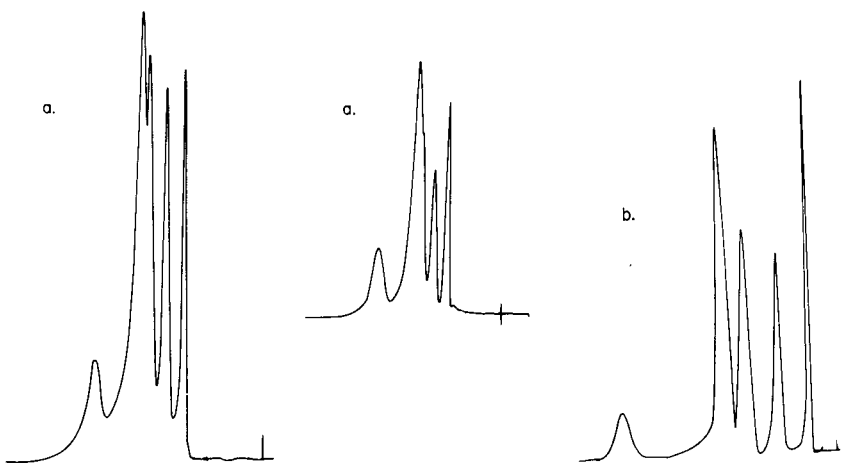


Fig. 2. Optimization of chromatographic conditions for multicomponent analysis: (a) test chromatograms (b) chromatogram obtained under optimal conditions of separation.

complexity of the pre-programmed functions that can be included are limited, because of the finite processor address space, the large variety of possible functions to calculate, and the desirability of maintaining ease of use. Improvements in microprocessor technology should lead to greater versatility in dealing with unusual situations.

REFERENCES

1 G. H. Stewart, Anal. Chem., 32 (1960) 1205.
2 Z. Bohm, J. Chromatogr., 3 (1960) 265.
3 C. H. Orr, Anal. Chem., 33 (1961) 158.
4 L. Chromy and Z. Hippe, Chem. Anal., 10 (1965) 629.
5 J. M. Gill, J. Chromatogr. Sci., 7 (1969) 731.
6 G. Schomburg and D. Henneberg, Anal. Chem., 42 (1970) 51A.
7 M. F. Burke and R. R. Thurmann, J. Chromatogr. Sci., 8 (1970) 39.
8 J. D. Hettinger, J. R. Hubbard, J. M. Gill and L. A. Miller, J. Chromatogr. Sci., 9 (1971) 710.
9 L. Klatt, P. Carr and R. J. Krusberg, Chem. Instrum., 3 (1972) 327.
10 J. E. Longbottom, J. Chromatogr. Sci., 11 (1973) 13.
11 I. Wehling, Chromatographia, 5 (1972) 197.
12 G. Schomburg, Angew. Techn., 84 (1972) 390.
13 K. Derge, Fette, Seiffen, Anstrichmittel, 6 (1973) 334.
14 Z. Hippe, B. Debska and A. Kerste, Cwiczenia z chemii fizycznej do obliczen na EMC, PWN, Warszawa, 1979, pp. 20—29.
15 Z. Hippe, A. Bierowska and T. Pietryga, Res. Mem., Techn. Univ., Rzeszów, 1976.
16 Z. Hippe, A. Bierowska and T. Pietryga, Res. Mem., Techn. Univ., Rzeszów, 1977.
17 F. Baumann, E. Herlicska and A. C. Brown, J. Chromatogr. Sci., 7 (1969) 680.
18 F. Hock, Chromatographia, 2 (1969) 334.
19 K. Kishimoto and A. Musha, J. Chromatogr. Sci., 9 (1971) 608.
20 P. Sutre and J. P. Malenge, Chromatographia, 5 (1972) 141.
21 A. H. Anderson, T. C. Gibb and A. B. Littlewood, Chromatographia, 2 (1969) 466.
22 A. H. Anderson, T. C. Gibb and A. B. Littlewood, Anal. Chem., 42 (1970) 434.
23 J. Novak, K. Petrovic and S. Wicar, J. Chromatogr., 55 (1971) 221.
24 A. Fozard, J. J. Frances and A. J. Wyatt, Chromatographia, 5 (1972) 130.
25 S. N. Chesler and S. P. Cram, Anal. Chem., 43 (1971) 1922.
26 M. Goedert and G. Guiochon, Chromatographia, 6 (1973) 76.
27 N. Guichard and G. Sicard, Chromatographia, 5 (1972) 83.
28 G. Schomburg and E. Ziegler, Chromatographia, 5 (1972) 96.
29 F. Caesar and M. Klier, Chromatographia, 7 (1974) 526.
30 P. C. Kelly and W. E. Harris, Anal. Chem., 43 (1971) 1170.
31 Z. Hippe, A. Bierowska and T. Pietryga, Res. Mem., Techn. Univ., Rzeszów, 1978.
32 H. G. Struppe, W. Saffer and H. J. Pommrich, Int. Conf. Applications of Computers in Analytical Chemistry, Leipzig, March, 1977.
33 Sigma 10 Chromatography Data Station, Perkin—Elmer, Norwalk, CT., 1978.
34 Z. Hippe, A. Bierowska and T. Pietryga, Polish—German Seminar on Applications of Computers in Processing of Physiochemical and Analytical Data, Rzeszów, May, 1978.
35 R. Hooke and J. A. Jeeves, J.A.C.M., 1961 pp. 212—229.
36 Z. Hippe, A. Bierowska and T. Pietryga, Res. Mem., Techn. Univ., Rzeszów, 1979.
37 S. N. Deming and S. L. Morgan, J. Chromatogr., 112 (1975) 267.
38 R. Kaiser, Gas Chromatographie, Geest and Portig, Leipzig, 1960, p. 33.
39 S. P. Cram and J. E. Leitner, Chromatographia, 9 (1974) 567.
40 J. J. Van Deemter, F. J. Zuiderweg and A. Klinkenberg, Chem. Eng. Sci., 5 (1956) 271.
41 P. C. Dryden, L. M. Altmayer and J. S. De Good, Pittsburg Conf. Analytical Chemistry of Applied Spectroscopy, Cleveland, March, 1979, Abstr. 166.

Analytica Chimica Acta, 122 (1980) 291—301
Computer Techniques and Optimization

MULTIPARAMETER MODELS AND STATISTICAL UNCERTAINTIES

LOWELL M. SCHWARTZ
Department of Chemistry, University of Massachusetts, Boston, MA 02125 (U.S.A.)
(Received 4th October 1979)

SUMMARY

When more than one parameter is found by a least-squares calculation, the statistical uncertainties of the parameters are generally interdependent. If the uncertainty of one such parameter is quoted as a confidence interval based on the standard error estimate of that parameter and the Student *t*-statistic, this interval tends to be an underestimate. A suggestion is made to quote more conservative parameter uncertainties as the extreme points on the 95% joint confidence ellipsoid and it is shown that these joint parametric uncertainties are easily calculated from the standard error estimates. Both linear and nonlinear multiple regression are discussed. Nonlinear parameter uncertainties are found after an iterative search for the minimum sum-of-squares of residuals; searches by the Gauss and simplex methods are considered. A joint parametric uncertainty calculation is illustrated by a four-parameter nonlinear regression involving a pH potentiometric titration.

Over two decades ago Mandel and Linnig [1] pointed out that when least-squares parameters of a straight line are calculated simultaneously by fitting the equation $y = a + bx$ to a set of data, the statistical uncertainties of the calculated parameters $(\hat{a},\hat{b})$ cannot be expressed properly independently of one another because of their statistical correlation. To report standard error estimates or confidence limits for $\hat{a}$ and $\hat{b}$ separately fails to convey the complete picture of the uncertainties of these determinations. Also in severe cases it has been shown [2] that this correlation may yield apparent but dubious relationships. In their Fig. 2, Mandel and Linnig illustrated the problem by showing the 95% confidence ellipse for the two parameters based on the ten data points given in their Table 2. Their ellipse is reproduced here in Fig. 1 and carries the following interpretation. If 100 replicate sets of ten data points each are taken by the same experimental procedure and a similar 95% confidence ellipse is drawn for each, then in approximately 95 of the 100 cases, the true parameters (a^*,b^*) will be represented by a point inside the ellipse and in about five cases the true parameter point will fall outside. The least-squares parameters calculated from the single data set are $\hat{a} = 6.99$ and $\hat{b} = 1.00765$, and this point is at the geometrical center. The negative slope of the major axis of the ellipse with respect to the (a,b) axes reflects the negative statistical correlation between $\hat{a}$ and $\hat{b}$.

Confidence limits for least-squares parameters are usually calculated independently from the formulas $\hat{a} \pm t_{\alpha/2}\mathrm{SE}(\hat{a})$ and $\hat{b} \pm t_{\alpha/2}\mathrm{SE}(\hat{b})$, where SE

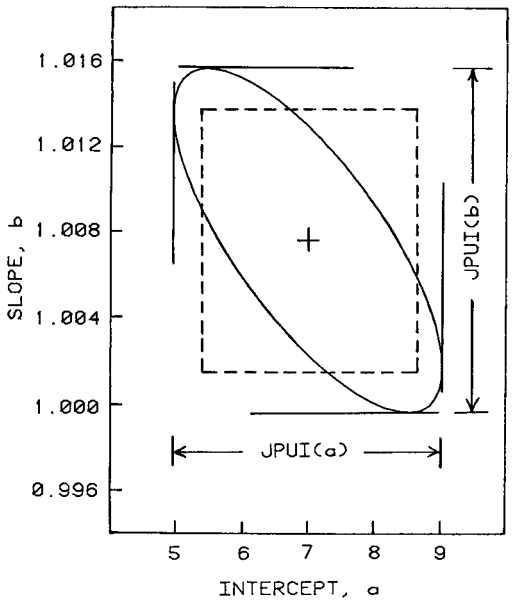


Fig. 1. 95% confidence ellipse for the straight-line example cited [1]. The central point corresponds to the least-squares parameter values, the sides of the broken rectangle are 95% non-simultaneous confidence limits, and tangents to the ellipse are joint parametric uncertainty limits (JPUL). The differences between these limits are the joint parametric uncertainty intervals (JPUI).

denotes the standard error estimate of the indicated parameter, and $t_{\alpha/2}$ is the Student t-statistic based on the relevant (data points-less-two) degrees of freedom and on a confidence level $(1 - \alpha)100\%$. If 95% confidence limits (CL) are calculated in this manner from Mandel and Linnig's data, we find $SE(\hat{a}) = 0.705$, $SE(\hat{b}) = 0.00271$, $t_{0.025} = 2.31$, $CL(\hat{a}) = 8.62, 5.37$ and $CL(\hat{b}) = 1.01391, 1.00139$, that is, confidence intervals of $CI(\hat{a}) = 3.25$ and $CI(\hat{b}) = 0.01252$. These limits are shown as vertical and horizontal broken lines defining a rectangle around the center of the ellipse in Fig. 1. To quote these two pairs of limits as statistical uncertainties for the parameters implies the mistaken impression that in some sense the rectangle defines a 95% confidence region. Clearly some (a,b) points are inside both the rectangle and the ellipse and some are not, but it is significant that points near the ends of the major axis are outside the rectangle. To the extent that these distant points are beyond the non-simultaneous confidence limits, the parameter uncertainties are underestimated; and, as will be shown, if the ellipsoidal representation is extended to more than two parameters (dimensions), the degree of underestimation increases.

Beyond least-squares straight lines there are many other data treatment problems of interest to analytical chemists [3] for which more than two parameters are calculated simultaneously. In some of these problems the

model equations are linear with respect to the parameters and for these the calculation procedure is available in any number of statistics texts [4] and is known as multiple linear regression. The joint confidence region for all the parameters becomes an ellipsoid in P-dimensional space for P parameters [5]. This representation is statistically rigorous but is cumbersome for expressing parametric uncertainties since the P-dimensional ellipsoid needs to be reduced to a two-dimensional basis to be placed on paper. This could be done by plotting ellipses for each pair of parameters, and so for $P = 3$, 4 and 5 the number of plots involved would be 3, 6 and 10 respectively. Even though these curves could be generated by computer plotting routines, many analytical chemists would be reluctant to report, say, four parameter values from an experiment and then append six elliptical plots to express the corresponding statistical uncertainties.

Faced with the unhappy choice of using the cumbersome but correct joint confidence ellipsoid or the misleading but conveniently calculated standard errors or non-simultaneous confidence limits, a suggestion is made here to report statistical uncertainties as the distance between the extreme points on the 95% joint confidence ellipsoid in the direction of each parameter axis. These points are defined by the planes tangential to the ellipsoid surface and perpendicular to the parameter axis in question. In Fig. 1, these are shown by the vertical and horizontal tangents to the ellipse. Since no concise terminology exists for such a measure, we will refer to the distance between these tangent planes as a "joint parametric uncertainty interval" (JPUI) and the uncertainty limits of the parameter involved as "joint parametric uncertainty limits" (JPUL). A calculated parameter, say, $\hat{a}$ is centered in the JPUI ($\hat{a}$) so that the parameter and its uncertainty limits might be reported as JPUL ($\hat{a}$) = $\hat{a} \pm$ JPUI ($\hat{a}$)/2. For the two-parameter line of Mandel and Linnig [1], their eqns. (8) and (9) yield JPUI ($\hat{a}$) = 4.14 and JPUI ($\hat{b}$) = 0.01591 which are represented by tangents in Fig. 1. Each of these intervals is some 27% greater than the corresponding non-simultaneous CI. It will be shown that calculating these more conservative limits is equally a simple matter for any number of parameters, and that for a single parameter the JPUI reduces to the ordinary 95% CI.

The JPUI measure is proposed for use by analysts who wish to report statistical uncertainties for one or more parameters of a multiparameter model and in such a way as to reflect the greater degree of uncertainty due to the multiple fitting calculation. The JPUI, however does not express all the information inherent in the full joint confidence ellipsoid and it does not carry a probabilistic interpretation similar to the one quoted in the introductory paragraph. In the absence of this interpretation, there seems little point in retaining an arbitrary confidence level such as is inherent in some other statistical measures of uncertainty. Therefore the suggestion was made above to adopt the 95% ellipsoid specifically as the basis for the JPUI. The 95% level is arbitrary but is commonly used for confidence limits in analytical chemistry. In a sense the JPUI is a compromise. It has the same conveni-

ence and calculation simplicity as the standard error or the non-simultaneous CI but it reflects the larger extent of the simultaneous confidence ellipsoid, albeit in a superficial manner.

A derivation of the formula by which the JPUI is calculated when three or more parameters are found does not seem to be available. There is little doubt than an exercise equivalent to this exists somewhere in the mathematical literature since it is a straightforward application of well known concepts. Nevertheless, a derivation is offered in Appendix A. There it is shown that if P parameters are estimated simultaneously then the JPUI for any particular parameter $\hat{b}_l$ is proportional to the standard error of $\hat{b}_l$. The relationship is eqn. (A9) and it is seen that the proportionality factor depends only on P and on the F_α-statistic which for N data points involves P and $N - P$ degrees of freedom and $\alpha = 0.05$, appropriate for the 95% confidence level.

If a linear regression problem is solved by using any of the digital computer programs available for this purpose, it is a simple matter to convert the standard error estimates for the parameters into JPUI uncertainties by using an F table. When $P = 1$, the JPUI equals the 95% CI because $F_\alpha^{1/2} = t_{\alpha/2}$, and the discrepancy between the JPUI and the non-simultaneous 95% CI for any one parameter increases with increasing number of parameters. With the aid of F and t tables this trend can be seen by calculating $\mathrm{JPUI}(\hat{b})/\mathrm{CI}(\hat{b}) = (PF_\alpha)^{1/2}/t_{\alpha/2}$ for any fixed N but with increasing P. This demonstrates that as P increases, the non-simultaneous CI becomes a smaller fraction of the more conservative JPUI.

NONLINEAR REGRESSION

There are perhaps even more applications of nonlinear regression than of multiple linear regression in analytical chemistry [3]. However, there are also a greater number of methods of data treatment to deal with problems of extracting parameter values from nonlinear model equations given a set of experimental data. A common characteristic of all these methods is that the calculation must be iterative in the sense that the optimum set of parameter values is approached by a stepwise algorithm starting from an initial set of approximate values. Bard [6] gives a comprehensive survey and mathematical account of these methods as developed up to 1972, but in this paper only two of the more widely used methods will be selected for discussion and both will rely on the principle of least squares for fitting the data.

A model equation relates the dependent variable y again having uniform variance s^2 to one or more independent variables x having negligible variances. This model equation contains P parameters $b_1, b_2, \ldots b_P$ to be denoted by the vector b such that the equation

$$y = f(x; \mathrm{b}) \tag{1}$$

is nonlinear with respect to the parameters. $N > P$ measurements (y_i, x_i) are made and parameter values are sought which minimize the sum-of-squares

(SS) of residuals of the y_i data from the y-values predicted by eqn. (1) from the corresponding x_i. A minimum is sought for

$$SS = \sum_{i}^{N} [f(x_i; \mathbf{b}) - y_i]^2 \tag{2}$$

Gauss method

The method of Gauss (Chapter 5 [6]) requires a set of initial parameter estimates $\mathbf{b}^0$ and then seeks corrections $\delta\mathbf{b}$ to each parameter such that SS is reduced. The refinements $\delta\mathbf{b}$ can be applied to $\mathbf{b}^0$ in one of a variety of ways to yield improved parameter values $\hat{\mathbf{b}}$. The simplest but not most efficient way is $\hat{\mathbf{b}} = \mathbf{b}^0 + \delta\mathbf{b}$. Regardless of how the refinements are applied, the improved set $\hat{\mathbf{b}}$ then serves as $\mathbf{b}^0$ estimates for the next iteration from which another set of refinements is computed. The iteration proceeds until some arbitrary criterion of convergence is satisfied. The $\hat{\mathbf{b}}$-values then correspond to a minimum in SS and these values may or may not be acceptable as physically meaningful. This brief summary ignores other practical problems which are not relevant to the topic under consideration. We assume that some variant of the Gauss method will succeed in locating a satisfactory minimum SS.

The set of refinements is calculated by replacing $f(x; \mathbf{b})$ in eqn. (2) by a linear approximation centered on the initial estimates $\mathbf{b}^0$, i.e.

$$f(x; \mathbf{b}) \approx f^0 + f_1' \delta b_1 + \ldots f_P' \delta\, b_P \tag{3}$$

where $f^0 = f(x; \mathbf{b}^0)$ and the derivatives $f_j' = (\partial f / \partial b_j)$ are understood to be functions of x and to be evaluated with $\mathbf{b}^0$ parameter values. After substitution of eqn. (3) into eqn. (2), the result is differentiated with respect to each refinement δb_j and each such derivative is equated to zero. These operations form a set of P linear equations for the P unknowns $\delta\mathbf{b}$

$$\begin{aligned} \delta b_1 \Sigma (f_1')^2 + \delta b_2 \Sigma f_1' f_2' + \ldots &= \Sigma f_1' (y_i - f^0) \\ \vdots \qquad\qquad \vdots \qquad\qquad &\vdots \\ \delta b_1 \Sigma f_P' f_1' + \delta b_2 \Sigma f_P' f_2' + \ldots &= \Sigma f_P' (y_i - f^0) \end{aligned} \tag{4}$$

where the summations are over the N data points. Comparing this set of equations with eqn. (A2), it is evident that this problem is one of multiple linear regression. The identification is made by taking $S_{jk} = \Sigma f_j' f_k'$, $S_{jy} = \Sigma f_y' (y_i - f^0)$ and $\mathbf{b} = \delta\mathbf{b}$. In direct analogy to linear regression, it can be shown (Chapter 7 [6]) that variance and covariance estimates of the optimum nonlinear parameters $\hat{\mathbf{b}}$ are calculated by multiplying the elements of the inverse matrix $\mathbf{S}^{-1}$ by s^2; here an element of $\mathbf{S}$ is $\Sigma f_j' f_k'$. The other developments outlined in Appendix A follow as well. In particular, the joint parametric uncertainty interval for a nonlinear parameter estimate $\hat{b}_l$ is approximately

$$\mathrm{JPUI}(\hat{b}_l) \approx 2(P F_\alpha)^{1/2} \mathrm{SE}(\hat{b}_l) \tag{5}$$

Discussion of the effect of the linearization approximation on this result will be deferred to a later section.

Simplex minimization method

When the derivatives f' of the model equations are difficult to evaluate, alternative "derivative-free" methods of seeking a minimum SS are available. Perhaps the most popular of these is that known as simplex optimization which has been adopted for many applications in analytical chemistry [7, 8]. In the application discussed here, P-values are sought to minimize the function SS of eqn. (2) and a set (simplex) of $P + 1$ parameter vectors b_j ($j = 1, 2 \ldots P + 1$) is initiated. These vectors, also called the vertices of the simplex, have different component parameter values and so each such vertex corresponds to a different value of the function SS being minimized. Following well established rules, new vertices are calculated corresponding to lesser SS values, and these replace unfavorable vertices corresponding to greater SS values. In this way, the simplex of $P + 1$ vertices gradually changes its constituent vectors in search of a local minimum in SS. Arriving at such a minimum, it condenses in size in the neighborhood of that minimum, still retaining a complement of $P + 1$ vertices. The exact procedural rules for seeking the minimum and the criteria for terminating the search vary somewhat among published accounts, but all variations arrive eventually at a set of $P + 1$ vertices all quite close to the one (unknown) vector whose elements are those parameters which truly minimize SS. The elements of the one parameter vector having the least value of SS are accepted as the solution to the problem.

In setting statistical uncertainties for these parameters, one possible approach is to make numerical evaluations of the derivatives f' at the minimum, and from these, to set up the matrix S required to solve the set of eqns. (4). The inverse of that matrix multiplied by s^2 has elements from which parameter uncertainties JPUI are found as described above, but if the analyst has turned to the simplex method to avoid calculating derivatives this approach may not be satisfactory. A more efficient approach suggested by Spendley [9] recognizes that enough information is already calculated in the final simplex from which the variance—covariance matrix of parameters can be found. Apparently this approach has not seen wide application to date but it is well suited to the present development. Therefore, Spendley's method is outlined in Appendix B to express it in terms of the notation used here, and to show how to calculate S^{-1} so that its elements may be used to calculate JPUI according to eqn. (5).

CHECKING THE LINEARIZATION APPROXIMATION AND A SAMPLE CALCULATION

The validities of nonlinear parameter uncertainties depend on the accuracy of the linear approximation of the model equation. This linearization is equivalent to the hypothesis that the function SS ($\hat{b}$) is quadratic with respect to the parameters (see Appendix B). Quadratic behavior very near the minimum is not in doubt, but may be suspect as far away as some of the JPUI values, which are several standard error distances from the minimum. Bard

(Chapter 7 [6]) suggests that the validity of such a parameter uncertainty should be checked by testing to see if the value of SS at such a point is reasonably close to the sum-of-squares calculated from the quadratic approximation to eqn. (2). If these two values are indeed nearly the same, then the uncertainty quotation can be accepted as accurate since it has been shown that the quadratic approximation is valid as far away from $\hat{\mathbf{b}}$ as implied by that quotation. However, if these two values differ greatly, the quoted uncertainties should be accepted only with some scepticism.

If $\mathbf{b}_{UL}$ is the vector of parameter values at some quoted uncertainty limit and $\hat{\mathbf{b}}$ is the vector at the minimum, then the quadratic approximation to SS is given by

$$SS_q = SS(\hat{\mathbf{b}}) + (\mathbf{b}_{UL} - \hat{\mathbf{b}})\,\mathbf{T}\,\mathbf{S}\,(\mathbf{b}_{UL} - \hat{\mathbf{b}}) \qquad (6)$$

and if this limit is a JPUL the second term according to eqn. (A3) is $PF_\alpha s^2$. This means that $SS_q = SS(\mathbf{b}) + PF_\alpha s^2$ is the same for all JPUL calculated simultaneously. The linearization approximation is checked by comparing this value with $SS(\mathbf{b}_{UL})$, which is eqn. (2) evaluated with the parameters $\mathbf{b}_{UL}$ substituted, and $SS(\mathbf{b}_{UL})$ has a different value at each individual confidence limit. The component parameters in $\mathbf{b}_{UL}$ are given by equations in Appendix A. If, say, the upper JPUL of $\hat{b}_l$ is being checked, the P components of that $\mathbf{b}_{UL}$ are $\hat{b}_l + \Delta b_l$, where Δb_l is the positive square root of eqn. (A7), and $P - 1$ components $b_j + \Delta b_j$ are from eqn. (A6) with Δb_l substituted.

An example will serve to illustrate a JPUL calculation and linearization check procedure. In a recent communication [10] a method of statistical treatment of pH potentiometric titration data was discussed and the results of a titration of aqueous potassium rhodizonate with hydrochloric acid were described. In that example, four parameters were determined simultaneously; two acid dissociation constants pK_1 and pK_2, the first equivalence point volume v_{ep}, and the pH meter calibration offset pH_{cal}. Those parameter values and associated standard error estimates have been reported (Table 1 [10]) and the relevant information is repeated here as part of Table 1. In that calculation, the variances of measurements, pH against volume of HCl added, were not uniform and neither measurement was much more precise than the other, but by using normalized weighting factors an effective variance of $s_m^2 = 0.8626 \times 10^{-4}$ was calculated [10]; this quantity plays the role of s^2 in the present paper. Also by using normalized weighting factors in all summations it was found that $SS(\hat{\mathbf{b}}) = 0.6842 \times 10^{-4}$ for the sum-of-squares at the minimum; the parameter values $\hat{\mathbf{b}}$ are listed in column 1 of Table 1. Based on 33 measurements, it is calculated that $F_{0.05} = 2.701$ for 4 and 29 degrees of freedom, $PFs^2 = 0.9321 \times 10^{-3}$, and $SS_q = 1.616 \times 10^{-3}$ from eqn. (6). The S^{-1} matrix at the minimum was

$$\begin{bmatrix} 0.2948 & -0.2644 & -0.08326 & -0.02344 \\ & 0.3274 & 0.07551 & 0.02852 \\ & & 0.05413 & 0.02013 \\ & & & 0.01127 \end{bmatrix}$$

TABLE 1

Joint parametric uncertainty limits (JPUL)[a] calculated for four parameters based on the titration experiment cited [10]

Parameter coordinates	$\hat{b}$[b]	pK_1		pK_2		v_{ep}		pH_{cal}		JPUI
		Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	
pK_1	4.377	4.394	4.361	4.363	4.391	4.366	4.388	4.371	4.384	0.033
pK_2	4.661	4.646	4.676	4.678	4.643	4.671	4.651	4.669	4.653	0.035
v_{ep}(ml)	2.053	2.048	2.058	2.057	2.049	2.060	2.046	2.059	2.047	0.014
pH_{cal}($\times 10^{-3}$)	−1.39	−2.70	−0.07	0.14	−2.91	1.26	−4.03	1.86	−4.63	6.5
$SS(b_{CL})$($\times 10^{-3}$)[c]		1.602	1.634	1.628	1.596	1.614	1.620	1.613	1.624	
% discrepancy[d]		0.9	−1.1	−0.7	1.3	0.1	−0.2	0.2	−0.5	

[a]Underlined entries. [b]Least-squares parameter values. [c]Sum-of-squares calculated from nonlinear model equations. [d]$100[SS_q - SS(b_{UL})]/SS(b_{UL})$ with $SS_q = 1.616 \times 10^{-3}$.

where omitted elements v_{jk} below the diagonal are the same as corresponding elements v_{kj} above because of symmetry. Each parameter is taken in turn, and upper and lower JPUL are calculated as extreme points on the ellipsoid. The coordinates of those points are the first four entries in columns 2—9 (Table 1) and the underlining denotes the JPUL values for each parameter. Each of the JPUI values (column 10) is $2(PF_\alpha)^{1/2} = 6.6$ times the corresponding standard error estimates quoted earlier [10]. To check the validity of the linearization approximation leading to each confidence limit, the sum-of-squares at each vector $\mathbf{b}_{UL}$ of parameters is calculated from the nonlinear equations and these sums are also listed. These are to be compared with $SS_q = 1.616 \times 10^{-3}$. The discrepancies given in the bottom row are all quite small and so the linearization approximation upon which the JPUI estimates are based can be accepted.

APPENDIX A

Joint parametric uncertainty intervals for parameters of a multiple linear regression model

A multiple linear regression model of the form

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p \tag{A1}$$

involves $P = p + 1$ unknown parameters $b_0, b_1, \ldots b_p$ which are to be estimated from $N > P$ data points y_i of uniform variance s^2 observed at $i = 1, 2, \ldots N$ settings of the independent variables $x_{1i}, x_{2i}, \ldots x_{pi}$, each of which has negligible variance relative to y_i. The principle of least squares leads to a set of P normal eqns. (A2) for the "best" values $\hat{\mathbf{b}}$ of the parameters

$$\begin{aligned} \hat{b}_0S_{00} + \hat{b}_1S_{01} + \ldots \hat{b}_pS_{0p} &= S_{0y} \\ \hat{b}_0S_{10} + \hat{b}_1S_{11} + \ldots \hat{b}_pS_{1p} &= S_{1y} \\ \vdots \quad\quad \vdots \quad\quad & \quad\quad \vdots \quad \vdots \\ \hat{b}_0S_{p0} + \hat{b}_1S_{p1} + \ldots \hat{b}_pS_{pp} &= S_{py} \end{aligned} \tag{A2}$$

where $S_{jk} = \sum_i x_{ji}x_{ki}$, $S_{jy} = \sum_i x_{ji}y_i$ and for symmetry of notation we have written $x_{0i} = 1$ so that, for example, $S_{00} = \sum_i 1^2 = N$ and $S_{0y} = \sum_i y_i$. Digital computer subroutines available for solving multiple linear regression problems frequently require that the input data be supplied as deviations from their respective means. With this transformation, the summations in eqn. (A2) become $S_{jk} = \sum_i (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$ or $S_{jy} = \sum_i (x_{ji} - \bar{x}_j)(y_i - \bar{y})$, and hence all terms involving b_0 as well as the first equation of the set disappear to reduce the set to $P = p$ equations for an equal number of unknowns. The following development is valid for either formulation.

In matrix notation, eqn. (A2) is $\mathbf{Sb} = \mathbf{y}$, where S is $P \times P$ and has elements S_{jk}, b is a P-element (column) vector of the parameters, and y is a P-element vector of elements S_{jy}. The P estimates of $\hat{\mathbf{b}}$ are found by calculating and premultiplication by the inverse $\mathbf{S}^{-1}$ so that $\hat{\mathbf{b}} = \mathbf{S}^{-1}\mathbf{y}$. The product $s^2\mathbf{S}^{-1}$ is called the variance–covariance matrix because v_{jj}, a diagonal element of that matrix, estimates the variance, var $\hat{b}_j$, and an off-diagonal element v_{jk} estimates the covariance, cov$(\hat{b}_j, \hat{b}_k)$.

For P parameters, the joint confidence region [5] is ellipsoidal in P-dimensional space and the boundaries of this region are given by the following quadratic equation expressed alternatively in matrix or algebraic forms.

$$\Delta\mathbf{b}^{\mathrm{T}}\mathbf{S}\Delta\mathbf{b} = \sum_{j=0}^{p}\sum_{k=0}^{p}(b_j - \hat{b}_j)(b_k - \hat{b}_k)S_{jk} = PF_\alpha s^2 \tag{A3}$$

Here the matrix $\Delta\mathbf{b}$ has elements $\Delta b_{jk} = (b_j - \hat{b}_j)(b_k - \hat{b}_k)$ and $\Delta\mathbf{b}^{\mathrm{T}}$ is its transpose. F_α represents the F-statistic with P and $N - P$ degrees of freedom and at a selected $(1 - \alpha)100\%$ level of confidence. The extreme positions of one of the parameters, say, b_l on the ellipse are characterized by its derivative with respect to all other parameters being zero, i.e. by $\partial b_l/\partial b_{j\neq l} = 0$. There are $P - 1$ such conditions leading to $P - 1$ equations:

$$\sum_{k=0}^{p} S_{jk}\Delta b_k = 0 \quad (j = 0, 1, \ldots p \text{ but } j \neq l) \tag{A4}$$

Considering again eqn. (A3), the left-hand side is a summation of all possible cross-products j,k and thus can be rewritten:

$$\Delta b_l \sum_{k=0}^{p} \Delta b_k S_{lk} + \sum_{j=0}^{p}{}' \Delta b_j \left[\sum_{k=0}^{p} \Delta b_k S_{jk}\right] = PF_\alpha s^2 \tag{A5}$$

where Σ' denotes the omission of $j = l$ from that summation. Recognizing that the quantity in brackets is zero according to eqn. (A4), only the first summation in eqn. (A5) is non-zero. Equations (A4) and (A5) together are P in number and may be solved for the P coordinates $\Delta b_j, j = 0, 1, \ldots p$ of the two extreme points in the l-direction on the joint confidence ellipse. This is most easily done by dividing eqn. (A5) by Δb_l so that the resulting left-hand side has the same form as eqn. (A4) and then expressing the full set as $\mathbf{S}\,\Delta\mathbf{b} = \mathbf{Z}$, where Z is a vector of P elements all of which are zero except for a single element $Z_l = PF_\alpha s^2/\Delta b_l$. The solution of this set is $\Delta\mathbf{b} = \mathbf{S}^{-1}\mathbf{Z}$ which can be expressed algebraically as

$$\Delta b_j = v_{jl} PF_\alpha s^2 / \Delta b_l \tag{A6}$$

for each coordinate j of the extreme points in the l-direction. In particular for the l-coordinate itself, eqn. (A6) becomes

$$\Delta b_l^2 = v_{ll} PF_\alpha s^2 = PF_\alpha \operatorname{var} \hat{b}_l \tag{A7}$$

The two roots of this equation give the extreme points in the l-direction on the joint confidence ellipsoid relative to $\hat{b}_l$ itself. Here with $\alpha = 0.05$ these are half the joint parametric uncertainty interval (JPUI) for $\hat{b}_l$. Since the standard error $\mathrm{SE}(\hat{b}_l)$ is the square root of var $\hat{b}_l$ and the (non-simultaneous) confidence interval (CI) for $\hat{b}_l$ is $2\, t_{\alpha/2}\mathrm{SE}(\hat{b}_l)$, there results

$$\mathrm{JPUI}(\hat{b}_l) = 2(PF_\alpha)^{1/2}\mathrm{SE}(\hat{b}_l) = [(PF_\alpha)^{1/2}/t_{\alpha/2})]\ \mathrm{CI}(\hat{b}_l), \qquad \alpha = 0.05 \tag{A8}$$

Once eqn. (A7) has been solved for the two roots Δb_l, each root when substituted in turn into eqn. (A6) yields the other $P - 1$ coordinates of an extreme point in the b_l direction.

APPENDIX B

Spendley's method [9] of calculating the variance-covariance matrix

A simplex search for a minimum in the sum-of-squares function SS of eqn. (2) terminates when the $P + 1$ vertices of the simplex are arbitrarily close to the minimum vertex or arbitrarily close to one another near that minimum. In either case, at that stage the computer will have found a single best vector $\hat{\mathbf{b}}$ of P parameters yielding the least value of SS and P other vectors $\mathbf{b}_p$, ($p = 1, 2 \ldots P$) each yielding slightly greater SS values. Recalling other notation associated with eqns. (1) and (2), each parameter vector has components b_j, ($j = 1, 2, \ldots P$) and the calculation is based on N data points (x_i, y_i)($i = 1, 2, \ldots N$). The symbols i, j and p therefore are indices of data points, parameters and simplex vertices, respectively. The nonlinear function of eqn. (1) is f.

Following Spendley, the best parameter value $\hat{b}_j$ is subtracted from each of the corresponding values b_{jp} of the other vertices to yield differences Δb_{jp}. These are set up in an array having a column for each vertex p and a row for each parameter j, thus forming a $P \times P$ matrix $\Delta\mathbf{b}$. Another matrix is set up by calculating the quantities M_{pi} from eqn. (1) for each data point i and for each set of parameters comprising vector p: $M_{pi} = f(x_i;\mathbf{b}_p) - f(x_i;\hat{\mathbf{b}})$. From these, a square $P \times P$ matrix Ω is constructed having elements $\Omega_{pq} = \sum_{i=1}^{N} M_{pi} M_{qi}$ where q, as well as p, is a vertex index. $\Delta\mathbf{b}$, its transpose $\Delta\mathbf{b}^{\mathbf{T}}$, Ω and its inverse Ω^{-1} are each $P \times P$ so that the operations $\Delta\mathbf{b}\ \Omega^{-1}\ \Delta\mathbf{b}^{\mathbf{T}}$ yield another $P \times P$ matrix which Spendley [9] shows to be $\mathbf{S}^{-1}$ of the equation $\hat{\mathbf{b}} = \mathbf{S}^{-1}\mathbf{y}$ (Appendix A), which if multiplied by s^2 is the variance—covariance matrix of the parameters.

This development is based on the assumption that a linear approximation to the nonlinear model equation is valid in the neighborhood of the optimum

set of parameters $\hat{b}$, or, equivalently, that SS($\hat{\mathbf{b}}$), which contains the square of that linearization, is quadratic with respect to the parameters in that same neighborhood. Clearly, this assumption will be false if one or more of the optimum parameters in $\hat{\mathbf{b}}$ is at one of its respective bounds. For example, perhaps one such $\hat{b} = \hat{b}_c$ is a chemical species concentration which obviously has a lower bound of zero and the model equations are fitted to data from experiments done with solutions containing none of that species. The optimum value of that $\hat{b}_c$ should be zero and in such a case SS($\hat{\mathbf{b}}$) is not quadratic in the immediate neighborhood towards lesser $\hat{b}_c$ values. Spendley [9] makes the following suggestion to deal with this problem. In the course of the simplex search, whenever a parameter steps outside its bound, the parameter value is reset to the boundary value and if such a parameter is found to be at this value at all $P + 1$ vertices, that parameter is eliminated from the search so that the problem reduces to P vertices and $P - 1$ parameters. Hence, $\mathbf{b}$ no longer contains the offending parameter causing failure of the quadratic assumption, $\Delta\hat{\mathbf{b}}$ no longer contains the corresponding row of zeros, and the matrix operations outlined above will be valid.

REFERENCES

1 J. Mandel and F. J. Linnig, Anal. Chem., 29 (1957) 743.
2 R. R. Krug, W. G. Hunter and R. A. Grieger, J. Phys. Chem., 80 (1976) 2335.
3 L. A. Currie, J. J. Filliben and J. R. DeVoe, Anal. Chem., 44 (1972) 497R.
4 N. R. Draper and H. Smith, Applied Regression Analysis, John Wiley, New York, 1966.
5 H. Scheffé, The Analysis of Variance, John Wiley, New York, 1959, Chapter 2.
6 Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974.
7 S. L. Morgan and S. N. Deming, Anal. Chem., 64 (1974) 1170.
8 S. N. Deming and L. R. Parker, J. Crit. Rev. Anal. Chem., 7 (1978) 187.
9 W. Spendley, in R. Fletcher (Ed.), Optimization, Academic Press, New York, 1969.
10 L. M. Schwartz and R. I. Gelb, Anal. Chem., 50 (1978) 1571.

Analytica Chimica Acta, 122 (1980) 303—313
Computer Techniques and Optimization

SPECTROPHOTOMETRIC DATA REDUCTION BY EIGENVECTOR ANALYSIS FOR EQUILIBRIUM AND KINETIC STUDIES AND A NEW METHOD OF FITTING EXPONENTIALS†

MARCEL MAEDER* and HARALD GAMPP
Institut für Anorganische Chemie, Universität Basel, CH-4056 Basel (Switzerland)
(Received 5th November 1979)

SUMMARY

The investigation of multicomponent mixtures by spectrophotometry is described. A desk computer, APPLE II, is used to calculate the number of absorbing species in a series of measured spectra by matrix rank analysis. Representation of the observed spectra as linear combinations of eigenvectors leads to significant reduction of the data set, so that a nonlinear least-squares fit based on the Newton—Gauss—Marquardt algorithm is possible on a small computer. As an example, the complexation of copper(II) with 1,4,7-triazaheptane (dien) was studied by combined spectrophotometric and pH titration. The spectra of 62 mixtures at 26 wavelengths were analysed; the number of absorbing species, their spectra and the underlying equilibrium constants were determined. Representation of kinetic curves as linear combinations of eigenvectors is described. It is shown that instead of finding the minimum of the square sum in a multidimensional rate constant space, these minima can be found in a one-dimensional space. Two examples are given: the first is theoretical whereas the second is based on the kinetics of dissociation of the μ-peroxo complexes formed between cobalt(II), oxygen and 4,7,10-triazatridecanedioic acid in acidic solution.

Spectrophotometry can be a good method for studying the structure and reactivity of transition metal coordination compounds in solution. Because of specific metal—ligand interactions, the different complexes usually have different and typical absorption spectra between 300 and 800 nm. However, the practical use of spectrophotometry has been limited by a number of difficulties [1, 2]. Problems arise from (a) overlapping of the spectra of similar species, and (b) the strong correlation between the molar absorptivities and equilibrium or rate constants of minor species. These difficulties can be overcome by using very precise data at several wavelengths. With the availability of inexpensive microprocessors and microcomputers, on-line data acquisition has become widely used. Equipment has been described [3] where absorbance readings with a standard error of about 1.5×10^{-4} absorbance unit were obtained at several wavelengths. Problem (b) can be overcome by eliminating the molar absorptivities, i.e., linear parameters, in the nonlinear regression analysis [4—6].

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

The main problem in the analysis of multicomponent mixtures thus lies in the numerical treatment of large amounts of data. For that purpose, several programs have been developed [2]. The most widely used program, SCOGS [7], is an ordinary Newton—Gauss program, which has been applied to many equilibrium studies. However, a large computer is necessary and good estimates for the linear parameters are needed, i.e. problem (b) is not solved. A program that has been successfully applied to complex equilibria was developed by Zuberbühler and Kaden [2]. It can be run on a desk computer and problem (b) is solved, but it cannot fit the data obtained at several wavelengths simultaneously. The program SQUAD published by Leggett and McBryde [8] is a modification of SCOGS, where point (b) has been overcome, but it still needs a large computer and has not been applied.

In this paper, a program that can be run on a desk computer (APPLE II) with 50 Kbyte memory is described. The program calculates the number of absorbing species in a series of measurements and reduces the data set. From the reduced set, the spectra of the absorbing species can be calculated by using a nonlinear least-squares analysis based on the Newton—Gauss—Marquardt algorithm [6].

The procedure is first illustrated by a chemical equilibrium study in which equilibrium constants and molar absorptivities of the complexes formed are determined. Application of the method of data reduction to kinetics reduces the task to finding the minima of the squares sum or rate constant data for one-dimensional rather than multi-dimensional rate constant space. Both theoretical and practical examples are presented.

MATHEMATICAL METHODS

It has been shown that it is convenient to treat spectra as vectors [5, 9—11]. Given a set of m solutions, for one particular solution i the measured absorbances at q different wavelengths λ_j are the components of the i^{th} spectrum vector $\mathbf{Y}_i$

$$\mathbf{Y}_i = \{Y_i(\lambda_1), Y_i(\lambda_2), \ldots Y_i(\lambda_q)\} \tag{1}$$

(Table 1 lists the meanings of the mathematical symbols.) Each of these $\mathbf{Y}_i$ is a vector in a q-dimensional space, and its direction is determined by the shape of the spectrum. The whole set of spectra can be written as a matrix $\mathbf{Y}$, the columns of which are the spectrum vectors $\mathbf{Y}_i$:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{bmatrix}^t = \begin{bmatrix} Y_1(\lambda_1) \ldots . Y_m(\lambda_1) \\ Y_1(\lambda_2) \ldots . Y_m(\lambda_2) \\ \vdots \quad\quad \vdots \\ Y_1(\lambda_q) \ldots . Y_m(\lambda_q) \end{bmatrix} = \begin{bmatrix} Y_{11} Y_{12} \ldots . Y_{1m} \\ Y_{21} Y_{22} \ldots . Y_{2m} \\ \vdots \quad \vdots \quad\quad \vdots \\ Y_{q1} Y_{q2} \ldots . Y_{qm} \end{bmatrix} \tag{2}$$

If there are n linearly independent species in the m solutions, each measured spectrum is a linear combination of the (unknown) spectra of these species. This means that each of the m vectors is a linear combination of the spectrum vectors of the n particles. Therefore all these m vectors lie in an n-dimensional subspace spanned by the unknown n component vectors.

TABLE 1

Mathematical symbols used in the text

Symbol	Meaning
q	number of wavelengths
m	number of spectra = number of measured solutions
n	number of particles = number of eigenvectors = number of exponentials + 1
Y	matrix ($q \times m$) of the spectra; the i^{th} column contains the spectrum of the i^{th} mixture
V	matrix ($q \times n$) of the eigenvectors; the i^{th} column contains the i^{th} eigenvector
E	matrix ($q \times n$) of the particle spectra; its columns are the spectra of the particles
C	matrix ($n \times m$) of the concentrations; its i^{th} column contains the concentrations of the n species in the i^{th} mixture
X	matrix ($q \times n$) of the linear coefficients belonging to Y
Z	matrix ($q \times n$) of the linear coefficients belonging to E
R	matrix ($q \times m$) of the residuals, i.e. the difference between Y_{cal} and Y (5)
F	matrix ($n \times m$) of exponentials (9)
A	matrix ($q \times n$) of the linear parameters (9)

The number n is equal to the rank of Y, which can be determined by Gaussian elimination. A much better method of calculating n is to compute the second-moment matrix M

$$M = 1/m\ (Y \cdot Y^t) \tag{3}$$

(where the dimension of M is $q \times q$). The rank of Y is equal to the rank of M, which is equal to the number of nonzero eigenvalues of M [9—11].

The eigenvalues and the eigenvectors are calculated by the method of vector iteration [4]. Beginning with an arbitrarily chosen vector, a sequence of iterated vectors V_k, $V_k = M \cdot V_{k-1}$ is calculated until two consecutive vectors become proportional. The factor is the corresponding eigenvalue. The resulting vector is the eigenvector to the greatest eigenvalue. The i^{th} eigenvector is found by choosing the starting vector orthogonal to the previously found $(i - 1)$ eigenvectors. Iterating this vector in the same way leads to the i^{th} eigenvector. Because of inevitable rounding-off errors, the vector must be re-orthogonalized from time to time. This method yields the eigenvectors in the order of decreasing eigenvalues.

Because M is a symmetric matrix, all the eigenvectors are orthogonal, and the set of n eigenvectors therefore forms an orthogonal basis of the n-dimensional subspace in which the spectrum vectors lie, i.e. the measured spectra as well as the particle spectra can be represented as linear combinations of these eigenvectors (see below).

For n absorbing species, the matrix M should give n non-vanishing eigenvalues. However, experimental errors have the consequence that all the eigenvalues are usually non-zero. Therefore statistical tests are used. The number n is equal to the minimum number of eigenvectors needed to represent the spectra vectors within the limits of significance of the absorbance measurements [5].

A program EV was written for an APPLE II desk computer. For 62 spectra at 26 wavelengths the computation time for 6 eigenvectors is about 30 min. An important feature of the method of representing the spectra as linear combinations of the eigenvectors is that once this representation has been found, the further procedure is greatly simplified. Each q-dimensional vector may be located in the n-dimensional subspace with only n coordinates, i.e. an n-dimensional vector replaces a q-dimensional one ($n \ll q$). As a consequence, calculation of equilibrium constants has no longer to be performed at all the $q \times m$ experimental points; all significant information is contained in the reduced set of $n \times m$ points. This aspect, which has not yet found great application in chemistry, will be made clear by the following considerations.

The $q \times m$ elements of the matrix $\mathbf{Y}$ are represented by $\mathbf{Y} = \mathbf{VX}$. The calculated values $\mathbf{Y}_{cal}$ are obtained via Beer's law: $\mathbf{Y}_{cal} = \mathbf{EC}$. The concentrations of the species are functions of the unknown nonlinear parameters, the equilibrium constants. The columns of $\mathbf{E}$ contain the spectra of the individual absorbing species; hence $\mathbf{E}$ can be represented by $\mathbf{E} = \mathbf{VZ}$, which is analogous to $\mathbf{Y} = \mathbf{VX}$. Combination of the expressions for $\mathbf{Y}_{cal}$ and $\mathbf{E}$ leads to

$$\mathbf{Y}_{cal} = \mathbf{VZC} \tag{4}$$

The matrix of the residuals becomes

$$\mathbf{R} = \mathbf{Y}_{cal} - \mathbf{Y} = \mathbf{V}(\mathbf{ZC} - \mathbf{X}) = \mathbf{VS} \tag{5}$$

$\mathbf{Z}$ and $\mathbf{C}$ are the unknowns to be determined. The sum of squares to be minimized is given by

$$QS = \sum_{i=1}^{m} \sum_{j=1}^{q} r_{ij}^2 = \text{trace } (\mathbf{R}^t\mathbf{R}) \tag{6}$$

$$\text{where } \mathbf{R}^t\mathbf{R} = (\mathbf{ZC} - \mathbf{X})^t(\underbrace{\mathbf{V}^t\mathbf{V}}_{\mathbf{1}})(\mathbf{ZC} - \mathbf{X}) = \mathbf{S}^t\mathbf{S} \tag{7}$$

and this reduces to

$$QS = \text{trace } (\mathbf{R}^t\mathbf{R}) = \text{trace } (\mathbf{S}^t\mathbf{S}) = \sum_{i=1}^{m} \sum_{k=1}^{n} s_{ik}^2 \tag{8}$$

This means that to calculate the sum of squares, it is necessary to sum over $q \times m$ elements in the traditional procedure (eqn. 6). If, however, the measured data are represented by $\mathbf{Y} = \mathbf{VX}$ it is necessary to sum over only $n \times m$ elements because of the orthogonality of the eigenvectors (eqn. 8).

In the Gauss–Newton minimization technique there are always expressions similar to eqn. (7) to be handled, which can all be simplified in the same way. Thus, in the minimization procedure, $\mathbf{R}^* = \mathbf{S}$ ($n \times m$ elements) can be used instead of $\mathbf{R} = \mathbf{VS}$ ($q \times m$ elements) without any loss of significant information.

A new application of the described methods treats kinetic curves, obtained by absorbance measurements, also as vectors. The absorbance matrix $\mathbf{Y}$ is built up as in eqn. (2); its columns are the spectrum vectors obtained at m

different times. Each row of Y corresponds to a kinetic curve at one particular wavelength.

Integration of the differential equations of reaction mechanisms with only first-order steps between n chemical species always leads to a sum of $(n-1)$ exponentials and a constant [12]. Therefore Y can be written as expression (9)

$$\begin{bmatrix} Y_{11} \dots Y_{1m} \\ \vdots \quad \vdots \\ Y_{q1} \dots Y_{qm} \end{bmatrix} = \begin{bmatrix} a_{11} \dots a_{1n} \\ \vdots \quad \vdots \\ a_{q1} \dots a_{qn} \end{bmatrix} \cdot \begin{bmatrix} e^{-k_1 t_1} \dots e^{-k_1 t_m} \\ \vdots \quad \vdots \\ e^{-k_n t_1} \dots e^{-k_n t_m} \end{bmatrix} \tag{9}$$

or $\mathbf{Y} = \mathbf{AF}$. In the above notation one of the k-values (e.g. k_n) is zero, thus yielding a constant vector, the spectrum of the solution at infinite time. This notation is the usual one for a system of $(n-1)$ parallel reactions. With other chemical models, the solution of the differential equations shows the functional relationship between the calculated k-values and the real first-order rate constants, and the relationship between the components of the matrix A and the first-order rate constants together with the molar absorptivities of the involved chemical species.

Since all the kinetic curves (measured absorbances at one wavelength) are linear combinations of the same n exponentials, these exponentials form the basis of the n-dimensional subspace. (The similar structure of the equations $\mathbf{Y}_{cal} = \mathbf{EC}$ and $\mathbf{Y} = \mathbf{AF}$ should be noted). For the kinetic investigation a new second-moment matrix M* is computed.

$$\mathbf{M}^* = 1/q\ (\mathbf{Y}^t\mathbf{Y}) \tag{3'}$$

(where the dimension of $\mathbf{M}^* = m \times m$). Its rank again is n and is equal to that of Y and M. This also leads to the same number of exponentials (see above). Computation of the first n eigenvectors as already mentioned, yields an orthonormal basis of the n-dimensional subspace.

So far, it has been established that each measured kinetic curve can be written as a linear combination of $(n-1)$ exponentials or of the basis eigenvectors. Each of the underlying exponentials is, however, also a linear combination of these eigenvectors. Yet, any other exponential cannot lie in that subspace, thus there remains a distance to that subspace

$$\mathbf{R}_i = \mathbf{Y}_i - \sum_{j=1}^{n} (\mathbf{Y}_i, \mathbf{V}_j) \cdot \mathbf{V}_j \tag{10}$$

This is shown in Fig. 1 for $n = 2$. This knowledge leads to a new method of determining the unknown exponents. For a given exponential, the distance $|\mathbf{R}|$ to the subspace is computed; this is a very easy calculation (eqn. 10) because of the orthogonality of the eigenvectors. This distance as a function of the rate constant k for an example of three exponentials is shown in Fig. 2. The problem of finding the minimum of a function, the sum-of-squares, in a three-dimensional rate constant space is reduced to finding three minima in a one-dimensional space. In the normal Newton—Gauss procedure, it became necessary to seek the minimum in a seven-dimensional space in the example

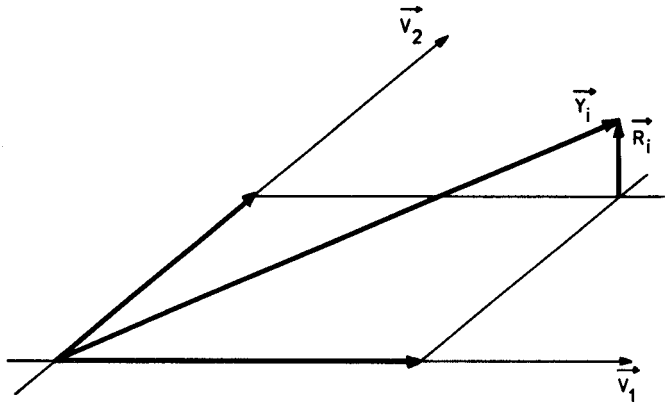


Fig. 1. |R| is the residual vector, the difference between the measured vector Y and its projection, calculated by eqn. (10) for $n = 2$.

from Fig. 2 [6]. This new method does not even need any initial estimates, a point that everybody who has done such calculations will appreciate [6].

EXPERIMENTAL

Spectrophotometric titrations were done on a Cary 118C spectrophotometer with the fully automatic set-up [3]. The complexation of copper(II) with 1,4,7-triazaheptane (dien) was studied at two total concentrations: (a) [dien. 3HCl] = 5.75×10^{-3} M, $[Cu^{2+}] = 5.46 \times 10^{-3}$ M; and (b) [dien. 3HCl] = 5.89×10^{-3} M, $[Cu^{2+}] = 2.80 \times 10^{-3}$ M. The ionic strength was 0.5 M (KNO_3): the temperature was 298 K. Aliquots (2 ml) of solution (a) or (b) were titrated with 31 0.01-ml portions of 0.2 M sodium hydroxide. After each addition of reagent, the spectrum at 26 wavelengths between 750 and 558 nm was recorded. The data were transferred to an APPLE II desk computer (50 Kbyte) and stored on a floppy disk for later use.

A Varian Techtron 635 spectrophotometer was equipped with a manual stopped-flow device to measure kinetic data. For a single run at one wavelength, 500 data points were collected by using a microprocessor system based on the Z80 and a suitable A/D converter. These data were stored on magnetic tape from which they were read directly into the memory of the APPLE II. In the computer, 50 points were selected, in order to obtain optimum information within a useful time.

As a chemical example, the acid hydrolysis of the μ-peroxo complex $[Co(TTDD)]_2O_2$ (TTDD = 4,7,10-triaza-tridecanedioic acid) was measured [14]. $[Co(TTDD)]_2O_2$ was obtained as a 4.17×10^{-4} M stock solution by bubbling oxygen through the cobalt(II) complex solution at pH 9.5 and ionic strength 1 M (KCl). This stock solution was mixed in the manual stopped-flow apparatus with equal volumes of 0.5 M acetate buffer of pH 4.52 (298 K) and the changes in absorbance were recorded at 13 wavelengths between 340 and 470 nm.

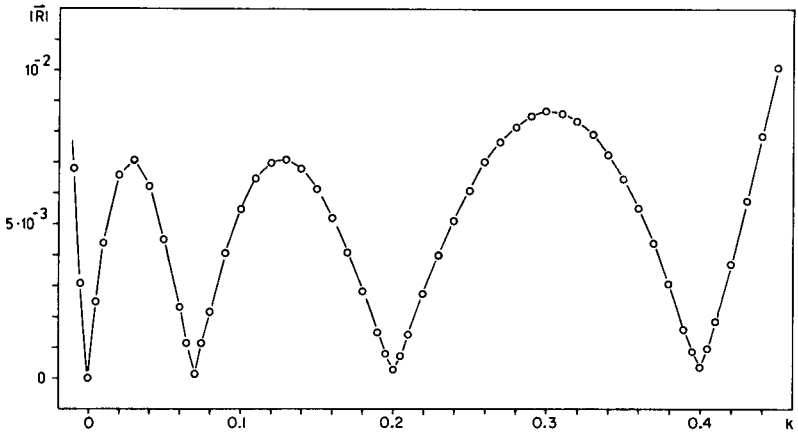


Fig. 2. The length of the residual vector |R| against the rate constant k. Data from the mentioned computer simulations, i.e., a set of three exponentials and a constant, were used.

Twice-distilled water was used for all experiments. Commercial chemicals of analytical grade were used without further purification.

RESULTS AND DISCUSSION

Equilibrium study: use of linear coefficients

The method was tested first on an equilibrium example, the complexation of copper(II) with dien. The spectra of 62 different solutions were recorded at 26 wavelengths between 750 and 550 nm and written as matrix **Y** (eqn. 2). Program EV was used to calculate the eigenvectors of the second-moment matrix **M** (eqn. 3) belonging to **Y**. The set of measured absorbances could be represented by five eigenvectors via $\mathbf{Y} = \mathbf{VX}$. The overall standard deviation in absorbance was found to be 1.55×10^{-4}. For the apparatus used, it has been shown [3] that the standard error of an individual absorbance measurement is about 1.5×10^{-4}, hence in this example five eigenvectors are necessary to represent the measured spectra, i.e. the rank of **Y** is 5. In other words, in the equilibrium system studied, five linearly independent absorbing species are expected. This is in accordance with an earlier study [13], where a model with Cu^{2+}, CuL^{2+}, $CuLH^{+}_{-1}$, CuL_2^{2+} and CuL_2H^{3+} (L = dien) was proposed.

Replacement of the measured absorbances by the calculated linear coefficients **X** in the Newton—Gauss—Marquardt minimization procedure made it possible to fit all the data simultaneously. With initial estimates of the parameters $\log K_1 = 5.88$ ($Cu^{2+} + LH^{+} \rightleftharpoons CuL^{2+} + H^{+}$), $\log K_2 = 9.39$ ($CuLH^{+}_{-1} + H^{+} \rightleftharpoons CuL^{2+}$), $\log K_3 = 3.58$ ($CuL^{2+} + LH^{+} \rightleftharpoons CuL_2H^{3+}$) and $\log K_4 = 4.92$ ($CuL^{2+} + L \rightleftharpoons CuL_2{}^{2+}$) [13], the minimum of the squares sum was found after 5 iterative cycles. The computing time was 20 min per cycle. The calculated $\log K$ values were 6.17 ±0.01, 9.20 ±0.03, 3.61 ±0.02 and

4.95 ±0.03 respectively. The overall standard deviation of an individual measurement was 9×10^{-4} absorbance unit, about 0.3% of the measured absorbances. This value is much higher than the standard deviation obtained by representing the spectra simply by $\mathbf{Y} = \mathbf{VX}$, where a standard deviation of 1.55×10^{-4} was found. This difference may be introduced by the uncertainties of the pH measurement or may reflect an incorrect choice of the chemical model. Additional experiments will be needed to decide this.

Thus, in quite a short time, it is possible to achieve a quality of fit as good as that obtained for the simplest case considered earlier [2], the titration of *p*-nitrophenol. In Fig. 3, a comparison between measured and calculated titration curves at 750, 654 and 550 nm (1:2 mixtures) shows the excellent fit of the data. In Fig. 4, the spectra of the chemical species in the system Cu^{2+}-dien are shown. All the complexes have similar spectra, i.e. it is almost impossible to follow the concentration of one particular species at a selected wavelength where the other species do not interfere. Simultaneous treatment of the data collected at several wavelengths is therefore essential.

Undoubtedly, in the field of equilibrium studies, the proposed method makes it possible to perform calculations in a short time with inexpensive desk computers. Such calculations previously required expensive computer time on large machines.

Kinetic studies: use of the linear search method

Computer simulation. As mentioned by Sylvestre et al. [5], the described method of reducing the data set by representing the spectra as linear combinations of the eigenvectors can be applied to kinetic systems. The concentrations of the reacting species are calculated not from stability constants but by the solution of the differential equations.

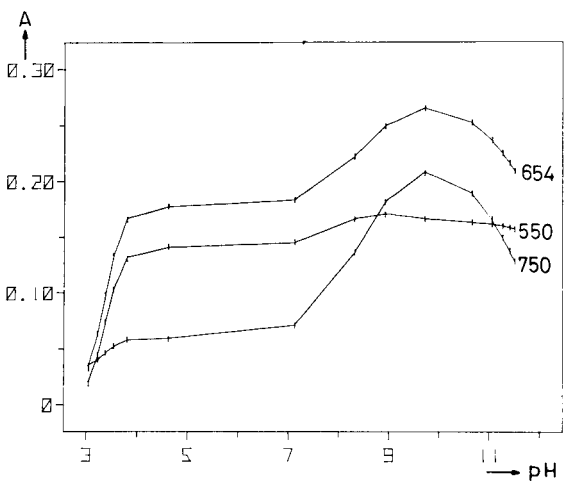


Fig. 3. Calculated (connected points) and experimental ($I = 6\sigma$) absorbances against pH at 550, 654 and 750 nm for the 1:2 mixtures in the Cu-dien system.

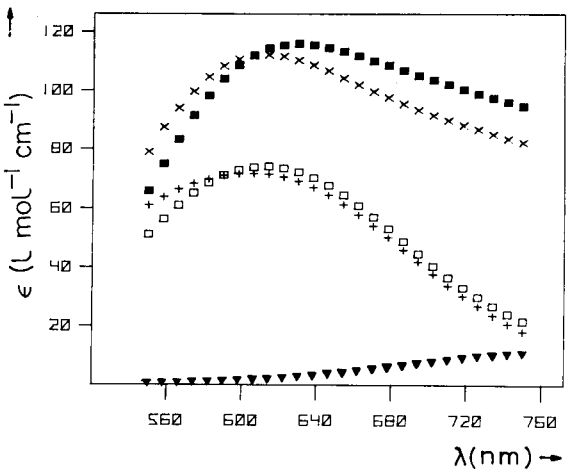


Fig. 4. Molar absorptivities against wavelength for the absorbing species in the Cu-dien system: (▾) Cu^{2+} (aq); (+) $CuLH_{-1}$; (□) CuL; (■) CuL_2; (×) CuL_2H.

The new method of one-dimensional search was tested with two systems, the first of which is a computer simulation of kinetic data. A matrix **Y** of dimensions 15 × 30 was calculated by applying expression (9). The rate constants k_1—k_3 were chosen as 0.07, 0.2 and 0.4, and the fourth, k_4, as 0 to give a constant vector. The difference between two measurements was one time unit throughout. The components a_{ij} of the matrix **A** were random numbers between −1 and +1 for $j \neq 4$ and between 0 and +1 for $j = 4$ (the final spectrum of a solution does not have negative absorbances, whereas the other linear parameters may be negative). A random noise with $\sigma = 10^{-4}$ was added to the components of the resulting matrix **Y**. This procedure simulated the measurement of a system with four reacting species; the spectrum was measured at 15 wavelengths and at 30 times ($t_i = i$).

Analysis of the eigenvectors of the second-moment matrix **M*** (eqn. 3′) showed the following decrease of the variances: 0.521, 1.51×10^{-3}, 6.67×10^{-5}, 9.99×10^{-9} and 9.09×10^{-9} for one to five eigenvectors. The expected variance of about 10^{-8} ($= \sigma^2$) was reached with four eigenvectors. This is the correct result for a set of three exponentials and a constant.

Plotting the length of the residual R as a function of k yielded Fig. 2. The minima of $|\mathbf{R}|$ were found for k 0.06998, 0.2001 and 0.3996, which is in excellent agreement with the true values. The fourth minimum for $k = 0$ is a consequence of the constant introduced by $k_4 = 0$. This minimum is shown only for completeness.

Chemical example. A second example of the one-dimensional search method is the acid dissociation of the μ-peroxo complex formed between Co^{2+}, TTDD and oxygen. The spectrum of the solution containing the μ-peroxo complex, $[Co(TTDD)]_2O_2$ in acetate buffer (pH 4.52) was measured at 13 wavelengths between 340 and 470 nm at 50 times.

The rank of the resulting matrix Y, computed as described, was found to be 3. Since the final spectrum is not zero at all wavelengths, this result means that a sum of two exponentials together with a constant must be used. The plot of |R| as a function of k (as in Fig. 2) showed two minima for two rate constants of 9.3×10^{-3} and 2.6×10^{-2}.

Evaluation by the first described method, based on linear coefficients as illustrated by the equilibrium system, yielded two rate constants: 9.0×10^{-3} and 2.6×10^{-2}. The agreement is not as good as in the theoretical example because (a) the standard deviation of a single absorbance measurement in the apparatus used is about 5×10^{-4}, and (b) choosing random numbers for A (eqn. 9) gives better results than similar a_{ij} as found in practice.

These results can be explained by the following mechanism: two isomers of the $[Co(TTDD)]_2O_2$ complex are formed and they react in acidic solution with different rate constants to yield the final products oxygen, cobalt(II) and the protonated ligand

Isomer α $[Co(TTDD)]_2O_2$ $\xrightarrow{H^+}$ $O_2 + Co^{2+} + TTDDH_3^+$

Isomer β $[Co(TTDD)]_2O_2$ $\xrightarrow{H^+}$ $O_2 + Co^{2+} + TTDDH_3^+$

Analysis of the linear parameters in A shows that the two isomers occur with a relative frequency of roughly 15% of the faster reacting isomer and 85% of the slower one. This statement is based on the assumption that both isomers have similar molar absorptivities. The similar ligand field for both isomers makes this assumption probable [15].

The one-dimensional search method needs kinetic measurements at several wavelengths. For determination of the rate constants of a well known system, it still may be accurate to use the classical methods at one particular wavelength, but if unknown systems with a greater degree of complexity are under investigation, measurements at several wavelengths are highly recommended.

First, the eigenvector analysis gives the "dimension" of the problem, i.e. the number of reacting species. Then, representation of the spectra or of the kinetic curves as linear combinations of the eigenvectors reduces the amount of data to be handled or, in the kinetic example, reduces the dimension of the parameter space to one. As a consequence of the speed and the numerical stability of the linear search, even starting values for the rate constants are unnecessary.

In conclusion, even inexpensive desk computers can be used successfully in kinetic or equilibrium studies if efficient mathematical methods are used.

REFERENCES

1 W. A. E. McBryde, Talanta, 21 (1974) 979.
2 A. D. Zuberbühler and T. A. Kaden, Talanta, 26 (1979) 1111 and references therein.
3 T. A. Kaden and A. D. Zuberbühler, Talanta, 26 (1979) 563.
4 H. R. Schwarz, H. Rutishauser and E. Stiefel, Numerik symmetrischer Matrizen, Teubner, Stuttgart, 1972.

5 E. A. Sylvestre, W. H. Lawton and M. S. Maggio, Technometrics, 16 (1974) 353.
6 H. Gampp, M. Maeder and A. D. Zuberbühler, Talanta (in press).
7 I. G. Sayce, Talanta, 15 (1968) 1379.
8 D. J. Leggett and W. A. E. McBryde, Anal. Chem., 47 (1975) 1085.
9 J. J. Kankare, Anal. Chem., 42 (1970) 1322.
10 T. W. Anderson, Ann. Math. Stat., 34, 1 (1963) 122.
11 D. E. Metzler, C. M. Harris, R. L. Reeves, W. H. Lawton and M. S. Maggio, Anal. Chem., 49 (1977) 864A.
12 W. Walter, Gewöhnliche Differentialgleichungen, Springer, Berlin, 1976.
13 T. A. Kaden and A. D. Zuberbühler, Helv. Chim. Acta, 54 (1971) 1361.
14 M. Maeder, Thesis, Univ. Basel, 1980.
15 H. Mäcke, M. Zehnder, U. Thewalt and S. Fallab, Helv. Chim. Acta, 62 (1979) 1804.

Analytica Chimica Acta, 122 (1980) 315—322
Computer Techniques and Optimization

LABORATORY COMPUTER SYSTEMS AND THE ROLE OF THE HUMAN INTERFACE

ENGELBERT ZIEGLER

Max-Planck-Institut für Kohlenforschung, D-4330 Mülheim a. d. Ruhr (F.R.G.)

(Received 24th January 1980)

SUMMARY

The importance of the "human interface" between the operator and a laboratory computer system is emphasized. Some rules for communication through dialogue programs and command languages as well as for the presentation of computer output are given. The degree of automation of the evaluation of analytical data versus interactive treatment is discussed.

When computer systems were first introduced in analytical laboratories about ten years ago, the main task for the designers of such systems was to interface the hardware of the analytical instruments with the computers and hopefully to arrive at programs that accomplished some useful work. Beyond this hardware interface, however, there exists a human interface (i.e. communication between the operator and his equipment) that was neglected in many of these early laboratory systems. This shortcoming must be blamed for the failure of many systems: working with them was a cumbersome, frustrating and error-prone affair, resulting in poor acceptance by the operators.

Working with a laboratory computer system, the operator must interact with the instrument and with the computer or, if the equipment is of modern design, with a combination of instrument and computer, where the latter is more or less integrated into the instrumental hardware. The operator has to set up the instrument, prepare the computer, run the experiment and finally obtain and interpret the results. The computer may participate in all these steps to varying degrees, depending on design philosophy and sophistication. One important facet in the design of a laboratory computer system is the question of for whom it is intended. How much does the operator need to know about its internals? Is it sufficient to know what to do for a certain type of measurement or some kind of data manipulation and to consider the rest as a black box? Or should the operator know about the algorithms used in the software? Should software modifications or additions be feasible through the user? Or, to go further, should the user have an education in information theory and computer science? Obviously, there cannot be a general answer to these questions. The answer depends on the environment,

the kind of work to be performed and on the type of people who are going to operate the system. Therefore, a system normally has to be designed not only for a particular analytical application but also for a range of humans with different backgrounds and divergent interests. It should certainly not be built for use through the programmer.

MANUAL INSTRUMENT PREPARATION VS. COMPUTER-CONTROLLED INSTRUMENTATION

After a sample has been submitted, the first step required from the operator is the preparation of the measuring equipment. For this purpose, conventional instruments are equipped with assorted dials and switches for the manual setting of instrumental conditions, such as voltages, pressures, time constants, etc. In newer instruments, the computer normally carries out part of this function. Instruments are available without any provisions for manual operation other than an on/off switch, and without any indicators; all parameters must be selected by typing information into the computer terminal via a keyboard. Similarly, the status of selected parameters may be requested; the display screen serves as a substitute for the indicators used in the conventional approach. This is certainly a very elegant way of maintaining control of the experimental parameters: the software always know all the instrumental conditions and therefore programmed algorithms that depend on some critical values of instrumental parameters can be made to work in all cases with the actually selected settings.

The operator, who may be a laboratory technician is, however, turned into a typist, and tends to lose contact with the physical process involved in the measurement. Even for each reading of an instrument parameter, some keyboard action is required, which is inherently different from the glance at various indicators required in a conventional instrument. This disadvantage becomes even more obvious if more than one person is involved in a measurement. Moreover, many people prefer analog meters to the digital representation of parameters.

To find the proper mixture of conventional manual actions and computer controls via keyboard typing is certainly not an easy task, but it is clear that some psychological considerations have to be taken into account, in addition to technical requirements and operational elegance.

PROGRAMMED DIALOGUE AND COMMAND LANGUAGES

Irrespective of how many operations are still performed manually, the interaction between the human operator and the computer program will be the main vehicle for exchange of information with the equipment. Therefore, a closer look at this form of communication is necessary. Three different approaches are possible and are implemented in modern systems: (1) the computer program conducts the operator through question-and-answer

sequences for setting the necessary parameters; (2) without being explicitly asked for type-in, the operator states the parameters he wants to set by typing specific commands taken from a set of commands within an application-oriented command language; and (3) the user identifies a predefined "method", for which all input parameters are already stored in the computer.

Rules necessary for a question-and-answer dialogue

There are several rules which should be observed in the design of a dialogue between the user and the computer software.

(1) Distinction between the occasional user and the routine user. The routine user is well acquainted with the various questions posed by the software and knows the parameters that have to be specified as well as their units and input formats. Long explanations, especially if printed at a slow terminal, will annoy the experienced user. Other desirable features for this type of user are as follows: the system should select default values or "last used" values for individual parameters or an entire set of parameters if the user wants to short-cut the dialogue; also the user may want to repeat selected parts of the dialogue in order to correct one or more parameters, in which case the system should not force him to go through a large sequence of questions again. In contrast, the first-time or occasional user of the computer system will not be familiar with the dialogue and therefore needs a more explanatory, extensive form of the questions asked. He should have the option to ask for HELP at the beginning of the dialogue as well as on individual questions. He should also be offered a set of default or recommended parameter values.

(2) Provision of understandable error-messages and warnings. It is not of much help to the user if his incorrect input is honored by a "?" or "Eh?"; of little more use are the messages "ERROR CODE 00703" or "LOOKUP (3)", even if there happens to be documentation for these error codes somewhere, explaining that the system has tried to read a non-existent data file. The program could explain instead: "No parameters specified yet for this instrument". Or instead of "ILL.MEM.REF. AT 03605", there could be a message such as: "More than 250 peaks found. Program GCMANY should be used in this case." Such comments would advise the user how to correct for the situation instead of just reporting an error.

Furthermore, an incorrect input should never be allowed to crash the program, resulting in an operating system error printout instead of producing an application-oriented error message.

(3) Check for input errors as early as possible. The validity of each input parameter should be checked immediately. It is a frustrating experience to learn, after a dozen or more questions, that the first one had already been answered incorrectly and to be forced to go through the entire procedure again.

(4) Avoidance of complex syntax for the data input. All too often the user is forced to queue up a string of numbers, such as "7,0,−1.,,3.1,7.,,3,....." (double comma means: default). This kind of parameter input is confusing

and error-prone. Normally, not more than 2 or 3 input numbers should be requested in one line; moreover, the user should not be expected to know whether the system expects a floating point number or an integer or alphameric input.

(5) Use of upper and lower case characters. Texts written with upper- and lower-case characters are more easily read than wholly upper-case texts. As almost all terminals available nowadays can print all ASCII characters, this capability should also be used by the software.

Application-oriented command languages

Several laboratory computer systems provide a special command language that is tailored for the application to which the computer system is dedicated. In this case, parameter input is not conducted through questions by the software. Such commands may look like:
".TEMP 103", ".PRESS 600", ".RANGE 29/600".

The use of such a command language may speed up communication with the computer: the only parameters referred to are those which the operator really wants to change. However, the user is expected to know all the necessary commands for correct operation of the equipment. Therefore, this approach is more oriented towards the daily routine operator and is less appropriate for the occasional user.

Of course, most of the rules outlined in the preceding section are also valid for the design of a command language. Explanatory help, for instance, could be requested by typing a question-mark after the keyword of the command (e.g., ."TEMP?"). The computer program then prints some help message: (e.g., "Temperature at the inlet in centigrade"). Other systems allow type-in of a prompting control character after the keyword: the computer then continues the line with explanatory text.

Use of predefined "methods"

If a certain type of analysis with a given set of parameters is performed frequently, the use of predefined "methods", characterized by this set of parameters is a convenient way of reducing the amount of keyboard type-in. The parameters are specified only once, either via a question-and-answer session or through a sequence of commands stored in an "indirect command file" and are then written into a disk file for long-term storage. This form of parameter specification appears to be hazardous in situations where the requirements change frequently: the operators tend to use "methods" that proved to be successful in previous measurements without checking individual parameter settings against the new requirements.

PRESENTATION OF INFORMATION

The output from a computer is provided for the user who should be able to read and comprehend it easily. In order to achieve this goal, some general rules should be followed.

(1) Restriction of the amount of information to matters of real interest. Programmers tend to output a lot of additional information just because it has been generated during the program execution and is readily available, or because they need it for debugging purposes. For the end user, too much information can be confusing and, if printed in hardcopy form, a waste of paper. Sometimes, however, the user may find that for a special case he really should have more data available. To be prepared for this case, the software should write more extensive reports into a disk file that can be listed if necessary. Examples of this are peak lists in chromatography where the extended list may include data such as half-widths of peaks, A/D sample numbers of peak start and peak end, absolute intensities, noise levels, etc., whereas the standard output may be restricted to retention times and peak-area percentages.

(2) Use of an easily recognizable form of presentation. Reports that are properly formatted enhance legibility. More important, however, is the choice of the right kind of presentation. A low-resolution mass spectrum, for instance, should not be presented as a table of mass numbers and intensities but as a bar graph of the type to which spectroscopists are accustomed.

The use of (analog) graphic output is often preferable to the (digital) table printout. Numerical information should, however, be inserted into the drawings where appropriate. For example, mass numbers should be written for the most significant masses of a spectrum; examples of computer-generated spectra plots have been published [1].

A difficult problem which has not yet been fully solved is the proper representation of structural formulas. Very often the name of a chemical substance is difficult to comprehend and the chemist requires to see a drawing of the structure. In many applications, structures are stored internally in the form of connection tables or in the form of a line notation. Programs have been developed that derive structure drawings from such internal representations [2]. In more complicated cases, the structures so generated will not resemble a conventional drawing, and are not easily comprehended (Fig. 1a). Better results are obtained [3], if the structure, as drawn by a chemist, is stored internally in a form that will exactly reproduce this structure (Fig. 1b).

(3) Provision of means for the interactive composition of reports. In order to produce analysis reports that are tailored to the chemical problem of interest, it should be possible to disregard all or most of the residual information. In gas chromatography, for instance, a chemist is often interested only in 2 or 3 components; in this case the report should not include some 50 other peaks identified by the computer, although it should state that 50 other peaks exist, together with their cumulative intensity. In mass spectrometry, the analyst may want to compose a report consisting of a graphical representation of the most relevant mass spectrum measured and of text stating his interpretation or his answer to the question the chemist had in mind.

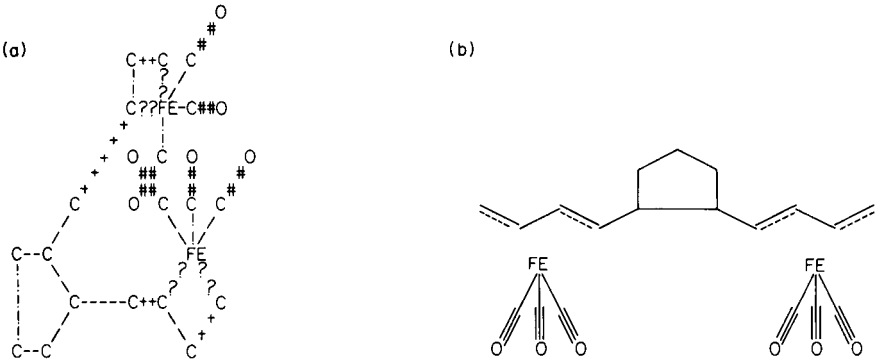


Fig. 1. Computer-generated structure drawings as derived from (a) a connection table representation (b) internally stored plot commands.

BALANCE OF AUTOMATED AND INTERACTIVE EVALUATION OF ANALYSIS

There are many laboratories where the same type of analysis for samples of very similar nature is used over and over again. In such a situation, the evaluation and interpretation of a measurement may be totally automated, i.e. no interaction by the analyst is necessary. Under other circumstances, however, e.g. in research laboratories, a changing variety of samples is presented for analysis and this requires flexible adaptation of evaluation procedures to the individual problem. Often enough, these procedures may not be selected without first looking at the measured data. In more complicated cases, the evaluation of data and subsequent interpretation of the analytical information constitutes a stepwise process where the operator must interact extensively with the computer system at all stages. The place of work for this operator is no longer his desk but the keyboard of the computer terminal. Only the development of more sophisticated computer programs that take over part of the interpretative work will return the analytical chemist to his desk or the laboratory in pursuit of more advanced work.

This change in style of work can be illustrated by the example of a mass spectroscopist who is involved in g.c.-m.s. First the repetitive scan method [4, 5] applied by laboratory computer systems freed him from closely watching the measurement, because it was no longer necessary to initiate a spectrum scan manually when a g.c. peak emerged, for example. The first steps of data evaluation have also been taken over by the computer, i.e., the detection of mass peaks and the assignment of mass numbers. After this step, the spectroscopist could use the computer interactively as a tool in the further evaluation process; using programmed dialogues, he could select the most interesting spectra, apply corrections for overlapping and background, etc. These were still very time-consuming processes even with the spectroscopist working at the computer terminal. With the further development of software, the next level of sophistication introduced programs that used

advanced algorithms to extract the relevant spectra automatically and apply the necessary corrections. With the development of suitable library search methods, for compound identification as well as for structure elucidation, the next level of automation became obvious: not only the extraction of relevant spectra but also their comparison to a large set of reference spectra is now done automatically, resulting in an identification or in a "hit-list" of compounds with similar spectra for each of the relevant spectra measured. After this (partial) interpretation by the computer has been accomplished, the spectroscopist returns to the terminal only in ambiguous cases or in order to accomplish special treatments of spectra. Working in this way is much more satisfactory than struggling through a dialogue with the computer system at all steps of data evaluation and interpretation. The computer should do as much as possible without human interaction but should provide means to allow such interaction at any stage of the analytical process.

Another example of this philosophy is the software for the processing of chromatograms at the Max-Planck-Institut in Mülheim [6]. The detection of g.c. peaks is evaluated by a sophisticated program in a uniform way for all chromatograms without the need for human interaction. For very difficult cases, e.g., extreme baseline drifts, trace analysis for components on the slope of solvent peaks, etc., it is possible to control the peak detection visually with a graphics display terminal. The operator can check for correct baseline reconstruction, redefine the baseline, select by means of a cross-hair cursor sections from a chromatogram for independent analysis, and request a rerun of the analysis program with modified peak-detection parameters. But all of this is necessary only in the rare cases where normal analysis with default peak-detection parameters is insufficient.

It may, however, be dangerous to automate only part of a task without fully replacing a person. Again, experience from the mass spectrometry laboratory of the Mülheim Institute may serve as illustration. In the original approach, the calibration of the mass scale had been automated: the technician just inserted the calibration sample, and called for the calibration program which, after he had pushed a button, established a mass number versus time relationship without further human interaction. The technician was in no way involved in the calibration process and did not know how the program worked, which contributed to his feeling of working at a very low level. The situation improved remarkably after redesign of this calibration process: now the technician has to assign mass numbers for two or three key peaks in critical parts of the calibration spectrum shown to him on the display screen. Furthermore he gets indications about the state of instrument, like noise levels, slope of scan, etc. and has to decide if the calibration is good enough. By this approach he is more involved in the measurement process, with increased responsibility, and so is more motivated.

CONCLUSION

Many years of experience in working with laboratory computer systems have stressed the importance of the human interface. As in probably all

cases of automation, the lower-level part of the human work is either entirely taken over by the computer or is simplified to a degree where it becomes difficult to find willing technicians. In contrast, the high-level part of the analytical work, i.e., the interpretation of results, becomes much more satisfactory. This statement will, however, only be true if the computerized tools can be applied easily and fit into the human style of doing analytical work. Costs of hardware are no longer a valid excuse for designing software with poor human interfaces and the programming tools to create the appropriate programs are available.

REFERENCES

1 H. Damen, D. Henneberg and B. Weimann, Anal. Chim. Acta, 103 (1978) 289.
2 R. J. Feldmann and S. R. Heller, J. Chem. Soc., 12 (1972) 48.
3 E. Ziegler and K. Boll, Anal. Chim. Acta, 103 (1978) 237.
4 R. A. Hites and K. Biemann, Anal. Chem., 40 (1968) 1217.
5 D. Henneberg, B. Weimann and E. Ziegler, Chromatographia, 7 (1974) 483.
6 G. Schomburg, F. Weeke, B. Weimann and E. Ziegler, Chromatographia, 7 (1974) 477.

Analytica Chimica Acta, 122 (1980) 323—326
Computer Techniques and Optimization

Short Communication

SEMIAUTOMATIC MICRODENSITOMETER—MINICOMPUTER SYSTEM FOR SPARK-SOURCE MASS SPECTROMETRY

M. VICZIÁN and P. A. PETIK

MS Laboratory of the Central Institute for Mining Development, H-1300 Budapest, Pf. 115 (Hungary)

(Received 26th October 1979)

SUMMARY

Quick and reliable evaluations of spectra produced by a spark-source mass spectrometer are possible with a system based on a Zeiss GII microdensitometer, a Texas Instruments TI-59 programmable calculator and a PC-100A printer. Only the selection of spectral lines requires operator control.

The main problems to be dealt with in this laboratory are analyses of a great variety of geological samples, high-purity materials, and inorganic residues of natural waters. Accordingly, the structure of the mass spectra (i.e. the main component and trace element ions, the kinds and quantities of multiple-charged ions and complex ions) may vary significantly from sample to sample. For this reason, total automation of the evaluation system with exclusion of an operator experienced in identification of elements (isotopes) in various spectra would require a complicated and uneconomic computer program. The much simpler system adopted is based on a Texas Instruments programmable calculator. Selection of spectral lines is done by the operator, but the detection of the peak maxima of the line "fractional blackening" [1] and the calculation of the element concentrations are done automatically [2—5].

Experimental

The outline of the microdensitometer—minicomputer system is shown in Fig. 1. The output voltage of the silicon photovoltaic cell of the Zeiss GII microdensitometer is proportional to the light intensity penetrating through the measured area of the photoplate selected for measurement. The higher the "fractional blackening" B, the smaller the voltage. The photovoltaic cell is followed by a unit converting the $B = 1 - T$ function, [1] i.e. producing an output voltage that follows the change of the inverse of the transparency T.

The signal obtained is supplied by the amplifier unit (60 dB gain, low

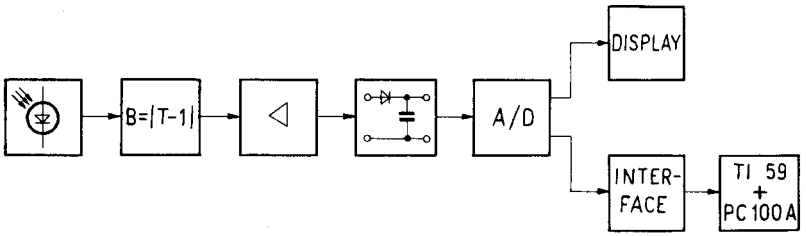


Fig. 1. Outline of the microdensitometer—minicomputer system.

output impedance) to the circuit that detects the maximum value; this is an analog network and gives higher measuring speed as well as simpler circuitry. This unit is followed by a 12-bit analog-to-digital (A/D) converter and a 7-segment decimal display. If the measured value is considered by the operator to be appropriate for evaluation, a start signal is given by a microswitch. The value in question, followed by an R/S (run—stop) command, is thus forwarded to the computer via the interface unit.

On the keyboard of the TI programmable minicomputer one contact belongs to each key forming a matrix with X_{jk} elements where j represents the row and k the column in which the contact is located; 14 wires (9 rows and 5 columns) join the contact matrix panel to the printed circuit panel inside the TI-59. These 14 points are wired to a 14-contact dual in-line IC socket which is built into the right-hand side of the cover of the calculator. Making contact between a row and a column simulates touching the particular key. Table 1 lists the contacts. The operation of the interface unit is explained by Fig. 2. The data entry cycle begins with the start signal, the rising edge of which starts the monoflop. Until its $\bar{Q}$ output is on level L, the start—stop oscillator runs and its output pulses are counted by a decimal counter. The first, third and fifth pulses open the AND gates of the 10^2, 10^1 and 10^0 decimal places respectively. The AND gates are followed by the BCD/DEC decoder, the ten outputs of which drive the matrix unit consisting of 11 reed relays. The R/S command generated by the seventh pulse of the start—stop oscillator is forwarded by the eleventh relay. This part of the interface unit is shown in detail in Fig. 3.

TABLE 1

List of contacts on the keyboard of the TI-59

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rows	1	—	7	6	—	4	5	—	2	9	1	3	8	—
Columns	—	*E*	—	—	*D*	—	—	*C*	—	—	*B*	—	—	*A*

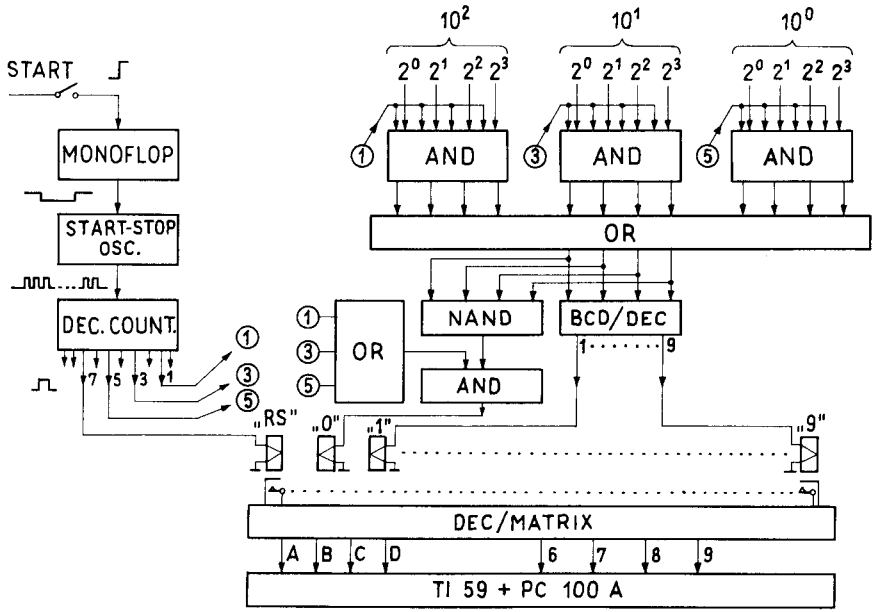


Fig. 2. Operation of the interface unit.

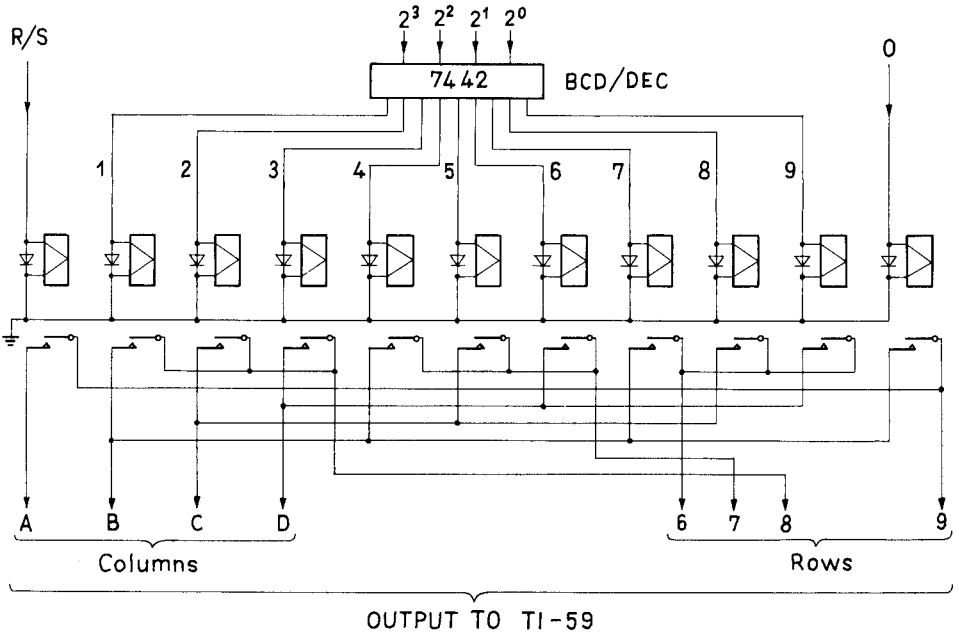


Fig. 3. The matrix unit.

Results and discussion

The basic equation for the determination of elemental concentrations by using a reference element [6] is

$$C_i = \frac{E_R}{E_X} C_S \frac{A_R}{A_X} \frac{M_X}{M_R} \frac{1}{S_r}$$

where C is the concentration, E is the estimated exposure at which an isotope of the element would produce a preselected line blackening, A is the isotope abundance, and M is the atomic mass. Subscripts R and X represent the reference element and analytical element respectively, and S is the relative sensitivity factor. E is calculated from the line blackening data (B) by using $E = E_k \times 10^{-W_k/\gamma}$, where E_k is the exposure applied, γ is the slope of the blackening curve, and W_k is the Seidel-transformed fractional blackening

$$W_k = \log\left(\frac{1}{1-B_k} - 1\right)$$

The $M_X/A_X S_r$ constants as well as the γ and E_k values are stored in the computer memory [7]. After the data have been entered and the R/S command given, the computer prints out the calculated element concentration. When the line fractional blackening has been measured at several exposures E_k, the results obtained can be averaged at a special command. Obviously, the use of this equipment provides all the advantages of on-line data systems, especially the availability of the final results within a few seconds.

With the help of the interface unit, the TI-59 can be used with any instrument having BCD data output. The various data-handling and evaluating programs can be stored on magnetic cards. The program can be changed very quickly as required. If another type of computer is used, the only modification necessary in the interface unit is to the connections of the contacts in the matrix unit (see Fig. 3). The data entry can be expanded by using a longer monoflop pulse and increasing the number of AND gates at the BCD input of the interface.

The authors thank Dr. I. Cornides for encouragement and advice.

REFERENCES

1 A. J. Ahearn, Trace Analysis by Mass Spectrometry, Academic Press, New York, 1972, p. 112.
2 E. J. Millett, J. A. Morice and J. B. Clegg, Int. J. Mass Spectrom. Ion Phys., 13 (1974) 1.
3 R. J. Conzemius, D. J. Adduci, G. O. Foss and H. J. Svec, Anal. Chem., 48 (1976) 1647.
4 L. S. Dale and R. N. Whittem, Appl. Spectr., 30 (1976) 461.
5 M. Viczián, Dissertation. Tech. Univ. Budapest (1978).
6 A. J. Ahearn, Mass Spectrometric Analysis of Solids, Elsevier, Amsterdam, 1966, p. 98.
7 M. Viczián, I. Kada and R. Rétháti, Publ. Hung. Mining Res. Inst., 19 (1976) 219.

Analytica Chimica Acta, 122 (1980) 327—330
Computer Techniques and Optimization

Short Communication

THEORY OF ERROR APPLIED TO FACTOR LOADINGS RESULTING FROM COMBINATION TARGET FACTOR ANALYSIS

EDMUND R. MALINOWSKI

Department of Chemistry and Chemical Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 (U.S.A.)

(Received 27th February 1980)

Summary. The theory of error for target factor analysis is used to derive a simple equation from which the root-mean-square error in the factor loadings can be calculated. The method is applied to a problem in gas—liquid chromatography and is shown to agree with errors estimated by the 'jackknife' method.

Factor analysis is a multivariant technique [1—4] which attempts to express a data matrix $[D]$ as a product of a row-factor matrix $[R]$, called the score matrix, and a column-factor matrix $[C]$, called the loading matrix

$$[D] = [R][C] \tag{1}$$

Since $[D]$ consists of r rows and c columns its size is $r \times c$. If there are n controlling factors, $[R]$ is an $r \times n$ matrix and $[C]$ is an $n \times c$ matrix. The first stage in factor analysis involves the use of a mathematical process called principal component analysis, which is used to define the coordinates of the factor space, and might more appropriately be called principal coordinate analysis. This process reveals the true dimensions of the factor space, n, and yields mathematical, abstract expressions for $[R]$ and $[C]$. The second stage involves transforming the abstract matrices into physically significant matrices, $[\bar{R}]$ and $[\bar{C}]$, which have real chemical meaning, so that $[D] = [\bar{R}][\bar{C}]$. This is achieved by employing target factor analysis, which focuses attention on the score matrix and provides a method for finding a target transformation matrix $[T]$ to convert the abstract row matrix into a real score matrix. Mathematically, this is achieved by carrying out the multiplication: $[\bar{R}] = [R][T]$. This same transformation matrix is used to obtain the real loading matrix $[\bar{C}]$ as follows

$$[\bar{C}] = [\bar{T}]^{-1}[C] \tag{2}$$

The recently developed theory of error for abstract factor analysis [5, 6] explains how the experimental error in the data matrix mixes into the abstract data reproduction process expressed by eqn. (1). When applied to target factor analysis [7], this theory shows how the data errors enter into the real score matrix. The present communication extends the theory for the purpose of determining the error in the factor loadings.

Theoretical derivation

Since the theory of error for abstract factor analysis [5, 6] forms the basis of the current investigation, the same notation will be used as in the early work. According to the theory, an element of the abstract column-factor matrix can be expressed by

$$c_{jk} = c^*_{jk}(r^*_{ik}/r_{ik}) + \sigma^{\neq}_{jk} \tag{3}$$

where the subscripts give the location (row and column) of the element in the respective matrix; the asterisks refer to pure data points free from experimental error and $\sigma^{\neq}_{jk}$ is the error contribution to the loading. If the errors are small, eqn. (3) reduces to $c_{jk} = c^*_{jk} + \sigma^{\neq}_{jk}$, which can be expressed in matrix form as

$$[C] = [C^*] + [E^{\neq}_c] \tag{4}$$

where $[C^*]$ is the hypothetically pure column-factor matrix and $\sigma^{\neq}_{jk}$ is an element of the error matrix $[E^{\neq}_c]$ associated with the column-factor matrix.

When the transformation described by eqn. (2) is applied to the column matrix expressed by eqn. (4), then

$$[\bar{C}] = [\bar{C}^*] + [T]^{-1}[E^{\neq}_c] \tag{5}$$

assuming that $[T]$ has little or no error so that $[T]^{-1}[C^*] = [\bar{C}^*]$, where $[\bar{C}^*]$ is the real loading matrix, free from error. Hence the error in the real loading matrix $[\bar{E}_c]$ is given by

$$[\bar{E}_c] = [\bar{C}] - [\bar{C}^*] = [T]^{-1}[E^{\neq}_c] \tag{6}$$

To convert the error matrix into root-mean-square errors, consider the product $[\bar{E}_c][\bar{E}_c]^T$, where $[\bar{E}_c]^T$ is the transpose of $[\bar{E}_c]$. According to eqn. (6)

$$[\bar{E}_c][\bar{E}_c]^T = [T]^{-1}[E^{\neq}_c][E^{\neq}_c]^T\{[T]^{-1}\}^T \tag{7}$$

Since $[E^{\neq}_c]$ is composed of random errors, positive and negative, $[E^{\neq}_c][E^{\neq}_c]^T$ is essentially a diagonal matrix. According to the theory of error for abstract factor analysis [5, 6] the j-th diagonal element of $[E^{\neq}_c][E^{\neq}_c]^T$ is given by $\Sigma_{k=1}^{k=c}(\sigma^{\neq}_{jk})^2 = c(RE)^2/\lambda^{\neq}_j$, where RE is the real error in the data matrix and $\lambda^{\neq}_j$ is the j-th primary eigenvalue. Hence

$$[E^{\neq}_c][E^{\neq}_c]^T = c\,(RE)^2\,[\lambda^{\neq}]^{-1} \tag{8}$$

where $[\lambda^{\neq}]^{-1}$ is a diagonal matrix consisting of the reciprocals of the primary eigenvalues. Placing eqn. (8) into eqn. (7) gives

$$[\bar{E}_c][\bar{E}_c]^T = c\,(RE)^2\,[\tilde{T}][\tilde{T}]^T \tag{9}$$

where $[\tilde{T}] = [T]^{-1}\,[\lambda^{\neq}]^{-1/2}$. The j-th diagonal element of $[\bar{E}_c]\,[\bar{E}_c]^T$ equals $c(EFL)^2_j$ where $(EFL)_j$ is the RMS of the error in the j-th factor loading. Hence,

$$(EFL)_j = RE(\tilde{T}_j \cdot \tilde{T}_j)^{1/2} \tag{10}$$

where $\tilde{T}_j$ is the j-th row of $[\tilde{T}]$ and $\tilde{T}_j \cdot \tilde{T}_j$ is the scalar product.

Equation (10) affords a method of calculating the error in the factor loadings since RE is readily available from abstract factor analysis and T_j is available from target factor analysis. A little afterthought shows that $(EFL)_j$ originates from the error in the data matrix. In essence, it is the analog of the real error in the predicted target vector (REP) (see eqn. (33) in [7]).

Application

Weiner et al. [8] developed a 'jackknife' method for determining the error in the factor loadings. The jackknife method consists of performing a complete combination target factor analysis on a series (r in number) of reduced data matrices, each reduced matrix consisting of the original data matrix minus a single row. Thus a total of r loading matrices are generated, each differing slightly because of the intrinsic nature of the errors in the data. The corresponding elements (i.e. loadings) in these matrices are then used to obtain standard deviations (and confidence limits, if so desired) for each loading.

Weiner et al. [8] treated by target factor analysis the retention volumes of a set of organic solutes (Table 1) which were subjected to gas—liquid chromatography on an aqueous electrolyte phase of tetraethylammonium bromide. As expected two principal factors emerged. Target transformation was used to identify the two controlling factors as being: (1) the surface area of the coated liquid phase per gram of packing; and, (2) the volume of stationary phase per gram of packing. Theoretically, the associated loading

TABLE 1

Loadings corresponding to adsorption (K_A) and partition (K_L) constants obtained by target factor analysis

Solute	K_A(× 10^{-5} cm)		K_L	
	Correlation[a]	Covariance[b]	Correlation[a]	Covariance[b]
Carbon tetrachloride	7.83 ± 0.12	7.88	3.40 ± 0.79	3.06
Methylene chloride	10.56 ± 0.22	10.55	25.51 ± 1.67	25.29
Chloroform	30.94 ± 0.25	31.01	20.36 ± 1.17	19.97
Benzene	26.06 ± 0.21	26.03	16.80 ± 1.08	16.39
Toluene	76.90 ± 0.58	76.86	21.44 ± 2.46	20.72
n-Hexane	5.11 ± 0.11	5.16	0.40 ± 0.62	0.13
Cyclohexane	4.32 ± 0.17	4.40	0.83 ± 1.33	0.26
n-Heptane	12.06 ± 0.11	12.32	4.41 ± 4.40	3.86
2-Methylheptane	27.48 ± 0.28	27.64	1.14 ± 2.43	−0.22
n-Octane	34.19 ± 0.10	34.19	1.23 ± 0.86	1.23
RMS error	±0.25[c]	±0.23[d]	±2.00[c]	±1.4[d]

[a]Obtained by Weiner et al. [8] using correlation about the origin and the jackknife method.
[b]Obtained in the present study by using covariance about the origin.
[c]Root-mean-square of the errors obtained by the 'jackknife' method.
[d]Obtained directly by means of eqn. (10).

factors should correspond to two distribution constants, K_A and K_L, respectively, where K_A is the adsorption constant at the gas—liquid interface and K_L is the bulk liquid distribution constant. The constants obtained by Weiner et al., using correlation about the origin, are shown in columns two and four of Table 1 together with the errors estimated by means of the jackknife method.

When the same data matrix was subjected to covariance about the origin, and target-transformed, the loading constants listed in columns three and five of Table 1 were obtained. Equation (10) was used to obtain the root-mean-square (RMS) errors in the factor loadings, shown at the bottom of Table 1. The fact that these errors compare so closely to the RMS errors obtained by the jackknife method gives credence to the jackknife method as well as to eqn. (10) derived from the theory of error.

REFERENCES

1 E. R. Malinowski and D. G. Howery, Factor Analysis in Chemistry, John Wiley, New York, 1980.
2 D. G. Howery, in R. F. Hirsch (Ed.), Statistics, 1977 Eastern Analytical Symposium, Franklin Institute Press, Philadelphia, 1978.
3 D. G. Howery, Am. Lab., 8 (1976) 14.
4 P. H. Weiner, Chem. Tech., (May) (1977) 321.
5 E. R. Malinowski, Anal. Chem., 49 (1977) 606.
6 E. R. Malinowski, in B. R. Kowalski (Ed.), Chemometrics: Theory and Applications, ACS Symp. Ser. 52, Am. Chem. Soc., Washington, DC, Chap. 3, 1977.
7 E. R. Malinowski, Anal. Chim. Acta, 103 (1978) 339.
8 P. H. Weiner, H. L. Liao and B. L. Karger, Anal. Chem., 46 (1974) 2182.

Analytica Chimica Acta, 122 (1980) 331—336
Computer Techniques and Optimization

Short Communication

THE POINT METHOD FOR ELECTROCHEMICAL DIGITAL SIMULATION

D. BRITZ

Department of Chemistry, Aarhus University, 8000 Aarhus C (Denmark)

(Received 8th October 1979)

Summary. The point method is recommended in preference to the box method in the development of programs for description of diffusion and convection models.

The extensive application of finite differences for the solution of electrochemical transport problems, i.e. digital simulation, began in 1964 [1, 2] although Randles was the first electrochemist to use the method [3]. Digital simulation consists of the (time-) stepwise solution of a discretized form of the continuous partial differential equation(s) describing a particular mass transport system, often simply a diffusion equation. There are two formally different approaches: the system is treated either by dividing the transport space into a number of small volume elements, each of which is assumed to be quite homogeneous, or by specifying space and time at a number of points. In the first case, the full Fick's diffusion equation

$$(\partial C/\partial t) = D(\partial^2 C/\partial x^2) \tag{1}$$

where C = concentration, t = time, x = distance and D is the diffusion coefficient, is never used. Instead, the more fundamental equation is preferred.

$$\mathrm{d}N/\mathrm{d}t = AD\ \mathrm{d}C/\mathrm{d}x \tag{2}$$

where N is the number of moles of substance diffusing across area A of a plane normal to x. In digital form, if we consider three adjacent slices of space, each δx thick and numbered $i-1$, i and $i+1$ with different concentrations (Fig. 1a), eqn. (2) can be applied to the pairs $(i-1, \mathrm{i})$ and $(i, i+1)$ to yield

$$\delta N_1 = AD[(C_i - C_{i-1})/\delta x]\,\delta t \text{ and } \delta N_2 = AD[(C_{i+1} - C_i)/\delta x]\,\delta t$$

Subtracting these two discrete equations gives the total number of moles flowing into element i and dividing by the volume, $A\delta x$, the concentration change in that element is explicitly obtained as

$$\delta C_i = (D\delta t/\delta x^2)(C_{i-1} - 2C_i + C_{i+1})$$

This sort of physical approach [4] has the advantage, in theory, that it is very close to the physical model and the full partial differential equation is not needed [5]. This is the method adopted by Feldberg who has been followed

by almost all electrochemists; it may be called the box method. In many papers dealing with simulation publications, a time error of $\frac{1}{2}\delta t$ is implied. Fortuitously, this often "improves" results and so may seem to be appropriate, but it also reflects the box philosophy; even time is divided into windows. The alternative, mathematical approach [4] proceeds from the full partial differential equation, and specifies concentrations at points in space (Fig. 1b). This was the approach used by Randles [3], based on a paper by Emmons [6], and seems to be adopted mainly by non-electrochemists. There is a basic set of discretization formulae for first and second derivatives to apply to the process: these can be arrived at by common sense or by Taylor's expansion plus truncation.

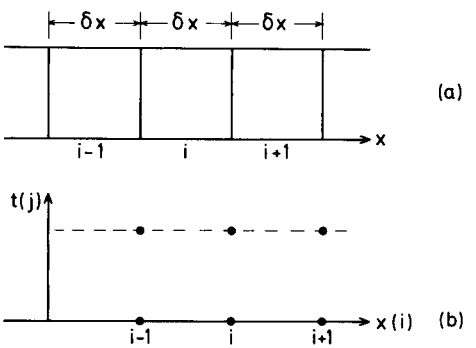
Given a C-set, at $t = (j-1)\delta t$, eqn. (1) can produce a new C_i for $t = j\delta t$ by discretizing it to

$$(C_i' - C_i)/\delta t = D(C_{i-1} - 2C_i + C_{i+1})/\delta x^2 \tag{3}$$

which is the same result. The discretization formulae used here are given in standard texts [4, 7—9]; $\partial C/\partial t$ was here approximated by a forward difference. The second (x-) derivative can be considered as the derivative of the first derivative which leads to

$$\frac{\partial^2 C}{\partial x^2} \approx \frac{1}{\delta x}\left(\frac{C_{i+1} - C_i}{\delta x} - \frac{C_i - C_{i-1}}{\delta x}\right) = \frac{1}{\delta x^2}(C_{i-1} - 2C_i + C_{i+1}) \tag{4}$$

This mathematical approach may be called the point method. One may ask why it is better, if the same result is obtained. The reason for the preference



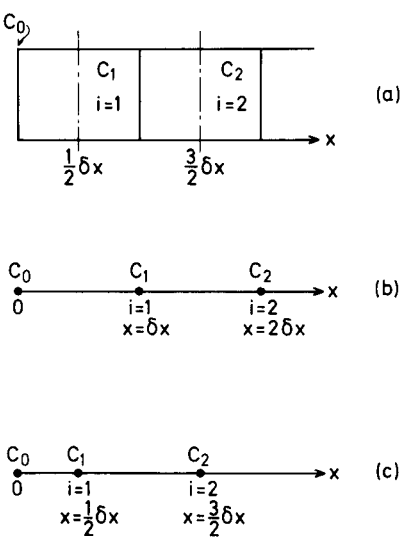


Fig. 1 (left). Illustration of the box and point models.

Fig. 2 (right). Selection of boxes.

is that, for problems involving geometric or hydrodynamic complications, the point method yields the discrete equation much more easily than the box method. Also, for most systems, fully worked-out partial differential equations are already available [10], e.g. the diffusion equation for cylindrical or spherical coordinates, and for convection components. Thus the physical box method has no great advantage. Practical examples are given below. First, however, a small problem must be dealt with. If, in testing the equivalence of the methods, two simple programs are written to compute e.g., current vs. time for a diffusion-controlled, potential-step experiment (simulating the Cottrell equation), then it will probably be found that the box method is more accurate for the same number of time steps and the same δx. The reason is that the natural first box (i.e. the box immediately next to the electrode with index $i = 1$) has its centre situated at $x = \frac{1}{2}\delta x$ (Fig. 2a), whereas the first point will probably be chosen at $x = \delta x$ (Fig. 2b). This produces two different discrete equations for the concentration change for $i = 1$: for the box method, the equation is

$$\delta C_1 = (D\delta t/\delta x^2)(2C_0 - 3C_1 + C_2) \tag{5}$$

and for the point method

$$\delta C_1 = (D\delta t/\delta x^2)(C_0 - 2C_1 + C_2) \tag{6}$$

where C_0, or $C(x = 0)$, is established by the particular system under scrutiny.

Joslin and Pletcher [11] suggested the use of unequal intervals, crowding closer together near the electrode, for better accuracy and/or for conserving computer time. The box method can be regarded as a crude beginning for this; it is, in fact, equivalent to moving the first point to $\frac{1}{2}\delta x$ (Fig. 2c). If this is done, then eqn. (5) is also obtained for the point method, with the resulting accuracy improvement. This has become standard practice in this laboratory. The advantage of the point method by reference to the following two situations will now be demonstrated.

Spherical diffusion

Figure 3(a) shows the box model for a spherical system; the boxes are thin shell elements of thickness δr. The concentration change in shell i is derived during time interval δt. For mole flux N_2 into shell i from shell $i + 1$, $\delta N_2 = D\delta t A_2(C_{i+1} - C_i)/\delta r$, and for flux N_1 from shell i into shell $i - 1$, $\delta N_1 = D\delta t A_1(C_i - C_{i-1})/\delta r$. Subtraction of the expression for δN_1 from δN_2 gives the net mole flux into shell i

$$N_i = (D\delta t/\delta r)[A_2(C_{i+1} - C_i) - A_1(C_i - C_{i-1})] \tag{7}$$

where A_1 and A_2 are the two areas of contact between the three elements. These are given by the approximate equations $A_1 = 4\pi r_{i-1}^2$ and $A_2 = 4\pi r_i^2$. The concentration change δC_i is obtained by dividing eqn. (7) by the volume δV_i of element i. This volume can best be approximated by the area of the spherical plane at the centre of element i multiplied by δr: $V_i = 4\pi\, \delta r\, [(r_{i-1} + r_i)/2]^2$. Combining these produces

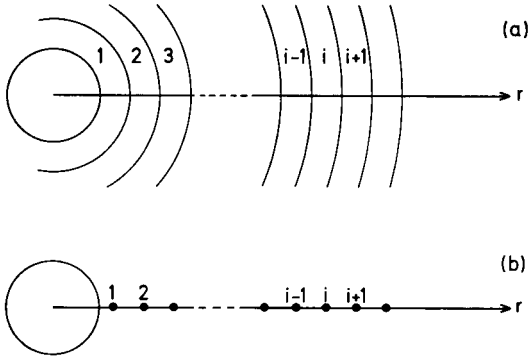


Fig. 3. The box and point models for a spherical system.

$$C_i = [4D\delta t/\delta r^2(r_{i-1} + r_i)^2]\,[r_i^2(C_{i+1} - C_i) - r_{i-1}{}^2(C_i - C_{i-1})] \tag{8}$$

Only simple algebra is needed to justify the approximate substitutions, $r_i^2 \approx r_{i-1}r_{i+1}$ and $(r_{i-1} + r_i)^2 \approx 4\,r_{i-1}r_i$. Substitution of these expressions into eqn. (8) and tidying up gives

$$\delta C_i = [D\delta t/\delta r^2 r_i]\,[r_{i+1}(C_{i+1} - C_i) - r_{i+1}(C_i - C_{i-1})] \tag{9}$$

$$\text{and}\,\delta C_i = [D\delta t/\delta r^2][C_{i-1} - 2C_i + C_{i+1} + (\delta r/r_i)(C_{i+1} - C_{i-1})] \tag{10}$$

Figure 3b is the scheme used for the point method. The diffusion equation in spherical coordinates is [10] $(\partial C/\partial t) = D(\partial^2 C/\partial r^2) + (2\,\partial C/r\,\partial r)$. The first term on the right-hand side is treated as for cartesian coordinates (eqn. 3). The second term can best be expressed by a central difference expression. This gives the discretized equation

$$C_i = D[(\delta t/\delta r^2)(C_{i-1} - 2C_i + C_{i+1}) + (2\delta t/r_i)(C_{i+1} - C_{i-1})/(2\delta r)]$$

$$= (D\delta t/\delta r^2)[C_{i-1} - 2C_i + C_{i+1} + (\delta r/r_i)(C_{i+1} - C_{i-1})] \tag{11}$$

which is identical with eqn. (10) derived by the box method. Clearly, the point method allows a far more rapid discretization process. The reason, of course, is that the long derivation for the box approach is analogous to the analytical derivation of the above-mentioned spherical diffusion equation, which is the starting point for the point method.

Moreover, in the discretization for the point method, little can go wrong except, perhaps, through neglect of a central-difference formula for the spherical term. In the box method, however, several steps require careful thought (e.g., choice of radial distances, means of approximating areas, volume etc). If the simplifying procedure that leads from eqn. (8) to eqn. (9) or (10) is not used, then eqn. (8) will need far more computer time than eqn. (11).

Convection terms

If the transport, besides diffusion, also involves convection, the electrolyte velocity must be considered. This arises, for example, in simulations involving the dropping mercury electrode (DME) or rotating disk and ring/disk electrodes, as well as hydrodynamic voltammetry. The procedure is to compute concentration changes caused by diffusion and solution movement separately and then add them. The box method computes the convection term for, say, box i in Fig. 1, by calculating from the velocity how far box i moves during time interval δt, in multiples of box thickness δx. The box is then located at that distance from box i and its concentration is assigned to box i. This can be refined by interpolating concentrations between box centres to achieve a more accurate value at a non-integral distance from the centre of box i. Obviously, this will be a somewhat involved procedure. The point method again starts from the partial differential equation, now containing a convection term [12]; e.g., in a one-dimensional problem with velocity $v(x)$ dependent on x:

$$\partial C/\partial t = [D(\partial^2 C/\partial x^2) - [v(x)(\partial C/\partial x)] \tag{12}$$

Rather similarly to the discretization of the spherical diffusion equation, the discretization of eqn. (12) is a simple one-step procedure, if $v(x)$ is known; again, a central-difference formula is best used to represent the $\partial C/\partial x$ term.

Implicit expressions

There are simulation problems when an implicit finite-difference technique is required. There are various implicit forms (see, e.g. Richtmyer and Morton [9]). This always leads to a system of simultaneous equations. In principle, this can be done for box elements also, but in practice the point method becomes inevitable, especially if a 'recipe' is taken from a textbook, as these are usually point-oriented.

Recently, it became necessary to write a simulation program for the diffusion current at a DME. The expanding plane model yields the differential equation [13]

$$(\partial C/\partial t) = [D(\partial^2 C/\partial x^2) + (2x\partial C/3t\partial x)] \tag{13}$$

which, again, is simple to discretize. In contrast, the box method runs into the convection-term difficulties already mentioned. The equation for a spherical DME is

$$(\partial C/\partial t) = [D(\partial^2 C/\partial x^2) + (2\partial C/r\partial x)] - v(r)(\partial C/\partial x) \tag{14}$$

where (as for the expanding-plane model), x is the distance from the growing drop. Here, the convection velocity $v(r)$ is a function of the radial distance from the drop and is, in fact, the only complication for the point method; it is not difficult, however, to derive the appropriate general expression. Both these programs worked well, showing no evidence of instability, and produced current values in excellent agreement with the Ilković equation

and that with Koutecký's [14] spherical correction applied. Feldberg [15] has achieved similar performance using the box method but with somewhat greater discretization difficulties. In conclusion, the point method should be readopted by electrochemists.

The author acknowledges with thanks the valuable and helpful correspondence with Dr. S. W. Feldberg.

REFERENCES

1 S. W. Feldberg and C. Auerbach, Anal. Chem., 36 (1964) 505.
2 S. W. Feldberg, in A. J. Bard (Ed.), Electroanal. Chem., 3 (1969) 199.
3 J. E. B. Randles, Trans. Faraday Soc., 44 (1948) 327.
4 W. F. Ames, Numerical Methods for Partial Differential Equations, Academic Press, New York, 1977.
5 N. Davids and R. L. Berger, Commun. A.C.M., 7 (1964) 547.
6 H. W. Emmons, Q. Appl. Math., 2 (1944) 173.
7 G. E. Forsythe and W. R. Wasow, Finite-Difference Methods for Partial Differential Equations, Wiley, New York, 1960.
8 A. R. Mitchell, Computational Methods in Partial Differential Equations, Wiley-Interscience, New York, 1969.
9 R. D. Richtmyer and K. W. Morton, Difference Methods for Initial-Value Problems, Interscience, New York, 1967.
10 J. Crank, The Mathematics of Diffusion, Clarendon Press, Oxford, 1975.
11 T. Joslin and D. Pletcher, J. Electroanal. Chem., 49 (1974) 171.
12 V. G. Levich, Physicochemical Hydrodynamics, Prentice-Hall, Englewood Cliffs, 1962.
13 K. J. Vetter, Elektrochemische Kinetik, Springer, Berlin, 1961.
14 J. Koutecký, Czech. J. Phys., 2 (1953) 50.
15 S. W. Feldberg, personal communication, 1979.

JOURNAL OF ANALYTICAL AND APPLIED PYROLYSIS

This new international journal brings together, in one source, qualitative and quantitative results relating to:

- Controlled thermal degradation and pyrolysis of technical and biological macromolecules;
- Environmental, geochemical, biological and medical applications of analytical pyrolysis;
- Basic studies in high temperature chemistry, reaction kinetics and pyrolysis mechanisms;
- Pyrolysis investigations of energy related problems, fingerprinting of fossil and synthetic fuels, coal extraction and liquefaction products.

The scope includes items such as the following:

1. Fundamental investigations of pyrolysis processes by chemical, physical and physicochemical methods.
2. Structural analysis and fingerprinting of synthetic and natural polymers or products of high molecular weight.
3. Technical developments and new instrumentation for pyrolysis techniques in combination with chromatographic or spectrometric methods, with special attention to automation, optimization and standardization.
4. Computer handling and processing of pyrolysis data.

Pyrolysis is applied in a wide range of disciplines. This journal is therefore of value to scientists in such diverse fields as polymer science, forensic science, soil science, geochemistry, environmental analysis, energy production, biochemistry, biology and medicine.

The journal publishes original papers, technical reviews, short communications, letters, book reviews and reports of meetings and committees. The language of the journal is English. Prospective authors should contact one of the editors.

Subscription Information:
1980: Volume 2 (in 4 issues),
US $ 80.00/Dfl. 156.00 including postage.

Full information on contents of Volume 1 (1979) and a free sample copy are available on request.

P.O. Box 211,
1000 AE Amsterdam
The Netherlands

52 Vanderbilt Ave
New York, N.Y. 10017

The Dutch guilder price is definitive.
US $ prices are subject to exchange rate fluctuations.

7162

ELSEVIER

STATISTICAL TREATMENT OF EXPERIMENTAL DATA

By J.R. GREEN, *Lecturer in Computational and Statistical Science, University of Liverpool, U.K. and* D. MARGERISON, *Senior Lecturer in Inorganic, Physical and Industrial Chemistry, University of Liverpool, U.K.*

PHYSICAL SCIENCES DATA 2

This book first appeared in 1977. In 1978 a revised reprint was published and in response to demand, further reprints appeared in 1979 and 1980. Intended for researchers wishing to analyse experimental data, this work will also be useful to students of statistics. Statistical methods and concepts are explained and the ideas and reasoning behind statistical methodology clarified. Noteworthy features of the text are numerical worked examples to illustrate formal results, and the treatment of many practical topics which are often omitted from standard texts, for example testing for outliers, stabilization of variances and polynomial regression.

What the reviewers had to say:

"The index is detailed; the format is good; the presentation is clear; and no mathematics beyond calculus is assumed".
–CHOICE

"A lot of thought has gone into this book and I like it very much. It deserves a place on every laboratory bookshelf".
–CHEMISTRY IN BRITAIN

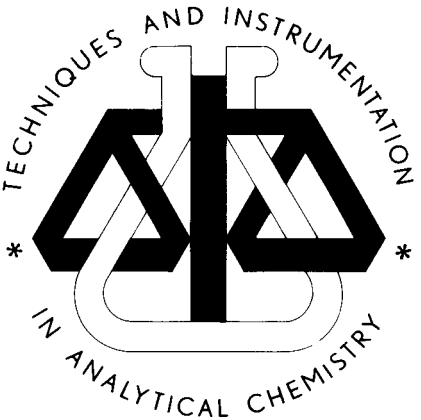
1977. Reprinted 1978, 1979, 1980.
xiv + 382 pages
US $39.25/Dfl.90.00
ISBN: 0-444-41725-7

ELSEVIER

P.O. Box 211, 1000 AE Amsterdam, The Netherlands.
52 Vanderbilt Ave., New York, NY 10017.

The Dutch guilder price is definitive. US$ prices are subject to exchange rate fluctuations.

;valuation and)ptimization of Laboratory ſlethods and ,nalytical Procedures



Survey of Statistical and athematical Techniques

L. MASSART, A. DIJKSTRA *and* L. KAUFMAN.

th contributions by S. Wold, B. Vandeginste *and* Y. Michotte

:chniques and Instrumentation in Analytical Chemistry - Volume 1

is book provides detailed treatment, in a single volume, of formal methods for timization in analytical chemistry. It is a comprehensive and practical handbook ıich no analytical laboratory will want to be without.

l aspects of optimization are discussed, from the simple evaluation of procedures the organization of laboratories or the selection of optimal complex analytical ɔgrammes. Quantitative discrete analysis as well as qualitative and continuous easurement techniques are evaluated.

ıe book consists of 30 chapters divided into 5 main parts.The main sections are: aluation of the Performance of Analytical Procedures, Experimental Optimization, ɔmbinatorial Problems, Requirements for Analytical Procedures, and Systems)proach in Analytical Chemistry.

ıis work will be of practical value not only to those involved with optimization ɔblems in analytical chemistry, but also to those in related fields such as inical chemistry or specialized fields such as chromatography. Because it scusses the application of many mathematical techniques in analytical chemistry, is book will also serve as a general introduction to the new field of Chemometrics.

78 1st Reprint 1979 xvi + 596 pages US $68.25 / Dfl. 140.00
BN 0-444-41743-5

ELSEVIER

P.O. Box 211,
1000 AE Amsterdam
The Netherlands

52 Vanderbilt Ave
New York, N.Y. 10017

ı Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

7113

VOL. 122, No. 3 ANALYTICA CHIMICA ACTA September 15, 1980

(Computer Techniques and Optimization, Vol. 4, No. 3)

CONTENTS

Printed in The Netherlands.