Vol. 103 No. 3 September 15, 1978
(Computer Techniques and Optimization, Vol. 2 No. 3)
ISSN 0378-4304

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 1978 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	96/1	96/2	97/1	97/2	98/1	98/2	99/1	99/2	100	101/1	101/2	102
Section on Computer Techniques and Optimization			103/1			103/2			103/3			103/4

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 to:
for *Analytica Chimica Acta:* 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.

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. 102, p. 253 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 8 volumes in 1978. The subscription for 1978 (Vols. 96–103) is Dfl. 1000.00 plus Dfl. 120.00 (postage) (Total approx. US $486.96). The subscription for the *Computer Techniques and Optimization* sections only (Vol. 103) is Dfl. 125 plus Dfl. 15.00 (postage) (Total approx. US $60.87). Journals are sent automatically by air mail 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. 1–87) are available at Dfl. 115.00 (plus postage).

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 our Journal Information Center, 52, Vanderbilt Avenue, New York, NY 10017. Tel: (212) 867-9040.

Analytica Chimica Acta, 103 (1978) 201—212
Computer Techniques and Optimization

PATTERN RECOGNITION AND BLIND ASSAY TECHNIQUES APPLIED TO FORENSIC SEPARATION OF WHISKIES

BO E. H. SAXBERG, DAVID L. DUEWER, JAMES L. BOOKER[‡]
and BRUCE R. KOWALSKI*

Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195 (U.S.A.)

(Received 10th May 1978)

SUMMARY

One problem in forensic science is the detection of counterfeit whisky (specifically, replacement of the contents of a bottle). A simple, inexpensive forensic method is required which would reliably distinguish between samples of Chivas Regal, as an example, and samples of less expensive whiskies. Pattern recognition techniques were applied to the results of gas chromatographic analysis according to the blind assay method. When the amounts of isoamyl alcohol and acetaldehyde present (as revealed in the chromatograms) were used perfect separation of the two classes of samples was possible.

The filling of a liquor bottle with any other liquor than that contained in the bottle at the time it received its tax stamp is a crime under the laws in many states and the U. S. Government (Section 5301, Title 26, U.S.C.). This crime represents consumer fraud and also causes loss of revenue to the distiller whose products are misrepresented. The most common form of violation is the refilling of a bottle containing well-known and expensive brands of liquor with less expensive liquors of lower quality. Successful prosecution of a violator requires that an adequate showing be made that the contents of the bottle are not of the brand of liquor represented by the label. Unfortunately, proving this crime is quite difficult because the only sample normally available to the forensic chemist is that which is purchased by the drink by an enforcement agent.

To establish a method whereby this showing may be made by using only the sample obtained in a field investigation, the techniques of pattern recognition [1] have been applied to one well-known brand of Scotch whisky. The method developed has two features which recommend it for crime laboratory use: first, the analysis of a single sample can be used to determine accurately that the sample is or is not the particular brand in question; secondly, this determination may be made on the basis of gas chromatograms which can be

[‡] Office of the Attorney General, Wyoming State Crime Laboratory, Cheyenne, Wyoming 82001, U.S.A.

easily obtained with unmodified equipment available in virtually any laboratory. In a blind assay [2], the amounts of various constituents in samples are initially determined without the identity of the constituents being known. The gas chromatogram (g.c.) lends itself to this technique, as the relative areas under the peaks provide a measure of relative quantities. With the aid of pattern recognition, it should then be possible to select the constituents relevant to the problem at hand. These constituents can then be studied extensively, including, if so desired, a determination of their exact composition and identity. A great advantage of the blind assay technique is that there is no need to spend time identifying constituents (i.e., peaks in the chromatogram) irrelevant to the problem.

For this particular forensic problem, blind assay is especially appropriate in that identification of the discriminatory compounds is not necessary. The chromatograms can be used as they are (without interpretation); in fact, once the important discriminatory features (peaks) are known, only those which have been selected as the most important features in separating classes of whiskies need be considered.

Because this study was designed for real forensic application, the complexity of the technique was an important consideration. If possible, injection on the g.c. column of a sample drawn straight from the bottle is preferred. This proved quite satisfactory, and it removes time and cost considerations involved in extensive concentration, extraction or other preparative techniques, which have little significance in the research laboratory, but would certainly affect the utility of the method in real applications.

Some freedom was given in the various parameters for obtaining the g.c. chromatogram, e.g. allowing the initial temperature to vary by as much as ±2°C). It was felt that if it was possible to obtain good results with these data, then any forensic laboratory (possibly even with a portable g.c.) could apply the method confidently and obtain as good, if not better, results.

EXPERIMENTAL

Materials

To establish the parameters which may be used to describe Chivas Regal Scotch whisky, control samples of 24 different blendings were obtained from Joseph E. Seagram & Sons. Samples of 34 brands of Scotch whisky (Table 1) were purchased from retail liquor stores for comparison with Chivas Regal. Only those whiskies which are less expensive than Chivas Regal (at Washington State Liquor Board prices as of August 1977) were considered, and selections were based on availability.

Nine samples were taken from an old bottle of Chivas Regal which had been opened many times to the air, so that only about 30 ml was left. Since all other samples were taken from previously unopened bottles, this "standard" category of 9 samples (one sample for each of the nine days that samples were analyzed) gives not only a check on the variability of the results for a single

TABLE 1

Description of Samples

Category	Number of Samples	Source
Standard (ST)	9	Single used bottle
Chivas (C)	24	New bottles from different batches } Chivas
Non-Chivas ($\bar{C}$)	34	New bottles of different non-Chivas brands

List of non-Chivas whiskies	Cost per oz. at August 1977 (¢)	List of non-Chivas whiskies	Cost per oz. at August 1977 (¢)
1. Bulloch Lade's B & L	26.6	18. MacArthurs	25.0
2. Churchill	23.0	19. Mackintosh	24.4
3. Cutty Sark	36.1	20. McGregor's Perfection	22.7
4. Dewars White Label	36.3	21. McMasters	26.2
5. E L Scotch	23.0	22. Monarch	23.0
6. Grand Macnish	27.4	23. Muirheads	23.8
7. House of Stuart	23.4	24. Old Mr. Boston	23.8
8. Hudson's Bay Best	25.0	25. Old Smuggler	26.2
9. Inver House Green	25.8	26. P & T Special Select	24.4
10. J. & B. Rare	36.3	27. Passport	29.1
11. J. W. Dants	24.4	28. Peter Dawson Special	27.4
12. John Begg Blue Cap	25.4	29. Plaid Piper	21.7
13. Johnnie Walker Red Label	36.6	30. Scoreby Rare	24.6
14. King George IV	24.4	31. Scottish Majesty	22.6
15. King James	24.4	32. Seagrams 100 Pipers	24.6
16. King William IV	26.0	33. Teachers Highland	36.3
17. Lauders	26.6	34. Ushers Green Stripe	28.0
Chivas Regal	53.1		

bottle, but also an indication of how "aging" affect the chromatogram. This difference must be compensated for, as in a forensic application one must take into account how long a bottle may have been in use. Table 1 summarizes the constituents of the different categories used in this study.

Gas chromatographic conditions

A 10 μl sample of whisky was injected onto a stainless steel column (6 ft. $\times$ 1/8 in. o.d.) packed with 10% Carbowax 20 M on 80/100 mesh HP Chromosorb W(AW-DWCS). A flame ionization detector was used on a Hewlett-Packard Model 402 gas chromatograph. The following temperature conditions were used: isothermal at 50°C for 5 min after injection, then temperature-programmed to 185°C at 15°C min^{-1}, and then isothermal at 185°C; the inlet temperature was 210°C, and the detector temperature 190°C. Nitrogen carrier gas flow was 75 cm^3 min^{-1} at 24°C.

Data processing

The heights and half-widths of the peaks in the g.c. patterns constituted the "measurements". The height was measured as the distance from the top

of the peak to the intersection of the vertical with the tangential baseline. The half-width was measured as the width of the peak at half the peak height. The peak areas were calculated as the product of peak height and peak width at half-height [3] and normalized to the total area under all measured peaks. Normalization serves two functions: it reduces any error from variations in the injection volume, and it reduces error from variations in the density of the whisky. The latter was found to be important, as room temperature varied from 21°C to 38°C, and a 50% alcohol—water mixture has a considerable change in density with temperature.

Each Scotch had 17 observable peaks in common (and no peaks unique to any category were observed), hence there were 17 "features" for each sample, each feature being a normalized peak area calculated as described above. In nearly every Scotch sample, the data are complete; no measurement had more than 5% missing values (these arose from temporary instrument failure) for each category (Chivas Regal, non-Chivas Regal, and Standard Chivas Regal). The missing data were filled in by the measurement mean taken over all the samples in the same category as the sample with the missing data.

Pattern recognition programs

The following is a description of the statistical and pattern recognition tools used, all of which are segments of the pattern recognition system ARTHUR [4].

Autoscale. Scaling of measurements (or features) is done by subtraction of the measurement mean and division by the square root of the measurement (or feature) scatter-about-the-mean [5]. This produces a feature with zero mean and a unit variance. If the measurements, or original features, are regarded as defining a (number-of-measurements)-space (n-space), this procedure may be thought of as adjusting the axes to a common origin and normalizing the the variance of the data along each axis to unity.

Fisher weight. This is a quantitative estimate of the utility of a given measurement for separating two categories [6]. The weight is the ratio between the square of the difference between the category means and the sum of the squared category standard deviations.

Karhunen—Loève (K—L) projection. This is the best (in the sense of a least-mean-squared-error) linear mapping onto a lower-order orthogonal space of a given degree [7]. The axes of the K—L space are obtained through principal component (PC) analysis of the data.

K-Nearest neighbor analysis (KNN). This classification method, which utilizes the distance between samples in the n-space as its decision criterion [8], is used to classify a sample in the category which contributes the greatest number of the k-nearest known samples. The Euclidean n-space distance was used in this study; however, many other definitions of distance can and have been used. Only the closest k samples are used in making any given classification. The importance of a given measurement (or feature) in making the decisions is proportional to its contribution to the distance calculation, (i.e., by the axis length fixed by prior scaling.

Least-squares discriminant analysis (LSDA). This classification method uses linear discriminant functions derived from least-squares fitting of the data [9]. One discriminant function is derived for each category studied by computing the least-squares discriminant through all non-members of the category (assigned a dependent-variable value of 0.0) to all category members (assigned a dependent-variable value of 1.0). Data points are classified into the category whose function generates the largest value. All the data intended for model construction are used in the derivation of each discriminant function; there is no exclusion based on category, etc. The importance of each independent variable (measurement or feature) is determined through the fitting procedure and is independent of prior linear scaling.

Statistical isolinear multicategory analysis (SIMCA). This classification method uses linear discriminant functions derived from disjointed principal component analysis of the data [10]. One set of functions is derived for each category studied by computing the category-mean and a specified number of the principal components (PCs). Samples are classified into the category whose PC model best reproduces the data. Only the data points which are members of a given category are used in determining the model functions for that category. The importance of each measurement or feature in classification is determined by its contribution to the category covariance matrices; this is a function both of prior scaling and intermeasurement/feature correlation.

RESULTS AND DISCUSSION

An example of a typical g.c. pattern showing the 17 peaks used in this study is given in Fig. 1. Table 2 gives the distribution of the data for each feature in each category. The aim is to find the features which are best at separating the Chivas Regal (C) from the non-Chivas ($\bar{C}$), but which do not separate C from the standard Chivas Regal category (ST). If a feature does separate C from ST, then it does so on the basis either of a difference in original composition, of "aging" (how long the bottle has been in use); both separations are undesirable, since both categories consist of samples from the same brand of whisky which should be identified as such. By finding features which will separate C and ST from $\bar{C}$, but which do not separate C from ST, features will be selected which will distinguish samples of Chivas Regal from many different non-Chivas Regal whiskies, even if the Chivas Regal sample comes from a well-used bottle. Thus, in a forensic application, this will reduce the contribution of previous use in explaining "misclassification" of a sample allegedly from the expensive brand.

The data were first autoscaled to zero mean and unit variance along each axis. By the Karhunen—Loève transformation, a projection (Fig. 2) is obtained that indicates a distinct difference between the ($\bar{C}$) and the C and ST whisky samples. This projection offers sufficient encouragement to proceed to attempt to identify exactly which features are most responsible for this separation.

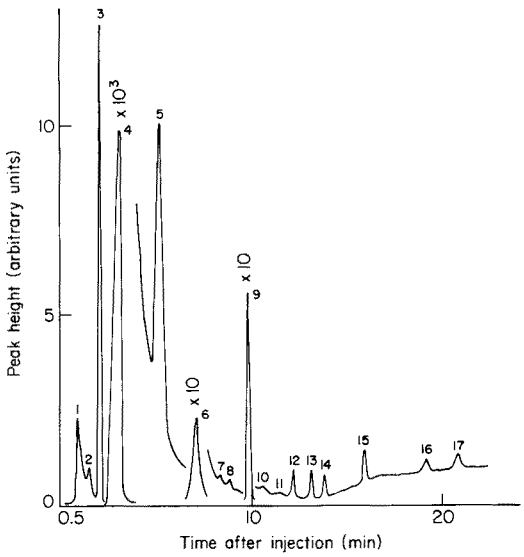


Fig. 1. An example of the g.c. pattern of a Scotch sample.

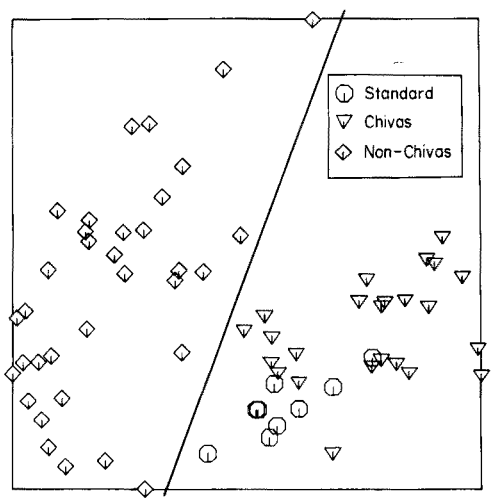


Fig. 2. Karhunen—Loève projection based on all 17 features (peak areas). The line drawn through the projection separates all members of C and ST from those of $\bar{C}$, and Fisher weights for this separation are listed below. Note that in a Karhunen—Loève projection the actual scale of the axis is irrelevant, because each axis represents a transformation from many features.

K—L Axis	Eigenvalue	% Variance	Fisher weight
1. Abscissa	8.2	48.0	1.5×10^{-1}
2. Ordinate	2.3	14.0	7.3×10^{-3}
3.	1.6	9.4	7.9×10^{-4}

TABLE 2

Data summary[a]

Feature	ST		C		$\bar{C}$	
	$\bar{X}$	$\sigma/\bar{X}$	$\bar{X}$	$\sigma/\bar{X}$	$\bar{X}$	$\sigma/\bar{X}$
1	1.1×10^{-4}	1.1×10^{-1}	1.3×10^{-4}	1.4×10^{-1}	7.0×10^{-5}	1.5×10^{-1}
2	7.1×10^{-6}	3.0×10^{-1}	9.4×10^{-6}	3.0×10^{-1}	7.1×10^{-6}	4.8×10^{-1}
3	3.8×10^{-4}	1.2×10^{-1}	4.4×10^{-4}	7.6×10^{-2}	3.6×10^{-4}	1.1×10^{-1}
4	9.9×10^{-1}	2.6×10^{-4}	9.9×10^{-1}	2.8×10^{-4}	9.9×10^{-1}	4.4×10^{-4}
5	6.2×10^{-4}	1.5×10^{-1}	6.2×10^{-4}	1.4×10^{-1}	7.6×10^{-4}	2.6×10^{-1}
6	1.7×10^{-3}	4.2×10^{-2}	1.8×10^{-3}	6.0×10^{-2}	1.5×10^{-3}	1.6×10^{-1}
7	9.1×10^{-6}	3.4×10^{-1}	7.9×10^{-6}	4.3×10^{-1}	1.0×10^{-5}	5.3×10^{-1}
8	6.5×10^{-6}	9.5×10^{-2}	7.4×10^{-6}	1.7×10^{-1}	4.4×10^{-6}	2.6×10^{-1}
9	1.7×10^{-3}	4.5×10^{-2}	1.8×10^{-3}	7.8×10^{-2}	1.2×10^{-3}	1.1×10^{-1}
10	4.5×10^{-6}	9.1×10^{-2}	4.0×10^{-6}	1.2×10^{-1}	2.3×10^{-6}	2.5×10^{-1}
11	1.1×10^{-6}	5.7×10^{-1}	2.4×10^{-6}	6.6×10^{-1}	1.5×10^{-6}	7.4×10^{-1}
12	1.9×10^{-5}	5.8×10^{-2}	1.8×10^{-5}	1.7×10^{-1}	1.3×10^{-5}	4.7×10^{-1}
13	1.5×10^{-5}	9.6×10^{-2}	2.1×10^{-5}	1.1×10^{-1}	1.1×10^{-5}	2.0×10^{-1}
14	1.6×10^{-5}	2.7×10^{-1}	1.4×10^{-5}	9.0×10^{-2}	1.5×10^{-5}	2.7×10^{-1}
15	2.8×10^{-5}	9.5×10^{-2}	3.9×10^{-5}	9.5×10^{-2}	2.7×10^{-5}	2.6×10^{-1}
16	2.2×10^{-5}	1.5×10^{-1}	2.5×10^{-5}	1.6×10^{-1}	2.2×10^{-5}	3.6×10^{-1}
17	3.1×10^{-5}	1.5×10^{-1}	3.1×10^{-5}	1.2×10^{-1}	2.3×10^{-5}	1.5×10^{-1}

[a] $\bar{X}$ is the mean of the feature taken over the indicated category.
σ is the standard deviation of the feature taken over the indicated category.

The Fisher weights in Table 3 quantify the utility of each feature to separate category members on a pair-wise category basis. Features with a high Fisher weight for separating C and $\bar{C}$, and ST and $\bar{C}$, but with a low weight for separating C and ST, are needed. To preserve as much information as possible, features with Fisher weights of as large a magnitude as possible for separating C and ST from $\bar{C}$ are desired. Feature 13 has a high weight for separating C and $\bar{C}$, but also a high weight for separating C from ST, rendering it unsuitable for the purpose. Features 1, 9, and 10 have the next-highest weights for separating C from $\bar{C}$, and the highest for separating ST from $\bar{C}$. Feature 9 has a very low weight for separating C from ST, indicating it to be an ideal discriminator according to the present criteria. Features 1 and 10 also have low weights for separating C from ST compared to their other weights indicating that they, too, are good choices. Though feature 17 has weights of generally low magnitude, the extremely low weight for separating C from ST indicates that most of the information it contains is in the other two separations, as desired.

Feature 3, having a high weight for separating C from ST, leaves 8 and 4 as the next possible choices. Because of the relatively large drop in weights between Feature 15 and Feature 6, and the small magnitude of the weights below 15, features with weights below that of 15 will be ignored as not contributing significant information; 15 itself is obviously unsuitable because of its high weight for separating C from ST. The six best features, in order of decreasing utility, are therefore: 1, 9, 10, 17, 8 and 4.

TABLE 3

Fisher weights[a]

C vs. $\bar{C}$		ST vs. $\bar{C}$		ST vs. C	
Feature	Weight	Feature	Weight	Feature	Weight
13	2.9×10^{-1}	10	3.9×10^{-1}	15	3.2×10^{-1}
1	2.9×10^{-1}	1	3.8×10^{-1}	13	2.0×10^{-1}
9	2.4×10^{-1}	9	3.2×10^{-1}	3	8.7×10^{-2}
10	1.8×10^{-1}	17	1.0×10^{-1}	10	3.8×10^{-2}
8	1.0×10^{-1}	8	8.4×10^{-2}	14	2.6×10^{-2}
3	7.9×10^{-2}	13	7.9×10^{-2}	11	2.5×10^{-2}
17	7.8×10^{-2}	4	4.7×10^{-2}	2	2.4×10^{-2}
4	7.0×10^{-2}	12	3.4×10^{-2}	16	2.2×10^{-2}
15	6.4×10^{-2}	5	1.5×10^{-2}	1	2.2×10^{-2}
6	2.2×10^{-2}	6	1.4×10^{-2}	8	1.8×10^{-2}
12	1.9×10^{-2}	14	8.8×10^{-3}	4	1.5×10^{-2}
5	1.4×10^{-2}	3	4.8×10^{-3}	6	1.0×10^{-2}
2	9.8×10^{-3}	11	2.7×10^{-3}	12	7.4×10^{-3}
11	8.7×10^{-3}	7	1.2×10^{-3}	7	4.0×10^{-3}
7	4.3×10^{-3}	15	7.0×10^{-5}	9	3.5×10^{-3}
16	4.0×10^{-3}	2	8.3×10^{-6}	17	1.3×10^{-4}
14	5.1×10^{-4}	16	6.9×10^{-7}	5	3.3×10^{-7}

[a]Features are listed in order of decreasing Fisher weights.

In Fig. 3, one sees that plotting 1 vs. 9 gives perfect separation of C and ST from $\bar{C}$, and 1 vs. 10 and 1 vs. 17 do nearly as well. Obviously, the problem is essentially solved at this point, because the differentiation between Chivas and non-Chivas samples can be done easily by inspection of Fig. 3. The diagonal nature of the separation indicates that both variables contribute significantly to the separation. It is clear from these plots that there is considerable correlation between these features (see the correlation matrix of the six best features in Table 4). Hence, it can be argued that there is only one real source of independent information responsible for the separation.

If the six best features are taken and the Karhunen—Loève transformation is applied only to these (Fig. 4), the separation between all Chivas (C and ST) and non-Chivas ($\bar{C}$) samples is enhanced compared with Fig. 2; in both cases, the two axes with the most information, i.e., variance, are plotted. The fact that the separation between all Chivas and non-Chivas samples occurs along a single axis reflects the earlier statement concerning the high correlation among "good" features, i.e., all highly correlated information can be thought of as essentially collapsed into a single axis.

To see how well the data could be classified more quantitatively, the classification techniques described above were applied. KNN, LSDA, and SIMCA were applied to the data to compare how these three methods would perform in assigning the various samples to their proper categories. C and ST were

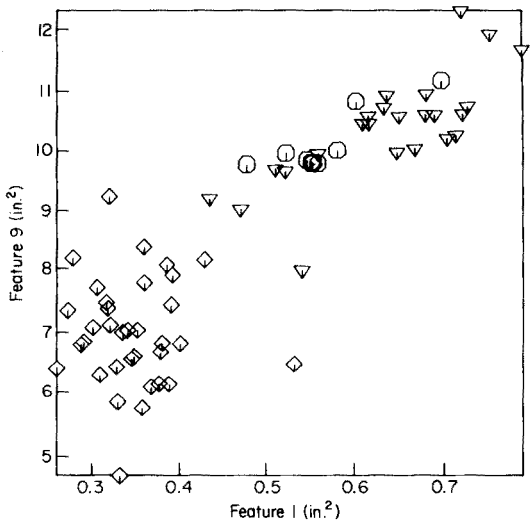


Fig. 3. Plot of data with Feature 1 as abscissa and Feature 9 as ordinate.

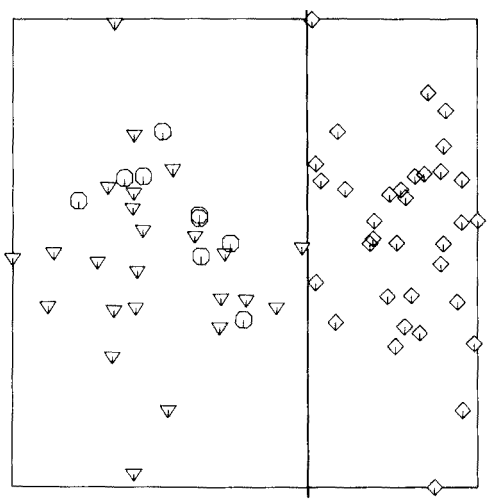


Fig. 4. Karhuner—Loève projection based on the 6 best features (1, 9, 10, 17, 8, 4). The line drawn through the projection separates members of C and ST from those of $\bar{C}$.

K—L Axis	Eigenvalue	% Variance	Fisher weight
1. Abscissa	4.8	80.	2.6×10^{-1}
2. Ordinate	0.42	7.1	4.8×10^{-4}
3.	0.34	5.7	2.0×10^{-3}

combined to make a single category for obvious reasons. The results for the three methods based on all 17 features are shown in Table 5; because the 17

TABLE 4

Correlation matrix

Feature	1	4	8	9	10
4	−0.76				
8	0.78	−0.67			
9	0.88	−0.79	0.78		
10	0.82	−0.67	0.67	0.86	
17	0.74	−0.74	0.59	0.82	0.70

TABLE 5

Classification results based on 17 Features, on the 6 best features, and on the 2 best features

Method	Number of misclassified points					
	17 features		6 features		2 features[a]	
	C and ST	$\bar{C}$	C and ST	$\bar{C}$	C and ST	$\bar{C}$
SIMCA (4 components)	0	2	0	4	1	0
LSDA	0	0	1	0	1	0
KNN						
1-NN	0	1	1	0	0	1
3-NN	0	1	1	0	1	0
5-NN	0	1	1	0	1	0
7-NN	0	1	1	0	1	0
9-NN	0	2	1	0	1	0

[a]1 Component for the SIMCA method.

features were normalized by their sum, the Karhunen—Loève transformation was used to rotate the data to the 16 independent features remaining before the classification routines were applied. The results obtained when the number of features was reduced to the best six mentioned above are also shown in Table 5. The improvement in results for the KNN method indicates amelioration of the information separating the two categories by distance. The fact that the performance of SIMCA is not greatly reduced indicates that most of the information separating the two categories has been retained, i.e. the loss of feature 13 was not serious. Reducing the number of features to two (see Table 5), gave essentially the same results, indicating that Features 8, 4, 10 and 17 are also not really necessary. (In fact, it is clear that only two features are needed to separate these two categories for these data, but the more "good" features retained, the less susceptible the analysis is to noise.) Even though these data indicate high correlation among these "good" features, one should bear in mind that in any extensive practical application of this method, any one of these features may prove critical in properly identifying

an unknown, e.g., the correlation observed with this data set may not extend to non-commercial "home-brew" whiskies or whiskeys.

The most important feature separating the C and ST samples from the $\bar{C}$ samples is Feature 1. As this is the first peak on the g.c. output, it would contain the more volatile (most likely polar since Carbowax is a non-polar column packing) components of the whisky. These components are clearly associated with the aroma of the whisky, such as low-boiling carbonyl compounds with probably lower molecular weight than ethyl acetate, which was identified as Feature 3. Acetaldehyde is probably the major component of Feature 1, for acetaldehyde is the major low-molecular-weight aldehyde present in whisky [11]. An almost equally good feature is Feature 9, which was identified as isoamyl alcohol. With these two features, as stated earlier, nearly-perfect classification is obtained.

It should also be noted that Feature 4 is the alcohol—water peak; there may be some correlation between peak area and the proof of the whisky which will render this feature of dubious forensic utility. In any case, the exact chemical composition of the six important features selected above is irrelevant to the problem at hand, though it is an interesting problem in its own right.

Conclusion

It is clear that the blind assay method employed here allowed the separation of the Chivas Regal and non-Chivas Regal samples. By building up a g.c. data library (possibly with columns and running conditions different from those used here as standards), a forensic laboratory could use this technique to establish a simple test to classify unknown samples. The advantages of the method are that no preparation of the sample is needed, and that it could easily be used by a mobile laboratory equipped with a portable g.c. The investigator need not know exactly what the distinctive components are chemically, only their relative amounts. Thus, considerable effort is eliminated.

It is obvious from the Figures presented here that the separation between the classes is not so distinct as to preclude the possibility of overlap. But with an "ideal case" situation, and with the parameters rigidly standardized, overlap seems improbable, giving a high degree of confidence in the identification of any unknown. However, the technique indicates that the magnitude of the variation within a single brand of whisky may be as severe as those between various brands of similar quality, as can be seen from the scatter of C compared to the scatter of ST or $\bar{C}$. Thus an attempt to distinguish between two whiskies of similar quality would not be successful for this coarse method of collecting data.

It is certainly feasible to separate samples of different batches of a single expensive whisky from samples of different, less-expensive whiskies by using pattern recognition techniques to define the features of interest obtained from the blind assay. The combination of these two methods reduces the time and cost of analysis considerably. Undoubtedly, this combination should be more widely utilized in future studies.

The authors thank Mr. Russell W. McLauchlan of Joseph E. Seagram and Sons, Inc., for providing the samples of Chivas Regal, and the National Science Foundation for partial financial support for Mr. Saxberg under the N.S.F. Undergraduate Research Participation program, Grant SMI 76-03095.

REFERENCES

1 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
2 C. H. Ho, W. H. Griest and M. R. Guerin, Anal. Chem., 48 (1976) 2223.
3 E. Cremer and R. Muller, Mikrochim. Acta, 36/37 (1951) 553.
4 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, ARTHUR, available from B. R. Kowalski, Department of Chemistry, BG-10, University of Washington, Seattle, WA 98195.
5 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.
6 R. A. Fisher, Ann. Eugen., 7 (1936) 179.
7 W. S. Meisel, Computer Oriented Approach to Pattern Recognition, Academic Press, New York, (1972).
8 T. M. Cover and P. E. Hart, IEEE Trans. Inform. Theory, IT-13 (1967) 21.
9 C. E. Liedtke and F. Torres, Int. J. Biomed. Comput., 6 (1975) 49.
10 S. Wold, Pattern Recog., 8 (1976) 127.
11 H. Suomalainen and L. Nykanen; Process Biochem., July (1970) 13.

Analytica Chimica Acta, 103 (1978) 213—221
Computer Techniques and Optimization

PRINCIPAL COMPONENT ANALYSIS OF THE INFRARED SPECTRA OF MIXTURES

GREGORY T. RASMUSSEN and THOMAS L. ISENHOUR*

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514 (U.S.A.)

STEVEN R. LOWRY

T. R. Evans Research Center, Diamond Shamrock Corporation, Painesville, Ohio 44077 (U.S.A.)

GARRY L. RITTER

Schering-Plough Corporation, 60 Orange Street, Bloomfield, New Jersey 07003 (U.S.A.)

(Received 25th April 1978)

SUMMARY

Principal component analysis of the infrared spectra of a series of related mixtures is used to determine the number of compounds present. The use of empirical error estimates makes it possible to determine correctly the number of components even when the spectra of the individual compounds are very similar.

With recent advances in automated instrumentation that can produce infrared spectral data relatively rapidly, the technique of combined gas chromatography—infrared spectroscopy (g.c.—i.r.) has moved from the realm of possibility towards that of practicality. In fact, the popularity of g.c.—i.r. is indicated by the commercial availability of such instrumentation. One goal of research in this area focuses on the development of an on-line spectral retrieval system capable of analyzing the data generated during the g.c.—i.r. experiment. One particular problem for such a search system arises when a mixture of compounds is incompletely resolved by the gas chromatograph. While a pure compound can be readily identified from its spectrum, compounds in a mixture are generally difficult to identify from a spectrum of the mixture. Because spectra of unresolved chromatographic peaks are spectra of mixtures, the compounds present can be difficult to identify by the customary search strategies. The first step in the solution of this problem is the determination of the number of compounds contained in the unresolved chromatographic peak. In previous work, principal component analysis of mass spectral data was employed to determine the number of compounds present in a series of mixtures [1, 2]. Presumably, the techniques used in that research are applicable to data from g.c.—i.r. experiments. Implicit in the application of principal component analysis to infrared spectral data is the assumption that

the observed absorbance of infrared radiation by a mixture approximates a linear combination of the observed absorption of the individual compounds present in the mixture. Although this assumption may not be completely valid, results indicate that it is a reasonable working hypothesis. Therefore given the infrared spectrum of each pure component, the spectrum of any mixture may be calculated.

Alternatively, in this study, the infrared spectra of a series of mixtures are given and the problem is to determine how many factors or components are necessary to reconstruct the infrared spectra of the mixtures. An infrared spectral data matrix is constructed which contains one row for each frequency at which data were collected and one column for each mixture spectrum. The rank of this matrix is the number of independent factors which must be used to reproduce the data matrix. In the ideal case, this is the number of components in the mixture if at least as many spectra as components are used. In a real system, which contains random error, the problem becomes slightly more difficult. In this case the number of independent factors which reproduce the data within allowed error must be found. Two conditions may lead to an incorrect estimate of the number of components. The number of components will be under-estimated if some component is not represented at any of the frequencies considered, or if the spectrum of one component is a linear combination of the others. Compounds of interest will customarily not be transparent over the full range of data acquisition, and the large number of points in a spectrum will virtually guarantee that the requirement for linear independence of the component spectra is met. Thus, the occurrence of either condition is unlikely, given the typical circumstances of the g.c.—i.r. experiment.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is one of the multivariate statistical techniques which are known collectively as factor analysis [3—6]. The applications of factor analysis to a number of diverse problems in chemistry have been recently reviewed [7]. As mentioned previously, in order to begin the analysis procedure, the infrared spectra of a series of unknown mixtures of the same components are first represented as a spectral system matrix containing n rows of frequencies and m columns of spectra. In this work n has a maximum value of 1500, but fewer frequencies may be used so long as care is taken to avoid the two conditions discussed above. The number of spectra used should equal or exceed the suspected number of components in the mixtures. The mean value of the entries in each column is subtracted from each entry in that column to produce a new matrix, A, which also contains n rows and m columns. Thus

$$\bar{a}_j = \frac{1}{n} \sum_{i=1}^{n} a'_{ij}; \; a_{ij} = a'_{ij} - \bar{a}_j; \; A = [a_{ij}]$$

where a'_{ij} is the absorbance at frequency i for spectrum j and a_{ij} is the corresponding normalized absorbance. The product moment coefficient matrix C is formed by premultiplying A by its transpose: $C = A^T A$

The resultant matrix, which is a symmetric matrix of order m, is the covariance about the mean. The problem now is to find the number of orthogonal factors necessary to account for this covariance. The number of such factors is easily determined by diagonalizing the covariance matrix and obtaining the eigenvalues and eigenvectors. These vectors, **X**, are chosen to satisfy the relation $C\mathbf{X} = \lambda\mathbf{X}$, where λ is a positive number. Then

$$\mathbf{X}^T C\mathbf{X} = \mathbf{X}^T \lambda\mathbf{X} = \lambda\mathbf{X}^T\mathbf{X} = \lambda$$

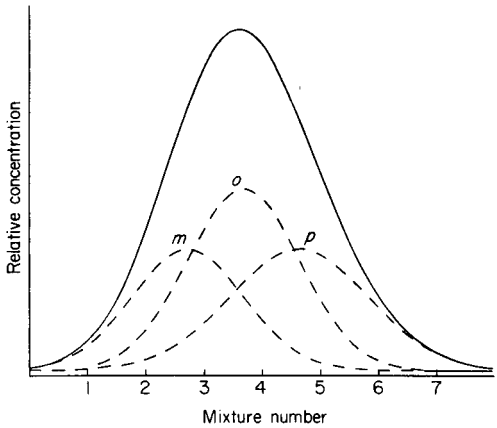
The vectors thus obtained are then ordered such that $\lambda_1 \geq \lambda_2 \ldots \geq 0$. Finally, it is necessary to determine how many of the eigenvalues (λ_n) are required to describe the system sufficiently and, hence, how many components the system contains.

EXPERIMENTAL

To investigate the usefulness of this method, infrared spectra were collected for (1) a series of mixtures of the xylene isomers and (2) a series of mixtures of nonane, decane and dodecane. All spectra were run with the sample mixtures contained in a liquid cell of 0.025-mm path length with KBr windows. The spectra were obtained on a Digilab FTS-14 Fourier transform spectrometer equipped with a dry nitrogen purge system. The digitized spectra as produced by the FTS-14 were used to compute absorbance spectra. An infrared spectrum for the region between 3500 cm^{-1} and 500 cm^{-1} with a resolution of approximately 4 cm^{-1} can be represented by 1500 data points. For each spectrum 20 scans were averaged to reduce noise.

The mixtures of the xylene isomers were made to simulate sampling across a hypothetical unresolved chromatographic peak. This peak and its three components are illustrated in Fig. 1. The seven points indicated in the figure reflect the times at which the spectra were collected during the elution of this hypothetical peak, in terms of the simulated experiment. The actual compositions of the xylene mixtures are listed in Table 1. In addition to the seven mixture spectra, two sets of reference spectra were collected to study possible sources of error in the model system. One set consisted of five reference spectra run after a lengthy purge of the spectrometer to approximate "ideal" conditions. The variance in these spectra should be due predominantly to instrumental noise. A similar series of reference spectra was obtained with the exception that the sample chamber was opened briefly and then closed for a few minutes prior to data acquisition, in order to simulate the effects of the changing spectrometer atmosphere that accompany the changing of the sample cell.

The compositions of the five mixtures of the three alkanes are listed in Table 2. The spectra of alkane mixtures were used to test the effectiveness



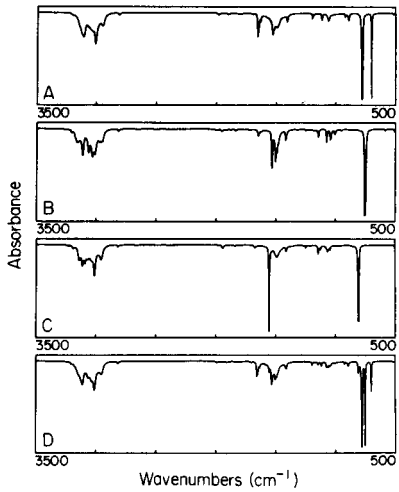


Fig. 1. Hypothetical chromatographic peak showing xylene components.

Fig. 2. Xylene spectra: (A) *m*-xylene; (B) *o*-xylene; (C) *p*-xylene; (D) mixture 2 of the xylenes.

TABLE 1

Compositions of the mixed xylene spectra

		Mixture number						
		1	2	3	4	5	6	7
Xylene isomer	*m*	0.831	0.645	0.380	0.153	0.038	0.004	0.001
	o	0.127	0.279	0.464	0.526	0.366	0.135	0.026
	p	0.042	0.076	0.156	0.321	0.596	0.860	0.973

TABLE 2

Compositions of the mixed alkane spectra

	Mixture number				
	1	2	3	4	5
C_9H_{20}	0.4	0.4	0.2	0.2	0.0
$C_{10}H_{22}$	0.4	0.2	0.4	0.6	0.2
$C_{12}H_{24}$	0.2	0.4	0.4	0.2	0.8

of the method when the spectra of the individual components in a mixture are very similar. One infrared spectrum was collected for each of the first four mixtures and five separate spectra of the fifth mixture were collected to provide an indication of the noise inherent in the experiment.

RESULTS AND DISCUSSION

The first step in performing the principal component analysis is the calculation of the covariance matrix. Next the matrix is diagonalized, and the eigenvalues are examined to determine the number of components in the mixtures. Table 3 shows the covariance matrix for the seven mixed xylene spectra in upper triangular form. It is important to note that all entries in the matrix have values of approximately the same magnitude, which indicates that the spectra are quite similar and that there will be one dominant direction of covariance in the data. The spectra of the three pure xylenes and of a mixture are shown in Fig. 2, and the results of the diagonalization are given in Table 4. The amount of variance along each eigenvector is reflected by the magnitude of the corresponding eigenvalue. In this case the first dimension accounts for more than 90% of the total variance in the data and the first three dimensions account for more than 99% of the variance. Figure 3 presents a graph of cumulative per cent variance as a function of the number of dimensions considered for the seven mixed xylene spectra. While the fact that 99% of the variance is covered by the first three dimensions suggests that these mixtures are composed of three components, one might argue that only two components are necessary to account for the "non-noise" variance in the data.

TABLE 3

Covariance matrix for mixed xylene spectra

1	2	3	4	5	6	7
564.13	528.67	507.13	491.18	464.81	455.70	458.56
	516.75	508.00	500.44	479.87	466.29	465.37
		511.25	513.15	495.26	477.84	474.47
			526.01	515.43	501.12	499.02
				525.96	529.44	536.33
					554.42	572.42
						597.67

TABLE 4

Results from diagonalization of covariance matrix for mixed xylene spectra

Eigenvalue	Eigenvectors						
3526.00	.372	.371	.374	.380	.380	.382	.387
213.17	−.527	−.370	−.248	−.725	.214	.430	.536
52.48	−.573	−.612	.326	.524	.323	−.135	−.406
3.39	−.462	.676	.260	−.503	−.063	.064	.037
0.67	.156	.041	−.304	−.303	.560	.411	−.554
0.30	−.086	.132	−.126	.330	−.611	.635	−.272
0.16	−.103	.496	−.719	.349	.118	−.273	.128

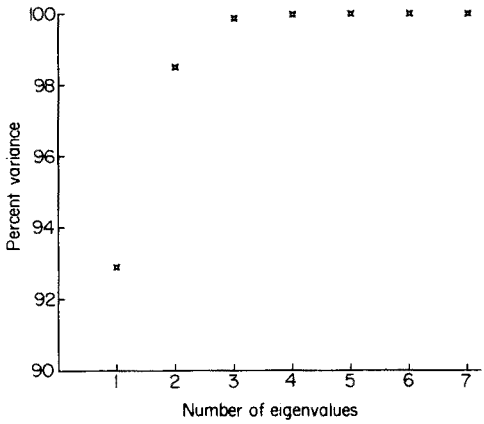


Fig. 3. Cumulative variance spanned by successive eigenvalues in the xylene analysis.

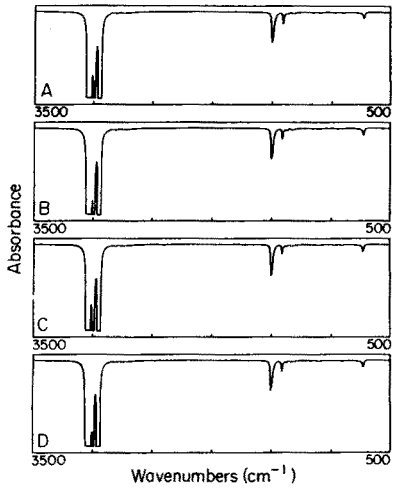


Fig. 4. Alkane spectra: (A) nonane, (B) decane, (C) dodecane, (D) mixture 1 of alkanes.

Another approach that should help to verify the number of components in the mixtures would be to identify the causes of the residual variance in the data. To study this question, principal component analysis was performed on the two sets of reference spectra. Because the sets consist of five reference spectra each, it is necessary for a proper comparison to analyze data for only the first five of the seven xylene mixtures. The results of the diagonalization of the covariance matrix for five mixtures are reported in Table 5 along with the results for the seven mixtures. In each case, the first three dimensions account for about the same relative amount of variance. Table 6 reports the results of the eigenanalysis of the covariance matrices for the two sets of reference spectra and the set of five xylene mixtures. Since the three sets of data contain the same number of spectra, the magnitudes of the eigenvalues can be compared directly. The small magnitudes of the eigenvalues for the

TABLE 5

Eigenvalues from analysis of five and seven mixed xylene spectra

5 Mixtures		7 Mixtures	
λ_i	$\%V$	λ_i	$\%V$
2528.00	91.63	3526.00	92.88
190.36	6.90	213.17	5.61
37.07	1.34	52.48	1.38
3.08	0.11	3.39	0.09
0.47	0.02	0.67	0.02
		0.30	
		0.16	

TABLE 6

Eigenvalues from analysis of (A) five mixed spectra, (B) "ideal" reference spectra, (C) reference spectra with limited purge time

Eigenvalues and per cent variance					
A (mixtures)		B (noise)		C (purge)	
λ_i	%V	λ_i	%V	λ_i	%V
2528.00	91.63	0.03481	67.71	3.057	96.42
190.36	6.90	0.00795	15.47	0.096	3.04
37.07	1.34	0.00346	6.72	0.008	0.26
3.08	0.11	0.00302	5.88	0.005	0.15
0.47	0.02	0.00217	4.21	0.004	0.13

"ideal" reference spectral data indicate that instrumental noise contributes an insignificant portion of the residual variance. In contrast, the analysis of the other reference spectra produces an eigenvalue of much greater magnitude. Thus the opening of the sample chamber in these experiments has added another direction of variance to the data. Significantly, the magnitude of the eigenvalue attributed to the opening of the sample chamber corresponds to the magnitude of the fourth eigenvalue from the analysis of the mixed xylene spectral data. Although variations in the purge of the spectrometer atmosphere may be detected in the data analysis, the magnitude of contributions from this source is small compared to the magnitude of contributions from real components of a mixture.

An alternative criterion for determining the number of components in a series of mixtures is illustrated by the analysis of the spectra of the mixtures of the three normal alkanes. The spectra of nonane, decane, and dodecane, and a mixture of the three are shown in Fig. 4. Infrared spectroscopists typically identify a "fingerprint" region where the most significant differences between spectra of similar compounds are observed. It seems reasonable that one way of meeting better the linear independence requirement for component spectra is to limit the data analysis to the fingerprint region of the spectra. Three definitions of the limits of the fingerprint region were taken from popular textbooks [8—10], and covariance matrices were calculated from the data corresponding to these segments of the alkane spectra. Table 7 shows the results of the diagonalization of the three pairs of covariance matrices. One pair exists for each definition of the fingerprint region, and each pair includes a covariance matrix computed for the five spectra of alkane mixtures and one computed for the five repeated spectra of the 4:1 mixture of dodecane and decane. Because the repeated spectra reflect the results obtained with a single component, any single pure alkane or mixture of alkanes could be used. Not surprisingly, given the similarity of the spectra, one component which corresponds to the first eigenvector accounts for over 99% of the variance in all cases. However, the repeated spectra should have only one real component,

TABLE 7

Eigenvalues and per cent variance from the analysis of the alkane spectra

Fingerprint region cm^{-1}	Different mixtures		Repetitions	
	λ_i	$\%V$	λ_i	$\%V$
	2.27817	99.72	2.59705	99.92
	0.00385	0.17	0.00083	0.03
1300—650	0.00145	0.06	0.00065	0.02
	0.00058	0.02	0.00036	0.01
	0.00040	0.02	0.00032	0.01
	λ_i	$\%V$	λ_i	$\%V$
	4.15564	99.88	3.77655	99.98
	0.00404	0.10	0.00040	0.01
1430—910	0.00075	<0.02	0.00015	<0.01
	0.00021	<0.01	0.00010	<0.01
	0.00007	<0.01	0.00009	<0.01
	$\lambda_i \times 10^2$	$\%V$	$\lambda_i \times 10^2$	$\%V$
	6.23725	99.06	5.83058	99.21
	0.17064	2.63	0.02061	0.35
1300—910	0.06824	1.05	0.01225	0.21
	0.01174	0.18	0.00719	0.12
	0.00505	0.08	0.00617	0.10

and therefore the second eigenvalue from the analysis of the repeated spectra reflects the level of error in the data. This empirical estimate of the error serves as a useful threshold to determine the number of components in the five spectra of different mixtures. Eigenvalues from the analysis of these spectra which have magnitudes greater than that of the second eigenvalue from the repeated spectra are attributed to real components in the mixture. Applying this criterion gives the result that the spectra of the five different mixtures contain three real components in each case.

Conclusion

These two examples demonstrate the feasibility of applying principal component analysis to infrared spectral data. The analysis of the spectra of the xylene mixtures shows that, for typical spectra of mixtures that simulate severely overlapped chromatographic peaks, the determination of the number of compounds present can be quite straightforward. By analyzing the two sets of reference spectra, it is possible to estimate the magnitude of the variance in infrared data that arises from sources such as instrumental noise and varying the atmosphere in the spectrometer. Finally, a technique for obtaining an empirical error estimate is used in the analysis of the alkane spectra. This

approach may be most relevant to the analysis of spectra generated by monitoring the effluent of a gas chromatographic column, where close control can be maintained over the conditions of data acquisition. The selection of data from the fingerprint region for principal component analysis may be useful as a general technique in studying the infrared spectra of mixtures. The fact that principal component analysis can be used to determine correctly the number of compounds in the alkane mixtures, where the spectra of the individual compounds are nearly identical, illustrates the effectiveness of the method.

The authors thank J. C. Marshall of St. Olaf College for helpful discussions during the course of this work.

REFERENCES

1 G. L. Ritter, S. R. Lowry, T. L. Isenhour, and C. L. Wilkins, Anal. Chem., 48 (1976) 591.
2 J. E. Davis, A. Shepart, N. Stanford, and L. B. Rogers, Anal. Chem., 46 (1974) 821.
3 H. H. Harman, Modern Factor Analysis, University of Chicago Press, Chicago, 1967.
4 L. L. Thurstone, Multiple Factor Analysis, University of Chicago Press, Chicago, 1967.
5 P. Horst, Factor Analysis of Data Matrices, Holt, Reinhart, and Winston, New York, 1965.
6 R. J. Rummel, Applied Factor Analysis, Northwestern University Press, Evanston, 1970.
7 P. M. Weiner, Chem. Technol, 7 (1977) 321.
8 H. H. Willard, L. L. Merritt, Jr., and S. A. Dean, Instrumental Methods of Analysis, D. van Nostrand, New York, 1974.
9 J. R. Dyer, Applications of Absorption Spectroscopy of Organic Compounds, Prentice-Hall, Englewood Cliffs, N.J., 1965.
10 R. M. Silverstein and G. C. Bassler, Spectrometric Identification of Organic Compounds, J. Wiley, New York, 1967.

Analytica Chimica Acta, 103 (1978) 223—235
Computer Techniques and Optimization

AUTOMATED EVALUATION OF PHOTOGRAPHICALLY RECORDED SPARK-SOURCE MASS SPECTRA

B. VANDERBORGHT* and R. VAN GRIEKEN
Department of Chemistry, University of Antwerp (U.I.A.), B-2610 Wilrijk (Belgium)
(Received 28th March 1978)

SUMMARY

A computer routine was developed for qualitative and quantitative analysis of photographically recorded spark-source mass spectra. Particular attention is given to the case of a graphite matrix. The program starts from the line intensities (expressed as Seidel values) and isotope masses calculated from the densitometer readings by a commercially available routine. From the intensities in the different exposures (typically 15 stages), it computes the parameters for the linear parts of the density curves for each ion. Taking into account mutual interferences of multivalent ions, isotope or C-clusters, oxide, carbide and dicarbide ions, the program automatically identifies and then quantifies the elements present. The precision of the results is around 5%. Reading and complete processing of one photoplate is achieved within 2—3 h.

Spark-source mass spectrometry (s.s.m.s.) is one of the few instrumental techniques which allows a panoramic multi-element analysis with comparable sensitivity for all elements from light to heavy over a wide concentration range. This makes s.s.m.s. very useful for survey analyses of, for example, environmental samples where unexpected toxic elements can be of primary importance. Preparation of samples from environmental origin for s.s.m.s. often results in a carbon matrix. For such analyses, photographic detection has definite advantages over electrical detection. First, the photoplate is an integrating detector over the whole mass spectrum while electrical detection usually measures only a few preselected ion masses. Secondly, the mass spectrum of an environmental sample, especially in a carbon matrix, can be so complex that the less advantageous resolution of electrical detection is inadequate. Manual analysis of photographically recorded mass spectra is, however, very time-consuming, and the precision and accuracy are not always satisfactory. This paper describes a simple computer system for the qualitative and quantitative evaluation of photographically recorded mass spectra with a precision of 5% over the full mass range for repeated analyses of the same plate. The procedure differs considerably from published sophisticated iterative programs [1—4] but yields a comparable precision.

EXPERIMENTAL

Apparatus

Ilford Q2 ion-sensitive plates (38 × 5 cm) are exposed to ions in a double-focusing, radio-frequency, spark-source mass spectrometer (JEOL JMS-01 BM-2). The automated microdensitometer is a single-beam instrument (JEOL JMD-2C) controlled by a JEOL JEC-6 electronic computer (8 K, 16 bits, magnetic drum memory of 8 K) provided with a teletypewriter and fast tape puncher. While a photoplate exposure is measured, the position of the x-axis is continuously moved by a screw at a rate of 1 mm s^{-1}, and the plate transmittance is measured at a frequency of 1000 Hz, resulting in a measurement every micrometer.

The output voltage of the transmittance measurement is digitalized by an analog—digital converter for real-time calculation of line profiles, intensities and positions. After measurement of an entire exposure, the microdensitometer moves back to the low mass end of the plate, shifts one exposure in the y-direction and restarts measurement. Primary data (line intensity in different exposures and ion mass) are punched on paper tape by a fast puncher. These tapes are processed off-line by a larger computer (Digital Equipment Corporation, PDP 11/45, 64 K). The system is outlined in Fig. 1. Instrumental parameters for the carbon matrix electrodes are given in Table 1.

Reagents

The graphite and metal oxide powders were all Specpure (Johnson-Matthey Company). Standard electrodes were prepared by adding standard solutions of metal oxides or salts in water to graphite, mixing with a minimum quantity of acetone as wetting agent, and drying in a film evaporator. In this apparatus, the glass flask containing the sample is continuously rotated so that the graphite

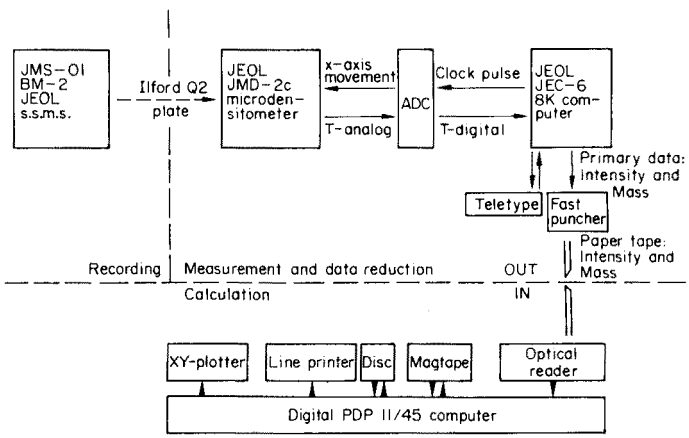


Fig. 1. Schematic representation of the hardware set-up.

TABLE 1

Instrumental parameters for the collection and evaluation of the spark-source mass spectra

Spark-source mass spectrometer			
Source vacuum	$<10^{-7}$ torr	Main slit width	15—20 μm
Analyser vacuum	$<10^{-6}$ torr	α-slit width	0.6 mm
Spark pulse duration	20 μs	β-slit width	0.8 mm
Spark repetition rate	1 kHz	Slit height	2 mm
Spark voltage	~40 kV	Number of exposures on each plate	15
Radiofrequency	1 MHz	Maximum exposure	~120 nC
Accelerating voltage	30 kV	Minimum exposure	~0.1 nC
		Mass range covered	10—250
Microdensitometer			
Slit width	15 μm	Slit height	1 mm

is well mixed with the solution which is evaporated under vacuum below 45°C. The homogeneously spiked graphite powder was compressed to cylindrical electrodes (12-mm length, 2-mm diameter) at 14,000 psi.

DATA ACQUISITION PROCEDURE

The complete JEOL "JMA-1340 Inorganic Photoplate Processing System" has been described in detail elsewhere [5]. Only basic principles will be discussed here. The major functions leading to the primary data output (line intensity and ion mass) are the scanning of the photoplate with the microdensitometer, the identification and mass calculation of each line and the combination of mass spectra in each exposure, to form a data matrix containing the intensities in all the exposures of all mass lines.

Line identification and intensity calculation

The real-time line-finding program is similar to that described by several authors (see, e.g. [1]). Several criteria must be fulfilled to identify a certain transmittance (T) profile as a mass line. As the scanning over the x-axis (function of m/e) proceeds, the absorbance must exceed a certain threshold over the base level, which is defined as the mean value over 32 μm where no mass line has been observed, and the gradient of the transmittance profile must exceed a certain value.

After passing the peak maximum, the absorbance drops below the threshold or shows a minimum between two doublet peaks. If the line width is greater than 16 μm and smaller than 300 μm and the criteria of threshold and gradient are fulfilled, the transmittance of each point is converted to the Seidel value, $S = \log\left(\frac{1}{T} - 1\right)$, and integrated between the points where the threshold level is reached. Net Seidel intensities are obtained by subtracting a background, interpolated from the base level on both sides of the line. They are stored in the magnetic drum memory of the JEC-6 computer.

Mass calculation

Before exposure, a black strip is mounted at the low mass end of the spectrum parallel to the mass lines (Fig. 2). Scanning of the spectrum starts on the black strip where the transmittance is zero. At the edge of the black strip the transmittance suddenly increases to almost 100%. The position where $T = 50\%$ is defined as the start level of the position measurement. The position of each point of the plate can be referred to this start level with a 1000-Hz clock pulse, and the position of each mass line is the distance, in μm, from the peak maximum to the start level. By referring to a few lines with known mass, the computer calculates a relation between plate distance and mass. With this semi-empirical relation the mass of each line can be computed. This mass calculation is done after measurement of each exposure, and line intensities of the same ion in subsequent exposures are combined to form the primary data output for every ion.

Data transfer

The primary data output, which is printed on teletype or on paper tape by a fast puncher contains the line number in the spectrum, the mass with three characters after the decimal point (in a.m.u.) and the net integrated Seidel value in each exposure, normalized between the arbitrary values 0 and 250.

Completely automatic quantitative and qualitative analysis of these primary data is impossible because of the limited size of the JEC-6 computer (8K core memory). The data are therefore transferred to a larger off-line computer where the paper tapes are first transduced to magnetic tape for faster assessment

In spark-source mass spectrometry of carbon electrodes, heavy secondary fogging is produced by impact of ^{12}C matrix ions on the insulating, gelatin-containing Ilford Q2 plates. This fogging can increase the background severely up to mass 70 or 80, hampering the determination of the concentration of elements lighter than bromine; it can be drastically reduced by cutting off a small piece of the plate so that the ^{12}C ions hit the conducting photoplate holder instead of the plate. Usually fogging is then very low above mass 35.

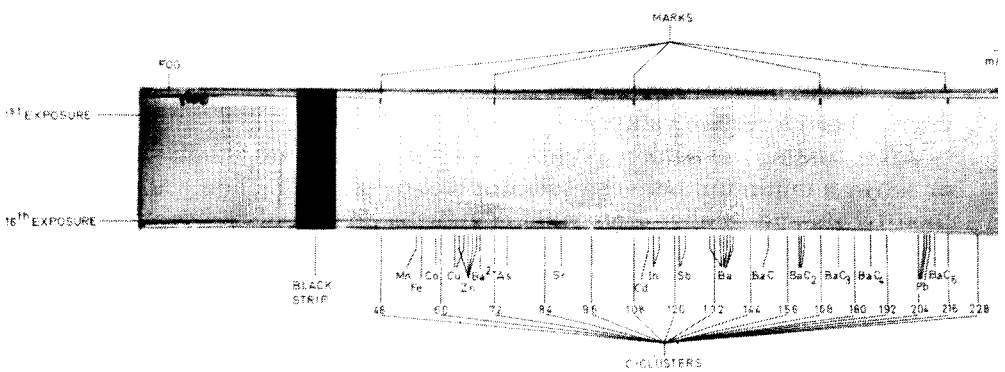


Fig. 2. Example of ion-sensitive s.s.m.s. plate from a graphite matrix sample.

As most elements which are of interest from an environmental viewpoint have a mass higher than chlorine, the end of the strip — and thus also the starting point for measurement — was placed at mass 37. With a scan speed of 1 mm s^{-1} and measurement of 15 exposures from mass 37 to 240, a photoplate can be processed by this procedure within 2 h.

CALCULATION OF CONCENTRATION

The integrated ion intensity (I) striking the photoplate is proportional to the total ion exposure (E) on the photoplate (measured on the beam monitor), the element concentration (C_i) in the electrodes, the isotopic abundance, the ionization efficiency and the transmittance through the mass spectrometer. The last three terms are contained in the proportionality factor k_i. This gives the equation $I = k_i C_i E$ or

$$\log I = \log E + \log k_i + \log C_i \tag{1}$$

If the linear relationships between log I and log E could be determined experimentally for an unknown element i and a reference element r, the concentration C_i could be calculated from these curves by reading off the relative exposures required to produce the same standard integrated intensity for both elements. Indeed, for $I_i = I_r$,

$$C_i = C_r G_{ir} E_r / E_i \tag{2}$$

where G_{ir} is the relative sensitivity factor including isotopic abundances.

Some authors (JEOL and program [6, 7]) use this formula for concentration calculations, while deriving E_r and E_i from the transmission areas of the densitometric measurements rather than from the ion-intensity peaks. With photographic detection, however, the ion intensity I is not directly accessible experimentally. The relation between the experimental transmittance T_x (transmittance at a certain position on the photoplate) and the ion intensity (I_x) is given by the Seidel equation which is a simplified form of the Hull relation:

$$I_x = \frac{1}{K_i'}\left[\frac{1}{T_x} - 1\right]^{\frac{1}{A'}} = \frac{A'}{K_i}\left[\frac{1}{T_x} - 1\right]^{\frac{1}{A'}} \tag{3}$$

where K_i is the photographic sensitivity constant for the ion, and A' is a constant dependent on the slope of the characteristic curve, which may be different for each ion.

In order to obtain the total ion intensity of a mass line i, eqn. (3) must be integrated over all the measurement values of the transmittance profile:

$$I_i = \sum I_x = \frac{A'}{K_i} \sum_x \left\{\left[\frac{1}{T_x} - 1\right]^{\frac{1}{A'}}\right\} \tag{4}$$

As the characteristic curve of the photoplate may be different for every ion, the parameters K_i and A' are not known at the time of the transmittance

measurement, and must be determined by iterations. This cannot be done in the real-time mode by a computer on-line with the densitometer. For integration the JEOL program assumes $A' = 1$.

After summation, the influence of the slope can be re-introduced in the form

$$I_i \cong \frac{A}{K_i}\left\{\sum\left[\frac{1}{T_x}-1\right]\right\}^{\frac{1}{A}} \cong \frac{A}{K_i} Z_i^{\frac{1}{A}} \tag{5}$$

and $\log I_i = \log (A/K_i) + (1/\mathrm{A}) \log Z_i$. Substitution in eqn. (1) yields

$$(1/A) \log Z_i = \log E_i + \log C_i + \log (k_i K_i/A) \tag{6}$$

This equation is experimentally accessible by dividing the $\log Z_i$ vs. $\log E_i$ curve, as obtained by the densitometric measurement with the JEOL program, by the slope value A. When the same "standard normalized logarithmic Seidel value" (i.e., the standard value of the logarithm of the Seidel function divided by the slope A) is used for the unknown ion i and the reference ion r, the concentration C_i can be calculated. Indeed, if $(\log Z_i)/A_i = (\log Z_r)/A_r$, then

$$\log C_i = \log (E_r/E_i) + \log C_r + \log (k_r K_r A_i/k_i K_i A_r) \tag{7}$$

Including k and K in the relative sensitivity factor R gives

$$C_i = C_r \frac{E_r}{E_i}\frac{A_i}{A_r}\frac{R_r}{R_i} \tag{8}$$

Usually the isotopic abundances θ are not included in the relative sensitivity coefficient. The isotopic masses M are taken into account in order to obtain concentration values in mass fractions rather than atom fractions. The relative sensitivity factor for the internal standard, R_r, is taken as equal to unity. Thus eqn. (8) becomes

$$C_i = C_r \frac{E_r}{E_i}\frac{\theta_r}{\theta_i}\frac{M_i}{M_r}\frac{A_i}{A_r}\frac{1}{R_i} \tag{9}$$

It is clear from eqn. (6) that the experimental density curve, $\log Z$ vs. $\log E$, will yield A as its slope. In practice, for each ion a straight line ($y = Ax + B$) is fitted by the least-squares method through the linear part of the experimental density curve. From eqn. (6), $B = A.\log (C_i \cdot k_i K_i/A)$. If the standard normalized logarithmic Seidel value is always taken as zero (for internal standard and ions), eqn. (6) becomes

$$(1/A) \log Z = 0 = \log E + B/A \tag{10}$$

and the exposure E yielding this zero normalized logarithmic Seidel value is calculated as

$$E = 10^{-B/A} \tag{11}$$

Substituting the A and E values for unknown ions and internal standards in eqn. (9) yields concentration values. This procedure differs from the iterative

procedures where ion-intensity calculation is done before integration of the line profile [1—3, 8] but it yields analogous results and also comparable precision [8] in a faster and simpler way.

THE COMPUTER PROGRAM

The actual program for complete spectrum identification and concentration calculation consists of a main program which links six subprograms (Fig. 3). The first subroutine finds the primary data from the spectrum on magnetic tape. The second calculates the parameters A and B (eqn. 10) of the calibration curve of each ion. Two subroutines execute qualitative and quantitative analysis, respectively. A fifth subroutine gives graphically the line intensity measured in each stage versus the exposure value together with the calculated least-squares fit for each mass line. A last subroutine plots the slope of this curve versus the ion mass. The isotopic table remains in overlay on disc while the lists of experimental masses and intensities are stored on tape, but are also transduced on disc for faster assessment when the spectrum has to be analysed.

Calibration curve calculation

The linear calibration curve log Z vs. log E (eqn. 6) is calculated for each mass line through least-squares fitting with two free parameters: slope and intercept. Negative aberrations of experimental data points from linearity can, however, occur on the high exposure side because of saturation or increased secondary fogging. Positive deviations at the low exposure side caused by underexposure of the emulsion are usually not observed. Erratic mistakes (dust, emulsion inhomogeneities, etc) can occur anywhere on the curve. One of the main objectives of this subprogram is to reject outlying points. Experience has shown that the following procedure yielded optimal results. The regression coefficient (r) is calculated. If $r \geqslant 0.975$ all data points are accepted;

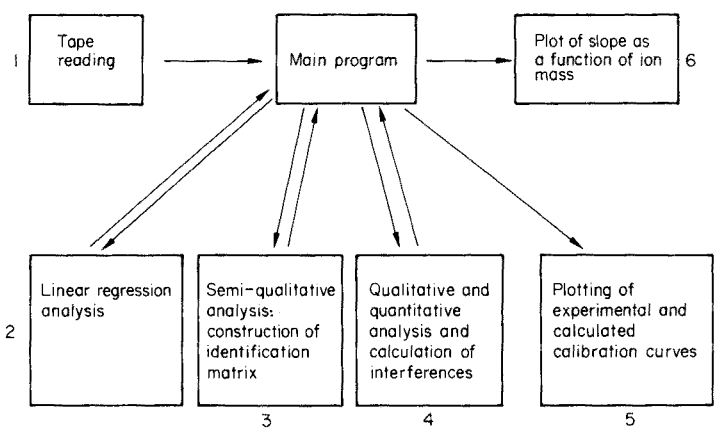


Fig. 3. Diagram of the subroutine program interaction.

if $r < 0.975$ the point that deviates most from linearity is rejected and the fitting is repeated until $r \geqslant 0.975$ or until no more than three points are retained; the value 0.975 was chosen on a purely pragmatic basis.

However, important differences in the slope of different mass lines could occur, and better precision in the concentration calculation was obtained if the variation of the slope was restricted within certain limits. The slope A of the density curve decreases significantly with ion mass [1] as shown in Fig. 4. It can be given by the relation

$$A = pM^q \tag{12}$$

where the factors p and q are strongly plate-dependent and $q < 0$. From the mass and calculated slope of a few ions, the factors p and q can be calculated for each plate. If for a particular mass line not enough data points are available for a reliable linear regression with two free parameters, or if the slope of the calibration curve of an ion deviates too much from the slope expected from eqn. (12) for a certain mass, the calibration curve is recalculated by using the least-squares fit with only one free parameter, B, and a fixed A. The values of B and A obtained in this subroutine form the basis for concentration calculations based on eqns. (9) and (11) in the subroutine for quantitative analysis.

Semi-quantitative analysis

In this subprogram an "identification matrix" is constructed which contains information on the possible presence of mono-, bi- and trivalent ions, mono- and di-carbides (MC^+ and MC_2^+) or oxides (MO^+) of an isotope or C-clusters (Table 2). At this stage of interpretation, no logical elimination of ions is done, e.g. on an isotopic pattern basis. Only mathematical comparison of masses is done.

In this subprogram the measured mass of a line is compared with the mass of all ions of the isotopic table. If both masses correspond within certain adaptable limits, e.g. 0.025 a.m.u., this ion is accepted as a possible identification. The elemental concentration is calculated from eqns. (9) and (11) and A and B values of the prior subroutine, assuming that the measured line is

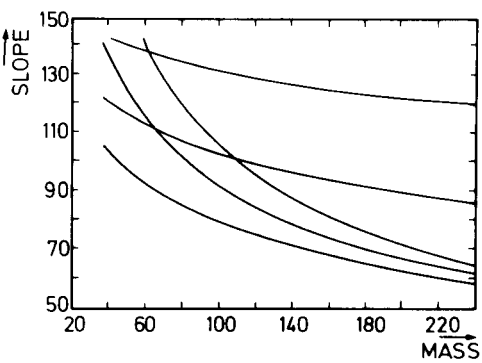


Fig. 4. Mass dependence of the slope of density curves for different photoplates.

TABLE 2

A simulated computer output for the semi-qualitative analysis subroutine

Line Number	Mass	Admitted Mass Error	Possible identification, and concentration (in ppm)							
...	...	...	...							
...	...	...	...							
...	...	...	...							
28	55.934	0.005	Fe	7						
		0.020	Cd2+	56	Sn2+	1396				
		0.025	CalC	−1[a]						
29	56.934	0.005	Fe	7						
		0.020	Cd2+	1	Sn2+	48				
		0.025	SclC	0						
30	57.469	0.005		0						
		0.020	In2+	1	Sn2+	266				
31	57.935	0.005	Fe	26	Ni	0				
		0.020	CalC	−1	TilC	1	Cd2+	2	Sn2+	1
...	...	...	...							
...	...	...	...							
...	...	...	...							

[a]—1 = concentration higher than admissible maximum.

interference-free. This concentration is inserted into the identification matrix in the column of the monovalent ions and the row of the proper isotope. If the two masses do not correspond, a possible identification as a polyvalent or molecular ion is checked. If such an ion is retained as a possibility, the elemental concentration is also calculated, assuming a sensitivity factor of one and no interference; this value is also stored in the matrix. The isotopic table is completely checked for each mass line of the spectrum. The identification matrix constructed in this way serves as a basis for the final qualitative analysis and is normally not printed for output.

Quantitative analysis

After the former subprogram has run for all mass lines of the spectrum, a first selection of ions is done. A requirement put forward by most authors is that the major isotope be present. However, this line may have been overlooked during the spectrum measurement by the microdensitometer, e.g. if it overlaps with an unresolved interference or is located on the background of an intense line. This algorithm can be too strict in certain cases. To retain an element as possibly present in the sample, at least one of the two major isotopes must be present in the spectrum; if neither is observed, all data for that element in the identification matrix are rejected and the element is considered as absent. The data in the identification matrix can thus be drastically reduced.

The final analysis is done sequentially for each element in the periodic table by using the data in the identification matrix. As shown in Table 3, when a few

TABLE 3

A simulated computer output for the qualitative and quantitative analysis subroutine

Element	Average concentration	% Standard deviation	-(% Abundance) Isotope mass -Element concentration in ppm -Interference free ions (NF = not interference free, XX = overexposed in 3rd exposure) -Intensity in exposure 5 calculated through the average concentration -Experimental intensity in exposure 5 calculated with least squares				
CHARGE = 1		CARBIDE or OXIDE: 0.000					
Fe			(5.8) 54	(91.7) 56	(2.2) 57	(0.3) 58	
	16.27	66.5	11.30	10.39	10.91	32.50	
	10.87	4.2				NF	
			18	171	−39	−144	
			20	169	−38	−83	
Co			(100) 59				
	8.24	0.0	8.24				
	8.24	0.0					
			156				
			156				
Ni			(67.9) 58	(26.2) 60	(1.2) 61	(3.7) 62	(1.1) 64
	0.16	0.0	0.16	0.00	0.00	0.00	0.00
	0.16	0.0	NF				
			−83				
			−83				
Cu			(69.1) 63	(30.9) 65			
	12.56	0.1	12.57	12.55			
			153	107			
			153	107			

lines are found whose mass corresponds to monovalent ions, the concentration of the element is calculated (from the A and B value of each isotope), as well as the mean concentration and relative standard deviation (r.s.d.). When the r.s.d. is less than 20%, all isotopes are accepted to be interference-free, and the next element of the periodic table is checked. When the r.s.d. is larger than 20% or the element is mono-isotopic, possible interferences (isobaric, polyvalent, molecular ions from other elements) are sought. The search for interferences is carried out sequentially for all isotopes, in the order of increasing isotopic abundance. First, a check is made as to which ions can interfere in view of their mass. If an interfering ion is divalent, the corresponding monovalent ion must be present. For a trivalent interference, both mono- and divalen ions must be present. For interfering molecular ions, the monovalent elemental ion must be present in the spectrum, and its intensity must exceed a preset level. If an isotope is found to suffer interference, a new average element concentration and r.s.d. are calculated excluding the interfered isotope. The

search for interferences is stopped after all isotopes have been checked or after the r.s.d. has dropped below 20%.

If no isotope is interference-free, the concentration of the element cannot be calculated accurately; only an upper limit can be determined. If an interference-free concentration value is available, the density curve for each ion is calculated followed by the corresponding expected line intensity for each isotope in an arbitrarily chosen exposure. This intensity can be compared by the operator with the printed intensity value calculated for the same exposure from the least-squares fit of the experimental density curve. In this manual interpretation of the computer analysis, the operator can examine the relative contributions of the different isotopes to one line, and decide whether errors are acceptable or not. The final interpretation of the analytical data can also be inspected by visually checking the quality of the fitted calibration curve and examining the computer selection and rejection of data points. After the whole periodic table has been scanned, the procedure is repeated for bi- and tri-valent ions, mono- and di-carbide ions and, if required, also for metal oxides.

After the whole photoplate has been analysed, the computer prints out which ions pertain to the lines, and the magnitude of the individual contribution of every ion to the B-value for each line. The experimentally determined intercept B, and its calculated value are also printed out. (The latter total B-value is not a linear addition of the contributions of each ion i, but it is given by $B_t = A \log \Sigma_{i=1}^{n} 10^{B_i/A}$, where B_i is the intercept of the density curve for every ion computed from the calculated average concentration.

A spark-source mass spectrum of a carbon matrix sample is characterized by the presence of lines of C-clusters. Mass lines of molecular ions are found over the whole ion-sensitive plate (see Fig. 2). In this procedure for spectrum analysis, these C-clusters are simply included as "elements" in the isotopic table data file on disc and are treated in the same way. The program can easily be adapted for other sample matrices by changing the cluster data of the isotopic table.

Organization and timing

Apart from the concentration of the internal standard and the mass of a few lines of the spectrum for mass calibration, an initial knowledge of the composition of the sample is not essential. Since most decisions can be taken by the computer, even completely unknown samples can be run.

The total elapsed time for calibrating, reading and calculating a spectrum can be broken down into 6 stages. First, the manual mass calibration of the ion-sensitive plate by the operator takes about 5 min. The microdensitometer scanning from mass 40 to 240 takes about 5.5 min per exposure. Usually 15 exposures are measured for a sample of environmental origin; this is also the maximum number on a plate and thus gives the maximum precision. Calculation of primary data after reading takes about 3 min while output on the fast paper-tape puncher usually requires 20 min. This results in a total of 2 h for the acquisition of the primary data of a plate reading. Unfortunately, the

conversion from these data to a compatible format on magnetic tape takes about 30 min. The calculation time of the final program is about 5 min for a spectrum of 150 mass lines. If plots of the calibration curves are wanted, the printing time becomes longer.

Of course, a single sample analysis by s.s.m.s. is seldom done. Usually a series of different samples is analysed in one run, and the total handling time of a sample in a series is much less than that of a single sample. With the proposed system, where quantitative information down to the 0.1-ppm level is required for a wide elemental range, usually no more than three plates can be exposed in the s.s.m.s. per day. Plate reading and calculation can be done at a frequency of about four plates a day. This results in a sample throughput of the whole analytical system of about 9 to 10 samples a week, which is very similar to the system of Millett et al. [1] and Wahlgren et al. [9].

Precision of the spectrum analysis

The instrumental precision of the densitometric measurement of photographically recorded mass spectra was checked by scanning one plate several times under identical conditions, after a sufficient stabilization time. The average standard deviation on the integrated intensity readings of the different scans was 3% [10]. Not only fluctuations in the electronic and optical densitometer components contributed to this variability; the grain structure of the plate also added to the uncertainty. From the data of Franzen et al. [11], it can be calculated that for a 0.015-mm^2 densitometer slit, this so-called "grain noise" alone would result in a minimum standard deviation of 2.6%.

One photoplate was analysed three times in three consecutive weeks, which implies slightly different settings in the densitometer of minimum and maximum transmission, plate focusing and alignment and ADC-amplifier parameters. (In particular, the setting of the maximal transmittance has an important effect on the calculation of relative sensitivity coefficients and thus on the reproducibility of the results of different analyses.) The average standard deviation on the results for 16 elements was 4.6%.

The calibration and calculation principles can be evaluated ideally through the known isotopic abundances of multi-isotopic elements. The isotope ratios are not influenced by sample heterogeneities, variations in spark conditions or mass spectrometer transmission. Also, plate sensitivities are nearly equal and plate heterogeneities are small for isotopes of the same element. The isotopic ratios of Ag, Sb, Re, Eu and Er in 16 different mass spectra were investigated. For each plate the spectrum analysis procedure was applied and the apparent concentrations of the elements were determined from every isotope. The average r.s.d. on the concentrations was 5.1% which is not significantly higher than the 4.6% precision of the densitometer reading. When measurements on spectra with a highly variable background and with doublet peaks were also taken into account, a 6% precision was observed. These values are in good agreement with the existing literature data [1—4].

CONCLUSION

A considerable commitment of time and effort has been devoted to the development of this program for qualitative and quantitative analysis of photographically recorded spark-source mass spectra. Yet, two years of experience with this system have shown that this pays off ultimately in the speed, accuracy and completeness achieved by this automated evaluation, especially for new types of spectra. In the analysis of carbon electrodes [12] with known composition, false identifications have practically never been observed. In those cases where the program cannot take a decision because too many possibilities of identification are open, enough information is printed out so that the operator can usually find other evidence for identification.

The system is readily comprehensible by any mass spectrometer operator and is sufficiently fast and flexible for use on a routine basis.

REFERENCES

1 E. J. Millett, J. A. Morice and J. B. Clegg, Int. J. Mass Spectrom. Ion Phys., 13 (1974) 1.
2 R. A. Burdo, J. R. Roth and G. H. Morrison, Anal. Chem., 46 (1974) 701.
3 J. Degrève and E. D. Champetier de Ribes, Int. J. Mass Spectrom. Ion Phys., 4 (1970) 125.
4 E. B. Owens and N. A. Giardino, Anal. Chem., 35 (1963) 1172.
5 E. Van Hoye, F. Adams and R. Gijbels, Bull. Soc. Chim. Belg., 84 (1975) 595.
6 W. W. Harrison and G. G. Clemena, Anal. Chem., 44 (1972) 940.
7 G. Vidal, P. Galmard and P. Lanusse, Rech. Aérosp., 132 (1969) 49.
8 P. Watson (Institute of Marine Environmental Research, Plymouth, England), private communication.
9 M. A. Wahlgren, D. N. Edgington and F. F. Rawlings, Proceedings of the ANS Topical Meeting on Nuclear Methods in Environmental Research, Columbia, Mo., August 23—24, 1971, p. 97.
10 A. Pilate, private communication.
11 J. Franzen, K. H. Maurer and K. O. Schuy, Z. Naturforsch., 21a (1966) 37.
12 B. Vanderborght and R. Van Grieken, Talanta, submitted.

Analytica Chimica Acta, 103 (1978) 237—243
Computer Techniques and Optimization

COMPUTER INPUT AND GRAPHICAL REPRODUCTION OF CHEMICAL STRUCTURES

E. ZIEGLER* and K. BOLL

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-4330 Mülheim/Ruhr (West Germany)

(Received 3rd May 1978)

SUMMARY

Software for the input of chemical structures into a computer system and for their pictorial reproduction is described. Structures are easily typed in through the keyboard of an inexpensive storage tube display terminal providing visual control of the input process. They can be drawn in the way the user wants to see them. When retrieved from a structure library, the reproductions are identical to the pictures generated during input. No specialized dedicated hardware is necessary. The software has been applied to a mass spectral library (since 1973) and to a ^{13}C-n.m.r. data collection.

During the past 10 years considerable progress has been achieved in the application of computers to chemical information systems. One of the main problems to be solved in this area has always been the computer storage, retrieval and visual reproduction of chemical structural information, for instance in connection with computerized collections of spectroscopic or literature data. Various approaches have been described, e.g. "line notations" (the "Wiswesser notation" is well known), topological codes in the form of connection tables, etc.

These methods suffer from either one or both of the following deficiencies. First, the pictorial formula normally used in organic chemistry has to be converted to some other form before it can be input into a computer. This conversion and the input itself are often cumbersome and error-prone. Secondly, the visual reproduction of a structural formula through a computer, e.g., as a result of a search, has to be derived from an internal representation such as a connection table or a line notation. Even with sophisticated conversion algorithms, computer programs generate a reproduction which is often difficult to comprehend, at least in the case of complicated molecular structures, and which differ from the formula a chemist would have drawn, especially if working in a field where specific drawing conventions have been developed (e.g. chemistry of sugars).

These shortcomings seriously restrict the routine use of structural information in connection with computerized collections of data. Therefore methods have been investigated to by-pass such difficulties. The IDC (Internationale

Dokumentationsgesellschaft für Chemie, Frankfurt/Main) has developed a dedicated computer system with expensive graphic terminals to facilitate the input of pictorial formulae by using the keyboard of a computer terminal for input and its display screen to control the input process visually [1]. The first deficiency mentioned above could be avoided by this method. Several other graphical systems have been described to facilitate the input of chemical structures [2—4]; these systems, however, do not offer satisfying solutions to the second deficiency, i.e. the proper reproduction of structural formulae.

For the purposes of this Institute, a specialized dedicated system such as the IDC system would have been far too expensive; furthermore, the structural information had to be integrated into existing collections of data and into software packages developed for the central DEC system-10 computer of the Institute, which is extensively used for the acquisition and interpretation of data from a variety of analytical instruments [5, 6].

The goal therefore has been to develop software (in the FORTRAN language) that avoids the shortcomings mentioned above, operates within the central in-house time-sharing system and applies inexpensive graphical displays (Tektronix 4010) that were already in use for several spectroscopic applications. The system described below has been in use since 1973 and has required only minor modifications since then.

The programs are written mainly in the FORTRAN language. Assembler language subroutines, however, are applied for the encoding and decoding of plot commands in 18-bit fields for compressed computer storage.

INPUT OF PICTORIAL FORMULA

The input of a molecular structure is accomplished via the typewriter keyboard of a display terminal (optionally supplemented by a separate numeric keypad). The program interprets the keystrokes as commands for plotting chemical bonds, atoms or text on the display for visual verification. For this purpose the viewable area of the scope is overlaid with a grid of 32 by 32 points (see Fig. 2). The atoms of the structure to be drawn occupy single points of the grid; consequently all chemical bonds start and end on points of the grid.

The interpretation of the keyboard commands depends on the mode of operation of the input program. Five different modes are implemented:

(1) "plot mode" to draw chemical bonds;

(2) "symbol mode" to insert alphanumeric symbols (e.g. atomic symbols) into the drawing of a formula;

(3) "library mode" to enter a formula from the structure library (a data file on disk memory) as substructure into the drawing;

(4) "replot mode" to regenerate (optionally stepwise) a drawing after corrections have been performed (because of the nature of the Tektronix storage tube, no partial erasing is possible; therefore, after the entire scope has been cleared, the structure has to be replotted from the internally stored plotting commands);

(5) "transfer mode" to store the current structure (i.e. the one viewable on the scope) in the library file on disk, or to delete structures from the library, etc.

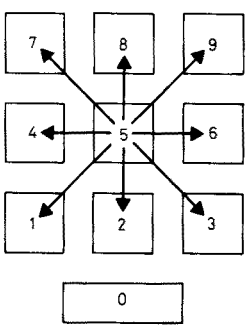
These different modes of operation will be described in more detail.

Operation in the plot mode

This is the main mode of operation to input a pictorial formula to be drawn onto the scope. All bonds (and "no bonds") are drawn from the grid point defined by the current position of a pointer — as identified by a blinking cursor on the screen — to a grid point defined by pressing a key of the numeric keypad. Drawing a bond changes the position of the pointer to the coordinates of the end of the bond. The assignment of keys to directions is shown in Figs. 1 and 2. Normally all bonds are drawn as "long bonds"; the eight most frequently used directions out of the total of 16 possible directions are identified by striking just one of the direction keys, 1, 2, 3, 4, 6, 7, 8 or 9. The remaining eight directions are identified by typing a '5' or '55' first, e.g. '59' or '559', respectively. When short bonds are plotted, the directions are identified as indicated in Fig. 2. Special characters ("L" and "K") have to be typed in order to switch from "short bonds" to "long bonds" or vice versa. The terms "long" and "short" bonds do not imply any chemical meaning; the two "bond lengths", however, provide more flexibility for drawing a picture.

The type of bonding is specified by typing an identifying character before pressing the direction keys, for instance a 'D' for a double bond, an 'S' for "special bond", and no character at all for a single bond.

In the "plot mode", the following bonds (or bond-like symbols) may be drawn (see Fig. 3): single, double, triple bond, dashed line ("special bond"), conjugation and arrow. The precise definition of those 'bonds' is left to the



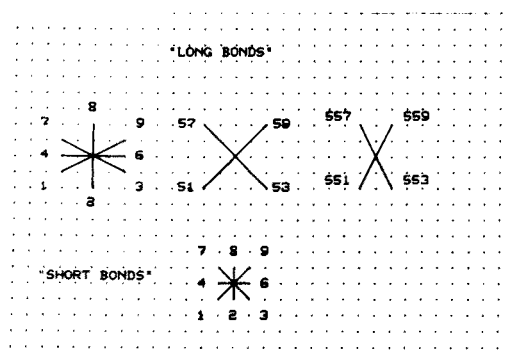


Fig. 1. The direction of a chemical bond to be drawn is specified by striking the corresponding key of the numeric keypad.

Fig. 2. With "long bonds" 16 different directions for drawing a bond are possible; 8 directions are possible with "short bonds".

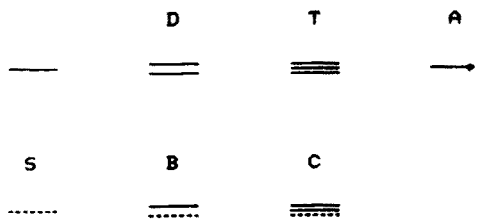


Fig. 3. Seven different types of bonds may be specified. The current implementation provides the bonding symbols shown here.

chemist who uses this software. For instance, the "arrow bond" can be used to write a chemical reaction, if the intention is to build a pictorial library of chemical reactions or just to prepare a slide for a lecture.

Other commands in the plot mode deal with non-drawing movements of the pointer: (1) relative to absolute coordinates, (2) by increments relative to the previous position, (3) to allow the pointer to be positioned via the reticle cursor under control of thumb-wheels, or (4) to allow the last bond drawn or any selected bond to be erased.

Input of atomic symbols and of any text (symbol mode)

The symbol mode is entered from the plot mode simply by typing a blank character; typing of a 'carriage return' then leads back to the plot mode. The symbol mode is used to enter any alphanumeric information starting from the current pointer position. If the length of the text string does not exceed two characters, it is interpreted as an atomic symbol. If, however, an additional quotation character (") is entered, the preceding symbol is considered a special isotope. It should be mentioned that it is not necessary to enter carbon atoms (C) explicitly. Wherever a change in type or direction of a bond occurs, the existence of a carbon atom is assumed, if another symbol is not entered. Hydrogen atoms and their corresponding bonds are normally not drawn at all.

Operation in the library mode

In this mode, a formula that is already stored in the structure library may be specified by an individually assigned number. This formula is then reproduced on the scope, and may be added to a formula already existing on the scope, or modified and then entered into the library as an additional new structure (see Transfer mode, below), or it may replace an already existing formula.

It is also possible to specify by abbreviated names any frequently used partial structures (e.g. TMS for trimethylsilyl-) to be added to the current drawing on the scope in order to save time during structure input.

Replot mode

Since the Tektronix storage tube does not allow for partial erasures, any

corrections that are not simple additions require erasure of the entire scope and the subsequent drawing of the corrected picture. This can be accomplished after switching to the replot mode.

A stepwise generation of the picture can be requested; this makes it possible to halt after each individual plot command and to disregard the rest of the commands.

Additions to the library (transfer mode)

After a picture has reached its intended form on the display screen, the operator can switch to the transfer mode (from the plot mode by typing a carriage return'), assign a unique number within the library to the new structure and transfer it to the library file.

User-dependent definitions

A user (or a group of users) of this software is free to define a set of rules and conventions for his own library. Not only can the interpretation of the bonding symbols be tailored to particular needs, but special conditions within a molecular structure can be marked through special drawing conventions. This is especially useful in sterically more complicated structures, e.g. in the case of stereo-isomerism. Furthermore, it is possible to mark uncertainties within a structure, e.g. uncertainty about *cis* or *trans*, or where branching occurs. A comprehensive set of such rules has been developed by the mass spectrometry group of this Institute.

STORAGE OF STRUCTURES IN THE LIBRARY

The internal representation of a structure is the sequence of plot commands given by the operator during structure input. These plot commands are internally encoded in a compressed way, namely, two plot commands into one computer word of 36 bits; 30 words per structure is the average length of an entry in the mass spectrometry file.

This kind of structure coding considerably reduces the necessary computation to reproduce a structure on the output device (display or plotter) in comparison with other methods. Far more important, however, is the fact that the molecular structure will be reproduced on output in exactly the same pattern as it was in the input process, i.e. in the way the chemist likes to see it.

While conforming to some not too restrictive rules for the drawing of structures, the applied form of internal coding does not prohibit the derivation of connection tables. A program exists to accomplish this. However, up to now no practical use has been made of this capability, and so this piece of software is currently not programmed to perform very efficiently.

EXPERIENCE WITH THE SYSTEM

Originally the effort necessary to input a large number of molecular structures was overestimated. For a trained person, the input of about 30 structures of medium complexity within one hour is not uncommon.

In this Institute two structural libraries have been built up — one in connection with a library of mass spectra containing 3200 structure entries, and another with a collection of ^{13}C-n.m.r. data containing 1500 entries. Each entry in the spectral library contains a reference to a structure number in the structure library file. Library searches therefore result in the output of compound names plus pictorial structures [7]. (In the case of mass spectrometry, structures are encoded for in-house spectra only. The size of the commonly available collection of spectra would require too much manpower to justify the encoding of structures for internal use only.) A spectrum may be extracted from the library and plotted together with the molecular structure. The output has so far been restricted to Tektronix terminals, which are connected to a hardcopy unit to allow copying of the screen contents, and to a software-compatible plotter with TTY-interface.

A computer output presented in this form is more easily read by the spectroscopist and the chemist than other forms where chemical compounds have to be identified by compound names or via line notations. It should be mentioned that during software development considerable effort was put into various cosmetic improvements in order to make the pictorial reproduction of structures as readable as possible and to avoid confusing details. Examples of computer-generated structural drawings are shown in Fig. 4.

Possible extensions. The software as currently implemented does not include structure or substructure search capabilities. The pictorial structure information is used only in addition to other information, e.g. mass spectra,

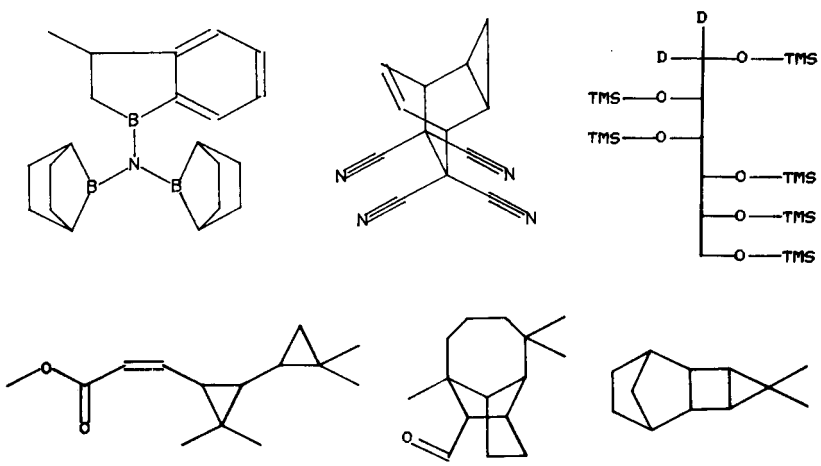


Fig. 4. Examples of computer-generated reproductions of structures stored in the library file.

to provide a clearer identification of chemical compounds; for instance, instead of, or as a supplement to, compound names which are often difficult to read, within a "hit-list" from a library search [7]. Another program, however, has been developed to derive from the structural information fragment codes tailored to specific spectroscopic methods. Because of a change in library search philosophy during recent years, no preselection based on chemical information, e.g. fragment codes, is applied in the present implementations of the search systems. Therefore no further effort went into this part of the software. The same is true for the derivation of connection tables from the encoded structural data.

The design of the software package described in this paper was stimulated by a demonstration of the IDC system by Dr. Kolb and Dr. Neubert in early 1973. Dr. J. Brandt (now at TU Munich) contributed helpful comments during the design phase. The practical experience accumulated by Dr. D. Henneberg and W. Joppek in the daily routine application of this software to mass spectroscopy has led to various improvements.

REFERENCES

1 H. Neubert, Conf. Proc. "24th Meeting of the AGARD Technical Information Panel", Oslo (1971).
2 E. J. Corey and W. T. Wipke, Science, 166 (1969) 178.
3 R. J. Feldmann and S. R. Heller, J. Chem. Doc., 12 (1972) 48.
4 J. E. Blake, N. A. Farmer and R. C. Haines, J. Chem. Inf. Comput. Sci., 17 (1977) 223.
5 E. Ziegler, D. Henneberg and G. Schomburg, Anal. Chem., 42 (8) (1970) 51A; Angew. Chem. Int. Ed. Engl., 11 (1972) 348.
6 E. Ziegler, Computer in der Instrumentellen Analytik, Akadem. Verlagsges., Frankfurt/Main, 1973, pp. 226.
7 H. Damen, D. Henneberg and B. Weimann, Anal. Chim. Acta, 103 (1978) 289.

Analytica Chimica Acta, 103 (1978) 245—251
Computer Techniques and Optimization

A UNIQUE COMPUTER REPRESENTATION FOR MOLECULAR STRUCTURES

CRAIG A. SHELLEY and MORTON E. MUNK*
Department of Chemistry, Arizona State University, Tempe, Arizona (U.S.A.)
ROGER V. ROMAN
Department of Mathematics, Arizona State University, Tempe, Arizona 85281 (U.S.A.)
(Received 3rd February 1978)

SUMMARY

Converting a non-unique connection table to a unique (canonical) name can be accomplished by assigning unique sequence numbers on the basis of topological (constitutional) properties. An algorithm is reported which performs this task by perceiving the topological symmetry of the molecule. A convention for assigning a single name to topologically unique structures which vary only in the position of π-electrons (resonance forms) is also presented.

An interactive program, CASE (Computer-Assisted Structure Elucidation) [1], is being developed to assist in reducing chemical and spectroscopic properties of biomolecules to their structural implications. As an integral part of this process, a molecule assembler, ASSEMBLE [2], constructs all molecular structures consistent with the structural features of an unknown. Although ASSEMBLE is guided by the perception of topological (constitutional) symmetry [3, 4], identical topological (structural) isomers may still be assembled. To provide an unduplicated list of topological isomers, the remaining duplicates must be removed by assigning to each isomer a unique name based on molecular topology.

Chemical structures can be represented in computer form by a connection table showing explicit connections, or in a multitude of different symbolic representations, e.g., the well-known Wiswesser Line Notation (WLN) [5]. For molecule assembly, the necessity of explicit connections makes the connection table format essential. Various algorithms produce unique computer representations from a connection table for molecular structures, e.g., the Morgan algorithm [6], CICLOPS [7], Randic's suggestions [8—10], Corneil's algorithm [11] and Tarjan's algorithm [12]. Converting a connection table to a unique name is a trivial problem if the atoms of the molecule can be assigned sequence numbers in a unique manner independent of the original numbering. The importance of a general scheme for assigning unique

sequence numbers by using a simple set of nomenclature conventions is readily apparent. In this paper, a general algorithm is described which produces a unique sequence numbering for molecules based exclusively on topological properties. Secondly, a form of the connection table is presented which uniquely describes all molecules with the exception of those related by delocalized π-electrons. This convention allows assignment of a single name to chemical structures differing only as resonance forms.

The Morgan algorithm [6] assigns sequence numbers by means of extended connectivity (a topological property), values of which are derived by repeatedly summing connectivity values of nearest neighbors. Although extended connectivity is an efficient method of differentiating atoms, the method fails to provide the maximum possible partitioning in many cases. Randic [9] has identified some of these difficulties as non-uniform convergence, oscillatory behavior and non-equivalent nuclei possessing the same extended connectivity values in every step.

Additional rules which have limited structural relevance and are not strictly based on topology are required in the Morgan algorithm to assign unique sequence numbers. Consequently, efficiency is reduced for highly symmetrical structures. Wipke and Dyott [13] have shown that the time required to assign sequence numbers is lowest for molecules with no topological symmetry and highest when the structure is highly symmetrical, e.g., cubane (I).

The approach used in CICLOPS [7] involves the arbitrary sequencing of atoms within a specifically chosen equivalence class. This procedure produces a unique name for most structures, but for certain highly symmetrical structures, e.g., (I), such an arbitrary assignment leads to an exceptionally large number of possibilities. In addition, the latest version of the algorithm [14], as its predecessor, fails to detect topologically non-equivalent atoms for some structures, e.g., (II). A similar approach by Jochum and Gasteiger [15] also fails to detect topologically non-equivalent atoms in compound (II).

 (I)  (II)

Randic [8] has suggested an algorithm based on connectivity. This scheme associates the numbering of atoms with the smallest binary code of the connectivity matrix. Although it is topologically sound, the procedure suggested for finding the smallest code is not perfect and can occasionally lead to local minima [16]. Recently, Randic suggested an alternative procedure for finding the smallest code [10]. Another suggestion was based on the eigenvector associated with the largest eigenvalue of a molecule [9], but as with the Morgan algorithm, maximum differentiation is not achieved by this technique.

The algorithm of Corneil and Gotlieb [11] is very effective at differentiating atoms, but the method is based on a conjecture which has since been

shown to be incorrect [17]. Tarjan has proposed a fairly complex algorithm which has not yet been programmed in its entirety [12].

TOPOLOGICAL SYMMETRY

Some of the available algorithms [6, 7, 10, 11, 15] determine the topological symmetry of a molecule while constructing a unique name. Thus, the relationship of topological symmetry perception to the assignment of a unique name is evident. Therefore, the perception of topological symmetry is the first step of the sequence numbering algorithm described in this paper.

An algorithm which perceives topological symmetry and which is similar, in part, to the algorithm of Corneil and Gotlieb [11], has been described in earlier publications [3, 4]. The approach consists of three steps.

Step 1. Partition the non-hydrogen atoms into classes by associating with each atom an ordered list of properties. For each atom in a molecule of n non-hydrogen atoms, the ordered property list consists of $n - 1$ integers which indicate the number of 2e covalent bonds joining an atom to non-hydrogen atoms, the element type and the number of distinct elementary cycles of each size in which the atom occurs, starting with those of length 3 and increasing to those of length $n - 1$. Integer class identifiers between one and the number of atoms with different property lists are assigned to each atom in sequential order where the atom with the smallest (lexicographically ordered) property list gets a class identifier of one.

Step 2. Within each equivalence class with more than one member, associate with each atom an ordered ascending property list of integers designating the equivalence classes of adjacent atoms. For each equivalence class, construct new classes by assigning atoms with identical property lists to the same class.

Step 3. Repeat step 2 if it results in more classes, otherwise stop.

The algorithm, although not rigorous in the graph theoretic sense [4], has correctly partitioned the atoms in each molecule of a substantial and diverse collection of organic chemical structures as well as numerous contrived graphs. In fact, some of the graphs successfully solved by this algorithm, e.g., (I) and (II), have served as counterexamples for other algorithms.

As an example, application of the algorithm to 1,2-dicyclopropylethane is shown in Table 1. Application of step 1 requires property lists of length 3 because there are no cycles in this molecule of length greater than 3. Consequently, the property list for step 1 contains three integers indicating the number of bonds to non-hydrogen atoms, the atom type (carbon = 1), and occurrences in cycles of length 3.

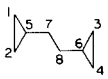
SEQUENCE NUMBER ASSIGNMENT

When a structure possesses no topological symmetry, the algorithm will associate each atom with a unique class. Because the classes, and therefore their atoms, are numbered consecutively from one, this results in the atoms

TABLE 1

Topological symmetry algorithm applied to 1,2-dicyclopropylethane

Sequential order



Atom No.	Step 1		Step 2	
	Property list	Class	Property list	Class
1	(2, 1[a], 1)	2	(2, 3)	2
2	(2, 1, 1)	2	(2, 3)	2
3	(2, 1, 1)	2	(2, 3)	2
4	(2, 1, 1)	2	(2, 3)	2
5	(3, 1, 1)	3	(1, 2, 2)	3
6	(3, 1, 1)	3	(1, 2, 2)	3
7	(2, 1, 0)	1	(1, 3)	1
8	(2, 1, 0)	1	(1, 3)	1

[a]Carbon = 1.

being ordered. The important property of the algorithm is that this ordering of the atoms will be the same regardless of their original ordering. This is true because at each point in the algorithm where class numbers are determined, the values assigned depend solely on the ordering of property lists which contain purely topological properties unrelated to the original numbering scheme.

When a structure possesses some symmetry, i.e., when at least one class contains more than one atom, the partial ordering imposed on the atoms by their class numbers is unique. This partial ordering may be extended to a linear order (each atom being assigned a unique number) by using the following algorithm.

Step 1. Let k be the largest class number among those classes possessing more than one atom. Increase the class number (m) of each class where $m > k$ to $m + 1$. With the exception of one arbitrarily chosen atom from class k, place all atoms currently in class k in a new class, $k + 1$.

Step 2. Within each equivalence class with more than one member, associate with each atom an ordered property list of integers designating the equivalence classes of adjacent atoms. For each equivalence class, construct new classes by assigning atoms with identical property lists to the same class.

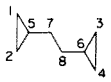
Step 3. Repeat step 2 if it results in more classes.

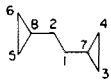
Step 4. If the number of classes is not equal to the number of atoms in the molecule then go to Step 1, otherwise stop.

This algorithm is applied to 1,2-dicyclopropylethane in Table 2. Because the partial ordering imposed by the topological symmetry algorithm is

TABLE 2

Sequence numbering algorithm applied to 1,2-dicyclopropylethane

Sequential order 

Unique sequential order 

Atom No.	Equivalence Class	Step 1	Step 2		Step 2		Step 1	Step 2		Step 1
		Class	Property list	Class	Property list	Class	Class	Property list	Class	Class
1	2	2	(2, 4)	4	(4, 6)	4	5	—	5	6
2	2	2	(2, 4)	4	(4, 6)	4	4	—	4	5
3	2	2	(2, 3)	3	(3, 5)	3	3	(3, 6)	3	4
4	2	2	(2, 3)	3	(3, 5)	3	3	(3, 6)	3	3
5	3	4	—	6	—	6	7	—	7	8
6	3	3	—	5	—	5	6	—	6	7
7	1	1	(1, 4)	2	—	2	2	—	2	2
8	1	1	(1, 3)	1	—	1	1	—	1	1

unique, class k, selected in Step 1, is unique. In Table 2, original atoms 5 and 6 will always be placed in equivalence class 3 and consequently k will always be 3 when Step 1 is initially performed. Although the arbitrary choice of an atom from class k does admit the possibility of non-unique ordering of the atoms with respect to their initial numbering, the connection tables associated with any of these potential orderings are identical.

RESONANCE

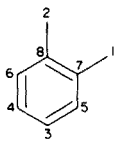
Although connection tables take a variety of forms, they possess similar characteristics: sequence numbers assigned to individual atoms, connectivity between atoms, and labels specifying atom properties. Bonds may also be labeled, e.g., single, double, aromatic and triple, but connection tables based on multiple connections for multiple bonds (e.g. *o*-xylene, Table 3) are most convenient for algorithms designed to assemble molecular structures. These unfortunately give rise to topological differences between π-resonance forms, e.g., the two Kekulé forms of *o*-xylene. The required redundancy in the connection tables of such systems can be attained, while at the same time retaining unambiguity for all other chemical systems*, by replacing all multiple connections with single connections without altering the value designating the number of 2e covalent bonds joining an atom to non-hydrogen atoms; *o*-xylene is also represented in this manner in Table 3. Although this representation may not be suitable for all graphs, the limited valency of the atoms found in chemical structures may account for its successful application in this instance.

*Conjugated cyclic structures which do not obey Hückel's rule ($4n + 2\pi$ electrons, $n = 0$, 1, 2,...) are identical when this convention is used.

TABLE 3

Connection tables of *o*-xylene

Unique sequential order



Atom No.[a]	Only connections given		Multiple bonds represented by multiple connections	
	Connected atoms	Connections	Connected atoms	Connections
1	(7)	1	(7)	1
2	(8)	1	(8)	1
3	(4, 5)	3	(4, 5, 5)	3
4	(3, 6)	3	(3, 6, 6)	3
5	(3, 7)	3	(3, 3, 7)	3
6	(4, 8)	3	(4, 4, 8)	3
7	(1, 5, 8)	4	(1, 5, 8, 8)	4
8	(2, 6, 7)	4	(2, 6, 7, 7)	4

[a]The atom type is C in all cases.

This convention has been incorporated into the molecule assembler (ASSEMBLE) of program CASE. It has given correct isomer counts for over 100 molecular formulae tested [2] and consistently eliminates identical resonance structures. This method represents a novel procedure of assigning a single name to chemical structures which differ only in the position of π-electrons.

The simple nomenclature conventions described for obtaining unique sequence numbers for structures of any complexity result in an efficient means of obtaining unique structural codes based entirely on topological properties.

The authors acknowledge the support of this project by the National Institutes of Health (GM21703) and Arizona State University Computer Services.

REFERENCES

1 C. A. Shelley, et al. in D. H. Smith (Ed.), Computer-Assisted Structure Elucidation, A.C.S. Symposium Series, Vol. 54, 1977, p. 92.
2 C. A. Shelley, M. E. Munk and R. V. Roman, Anal. Chim. Acta, 103 (1978) 00.
3 C. A. Shelley and M. E. Munk, J. Chem. Inf. Comp. Sci., 17 (1977) 110.
4 C. A. Shelley, M. E. Munk and R. V. Roman, J. Chem. Inf. Comp. Sci., submitted.
5 W. J. Wiswesser, Comput. Autom., 19 (1970) 2.

6 H. L. Morgan, J. Chem. Doc., 5 (1965) 107.
7 J. Blair, et al., Tetrahedron, 30 (1974) 1845.
8 M. Randic, J. Chem. Phys., 60 (1974) 3920.
9 M. Randic, J. Chem. Inf. Comp. Sci., 15 (1975) 105.
10 M. Randic, J. Chem. Inf. Comp. Sci., 17 (1977) 171.
11 D. G. Corneil and C. C. Gotlieb, J. Assoc. Comp. Mach., 17 (1970) 51.
12 R. E. Tarjan in R. E. Christoffersen (Ed.), Algorithms for Chemical Computations, A.C.S. Symposium Series, Vol. 46, 1977, p. 1.
13 W. T. Wipke and T. M. Dyott, J. Am. Chem. Soc., 96 (1974) 4834.
14 W. Schubert and S. Ugi, J. Am. Chem. Soc., 100 (1978) 37.
15 C. Jochum and J. Gasteiger, J. Chem. Inf. Comp. Sci., 17 (1977) 113.
16 A. L. Mackay, J. Chem. Phys., 62 (1975) 308.
17 D. G. Corneil, Tech. Rep. No. 65, Dept. of Comp. Sci., Univ. of Toronto, Toronto, Ontario, Canada, 1974.

Analytica Chimica Acta, 103 (1978) 253—262
Computer Techniques and Optimization

RELATIONSHIP BETWEEN THE AUTOCORRELATION TECHNIQUE AND AN ANALYSIS OF VARIANCE SCHEME IN TIME SERIES ANALYSIS OF FIRST-ORDER AUTOREGRESSIVE STOCHASTIC STATIONARY PROCESSES

C. B. G. LIMONARD* and F. W. PIJPERS

Department of Analytical Chemistry, Catholic University of Nijmegen, Toernooiveld, Nijmegen (The Netherlands)

(Received 25th April 1978)

SUMMARY

Autocorrelation analysis of time series yields information on the transient behaviour of such series. An appropriately designed scheme for the analysis of variance can, in principle, give the same information. A relationship between the two techniques, with regard to their ability to verify and quantify a time constant of first-order autoregressive stochastic stationary processes, is described. Monte Carlo simulations were performed to test the validity and applicability of the theory. A practical example of the procedure is included.

Quality-control procedures are indispensable for maintaining working standards in clinical chemistry laboratories. The goal is to assess the accuracy and precision of analytical procedures and to detect deviations from a target value. Any deviations can then be suppressed so that their accumulation does not cause the analytical procedure to run out of control. In clinical chemistry, quality-control standard sera containing analytes such as calcium, urea, cholesterol, etc. are used. Samples of these standards are included in series of unknown samples which are analyzed for the particular analyte so that the performance of the analytical procedure during the working time can be assessed. Statistical analysis of the quality-control data then indicates the accuracy and precision of the analytical procedure.

Dynamic aspects of these quality-control systems are under current investigation [1]. The underlying hypothesis is that an analytical procedure can be represented as a first-order autoregressive stochastic stationary process. Figure 1 shows the analytical procedure as the process, with underlying disturbances $Z(t)$, that converts a physical or chemical quantity, i.e. some property of the samples, to a value representative of the sample. The quality control data, measured at equally spaced time intervals T_A, contain the information on the behaviour of the process. These data are assembled by the measuring unit M, which in turn is subject to measuring errors, $\nu(t)$. Interventions initiated by control unit C, based on the quality control data, are intended to neutralize

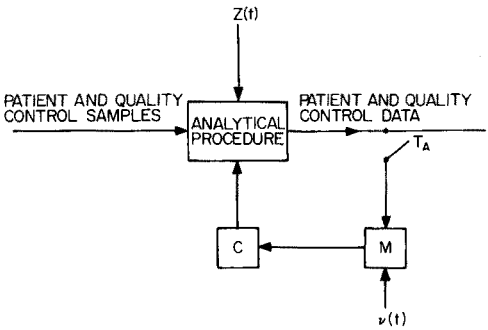


Fig. 1. Schematic diagram of an analytical procedure as a process to be controlled. M, measuring unit; C, control unit; T_A, sampling frequency of quality control data; $\nu(t)$, measuring errors; $Z(t)$, disturbances causing the process to run out of control.

the effects causing the discrepancies from the target value. For appropriate interventions to be made, the measurements must conform to the actual process value as closely as possible. The efficiency of the control unit (Fig. 1) can be derived from the measurability and controllability rules of Van der Grinten [2].

The practical application of these rules is that the efficiency of various control systems can be quantified and compared. The important characteristics are the analysis time, the sampling frequency, the standard deviation and time constant of the process, and the measuring errors $\nu(t)$. The time constant of the process, T_x, is inversely proportional to the rate of the process fluctuations.

Autocorrelation analysis makes it possible to identify processes from discrete time series, in this case quality control data over a prolonged period of time. The technique allows quantification of T_x. A necessity for application of the technique is that data over a large time span are available or can be collected.

During the project mentioned [1], situations arose where these data were not available from files documenting the recent history of a laboratory. However, some laboratories could retrieve the results of analysis of variance schemes, because such schemes formed part of their laboratory control systems. The basic idea behind the present investigation is that both autocorrelation and analysis of variance give information on the transient behaviour of a process. It is demonstrated that the time constant T_x of a first-order autoregressive stochastic stationary process can be estimated from the results of an appropriate analysis of variance scheme. A functional relationship between autocorrelation and analysis of variance for such processes is derived. Monte Carlo simulations were done to test the validity and applicability of the theory. A practical example is given.

Autocorrelation

The processes investigated are considered to be first-order autoregressive stochastic and stationary. This means that the process value x_t at time t is not fixed by a mathematical model, but is defined only as probability distributions $\rho(x_t)$. For Gaussian processes this distribution is determined by the mean value $\epsilon(x_t)$ and the standard deviation σ_{x_t} [3]:

$$\epsilon(x_t) = \int_{-\infty}^{+\infty} x_t \,.\, \rho(x_t)\, dx_t \text{ and } \sigma_{x_t}^2 = \int_{-\infty}^{+\infty} \{x_t - \epsilon(x_t)\}^2\, \rho(x_t)\, dx_t$$

If $\epsilon(x_t)$ and σ_{x_t} are constant in time, the process is called stationary. In this case $\epsilon(x_t)$ and σ_{x_t} can be estimated from a limited time series, consisting of N values observed at constant time intervals. Thus

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

The correlations between successive observations are calculated by computing the autovariance estimates, c_τ, from

$$c_\tau = \sum_{i=1}^{N-\tau} (x_i - \bar{x}) \,.\, (x_{i+\tau} - \bar{x})/(N-\tau) \quad (\tau = 0, 1, \ldots\ldots m) \tag{1}$$

where τ is the time lag expressed in units of sampling interval and N is the length of the series in the same units. Autocorrelation estimates, r_τ, are then obtained from $r_\tau = c_\tau/c_0$, where $\tau = 0, 1, \ldots\ldots m$. For first-order stochastic stationary processes, the autocorrelation function r_τ is a continuously decreasing function described by $r_\tau = \exp(-\tau/T_x)$, in which T_x, the time constant of the process, is a measure of the frequency of the process fluctuations. For $T_x = 0$ all frequencies occur and such a process is called "white noise".

Analysis of variance

The analysis of variance (anova) technique permits estimation of one or more factors suspected of contributing significantly to the total uncertainty of a measured quantity. Generally, a scheme is designed which allows the mean square between groups, s_B^2, to be estimated in relation to s_w^2, the mean square within groups of time series.

Here the variance of the measured quantity with respect to its average value over a considerable period of time, is compared with the variance of the same quantity over a relative short period. For a first-order autoregressive stochastic stationary process (henceforth called process), with $T_x = 0$, this would result in equal expected values for both variance estimates, whereas a process with $T_x \neq 0$ (provided that the sampling frequency is greater than T_x^{-1}), should show a smaller variance for the short period than for the total period, because of the correlation between successive measurements.

THEORETICAL MODEL

According to this anova scheme, the ratio s_w^2/s_b^2 is given by

$$\frac{s_w^2}{s_b^2} = \frac{(k-1)}{nk\,(n-1)} \cdot \frac{\left\{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{i,j}-\bar{x}_i)^2\right\}}{\left\{\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2\right\}} \tag{2}$$

in which k is the number of groups; n is the number of measurements $x_{i,j}$ per group; $\bar{x}_i$ is the mean value of the n measured quantities of group i; and $\bar{x}$ is the mean value of all $n.k$ $(=N)$ measurements. After the process has been sampled at a constant sampling rate, the time series consists of N observations. Groups are formed from equal sets of consecutive measurements and the ratio shown in eqn. (2) is calculated. The autocorrelation estimate, r_1, which is the maximum likelihood estimate of T_x, is also computed.

In order to relate the $1/T_x$ value of this calculated estimate to the results of the anova scheme, the summations in eqn. (2) are rearranged to give

$$\frac{s_w^2}{s_b^2} = \frac{(k-1)}{k\,(n-1)} \cdot \frac{\left\{n\sum_{i}^{k}\sum_{j}^{n}x_{i,j}^2 - \sum_{i}^{k}\left(\sum_{j}^{n}x_{i,j}\right)^2\right\}}{\left\{\sum_{i}^{k}\left(\sum_{j}^{n}x_{i,j}\right)^2 - 1/k\left(\sum_{i}^{k}\sum_{j}^{n}x_{i,j}\right)^2\right\}} \tag{3}$$

with $\Sigma_i^k = \Sigma_{i=1}^k$ and $\Sigma_j^n = \Sigma_{j=1}^n$, or

$$\frac{s_w^2}{s_b^2} = \frac{(k-1)}{k\,(n-1)} \cdot \frac{\left\{n\sum_{i}^{k}\sum_{j}^{n}(x_{i,j}-\bar{x})^2 - \sum_{i}^{k}\left\{\sum_{j}^{n}(x_{i,j}-\bar{x})\right\}^2\right\}}{\left\{\sum_{i}^{k}\left\{\sum_{j}^{n}(x_{i,j}-\bar{x})\right\}^2\right\}} \tag{4}$$

In eqn. (4), all x values have been diminished by $\bar{x}$, in accordance with the autocorrelation technique. The summation of the denominator can be expande

$$\sum_{i}^{k}\left\{(x_{i,1}-\bar{x})^2+(x_{i,2}-\bar{x})^2+\ldots.+(x_{i,n}-\bar{x})^2\right\}$$

$$+2\left\{\sum_{i}^{k}(x_{i,1}-\bar{x})(x_{i,2}-\bar{x})+\ldots.+\sum_{i}^{k}(x_{i,n-1}-\bar{x})(x_{i,n}-\bar{x})\right\}$$

$$+2\left\{\sum_{i}^{k}(x_{i,1}-\bar{x})(x_{i,1+2}-\bar{x})+\ldots.+\sum_{i}^{k}(x_{i,n-2}-\bar{x})(x_{i,n}-\bar{x})\right\}$$

$$+2\left\{\sum_{i}^{k}(x_{i,1}-\bar{x})(x_{i,n}-\bar{x})\right\}$$

The expression in the first set of curly brackets is an estimate of nc_0 (eqn. 1 for $\tau = 0$); in the second set, $c_0\,(n-1)\exp(-1/T_x)$, (eqn. 1 for $\tau = 1$); in the third set $c_0\,(n-2)\exp(-2/T_x)$; and in the last $c_0\,\{n-(n-1)\}\exp(-(n-1)/T_x)$.

Substitution in eqn. (4) and some rearrangement yields

$$\frac{s_w^2}{s_b^2} = \frac{(k-1)}{k(n-1)} \cdot \frac{\left\{n - 1 - 2\sum_j^n (n-j)/n \exp(-j/T_x)\right\}}{\left\{1 + 2\sum_j^n (n-j)/n \exp(-j/T_x)\right\}} \tag{5}$$

In eqn. (5), s_w^2/s_b^2 is used instead of s_b^2/s_w^2, the more commonly used expression in the anova scheme, because of the limiting values of the former expression, which are 0 and $(k-1)/k$ for $T_x \to \infty$ and $T_x \to 0$, respectively.

In the derivation of eqn. (5), it was assumed that the time series are a sample from an infinite population of such series, representative of a first-order autoregressive stochastic stationary process.

Simulation model

First-order autoregressive stochastic stationary processes were simulated with a discrete "white-noise" generator from the IBM Library. With u representing a discrete white-noise process with zero mean, variance σ_u^2 equal to one, and $E(u_n \cdot u_m) = 0$ for $m \neq n$, a Markov process can be generated [4]:

$$x_{t+1} = ax_t + bu_t \tag{6}$$

where $a = \exp(-1/T_x)$ and $\sigma_x = b(1-a^2)^{-\frac{1}{2}}$. Time series of 1024 observations with various $1/T_x$ values were generated and s_w^2/s_b^2 values were calculated. The ratio of the number of measurements within groups (n) to the number of groups (k) was varied from 2:512 to 512:2 in multiples of four for each simulation.

RESULTS AND DISCUSSION

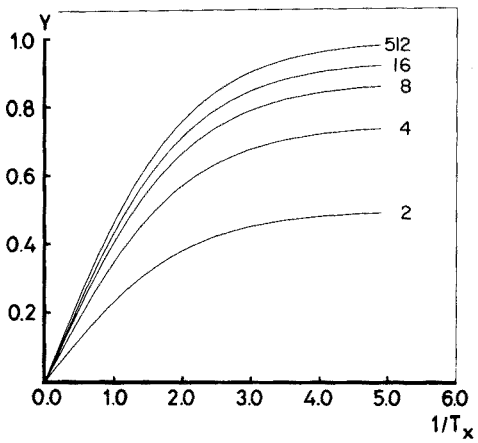
Highly correlated x_t values were obtained upon substitution of $1/T_x$-values of approximately zero in eqn. (6). In such time series, s_w^2 was small, approaching zero as $T_x \to \infty$. As s_b^2 had higher values than s_w^2, the ratio s_w^2/s_b^2 approached zero for all ratios of n/k.

Generating a process with high $1/T_x$ values gave time series with practically uncorrelated x_t values. Here the ratio s_w^2/s_b^2 depended on the number of groups k only and approached the asymptotic value $(k-1)/k$.

These observations are in close agreement with the graphical representation of eqn. (5) shown in Fig. 2, where s_w^2/s_b^2 is plotted as a function of $1/T_x$ for various values of k. Here T_x was computed from r_1, which is the maximum likelihood estimate of this parameter, from the equation $r_\tau = \exp(-\tau/T_x)$, and the value was substituted in eqn. (5).

Table 1 gives a comparison between the experimental s_w^2/s_b^2 ratios calculated from eqn. (5), and those predicted by calculation by anova from the generated time series.

An increase in n, which directly implies a decrease in k, results in larger deviations between predicted and simulated values. To test the dependence of



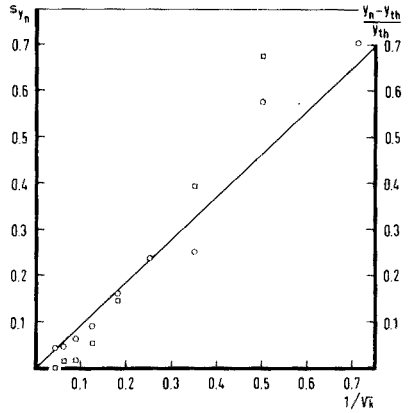


Fig. 2. s_w^2/s_b^2 (= y) ratio as a function of $1/T_x$ according to eqn. (5) for various numbers of groups k.

Fig. 3. Standard deviation of eight s_w^2/s_b^2 ratios from independently simulated time series with $\bar{x} = 1.0$, $s_x = 1.0$, and $T_x = 1.0$ as a function of $1/k^{-\frac{1}{2}}$, where k is the number of groups, compared with relative differences between calculated and simulated s_w^2/s_b^2 ratios. (○) s_{y_n}; (□) $(y_n - y_{th})/y_{th}$.

TABLE 1

Comparison between experimental and predicted s_w^2/s_b^2 values calculated from eqn. (5) for various $1/T_x$ values and $k = 1024/n$

T_x^{-1}	0.36 s_w^2/s_b^2		4.84 s_w^2/s_b^2		5.10 s_w^2/s_b^2	
n	Exp.	Calc.	Exp.	Calc.	Exp.	Calc.
2	0.18	0.18	1.04	0.98	1.05	0.99
4	0.16	0.16	0.99	0.98	1.00	0.98
8	0.15	0.16	0.98	0.98	0.99	0.98
16	0.17	0.16	0.99	0.97	0.99	0.97
32	0.26	0.16	1.10	0.95	1.10	0.96
64	0.30	0.16	1.16	0.92	1.17	0.93
128	0.17	0.16	1.43	0.86	1.44	0.86
256	0.19	0.16	0.99	0.74	0.99	0.74
512	0.09	0.12	0.36	0.49	0.36	0.49

the predicted ratios on k, at constant total time, eight mutually independent simulations were performed, all with $T_x = 1.0$, $s_x = 1.0$ and $\bar{x} = 1.0$ (Table 2). The last two columns show that both $\bar{x}$ and s_x values agree well with the input values. The bottom two lines demonstrate that $\bar{y}_n$, the average over eight measurements of s_w^2/s_b^2, and s_{y_n} increase with increasing n. This is in accordance with the results obtained for other values of $1/T_x$. The standard deviation for

TABLE 2

Eight independent simulations with $T_x = 1.0$, i.e. $1/T_x = 1.0$, $s_x = 1.0$ and $\bar{x} = 1.0$

	$s_w^2/s_b^2 = y_n$										
n =	2	4	8	16	32	64	128	256	512	$\bar{x}$	s_x
a	0.493	0.480	0.489	0.556	0.799	1.000	0.678	0.712	0.316	1.062	0.983
b	0.515	0.449	0.504	0.499	0.536	0.495	0.742	0.463	1.960	1.018	1.010
c	0.436	0.422	0.384	0.358	0.311	0.258	0.197	0.099	0.034	1.037	1.040
d	0.396	0.364	0.387	0.378	0.295	0.290	0.354	0.216	0.254	0.977	1.018
e	0.517	0.514	0.495	0.498	0.546	0.508	0.535	0.311	0.184	0.925	0.966
f	0.443	0.472	0.530	0.588	0.460	0.596	0.570	0.550	0.528	0.936	0.994
g	0.446	0.468	0.457	0.529	0.498	0.455	0.413	0.366	0.250	0.956	1.008
h	0.432	0.444	0.372	0.369	0.617	0.695	1.013	1.919	1.483	1.015	1.018
$\bar{y}_n$	0.460	0.451	0.452	0.472	0.508	0.537	0.562	0.580	0.626		
s_{y_n}	0.043	0.045	0.062	0.091	0.163	0.236	0.253	0.574	0.702		

TABLE 3

Comparison between calculated and predicted s_w^2/s_b^2 values as a function of the number of groups k

k	y_{th}[a]	$\bar{y}_k$[b]	s_{y_k}[c]	$k^{-\frac{1}{2}}$	$\lvert y_{th} - \bar{y}_k \rvert / y_{th}$
512	0.461	0.460	0.043	0.044	0.002
256	0.444	0.451	0.045	0.063	0.016
128	0.445	0.452	0.062	0.088	0.016
64	0.447	0.472	0.091	0.125	0.056
32	0.443	0.508	0.162	0.177	0.147
16	0.431	0.537	0.236	0.250	0.246
8	0.403	0.562	0.253	0.353	0.395
4	0.346	0.580	0.574	0.500	0.676
2	0.231	0.626	0.702	0.707	1.710

[a] Calculated s_w^2/s_b^2 from eqn. (5) with $T_x = 1.00$.
[b] Mean s_w^2/s_b^2 from eight independent simulations with $T_x = 1.00$ (Table 2).
[c] Standard deviations of eight s_w^2/s_b^2 values (Table 2).

all the values of $\bar{x}$, $s_{\bar{x}}$ is 0.05. This is in accordance with a reduction of the input value of $s_x = 1.0$ by averaging 1024 data ($2.0/(1024)^{\frac{1}{2}} \cong 0.06$).

The increase in s_{y_n} as a function of n is due to a reduction in the number of groups k; its influence on the variance of s_b^2 is illustrated in Fig. 3. With increasing s_{y_n}, the predictability of s_w^2/s_b^2 decreases; this is illustrated in the same figure by plotting $(\bar{y}_n - y_{th})/y_{th}$ as a function of $1/k^{\frac{1}{2}}$.

For the above example, it is clear that prediction of a y-value with an accuracy > 85% from a time series with $s_x = 1.0$, the value of $1/k^{-\frac{1}{2}}$ should be less than 0.2, i.e. more than 25 groups are required. (See numerical example in Table 3.)

TABLE 4

Comparison between T_x values for analysis series calculated from eqn. (7) for $n = 2$, with substitution of s_w^2/s_b^2, directly computed from the anova scheme, and those derived from the autocorrelation estimate r_1 of the same time series

Run	Total time k (days)	s_w^2/s_b^2 (y)	T_x from eqn. (7)	T_x from r_1	$1/k^{\frac{1}{2}}$ (%)
a	288	0.284	1.7	2.0	6
b	146	0.246	2.0	1.9	8
c	228	0.312	1.5	—	7
d	283	0.353	1.4	—	6

Example of application to clinical data

Table 4 gives an example of the application of eqn. (5) to clinical chemical data. During the period 1974/1975, four control sera were used in a survey of urea determinations. The total time during which one of these sera was used is listed in column 2. For the more recently employed control sera (groups a and b) actual analytical results as well as results from a one-way analysis of variance [5] were available. The anova scheme permitted the calculations of between-day and within-day variance contributions and thus s_w^2/s_b^2 (eqn. 7, see below). Laboratory organization was such that every analysis series included one control serum sample to check the performance of the procedure; the quality control samples, within a working day, were used at equally spaced intervals. Various analysis series were run per day. For the control sera (groups a and b), the autocorrelation estimate r_1 was calculated from $r_\tau = c_\tau/c_0$, and T_x was estimated. T_x was also computed from the anova results obtained from the measurements of these sera by eqn. (7) on substitution of the s_w^2/s_b^2 ratio from the laboratory files. It can be seen that the result of the two methods agree well. For the older groups (c and d) only anova results were available. The T_x estimates listed in column 4 of Table 4 stem from these data and were obtained by application of eqn. (7) after substitution of the s_w^2/s_b^2 ratio from the anova results.

The values of T_x are expressed in units of "time required to process one complete analysis series of patient and quality control samples". This time unit depends on the number of analysis performed per series and may vary between 40 and 50 min. Up to twelve analysis series per day were available. However, because extra series did not change the numerical value of the s_w^2/s_b^2 ratio (eqn. 5) significantly, the ratio was computed from two equally spaced quality control results, which were obtained in two successive analysis series.

In such a situation, when $n = 2$, eqn. (5) reduces to

$$y = \frac{s_w^2}{s_b^2} = \frac{k-1}{k} \cdot \left\{ \frac{1 - e^{-1/T_x}}{1 + e^{-1/T_x}} \right\}$$

or

$$\frac{1}{T_x} = \ln\left\{\frac{1-k\,(y+1)}{1+k\,(y-1)}\right\} \qquad (7)$$

where k is the total time expressed in days. The results listed in column 4 of Table 4 were calculated using this equation.

Conclusions

Based on the theory of first-order autoregressive stochastic stationary processes, the relationship derived between a one-way analysis of variance scheme and the autocorrelation technique is shown to be valid and applicable for estimation of the time constant of first-order autoregressive stochastic stationary processes from discrete time series.

In situations where original data are not available but analysis of variance results are, the time constant can be estimated after it has been verified that the processes under investigation are indeed first-order autoregressive stochastic and stationary. Verification of such a model from the analysis of variance results alone is impossible. Consequently, it is necessary to ascertain if the process investigated is of this kind.

The project investigating the dynamic aspects of quality control systems in clinical chemistry laboratories was based on the hypothesis that analytical procedures can be represented as first-order autoregressive stochastic stationary processes. The results of autocorrelation analysis applied to the data of one of the participating laboratories confirmed this assumption [1]. Consequently the relationship derived between the output of a one-way analysis of variance scheme and the results of autocorrelation analysis was investigated for this practical example.

The results indicate the validity of the procedure with discrepancies of +15% and −5%, respectively, when both techniques were applied to the same discrete time series. These discrepancies do not hamper practical applications of the time constant in measurability and controllability rules for quantifying the efficiency of the quality control systems of a laboratory.

The estimated time constants (1.5 and 1.4) of the series where only analysis of variance results were available, confirmed the behaviour of the process, i.e. a time constant between 1 and 2 analysis series. The merit of the derived relationship between analysis of variance and autocorrelation is that although original data were not available, time constants could still be estimated.

The autocorrelation technique is preferable for verifying the time constant of a process, because it also indicates the correctness of an applied model. However, where this is impossible and where it can be presumed that the data under consideration stem from a first-order autoregressive stochastic stationary process, the derived relationship gives insight into the transient behaviour of the process.

REFERENCES

1 C. B. G. Limonard, J. Clin. Chem. Clin. Biochem., 15 (1977) 172.
2 P. M. E. M. van der Grinten, Statist. Neerland., 22 (1968) 43.
3 G. E. P. Box, G. M. Jenkins, Time series analysis: Forecasting and Control, Holden-Day, San Francisoo, 1970, p. 26.
4 T. H. Naylor, J. L. Balintfy and D. S. Burdick, Computer simulation techniques, J. Wiley, New York, 1966, p. 43.
5 O. L. Davies and P. L. Goldsmith, Statistical Methods in Research and Production, 4th edn., Oliver and Boyd, Edinburgh, 1972, p. 125.

Analytica Chimica Acta, 103 (1978) 263—267
Computer Techniques and Optimization

Short Communication

A METHOD FOR THE REFINEMENT OF INITIAL PARAMETERS IN THE RESOLUTION OF OVERLAPPING SPECTRAL BANDS BY LEAST-SQUARES PROCEDURES

F. GÓMEZ-BELTRÁN* and A. SALAS

Departamento de Química-Física, Universidad de Oviedo, Oviedo (Spain)

A. VALERO

Escuela Técnica Superior de Ingenieros Industrials, Universidad de Zaragoza, Zaragoza (Spain)

(Received 28th June 1978)

The spectral band decomposition problem has been widely investigated, and many methods have been developed for resolution of the envelope of a band into the single absorptions of its components. Normally, the final solution is attained through a least-squares procedure which optimizes what is considered as the "best starting hypothesis". Various computer programs (MROCOS, BANDAN, SUAN, GSAN, LOGFIT, etc.) have been developed which solve the problem more or less successfully [1—13].

When severe overlap among the component bands does occur, the problem remains far from being satisfactorily solved. In this situation the most important question is the unicity of the final result [3, 4, 9—11, 14—16]. From a purely mathematical point of view, the solution obtained is unique indeed [17], but sometimes these mathematically acceptable final solutions resulting from an apparently good starting hypothesis are physically meaningless. In these cases, a very common "test" practice is to estimate different sets of starting parameters, use them to calculate the corresponding final solutions, and choose from these the result which is best suited to mathematical and physical criteria.

In the search for objective automatic procedures to solve the spectral decomposition problem, attempts have been made to replace this "test" method by refinement procedures which work on the starting parameters to drive them to their "best" condition as a starting point for the subsequent least-squares treatment. Such a refinement procedure must conserve the physical meaning of these parameters throughout the calculations.

Lischka and Derflinger [13] have reported a computer program (BKORR) which refines band widths in the case of Gaussian absorptions. This program is based on the second derivative of the experimental contour. However, the basic idea of the refinement was not fully developed and therefore the BKORR method gives only partially good results.

The present communication reports a refining method which includes simultaneous treatment of band positions (λ_0), maximum heights (ϵ_0) and band widths (b), and utilizes the experimental contour and its first and second derivatives to obtain significant improvements of the starting parameters of the bands. Afterwards, if necessary, a final least-squares treatment can be used to finish the calculation by optimizing the nearly correct set of parameters obtained by the proposed method.

Mathematical background

The method of calculation refines the band parameters by applying the "consistency condition" of one particular set of bands to the experimental data. When this condition is applied to the contour ($\epsilon_{t,i}$) and its first ($\epsilon'_{t,i}$) and second ($\epsilon''_{t,i}$) derivatives in the positions of the assumed maxima, the equations are

$$(d^k\epsilon/d\lambda^k)_{t,i} = (d^k\epsilon/d\lambda^k)_{0,i} + \sum_{i \neq j}^{N} (d^k\epsilon/d\lambda^k)_{j,i}; \; k = 0, 1, 2 \tag{1}$$

where N is the number of bands, $(d^k\epsilon/d\lambda^k)_{t,i}$ is the k-order derivative value corresponding to the experimental contour at the maximum of the i-th band ($\lambda_{0,i}$). Subscripts "o" and "j" relate to the contribution, at this position, of the i-th and j-th bands, respectively. For example, if the energy-symmetric Gaussian function is used as the band model ($\epsilon = \epsilon_0 \exp\{[(\lambda_0/\lambda)-1]^2/b^2\}$), then for an isolated band: $\lambda_{0,i} = \lambda + \epsilon'_i(\epsilon''_{0,i})^{-1}$, and $b_i = (1/\lambda_i)[2\epsilon_{0,i}(\epsilon''_{0,i})^{-1}]^{\frac{1}{2}}$, where λ represents a value which is close enough to the i-th maximum position ($\lambda_{0,i}$) that $\epsilon = \epsilon_0$. ϵ'_i is the first derivative value at this position.

Finally the following system of equations can be developed:

$$\epsilon_{0,i} = \epsilon_{t,i} - \sum_{i \neq j}^{N} \epsilon_{j,i} \tag{2.a}$$

$$\lambda_{0,i} = \lambda_i + \left(\epsilon'_{t,i} - \sum_{i \neq j}^{N} \epsilon'_{j,i}\right)\left(\epsilon''_{t,i} - \sum_{i \neq j}^{N} \epsilon''_{j,i}\right)^{-1} \tag{2.b}$$

$$b_i = (1/\lambda_i)\left[2\left(\epsilon_{t,i} - \sum_{i \neq j}^{N} \epsilon_{j,i}\right)\left(\epsilon''_{t,i} - \sum_{i \neq j}^{N} \epsilon''_{j,i}\right)^{-1}\right]^{\frac{1}{2}} \tag{2.c}$$

where λ_i is the initially estimated value for the i-th maximum position. Peak-search methods [1, 4, 9, 15, 16, 18—24] allow sufficiently precise estimates of $\lambda_{0,i}$ values.

Experimental profile derivatives are calculated by the Savitzky and Golay method [23] by using a variable sampling interval appropriate to the specific requirements of each region of the spectrum. Interpolation is needed to calculate these derivatives in intermediate positions.

The final system is solved by an iterative procedure which has two different stages. In the first, the program refines the maximum heights and band widths

until convergence is achieved. In the second, these quantities are fixed and the peak positions are refined. Then, the program returns to the first stage and iterations begin. In this way, a particular initial estimate can be improved as far as is possible given the experimental accuracy, the quality of the calculated derivatives and the fit of the band model.

This method provides an objective way of resolving a spectrum into its component bands, because all the information required for starting can be obtained from the experimental contour and its derivatives. The introduction of contour derivatives imposes a larger number of mathematical conditions than is found in the current least-squares methods. These additional conditions efficiently restrain and force the system to generate a physically meaningful solution.

Results and conclusions

Results obtained from simulated and experimental spectra show that the refinement procedure described above converges better and faster to the final result than do the least-squares methods. For comparison, the powerful and extensively used Marquardt method [24] was selected.

The refinement procedure requires some precision in the estimates of peak positions. Its final result does not depend on the initial values of ϵ_0 and b. In contrast, the least-squares method depends strongly on all these values, which causes many problems including divergence, slow convergence, appearance of local minima, etc.

The quality of the starting parameters always affects the computation time. For the least-squares method, this time depends on two factors: the number of measured points and the number of bands present. In the proposed refinement method, the computation time depends only on the second factor. Accordingly, a considerable amount of computation time is saved in cases of intense band overlapping where the spectra must be represented by a large number of points.

The average computation times by iteration or cycle with the Marquardt least-squares method and by the refinement method proposed here were compared. The time required for the Marquardt method was approximately six times longer when the two methods were applied to one spectrum represented by 426 data points and consisting of 9 severely overlapped bands.

Figure 1 shows the refinement evolution in a simulated spectrum formed by a total of five overlapped symmetric Lorentzian functions ($\epsilon = \epsilon_0 \{1 + [(\lambda_0/\lambda)-1]^2/b^2\}$) and represented by 150 data points. This figure shows how the same result can be achieved from three different sets of parameters. The starting set and the two first refinement cycles are represented for the three cases. In case number 3, the Marquardt method cannot reach the minimum without previous refinement.

A computer program (ANALBANDAS) which includes the above-mentioned procedures for spectral decomposition has been developed, and is available from the authors on request. The program allows automatic band

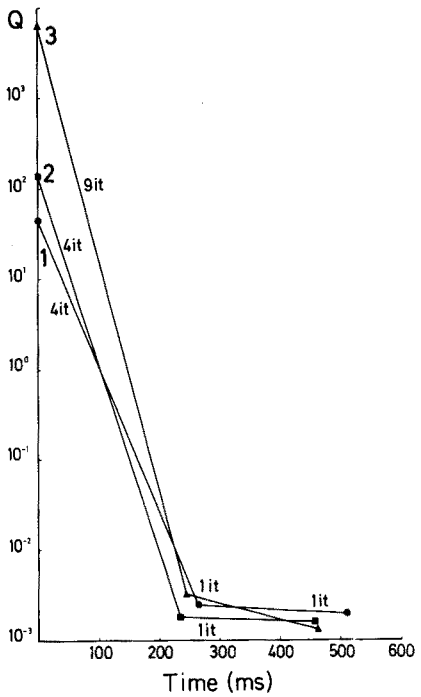


Fig. 1. Quadratic deviation sum vs. computation time. The number of iterations in each refinement cycle is indicated.

detection by analysis of the second derivative of the experimental contour. It will be described in detail in a later paper.

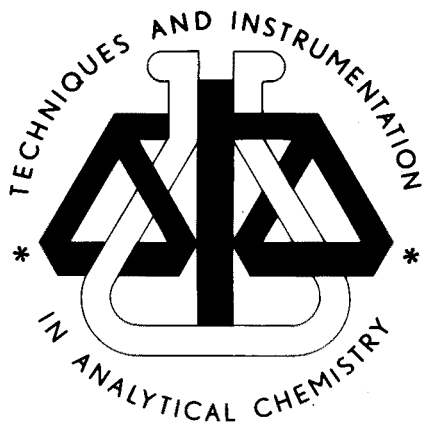
The authors thank Dr. L. Pueyo for many valuable suggestions.

REFERENCES

1 J. S. Challice, Spectrochim. Acta, 20 (1964) 765.
2 J. T. Bell and R. E. Biggers, J. Mol. Spectrosc., 18 (1965) 247.
3 R. D. B. Fraser and E. Suzuki, Anal. Chem., 38 (1966) 1770.
4 J. Pitha and R. N. Jones, Can. J. Chem., 44 (1966) 3031.
5 J. T. Bell and R. E. Biggers, J. Mol. Spectrosc., 22 (1967) 262.
6 G. Derflinger and H. Lischka, Monatsh. Chem., 99 (1968) 1851.
7 J. Lang and R. Müller, Comput. Phys. Commun., 2 (1971) 79.
8 L. M. Schwartz, Anal. Chem., 43 (1971) 1336.
9 J. Carlier, Radiochem. Radioanal. Lett., 10 (1972) 19.
10 K. V. Schwartz, Comput. Prog. in Biomed., 2 (1972) 257.
11 B. Klabuhn, D. Spindler and H. Goetz, Spectrochim. Acta, Part A, 29 (1973) 1283.
12 B. E. Barker, M. F. Fox, A. Walton and E. Hayon, J. Chem. Soc. Faraday Trans. 1, 72 (1976) 344.
13 H. Lischka and G. Derflinger, Monatsh. Chem., 99 (1968) 2450.
14 B. G. M. Vandeginste and L. De Galan, Anal. Chem., 47 (1975) 2124.

15 P. Gans, Coord. Chem. Rev., 19 (1976) 99.
16 C. L. Lawson and R. J. Hanson, Solving Least-squares Problems, Prentice-Hall, New Jersey, 1974, p. 7.
17 J. S. Challice and G. M. Clarke, Spectrochim. Acta, 22 (1966) 63.
18 J. R. Morrey, Anal. Chem., 40 (1968) 905.
19 J. W. Perram, J. Chem. Phys., 49 (1968) 4245.
20 E. Bernal, J. Chem. Phys., 55 (1971) 2538.
21 H. S. Gold, C. E. Rechsteiner and R. P. Buck, Anal. Chem., 48 (1976) 1540.
22 A. Schmitt, Quim. Ind., 23 (1977) 471.
23 A. Savitzky and M. J. E. Golay, Anal. Chem., 36 (1964) 1627.
24 P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969, p. 236.

valuation and ptimization of Laboratory lethods and nalytical Procedures



urvey of Statistical and thematical Techniques

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

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

chniques and Instrumentation in Analytical Chemistry - Volume 1

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

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

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

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

. 1978 xvi + 596 pages US $57.75/Dfl. 130.00 ISBN 0-444-41743-5

ELSEVIER

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

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

52 Vanderbilt Ave
New York, N.Y. 10017

7113

VOL. 103, No. 3 ANALYTICA CHIMICA ACTA September 15, 1978

(Computer Techniques and Optimization, Vol. 2, No. 3)

CONTENTS

Printed in The Netherlands