Vol. 112 No. 4 December 15, 1979 ISSN 0378-4304
(Computer Techniques and Optimization, Vol. 3 No. 4)
Completing Vol. 112

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 1979 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	104/1	104/2	105	106/1	106/2	107	108	109/1	109/2	110/1	110/2	111
Section on Computer Techniques and Optimization			112/1			112/2			112/3			112/

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

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 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 9 volumes in 1979. The subscription for 1979 (Vols. 104–112) is Dfl. 1179.00 plus Dfl. 135.00 (postage) (total approx. U.S. $641.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 112) is Dfl. 131.00 plus Dfl. 15.00 (postage) (total approx. U.S. $71.00). 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. 1–103) except Vols. 23 and 28 are available at Dfl. 150.00 (U.S. $73.20), plus Dfl. 10.00 (U.S. $4.90) postage and handling, per volume.

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

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

Analytica Chimica Acta, 112 (1979) 321—340
Computer Techniques and Optimization

PRINCIPLES AND APPLICATIONS OF A RESEARCH-ORIENTED GAS CHROMATOGRAPHY—MASS SPECTROMETRY DATA SYSTEM

J. E. CAMPANA*, T. H. RISBY** and P. C. JURS

Department of Chemistry, The Pennsylvania State University, University Park, PA 16802 (U.S.A.)

(Received 30th May 1979)

SUMMARY

A research-oriented gas chromatography—mass spectrometry data system for a quadrupole mass spectrometer has been developed based on a centrally located departmental computer facility. An overview of the hardware and software system is presented, emphasizing the important aspects of on-line computer data acquisition and control and the design philosophy used in the development of the system. The application of the system is demonstrated by the g.c.—m.s. analysis of a mixture of four transition metal β-diketonates (Al, Cr, Rh, and Ru tris-1,1,1-trifluoro-pentane-2,4-dionate). This analysis involved vacuum gas chromatography with a support-coated open tubular column and detection of the eluent by chemical ionization mass spectrometry. The results demonstrate the data system capabilities and indicate the utility of the combined methodologies.

The advantages of on-line computer data acquisition and control in the laboratory cannot be overemphasized. Three prominent advantages are as follows: (1) the elimination of the scientist as the link between the experiment and the off-line computer configuration; (2) elimination of manual interpretations of data and the errors associated with such data interpretations; and (3) the direct computer control of the experiment including closed-loop control, automation, and optimization. Many reviews have appeared in the chemical literature on this subject illustrating its importance and breadth [1—6].

Computer data acquisition and control is most advantageous in experimental applications that generate large data sets quickly. It is not surprising that the first chemical applications of computer-compatible electronic recording of data were in the area of mass spectrometry in the middle-1960s. The need for fast electronic recording of data was realized by those working in the field of fast-scan high-resolution mass spectrometry. These workers

*Author for correspondence. Present address: Department of Pharmacology and Experimental Therapeutics, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205, U.S.A.
**Present address: Department of Environmental Health Sciences, The Johns Hopkins University, School of Hygiene and Public Health, 615 North Wolfe Street, Baltimore, MD 21205, U.S.A.

were interested in accurate mass measurements for the determination of elemental compositions of organic compounds, which were shown to be feasible in 1959 by Beynon [7]. The processing of g.c.—m.s. data has proved the greatest demand for the on-line recording of data, because of the thousands of spectra that can be generated in a single sample analysis.

Off-line analog recording of fast-scan high-resolution mass spectral data on magnetic tape for later computer processing was first demonstrated in 1964 [8—11]. The utility of fast recording of data was shown, but limitations arose from the available commercial recording systems. The Biemann group demonstrated the more direct approach of recording mass spectral data on a digital-tape system for later off-line computer processing [12]. The first real-time data acquisition system for a quadrupole mass spectrometer was reported by Reynolds et al. [13] and for high-resolution mass spectrometry by Burlingame et al. [14].

Since then the growth of m.s. and g.c.—m.s. data systems has been tremendous. Various mass spectral processing software systems (including archival documentation and retrieval) [15—31], g.c.—m.s. data systems [32—44] and their applications [45—47], commercial systems [48, 49], and general reviews [50, 51] have been published. A book emphasizing the software aspects has been published [52]. This is not an exhaustive list, but it demonstrates the diversity of research in m.s. data systems.

Recently, microprocessor-based systems have been reported [53, 54]; however, the minicomputer-based systems still provide the flexibility and simplicity needed for designing a research-oriented data acquisition and control system in the experimental laboratory.

The idea of the coupling of computers to instrumentation has led to the development of general-purpose data acquisition and control systems for various experiments [5]. Such systems have been developed in industrial [55], governmental [56, 57], and academic laboratories [58]. The system developed at the Chemistry Department of Pennsylvania State University is described below.

COMPUTER FACILITIES

A computer system was purchased in 1973 for the purpose of developing a department-wide general-purpose data acquisition and control system. It was found that development of a g.c.—m.s. data system using the department's facilities had many merits over purchasing any of the systems commercially available at that time which were expensive ($20,000—$125,000) [48]; also processing and data acquisition and control software were limited to whatever was offered by the vendor. Recently, commercial systems with capabilities to accommodate user-developed processing software have become available.

"In-house" development of data acquisition and control systems is often hampered by development [13, 48] and limitations of the software and

hardware specialist's support during various phases of system evolution. The system described here was developed in about three man-years and provides adequate software for the present g.c.—m.s. research demands.

This Department's Lab Box system consists of a centrally located MODCOMP II/25 computer system (Modular Computer Systems, Inc., Ft. Lauderdale, FL 33309) and a logic interface named "MAX Central". "MAX Central" determines which Lab Box is to be put on-line (enabled), when a request signal is generated by an experimenter. Lab Boxes (interfacing hardware) are located in several laboratories in the department. Only one Lab Box is permitted to be on-line at a time although several experiments can share one Lab Box. Differential digital-data transmission is used for all communication within the system. Figure 1 presents a block diagram describing the general Lab Box system.

The MODCOMP II/25 computer operates under the MODCOMP MAX II/III Disc Operating System. A multiprogramming environment, which supports multi-task handling and batch processing, has enabled batch processing to continue during data acquisition and control and has allowed data acquisition and control features such as a dynamic spectra display during collection of g.c.—m.s. data.

The criteria used in choosing the Lab Box hardware configuration were (a) to design a general-purpose digital interface capable of high data acquis-

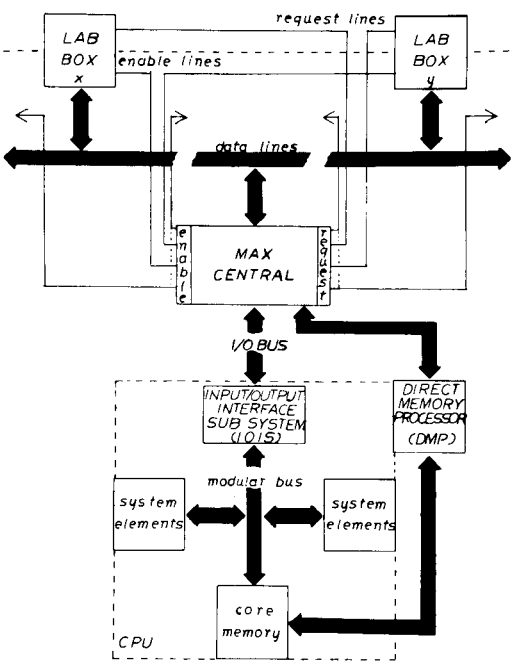


Fig. 1. The MODCOMP II/25 Computer System and the Lab Box System.

ition rates (ca. 40 kHz), and (b) to use minimal software supervision for data acquisition and control. The first specification meant choosing the hardware configuration judiciously, while the second meant making use of external processors, computer hardware interrupts, and hardwiring controls as much as physically and economically possible. The result was a minimization of timing and collection error and an optimization of computation time.

Each Lab Box is a self-contained laboratory computer interface. It contains a positive 5-V power supply, positive and negative 15-V power supplies, and also optical isolation circuitry for computer protection and for the elimination of ground loops. The Lab Box contains a successive approximation analog-to-digital converter (ADC) with a 24-μs conversion rate (Teledyne Philbrick 4129QZ) with a high-speed sample-and-hold amplifier (SHA) with a 1-μs acquisition time and 300-ns settling time (Teledyne Philbrick 4853) connected to the input of the ADC [59]. This circuitry converts the electrometer signal to a digital format during data acquisition. The sample/hold pin of the SHA is connected to the status pin of the ADC. Therefore, whenever the ADC is converting a voltage, the SHA is holding that same electrometer voltage, and when the ADC is idle the SHA is tracking the electrometer voltage. This standard ADC configuration allows the digital-to-analog converter (DAC) control voltage (mass spectrometer control signal) to be changed during ADC conversion without affecting the integrity of the ADC conversion.

The SHA inverts the electrometer signal so that it was necessary to precede it with an operational amplifier (Teledyne Philbrick 1027) in an unity-gain inverting configuration to maintain the polarity of the signal. This operational amplifier was also configured as a low pass filter (cutoff above 30 kHz) to eliminate high-frequency noise pickup from the radio-frequency (r.f.) voltages generated by the quadrupole controller system. This operational amplifier could also be used as an inverting integrator (with low pass filtering) to integrate the electrometer signal for a chosen integration interval; however, this feature is not required by current research demands.

The quadrupole mass spectrometer (Scientific Research Instruments Corporation BIOSPECT System) gives reasonable resolution up to about m/z 1200. In order to make this entire m/z 1200 region accessible during data acquisition, a DAC with a wide dynamic range was employed. The quadrupole controller (Extranuclear Laboratories, Model QPS) could be modified so that any m/z range would be accessible with a 0—10-V control voltage. A 16-bit DAC with 30-μs conversion time full scale (Datel Systems Inc. DAC-169) was employed because of its economy. The most significant 14 bits of this converter were used and provided the dynamic range needed to acquire data over the m/z 1200 range. The DAC conversion time would typically be less than 30 μs because the typical control-voltage step size is a fraction of the full scale conversion. This insures that data acquisition rates of about 40 kHz can be supported with this hardware configuration. Additionally, the DAC digital increment is under operator control (i.e., the

demands of the particular experimental situation) so that high data acquisition rates are still possible. The timing diagram corresponding to the maximum acquisition rate of this hardware configuration is shown in Fig. 2. All the hardware described is capable of converting or following voltages to at least 0.01% full scale (full scale is 10 V).

Two multiplexed 10-bit DACs with 5-μs conversion times (Teledyne Philbrick 4023) are used to control an analog recorder for laboratory display of data. One channel may be used for real-time display of spectra on the m.s. oscilloscopic-display system. A graphics terminal (Tektronix 4006-1) provides communication and graphics display capabilities.

THE DATA ACQUISITION HARDWARE CONFIGURATION

This paper describes a data acquisition and control system designed for a quadrupole mass spectrometer. Similar systems have been reported by other workers [13, 60, 61]. This system can control other types of mass analyzers by modification of the spectrometer control hardware and software. In the off-line mass spectrometer configuration, the quadrupole controller permits the detection of ions in any mass range within three to about m/z 1600 with a 0—10-V sawtooth control voltage. This quadrupole controller supplies two summed combination r.f. and d.c. voltages of equal magnitude but opposite signs to the quadrupole mass filter that results in the detection of a particular m/z ion. The detector current is converted by the electrometer circuit to a voltage. This voltage is amplified and displayed on an oscilloscope as a function of the control voltage.

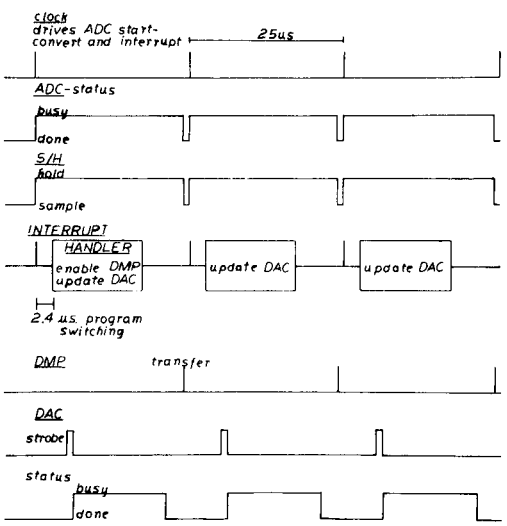


Fig. 2. Timing diagram for the Lab Box hardware at the maximum data acquisition rate of 40 kHz.

The on-line mass spectrometer—computer configuration is shown in Fig. 3. The design goal of using minimal software supervision for data acquisition and control is illustrated and discussed here. The quadrupole controller voltage is supplied by the DAC, while the electrometer voltage is followed and converted to a digital format by the SHA—ADC circuitry. The low pass filter, unity gain, inverting operational amplifier is not shown, but resides between the electrometer and SHA.

The operation of the SHA—ADC circuitry has been discussed above. The status or end-of-conversion (EOC) pin of the ADC is connected directly to the transfer-control pin of the MODCOMP direct memory processor (DMP; a direct memory access device). When an ADC conversion is complete, an EOC signal is generated and triggers the DMP to make a datum transfer automatically into core memory without the execution of any software. The DMP operates on a cycle-stealing basis (330-kHz transfer rate) and is programmable from software. A clock (interval timer, 1-μs interval, MODCOMP model 4701-1), which is a 16-bit counter that can be set under program control, provides a clock pulse when countdown is complete. The clock output signal is hardwired to both a high-priority interrupt (level three) and the "start conversion" pin of the ADC. An interrupt-handler program updates the DAC voltage value, transfers the previous DMP core

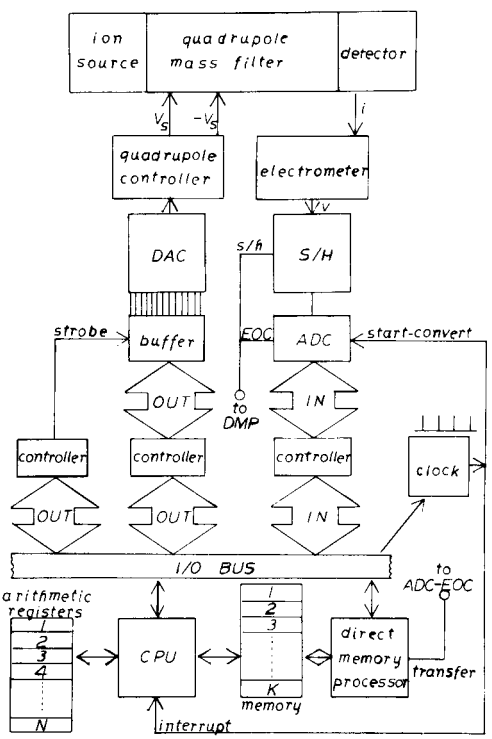


Fig. 3. On-line computer—mass spectrometer configuration.

memory-transferred datum to a predetermined array location, and performs other elementary manipulations depending on the nature of the data acquisition mode (ensemble averaging, g.c. data collection, etc).

Data acquisition and control proceed as follows.

1. A starting DAC voltage is initialized from software.
2. The clock is programmed at a rate corresponding to the chosen data acquisition rate.
3. The computer hardware interrupt level connected to the clock is enabled.
4. When the next clock pulse (interval timer) is generated, the ADC begins converting the SHA voltage, which the SHA was previously following, but is now holding as a result of the ADC status signal switching it into the hold state. Simultaneously, a computer hardware interrupt is generated.
5. An interrupt-handler program is entered, which updates the DAC voltage, resulting in a new electrometer voltage signal (a new m/z ion for detection). Program control is returned to the computer system. Batch processing or other task handling may proceed following this intermittent period.
6. When ADC conversion is complete, a data transfer into core memory is made automatically by the DMP and simultaneously the SHA begins to follow the current electrometer voltage.
7. Steps 4—6 are repeated during data acquisition and control.

The advantage of this procedure is that the only software supervision during acquisition and control is updating the DAC voltage. The time between the end of this software process and the next hardware interrupt is available for elementary manipulations of data within the interrupt-handler program and/or for the multiprogramming environment of the computer system to initiate the execution of other multi-tasks or batch processing.

Typical data acquisition rates (clock pulse rates) are of the order of 1—2 kHz which corresponds to one interrupt being generated about every 500 μs. The interrupt-handler program executes in about 50 μs, which means that 90% of the time during data acquisition and control is available for the multiprogramming environment of the computer system.

THE CAD SOFTWARE SYSTEM

The software package, named CAD, was developed from a primitive software system designed to collect, ensemble average, and display mass spectrometric data.

The criteria used for software development were to preserve the multiprogramming environment of the computer system and to accommodate simple and fast program modification and additions in a modular software system. These requirements meant a system of overlays; the net result was an apparent reduction in program size, a fast and simplified editing procedure and modularity.

The concept of using overlays in data acquisition and control is not new [14, 33]. It has been described [14] as follows: "overlay" means that various programs can be executed sequentially in the same area of memory while simultaneously using undistributed other areas of memory for communication linkage and/or data storage. The impetus for using overlays here was the preservation of the multiprogramming environment of the small computer system by the apparent reduction of program size and making the software system modular. The utilization of the program design philosophies and methods of top-down program design and modular programming [62] in this system has contributed to the success and rapid development of the software system.

The main program, CAD, contains COMMON blocks for the spectral data and the relevant data acquisition information (spectral range(s), title, etc.) and the frequently used system and user library subroutines. A communications overlay named COM is then entered, which allows the experimenter to choose various functions, called modules, to be executed. The overlay system is shown in Fig. 4.

The mass spectrometer computerization involved the following basic types of processing: (1) data acquisition under computer control; (2) data manipulation and display; (3) storage and retrieval of data. The basic modules allow for instrument calibration, data acquisition and control, data manipulation and display, storage and retrieval of data, diagnostic, and utility programs.

The communications module serves as a link between the researcher and the other various modules. A given module may be associated with more than one processing operation. The intricate relationship between the communications module and the other modules is detailed in Fig. 5. Overlay COLLECT is a module for data acquisition and control and is called by the "EX" (experiment) directive. Overlay PLOT is the module for data manipulation and display. Processing of a single ensemble-averaged spectrum is accomplished by entering the PLOT module with the "PL" directive. Notice

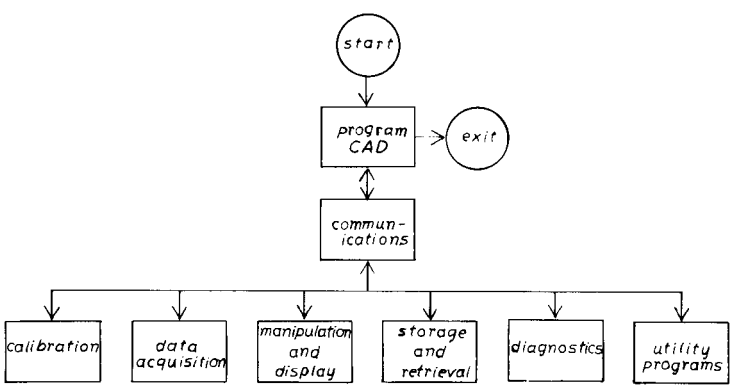


Fig. 4. The CAD overlay system.

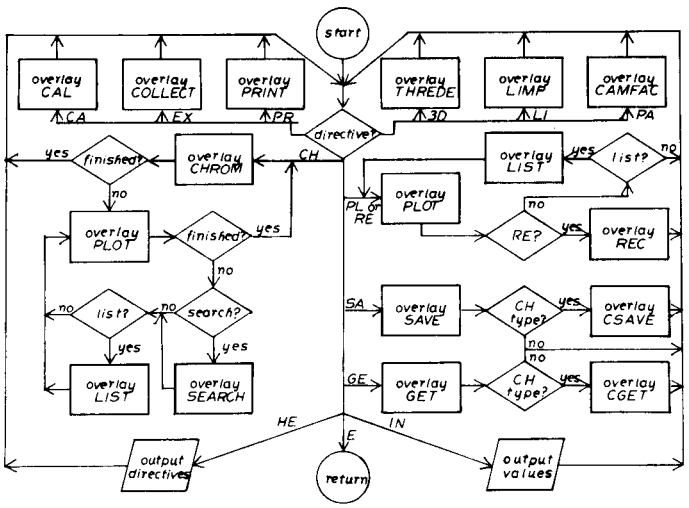


Fig. 5. Flowchart of the communication module.

that for the processing of continuous repetitive measurement of spectra (CRMS) data, a "CH" (chromatographic data) directive is entered and overlay CHROM is executed; then, overlay PLOT is executed for the plotting and the reduction of the CRMS data.

All modules share a common set of directives for simplicity although each module may have additional unique directives. A brief summary of each module and the important features associated with each is presented below. The organization is according to module function.

Mass spectrometer — computer calibration

A calibration module named CAL allows the researcher to calibrate the voltages generated by the quadrupole controller to the DAC control voltages. This method [13] enables the m/z to be a function of DAC control voltage. All subsequent m/z assignments are determined by this relationship. Typically, one of the chemical ionization reactant ions such as $[CH_5]^+$ (m/z 17), $[H_3O]^+$ (m/z 19) or $[C_2H_5]^+$ (m/z 29) is used as a low-mass calibrant while protonated methyl stearate (m/z 299), $[Cr(tfa)_3H]^+$ (m/z 512), or an appropriate lanthanide β-diketonate is used for high-mass calibration. This method does not account for the mass defect of the calibrant species nor does it rely on the stability of the quadrupole-controller electronics. However, it is simple, adequate, and reproducible for mass determination with this particular quadrupole mass spectrometer. The calibration module allows the researcher to choose the number of data to be collected per unit m/z region.

Data acquisition and control

COLLECT is the name of the data acquisition and control module. This module allows the operator to enter the data title, file name, and the *m/z* ranges for data acquisition along with the integration time per *m/z* region. Eight different ranges can be accommodated; integration time actually refers to the period of the acquisition rate (as discussed previously, the hardware does not perform integration in its present configuration). The type of experiment to be done is then entered. Some of the types of data acquisition modes will be described briefly.

Ensemble-averaging mode. 1—32,767 spectra can be averaged. A selected ion monitoring option is present.

CRMS mode. The time duration of data collection and the time interval between each scan is specified. Spectra are collected as a function of time and each individual spectrum is saved. A selected ion monitoring option is present.

Stop-scan mode. Individual spectra are collected and saved as a function of some experimental parameter such as temperature, pressure, lens voltage, etc. After each spectral scan, the experiment is stopped automatically. The researcher can then alter an instrumental parameter and enter the value of the new experimental parameter for storage with its corresponding spectrum, before initiating another spectral scan. Spectra are collected and stored as CRMS data. A selected ion monitoring option is present.

Tune mode. This mode allows consecutive and repetitive scanning of a low *m/z* and a high *m/z* region (or several *m/z* regions) with simultaneous display of the spectra on an oscilloscopic system (real-time display). This feature was reported on the first high-resolution mass spectrometry data system [14] and was considered to be one of the most important parts of the system. The present research is in agreement with these views because of the capability to view several spectral regions in real time and because of the severe mass discrimination inherent in quadrupole mass filters. This feature allows a wide *m/z* region to be tuned for unit resolution and intensities of the accepted isotopic abundances in a short period of time. The current movement toward standard-reference compounds to calibrate mass spectrometers for ion abundance measurements [63, 64] indicates the importance of such a real-time display mode in future systems.

Other features in this module are options that allow a time delay between scans to be specified, the output of some physical parameters regarding data acquisition, and diagnostic collection modes.

There are three other tasks that are associated with COLLECT. These are three interrupt-handler programs; one each for the ensemble-averaging mode, the CRMS mode, and the tune mode. Another task named CHUPD for chromatography update allows dynamic display of the last spectrum collected and/or an updated total ion current profile during the data acquisition of CRMS data. This task runs independently of the CAD system in the multiprogramming environment. The interrupt-handler programs run at the

highest practical hardware interrupt level that is available on the system. The CHUPD module runs at a software priority level between the interrupt handlers and CAD programs, so that if control commands are not being made, a dynamic display of data is presented if desired. The advantages of the ensemble-averaging mode [65, 66] and the selected ion monitoring option [67, 68] present in this module have been discussed elsewhere.

Data manipulation and display

Overlay PLOT. This module fills arrays with ensemble-averaged data and performs elementary operations on all data to be plotted graphically or listed. The elementary processing operations include the capabilities to threshold, normalize, subtract background, and smooth by least-squares polynomial with either bar or continuous spectral display. Data manipulation and display of several types exist including selected ion current profiles, total ion current profiles, and extracted ion current profiles [69].

Overlay LIST. This module lists spectral data or CRMS data on the graphics unit or line printer. Lists of mass, intensity and percent total ion current and scan number, intensity, percent total ion current are available.

Overlay CHROM. CHROM is named for chromatographic data (CRMS data) and processes chromatographic and non-chromatographic CRMS data. This module puts CRMS data in a format that can be processed by overlay PLOT. Ion current (total, extracted and selected) profiles and extracted individual spectra are put in an acceptable format by the CHROM module for the PLOT module. Additionally, any individual spectrum from CRMS data can be transferred to the background file for later background subtraction in the PLOT processor or for archival storage of that individual spectrum with the SAVE module.

Overlay REC. This module allows bar spectra to be plotted on an analog *x-y* recorder in the laboratory.

Overlay SEARCH. This module simply allows for the identification of a particular spectrum index number contributing to the ion current at any point along an ion current profile (total, extracted or selected).

Overlay PRINT. This module prints the raw data of a single spectrum on a line printer for diagnostic purposes.

Overlay THREDE. THREDE plots CRMS data in three dimensions. The unique features of this module are variable viewpoint, a hidden-line algorithm, and least-squares polynomial smoothing.

Overlay LIMP. LIMP is named for list intensity maximum of peaks. This module orders the *m/z* peaks by their intensity contribution to the corresponding extracted or total ion current profile. This method has been described by Hites and Biemann [70].

Data storage and retrieval

All modes use temporary-disc storage for data acquisition, manipulation and display. Single spectra are stored archivally on disc, while sets of spectra (CRMS data) are stored on magnetic tape.

Overlay SAVE. This module determines by query whether data which are to be saved are single spectrum or CRMS data. If the latter, then the CSAVE module is entered. The data are saved on a directorized load module file according to the MODCOMP Disc Cataloger Processor format.

Overlay CSAVE. This module saves CRMS data archivally on magnetic tape.

Overlay GET. This module determines by query whether the data to be retrieved are a single spectrum or CRMS data. If the latter, the CGET module is entered. After the three-character spectrum name has been entered, the data are transferred from the directorized archival file to the specified-disc file for processing.

Overlay CGET. This module retrieves CRMS from its archival magnetic-tape storage.

Utility modules

CAMFAC is named for calculating abundances and masses for a chemical compound. This program, when given a molecular formula, calculates the isotopic peak contributions of the parent molecular ion. Because very little fragmentation is observed in chemical ionization mass spectrometry (c.i.m.s.) for particular classes of compounds, this program has been useful as a secondary verification of the characterization and analysis of many organometallic complexes by c.i.m.s. Plots and lists of m/z, relative abundance, and percentage abundance are available. This processor is based on the multinomial distribution.

To summarize, the system of overlays makes addition to the software package easy and has added to the rapid development of the system. The preservation of the multiprogramming environment has led to more efficient computer use and has made possible such system features as dynamic display in the CRMS mode.

The software directives are easy and command errors are detected and reported, so that the researcher can learn interactively with the system. All numeric data entry is as format-free integers for simplicity.

APPLICATION

To demonstrate the application, use, and performance of this g.c.—m.s. data system, a g.c. analysis of a trace solution of a mixture of several transition metal β-diketonates was performed.

The environmental and biological significance of trace and ultra-trace analytical methodologies for metals is well appreciated [71]. Many workers have demonstrated g.c. separations of various mixtures of the β-diketonates.

Attempts to reproduce elutions of published results for various β-diketonates by using mass spectrometric detection have been only partially successful in this laboratory. Failure to obtain these separations could be attributed to the lack of reproducibility of chromatographic columns.

C.i.m.s. is an ideal chromatographic detector for the analysis of some g.c. eluents because of the specificity of mass spectrometry and the sensitivity of the chemical ionization methodology for various classes of compounds. The sensitivity of c.i.m.s. has been demonstrated for the metal β-diketonates [72, 73]. Many metal β-diketonates that cannot be separated chromatographically can possibly be separated in mass by the m.s. detector.

In order to eliminate problems associated with the packed column, such as reproducibility of its packing and the high flow rates needed for column efficiency, a new technique, that of vacuum gas chromatography (v.g.c.) based on open tubular columns, combined with mass spectrometry was employed [74].

Experimental

A Varian Aerograph series 1200 gas chromatograph was interfaced to the Scientific Research Instruments Corporation BIOSPECT mass spectrometer. The Varian instrument was modified by replacing the flame ionization detector with a 3-mm bulkhead union (all fittings used were Swagelok). A 25-cm length of glass-lined stainless steel tube (3-mm o.d.: Supelco, Bellefonte, PA), packed with silanized glass wool (Applied Science, State College, PA) to minimize dead volume, ran directly from the bulkhead union into the mass spectrometer source via a vacuum feedthrough. This interface was heated inside and outside the vacuum housing. The other end of the bulkhead union was coupled to a 1.6/3-mm reducer. A 15-cm length of glass-lined stainless steel tubing (3-mm o.d.) ran from the septum into the oven; it was coupled to a 6/1.6-mm reducer via a 6/3-mm ferrule (Vespel). The latter reducer was modified by drilling a 1.6-mm hole, inserting a length of 1.6-mm stainless steel tubing and silver-soldering the junction. A fine metering valve (Nupro SS-4SG) was connected to the tubing to control the split ratio. Layered septums (760-03, Hamilton) were used and sustained a moderate lifetime under v.g.c. conditions. A 6.5-m support-coated open tubular (SCOT) column (0.64-mm i.d. × 0.97-mm o.d., 0.25% SE-30, silica gel support) was used for separation of the β-diketonates. The injection port, column, interfaces, and mass spectrometer source temperatures were maintained at 140, 90, 190, and 190°C, respectively, for the final analysis. The mass spectrometer analyzer pressure was 3.5×10^{-5} mm Hg. The column outlet pressure was measured by placing a 3-mm union T-joint between the 3/1.6-mm reducing union and the 3-mm bulkhead union for manometer pressure measurement. The column-outlet pressure was found to be 400-mm Hg. The mass spectrometer parameters were optimized in the mass ranges of interest by using the CAD system's tune mode. The electron emission current was regulated to 0.2 mA. Methane (99.99% Airco Products) was used as the carrier and the chemical ionization reactant gas.

Rhodium tris-trifluoroacetylacetonate [$Rh(tfa)_3$] was prepared by the method of Fay and Piper [75, 76], $Cu(tfa)_2$ was purchased (Pierce Chemical, Co., Rockford, IL) and other metal chelates were prepared by Prescott [77].

Results and discussions

Solutions of $Al(tfa)_3$, $Co(tfa)_3$, $Cr(tfa)_3$, $Cu(tfa)_2$, $Fe(tfa)_3$, $Ni(tfa)_2$, $Rh(tfa)_3$, $Ru(tfa)_3$, $VO(tfa)_2$ in toluene (100 ppm) were subjected to g.c.—m.s. analysis under various chromatographic conditions. Only the complexes of Al, Cr, Rh, and Ru had protonated-parent ions which were detected by c.i.m.s. A solution containing these four compounds (100 ppm) in toluene was subjected to chromatography. The CAD system monitored only the protonated-parent mass spectral regions of the four compounds (*m/z* 484—492, 509—518, and 555—569) in the CRMS mode of data collection. The selected ion-monitoring CRMS mode of data collection could have been used to monitor only the protonated parent ion of the four species. Data were collected for 8 min beginning after the emission current and analyzer pressure had settled to their initial values after injection of the solution (0.1-μl) into the gas chromatograph. Eleven datum points were collected per *m/z* region at a rate of 0.314 kHz which corresponds to acquiring one spectrum (over the three ranges) in 1.1 s followed by a 0.65-s delay between scans. A total ion current profile of the raw data is shown in Fig. 6. The *x*-axis represents scan number which is proportional to time; 274 scans were made during the 8-min period of data acquisition. This total ion current profile is analogous to a gas chromatogram and is essentially the same as would be obtained with a typical chromatographic detector. Figure 7(a—c) shows extracted ion current profiles of the *m/z* 484—492, 509—518, and 555—569 regions, respectively. This illustrates how each peak in the profile (Fig. 6) can be assigned to a particular *m/z* range by plotting only the ion current in a chosen mass range as a function of spectrum number (time). The shoulder on the peak corresponding to the $[Cr(tfa)_3H]^+$ (Fig. 7b) demonstrates the chromatographic resolution on the *cis* and *trans* geometrical isomers; the shoulder corresponds to the less abundant *trans* isomer. A plot of the spectrum corresponding to the maxima in each of the three extracted ion current profiles will identify the eluting components by their mass spectra. The mass spectra of these are shown in Fig. 8(a—d), and represent $Al(tfa)_3$, $Cr(tfa)_3$, $Rh(tfa)_3$, and $Ru(tfa)_3$, respectively. All the above plots are background-corrected and normalized to the most intense peak, and the spectral plots have been smoothed by a five-point least-squares polynomial algorithm with three smoothing repetitions.

The three-dimensional plotting capability of the system is illustrated by the following plots. The raw data of the *m/z* 555—569 region — spectrum numbers 190—274 corresponding to the elution of $Rh(tfa)_3$ and $Ru(tfa)_3$ — are plotted by the THREDE module (Fig. 9A). Figure 9B shows a plot of the same data rotated 90°, with background correction, a 1% threshold, and a five-point least-squares polynomial smooth with three smoothing repetitions.

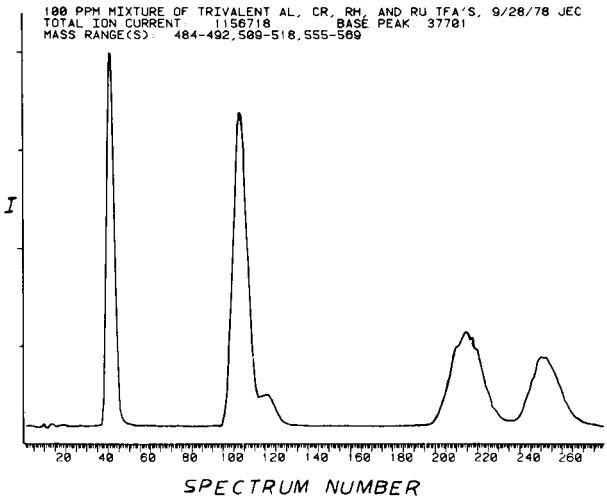


Fig. 6. Total ion current profile of $Al(tfa)_3$, $Cr(tfa)_3$, $Rh(tfa)_3$, and $Ru(tfa)_3$ metal chelate mass spectral data. The 274 spectra were collected over a period of approximately 8 min.

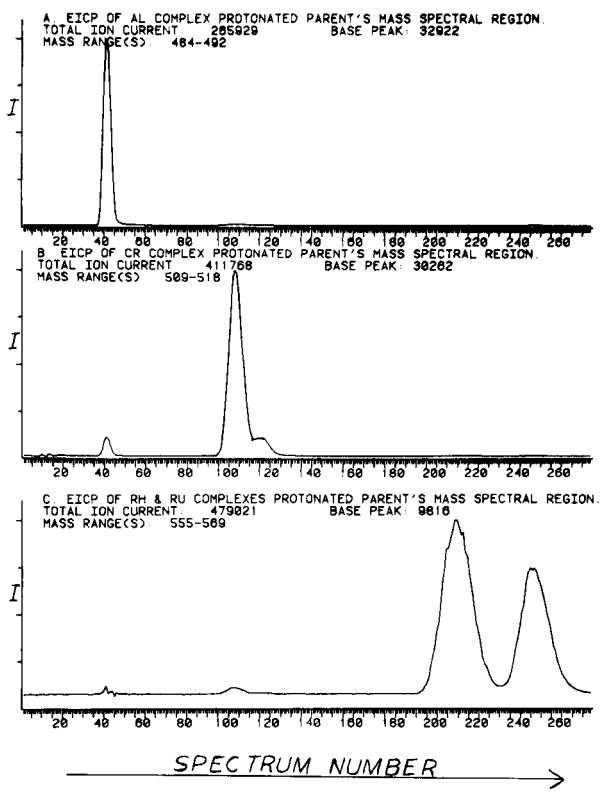


Fig. 7. Extracted ion current profile of the mass spectral regions of (A) $[Al(tfa)_3H]^+$; (B) $[Cr(tfa)_3H]^+$; (C) $[Rh(tfa)_3H]^+$ and $[Ru(tfa)_3H]^+$.

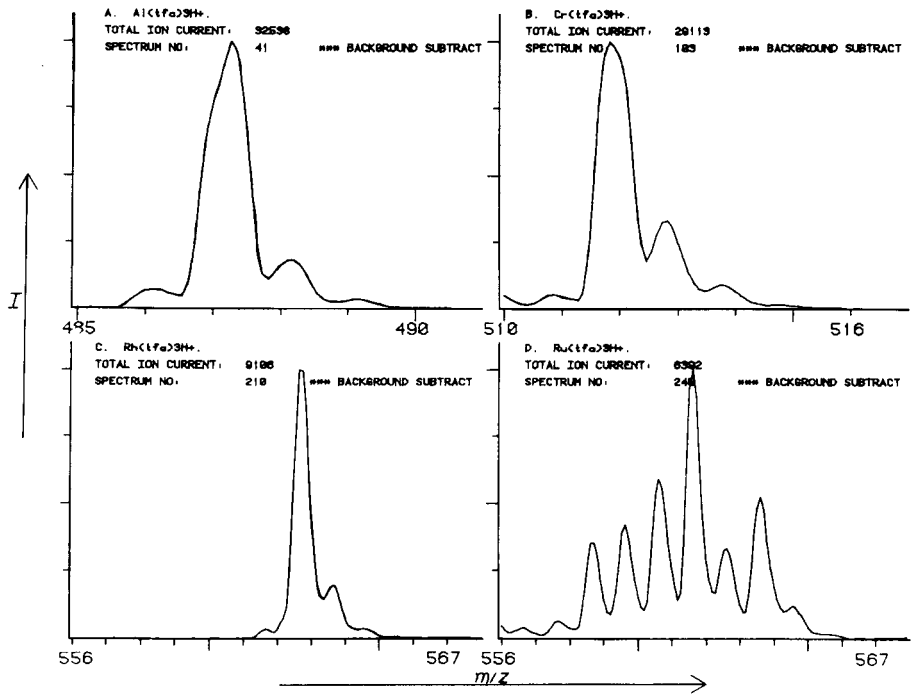


Fig. 8. Chemical ionization mass spectra of (A) $Al(tfa)_3$; (B) $Cr(tfa)_3$; (C) $Rh(tfa)_3$; (D) $Ru(tfa)_3$ in the protonated parent MASS regions.

One can imagine from these data the distinct advantage of the rotational featuré, hidden-line algorithm, background correction and threshold options of this module for analyzing very complex g.c.—m.s. data. Note that the rhodium isotopic peak at m/z 564 which was hidden in Fig. 9A is apparent in the rotated plot. This rotational feature should be of advantage for analysis of complex samples.

These studies show that the v.g.c. method with a SCOT column is very compatible with the c.i.m.s. methodology. The failure to elute several of the metal chelates can be attributed to several possibilities: the wrong column used for the analysis, failure to investigate exhaustively the chromatographic conditions, degradation of the glass-lined stainless steel interface, and failure to mass-analyze for decomposition products. Past experience has indicated that glass-lined stainless steel sometimes degrades after extended use. A more reliable interface could be made from capillary-bore quartz or possibly the capillary column could be directly linked to the m.s. source. The frequency response of the Extranuclear Laboratories (ENL) 031-5 electrometer limited the data acquisition to about 0.3 kHz without severe loss of resolution. This low rate of data acquisition makes the examination of large spectral ranges for the monitoring of decomposition products impossible. This problem will be rectified in the future by replacement of the standard ENL operational amplifiers with faster versions so that acquisition rates approaching 1000

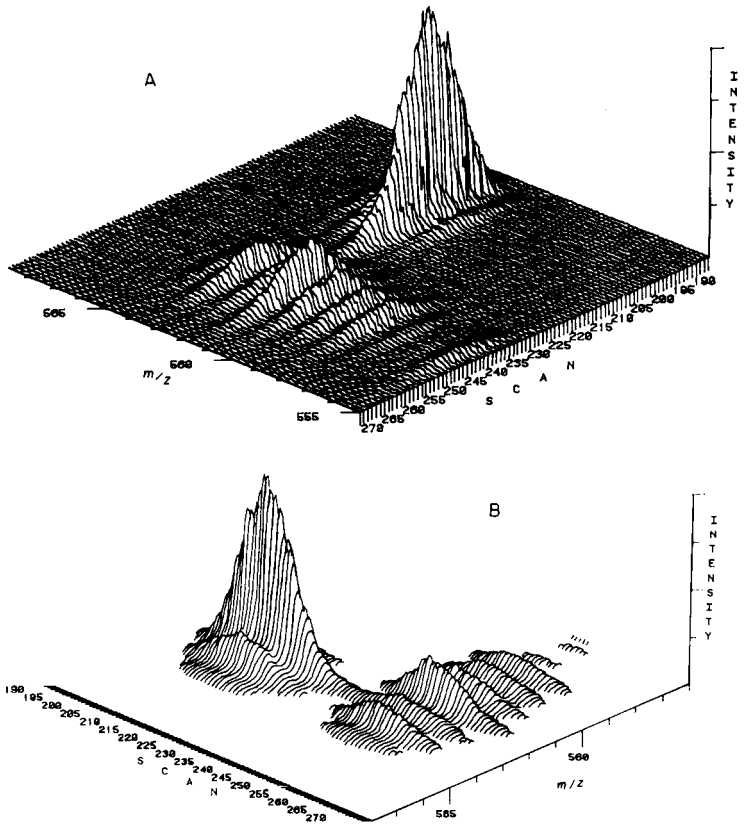


Fig. 9. Three-dimensional plots of $Rh(tfa)_3$ and $Ru(tfa)_3$. (A) Raw data; (B) massaged data.

integral m/z units per second can be obtained while operating at 10^8 electrometer gain. It is apparent that monitoring some of the fragment ions or decomposition products of the other β-diketonates by mass spectrometry might prove useful in the g.c.—m.s. analysis of these complexes.

CONCLUSIONS

The research-oriented g.c.—m.s. data system for a quadrupole mass spectrometer is very flexible. Hardware and software are provided that meet current research demands while the system can be expanded to accommodate future research demands. The software system is structured in such a way that user software additions and modifications are fast and easy. The system preserves the multiprogramming environment of the computer and executes in 28K of memory, which could be decreased if necessary. An operation manual of the system including hardware and software documentation has been prepared to serve as an instruction guide to users and to guarantee the longevity of the software system [78].

The advantages and power of on-line data acquisition and control of laboratory instrumentation has been well realized by the users of this system through the increased possibility of instrument control and the speed of data presentation, features not present in the off-line instrumentation. The advantage of interpretation of data by increased graphic capabilities that can be present in on-line computer configurations is a distinct advantage that is virtually impossible in off-line experimental configurations.

This work was supported by a grant from the U.S. Environmental Protection Agency (Grant R-803651). We thank E. I. DuPont de Nemours and Co. for donation of a computer graphics system and W. E. Brugger, M. A. Henry, D. E. Klees, A. L. McIntosh, and J. A. Yergey for their contributions in the development of the data system. The MODCOMP II/25 computer system was purchased with the partial financial support of the National Science Foundation.

REFERENCES

1 R. J. Spinrad, Science, 158 (1967) 55.
2 J. W. Frazer, Anal. Chem., 40 (1968) 26A.
3 C. W. Childs, P. S. Hallman and D. D. Perrin, Talanta, 16 (1969) 629.
4 R. Venkataraghavan, R. J. Klimowski and F. W. McLafferty, Acc. Chem. Res., 3 (1970) 158.
5 S. P. Perone, Anal. Chem., 43 (1971) 1288.
6 J. W. Frazer, Acc. Chem. Res., 7 (1974) 141.
7 J. H. Beynon, in J. D. Waldron (Ed.), Advances in Mass Spectrometry, Vol. 1, Pergamon, London, 1959.
8 C. Merritt, Jr., presented at the 3rd Annual Meeting, ASTM Committee E-19, Houston, Texas, October 1964.
9 P. Issenburg, M. L. Bazinet and C. Merritt, Jr., Anal. Chem., 37 (1965) 1074.
10 C. Merritt, Jr., P. Issenburg, M. L. Bazinet, B. N. Green, T. O. Merron and J. G. Murray, Anal. Chem., 37 (1965) 1037.
11 W. J. Murray, B. N. Green and S. R. Lipsky, Anal. Chem., 38 (1966) 1194.
12 R. A. Hites and K. Biemann, Anal. Chem., 39 (1967) 965.
13 W. E. Reynolds, V. A. Bacon, J. C. Bridges, T. E. Coburn, B. Halpern, J. Lederburg, E. C. Levinthal, E. Steed and R. B. Tucker, Anal. Chem., 42 (1970) 1122.
14 A. L. Burlingame, D. H. Smith and R. W. Olsen, Anal. Chem., 40 (1968) 13.
15 J. Braumann, Anal. Chem., 38 (1966) 607.
16 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1466.
17 D. D. Tunnicliff and P. A. Wadsworth, Anal. Chem., 40 (1968) 1826.
18 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 690.
19 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
20 P. C. Jurs, Anal. Chem., 43 (1971) 22.
21 H. S. Hertz, R. A. Hites and K. Biemann, Anal. Chem., 43 (1971) 681.
22 S. L. Grotch, Anal. Chem., 43 (1971) 1362.
23 L. R. Crawford and J. D. Morrison, Anal. Chem., 43 (1971) 1790.
24 M. W. Bell, presented at the 21st Annual Conference on Mass Spectrometry and Allied Topics, San Francisco, May 1973.
25 J. E. Biller and K. Biemann, Anal. Lett., 7 (1974) 515.

26 P. R. Naegelli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
27 S. R. Heller, D. A. Koniver, H. M. Fales and G. W. A. Milne, Anal. Chem., 46 (1974) 947.
28 G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
29 R. G. Dromey, M. J. Stefik, T. C. Rindfleisch and A. M. Duffield, Anal. Chem., 48 (1976) 1368.
30 G. W. A. Milne and S. R. Heller, Am. Lab., (Sept. 1976) 43.
31 D. W. Kuehl, Anal. Chem., 49 (1977) 521.
32 R. A. Hites and K. Biemann, Anal. Chem., 40 (1968) 1217.
33 R. J. Klimowski, R. Venkataraghavan, F. W. McLafferty and E. B. Delany, Org. Mass Spectrom., 4 (1970) 17.
34 R. Knutti and R. E. Buhler, Chemica, 24 (1970) 437.
35 C. R. Langergren and J. J. Stoffels, Int. J. Mass Spectrom. Ion Phys., 3 (1970) 429.
36 C. C. Sweeley, B. D. Ray, W. I. Wood, J. F. Holland and M. I. Krichevsky, Anal. Chem., 42 (1970) 1505.
37 D. H. Smith, R. W. Olsen, F. C. Walls and A. L. Burlingame, Anal. Chem., 43 (1971) 1796.
38 J. W. Frazer, L. R. Carlson, A. M. Krag, M. R. Bertoglio and S. P. Perone, Anal. Chem., 43 (1971) 1479.
39 J. R. Plattner and S. P. Markey, Org. Mass Spectrom., 5 (1971) 463.
40 J. F. Holland, C. C. Sweeley, R. E. Thrush, R. E. Teets and M. A. Bieber, Anal. Chem., 45 (1973) 308.
41 R. M. Hilner and J. W. Taylor, Anal. Chem., 45 (1973) 1031.
42 J. T. Watson, D. R. Pelster, B. J. Sweetman, J. C. Frolich and J. A. Oates, Anal. Chem., 45 (1973) 2071.
43 B. Hedfjall and R. Ryhage, Anal. Chem., 47 (1975) 666.
44 R. A. W. Johnstone, F. A. Mellon and S. D. Ward, Int. J. Mass Spectrom. Ion Phys., 5 (1970) 241.
45 F. W. McLafferty, R. Venkataraghavan, J. E. Coutant and B. G. Giessner, Anal. Chem., 43 (1971) 967.
46 H. Nau and K. Biemann, Anal. Chem., 46 (1974) 426.
47 S. C. Gates, M. J. Smisko, C. L. Ashendel, N. D. Young, J. F. Holland and C. C. Sweeley, Anal. Chem., 50 (1978) 433.
48 F. W. Karasek, Anal. Chem., 44 (1972) 32A.
49 R. S. Gohlke, G. P. Happ, D. P. Maier and D. W. Stewart, Anal. Chem., 44 (1972) 1484.
50 S. D. Ward, in D. H. Williams (Senior Reporter), Mass Spectrometry, Vol. 1, Specialist Periodical Report, The Chemical Society, London, 1970; see also subsequent volumes.
51 A. L. Burlingame and G. A. Johnson, Anal. Chem., 44 (1972) 337R; also subsequent biennial reviews.
52 J. R. Chapman, Computers in Mass Spectrometry, Academic Press, New York, 1978.
53 G. A. Schaeffer and W. Huebsch, presented at the 25th Annual Conference on Mass Spectrometry and Allied Topics, Washington, D.C., May 1977.
54 R. D. Friesen, R. S. Newbury and R. J. Dupzyk, presented at the 25th Annual Conference on Mass Spectrometry and Allied Topics, Washington, D.C., May 1977.
55 G. Lauer and R. A. Osteryoung, Anal. Chem., 39 (1967) 765A.
56 E. Ziegler, D. Henneburg and G. Schomburg, Anal. Chem., 42 (1970) 51A.
57 J. W. Frazer and F. W. Kunz (Eds.), Computerized Laboratory Systems, American Society for Testing and Materials, Publication STP-578, Philadelphia, 1975.
58 L. Ramaley and G. S. Wilson, Anal. Chem., 42 (1970) 606.
59 D. H. Sheingold (Ed.), Analog–Digital Conversion Handbook, Analog Devices, Norwood, Mass., 1972, p. I-21.
60 N. A. Jones, R. D. Friesen and J. W. Pyper, Rev. Sci. Instrum., 41 (1970) 1828.

61 J. Houseman and F. W. Hafner, J. Phys. E, 4 (1971) 46.
62 E. Yourdon, Techniques of Program Structure and Design, Prentice-Hall, New Jersey, 1975.
63 J. W. Eichelberger, L. E. Harris and W. L. Budde, Anal. Chem., 47 (1975) 995.
64 W. L. Budde and J. W. Eichelberger, presented at the 26th Annual Conference on Mass Spectrometry and Allied Topics, St. Louis, May 1978.
65 F. J. Biros, Anal. Chem., 42 (1970) 537.
66 A. L. Burlingame, in K. Ogata and T. Hayakawa (Eds.), Recent Developments in Mass Spectrometry, University Park, Baltimore, MD, 1970.
67 F. C. Falkner, B. J. Sweetman and J. T. Watson, Appl. Spectrosc. Rev., 10 (1975) 51.
68 W. L. Budde and J. W. Eighelberger, J. Chromatogr., 134 (1977) 147.
69 A. L. Burlingame, C. H. L. Shackleton, I. Howe and O. S. Chizhov, Anal. Chem., 50 (1978) 346R.
70 R. A. Hites and K. Biemann, Anal. Chem., 42 (1970) 855.
71 T. H. Risby (Ed.), Ultratrace Metal Analysis in Biological Sciences and Environment, American Chemical Society, Washington, D.C., 1979.
72 S. R. Prescott, J. E. Campana, P. C. Jurs, T. H. Risby and A. L. Yergey, Anal. Chem., 48 (1976) 829.
73 S. R. Prescott, J. E. Campana and T. H. Risby, Anal. Chem., 49 (1977) 1501.
74 F. W. Hatch and M. E. Parrish, Anal. Chem., 50 (1978) 1164.
75 R. C. Fay and T. S. Piper, J. Am. Chem. Soc., 84 (1962) 2303.
76 R. C. Fay and T. S. Piper, J. Am. Chem. Soc., 85 (1963) 500.
77 S. R. Prescott, Ph.D. Dissertation, The Pennsylvania State University, Department of Chemistry, University Park, PA (1979).
78 J. E. Campana, CAD: A Gas-Chromatography—Mass Spectrometer—Computer Data Acquisition System for a Quadrupole Mass Spectrometer, The Center for Air Environment Studies, Publication No. 456-76, The Pennsylvania State University, University Park, PA, 1976.

Analytica Chimica Acta, 112 (1979) 341—360
Computer Techniques and Optimization

CORRELATION CHROMATOGRAPHY AND NOISE. THEORETICAL AND PRACTICAL CONSIDERATIONS ON VARIOUS TYPES OF CORRELATION NOISE

T. T. LUB and H. C. SMIT*

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

(Received 17th April 1979)

SUMMARY

The influences of the most important noise sources on the decrease of the detection limit that can be ultimately reached by correlation chromatography are evaluated. Expressions are derived for the influences of system noise as well as quantization noise that originates from the analog-to-digital conversion, which is an essential operation necessary in a computer-based analytical method. The contributions of both kinds of noise approach zero with increasing correlation time, in contrast to the correlation noise caused by non-stationarity of the system. The latter can be separated from the stationary part of the correlogram, if the non-stationarity is known. This is verified by simulation on a digital computer. The implications for practical correlation chromatography are evaluated.

In correlation chromatography, the impulse response (chromatogram) of a chromatographic system is not determined in the usual way by applying an impulse-shaped excitation at the input and measuring the response at the output. Instead, the input excitation is a random signal, and the system responds with a random output. The cross-correlation function of the input and output is identical with the impulse response if the input satisfies certain conditions. System noise is not correlated with the input; its contribution to the cross-correlogram converges to zero with increasing correlation time. A considerable improvement of the signal-to-noise ratio can be reached in a relatively short time, e.g., one hundred times in a correlation time during which only sixteen normal chromatograms could have been recorded [5]. Applications to gas chromatography as well as high-performance liquid chromatography have been reported [1—5]. The second part of this paper will deal with the effect of system noise on the correlogram.

Another source of noise is the quantization of the output of the chromatographic system. The analog-to-digital conversion introduces an error into each digital sample of the output; the effect of this quantization error is discussed in the third part of this paper.

A third and potentially most dangerous kind of correlation noise is caused by the non-stationary nature of the system. It may be due to variations in in-

put concentration, variations of the flow, etc. It is the only kind of correlation noise that does not converge to zero with increasing correlation time; it will be proved in this paper that the stationary part of the correlogram can be separated from the non-stationary part. Finally, some implications for practical correlation chromatography will be discussed.

THE EFFECT OF SYSTEM NOISE ON THE VARIANCE AND COVARIANCE OF THE COEFFICIENTS OF THE CORRELOGRAM

The input of the system is $x(t)$. By taking samples of $x(t)$ at discrete time intervals Δt, $x(t)$ is transformed to the vector $\mathbf{X}$

$$x(t) \rightarrow \mathbf{X} = (x_{-\infty}, \ldots\ldots, x_{-1}, x_0, x_1, \ldots\ldots, x_{+\infty}) \tag{1}$$

with $x_j = x(j\Delta t)$. The input of the system is considered to be defined at any point in time. The output of the system, $z(t)$, is considered to consist of: (a) $y(t)$, the response of the system to $x(t)$; (b) $n(t)$, the noise generated by the system; and (c) $d(t)$, the noise generated by the quantization (or digitalization) Samples of $z(t)$ are taken at time intervals Δt apart. The samplings of $x(t)$ and $z(t)$ are synchronized.

$$z(t) \rightarrow \mathbf{Z} = (z_0, z_1, z_2, \ldots\ldots, z_{M-2}, z_{M-1}) \tag{2}$$

with $z_k = z(k\Delta t)$, and

$$\mathbf{Z} = \mathbf{Y} + \mathbf{N} + \mathbf{D} \tag{3}$$

The output of the system is measured only in the interval $\{0, (M-1)\Delta t\}$. During this interval, M samples are taken. The cross-correlation function of $\mathbf{X}$ and $\mathbf{Z}$ is

$$R_{xz}(n,M) = \frac{1}{M}\sum_{k=0}^{M-1} x_{k-n}\, z_k = \frac{1}{M}\sum_{k=0}^{M-1} x_{k-n}\,(y_k + n_k + d_k)$$

$$= R_{xy}(n,M) + R_{xn}(n,M) + R_{xd}(n,M) \tag{4}$$

$R_{xy}(n)$ will be discussed later; it contains the analytical information. It is a definite function of n and M if $\mathbf{X}$ satisfies certain conditions. R_{xn} and R_{xd} are stochastic functions of n; because they are not cross-correlated, their contributions to the statistical characterization of R_{xz} can be treated separately.

The power spectrum of $\mathbf{X}$ is

$$\Gamma_{xx}(m) = \sum_{-\infty}^{+\infty} \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} (x_{k-n} - \mu_x)(x_k - \mu_x)\cos(2\pi nm\, K^{-1})$$

$$= \lim_{K\to\infty}\left[\left|\sum_{k=0}^{K-1}(x_k - \mu_x)\exp(-2\pi jkm\, K^{-1})\right|^2 K^{-1}\right] \tag{5}$$

where $\mu_x = E\{x_k\}$, the mean value of $\mathbf{X}$.

$\Gamma_{nn}(m)$ is defined in a similar way.

Equation (5) implies that the power spectrum can be obtained either by direct Fourier transformation, and taking the modulus, or by transformation of the auto-covariance function (Wiener—Khinchine theorem).

Stationarity is a necessary condition. It can be shown (see Appendix) that, if the cross-correlation function is computed by using a record length M for the N record and a length ∞ for the X record, the expected value of a power spectrum coefficient $G'_{xn}(m,M)$ of this cross-correlation function is

$$E\{G'_{xn}(m,M)\} = \frac{1}{M}\Gamma_{xx}(m)\,\Gamma_{nn}(m) \tag{6}$$

provided that X and N are stationary, mutually uncorrelated time series.

The auto-covariance function of a record is the inverse Fourier transform of its power spectrum. This implies that a multiplication of the power spectra of two records corresponds to a convolution of their auto-covariance functions, so the expected value of an auto-covariance coefficient of the cross-covariance function is

$$\begin{aligned} E\{R_{RR}(l,M)\} &= E[\{R_{xn}(m-l,M)-\mu_{R_{xn}}\}\{R_{xn}(m,M)-\mu_{R_{xn}}\}] \\ &= \frac{1}{M}\sum_{-\infty}^{+\infty}\varphi_{xx}(l-m)\,\varphi_{nn}(m) \end{aligned} \tag{7}$$

where

$$\varphi_{xx}(l) = E\{(x_{k-l}-\mu_x)(x_k-\mu_x)\} = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}(x_{k-l}-\mu_x)(x_k-\mu_x) \tag{8}$$

and $$\mu_{R_{xn}} = \mu_x\mu_n \tag{9}$$

$\varphi_{nn}(l)$ is defined in a similar way.

Until here, all estimators relate to records that have been corrected for their mean values. $R_{xn}(n,M)$ has been defined without this correction. The definition of R_{xn} with the correction for the mean value is

$$\begin{aligned} R^{\mathrm{corr}}_{xn}(n,M) &= \frac{1}{M}\sum_{k=0}^{M-1}(x_{k-n}-\mu_x)(n_k-\mu_n) \\ &= \frac{1}{M}\sum_{k=0}^{M-1}(x_{k-n}n_k-\mu_x n_k-\mu_n x_{k-n}+\mu_x\mu_n) \end{aligned} \tag{10}$$

If X is a pseudo-random binary sequence [6, 7] with a sequence length N, and $m = iN$ (i positive integer), then

$$R^{\mathrm{corr}}_{xn}(n,M) = R_{xn}(n,M) + C \tag{11}$$

(C constant). This is caused by the independence of $\Sigma_{k=0}^{M-1}\,x_{k-n}$ on n. If the

PRBS has the levels (−1) and (+1), and i digital samples are taken during each clockperiod, then

$$\left.\begin{aligned}\gamma_{xx}(m) &= 1 - \frac{|m - ijN|}{i} \cdot \frac{N+1}{N}, \text{ if } i(jN-1) < m < i(jN+1) \\ &\qquad \text{(any integer } j\text{)} \\ &= -\frac{1}{N}, \text{ elsewhere}\end{aligned}\right\} \tag{12}$$

The definition of $\gamma_{xx}(m)$ is

$$\gamma_{xx}(m) = E\{x_{k-m}x_k\} = \lim_{M\to\infty} \frac{1}{M} \sum_{k=0}^{M-1} x_{k-m}x_k \tag{13}$$

$$\gamma_{xx}(m) = \varphi_{xx}(m) + \mu_x^2 \tag{14}$$

In the case of a PRBS, $\mu_x = -1/N$, and for $i = 1$, eqn. (7) yields:

$$E\{R_{RR}(l,M)\} = \frac{1}{M} \sum_{m=-\infty}^{+\infty} \left[\left\{\gamma_{xx}(l) - \frac{1}{N^2}\right\}\varphi_{nn}(m)\right]$$

$$= \frac{1}{M}\left[\frac{N+1}{N} \sum_{m=-\infty}^{+\infty} \{\varphi_{nn}(l - mN)\} - \frac{N+1}{N^2}\Gamma_{nn}(0)\right] \tag{15}$$

For large N ($N \geqslant 127$), this equation may be approached by:

$$E\{R_{RR}(l,M)\} = \frac{1}{M} \sum_{m=-\infty}^{+\infty} \varphi_{nn}(l - mN) \tag{16}$$

$E\{R_{RR}(l,M)\}$ has a period iN; this is caused by the periodicity of $R_{xn}(m,M)$, which originates from the periodic time series X. This implies that the power spectrum of $R_{RR}(l,M)$ does not contain any frequency below $1/iN$. Equation (6) also elucidates this marked high-pass filtering effect of the correlation operation: If X is a PRBS, Γ_{xx} is a line spectrum with a lowest frequency $1/iN$

THE EFFECT OF THE QUANTIZATION OF THE OUTPUT OF THE SYSTEM

The continuous output of the system is quantized by an analog-to-digital (A/D) converter with a uniform quantizer level separation Δ. The input of the A/D converter is

$$g(t) = y(t) + n(t) \tag{17}$$

The output of the A/D converter is

$$z(t) = y(t) + n(t) + d(t) \tag{18}$$

$z(t)$ is sampled at time intervals Δt, which operation converts $z(t)$ to the time

series Z. Δt is assumed to be in accordance with the Shannon—Nyquist sampling theorem. The probability density function (pdf) of $g(t)$ is $p(g)$; the characteristic function of $g(t)$ is

$$\chi_g(q) = E\{\exp(jqg)\} = \int_{-\infty}^{+\infty} \exp(jqg)p(g)\mathrm{d}g \tag{19}$$

It can be shown [8, 9] that, if

$$\chi_g(q) = 0 \text{ for any } |q| \geqslant 2\pi/\Delta \tag{20}$$

then the following three equations are valid:

$$E(g) = E(z) \tag{21}$$

$$E(g^2) = E(z^2) - \frac{1}{12}\Delta^2 \tag{22a}$$

$$E(d^2) = \sigma_d^2 = \frac{1}{12}\Delta^2 \tag{22b}$$

Furthermore, every statistical moment of $g(t)$ is fully determined by the statistical moments of $z(t)$. The quantization noise $d(t)$ is uniformly divided on the interval $(-\frac{1}{2}\Delta, +\frac{1}{2}\Delta)$.

A Gaussian pdf cannot possibly satisfy condition (20) because its characteristic function is not bounded on an interval in the q domain. For any $\Delta \leqslant \sigma_g$, however, condition (20) is sufficiently approximated, and eqns. (21) and (22) are valid. The ACVF of the quantization noise is dependent on the ACVF of the input of the converter; for instance if $\Delta = \sigma_g$, then the normalized autocorrelation function $\rho_{dd}(n) = \phi_{dd}(n)/\sigma_d^2 \neq 0$ for those values of n where $\rho_{gg}(n) > 0.9$ [8]. For most practical purposes, this means that the quantization noise may be considered non-autocorrelated. For smaller values of Δ, the approximation of uncorrelated noise will even be better. This leads to:

$$E\{R_{R_{xd},R_{xd}}(l,M)\} = \frac{1}{M}\sum_{m=-\infty}^{+\infty} \gamma_{xx}(l-m)\,\gamma_{dd}(m) = \frac{\Delta^2}{12M}\gamma_{xx}(l) \tag{23}$$

An estimate of σ_g has to be made in order to check the validity of eqns. (20)—(22). Because $y(t)$ and $n(t)$ are uncorrelated:

$$\sigma_g^2 = \sigma_y^2 + \sigma_n^2 \tag{24}$$

σ_n^2 can be estimated from the ACVF of N. σ_y^2 can be estimated from the impulse response $h(t)$ of the system [10]:

$$R_{yy}(\tau) = E\{y(t-\tau)\,y(t)\} = \int_0^{\infty} \gamma_{hh}(\lambda)\,\gamma_{xx}(\tau-\lambda)\,\mathrm{d}\lambda \tag{25}$$

where

$$\gamma_{hh}(\lambda) = \int_0^{\infty} h(\tau)\,h(\tau+\lambda)\,\mathrm{d}\tau \tag{26}$$

$$E(y^2) = \int_0^\infty \gamma_{hh}(\lambda)\, \gamma_{xx}(-\lambda)\, d\lambda = \int_0^\infty \gamma_{hh}(\lambda)\, \gamma_{xx}(\lambda)\, d\lambda \tag{27}$$

If X is a PRBS with sequence length N and a clock period $\Delta' t$, and $\gamma_{hh}(\lambda) = 0$ for $\lambda > (N-1)\Delta' t$, then

$$E(y^2) = \int_0^{\Delta' t} \left(1 - \frac{\lambda}{\Delta' t} \cdot \frac{N+1}{N}\right) \cdot \gamma_{hh}(\lambda)\, d\lambda - \frac{1}{N} \int_{\Delta' t}^{(N-1)\Delta' t} \gamma_{hh}(\lambda)\, d\lambda \tag{28}$$

$$E(y) = E(x) \cdot \int_0^\infty h(\lambda)\, d\lambda = -\frac{1}{N} \int_0^\infty h(\lambda)\, d\lambda \tag{29}$$

$$\sigma_y^2 = E(y^2) - \{E(y)\}^2 = \int_0^{\Delta' t} \left(1 - \frac{\lambda}{\Delta' t} \cdot \frac{N+1}{N}\right) \gamma_{hh}(\lambda)\, d\lambda - \frac{1}{N} \int_{\Delta' t}^{(N-1)\Delta' t} \gamma_{hh}(\lambda)\, d\lambda$$

$$- \frac{1}{N^2} \left\{\int_0^\infty h(\lambda)\, d\lambda\right\}^2$$

$$= \int_0^{\Delta' t} \frac{N+1}{N} \left(1 - \frac{\Delta}{\Delta' t}\right) \gamma_{hh}(\lambda)\, d\lambda - \int_0^{(N-1)\Delta' t} \frac{1}{N}\, \gamma_{hh}(\lambda)\, d\lambda - \frac{1}{N^2} \left\{\int_0^\infty h(\lambda)\, d\lambda\right\}^2 \tag{30}$$

For large N ($N \geqslant 127$) this approaches:

$$\lim_{N \to \infty} \sigma_y^2 = \int_0^{\Delta' t} \left(1 - \frac{\lambda}{\Delta' t}\right) \gamma_{hh}(\lambda)\, d\lambda \tag{31}$$

If $\gamma_{hh}(\lambda)$ does not change appreciably on $(0, \Delta' t)$:

$$\lim_{N \to \infty} \sigma_y^2 \simeq \frac{1}{2} (\Delta' t)^2\, \gamma_{hh}(0) = \frac{1}{2} (\Delta' t)^2 \int_0^\infty h^2(\lambda)\, d\lambda \tag{32}$$

CROSS-CORRELATION OF PRBS INPUT AND RESULTING RESPONSE OF STATIONA AND NON-STATIONARY LINEAR SYSTEMS

Each linear system can be characterized by its impulse response $h(t,t')$. t is the absolute time parameter; if an impulse is applied to the input of the system at a point t_0 in absolute time, $h(t_0,t') = h(t_0, t - t_0)$ describes the response of the system departing from t_0 (see also Fig. 1).

The input of the system is $x(t)$, the output is the response $y(t)$ to the input $x(t)$. System noise and quantization noise are not considered here. $x(t)$

and $y(t)$ are transformed to the time series X and Y; $h(t,t')$ is transformed to a matrix $\{H\}$; an element of this matrix is $h_{m,n} = h(m\Delta t, n\Delta t)$. The number of elements in Y is M; it is assumed that $h_{m,n} = 0$ for $n < 0$ and $n > M - 1$.

An element y_k of the output Y is composed of earlier elements of the input, multiplied by the appropriate elements of the two-dimensional impulse response $\{H\}$:

$$y_k = x_k\, h_{k,0} + x_{k-1}\, h_{k-1,1} + \cdots\cdots + x_{k-M+1}\, h_{k-M+1,M-1} = \sum_{l=0}^{M-1} x_{k-l}\, h_{k-l,l} \tag{33}$$

This is the discretized convolution of $x(t')$ with $h(t,t')$.

Another way to write this equation is $Y = \{H\}\{X\}$, where

$$\{X\} = \begin{Bmatrix} x_{M-1}\, x_{M-2} \cdots\cdots x_0 \\ x_{M-2}\, x_{M-3} \cdots\cdots x_{-1} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ x_0 \quad x_{-1} \quad x_{1-M} \end{Bmatrix} \text{ and } \{H\} = \begin{Bmatrix} h_{M-1,0}\, h_{M-2,0} \cdots\cdots h_{0,0} \\ h_{M-2,1}\, h_{M-3,1} \cdots\cdots h_{-1,1} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ h_{0,M-1}\, h_{-1,M-1} \cdots\cdots h_{1-M,M-1} \end{Bmatrix}$$

The cross-correlation function of X and Y is

$$R_{xy}(n,M) = \frac{1}{M} \sum_{k=0}^{M-1} x_{k-n}\, y_k \tag{34}$$

Substitution of eqn. (33) into (34) yields:

$$R_{xy}(n,M) = \frac{1}{M} \sum_{k=0}^{M-1} x_{k-n} \sum_{l=0}^{M-1} x_{k-l}\, h_{k-l,l} = \frac{1}{M} \sum_{l=0}^{M-1} \sum_{k=0}^{M-1} h_{k-l,l}\, x_{k-n}\, x_{k-l} \tag{35}$$

Cross-correlation of the input and the output of a stationary linear system

If the system is stationary, $h_{k,l}$ is independent of k. All columns of $\{H\}$ are identical, and eqn. (35) reduces to

$$R_{xy}(n,M) = \frac{1}{M} \sum_{l=0}^{M-1} h_l \sum_{k=0}^{M-1} x_{k-n}\, x_{k-l} = \sum_{l=0}^{M-1} h_l\, R_{xx}(n,l,M) \tag{36}$$

where $R_{xx}(n,l,M) = 1/M \sum_{k=0}^{M-1} x_{k-n}\, x_{k-l}$. If $R_{xx}(n,l,M) \equiv R_{xx}(n-l,M)$ for all n and l, then:

$$R_{xy}(n,M) = \sum_{l=0}^{M-1} h_l\, R_{xx}(n-l,M) \tag{37}$$

which is the discretized convolution of $h(t')$ with $R_{xx}(t',T)$; $T = (M-1)\Delta t$.

The ideal $R_{xx}(n-l,M)$ would be zero for all $n \neq l$ and 1 for $n = l$. In that

case, eqn. (37) would reduce to $R_{xy}(n,M) = h_n$. The cross-correlogram would be an exact replica of the impulse response. It is possible to construct time se series that have such an ACVF, but a series can be chosen that is more suitable from a computational point of view: the PRBS. A PRBS is considered with a sequence length N and a clock period $\Delta' t$; $\Delta' t = i\Delta t$, i integer > 0. This implies that $M = iN$. Because a PRBS is periodic with a period M, $R_{xx}(n,l,M) = R_{xx}(n-l,M)$ and $R_{xx}(n,jM) = \gamma_{xx}(n)$ (j integer > 0).

This means that the estimation of the auto-covariance function of the PRBS, if computed over an integral number of sequences, is exactly equal to the auto-covariance function. For γ_{xx}, see eqn. (15). Equations (15) and (37) can now be combined to

$$R_{xy}(n,M) = \sum_{l=0}^{M-1} h_l \, \gamma_{xx}(n-l) \tag{38}$$

If $i = 1$, eqn. (38) reduces to

$$R_{xy}(n,N) = h_n - \frac{1}{N}\left\{\left(\sum_{l=0}^{N-1} h_l\right) - h_n\right\} \tag{39a}$$

which can be rewritten as

$$h_n = \frac{N}{N+1} R_{xy}(n,M) + \frac{1}{N+1} \sum_{l=0}^{N-1} h_l \tag{39b}$$

This equation shows a linear relation between $R_{xy}(n,N)$ and h_n.

If an impulse response obtained by impulse excitation is compared with an impulse response obtained by cross-correlation, the sequence length should be taken into account.

If more than 1 digital sample is taken per clock period ($i > 1$), no straightforward relation exists between $R_{xy}(n,m)$ and h_n. Because of the convolution of $h(t')$ with $R_{xx}(t',T)$, $R_{xy}(n,M)$ depends not only on h_n, but also on its neighbours within a distance $\pm (i-1)$. The sum of the coefficients of R_{xy}, which is a measure for the surface of the impulse response, is

$$\begin{aligned}
\sum_{n=0}^{M-1} R_{xy}(n,M) &= \sum_{n=0}^{M-1} \sum_{l=0}^{M-1} h_l \, \gamma_{xx}(n-l) \\
&= \sum_{n=0}^{M-1} \left[-\frac{1}{N} \sum_{l=0}^{M-1} h_l + \sum_{l=n-i+1}^{n+i-1} h_l \left\{ \frac{N+1}{N} \left(1 - \frac{|n-l|}{i}\right)\right\}\right] \\
&= -i \sum_{l=0}^{M-1} h_l + \sum_{l=0}^{M-1} h_l \cdot \sum_{n=1-i}^{i-1} \frac{N+1}{N} \left(1 - \frac{|n|}{i}\right)
\end{aligned} \tag{40}$$

This sum approaches zero with increasing N, which is the consequence of the fact that μ_x approaches zero with increasing N. The sum corrected for the negative baseline offset is

$$\sum_{n=0}^{M-1} R_{xy}^{\text{corr}}(n,M) = \sum_{n=0}^{M-1} \sum_{l=n-i+1}^{n+i-1} h_l \left\{ \frac{N+1}{N} \left(1 - \frac{|n-l|}{i}\right) \right\}$$

$$= \sum_{l=0}^{M-1} h_l \sum_{n=1-i}^{i-1} \frac{N+1}{N} \left(1 - \frac{|n|}{i}\right) = i \left(\frac{N+1}{N}\right) \sum_{l=0}^{M-1} h_l \qquad (41)$$

Equation (41) shows a linear relation between $\Sigma R_{xy}^{\text{corr}}$ and Σh_l. Though for $i > 1$ the simple relation between $R_{xy}(n)$ and h_n is lost, a proportionality remains between ΣR_{xy} and Σh_n.

Cross-correlation of the input and the output of a non-stationary linear system

If the linear system is not stationary, the columns of $\{H\}$ are not equal. The coefficients $h_{k-l,l}$ in eqn. (35) depend on k and they cannot be moved to a position before the Σ_k sign. The ACVF $R_{xx}(n-l,M)$ cannot be separated from eqn. (35). However, by developing $h(t,t')$ in a Taylor series departing from $t = 0$, it is possible to split eqn. (35) into higher-order terms:

$$h(t,t') = \sum_{i=0}^{\infty} \frac{t^i}{i!} \frac{\partial^i h(0,t')}{\partial t^i} \qquad (42)$$

$$h_{k-l,l} = \sum_{i=0}^{\infty} \frac{\{(k-l)\Delta t\}^i}{i!} h_{0,l}^{(i)} \qquad (43)$$

where $h_{0,l}^{(i)} = \partial^i h(0,l\Delta t)/\partial t^i$.

Other expansions of $h(t,t')$ into a series might also be considered. The Taylor series expansion describes the behaviour of each coefficient of the impulse response in absolute time, departing from $t = 0$. Substitution of eqn. (43) into eqn. (35) and reversion of the order of summation yields

$$R_{xy}(n,M) = \frac{1}{M} \sum_{i=0}^{\infty} \sum_{l=0}^{M-1} \sum_{k=0}^{M-1} \frac{\{(k-l)\Delta t\}^i}{i!} h_{0,l}^{(i)} x_{k-n} x_{k-l}$$

$$= \sum_{i=0}^{\infty} \sum_{l=0}^{M-1} C_{i,l} R_{xx}^{(i)}(n,l,M) \qquad (44)$$

$$\text{with } C_{i,l} = \frac{(\Delta t)^i}{i!} h_{0,l}^{(i)} \qquad (45)$$

$$\text{and } R_{xx}^{(i)}(n,l,M) = \frac{1}{M} \sum_{k=0}^{M-1} (k-l)^i x_{k-n} x_{k-l} \qquad (46)$$

Equation (44) reduces to (36) if all terms vanish for $i \neq 0$ (stationary case).

Even if X is a stationary time series with $\mu_x = 0$, the higher-order terms of eqn. (44) do not converge for $M \to \infty$. The same is true when X is a PRBS. A PRBS, however, is not a truly random series but a well defined function of

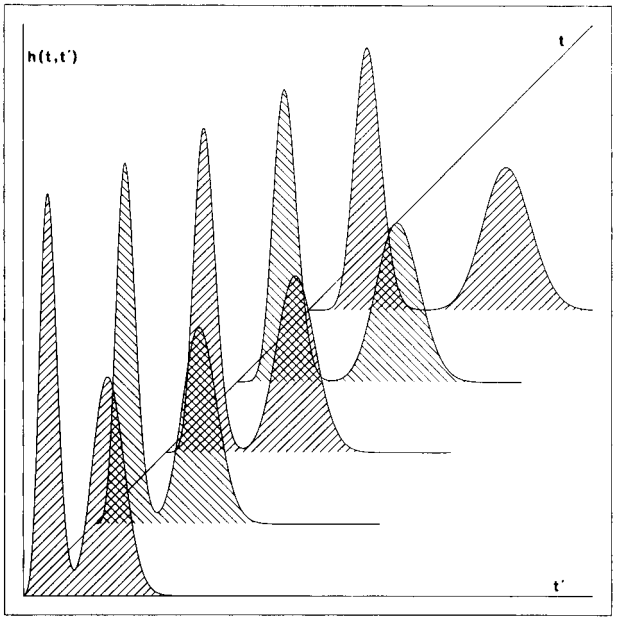


Fig. 1. Two-parameter impulse response of a non-stationary linear system.

time, and the higher-order terms of eqn. (46) can be computed for a given M, N, i, n, and l. If the impulse response is also known (both in the absolute and the relative time domain), all terms of eqn. (44) can be computed, and it becomes possible to subtract the higher-order terms from this equation to obtain the stationary R_{xy}.

DC Shift between the input applied to the system and the input used for the computation of the correlogram

For various reasons, it may be desirable to execute the computation of the correlogram with a PRBS that is DC-shifted (offset) with respect to the PRBS that was actually applied to the system. Many systems only allow a one-sided input PRBS (levels 0 and +2). An example is the chromatograph: input and output both are concentrations. However, the DC components of input and output will be multiplied in the correlogram. The resulting DC offset $\mu_x\mu_z$ must be subtracted from the correlogram, giving rise to an increased arithmet error. The DC offset might even result in an arithmetic overflow of the computing device used. In this case, it is advisable to use a symmetrical PRBS (levels −1 and +1) for the computation, because in this case $\mu_x = -1/N \simeq 0$.

If the computing device allows it, however, it may be advantageous to use a one-sided PRBS for the computation. The computation of R_{xy} by using a one-sided PRBS needs only half the number of additions needed for a symmetrical PRBS, thus reducing the time needed by 50%. The effect of a DC shift applied to X is

$$R'_{xy}(n,M) = \frac{1}{M}\sum_{k=0}^{M-1}(x_{k-n}+C)y_k = R_{xy}(n) + \frac{C}{M}\sum_{l=0}^{M-1}\sum_{k=0}^{M-1} h_{k-l,l}\, x_{k-l}$$

$$= R_{xy}(n) + \frac{C}{M}\sum_{i=0}^{\infty}\sum_{l=0}^{M-1}\sum_{k=0}^{M-1}\frac{\{(k-l)\Delta t\}^i}{i!}\, h_{0,l}^{(i)}\, x_{k-l} = R_{xy}(n) + C' \quad (47)$$

In the stationary case:

$$C'_{\text{stat}} = \frac{C}{M}\sum_{l=0}^{M-1} h_l \sum_{k=0}^{M-1} x_{k-l} \quad (48)$$

When X is a PRBS with sequence length N and levels (a,b):

$$C'_{\text{stat,PRBS}} = \frac{C}{N}\left(\frac{N-1}{2}\,b - \frac{N+1}{2}\,a\right)\sum_{l=0}^{M-1} h_l \quad (49)$$

In the non-stationary case, the value of C' cannot easily be predicted, but it is also independent of n. So, in both cases the application of a DC shift to X only results in a DC shift of the correlogram.

Digital simulation

In order to check the validity of eqns. (44) and (47), a non-stationary system was simulated on a minicomputer (Varian V76 operating under Vortex-II). The simulation program computes the non-stationary R_{xy} in two different ways: one using eqns. (33) and (34), the other using eqn. (44) and, if necessary, eqn. (47). One sequence of the output, Y, is computed using two sequences of the input X in eqn. (33). The cross-correlation function is computed over one sequence, using the same two sequences of the input and the resulting sequence of the output. It is assumed that $M = N$. No extraneous noise is added and the quantization noise is not taken into account. Inputs to the program are: the sequence length; the choice of PRBS levels (−1, +1) or (0, +2); the type of the impulse response (a Poisson or a Gaussian peak); the type of the non-stationarity (linear increase or decrease of the amplitude of the impulse response with t, exponential increase or decrease with t, or linear change with t of the first time moment of the impulse response); the amount of change per clock period; and the number of higher-order terms calculated (maximum 6). Outputs of the program are: R_{xy} computed both ways; the record of differences between these two functions; the contribution of the non-zero order terms to R_{xy}; R_{xy} corrected for non-zero order terms; and the impulse response at $t = 0$.

Results of simulation

The two cross-correlograms obtained both ways were equal to each other within computational error. Figure 2 shows a non-stationary simulated correlogram (obtained either way). In Fig. 3 the correlogram is split into a stationary

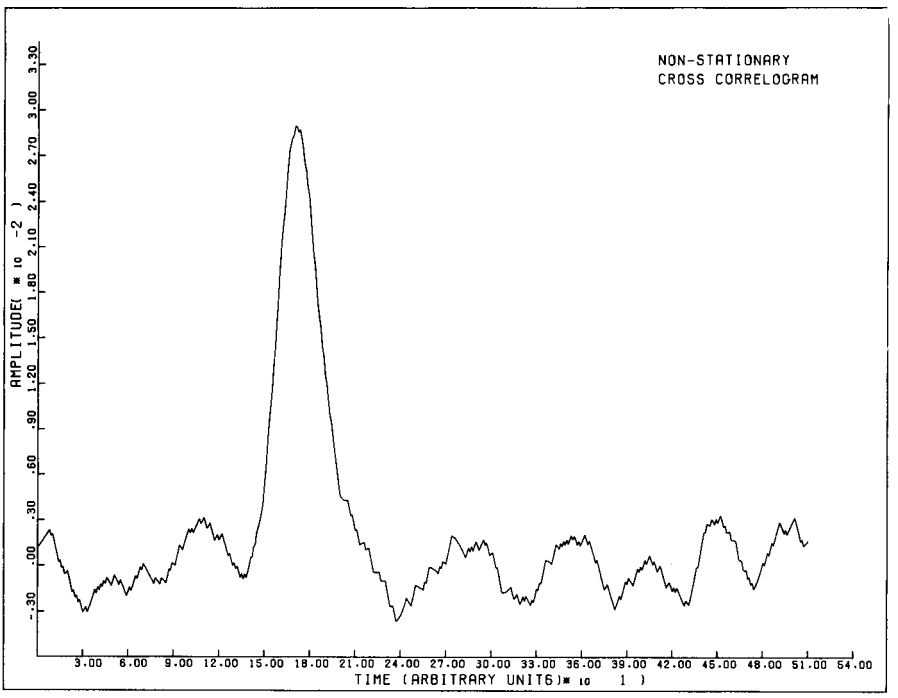


Fig. 2. Cross-correlogram of a non-stationary linear system. Sequence length = 511; number of digital samples per clock period = 1; levels of input PRBS are (0) and (+2); levels of PRBS used for correlation are (—1) and (+1); the impulse response is a Poisson distribution the first time moment shifting +0.03 clock period per clock period.

and a non-stationary part. Figure 4 shows the power spectrum of the non-stationarity "noise" of Fig. 3; it can be seen that only a small part of the "noise" could be removed by low-pass filtering, because most of the power is located in the same band as the power of the stationary part. The spectra were obtained by a user-oriented fast Fourier program, developed in this laboratory [15].

IMPLICATIONS FOR PRACTICAL CORRELATION CHROMATOGRAPHY

In addition to some optimal conditions mentioned in an earlier paper [5], the following observations can be made.

Filtering of the output of the system

Low-pass filtering has already been discussed [5]. From eqn. (7) it can be seen that noise frequencies below $1/M\Delta t$ are not expected to be present in the spectrum of R_{xn}. Nevertheless, they can contribute to $G'_{xn}(0,M)$ and $G'_{xn}(1,M)$ by leakage [11, 12]. This is a consequence of the finite record

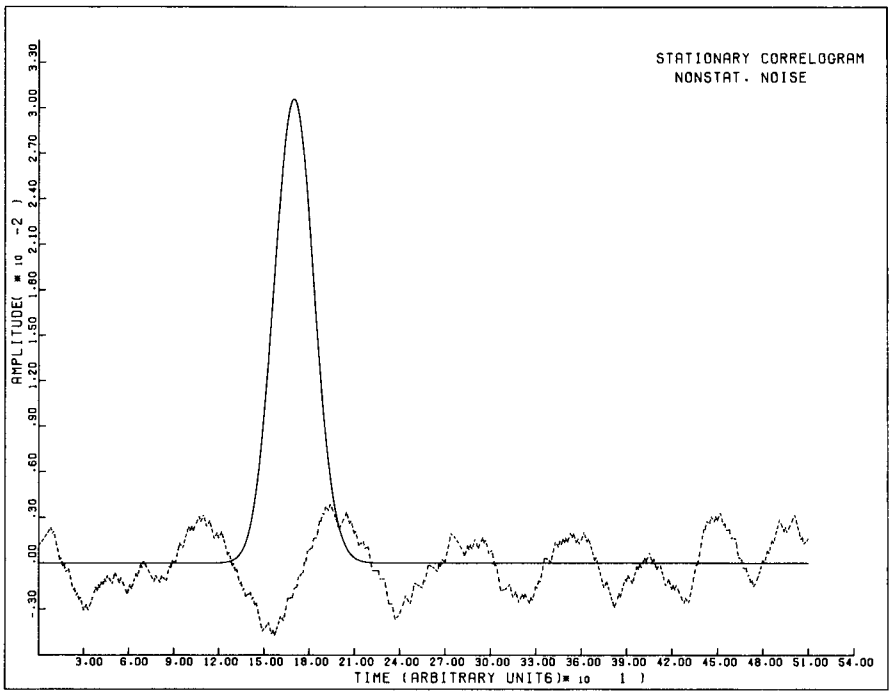


Fig. 3. Same as Fig. 2; the non-stationary correlogram is split into a stationary (zero-order) part (——) and a non-stationary (higher-order) part (-----). Higher-order terms were computed up to 6th order.

length of N. Therefore, it may be of use to pass $g(t)$ through an analog high-pass filter, or Z through a digital one, with a cut-off frequency $1/M\Delta t$. A considerable loss of time has to be taken into account, however, because these filters need a very long time to be free of transients (2 or 3 times $M\Delta t$).

Prediction of the signal-to-noise-ratio improvement as a function of the correlation time, the resolution of the A/D conversion, and the integration time of the chromatographic peak

In chromatographic practice, quantitative analytical information is often obtained by integration of a chromatographic peak. The noise is integrated together with the analytical signal, giving rise to a variance σ_I^2 in the integral I. In this section, the integrated noise of the normal chromatogram is compared with the integrated noise of the correlation chromatogram, the integration (actually a summation) being carried out over the same integration time $K\Delta t$. If the injection volume of the normal chromatogram equals the virtual injection volume [5] of the correlation chromatogram, and if the input sample concentrations are the same, then the signal-to-noise-ratio improvement equals the reciprocal of the square root of the ratio of the variances of the integrated noise.

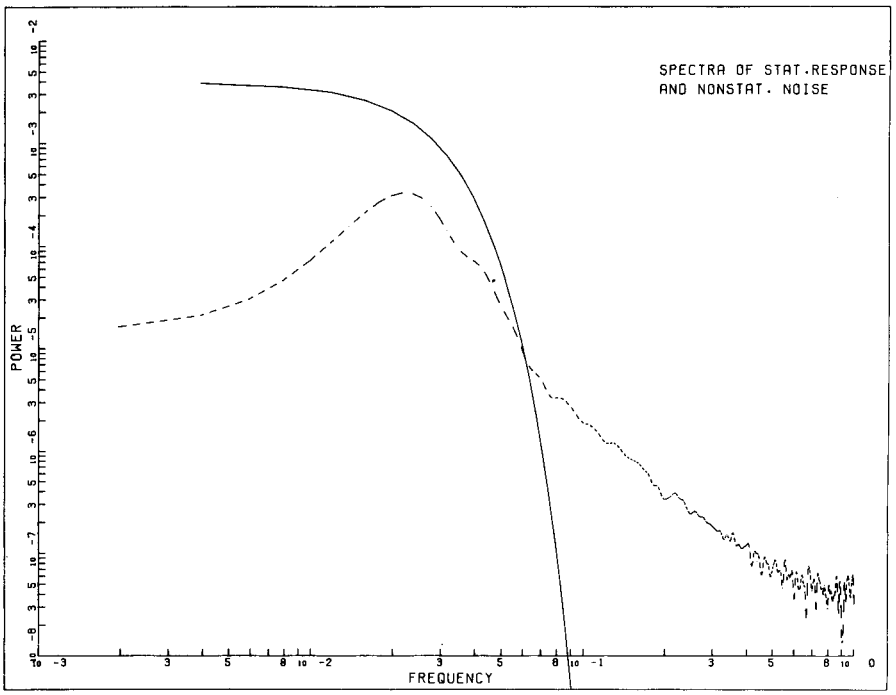


Fig. 4. Spectra of stationary part (——) and non-stationary part (-----) of Fig. 3. The absolute sizes of both spectra may not be compared with each other.

The normal chromatogram is quantized with a level-distance Δ_1, the output $g(t)$ of the correlation system with Δ_2. Δ_1 and Δ_2 are both assumed to satisfy eqn. (20), which means that the quantization noise is not autocorrelated, and its variance $\sigma_d^2 = \Delta^2/12$. For a range of summation K, the variance of the integral of the noise in a normal chromatogram is:

$$\sigma_{I,K}^2 = K\left(\sigma_n^2 + \frac{\Delta_1^2}{12}\right) + 2\sum_{k=1}^{K-1}(K-k)\phi_{nn}(k) \tag{50}$$

This is the discretized form of the relation derived in [13]. In practice, the part of the chromatogram that is going to be integrated is corrected for the DC term and for linear drift. Therefore, ϕ_{nn} should be estimated after the application of a similar correction to each interval used in the computation of the estimator. A paper on this subject is in preparation [14].

The variance of the integrated noise of the correlogram is

$$\sigma_{R,I,K}^2 = K\phi_{RR}(0) + 2\sum_{k=1}^{K-1}(K-k)\phi_{RR}(k)$$

$$= \frac{1}{M}\left[K\left\{\sigma_n^2 + \frac{\Delta_2^2}{12} + 2\sum_{m=1}^{\infty}\varphi_{xx}(m)\,\varphi_{nn}(m)\right\}\right.$$

$$\left. + 2\sum_{k=1}^{K-1}(K-k)\sum_{m=-\infty}^{+\infty}\left\{\varphi_{xx}(k-m)\,\varphi_{nn}(m)\right\}\right] \tag{51}$$

It is assumed that X is a PRBS. If eqn. (16) is valid, and φ_{nn} is 0 outside $(1-N, N-1)$, then

$$\sigma^2_{R,I,K} = \frac{1}{M}\left[K\left(\sigma_n^2 + \frac{\Delta_2^2}{12}\right) + 2\sum_{k=1}^{K-1}(K-k)\,\phi_{nn}(k)\right] \tag{52}$$

$\sigma^2_{R,I,K}$ could also be improved by a linear correction of the range of summation. In that case, however, there is no obvious relation between $\sigma^2_{I,K}$ and $\sigma^2_{R,I,K}$.

The S/N ratio improvement related to the peak amplitude (no integration) is:

$$[\sigma^2(n+d)/\sigma_R^2]^{\frac{1}{2}} = M\left\{\sigma_n^2 + \frac{\Delta_1^2}{12}\right\}\Big/\left\{\sigma_n^2 + \frac{\Delta_2^2}{12} + 2\sum_{m=1}^{\infty}\varphi_{xx}(m)\,\varphi_{nn}(m)\right\} \tag{53}$$

If the conditions leading to eqn. (52) are satisfied, then:

$$[\sigma^2(n+d)/\sigma_R^2]^{\frac{1}{2}} = \left[M\left(\sigma_n^2 + \frac{\Delta_1^2}{12}\right)\Big/\left(\sigma_n^2 + \frac{\Delta_2^2}{12}\right)\right]^{\frac{1}{2}} \tag{53a}$$

Choice of the resolution of the AD converter

$\sigma^2_{R,I,K}$ increases with Δ_2. An ideal quantization would have no effect on the uncertainty in the correlogram. Let us assume that the contribution of the quantization to the variance of the correlogram should not be more than $a\%$ of the contribution of the system noise; this is the case if $\Delta_2^2 \leqslant 12a\sigma_n^2/100$. This criterion dictates the level separation of the quantization; the minimal number of levels required is dictated by the separation level Δ_2 and the expected signal range; if $g(t)$ is Gaussian, the range of $g(t)$ will be about $\mu_g \pm 4\sigma_g$. If the required number of levels is L, then $L\,\Delta_2 \geqslant 8\sigma_g$. If the S/N ratio at the output of the system is b, which means $\sigma_y = b\sigma_n$, and $\sigma_y^2 = (1+b^2)\sigma_n^2$, then:

$$L \geqslant 8[(1+b^2)\sigma_n^2]^{\frac{1}{2}}/[0.12\,a\sigma_n^2]^{\frac{1}{2}} = 8[(1+b^2)/0.12\,a]^{\frac{1}{2}} \tag{54}$$

It is important that σ_n^2 is estimated including the very low noise frequencies (usually called drift). Those noise components may contribute considerably to σ_n^2, which is another possible reason for high-pass filtering: a reduction of L.

Injection type and number of digital samples per clock period

Suppose the levels of the input PRBS X are a and b. Two types of input excitation are possible: (a) synchronous with the digital sampling, impulses

with an amplitude $a\Delta t$ (if $x_k = a$), or $b\Delta t$ (if $x_k = b$) are applied to the input, or (b) the input is switched between the levels a and b. The discretized γ_{xx} of both processes are the same, but the continuous $\gamma_{xx}(\tau)$ are not (Fig. 5). If it can be realised, method (a) should be preferred. In that case, one injection per clock period should be applied; eqn. (39) can be applied, and R_{xy} is proportional to the true impulse response. If more points in the correlogram are desired, they should be obtained by an interpolation (or fitting) procedure rather than by an increase of the number of impulse injections in each clock period, because in this case R_{xy} will be a convolution of the impulse response with a triangular γ_{xx}. It is also possible to take more than one sample per clock period, using only one of them for a synchronous injection.

If method (b) is used, the number of samples per clock period is not important. R_{xy} will always be a digitized convolution of $\gamma_{xx}(\tau)$ and $h(\tau)$. The number of points in the correlogram may be increased indefinitely by taking more samples per clock period.

In practice, the clock period chosen will mostly be very small compared with the impulse response of the system; in this case there will be no considerable difference between method (a) and method (b). Method (b) should even be preferred if there is a risk that the system will be overloaded by large input amplitudes (i.e., the system is non-linear for large amplitudes).

Correction of the correlation chromatogram for non-stationarity

It is obvious that an optimal design with respect to stationarity is the best wa to avoid the ill effects of non-stationarity on the correlogram. The input con-

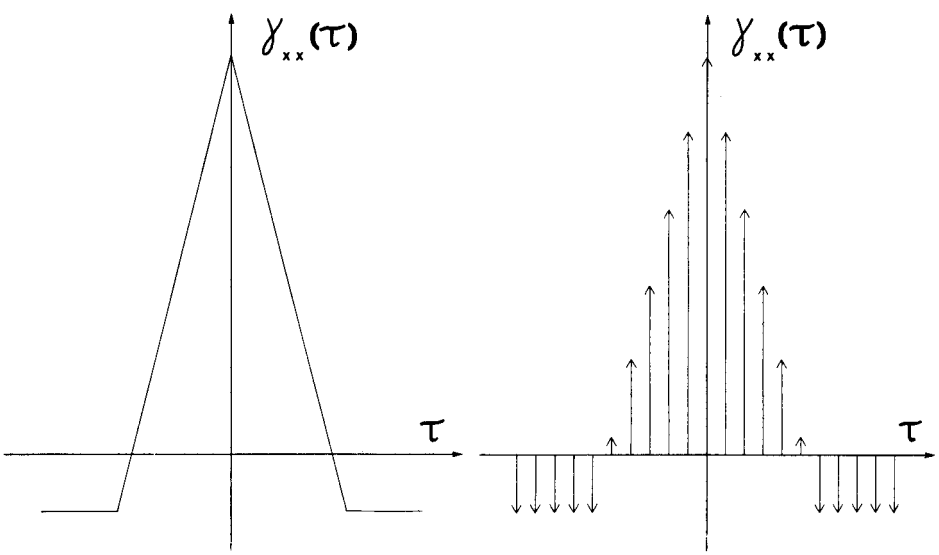


Fig. 5. (a) $\gamma_{xx}(\tau)$ of a PRBS-type continuous injection. (b) $\gamma_{xx}(\tau)$ of a PRBS-type impulse train injection.

centration and the flow rate of the mobile phase should be kept as constant as possible. A constant temperature of the column should result in stable retention indices of the components of the sample.

If, nevertheless, some non-stationarity does occur, an attempt can be made to correct the correlogram during computation. In that case, $\{H\}$ has to be known. In practice, only a limited number of peaks in the chromatogram will be known before the correlation is carried out; those are peaks that can be found by straightforward chromatography. Because they are the largest peaks they will cause most of the non-stationarity "noise". Each coefficient of $\{H\}$ is considered as a sum of the contributions from J different components of the sample:

$$h_{k,l} = {}^1h_{k,l} + {}^2h_{k,l} + \cdots\cdots {}^{J-1}h_{k,l} + {}^Jh_{k,l} \tag{55}$$

$$\text{or } \{H\} = \sum_{j=1}^{J} {}^j\{H\} \tag{56}$$

$R_{xy}(n)$ can be considered as composed of J parts, each originating from one component:

$$R_{xy}(n) = \frac{1}{M} \sum_{i=0}^{\infty} \sum_{l=0}^{M-1} \sum_{k=0}^{M-1} \sum_{j=1}^{J} \frac{\{(k-l)\Delta t\}^i}{i!} {}^jh_{0,l}\, x_{k-n}\, x_{k-l}$$

$$= \frac{1}{M} \sum_{i=0}^{\infty} \sum_{j=1}^{J} \sum_{l=0}^{M-1} \sum_{k=0}^{M-1} \frac{\{(k-l)\Delta t\}^0}{i!} {}^jh_{0,l}\, x_{k-n}\, x_{k-l} = \sum_{i=0}^{\infty} \sum_{j=1}^{J} {}^jR_{xy}^{(i)}(n) \tag{57}$$

If the system parameters from which the non-stationarity arises are monitored, and the relations between those parameters and a number of the matrices ${}^j\{H\}$ are known, then a number of the ${}^jR_{xy}^{(i)}$ can be computed for $i \neq 0$. They can be computed in one run by adding all known ${}^j\{H\}$. They are subtracted from R_{xy}. If new peaks emerge from the now reduced baseline noise, the procedure can be repeated with the newly found ${}^j\{H\}$, and so on, until no further enhancement can be achieved.

The most frequent sources of non-stationarity are variations of the input concentration and variations of the flow. The input concentration can be monitored by some overall concentration measurement (e.g. u.v. absorption, conductivity, dielectric constant, etc.). The input amplitude of each component at a point in absolute time is proportional to the total concentration. The flow can be measured directly (drop counter, flow meter) or indirectly from the correlograms of subsequent sequences. The unretarded peaks in the correlograms of separate sequences indicate the mean flows during the input sequences; the flow at each point in time can be approached by fitting a curve through the mean flows of all sequences. Because the variations will be relatively small, it will be a good approximation to assume that all peaks wander without changing their shapes.

APPENDIX

The power spectrum of the cross-covariance function of two mutually uncorrelated stationary time series

Two stationary time series **X** and **Y** are considered: $\mu_x = \mu_y = 0$ and $E\{x_{k-n}, y_k\} = 0$ (all k and n). From both time series a sample of M elements is taken:

$$\left.\begin{aligned} x_M(k) &= x_k, \text{ if } 0 \leqslant k < M \\ &= 0, \text{ elsewhere} \end{aligned}\right\} \text{ and } \left.\begin{aligned} y_M(k) &= y_k, \text{ if } 0 \leqslant k < M \\ &= 0, \text{ elsewhere} \end{aligned}\right\}$$

The Fourier transforms of both finite records are defined [16]:

$$X_M(m) = \sum_{k=-\infty}^{+\infty} x_M(k) \exp(-2\pi ikm\, M^{-1}) = \sum_{k=0}^{M-1} x_k \exp(-2\pi ikm\, M^{-1})$$

$$Y_M(m) = \sum_{k=-\infty}^{+\infty} y_M(k) \exp(-2\pi ikm\, M^{-1}) = \sum_{k=0}^{M-1} y_k \exp(-2\pi ikm\, M^{-1})$$

The auto-covariance functions of both finite records are defined:

$$R_{xx}(n,M) = \frac{1}{M} \sum_{k=-\infty}^{+\infty} x_M(k-n)\, x_M(k) \text{ and } R_{yy}(n,M) = \frac{1}{M} \sum_{k=-\infty}^{+\infty} y_M(k-n)\, y_M(k)$$

The power spectra of both finite records are defined:

$$G_{xx}(m,M) = \sum_{n=-\infty}^{+\infty} R_{xx}(n,M) \exp(-2\pi inm\, M^{-1}) = \frac{1}{M} \cdot X_M(m) \cdot X_M^*(m)$$

$$G_{yy}(m,M) = \sum_{n=-\infty}^{+\infty} R_{yy}(n,M) \exp(-2\pi inm\, M^{-1}) = [Y_M(m) \cdot Y_M^*(m)]/M$$

The cross-covariance function of the finite records is defined:

$$R_{xy}(n,M) = [\sum_{k=-\infty}^{+\infty} x_M(k-n)\, y_M(k)]/M$$

The cross-spectrum of the two finite records is defined:

$$G_{xy}(m,M) = \sum_{n=-\infty}^{+\infty} R_{xy}(m,M) \exp(-2\pi inm\, M^{-1}) = [-X_M^*(m) \cdot Y_M(m)]/M$$

The cross-covariance function can also be defined if only one of the two records is limited to the interval $(0, M-1)$:

$$R'_{xy}(n,M) = \frac{1}{M} \sum_{k=-\infty}^{+\infty} x_{k-n}\, y_M(k)$$

$R_{xy}(n,M)$ is 0 for $|n| > M-1$, while $R'_{xy}(n,M)$ is not. $R'_{xy}(n,M)$ can be limited on the interval $(0,M-1)$:

$$\left.\begin{aligned} R'_{M,xy}(n,M) &= R'_{xy}(n,M), \text{ if } 0 \leqslant n \leqslant M-1 \\ &= 0, \text{ elsewhere} \end{aligned}\right\}$$

The Fourier transform of $R'_{M,xy}(n,M)$ is:

$$G'_{xy}(m,M) = \sum_{n=-\infty}^{+\infty} R'_{M,xy}(n,M) \,.\, \exp(-2\pi inm\, M^{-1})$$

The auto-covariance function of $R'_{M,xy}(n,M)$ is:

$$R_{RR}(l,M) = \frac{1}{M} \sum_{n=-\infty}^{+\infty} R'_{M,xy}(n-l,M)\, R'_{M,xy}(n,M)$$

The power spectrum of $R'_{M,xy}(n,M)$ is:

$$G_{RR}(m,M) = \sum_{l=-\infty}^{+\infty} R_{RR}(l,M) \exp(-2\pi ilm\, M^{-1}) = \frac{1}{M} \cdot G'_{xy}(m,M) \cdot G'^{*}_{xy}(m,M) \tag{A-1}$$

If something is to be said about $G_{RR}(m,M)$, the properties of $G'_{xy}(m,M)$ should be known. For this purpose, $G'_{xy}(m,M)$ is compared with $G_{xy}(m,M)$; the properties of the latter are known.

$$G'_{xy}(m,M) = \sum_{n=0}^{M-1} \frac{1}{M} \sum_{k=-\infty}^{+\infty} x_{k-n}\, y_M(k) \exp(-2\pi inm\, M^{-1}) \tag{A-2}$$

$$\begin{aligned} G_{xy}(m,M) &= \sum_{n=-\infty}^{+\infty} \frac{1}{M} \sum_{k=-\infty}^{+\infty} x_M(k-n)\, y_M(k) \exp(-2\pi inm\, M^{-1}) \\ &= \frac{1}{M} \sum_{k=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} x_M(k-n)\, y_M(k) \exp(-2\pi inm\, M^{-1}) \\ &= \frac{1}{M} \sum_{k=-\infty}^{+\infty} \sum_{n=k}^{k+M-1} x_{k-n}\, y_M(k) \exp(-2\pi inm\, M^{-1}) \end{aligned} \tag{A-3}$$

$x_k \overset{\text{div}}{=} x_0$ (X is a stationary time series); here $\overset{\text{div}}{=}$ stands for "has the same probability function as".

$x_k y_M(k) \overset{\text{div}}{=} x_0 y_M(k)$ (X and Y are mutually independent) (A-4)

It follows from eqn. (A4) that:

$$\begin{aligned} G_{xy}(m,M) &= \frac{1}{M} \sum_{k=-\infty}^{+\infty} \sum_{n=k}^{k+M-1} x_{k-n}\, y_M(k) \exp(-2\pi inm\, M^{-1}) \\ &\overset{\text{div}}{=} \frac{1}{M} \sum_{k=-\infty}^{+\infty} \sum_{n=0}^{M-1} x_{k-n}\, y_M(k) \exp(-2\pi inm\, M^{-1}) \\ &= \sum_{n=0}^{M-1} \frac{1}{M} \sum_{k=-\infty}^{+\infty} x_{k-n}\, y_M(k) \exp(-2\pi inm\, M^{-1}) = G'_{xy}(m,M) \end{aligned} \tag{A-5}$$

Summarizing:

$$G'_{xy}(m,M) \overset{\text{div}}{=} G_{xy}(m,M) = \frac{1}{M}\{-X^*_M(m) \,.\, Y_M(m)\} \tag{A-5a}$$

For G_{RR} follows:

$$G_{RR}(m,M) = \frac{1}{M}\{G'_{xy}(m,M) \,.\, G'^*_{xy}(m,M)\} \overset{\text{div}}{=} \frac{1}{M}\{G_{xy}(m,M) \,.\, G^*_{xy}(m,M)\}$$

$$= \frac{X_M(m) \,.\, X^*_M(m) \,.\, Y_M(m) \,.\, Y^*_M(m)}{M^3} = \frac{G_{xx}(m,M) \,.\, G_{yy}(m,M)}{M} \tag{A-6}$$

It follows from eqn. (A-6) that:

$$E\{G_{RR}(m,M)\} = \Gamma_{RR}(m) = \frac{\Gamma_{xx}(m)\,\Gamma_{yy}(m)}{M} \tag{A-7}$$

and

$$E\{R_{RR}(l,M)\} = \varphi_{RR}(l) = \frac{1}{M}\sum_{n=-\infty}^{+\infty} \varphi_{xx}(l-n)\,\varphi_{yy}(n) \tag{A-7}$$

REFERENCES

1 H. C. Smit, Chromatographia, 3 (1970) 515.
2 R. Annino and L. E. Bullock, Anal. Chem., 45 (1973) 1221.
3 R. Annino, J. Chromatogr. Sci., 14 (1976) 765.
4 S. B. Philips and M. F. Burke, J. Chromatogr. Sci., 14 (1976) 495.
5 T. T. Lub, H. C. Smit and H. Poppe, J. Chromatogr., 149 (1978) 721.
6 G. A. Korn, Random Process Simulation and Measurements, McGraw-Hill, New York, 1966, p. 83.
7 A. C. Davies, IEEE Trans. Comput., C20 (1971) 270.
8 B. Widrow, IRE Trans. Circuit Theory, (Dec. 1956) 266.
9 See ref. 6, p. 131.
10 See ref. 6, p. 7.
11 E. O. Brigham, The Fast Fourier Transform, Prentice-Hall, New Jersey, 1974, p. 105.
12 G. M. Jenkins and D. G. Watts, Spectral Analysis and its Applications, Holden-Day, San Francisco, 1968, p. 282.
13 H. C. Smit and H. L. Walg, Chromatographia, 8 (1975) 311.
14 R. P. J. Duursma and H. C. Smit, in preparation.
15 H. L. Walg and H. C. Smit, Anal. Chim. Acta, 103 (1978) 43.
16 J. S. Bendat, Principles and Applications of Random Noise Theory, Wiley, New York, 1958, p. 63.

Analytica Chimica Acta, 112 (1979) 361—374
Computer Techniques and Optimization

MULTICOMPONENT-ANALYSIS COMPUTATIONS BASED ON KALMAN FILTERING

H. N. J. POULISSE

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

(Received 31st May 1979)

SUMMARY

A theoretical introduction to the use of Kalman filtering in analytical chemistry is based on multicomponent-analysis computations with the non-recursive least-squares estimation method as a starting point. An initial value for the computation of the error covariance matrix is given and some new practical applications (determination of number of components, estimation of constant systematic error) are derived and demonstrated. Theory and practice suggest a new possible design for experimental measurements and novel applications of on-line computation and computer control. The excellent performance of the Kalman filter algorithm is demonstrated.

The handling of data obtained from measurements plays a crucial role in analytical chemistry and is normally a problem of a statistical nature, because in practice the signals observed are corrupted by noise. Wiener [1] showed that for a stationary situation the separation of signal and noise leads to the so-called Wiener—Hopf equation. Computational difficulties have prevented the Wiener filter from becoming popular. The solution to this problem formulated by Kalman [2] does not show such difficulties. The Kalman filter, in contrast to the Wiener filter, is recursive in nature and applicable to non-stationary situations. It has been used in several fields [3—5] but has rarely been applied in analytical chemistry. One exception is the papers by Seelig and Blount, who have described a successful application of the Kalman filter in anodic stripping voltammetry [6] and have recently compared the Kalman filter with non-recursive methods of estimation [7]. Another exception is an application to surveillance of surface water quality by Müskens [8].

In the present paper, the non-recursive least-squares algorithm, a very well known technique in analytical chemistry [9], is used as a starting point for applications of Kalman filtering to multicomponent analysis. Some new applications of Kalman filtering in analytical chemistry will be described, as well as a new criterion for the initial estimate for the error covariance matrix, which is of interest in many analytical applications. Consequences for the design of analytical experiments based on known theoretical results in Kalman filter theory are considered.

THEORY

A suitable analytical procedure is used to determine the composition of a sample, consisting of n components. The contribution of each of the components to the measured signal y is weighted by coefficients c_i, $i = 1, 2, \ldots n$; in the case of spectrophotometric measurements, for example, the c_i coefficients are molar absorptivities:

$$y = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n + v \tag{1}$$

where x_i ($i = 1, 2, \ldots, n$) is the unknown concentration of the component and v is the noise. Equation (1) can be written more concisely as

$$y = \mathbf{c}^T \mathbf{x} + v \tag{2}$$

where $\mathbf{x}$ is the n column vector with coefficients x_i, and $\mathbf{c}^T$ is the n row vector, the transpose of the column vector $\mathbf{c}$ with coefficients c_i. To determine the x_i values, m measurements are required:

$$y(k) = \mathbf{c}^T(k)\,\mathbf{x} + v(k) \quad (k = 1, 2, \ldots, m; m > n) \tag{3}$$

These m equations can be summarized in a single matrix vector equation:

$$\mathbf{Y}(m) = C(m)\,\mathbf{x} + \mathbf{V}(m) \tag{4}$$

where $\mathbf{Y}(m)$ is the m column vector with coefficients $y(k)$, $\mathbf{V}(m)$ is the m column vector with coefficients $v(k)$, and $C(m)$ is the $m \times n$ matrix with rows $\mathbf{c}^T(k)$.

The problem of finding the composition x_i — $i = 1, 2, \ldots, n$ — can be formulated as follows: given the measurements $y(1), \ldots, y(m)$, find an estimate $\bar{\mathbf{x}}$ of the parameter vector $\mathbf{x}$. The solution to this problem depends on the assumptions made, a priori statistical knowledge and the definition of a suitably chosen optimization criterion, such as cost function or performance index.

One assumption has already been made tacitly: the starting point is the linear eqn. (1). In system identification terms, this means that the model has been chosen [10]. It will also be assumed that the zero mean measurement noise $v(k)$ is statistically independent [11] of any of the other quantities involved. A data set $\{\mathbf{c}(k)\}$ obtained by some standard procedure is, of course, essential [12]. There are many situations in analytical chemistry which can be described adequately by this model. The least-squares method is the most extensively used method of data handling in analytical chemistry. Although a problem of a statistical nature is discussed here, statistics are essentially ignored. The problem formulated above is treated as a deterministic curve-fitting problem.

In view of eqn. (3) an obvious condition for a useful estimate $\bar{\mathbf{x}}$ of $\mathbf{x}$ will be that the error $e(k)$

$$e(k) \triangleq y(k) - \mathbf{c}^T(k)\bar{\mathbf{x}} \tag{5}$$

i.e. the difference between the measurement $y(k)$ and the "estimated measurement" $\mathbf{c}^T(k)\bar{\mathbf{x}}$, should be as small as possible for all k values. This can be achieved by choosing a weighted sum of the products $e(i)\,e(j)$, where $i, j = 1, \ldots, m$, as the cost function. The least-squares estimate $\bar{\mathbf{x}}(m)$ of $\mathbf{x}$, based on m measurements, minimizes this cost function, leading to the familiar normal equation [9]:

$$\bar{\mathbf{x}}(m) = \{C^T(m)R^{-1}(m)C(m)\}^{-1}\, C^T(m)R^{-1}(m)\mathbf{Y}(m) \tag{6}$$

(notation is explained in the Appendix) where $R^{-1}(m)$ is an $m \times m$ weighting matrix; an element $a_{ij}(m)$ of this matrix is the weight attached to the product $e(i)\,e(j)$. If

$$P(m) = \{C^T(m)R^{-1}(m)\,C(m)\}^{-1} \tag{7}$$

where $P(m)$ is an $n \times n$ matrix, eqn. (6) can be rewritten as

$$\bar{\mathbf{x}}(m) = P(m)\,C^T(m)R^{-1}(m)\mathbf{Y}(m) \tag{8}$$

From the above, it can be concluded that it will only be possible to "observe $\mathbf{x}$ through the measurement $\{y(k)\}$" if the inverse $P(m)$ exists.

Suppose that an additional measurement of $\mathbf{x}$ is obtained, given by

$$y(m+1) = \mathbf{c}^T(m+1)\,\mathbf{x} + v(m+1) \tag{9}$$

When this new observation is added to the existing data set, a new estimate $\bar{\mathbf{x}}(m+1)$ of $\mathbf{x}$ is obtained, based on $(m+1)$ measurements:

$$\bar{\mathbf{x}}(m+1) = P(m+1)\,C^T(m+1)\,R^{-1}(m+1)\,\mathbf{Y}(m+1) \tag{10}$$

$$P(m+1) = \{C^T(m+1)\,R^{-1}(m+1)\,C(m+1)\}^{-1} \tag{11}$$

where $R^{-1}(m+1)$ is the new weighting matrix for the $(m+1)$ observation problem. Again an $n \times n$ matrix must be inverted (eqn. 11), and none of the above calculations can be employed to facilitate the job. This waste of effort can be avoided by using a scheme of the structure

$$\text{new estimate} = (\text{known function of}) \text{ old estimate} + \text{correction} \tag{12}$$

The "old estimate" is the estimate based on m measurements, the "new estimate" is the estimate based upon $(m+1)$ measurements, while the "correction" is calculated on the basis of the new information supplied by the additional measurement. This recursive structure (12) can be derived from the non-recursive equations (10, 11) if it is assumed that the weighting matrix $R^{-1}(k)$ is a diagonal matrix for all k values, i.e. that there is no weighting of the errors between the "old" observations and the "new" observation. With this additional assumption the following recursive weighted least squares (RWLS) algorithm can be derived (see, e.g. [13]) from eqns. (11) and (12):

$$\bar{\mathbf{x}}(k+1) = \bar{\mathbf{x}}(k) + \mathbf{k}(k+1)\,\{y(k+1) - \mathbf{c}^T(k+1)\,\bar{\mathbf{x}}(k)\} \tag{13}$$

$$\mathbf{k}(k+1) = P(k)\,\mathbf{c}(k+1)\,\{r(k+1) + \mathbf{c}^{\mathrm{T}}(k+1)P(k)\,\mathbf{c}(k+1)\}^{-1} \tag{14}$$

$$P(k+1) = P(k) - \mathbf{k}(k+1)\,\mathbf{c}^{\mathrm{T}}(k+1)P(k) \tag{15}$$

(for $k = 0, 1, 2, \ldots, m$) where $r(k+1)$ is the weighting coefficient for the error in the $(k+1)$th measurement. (The vector product $\mathbf{k}(k+1)\,\mathbf{c}^{\mathrm{T}}(k+1)$ is explained in the Appendix.)

Instead of eqn. (15) the following equation is often used:

$$P(k+1) = \{1 - \mathbf{k}(k+1)\,\mathbf{c}^{\mathrm{T}}(k+1)\}\,P(k)\,\{1 - \mathbf{k}(k+1)\,\mathbf{c}^{\mathrm{T}}(k+1)\}^{\mathrm{T}} + \mathbf{k}(k+1)\,r(k+1)\,\mathbf{k}^{\mathrm{T}}(k+1) \tag{15a}$$

The numerical stability of eqn. (15a) is greater than that of eqn. (15) [15]. Equation (13) is of the form of eqn. (12). If

$$r(k) = 1 \text{ for all } k \tag{16}$$

(i.e., all measurements have weight "1"), this results in the important recursive ordinary least squares (ROLS) algorithm, described by Brubaker et al. [14]. Inspection of eqns. (13)—(15) shows that there is no longer a matrix inversion involved: $r(k+1) + \mathbf{c}^{\mathrm{T}}(k+1)P(k)\,\mathbf{c}(k+1)$ is a scalar. Moreover, the need for a growing computer memory is avoided. These features provide a strong argument in favour of the recursive algorithm.

By making the following identifications [15], $\mathbf{x} \approx$ random variable, and $r(k) \approx$ variance of the measurement noise $v(k)$, the recursive algorithm (13)—(16) represents the Kalman filter algorithm for the given problem of estimation. In this probability context, the matrix $P(k)$ is identified as the error covariance matrix of the difference between x and the estimate $\bar{\mathbf{x}}(k)$:

$$P(k) \triangleq E\{[\mathbf{x} - \bar{\mathbf{x}}(k)]\,[\mathbf{x} - \bar{\mathbf{x}}(k)]^{\mathrm{T}}\} \tag{17}$$

where $E\{\ \}$ denotes the expectation of the quantities within the brackets. This interpretation of the above least-squares assumption that there is no weighting between the old observations and the new observations means that the measurement noise is not correlated in time, i.e. it has no time history [15] and is white noise. The correction term in eqn. (13)

$$v(k+1) = y(k+1) - \mathbf{c}^{\mathrm{T}}(k+1)\,\bar{\mathbf{x}}(k) \tag{18}$$

is known as the innovation, in control literature [16]; heuristically it can be interpreted as the new information contained in the measurement $y(k+1)$.

If $\mathbf{x}$ and $v(k)$ are normally distributed, the Kalman filter gives the optimal estimate $\bar{\mathbf{x}}$ of $\mathbf{x}$. In the non-Gaussian case the Kalman filter gives the best linear estimate; optimal estimates will generally be obtained by a non-linear filter in this situation [17].

In contrast to the non-recursive algorithm, the recursive algorithm needs initial values to start the procedure: thus $\bar{\mathbf{x}}(0)$ is the estimate of $\mathbf{x}$ before processing of the measurements starts, and $P(0)$ is the error covariance matrix of the difference between $\mathbf{x}$ and $\bar{\mathbf{x}}(0)$, i.e. $E\{[\mathbf{x} - \bar{\mathbf{x}}(0)]\,[\mathbf{x} - \bar{\mathbf{x}}(0)]^{\mathrm{T}}\}$.

In absence of prior information, it is assumed that the filter is initiated with $\bar{\mathbf{x}}(0) = 0$ and $P(0) = \sigma_0^2 I_n$, where I_n is the $n \times n$ identity matrix [18, 19]. These generally incorrect initial statistics will cause the estimates $\{\bar{\mathbf{x}}(k)\}$ to be biased [13]. However, it can be shown [20] that this bias is asymptotically zero and smaller for larger initial variance σ_0; thus it is usually advisable to choose a large value for σ_0. If a particular choice of σ_0 leads to the inequality

$$\mathbf{c}^T(k+1)P(k)\,\mathbf{c}(k+1) \gg r(k+1) \tag{19}$$

for the first stage(s), i.e. if the numerical value of $r(k+1)$ can be completely neglected with respect to the numerical value of $\mathbf{c}^T(k+1)P(k)\,\mathbf{c}(k+1)$, it can be shown (see Appendix) that the matrix $P(k)$ becomes singular, i.e., it loses its statistical significance as a covariance matrix [13, 15]. This situation may occur if the coefficients $c_i(k)$ are very much larger than the true values x_i. Choice of the largest possible value for σ_0 is advocated in this case. Satisfactory results were obtained by choosing σ_0 according to

$$\sigma_0 = \alpha\,\{r(1)/[\mathbf{c}^T(1)\mathbf{c}(1)]\}^{\frac{1}{2}} \tag{20}$$

An example is given below. The coefficient α is directly proportional to the accuracy of the available computer and generally lies in the range 10—100.

In many practical situations, it is very difficult to account for systematic errors in the measurements [12]. A new method for estimating a constant systematic error b will be derived here; this application has been mentioned previously [7]. The original parameter vector $\mathbf{x}$ of dimension n is augmented to the $(n+1)$ dimensional parameter vector $\mathbf{x}^a$ as follows:

$$\mathbf{x}^a = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ b \end{pmatrix} \tag{21}$$

The n vector $\mathbf{c}(k)$ in the measurement eqn. (3) is augmented to $\mathbf{c}^a(k)$ as follows

$$\mathbf{c}^a(k) = \begin{pmatrix} c_1(k) \\ \vdots \\ c_n(k) \\ 1 \end{pmatrix} \tag{22}$$

The Kalman filter eqns. (13)—(15) are applied to the augmented system to estimate the composition x_i and the constant systematic error b simultaneously.

Situations often arise in which it is uncertain how many of n candidate components are present in the sample. Equations (13)—(15) for the n-component system are applied to the available measurements. If one of the components, e.g. the ith component is not present in the sample, the filter will produce a very small value for the ith coefficient (with respect to the other coefficients) of the estimated parameter vector $\bar{\mathbf{x}}(k)$.

Examples of these aspects are given under Results.

If the variance of the measurement noise is known for all k values, e.g.,

in the practically important case where the noise $v(k)$ can be considered to be a stationary process with known variance, then from eqns. (14) and (15) it can be seen that the Kalman gain vectors $\{\mathbf{k}(k)\}$ and the error covariance matrices $\{P(k)\}$ (for $k = 1, 2, \ldots, m$) can be calculated before the measurement process has actually started. $P(k)$ being a covariance matrix can serve as a measure of the quality of the estimate $\bar{\mathbf{x}}(k)$. The diagonal elements of $P(k)$ should be the variances in the estimates of the coefficients x_i of the parameter vector $\mathbf{x}$. Depending on the desired accuracy the number of measurements may be computed beforehand. However, it turns out in practice that $P(k)$ is only a boundary on the real error covariance matrix. It can be shown, e.g. [15, 21, 22], that by setting both the noise variance $r(k)$ (if $r(k)$ is not known in advance) and the initial error covariance matrix $P(0)$ greater than or equal to the true $r(k)$ and $P(0)$, respectively (cf. Appendix), then the computed $P(k)$ will be greater than or equal to the actual $P(k)$ for all k values. Based on the computed $P(k)$, a pessimistic answer should result concerning the required number of measurements in a certain experiment, if the selected values of both $P(0)$ and $r(k)$ are greater than their actual values. For more quantitative statements, this subject needs further research.

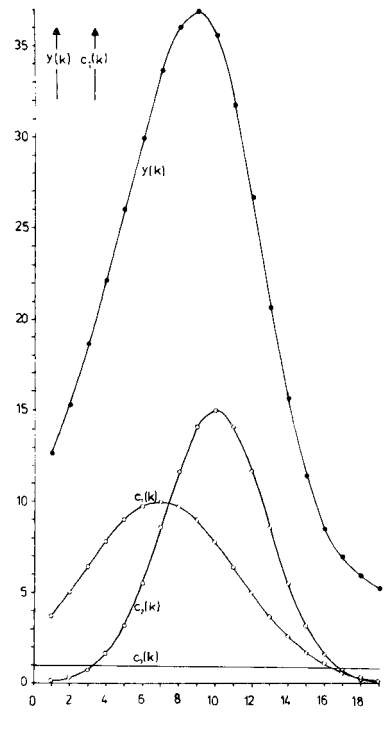
RESULTS

Example 1

Figure 1 shows a simulated spectrum of two components with unknown concentrations x_1 and x_2, as well as the coefficients $c_1(k)$ and $c_2(k)$ in eqn. (3). Although $c_1(19)$ and $c_2(19)$ are very small, $y(19)$ still has a significant value; thus a definite "background" b (systematic error) must be present. Hence the 2-dimensional parameter vector $\mathbf{x}$ is augmented to the 3-dimensional parameter vector $\mathbf{x}^a$, where the coefficient $x_3 = b$ represents the systematic error. The vector $\mathbf{c}(k)$ is augmented to $\mathbf{c}^a(k)$, where $c_3(k) = 1$ for all k values. The measurement noise is assumed to be stationary, zero mean, normally distributed white noise with variance 10^{-2}. The true value of $\mathbf{x}$ is $\mathbf{x}^T = (2.0, 1.0, 5.0)$.

By using eqns. (13)—(15), the concentrations x_1 and x_2 and the background b were estimated simultaneously; it was assumed that the background is constant during the measurements. The estimates $\bar{x}_i(k)$ for $i = 1, 2, 3$ and $k = 1, 2, \ldots, 19$, are shown in Fig. 2. $\bar{\mathbf{x}}(0) = 0$ and $P(0) = I_3$ (the 3×3 identity matrix) were chosen as initial values.

Tables 1 and 2 show some computed results. Table 2 shows that $k_1(1)$ is much larger than the other two coefficients $k_2(1)$ and $k_3(1)$. This corresponds to the fact that the relative contribution of the first component to the measurement $y(1)$, given by $c_1(1)$, is much greater than the relative contributions of both the second component and the systematic error, given by $c_2(1)$ and $c_3(1) = 1$; the height of the separate "peaks" $x_1c_1(1)$, $x_2c_2(1)$ and $x_3c_3(1) = b$ is of no importance. In accordance with this, Table 2 shows that



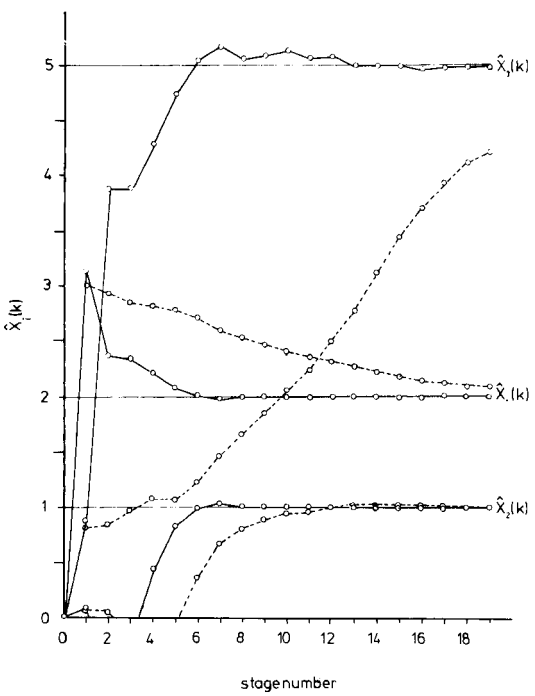


Fig. 1. Simulated spectrum and measurement coefficients. True value x: $x^T = (2.0\ 1.0\ 5.0)$.

Fig. 2. Estimated concentrations $\bar{x}_i(k)$ by using Kalman filtering (—○—) and ROLS (--○--) initial values: $\bar{x}(0) = 0$; $P(0) = I_3$. Measurement noise variance: $r(k) = 10^{-2}$ for all k. The true values are indicated by the horizontal lines.

TABLE 1

Estimates ($\bar{x}(k)$), measurement coefficients ($c_i(k)$) and measurements ($y(k)$) for Example 1 ($c_3(k) = 1$ for all k)

k	$\bar{x}_1(k)$	$\bar{x}_2(k)$	$\bar{x}_3(k)$	$c_1(k)$	$c_2(k)$	$y(k)$
0	0	0	0			
1	$0.3189*10^1$	$0.7799*10^{-1}$	0.8666	$0.3680*10^1$	$0.9000*10^{-1}$	$0.1262*10^2$
2	$0.2372*10^1$	$-0.1285*10^1$	$0.3875*10^1$	$0.4990*10^1$	0.2700	$0.1527*10^2$
3	$0.2333*10^1$	−0.2441	$0.3878*10^1$	$0.6410*10^1$	0.7100	$0.1870*10^2$
17	$0.2018*10^1$	0.9962	$0.4923*10^1$	0.6200	0.7100	$0.6942*10^1$
18	$0.2017*10^1$	0.9960	$0.4939*10^1$	0.3500	0.2700	$0.5976*10^1$
19	$0.2019*10^1$	0.9963	$0.4919*10^1$	0.1800	$0.9000*10^{-1}$	$0.5276*10^1$

the reduction in the estimated variance in the different estimates $\bar{x}_i(1)$, i.e., the diagonal elements $p_{ii}(1)$ of $P(1)$, is much larger for the first component than for the other two parameters. The estimates resulting from the ROLS

TABLE 2

Kalman gain coefficients ($k_i(k)$) and estimated variances ($p_{ii}(k)$) for Example 1

k	$k_1(k)$	$k_2(k)$	$k_3(k)$	$p_{11}(k)$	$p_{22}(k)$	$p_{33}(k)$
0				1	1	1
1	0.2527	$0.6181*10^{-2}$	$0.6868*10^{-1}$	$0.6992*10^{-1}$	0.9994	0.9313
2	0.5350	0.8923	$-0.1970*10^{1}$	$0.2176*10^{-1}$	0.8654	0.2783
3	$-0.7287*10^{-1}$	$0.1946*10^{1}$	$0.5600*10^{-2}$	$0.2109*10^{-1}$	0.3920	0.2782
17	$-0.2851*10^{-1}$	$-0.4015*10^{-2}$	0.2498	$0.7709*10^{-4}$	$0.2706*10^{-4}$	0.2728*
18	$-0.2376*10^{-1}$	$-0.3714*10^{-2}$	0.2085	$0.7004*10^{-4}$	$0.2689*10^{-4}$	0.2185*
19	$-0.2031*10^{-1}$	$-0.3162*10^{-2}$	0.1767	$0.6505*10^{-4}$	$0.2676*10^{-4}$	0.1808*

algorithm are also shown in Fig. 2. These estimates tend to be more conservative than those obtained by the Kalman filter. This can be explained by regarding the ROLS algorithm as a Kalman filter algorithm, where a noise variance of "1" instead of "10^{-2}" has been used; this means that the filter has been told that there is considerably more uncertainty in the measurements than there actually is.

Figure 3 shows the course of two diagonal elements of $P(k)$, $p_{11}(k)$ and $p_{22}(k)$, i.e. the estimated variances in the estimates $\bar{x}_1(k)$ and $\bar{x}_2(k)$ respectively, as a function of k, when two different initial values ($\sigma_0^2 = 1$ and $\sigma_0^2 = 50$) were used for $P(0)$ (cf. $P(0) = \sigma_0^2 I_n$). After only 7 measurements, the effect of the initial value has been damped out, showing the robustness of the algorithm for the initial estimate $P(0)$.

The reduction of the variance in $\bar{x}_2(k)$, the estimated concentration of the

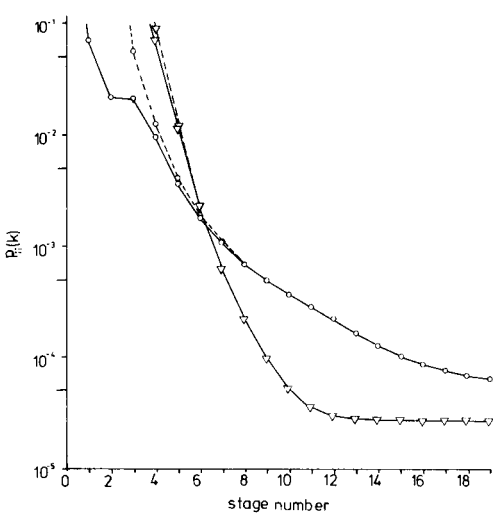


Fig. 3. Demonstration of effect of initial value for the estimated variances $p_{11}(k)$ (○) and $p_{22}(k)$ (▽) in the estimates $\bar{x}_1(k)$ and $\bar{x}_2(k)$, respectively. (—) $P(0) = I_3$ (---) $P(0) = 50 \times I_3$.

second component after 19 measurements, is larger than the reduction in the variance in the estimate $\bar{x}_1(k)$. This is because the overall relative contribution of the second component to the measured signal $y(k)$ is greater than the contribution of the first component; the separate peaks $x_1c_1(k)$ and $x_2c_2(k)$ are of no importance.

Example 2

Figure 4 shows the u.v. spectrum of a mixture of aniline (concentration x_1, absorptivity $c_1(\lambda)$), nitrobenzene ($x_2, c_2(\lambda)$), azobenzene ($x_3, c_3(\lambda)$) and azoxybenzene ($x_4, c_4(\lambda)$). It is known that the spectrophotometer produces a signal that is corrupted by stationary, zero-mean white noise with variance 10^{-6}. Because the absorptivities are of order 10^{+4} (see Figs. 4, 5) while the noise variance is 10^{-6}, this is a typical situation in which eqn. (20) has to be applied. If α in eqn. (20) is selected as 10, then $P(0) \approx 10^{-12} \times I_4$, where I_4 is the 4×4 identity matrix. Measurements were made every 2 nm.

Table 3 shows the estimates of the concentrations of the four components at different wavelengths. Figure 5(A) shows the course of the estimated concentrations $\bar{x}_i(\lambda)$ as a function of λ. The estimated variances $p_{ii}(\lambda)$ in the estimates $\bar{x}_i(\lambda)$ — $i = 1, 2, 3, 4$, are also given in Table 3. An interesting feature of the recursive estimation method becomes clear on comparing the two parts of Fig. 5. The estimation is "faster" when the u.v. spectrum is scanned from lower to higher wavelengths, than in the conventional reverse direction, mainly because in the first case all the components contribute immediately to the measured signal $y(\lambda)$; cf. the absorptivities in Fig. 4. Figure 5B also

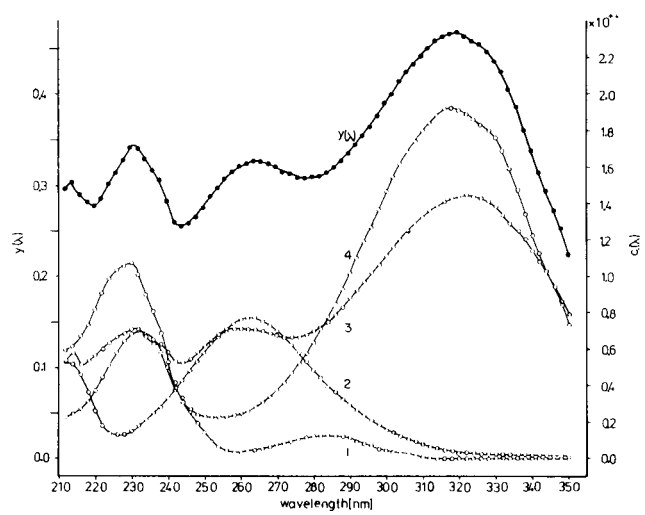


Fig. 4. U.v. spectrum of a mixture of aniline (1), nitrobenzene (2), azobenzene (3) and azoxybenzene (4). $c_i(\lambda)$ are the absorptivities of these components expressed in l mol^{-1} cm^{-1}. Curves 1—4 refer to the right-hand axis.

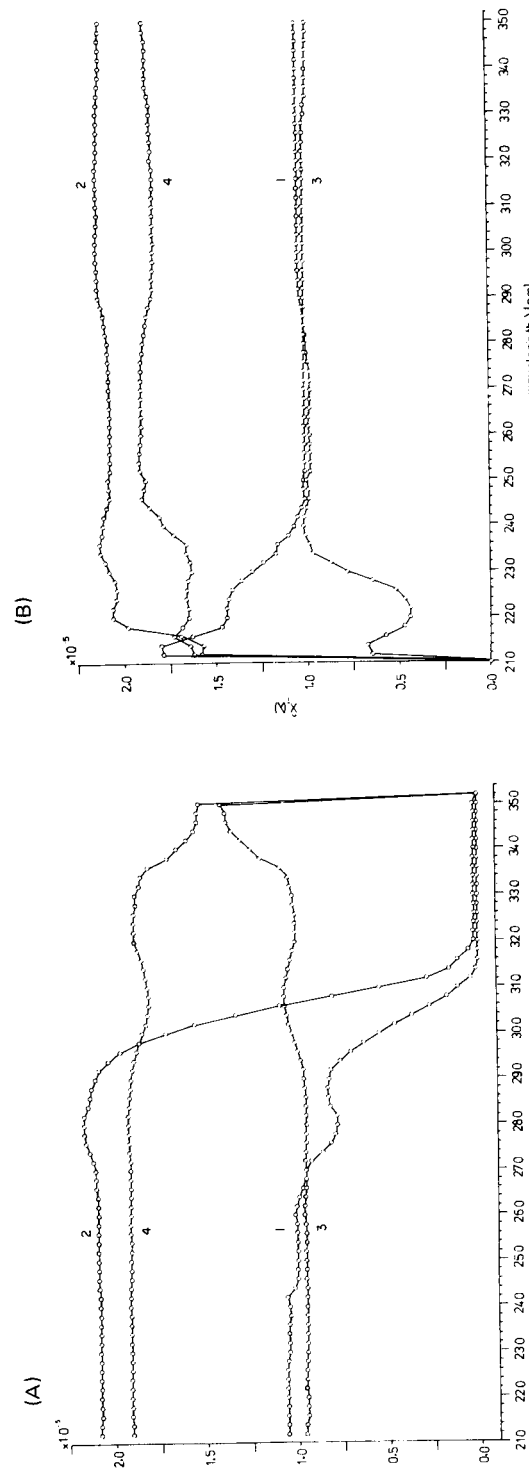


Fig. 5. Estimated concentrations ($\bar{x}_i(\lambda)$ in mol l^{-1}) of aniline (1), nitrobenzene (2), azobenzene (3) and azoxybenzene (4). (A) Measurements processed from higher to lower wavelengths. Initial estimates: $\bar{\mathbf{x}}(0) = \mathbf{0}$; $P(0) = 10^{-12} I_4$. Noise variance $r(k) = 10^{-6}$ for all λ. (B) Measurements processed from lower to higher wavelengths. Initial estimates: $\bar{\mathbf{x}}(0) = \mathbf{0}$; $P(0) = 10^{-12} I_4$. Noise variance $r(k) = 10^{-6}$ for all λ.

TABLE 3

Estimated concentrations of aniline ($\bar{x}_1(\lambda)$), nitrobenzene ($\bar{x}_2(\lambda)$), azobenzene ($\bar{x}_3(\lambda)$) and azoxybenzene ($\bar{x}_4(\lambda)$) at different wavelengths, and the corresponding estimated variances ($p_{ii}(\lambda)$)

λ	$\bar{x}_1(\lambda)$	$\bar{x}_2(\lambda)$	$\bar{x}_3(\lambda)$	$[\bar{x}_3(\lambda)]$[a]	$\bar{x}_4(\lambda)$	$p_{11}(\lambda)$	$p_{22}(\lambda)$	$p_{33}(\lambda)$	$p_{44}(\lambda)$
352	0	0	0	[0]	0	10^{-12}	10^{-12}	10^{-12}	10^{-12}
350	$-0.2745*10^{-7}$	$0.1764*10^{-6}$	$0.1396*10^{-4}$	$[0.9553*10^{-5}]$	$0.1517*10^{-4}$	$0.1010*10^{-11}$	$0.1010*10^{-11}$	$0.5508*10^{-12}$	$0.4679*10^{-12}$
348	$-0.1347*10^{-7}$	$0.1694*10^{-6}$	$0.1375*10^{-4}$	$[0.8890*10^{-5}]$	$0.1528*10^{-4}$	$0.1010*10^{-11}$	$0.1010*10^{-11}$	$0.5042*10^{-12}$	$0.4563*10^{-12}$
346	$-0.1474*10^{-7}$	$0.1696*10^{-6}$	$0.1371*10^{-4}$	$[0.8403*10^{-5}]$	$0.1531*10^{-4}$	$0.1010*10^{-11}$	$0.1010*10^{-11}$	$0.4576*10^{-12}$	$0.4329*10^{-12}$
344	$-0.2484*10^{-7}$	$0.1551*10^{-6}$	$0.1343*10^{-4}$	$[0.7828*10^{-5}]$	$0.1552*10^{-4}$	$0.1010*10^{-11}$	$0.1010*10^{-11}$	$0.4268*10^{-12}$	$0.4149*10^{-12}$
218	$0.1058*10^{-4}$	$0.2079*10^{-4}$	$0.9532*10^{-5}$	$[-0.8295*10^{-7}]$	$0.1911*10^{-4}$	$0.4163*10^{-14}$	$0.7890*10^{-14}$	$0.8907*10^{-14}$	$0.1534*10^{-13}$
216	$0.1058*10^{-4}$	$0.2079*10^{-4}$	$0.9535*10^{-5}$	$[-0.8515*10^{-7}]$	$0.1911*10^{-4}$	$0.4134*10^{-14}$	$0.7283*10^{-14}$	$0.8301*10^{-14}$	$0.1427*10^{-13}$
214	$0.1057*10^{-4}$	$0.2085*10^{-4}$	$0.9584*10^{-5}$	$[-0.3893*10^{-7}]$	$0.1904*10^{-4}$	$0.4122*10^{-14}$	$0.6956*10^{-14}$	$0.8049*10^{-14}$	$0.1383*10^{-13}$
212	$0.1057*10^{-4}$	$0.2085*10^{-4}$	$0.9583*10^{-5}$	$[-0.4313*10^{-7}]$	$0.1904*10^{-4}$	$0.4091*10^{-14}$	$0.6526*10^{-14}$	$0.7687*10^{-14}$	$0.1319*10^{-13}$

[a]Estimated concentration of azobenzene when it was not present in the sample.

shows that measurements made above 290 nm have not changed the estimates $\bar{x}_i(\lambda)$ very much. This again shows that only the relative contributions of the components (given by the $c_i(\lambda)$ values) are of importance. The estimates $\bar{x}_i(\lambda)$ in Fig. 5 do not differ significantly, in the statistical sense. That the actual numerical values slightly differ is chiefly due to the fact that different independent sets of measurements $\{y(\lambda)\}$ were used for Fig. 5A and 5B to avoid the introduction of any dependence between the two sets of measurements. Another reason for this numerical difference is that the Kalman filter is sensitive to computer round-off errors [4].

Further, the Kalman filter was applied to a data set $\{y(\lambda)\}$, which was obtained from a sample that possibly contained all four components. Figure 6 shows that the third component (azobenzene) was not present in the sample. The numerical values for the estimated concentrations of azobenzene ($\bar{x}_3(\lambda)$) obtained in this case are given in brackets in Table 3. Because the variance in the measurement noise was again 10^{-6} and the same initial value was used for $P(0)$ ($10^{-12}\, I_4$) the estimated variance in this estimate can be found in Table 3 ($p_{33}(\lambda)$).

DISCUSSION

The estimates converge rapidly to invariant values (Figs. 2, 5, 6). By taking the square root of the estimated variances $p_{ii}(19)$ ($i = 1, 2, 3$) given in Table 2, the estimated standard deviations $\sigma_{ii}(19)$ ($i = 1, 2, 3$) are obtained. A comparison between the estimates $\bar{x}_i(19)$ and the true values x_i (Table 1) reveals that $\bar{x}_1(19)$ and $\bar{x}_3(19)$ are not within the $1\sigma_{11}(19) - 1\sigma_{33}(19)$ limit, respectively, but within the respective $2\sigma_{ii}$ limits. As the initial value effect has

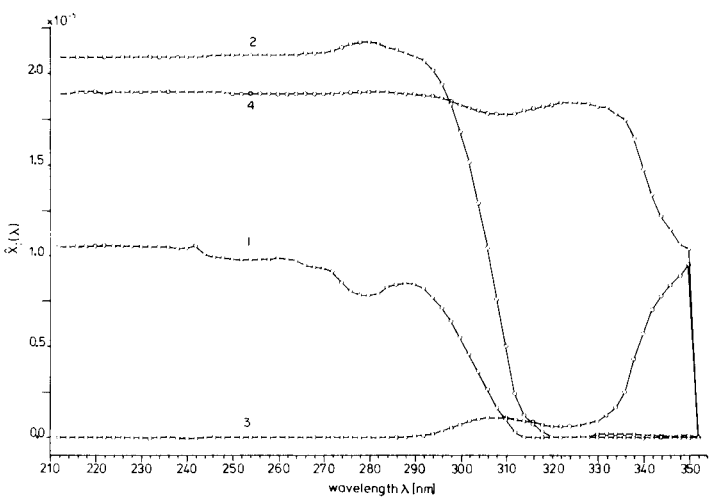


Fig. 6. Determination of number of components; azobenzene (estimated concentration $\bar{x}_3(\lambda)$) was not present in the sample. Symbols, initial estimates and noise variance as in Fig. 5.

been damped out (Fig. 3), this probably happens because x is not normally distributed [17]. Further research is necessary on this point.

The examples show that for separation of the peaks, geometrical (visual) criteria are not necessary, e.g. the simulated spectrum of Fig. 1 shows only one peak, which is a combination of two peaks, but excellent separation is still obtained (Fig. 2). For separation, it is necessary that the data set $\{c(k)\}$ contains a sufficient number of linear independent (see Appendix) subsets; this is obviously an algebraic criterion. In control engineering terms, this condition leads to the "observability condition" [13, 15], for which measurements $\{y(k)\}$ are of no importance compared to visual criteria. This statement agrees with the fact that neither the updating equation for the error covariance matrix (eqn. 15) nor the equation for the Kalman gain (eqn. 14) is a function of the measurements $\{y(k)\}$. The existence of the inverse $P(m)$ (eqn. 7) is another way of formulating the observability condition.

Theory and experimental results clearly indicate that the Kalman filter algorithm may furnish a new approach to experimental measurement design: estimation of the required number of measurements, and scanning that part of the spectrum where all components contribute to the signal, as indicated by the coefficients $c_i(k)$. Example 2 shows that these parts are not necessarily the high peaks of the spectrum.

Finally, the Kalman filter algorithm, being recursive,calls for on-line computation, i.e. a direct coupling of measurement device and computer. When the desired accuracy is reached, measurements can be stopped. It would be convenient if the computer controlled the measurement device in this way.

APPENDIX

Transposition. The superscript "T" denotes the transpose of a matrix or vector. This linear operation can be formulated as a "changing of rows and columns". A column vector becomes a row vector, and an $n \times m$ matrix, an $m \times n$ matrix on transposition.

Inversion. The superscript "—1" denotes the inverse of an $n \times n$ matrix. The inverse is only defined for square matrices, having linear independent (see below) columns (rows). The inverse of an $n \times n$ matrix A is given by $AA^{-1} = A^{-1}A = I_n$.

Vector products. If x and y are n column vectors with coefficients x_i and y_i, respectively, then $\mathbf{x}^T \mathbf{y} \triangleq \sum_{i=1}^{n} x_i y_i$. This equation is a scalar

$$\mathbf{x}\mathbf{y}^T = \begin{pmatrix} x_1 y_1 & \dots & x_1 y_n \\ \vdots & & \\ x_n y_1 & \dots & x_n y_n \end{pmatrix}$$

which is an $n \times n$ matrix.

Linear independence. A set of n column vectors $\{\mathbf{x}i\}$ is said to be linearly

independent if $\alpha 1 \mathbf{x} 1 + \alpha 2 \mathbf{x} 2 + \ldots + \alpha n \mathbf{x} n = 0$ implies $\alpha i = 0$ for $i = 1, 2, \ldots, n$.

Symmetry. An $n \times n$ matrix A is called symmetric if $A = A^{\mathrm{T}}$.

Inequality for matrices. If A and B are two $n \times n$ matrices, then $A > B$ means that the matrix $(A - B)$ is positive definite [23], i.e. $\mathbf{x}^{\mathrm{T}}(A - B)\,\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$.

Singularity of P(k). Combining eqns. (14) and (15) and assuming (19) holds, gives:

$$P(k+1) = P(k) - P(k)\,\mathbf{c}(k+1)\,\{c^{\mathrm{T}}(k+1)\,P(k)\,\mathbf{c}(k+1)\}^{-1}\,\mathbf{c}^{\mathrm{T}}(k+1)\,P(k)$$

From this expression it is easy to show that $\mathbf{c}^{\mathrm{T}}(k+1)\,P(k+1)\,\mathbf{c}(k+1) = 0$. Hence the symmetric matrix $P(k+1)$ must be singular [23]. This also implies that the Kalman filter cannot be used if measurement noise is absent.

The author thanks Mr. C. Didden for his computational assistance and his suggestions concerning formula (20), and Prof. drs. G. Kateman for his critical remarks.

REFERENCES

1 N. Wiener, Time Series, MIT Press, Cambridge, 1975.
2 R. E. Kalman, J. Basic Eng., 82 (1960) 35.
3 D. E. Seborg, D. Grant Fischer and J. C. Hamilton, Automatica, 11 (1975) 351.
4 G. J. Bierman and C. L. Thornton, Automatica, 13 (1977) 23.
5 A. K. Sinha and J. P. Sharma, Int. J. Syst. Sci., 6 (1975) 681.
6 P. F. Seelig and H. N. Blount, Anal. Chem., 48 (1976) 252.
7 P. F. Seelig and H. N. Blount, Anal. Chem., 51 (1979) 327.
8 P. J. W. M. Müskens, Anal. Chim. Acta, 103 (1978) 445.
9 See, e.g., D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
10 P. Eijkhoff, System Identification, Wiley, London, 1974.
11 W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, Wiley, New York, 1968.
12 J. R. DeVoe (Ed.), Validation of the Measurement Process, ACS Symp. Ser. 63, American Chemical Society, New York, 1977.
13 A. P. Sage and J. L. Melsa, Estimation Theory with Applications to Communications and Control, McGraw-Hill, New York, 1971.
14 T. A. Brubaker, R. Tracy and C. L.Pomernacki, Anal. Chem., 50 (1978) 1017.
15 A. H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.
16 T. Kailath, IEEE Trans. Autom. Control, AC-13 (1968) 646.
17 See, e.g., S. K. Park and D. G. Lainoitis, Int. J. Control, 16 (1972) 1029.
18 Vincent J. Aidala, IEEE Trans. Autom. Control, AC-22 (1977) 471.
19 D. Graupe, Identification of Systems, R. E. Krieger, New York, 1976.
20 Allen Klinger, IEEE Trans. Autom. Control, AC-13 (1968) 102.
21 T. Nishimura, IEEE Trans. Autom. Control, AC-11 (1966) 197.
22 T. Nishimura, IEEE Trans. Autom. Control, AC-12 (1967) 123.
23 F. R. Gantmacher, The Theory of Matrices, Vol. 1, Chelsea, New York, 1960.

Analytica Chimica Acta, 112 (1979) 375—383
Computer Techniques and Optimization

AN ALGORITHM FOR THE COMPUTATION OF AQUEOUS MULTICOMPONENT, MULTIPHASE EQUILIBRIA

GUNNAR ERIKSSON

Department of Inorganic Chemistry, University of Umeå, S-901 87 Umeå (Sweden)

(Received 22nd June 1979)

SUMMARY

The free-energy minimization method is applied to systems containing liquid solutions with a solvent of unit activity, a gas phase of constant volume, and solid phases of invariant stoichiometry. The equations derived are employed in a computer program, SOLGASWATER.

The existing general procedures for the computation of chemical equilibrium compositions are traditionally divided into two main categories: optimization techniques, including free-energy minimization, and techniques utilizing non-linear equations such as those based on the law of mass action. The mathematical structures of various algorithms belonging to these groups have been reviewed in a monograph by van Zeggeren and Storey [1], and also recently by Smith [2]. These algorithms are primarily concerned with chemical equilibrium within a homogeneous gaseous mixture or, in some instances, between a gas phase and pure condensed phases. SOLGASMIX [3] is an example of a computer program for the calculation of the composition of an equilibrium mixture. In SOLGASMIX, the total free energy of a system containing a gas phase and condensed phases is minimized according to the method of steepest descent as originally formulated by White et al. [4]. SOLGASMIX is used extensively, and has been found fast and reliable for complicated multiphase high-temperature systems.

Since the work of Sillén and co-workers [5] on the computation of the equilibrium composition of liquid solutions, there has been increasing interest in this field, with particular emphasis on systems of analytical and geochemical importance. In the papers published so far, algorithms based on mass action expressions are the most common (see, e.g. [5—11]). Free-energy minimization seems to have been used only by Karpov and Kaz'min [12] and by Gautam and Seider [13]. In the two latter investigations, a liquid solution is considered as a phase in which the activity for the solvent is related to a mole fraction scale, whereas the activities for the solutes are related to a molality scale.

In an alternative model, which is usually applicable when dealing with dilute solutions or solutions in an ionic medium, the solvent is assumed to be of constant unit activity and need not be regarded as an independent component of the system, even if it takes part in chemical reactions. Although it can be applied to most liquid solutions encountered in practice, this solution model has not yet been used as the basis of a free-energy minimization algorithm. It was therefore thought worthwhile to formulate the minimization method for systems containing one or several liquid solutions, a gas phase of constant volume, and solid phases, and to assume unit activities for the solvents and with activities for the solutes proportional to their molar concentrations. This formulation is computationally more efficient than those described earlier [12, 13] because of the smaller number of variables resulting when the mole fraction of the solvent is constant. Several sources of equilibrium data for association of inorganic and organic ligands with metallic ions and the hydrogen ion in solutions where the model is applicable are available [14—16]; these compilations find the most widespread use among solution chemists.

The equations derived here are implemented in the computer program SOLGASWATER.

COMPUTATIONAL PROCEDURE

The free-energy minimization method is based on the thermodynamic criterion that the total free energy of a system at constant temperature and pressure has a minimum value at equilibrium. In the following treatment, multiphase systems with one or several liquid solutions are considered, each of which contains a solvent of unit activity and a variable number of solutes, a gas phase of constant volume, and a number of solid phases. The solutes are divided into components and complexes, where the components must be chosen so that the composition of a reactive system is unambiguously determined from equilibrium constants of formation and the component concentrations at equilibrium. The minimization is constrained in that the equilibrium concentrations must be non-negative and must satisfy the mass-balance equations. Without regard to the former constraint, the problem is to minimize an objective function with linear equality constraints, and the Lagrange method of undetermined multipliers is suitable for this purpose.

Basic equations

In the first description of the free-energy minimization method, White et al. [4] pointed out that a Lagrangian multiplier π_j, of which there is one for each of the N_c components, represents the free-energy contribution of the presence of b_j mol of the jth component. The total free energy G can thus be expressed as a non-constrained function

$$G = RT \sum_{j=1}^{N_c} \pi_j b_j \qquad (1)$$

where R is the gas constant and T is the absolute temperature. The mass-balance equations, which serve as subsidiary conditions in the minimization, can be written

$$b_j = \sum_{i=1}^{N_t} a_{ij} c_i \qquad (j = 1, 2, \ldots, N_c) \tag{2}$$

where N_t equals the sum of N_f and N_s, N_f is the total number of solutes in the fluid (liquid and gaseous) phases, N_s is the number of solid phases assumed to be present at equilibrium, and a_{ij} is the amount of the jth component in 1 mol of the ith species. In this notation, c_i is used for the amount of a solute or gaseous species in a fluid phase of volume 1 dm^3 or the amount of a solid phase in equilibrium with fluids of the same volume. The absolute values for the volumes are only of importance in connection with the prediction of distribution equilibria. Combination of eqns. (1) and (2) gives

$$G = \sum_{i=1}^{N_t} \left(RT \sum_{j=1}^{N_c} \pi_j a_{ij} \right) c_i \tag{3}$$

The total free energy and the molar concentrations are also related by the basic thermodynamic function

$$G = \sum_{i=1}^{N_t} (\mu_i^0 + RT \ln a_i) c_i \tag{4}$$

where μ_i^0 denotes the chemical potential in the standard state and a_i is the relative activity. The activity for a solute (or a gaseous) species is by definition equal to $f_i \cdot c_i$, where f_i is an activity coefficient which approaches unity when the composition of the solution approaches that of the solvent. The activities for the solid phases, which are considered to be of invariant stoichiometry, are equal to unity. Comparison of eqns. (3) and (4) leads to the generally valid expressions

$$\sum_{j=1}^{N_c} \pi_j a_{ij} - (\mu^0/RT)_i - \ln f_i - \ln c_i = 0 \qquad (i = 1, 2, \ldots, N_f) \tag{5}$$

$$\sum_{j=1}^{N_c} \pi_j a_{ij} - (\mu^0/RT)_i = 0 \qquad (i = N_f + 1, N_f + 2, \ldots, N_t) \tag{6}$$

where $(\mu^0/RT)_i$ is a thermodynamic quantity for which the value must be available. Parametrized activity coefficients for the species in the non-ideal phases are supplied by the user.

The system consisting of the $(N_c + N_t)$ equations (2), (5) and (6) with the variables π_j $(j = 1, 2, \ldots, N_c)$ and c_i $(i = 1, 2, \ldots, N_t)$ is non-linear, and the next step is a linearization of eqn. (5) by expansion in a Taylor series around an estimated equilibrium composition y_i $(i = 1, 2, \ldots, N_t)$ up to and including the term of the first order. This is equivalent to making a quadratic

approximation to the free-energy surface. The y_i values must be non-negative, but need not satisfy the mass-balance constraints. On the assumption that f_i is independent of c_i, the expansion results in the equation

$$c_i = -\psi_i + y_i \sum_{j=1}^{N_c} \pi_j a_{ij} \qquad (i = 1, 2, \ldots, N_f) \tag{7}$$

where $\psi_i = y_i \left[(\mu^0/RT)_i + \ln f_i + \ln y_i - 1\right]$

The partial differentiation of eqn. (5) with respect to c_i is approximate for solutes with non-constant activity-coefficient expressions. Equation (7) becomes, however, identical with eqn. (5) when the y_i values correspond to a free-energy minimum. Incorporation of eqn. (7) into eqn. (2), which reduces the number of equations in the system to $(N_c + N_s)$, gives

$$\sum_{k=1}^{N_c} r_{jk}\pi_k + \sum_{i=N_f+1}^{N_t} a_{ij}c_i = b_j - \sum_{i=1}^{N_f} a_{ij}\psi_i \qquad (j = 1, 2, \ldots, N_c) \tag{8}$$

where $r_{jk} = \sum_{i=1}^{N_f} a_{ij}a_{ik}y_i \qquad (j, k = 1, 2, \ldots, N_c)$

By using eqn. (7) and the numerical solution of the system of linear equations (6) and (8), values of c_i $(i = 1, 2, \ldots, N_t)$ are obtained. The free-energy surface approximation is implicit in the iterative algorithm and, if positive, the c_i values are used as improved estimates in the subsequent iteration cycle.

In the computational procedure outlined, the solvents are considered as species with unit activity which do not participate in the mass balances. Neither of the total number of moles in the fluids and π_j for the solvent components are unknowns (cf. [12, 13]), and the number of simultaneous linear equations is consequently reduced by twice the number of fluids assumed to be present at equilibrium.

The iteration process

In the early stage of the iteration process and especially when the estimated equilibrium concentrations are poor, negative c_i values can occur frequently. Because of the term $y_i \ln y_i$ in eqn. (7), the calculated concentrations must be transformed before being used as the starting-point for a new Taylor expansion. Improved non-negative estimates which satisfy the mass-balance constraints, provided that the y_i values do, are calculated by the relationship

$$y_i' = y_i + \lambda_i (c_i - y_i) \qquad (i = 1, 2, \ldots, N_f) \tag{9}$$

where λ_i is a relaxation parameter. The non-negativity constraints are satisfied for all values of λ between zero and λ', where λ' is the minimum value of the quotient $y_i/(y_i - c_i)$ for the solutes with negative c_i values.

In the computer program SOLGASWATER, λ is replaced with the value of the arithmetic expression $0.999\lambda' (1 - 0.5\lambda')$, i.e., a y_i value can decrease maximally three powers of ten between two successive iteration cycles. With the object of avoiding excessively long computing times caused by bad initial estimates, a lowest allowed y_i value is adopted, viz. 10^{-14} mol dm^{-3}. A solute for which the concentration becomes less than this value does not significantly affect the mass balances, and is temporarily removed from the calculation. Since the numerical solution of the simultaneous linear equations is independent of the y_i values for the solid phases, these are put equal to the absolute value of c_i.

The solids included in the initial estimate need not necessarily be the correct ones of the final equilibrium state. Another phase combination might yield a lower free energy, and solids must be withdrawn from or added to the previous set until the equilibrium phases are found. This set has the characteristic feature that the value for the left-hand side of eqn. (6) is less than zero or, equivalently, that the solubility product is not exceeded for the omitted phases, and the search for the equilibrium phases proceeds until these inequalities are satisfied. The only variable quantity in eqn. (6) is π_j and the validity of the inequalities is therefore most appropriately tested as soon as the π_j values are essentially unchanged in two successive iteration cycles. As a result of the test, the solid phase with the maximum positive value of the left-hand side of eqn. (6) is added to the old set. A solid phase is withdrawn if its amount becomes equal to zero or persists in going negative in the course of the iteration process. Furthermore, violation of the Gibbs phase rule also induces withdrawal of phases. When the phase assembly is changed, all temporarily removed species in the fluids are reconsidered. Tests are also made to ascertain whether or not the current combination of solids has been previously considered, otherwise the calculation may enter a mode where it operates in the form of an infinite loop.

The iteration process is terminated when the equilibrium phases are found and when the concentrations for all included species are unchanged and positive in two consecutive iteration cycles. For the purpose of obtaining a complete equilibrium composition, the concentrations for the trace constituents removed are calculated by using eqn. (5) and the equilibrium π_j values. A Newton—Raphson iteration is employed if the activity-coefficient expression f_i is dependent on the value of c_i.

Experience with this algorithm for handling the solid phases has shown it to work successfully, even for bad initial estimates and complicated systems. In this respect, the shape of the free-energy surface and the existence and uniqueness of a solution are important considerations. This topic has been reviewed by Smith [2]. For systems containing ideal fluid and solid phases it can be established that the free-energy surface does not possess local minima, and that a computed equilibrium composition corresponds to a global minimum. For non-ideal systems, however, activity-coefficient expressions entered through a user-written subroutine are used, and uniqueness cannot be guaranteed.

Input variables

The solutes which can occur in significant concentrations at equilibrium and the solid phases which may appear must be forecast prior to the calculation. As the number of solutes does not influence the number of variables in the system of linear equations, it is recommended that too many solutes be considered rather than too few. An increased number of solid phases strongly reduces the probability of the initial estimate containing the correct phases, and hence the average number of iteration cycles necessary to obtain an equilibrium composition as well as the computing time will increase. A certain caution in the inclusion of extra solids must therefore be observed.

It is not possible to give explicitly an absolute value of the chemical potential. In order to assign numerical values to the quantity (μ^0/RT), a convention must therefore be adopted. The (μ^0/RT) values for the components can, as a matter of principle, be chosen arbitrarily since only differences between (μ^0/RT) values have significance when a formation reaction is considered. However, it is most practical to choose (μ^0/RT) to be zero for all components at all temperatures. Since $\mu^0 = G^0$ for a pure compound, it can be shown that $\Delta(\mu^0/RT)$ for the formation of a species from its components, and thus also (μ^0/RT), equals $-\ln \beta$ where β is a cumulative stability constant. To all species must be assigned properly selected stability constants because of their dependence on the temperature, pressure, solvent, and the concentration of a medium. A complex simultaneously present in two immiscible phases must be regarded as two distinct chemical species. If a stability constant is unknown for a species it must be estimated, or the species disregarded.

When distribution equilibria are considered, the coexisting phases are assumed to have equal volumes. Should the volumes be unequal, say V and V', the mass-balance equations are influenced and the β values for all species in the primed phase as well as the total concentrations for all components in the primed phase must be multiplied by the ratio V'/V. For a species in an ideal gas phase of constant volume, present together with a liquid phase, the β value can be given either in concentration or pressure units. If given in pressure units, it must be divided by RT to make the dimensions in the mass balances consistent.

The total concentrations can be given explicitly or calculated from titration data. If the total concentrations satisfy the electroneutrality equation, so do the equilibrium concentrations. In order to maintain electroneutrality in a system where redox equilibria take place, the total concentration of the electrons must be ultimately adjusted to zero. Sometimes, a free-component activity is known instead of the total concentration, e.g. pH instead of the total concentration for H^+. As indicated by eqn. (5), a known activity for the jth component corresponds to a known value of π_j and the number of variables in the system of equations is accordingly reduced. It should be noted that a solid phase cannot appear at equilibrium if the activities of its components are preset to particular values (cf. eqn. 6).

SOME COMMENTS ON SOLGASWATER

SOLGASWATER is a comparatively small, publicly available program which is written in FORTRAN IV for a computer CDC CYBER 172, in such a way that only minor changes will be necessary to run the program on other machines. The program contains totally around 900 statements in one main program and two subroutines, and is dimensioned for a maximum of nine components, 50 species, and 99 equilibrium compositions in sequence. The main program and one subroutine contain all input/output statements, of which the plotting output is designed for a Benson plotter. The other subroutine contains the search procedure for the equilibrium phases and the procedures for the numerical solution of the simultaneous linear equations: Choleski factoring if no solids are included in the initial estimate and Gaussian elimination with partial positioning for size if solids are included.

Program output

Once the equilibrium compositions corresponding to a group of points of input concentrations are computed, the results can be presented either in the form of a table or as diagrams. The table is edited by the user himself and can contain, besides equilibrium concentrations (c) and activities (a) and the logarithms of these quantities, also some derived quantities. These include:
(a) the percentage distribution of a component between the species in the fluid phases (F);
(b) the total concentration for a component (T_c) or the total concentration in the fluid phases for a component (T_f) or the logarithm of their absolute values;
(c) the logarithm of the partial pressure for a gaseous species in an ideal gas phase of volume 1 dm^3 at a specified temperature;
(d) the average amount of one component, regarded as a ligand, bound to unit amount of another component, regarded as the central atom, in a specified phase ($\bar{n}$);
(e) the η value for a component ($\eta = \log (T_f/c)$, c = concentration for the free component species).
The diagrams are generated by the computer via the plotting procedure. The variables which can be plotted along the ordinate are: F values, the accumulated sum of F values, calculated log a values (log a is put equal to log c for a solid phase), $\bar{n}$ values, and η values. The variables which can be plotted along the abscissa are: input values of $\pm \log a$ for a free component species (e.g. log $[H^+]$ or pH), input values of $\pm T_c$ or $\pm \log T_c$, volume added in titration, and calculated log c values for a species. For the construction of diagrams of various types, such as distribution diagrams, pH diagrams, solubility diagrams, $\bar{n}$ diagrams, titration curves, and so forth, the proper dependent and independent variables are combined. An $\bar{n}$ diagram can contain information from fifteen consecutive groups of points, whereas the remaining diagrams can contain information from one group only.

An alternative possibility for presenting calculated results in a compact form is to let the computer sketch a predominance area diagram with ± log a or ± log T_c as axes. These diagrams indicate the solid phases present for every considered x—y pair or, if no solids are present, the predominating species containing a specified component.

Computational efficiency

The computer time needed to calculate an equilibrium composition depends on the complexity of the system and then primarily on the total number of solid phases considered. The greater this number, the more combination alternatives of solid phases have on average to be investigated until the correct one is found, and the greater will be the variance of the computing time. For a very simple system with two components and six solutes, the computer (CDC CYBER 172) calculates around 40 equilibrium compositions per second. For a more complicated system with nine components and 39 solutes, the capacity is reduced to around two equilibrium compositions per second. Replacement of some of the solutes with solids would lead to an increased computing time and the figures given then represent a measure of the minimum computing time for systems of corresponding complexity.

Applications

Because of its flexibility and generality, SOLGASWATER provides a valuable tool for the study of the chemistry of multicomponent, multiphase aqueous systems. In connection with the evaluation of several ternary complex models in equilibrium analysis, the plotting routine of SOLGASWATER has been utilized to draw different kinds of diagrams, e.g., $\bar{n}$ diagrams distribution diagrams, and predominance area diagrams (cf. Forsling [17]). Such diagrams may facilitate experimental planning, partly by giving a concise picture of the fit to experimental data of the complex model assumed and partly by visualizing the species distributions at various conditions. A melt might also be regarded as a phase having as constituents a number of solute species or clusters. Tegman [18] has used SOLGASWATER to calculate the distribution of the clusters assumed, S_i^{2-} (i = 1, 2, . . . , 6, 8), in sodium polysulfide melts at various temperatures.

An algorithm, such as the one described in this paper, should find increasing applications within the field of geoscience as more reliable thermodynamic data are produced for minerals in aqueous solutions, e.g. for the simulation of geochemical processes or natural water systems. The possibility of using equilibrium models for such complicated systems by using SOLGASWATER has been demonstrated by Ingri [19]. Some of the problems considered include the solubility of amorphous silica in water and in different ionic sodium media and the solubility and phase relations in the kaolinite and chrysolite systems. Also included is a preliminary study of the speciation in the chemical weathering process of a feldspar. More detailed discussions of this type of computer calculation and its possibilities will be documented in a forthcoming paper by Ingri et al. [20].

The author is greatly indebted to Professor Nils Ingri who was the main initiator of this work. Thanks are also due to Professor Ingri and Dr. Staffan Sjöberg for many suggestions and discussions. Financial support for this work by the Swedish Natural Science Research Council is gratefully acknowledged.

REFERENCES

1 F. van Zeggeren and S. H. Storey, The Computation of Chemical Equilibria, Cambridge University, New York, 1970.
2 W. R. Smith, Ind. Eng. Chem., Fundam., in press.
3 G. Eriksson, Chem. Scr., 8 (1975) 100.
4 W. B. White, S. M. Johnson and G. B. Dantzig, J. Chem. Phys., 28 (1958) 751.
5 N. Ingri, W. Kakolowicz, L. G. Sillén and B. Warnqvist, Talanta, 14 (1967) 1261.
6 M. Bos and H. Q. J. Meershoek, Anal. Chim. Acta, 61 (1972) 185.
7 Ting-Po I and G. H. Nancollas, Anal. Chem., 44 (1972) 1940.
8 H. S. Dunsmore and D. Midgley, Anal. Chim. Acta, 72 (1974) 121.
9 D. A. Crerar, Geochim. Cosmochim. Acta, 39 (1975) 1375.
10 T. J. Wolery and L. J. Walters, Jr., J. Int. Assoc. Math. Geol., 7 (1975) 99.
11 T. M. L. Wigley, Tech. Bull., Br. Geomorphol. Res. Group, 20 (1977) 1.
12 I. K. Karpov and L. A. Kaz'min, Geokhimiya, 4 (1972) 402.
13 R. Gautam and W. D. Seider, AIChE J., in press.
14 L. G. Sillén and A. E. Martell, Stability Constants of Metal—Ion Complexes, The Chemical Society, London, 1964 (Spec. Publ. No. 17) and 1971 (Spec. Publ. No. 25).
15 A. E. Martell and R. M. Smith, Critical Stability Constants, Vols. 1—4, Plenum, New York, 1974—1977.
16 D. D. Perrin, Stability Constants of Metal—Ion Complexes: Part B, Pergamon, Oxford, 1979.
17 W. Forsling, Acta Chem. Scand., Ser. A, 32 (1978) 857.
18 R. Tegman, Chem. Scr., 9 (1976) 158.
19 N. Ingri, in G. Bendz and I. Lindqvist (Eds.), Biochemistry of Silicon and Related Problems, Plenum, New York, 1978, p. 3.
20 N. Ingri, G. Eriksson and S. Sjöberg, to be published.

Analytica Chimica Acta, 112 (1979) 385—396
Computer Techniques and Optimization

DIE ANWENDUNG DER INFORMATIONSTHEORIE ZUR BEWERTUNG VON COMPUTERGESTÜTZTEN SPEKTRENSUCHSYSTEMEN

KARL SCHAARSCHMIDT

Sektion Chemie der Technischen Universität Dresden, 8027 Dresden (D.D.R.)

(Eingegangen den 18 Mai 1979)

ZUSAMMENFASSUNG

Die Anwendung der Informationstheorie gestattet eine objektive Bewertung der Leistungsfähigkeit computergestützter Spektrensuchsysteme. Dazu ist eine signifikante Zahl von Suchprozessen auszuwerten. Der Informationsgewinn durch den Computereinsatz wird als Differenz der Entropie des Datenbestands und einer bedingten Entropie, die vom Anteil der erfolglosen Suchprozesse und dem anfallenden Ballast abhängt, betrachtet. Der Einfluß folgender Faktoren kann abgeschätzt werden: Umfang, Struktur und Qualität der gespeicherten Spektrensammlung, Effektivität der Codiervorschrift und des Vergleichsalgorithmus, subjektive Fehler bei der Codierung der Spektren. Die abgeleiteten Beziehungen wurden auf zwei bereits früher publizierte Speicher- und Suchsysteme für Infrarotspektren angewendet.

SUMMARY

(Evaluation of the efficiency of computer-aided spectra search systems based on information theory.)

Application of information theory allows objective evaluation of the efficiency of computer-aided spectra search systems. For this purpose, a significant number of search processes must be analyzed. The amount of information gained by computer application is considered as the difference between the entropy of the data bank and a conditional entropy depending on the proportion of unsuccessful search processes and ballast. The influence of the following factors can be estimated: volume, structure, and quality of the spectra collection stored, efficiency of the encoding instruction and the comparing algorithm, and subjective errors involved in the encoding of spectra. The relations derived are applied to two published storage and retrieval systems for infrared spectra.

Die rasche Entwicklung des Computereinsatzes in der analytischen Chemie spiegelt sich u.a. auch wider in zahlreichen Veröffentlichungen über den rechnergestützten Spektrenvergleich zum Zwecke der Substanzidentifizierung. Allein für die automatische Durchmusterung entsprechend verschlüsselter und abgespeicherter Sammlungen von Infrarotspektren gibt es die unterschiedlichsten Lösungsvorschläge, die zu mehr oder weniger einsetzbaren Suchsystemen führten [1—24].

Die Vielfalt der Vorschläge ergibt sich aus der Kompliziertheit der Problemstellung. Es gilt, einen Algorithmus zu entwickeln, der die Identität bzw.

Nicht-Identität von Spektren festzustellen gestattet und dabei die wohlbekannte Tatsache berücksichtigt, daß Probenpräparation und Aufnahmebedingungen nicht ohne Einfluß auf das Aussehen des Spektrums sind. Demnach ist in allgemeiner Form zu entscheiden, welche spektralen Parameter als wesentlich für die Identifizierung anzusehen sind und welche Abweichungen als spektroskopisch begründet zu tolerieren sind. Es ist offensichtlich, daß bei der Lösung des Problems Vergleichsalgorithmus und Codiervorschrift (für das unbekannte und jedes Dateispektrum) eine Einheit bilden. Leider enthalten die meisten der zitierten Veröffentlichungen nur spärliche oder gar keine Angaben über die Leistungsfähigkeit der entwickelten Programme. Die jeweils angeführten Beispiele sind eher geeignet, den Vergleichsalgorithmus zu illustrieren, als eine signifikante Anzahl von Suchprozessen in ihrer Gesamtheit zu charakterisieren.

Eine quantitative Bewertung dreier Suchsysteme wurde von Erley [4] durchgeführt. Dabei erwies es sich als vorteilhaft, daß alle drei Systeme auf demselben Datenbestand, dem von ASTM—Wyandotte basierten und die gefundenen Referenzen in vergleichbarer Weise vom Rechner ausgegeben wurden, nämlich in einer nach einem Wahrscheinlichkeitsmaß geordneten Liste. Wesentlich komplizierter wird ein Vergleich von Suchsystemen, die verschiedene Datenbestände benutzen und bei denen die Kriterien für Identität bzw. Ähnlichkeit der zu vergleichenden Spektren auf recht unterschiedliche Weise vom Benutzer zu handhaben sind. Im folgenden wird eine Methode vorgeschlagen, die geeignet sein sollte, in allgemeiner Form die Leistungsfähigkeit verschiedener Suchsysteme einzuschätzen und miteinander zu vergleichen. Sie besteht darin, daß man auf eine genügend große Anzahl von Suchprozessen die Beziehungen der Informationstheorie anwendet.

Die Informationstheorie als ein Teilgebiet der Kybernetik wird in zunehmendem Maße zur Beurteilung chemischer Analysenverfahren — unabhängig davon, ob dabei nachrichtentechnische Mittel verwendet werden oder nicht — eingesetzt [25—33]. Die Durchführung einer Analyse wird dabei als informationsliefernder Prozeß betrachtet, der eine — quantitativ zu erfassende — Ungewißheit beseitigt. In diesem Sinne wird abzuschätzen sein, welchen Informationsgewinn der Rechnereinsatz im Mittel bringt, wenn aus einem gegebenen Datenbestand bestimmte Referenzen, welche Lösungen von Identifikationsproblemen darstellen, herausgesucht werden.

Mit dieser Zielstellung grenzt sich dieses Verfahren gegenüber einem kürzlich von Dupuis und Dijkstra publizierten [32] ab, bei dem der Informationsgehalt der einzelnen spektralen Merkmale (Auftreten von Banden in bestimmten Bereichen) abgeschätzt wird [33].

INFORMATIONSGEWINN DURCH RECHNEREINSATZ BEI DER SPEKTRENSUCHE

Folgende Voraussetzungen gelten. (a) Zum Testen werden nur solche Substanzen verwendet, die in dem betrachteten Datenbestand vorhanden sind. (Das ist nur eine Einschränkung im Hinblick auf die folgenden Betrachtungen. Bereits früher war gezeigt worden, daß der Einsatz der rechnergestützten

Spektrensuche auch dann erfolgreich sein kann, wenn die gesuchte Substanz nicht in der Datei vorhanden ist, wohl aber ähnliche [34].) (b) Ihre Auswahl erfolgt zufällig. (c) Die verwendeten Spektren der ausgewählten Substanzen müssen anderen Ursprungs sein als die zum Aufbau der Datei benutzten. Die Wahl der Bezeichnungen in diesem Abschnitt erfolgt in enger Anlehnung an [35]. Von dort wurden auch die allgemeinen informationstheoretischen Beziehungen übernommen.

Der Datenbestand umfasse insgesamt N Referenzen. Im Hinblick auf das Folgende ist es zweckmäßig, grundsätzlich die Substanzen als Referenzen aufzufassen, nicht die Spektren. Unter den angenommenen Voraussetzungen hat dann jede Identifizierungsaufgabe, die als Zufallsversuch β bezeichnet werden soll, eine der N möglichen Lösungen $B_j (j = 1, 2, \ldots N)$. Die Zufälligkeit des Ausgangs von β besteht demnach darin, welche analytische Aufgabe gestellt ist. Sie wird durch die Wahrscheinlichkeiten $p(B_j)$ ausgedrückt. $p(B_j)$ ist die Wahrscheinlichkeit dafür, daß nach der Referenz j gesucht wird. Es gilt somit $\sum_{j=1}^{N} p(B_j) = 1$. Nach Shannon ist die Entropie eines solchen Versuchs, auch mittlere Informationsmenge genannt,

$$H(\beta) = -\sum_{j=1}^{N} p(B_j) \operatorname{ld} p(B_j) \tag{1}$$

(ld = Logarithmus zur Basis 2). Problematisch ist die Abschätzung der Werte von $p(B_j)$. Aus ihrer Definition folgt, daß man dazu feststellen müßte, wie oft nach jeder Referenz j gesucht wird. Das ist praktisch undurchführbar. Geht man aber davon aus, daß alle Referenzen im Mittel etwa die gleiche Chance haben, gesucht zu werden (siehe auch unten), dann gilt

$$p(B_j) = 1/N \tag{2}$$

und man erhält den Maximalwert der Entropie für den Versuch β:

$$H_{\max}(\beta) = \operatorname{ld} N \tag{3}$$

Dabei ist es im Prinzip gleichgültig, ob das Heraussuchen der richtigen Referenz per Hand oder per Computer geschieht. $H_{\max}(\beta)$ in Gl. (3) ist ausschließlich durch Umfang und Struktur des Datenbestands bestimmt und ist demzufolge als maximale Entropie des Datenbestands zu bezeichnen.

Der Einsatz des Rechners führt im allgemeinen zu einer mehr oder weniger großen Auswahl von Referenzen als möglichen Lösungen, aus denen der Benutzer des Suchsystems durch visuellen Spektrenvergleich die richtige heraussucht. (Es gibt gelegentlich Meinungen und Forderungen, der Rechner müsse eine eindeutige Lösung bringen. Das ist aber in der Praxis nicht haltbar, und fast alle Suchsysteme berücksichtigen diesen Umstand.) Im Sinne der Informationstheorie ist der Rechnereinsatz demnach als Vorversuch aufzufassen, der die Ungewißheit über den Ausgang des Versuchs β um einen möglichst großen Wert verringert. Er soll als Versuch α bezeichnet werden. Zweckmäßigerweise definiert man als Ausgang A_i des Versuchs α einen solchen, bei dem sich die richtige Referenz unter insgesamt N_i vom Rechner herausgesuchten befindet.

Von Interesse sind die Ausgänge mit kleinen Werten von N_i und die mit $N_i \approx N$ Letztere sind die negativen Ergebnisse des Rechnereinsatzes: nach dem Durchmustern der herausgesuchten Referenzen zeigt sich, daß die richtige — unter den Voraussetzungen dieser Betrachtungen — im großen Rest enthalten ist. Damit sind Erfolge und Mißerfolge formal auf einen der möglichen Ausgänge von α zurückgeführt, und es gilt $\sum_{i=1}^{N} p(A_i) = 1$. Die Verteilung der $p(A_i)$ ist durch eine größere Anzahl von Suchprozessen empirisch abzuschätzen. Sie wird benötigt, um die Information zu berechnen, die α über β bringt. Folgende Schritte führen dazu.

Ein spezieller Ausgang A_i des Versuchs α ändert die Ungewißheit in Bezug auf den Ausgang von β. An die Stelle der Entropie der Gl. (1) tritt die bedingte Entropie

$$H_{A_i}(\beta) = -\sum_{j=1}^{N} p_{A_i}(B_j) \operatorname{ld} p_{A_i}(B_j) \tag{4}$$

($p_{A_i}(B_j)$ bedingte Wahrscheinlichkeit, d.h. Wahrscheinlichkeit von B_j unter der Bedingung, daß α den Ausgang A_i hat, also $\sum_{j=1}^{N} p_{A_i}(B_j) = 1$.) Als Mittelwert ergibt sich die bedingte Entropie des Versuchs β unter der Bedingung α:

$$H_\alpha(\beta) = p(A_1)\, H_{A_1}(\beta) + p(A_2)\, H_{A_2}(\beta) + \cdots + p(A_N)\, H_{A_N}(\beta) \tag{5}$$

Die Differenz $H(\beta) - H_\alpha(\beta)$ ist ein Maß für die Verringerung der Unsicherheit, welche die Durchführung des Versuchs α (Einsatz des Rechners) in Bezug auf den Versuch β (das Auffinden der richtigen Referenz) bringt. Es ist

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta) \tag{6}$$

die in α über β enthaltene Information, gemessen in bit, und es gilt $0 \leqslant I(\alpha, \beta) \leqslant H(\beta)$. Nur im Idealfall, wenn der Rechner stets die richtige und nur die richtige Referenz herausfindet, wird der Maximalwert von $I(\alpha, \beta)$ erreicht. Jedes negative Ergebnis und jeglicher Ballast im Rechnerausdruck vermindern die durch den Rechnereinsatz zu erhaltende Information.

Zur praktischen Benutzung der gegebenen Beziehungen ist noch die Berechnung der bedingten Entropien $H_{A_i}(\beta)$ nötig. Anschaulich ausgedrückt ist es jeweils die Ungewißheit, welche unter den N_i herausgesuchten Referenzen die richtige ist. Entsprechend den Erläuterungen zu den Gl. (2) und (3) ist zu schreiben $p_{A_i}(B_j) = 1/N_i$ für jede der N_i herausgesuchten Referenzen und $p_{A_i}(B_j) = 0$ für alle übrigen. Somit wird

$$H_{A_i}(\beta) = \operatorname{ld} N_i \tag{3'}$$

Daraus ergibt sich die für die Auswertung benötigte Gleichung

$$I(\alpha, \beta) = \operatorname{ld} N - \sum_{i=1}^{N} p(A_i) \operatorname{ld} N_i \tag{7}$$

DURCHFÜHRUNG DER UNTERSUCHUNGEN AN ZWEI SUCHSYSTEMEN

Die zur Testung benutzten Suchsysteme SSU 2 und SPEKSU wurden berei

beschrieben [17, 24]. Für beide sind die Vergleichsspektren in Form von Datensätzen gespeichert, die aus einer geordneten Folge von acht (aus 47 möglichen) Codezahlen, den "spektralen Merkmalen" bestehen. Im Hinblick auf das Folgende ist vor allem von Interesse, daß bei beiden Systemen der Benutzer die Möglichkeit hat — wie bei den meisten anderen Systemen auch — die Präzision der Suchanfrage zu variieren, d.h. unterschiedliche Anforderungen an den Grad der Ähnlichkeit von Spektren zu stellen. Bei SSU 2 hat er zu bestimmen, wieviele Merkmale übereinstimmen müssen (maximal 8) und ob dabei ihre Reihenfolge berücksichtigt wird (Suchvariante 2) oder nicht (SV 1). Bei SPEKSU gibt er eine obere Schranke (MAXDIF) für eine Kennzahl an, die als "Differenz" zwischen je zwei Datensätzen vom Programm berechnet wird und die bei völliger Übereinstimmung den Idealwert Null besitzt.

Der Datenbestand für beide Systeme ist der gleiche. Für die Testung wurde nur der Anteil der niedermolekularen Verbindungen benutzt. Das waren zum Zeitpunkt der Testläufe etwa 900 Referenzen aus der Sammlung von Hummel und Scholl [36], und etwa 1500 andere Referenzen, vorzugsweise aus der DMS-Kartei [37], vgl. [34].

Die Testsubstanzen wurden mit Hilfe eines Zufallszahlenprogramms unter den im Datenbestand vorhandenen ausgewählt. Die zugehörigen Testspektren wurden vor allem der Sammlung der Standard-Gitterspektren von Sadtler [38] entnommen. Sie stammten auf jeden Fall aus anderen Registrierungen als die beim Aufbau der Datei verschlüsselten. Insgesamt wurde nach 43 Testsubstanzen als "Unbekannten" gesucht.

Die Suchanfragen wurden entsprechend den Benutzungsanweisungen für beide Systeme formuliert. Bei SSU 2 wurden grundsätzlich je 10 Anfragen eingegeben, nämlich nach SV 1 und SV 2 mit je 4 bis 8 Suchmerkmalen. Bei SPEKSU wurde mit unterschiedlichen Werten von MAXDIF gesucht. Damit ist für beide Systeme die Möglichkeit gegeben, den Einfluß der Präzision der Fragestellung auf die Antworten des Rechners systematisch zu untersuchen.

ERGEBNISSE

Erwartungsgemäß unterscheiden sich die Ergebnisse der Suchen nach den einzelnen Testsubstanzen beträchtlich. Eine ideale Übereinstimmung zwischen Test- und Vergleichsspektrum würde sich darin ausdrücken, daß die gesuchte Referenz auch bei einer Fragestellung höchster Präzision gefunden würde, bei SSU 2 also mit Suchvariante 2 und 8 Merkmalen und bei SPEKSU mit MAXDIF = 0. Das wird nur selten erreicht. Vielmehr bedingen die bereits einleitend erwähnten Einflüsse von Probenpräparation und Aufnahmebedingungen gewisse Unterschiede, die sich dadurch in den Suchergebnissen ausdrücken, daß die gesuchte Referenz erst bei geringeren Anforderungen an die Identität anfällt. Damit erhöht sich natürlich auch der Umfang des Ballasts.

In Tabelle 1 ist für SSU 2 der nach Gl. (7) berechnete Informationsgewinn, gemittelt über alle Suchprozesse nach jeweils gegebener Suchvariante und Anzahl von Suchmerkmalen, zusammengestellt. Aus Gründen, die weiter

TABELLE 1

Informationsgewinn (in bit) durch Anwendung des Suchprogramms SSU 2 mit unterschiedlicher Präzision der Fragestellung

Suchstrategie (SV/M)	I(ges)	I(HS)	I(Rest)
1/4	4,13	4,98	3,84
1/5	4,96	6,43	4,47
1/6	5,52	6,99	5,01
1/7	3,78	2,52	4,21
1/8	2,12	2,17	2,11
2/4	4,74	5,32	4,54
2/5	4,64	6,17	4,11
2/6	4,39	5,58	3,99
2/7	2,64	4,19	2,12
2/8	1,42	2,49	1,05

unten zu diskutieren sind, wird dabei unterschieden zwischen I(ges), I(HS) und I(Rest), je nachdem ob die Mittelung über die Suchen nach allen 43 Referenzen oder nach denen des Hummel/Scholl oder nach den übrigen durchgeführt wurde. Dabei wurde stets im ganzen Datenbestand gesucht, d.h. der maximale Informationsgewinn beträgt nach Gl. (2) immer 11,23 bit. In gleicher Weise enthält Tabelle 2 die entsprechenden Werte für den Einsatz von SPEKSU mit unterschiedlichen Werten von MAXDIF.

Bei der Berechnung der Tabellen 1 und 2 sind aus Gründen der Systematik auch solche Suchergebnisse mit berücksichtigt, die beim Einsatz von Suchsystemen in der analytischen Praxis als nicht auswertbar betrachtet werden. Wenn beispielsweise bei Suchvariante 1 und 4 Suchmerkmalen in einzelnen Fällen mehr als 100 Referenzen anfallen, bei der Suche nach derselben Unbekannten

TABELLE 2

Informationsgewinn (in bit) durch Anwendung des Suchprogramms SPEKSU mit unterschiedlichen Schranken MAXDIF

Suchstrategie (MAXDIF)	I(ges)	I(HS)	I(Rest)
12	5,81	6,85	5,45
10	5,33	7,05	4,74
8	4,84	6,67	4,21
6	4,42	6,54	3,60
5	3,55	5,80	2,78
4	3,03	4,71	2,46
3	2,83	4,94	2,11
2	2,12	4,28	1,37
1	0,65	2,53	0
0	0,26	1,02	0

mit SV 2 und 5 Merkmalen nur zwei oder drei, dann betrachtet man in praxi selbstverständlich nur die Listen geringeren Umfangs, und erst dann, wenn das ohne Erfolg bleibt, geht man zu den umfangreicheren Antworten über. (Hier und im Folgenden wird als "Antwort" die Gesamtheit der Referenzen betrachtet, die der Rechner auf eine bestimmte "Anfrage", charakterisiert durch Suchvariante und eingegebene Suchmerkmale, ausgibt. Die Begriffe sind entsprechend auf SPEKSU zu übertragen.) In diesem Sinne ist in den Tabellen 3 und 4 angegeben, wie oft (Spalte 2) Fragestellungen einer bestimmten Präzision (Spalte 1) diejenigen Antworten lieferten, in denen die jeweils gesuchte Referenz unter möglichst wenigen anderen enthalten war. Der minimale (Spalte 3), maximale (Spalte 4) und durchschnittliche (Spalte 5) Umfang dieser günstigsten Antworten ergibt eine unmittelbare, anschauliche Vorstellung von dem bei der Auswertung des Rechnerausdrucks zu betreibenden Aufwand. Demnach enthalten die Tabellen 3 und 4 die Angaben über die Leistungsfähigkeit der Systeme unter dem Gesichtspunkt der praktischen Anwendung. Auch auf diese Daten, d.h. auf den Umfang der günstigsten Antworten ist Gl. (7) anzuwenden. Man erhält damit (Tabelle 5) einen Informationsgewinn, der sich auf die jeweils günstigsten Antworten bezieht und deshalb hier als "günstigster Informationsgewinn" bezeichnet werden soll. Wiederum wird unterschieden, ob über die Gesamtheit der gesuchten Referenzen gemittelt wird oder über diejenigen, die sich im Hummel/Scholl bzw. dem "Rest" befinden.

DISKUSSION

Die hier beschriebene Testung erfaßt, indem sie die erhaltenen Ergebnisse analysiert, alle Faktoren, welche die Leistungsfähigkeit eines Suchsystems beeinflussen: Vergleichsalgorithmus und Codiervorschrift, Umfang und

TABELLE 3

Verteilung und Charakterisierung der günstigsten Antworten für SSU 2[a]

Suchstrategie (SV/M)	Anz. günstige Antworten	Anz. ausgegebene Spektren progünst. Antwort		
		Minimum	Maximum	Mittel
1/4	1	103	103	103
1/5	1	5	5	5
1/6	3	1	10	4
1/7	1	3	3	3
1/8	3	1	1	1
2/4	2	1	13	7
2/5	3	1	48	17
2/6	6	1	3	1,5
2/7	6	1	4	2
2/8	6	1	13	3,5

[a]11 Referenzen wurden nicht gefunden.

TABELLE 4

Verteilung und Charakterisierung der günstigsten Antworten für SPEKSU[a]

Suchstrategie (MAXDIF)	Anz. günstige Antworten	Anz. ausgegebene Spektren progünst. Antwort		
		Minimum	Maximum	Mittel
17	2	10	13	11,5
16	3	2	2	2
15	1	1	1	1
14	1	1	1	1
13	1	2	2	2
12	3	2	4	3
11	1	3	3	3
10	2	1	2	1,5
9	1	1	1	1
7	3	1	3	2
6	3	1	8	3
5	3	1	30	11
4	1	1	1	1
3	4	1	2	1,5
2	5	1	4	1,5
1	2	3	10	6,5
0	1	1	1	1

[a]7 Referenzen wurden nicht gefunden.

TABELLE 5

Günstigster Informationsgewinn I_g (in bit) bei der Anwendung der beiden Suchsysteme

	I_g(ges)	I_g(HS)	I_g(Rest)
SSU 2	7,48	9,43	6,81
SPEKSU	8,83	9,83	8,43

Zusammensetzung des Datenbestands, aufnahme- und präparationsbedingte Unterschiede zwischen Spektren ein und derselben Substanz, Codier- und sonstige subjektive Fehler beim Aufbau der Datei und bei der Durchführung der Suche. Die rein rechentechnischen Aspekte werden dabei bewußt unberück sichtigt gelassen.

Die oben aufgestellte Forderung, zur Testung nur unabhängige Spektren einzusetzen, ist dabei von grundsätzlicher Bedeutung. Wenn beispielsweise Erley [4] bei seinen vergleichenden Untersuchungen mit Spektren testete, von denen etwa zwei Drittel mit den in der Datei enthaltenen identisch waren, so erfaßte er im Wesentlichen nur die subjektiven Fehler, nicht aber die bedeutungsvolleren Einflüsse von Präparations- und Aufnahmetechnik.

Die Bedeutung der zuletzt genannten Faktoren darf nicht unterschätzt werden. Es ist bemerkenswert, daß nur selten eine völlige Übereinstimmung zwischen Test- und Vergleichsspektrum, ausgedrückt in der Differenz Null

bzw. Übereinstimmung bis zu acht Merkmalen nach Suchvariante 2, gefunden wurde, selbst dann nicht, wenn die gesuchte Referenz nach dem Rechnerausdruck als die mit höchster Wahrscheinlichkeit zutreffende aufzufassen ist. Das unterstreicht die Forderung an den Vergleichsalgorithmus, bei der Feststellung der Identität gewisse Abweichungen zu tolerieren.

Wie bereits erwähnt, haben die meisten Systeme die Möglichkeit, schon von der Eingabe her die Vergleichskriterien zu variieren. Eine Vergrößerung der Toleranzen erhöht zwar die "Erfolgsquote", d.h. die Wahrscheinlichkeit, die richtige Referenz zu finden, doch fällt dann auch mehr Ballast an. Das ist übrigens ein generelles Problem für jede Art von Informationsrecherchen [39]. Nach Gl. (7) wird der maximal mögliche Informationsgewinn sowohl durch erfolglose Suchen als auch durch umfangreichen Ballast vermindert. Die Tabellen 1 und 2 zeigen, wie die Präzision der Fragestellung den durch den Rechnereinsatz zu erhaltenden Informationsgewinn beeinflußt. Besonders auffällig sind die kleinen Werte bei sehr hohen Ansprüchen an die Übereinstimmung zweier Spektren. Bei einer Verminderung der Anzahl der Suchmerkmale bzw. einer Erhöhung von MAXDIF steigt der Informationsgewinn zunächst stetig an. Das zu erwartende Maximum, das durch eine relativ hohe Erfolgsquote bei vertretbarem Ballast bedingt ist, liegt aber nur teilweise innerhalb des untersuchten Variationsbereichs, ein Umstand, auf den noch zurückzukommen sein wird.

Insgesamt sind die Werte der Tabellen 1 und 2 wesentlich kleiner als der nach Gl. (3) zu berechnende Maximalwert von 11,23 bit. Das ist darin begründet, daß bei der Mittelung schematisch alle Ergebnisse für eine vorgegebene Kombination SV/M bzw. einen Wert von MAXDIF berücksichtigt wurden. Wie die Tabellen 3 und 4 zeigen, fallen die für die praktische Auswertung günstigsten Antworten bei Fragestellungen recht unterschiedlicher Präzision an. Die in Tabelle 5 enthaltenen Werte geben deshalb ein realeres Bild von der Leistungsfähigkeit der beiden Systeme und ermöglichen somit einen Vergleich. Es zeigt sich, daß SPEKSU einen größeren Informationsgewinn bringt als SSU 2. Dabei ist nicht ohne Bedeutung, in welchem Teil des Datenbestands die gesuchte Referenz vorhanden ist. Offensichtlich hängt das mit der Spektrenqualität zusammen, einer Frage, die im Zusammenhang mit Suchsystemen häufig diskutiert, hier aber quantitativ erfaßt ist.

Die Spektren des Hummel/Scholl wurden mit modernen Gittergeräten in einem Laboratorium unter den einheitlichen Gesichtspunkten der Spektrendokumentation aufgenommen. Demgegenüber sind in der DMS-Kartei Spektren der unterschiedlichsten Herkunft und Qualität vereinigt, die teilweise keineswegs im Hinblick auf die Dokumentation, sondern unter sehr speziellen spektroskopischen Fragestellungen aufgenommen wurden. Obwohl die Herausgeber eine kritische Auswahl getroffen haben, wurden bereits bei Voruntersuchungen [40] Unzulänglichkeiten (ungünstige Intensitätsverhältnisse, steigender Untergrund, größere Wellenzahlabweichungen) festgestellt. So konnte eine ganze Reihe an sich wünschenswerter Referenzen von vornherein nicht in die Datei aufgenommen werden. Immer noch bleiben die Suchergebnisse nach

diesen Spektren eindeutig hinter denen nach den Spektren aus dem Hummel/Scholl zurück. Die Qualitätsunterschiede zeigen sich bereits bei einer ersten Überprüfung der Suchergebnisse. Während die Hummel/Scholl-Spektren durch beide Systeme immer gefunden wurden, blieb die Suche nach den übrigen mit SSU 2 in 11 Fällen, mit SPEKSU in 7 Fällen erfolglos. Der Durchschnittswert der bei SPEKSU berechneten Differenz beträgt für erstere 4,3, für letztere 8,9 — wobei in letztere Mittelwertsbildung nur die bei SPEKSU erfolgreichen Suchen eingehen. Tabelle 5 zeigt die Qualitätsunterschiede noch deutlicher. Im Hinblick auf den Vergleich der Systeme ist von Interesse, daß beide bei der Suche nach Hummel/Scholl-Spektren etwa denselben Informationsgewinn erbringen (Spalte 2). Der Vorteil von SPEKSU zeigt sich vor allem bei der Suche nach Spektren, die aus den oben genannten Gründen im Mittel größere Unterschiede zwischen Test- und Dateispektren zeigen (Spalte 3).

Die angeführten Qualitätsunterschiede machen auch die Werte der Tabellen 1 und 2 verständlicher. Bei der Suche nach Hummel/Scholl-Spektren ist nicht nur der mittlere Informationsgewinn höher, sondern man beobachtet auch stets das zu erwartende Maximum, wenn man MAXDIF bzw. bei gegebener Suchvariante die Zahl der Suchmerkmale variiert. Bei der Suche nach den übrigen Spektren ist ein solches Maximum nur für SSU 2 und der wenig selektiven Suchvariante 1 zu finden, sonst liegt es offenbar außerhalb des betrachteten Variationsbereichs. Die Lagen der beobachteten Maxima selbst entsprechen etwa den Werten, die man nach einiger Erfahrung beim praktischen Einsatz der Systeme anwendet.

Es ist bemerkenswert, daß die Anwendung der Erley'schen Kriterien [4] eine solch differenzierte Beurteilung nicht gestattet. Die Berechnung der "accuracy" A, der "precision" P und der "performance" Q aus den Ergebnissen der Testung ergibt für die Suche

mit SSU 2 nach allen Spektren	$A = 0{,}70$	$P = 0{,}70$	$Q = 0{,}49$
nach den HS-Spektren	0,91	0,51	0,46
nach dem Rest	0,62	0,80	0,50
mit SPEKSU nach allen Spektren	0,84	0,68	0,57
nach den HS-Spektren	0,91	0,60	0,55
nach dem Rest	0,81	0,71	0,57

Die Werte von Q, die als Gütekennzahlen der Systeme aufgefaßt werden, zeigen zwar ebenfalls eine höhere Bewertung von SPEKSU, ergeben aber keinerlei Aussage über den Einfluß der durchschnittlichen Spektrenqualität.

Die Erley'schen Kriterien sind auch grundsätzlich nicht in der Lage, den Umfang der Datei zu bewerten. Nach einer Erweiterung des Datenbestands würde eine Wiederholung der Testläufe Q bestenfalls konstant lassen, meist aber wohl wegen des höheren Ballasts, der zu kleineren P-Werten führt, verringern. Die hier vorgeschlagene Methode berücksichtigt dagegen durch den Ausdruck ld N in Gl. (7) den Umfang des Datenbestands. Ob die durch Rechnereinsatz erhältliche Information bei Vergrößerung der Datei wirklich wächst, hängt davon ab, wie sich dann der Ballast vergrößert, und das ist in dem zweiten Term von Gl. (7) enthalten.

Wie bereits bei der Ableitung von Gl. (3) bemerkt wurde, ist in dem Ausdruck ld N jedoch nicht nur der Umfang der Datei enthalten, sondern auch eine Annahme über die Häufigkeitsverteilung der Ausgänge B_j, vgl. Gl. (2). Entsprechend den allgemeinen Prinzipien der Informationstheorie wird die nach Gl. (3) erreichbare maximale Entropie durch solche Referenzen vermindert, nach denen nie oder aber sehr oft gesucht wird. In diesem Sinn ist Gl. (2) als Forderung an die Zusammensetzung des Datenbestands aufzufassen. Die N Referenzen der Datei müssen a priori jeweils dieselbe Wahrscheinlichkeit haben, jemals als analytisches Problem aufzutauchen. Das mag im Hinblick auf kleine, spezialisierte Sammlungen als trivial erscheinen, für große dagegen zu allgemein. Angesichts des großen Aufwands, der beim Aufbau der Datei zu betreiben ist, sollte diese Forderung jedoch unbedingt beachtet werden. In den großen kommerziellen Sammlungen gibt es nicht wenige Spektren, die im Hinblick auf sehr spezielle Problemstellungen bzw. unter extremen Bedingungen aufgenommen wurden und somit keineswegs den Gesichtspunkten der Routineanalytik entsprechen. Auch sind gewisse Substanzklassen als homologe Reihen sehr reichlich vertreten, andere dagegen nur spärlich. Auf die Notwendigkeit eines ausgewogenen Datenbestands in computerisierten Spektrensammlungen wurde bereits in anderem Zusammenhang verwiesen [10, 35, 41]. Die Informationstheorie begründet, wie oben gezeigt wird, diese Notwendigkeit in allgemeiner Form.

Es ist von Bedeutung, daß das hier vorgestellte Verfahren in völligem Einklang steht mit Vorstellungen und Schlußfolgerungen, die durch längere Erfahrungen bei der praktischen Anwendung von Spektrensuchsystemen gewonnen werden. Während derartige Erfahrungen jedoch mehr subjektiver und qualitativer Natur sind, gestattet die Anwendung der Informationstheorie eine objektive und quantitative Beurteilung der Faktoren, welche die Leistungsfähigkeit eines Suchsystems von der spektroskopisch-analytischen Seite her bestimmen: Umfang, Struktur und Qualität des Datenbestands, Effektivität von Codiervorschrift und Vergleichsalgorithmus, subjektive Fehler beim Codieren der Spektren.

Für die wertvollen Diskussionen bei der Abfassung des Manuskripts danke ich Herrn Prof. Dr. habil. E. Steger, Sektion Chemie der TU Dresden, Herrn Dr. H. Mühlig, Sektion Mathematik der TU Dresden, Herrn Dr. K. Danzer, Sektion Chemie/Werkstofftechnik der TH Karl-Marx-Stadt.

LITERATUR

1 D. H. Anderson und G. L. Covert, Anal. Chem., 39 (1967) 1288; dort auch Hinweis auf einige frühere Arbeiten.
2 L. H. Cross, J. Haw und D. J. Shields, in P. Hepple (Ed.), Molecular Spectroscopy, Institute of Petroleum, London, 1968.
3 D. S. Erley, Anal. Chem., 40 (1968) 894.
4 D. S. Erley, Appl. Spectrosc., 25 (1971) 200.
5 Ju. P. Drobischev, R. S. Nigmatullin, W. I. Lobanov, I. K. Korobeinitscheva, W. S. Botschkarev und W. A. Koptjug, Mitt. Akad. Wiss. UdSSR, 40 (1970) 75.

6 Ju. P. Drobischev, R. S. Nigmatullin, W. I. Lobanov, I. K. Korobeinitscheva, W. S. Botschkarev und W. A. Koptjug, Nachr. Sibir. Abt. Akad. Wiss. UdSSR, Ser. Chem. Wiss., (1972) 108.
7 F. E. Lytle, Anal. Chem., 42 (1970) 355.
8 C. S. Rann, Anal. Chem., 44 (1972) 1669.
9 R. W. Sebesta und G. G. Johnson, Jr., Anal. Chem., 44 (1972) 260.
10 F. Erni und J. T Clerc, Helv. Chim. Acta, 55 (1972) 489.
11 F. Erni, Dissertation, ETH Zürich (1972).
12 M. M. Noone, in E. Hettamann (Ed.), Modern Methods of Steroid Analysis, Academic Press, New York, 1973.
13 J. Zupan, D. Hadzi und M. Penca, Kem. Ind., 23 (1974) 275.
14 E. C. Penski, D. A. Padowski und J. B. Bouck, Anal. Chem., 46 (1974) 955.
15 D. G. Strauss, J. Antibiotics, 27 (1974) 805.
16 K. Tanabe, S. Saeki und T. Tamura, J. Nat. Chem. Lab. Ind. Tokyo, 69 (1974) 487; Jpn. Anal., 23 (1974) 626.
17 K. Schaarschmidt, R. Riemer und E. Steger, Z. Chem., 14 (1974) 374.
18 K. Tanabe und S. Saeki, Anal. Chem., 47 (1975) 118.
19 E. M. Kirby, R. N. Jones und D. G. Cameron, CODATA Bull., 21 (1976) 18.
20 R. C. Fox, Anal. Chem., 48 (1976) 717.
21 Z. Hippe, Persönliche Mitteilung (1976).
22 J. Zupan, M. Penca, D. Hadzi und J. Marsel, Anal. Chem., 49 (1977) 2141.
23 W. H. Littke und Ch. Jähn, Z. Chem., 17 (1977) 169.
24 K. Schaarschmidt, Z. Chem., 18 (1978) 337.
25 K. Doerffel und W. Hildebrandt, Wiss. Z. Techn. Hochsch. Chem. Carl Schorlemmer Leuna Merseburg, 11 (1969) 30.
26 K. Eckschlager, Collect. Czech. Chem. Commun., 36 (1971) 3016; 37 (1972) 137, 1486; 38 (1973) 1330; 39 (1974) 1426; 41 (1976) 1875; Z. Chem., 16 (1976) 111; Anal. Chem., 49 (1977) 1265.
27 K. Danzer, Z. Chem., 13 (1973) 20, 69, 229; 14 (1974) 73; 15 (1975) 326.
28 K. Doerffel, Chem. Tech., 25 (1973) 94.
29 K. Danzer, E. Than und D. Molch, Analytik, Akademische Verlagsgesellschaft, Geest und Portig K.-G., Leipzig, 1976.
30 H. Malissa und J. Rendl, Z. Anal. Chem., 272 (1974) 1.
31 F. Dupuis und A. Dijkstra, Anal. Chem., 47 (1975) 379.
32 P. F. Dupuis und A. Dijkstra, Z. Anal. Chem., 290 (1978) 357.
33 P. F. Dupuis, A. Dijkstra und J. H. van der Maas, Z. Anal. Chem., 291 (1978) 27.
34 K. Schaarschmidt, Z. Chem., 17 (1977) 299.
35 A. M. Jaglom und I. M. Jaglom, Wahrscheinlichkeit und Information, VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.
36 D. O. Hummel und F. Scholl, Atlas der Kunststoffanalyse, Bd. 2, Zusatzstoffe und Verarbeitungshilfsmittel, Carl Hauser/Verlag Chemie, München/Weinheim, 1973.
37 Dokumentation der Molekülspektroskopie, Kartei der Infrarotspektren, Verlag Chemie/ Butterworths, Weinheim/London.
38 Sadtler Research Laboratories, Philadelphia, U.S.A., Standard Spectra — Infrared Grating Spectra.
39 G. Salton, Automatic Information Organisation and Retrieval, McGraw-Hill, New York, 1968.
40 E. Knabe, Diplomarbeit, TU Dresden, 1971.
41 R. Büchi, J. T. Clerc, Ch. Jost, H. Koenitzer und D. Wegmann, Anal. Chim. Acta, 103 (1978) 21.

Analytica Chimica Acta, 112 (1979) 397—405
Computer Techniques and Optimization

AUTOMATED POTENTIOMETRIC ANALYSIS WITH SELECTIVE ELECTRODES

L. P. RIGDON, C. L. POMERNACKI, D. J. BALABAN and J. W. FRAZER*
Lawrence Livermore Laboratory, University of California, Livermore, CA 94550 (U.S.A.)
(Received 8th July 1979)

SUMMARY

An automated potentiometric analyzer based on a residual-chlorine electrode has been developed that can determine chlorine concentrations of a 4—75 ppb (± 2 ppb, standard deviation 0.6—1.9 ppb). A microcomputer-controlled digital buret delivers a predetermined number of aliquots of standard, and the electrode measures the potential of the test solution after each addition. An arithmetic processing unit transforms the acquired potentiometric data to the Gran (antilog) domain; the microcomputer then calculates a set of equivalence-point estimates and, using an error function criterion, selects the best estimate. The instrument includes a digital buret, 8080 microcomputer, arithmetic processing unit, real-time clock, video display, and 16-character key pad; it uses a modified version of the computer language BASIC compiled and programmed into Programmable Read-Only Memory (PROM). Although this system was designed to assay residual chlorine in water, the analyzer can also function as a digital pH (or millivolt) meter or, with minor software modifications, serve as a general-purpose instrument using any appropriate electrode system.

Minicomputers have been used successfully to improve the accuracy of potentiometric analyses with ion-selective electrodes (ISE) [1—3]. The same tasks and functions achieved by minicomputers can now be performed by a microcomputer-controlled potentiometric analyzer, a less expensive instrument designed and built at this laboratory from commercially available components.

The techniques used to increase the accuracy of these analyses, originally proposed by Gran [4] and refined by Frazer et al. [1, 2], manipulate 100—500 potential—volume data pairs to determine the concentration of the constituent of interest. Because the titrant addition volumes are small and because the electrochemical potential is proportional to the logarithm of the concentration (Nernst equation), the change of potential from one addition of titrant to the next is small for most of the data set. Very precise measurements of the potential, with resolution of about 10 μV, are therefore required. Precise timing of titrant addition is also necessary because the electrochemical cell usually cannot equilibrate in the short intervals (0.3—5 s) between additions. Calculation errors introduced by rounding and truncation of significant numbers can also result in serious errors in the assay values. These problems have been minimized in recent years, first by using minicomputers and now microcomputers.

During an analysis by the instrument described here, a digital buret delivers a predetermined number of titrant additions of known volume, and a microcomputer records the electrochemical potential of the test solution immediately before each subsequent addition. After the required values have been accumulated, the microcomputer transforms them to the Gran (antilog) domain, smooths the transformed values, calculates a set of concentration estimates, and chooses the best value by means of an error-function technique.

The current software package is dedicated primarily to performing the known standard-addition assay for active chlorine in water [3] with the Orion 97-70 residual chlorine electrode. The instrument can also function as a digital pH or millivoltmeter (DVM) without modification of the software.

EXPERIMENTAL

Hardware

The potentiometric analyzer and the associated ancillary equipment include a digital buret (Mettler Instruments Model DV10) and an electrochemical cell with stirrer and a residual-chlorine electrode (Orion Model 97-70) [1—3]. A block diagram (Fig. 1) describes the functions and interactions of the principal components and traces the signal from the electrode through the hardware elements to the CRT where it is displayed. The analyzer is controlled by the operator using a 16-character BCD key pad.

During an analysis, the digital buret, controlled by the microcomputer and driven by the I/0 interface, adds fixed volumes of reagent to the stirred solution in the electrochemical cell. The potential difference measured, which is proportional to the species concentration, is amplified, filtered, sampled at

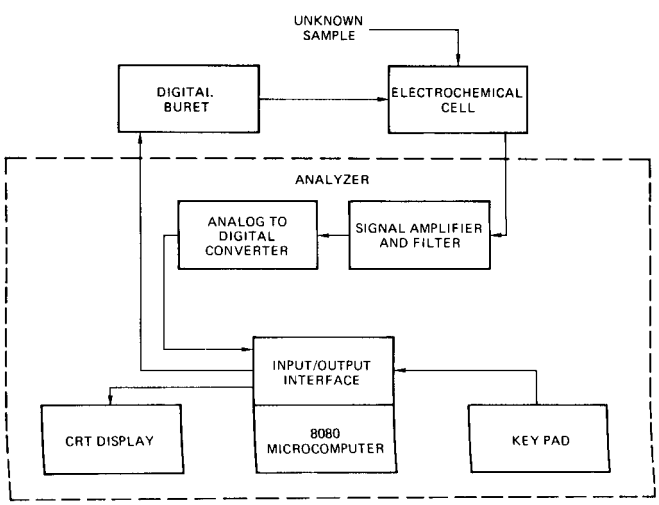


Fig. 1. Principal hardware elements of the Potentiometric Analyzer showing information flow.

a rate specified by the software program, and converted from analog to digital form. The microcomputer collects the digitized data through the I/0 interface, performs the required calculations, and displays the results on a CRT.

Figure 2 shows the analog signal conditioning and switching elements. A unity-gain amplifier with high input impedance (10^{14} ohm) and low bias current (10^{-14} A) minimizes loading of the electrochemical cell. If the analyzer is operated in the ISE mode, which is the one used for the residual-chlorine electrode, the signal is offset 0.5 V to reduce the input voltage level and is amplified 50-fold; the residual-chlorine electrode develops a potential of 510—680 mV. The offset can be changed manually to accommodate ion-selective electrodes that produce significantly different output voltages. If the analyzer is operated in the pH mode, the signal is amplified 5-fold. An analog switch, controlled by the microcomputer, transmits the signal from the appropriate amplifier to a 5-Hz Bessel filter for noise rejection. The filtered analog signal is attenuated 5-fold and sent to the analog-to-digital converter (ADC). The ADC is a panel meter (Analog Devices Model 2004) employing dual slope integration with 100-μV resolution and a range of 0—1.9999 μV. The offset and amplification portions of the ISE circuit increase the resolution of the signal to 10 μV.

Figure 3 describes the microcomputer and interface hardware elements in greater detail. The computer consists of an 8080 microprocessor with 4K of read-and-write memory (RAM) and 12K of read-only memory (ROM), and an arithmetic processing unit (APU; Advanced Micro Devices AM9511). The APU performs standard arithmetic operations in fixed- and floating-point formats, computes the transcendental functions and their inverses, and

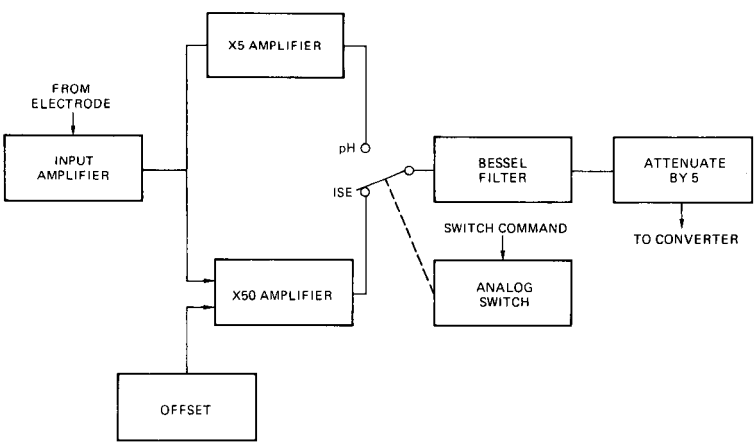


Fig. 2. Analog signal conditioning and switching elements from the electrode to the analog-to-digital converter.

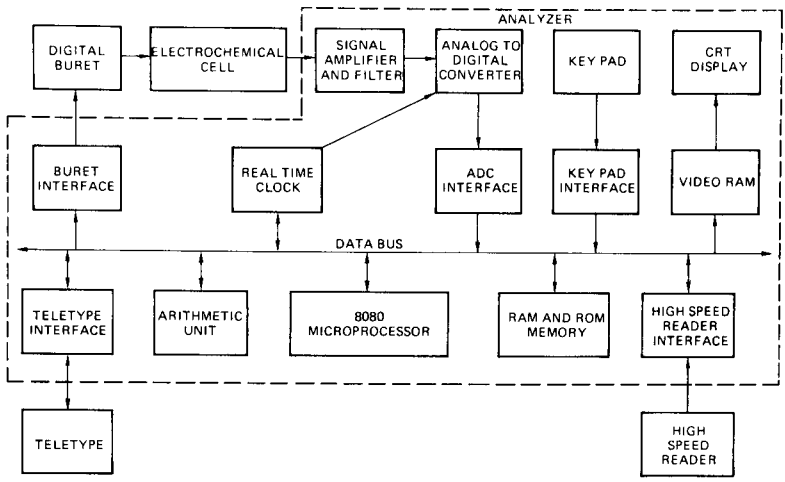


Fig. 3. Microcomputer and hardware-interface elements showing information flow.

performs control and type conversion. The APU is treated as a memory device and is addressed at a special memory location; operation commands and data are transferred between the central processing unit (CPU) and APU by a sequence of bytes via the 8-bit data bus. The CPU monitors the APU during periods of APU activity, and, when the APU computations are complete, the CPU acquires the numerical results and continues in the BASIC program.

A real-time 60-Hz a.c. clock provides the precision timing required to operate the digital buret and to sample the potentiometric data from the ADC. The clock frequency is divided by 6 to yield 100-ms periods (0.1 Hz). Sampling intervals of 0.2—25 s in steps of 0.1 s are thus possible. Initiation of data sampling by the ADC is determined by a delay factor chosen by the operator. After the sampling interval has been established by program control and transferred to a hardware buffer, the ADC is driven independently by precise timing signals from the timer, which also controls the buret, and the repetition rate of the pulsed signal is generated by a software routine in assembly-level language.

Communication between the operator and the analyzer is achieved by using the key pad for input and a CRT for output; a video RAM (a display memory unit) drives the CRT display monitor. The teletype and high-speed paper reader shown in Fig. 2 were used for program development and debugging but are not part of the instrument.

Software

The microcomputer was programmed with BASIC II, a modified version of BASIC with real-time data acquisition capabilities, and a BASIC II compiler [5, 6]. These provide microprocessor users with the conveniences

of a high-level interactive computer language for program development and the computational efficiencies of a compiled program. [BASIC II is available from Argonne Code Center, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL. BASIC II programs can be compiled and the compiled instructions loaded into PROM. The BASIC II compiler is available from the authors.]

The value of the APU for these analyses was demonstrated by comparing two mathematical routines of the analytical program. One program used only arithmetic control routines of the 8080 software; the other program relied on the APU to perform all arithmetic and transcendental functions. Both routines were implemented in interpretative BASIC. To complete a typical analysis, the latter program with the APU required 20% less memory and was 10 times faster than the 8080 version that employed only software subroutines. By eliminating the interpreter and loading the compiled program into PROM, the computational time was further reduced by a factor of 10. The analyzer incorporates both of these techniques, and the computations for each analysis require about 10 s.

Although the software is intended primarily for potentiometric analysis of low levels of chlorine in water [3], two other programs are available for convenience. The first is a very simple routine that reads the voltage of any electrode every 0.5 s and displays the values on the CRT. Each new reading replaces the oldest value so that the 15 most recent values are always retained in chronological sequence on the CRT.

The second program calculates the pH of a solution every 0.5 s and displays the most recent value. The pH electrodes are first calibrated by measuring the potential of two buffer solutions with different pH values. The pH is then computed by using two parameters (pH_0 and $1/S$) and conventional equations:

$$pH = (pH_0) - (1/S)\,E \tag{1}$$

$$pH_{low} = (pH_0) - (1/S)\,E_{low} \tag{2}$$

$$pH_{high} = (pH_0) - (1/S)\,E_{high} \tag{3}$$

Solving the linear eqns. (2) and (3) for pH_0 and $1/S$ yields

$$pH_0 = (E_{high} \cdot pH_{low} - E_{low} \cdot pH_{high})/(E_{high} - E_{low}) \tag{4}$$

$$1/S = (pH_{high} - pH_{low})/(E_{low} - E_{high}) \tag{5}$$

The microcomputer retains the values calculated for pH_0 and $1/S$ and uses them in eqn. (1) to calculate a new pH value each time the potential is measured. This process continues until the electrodes are recalibrated or the program is stopped.

The ISE software program can be used without modification for any analysis in which the known-standard-addition technique is used, but requires minor changes if more than 120 additions are made per analysis or if titrations are conducted. The ISE program, longer and more complex than the pH

or DVM programs, consists of six operations: (1) the program is initialized to define the run and calculation parameters; (2) the data are acquired, digitized, and smoothed; (3) the smoothed data are transformed to the Gran (antilog) domain, and the transformed data are smoothed; (4) a set of estimates of the unknown concentration based on the transformed data is made; (5) the differences between successive estimates are calculated; and (6) the estimate with the smallest difference from the previous estimate is selected and displayed.

Before an analysis begins, the microcomputer requests the necessary information on the CRT, and the operator answers using the key pad. The operator enters: (1) the concentration of the standard solution; (2) the number of additions of standard; (3) the volume of each addition; (4) total time to add the standard; (5) the temperature of the test solution; and (6) the volume of the test solution being assayed. The same run parameters can be used repeatedly without re-initialization.

The analysis is started by depressing any key. The program calculates the time interval between each addition and measurement of potential and determines the number of steps of the buret motor required for each titrant addition. The microcomputer drives the buret automatically and accumulates data until the preselected run conditions are satisfied.

Data are acquired by reading the digital panel meter and are smoothed by a 13-point quadratic-polynomial filter; the jth smoothed datum is computed as follows:

$$S_j = C_0 d_j + \sum_{i=1}^{6} C_i(d_{j-i} + d_{j+i}) \tag{6}$$

where d_j are the unsmoothed data and S_j are the smoothed data. This filter is the weighted moving-average filter described by Savitzky and Golay [7] who gave values for the coefficients C_0 and C_i.

The concentration for each addition is calculated by

$$X_i = (V_0 X_0 + i\Delta V X_s)/(V_0 + i\Delta V) \tag{7}$$

where X is the chemical species being determined, X_0 is the initial unknown concentration of X, X_s is the concentration of X in the standard, X_i is the concentration of X after the ith standard addition, V_0 is the initial volume of the test solution, and ΔV is the volume added for each addition.

The simple Nernst equation relating concentration to the potential is then given by $E_i = E^0 + RT \ln(X_i)/nF$, and the Gran transformation is

$$G_i = K(V_0 + i\Delta V) \exp(E^0 nF/RT) \tag{8}$$

where E_i is the potential at the ith addition, K is a scaling constant, and other symbols have their conventional definitions. Substitution of E_i into eqn. (8) yields

$$G_i = K \exp(E^0 nF/RT)/(V_0 + i\Delta V) X_i \tag{9}$$

Letting $K' = K \exp(E^0 nF/RT)$ and substituting X_i (eqn. 7) into eqn. (9) yields

$$G_i = K'X_s i\Delta V + K'V_0X_0. \tag{10}$$

A nearly linear plot is obtained when G_i is plotted as a function of i with slope $m = K'X_s\Delta v$ and intercept $b = K'V_0X_0$. The unknown concentration of the species X_0 in the solution is calculated by solving the equations of the slope and intercept for X_0:

$$X_0 = bX_s\Delta V/mV_0 \tag{11}$$

Plotting G_i vs. i (eqn. 10) would give a straight line if there were no electrical noise superimposed on the signal and if the electrochemical cell followed the Nernst model exactly. In actuality, however, neither condition applies. It is therefore beneficial to reduce noise from the transformed data by using the 13-point filter (described in eqn. 6), and to find the most linear segment of the Gran plot by using an error-function criterion.

This method [2, 3] proceeds as follows: the CPU calculates a set of equivalence-point estimates from subsets of the potentiometric data file. Each estimate is based on a linear least-squares fit of 35 data points; the first estimate uses the points 1—35, the second uses points 6—40, the third uses points 11—45, etc. The estimate that varies the least from the previous estimate is chosen, and the concentration of the unknown species is calculated and displayed on the CRT.

Least-squares computation

The concentration estimates are based on linear least-squares fits of data subsets computed in turn by finding the differences between the sums of the squares of data subsets, usually large, nearly equal numbers. Restrictions imposed by the CPU word length require truncation of significant digits, however, and the use of only seven decimal digits of accuracy to determine the slope of the line can result in frequent erroneous estimates. For example, the use of three 8-bit words gives an effective mantissa of 24 bits (equivalent to slightly more than seven decimal digits), but after the data have been transformed to the antilog domain, the information needed is generally contained in the fourth to tenth digits.

The following procedure minimizes this problem by, in effect, shifting the data so that only the differences required for calculation are retained. If the equation describing the desired line is $y = b_0 + b_1 X$, then the usual solution for the y intercept is

$$b_0 = \frac{1}{n}\sum Y_i - \frac{b_1}{n}\sum X_i \tag{12}$$

and the slope is

$$b_1 = [n\Sigma X_iY_i - (\Sigma X_i)(\Sigma Y_i)]/[n\Sigma X_i^2 - (\Sigma X_i)^2] \tag{13}$$

where all summations run from 1 to n. Let D and E equal the first values in the X and Y data arrays, respectively. When these values are subtracted from the data, eqn. (13) becomes

$$b_1 = \{n\,\Sigma\,(X_i - D)\,(Y_i - E) - [\,\Sigma\,(X_i - D)]\,[\,\Sigma(Y_i - E)]\}/$$
$$n\,\Sigma\,(X_i - D)^2 - [\,\Sigma\,(X_i - D)]^2 \qquad (14)$$

Table 1 contains a set of potentiometric data obtained with a bromide-selective electrode and processed by a PDP11/45 computer. The mantissas are the same as those used in the 8080 microcomputer. Using these data in eqn. (13) gave $b_1 = 1.2/0$, an unusable result; using the values of $X_i - D$ and $Y_i - E$ gave $b_1 = -1.096293/0.2565266 \times 10^{-2} = -4.273604$. This value can be used in eqn. (12) to calculate the intercept and, subsequently, an estimate of the concentration of the chemical species.

PERFORMANCE EVALUATION

The analyzer was tested by assaying aqueous samples of known concentration containing 3.91—75.45 ppb of chlorine. The mock samples and standard addition solution were prepared with chloramine-T as described previously [3]. Two operators using two different Orion 97-70 residual-chlorine electrodes assayed 95 test samples. Standard solution was added and the potential measured at 1-s intervals to generate 126 data points for each

TABLE I

Comparison of numbers retained before and after subtraction of the first member of each data pair from each member of a typical experimental data set

	X_i	$X_i - D$	Y_i	$Y_i - E$
1	−5.629047	0	166.1560	0
2	−5.628498	0.5497932E-03	165.8569	−0.2990723
3	−5.627949	0.1098156E-02	165.5823	−0.5737305
4	−5.627401	0.1646042E-02	165.2588	−0.8972168
5	−5.626854	0.2192974E-02	164.9597	−1.196289
6	−5.626309	0.2738476E-02	164.7827	−1.373291
7	−5.625764	0.3283501E-02	164.5020	−1.654053
8	−5.625219	0.3828049E-02	164.2029	−1.953125
9	−5.624676	0.4371166E-02	163.9954	−2.160645
10	−5.624134	0.4913330E-02	163.7512	−2.404785
11	−5.623592	0.5455017E-02	163.5864	−2.569580
12	−5.623052	0.5995274E-02	163.2996	−2.856445
13	−5.622512	0.6535053E-02	163.1348	−3.021240
14	−5.621974	0.7073879E-02	162.9211	−3.234863
15	−5.621436	0.7611752E-02	162.7075	−3.448486
16	−5.620899	0.8148670E-02	162.5366	−3.619385
17	−5.620362	0.8685112E-02	162.3657	−3.790283
18	−5.619827	0.9220123E-02	162.2437	−3.912354

assay. Eighteen concentration estimates were made for each run, and the error-function minimum was selected as the assay result. The results (Table 2) show that an accuracy of ± 2 ppb or better was obtained.

TABLE 2

Performance evaluation data of the analyzer for assay of chlorine in water

No. of runs	Active chlorine		Error	
	Taken (ppb)	Found (ppb)	ppb	RSD
24	3.91	3.91	0	0.60
25	15.56	15.76	+0.20	0.98
36	38.46	38.57	+0.11	0.98
10	75.45	77.32	+1.87	1.14

The authors thank K. Gertz for his help in constructing the hardware and H. Brand for helpful suggestions on software development. This work was performed under the auspices of the U.S. Department of Energy under contract No. W-7405-Eng-48 and the U.S. Environmental Protection Agency under interagency agreement D7-0321.

REFERENCES

1 J. W. Frazer, A. M. Kray, W. Selig and R. Lim, Anal. Chem., 47 (1975) 869.
2 J. W. Frazer, W. Selig and L. P. Rigdon, Anal. Chem., 49 (1977) 1250.
3 L. P. Rigdon, G. J. Moody and J. W. Frazer, Anal. Chem., 50 (1978) 465.
4 G. Gran, Analyst (London), 77 (1952) 661.
5 T. Allison, R. Eckard and J. Barber, User's Guide to the LLL Basic Interpreter, Lawrence Livermore Laboratory, Livermore, Calif., UCID-17090 Rev. 1, 1977.
6 P. R. McGoldrick, J. Dickinson and T. G. Allison, LLL 8080 BASIC II, Interpreter User's Manual, Lawrence Livermore Laboratory, Livermore, Calif., UCID-17752, 1978.
7 A. Savitzky and M. J. E. Golay, Anal. Chem., 36 (1964) 1627.

Analytica Chimica Acta, 112 (1979) 407—416
Computer Techniques and Optimization

THE ROLE OF PATTERN RECOGNITION IN THE COMPUTER-AIDED CLASSIFICATION OF MASS SPECTRA

WILLIAM S. MEISEL and MATT JOLLEY
Technology Services Corporation, Santa Monica, CA 90403 (U.S.A.)
STEPHEN R. HELLER*
Environmental Protection Agency, PM-218, Washington, DC 20460 (U.S.A.)
GEORGE W. A. MILNE
NHLBI, NIH, Bethesda, MD 20205 (U.S.A.)
(Received 28th June 1979)

SUMMARY

The requirements for the use of pattern recognition techniques as an aid in the identification of chemical substances from their mass spectra are reviewed. Decision-tree pattern recognition is recommended as potentially satisfying these requirements. Examples of this approach using a large data base of mass spectra are provided.

In recent years, there has been considerable activity in the area of identification of unknowns from their spectral features. Mass spectrometry, infrared, proton and carbon-13 nuclear magnetic resonance have been the primary tools used [1]. Early approaches involved the establishing of as large a library of spectral data as possible and use of one of a number of algorithms to compare an unknown with the contents of the library [2]. Later, combined library searching, based on combinations of spectral data was introduced [3], and finally, as libraries grew larger and comparison algorithms seemed to have reached their limits of usefulness, pattern recognition techniques appeared [4].

While all these approaches have virtues based on practical needs, they also all leave much to be desired. In particular, as libraries grow bigger, search times (costs) also grow; and pattern recognition techniques, for all their sophistication, have rarely been able to match the accuracy of simple library-search algorithms in the area of structure elucidation. Classification to the level of a particular chemical structure will probably continue to be best performed by a library search, provided that the particular structure is in the library.

The role that is envisioned for pattern recognition techniques in structure elucidation by mass spectrometry is twofold. First, pattern recognition may be used as a supplement to library search to signal a possible error in the final match. For example, if pattern classification algorithms indicate a lack

of nitrogen in the compound with high confidence, and the library search matches with a nitrogen-containing compound, then the results must be viewed with caution. Secondly, pattern recognition techniques can aid identification of an unknown when it is not in the library by indicating the probability of the presence or absence of a certain chemical structure or substructure. This information can guide classical approaches based on experience or can be used as a constraint for a computer program which generates alternative structures [5].

The form of the results of such a pattern recognition algorithm, which uses a mass spectrum as input data, is a list of potential characteristics of the apposite functional groups (e.g. "nitrogen present", "aromatic rings present", etc.) along with a measure of the probability that the given characteristic applies to the unknown chemical. The probability that a characteristic is present can be stated both in relative and absolute terms; i.e. one can indicate the probability that the characteristic is present, either by including consideration of the frequency with which the characteristic occurs among representative chemicals or by ignoring such a consideration. In addition to an alphabetical list of characteristics, the characteristics should be listed with the most probable characteristics first and the least probable last. The ranked or ordered list gives a quick view of the most likely and least likely characteristics of the unknown.

In view of these aims, a classification algorithm should have four features. First, the algorithm should yield a measure of the probability that a characteristic is present, not simply a yes/no "best guess". Many classical approaches to pattern recognition simply yield an unequivocal classification of the unknown, with no indication as to the confidence in that classification. Secondly, the operational algorithm should be computationally efficient to minimize expense and to allow a large number of characteristics to be tested. (This does not require that the derivation of the algorithm be computationally efficient. Often the accuracy and efficiency of the operational algorithm are directly related to the effort in deriving the algorithm.) Thirdly, the algorithm should allow significant discrimination among classes, i.e., it should classify accurately when possible. An algorithm which always indicated that a characteristic has a 50:50 chance of being present would be of no utility. Fourthly, the algorithm should, to some degree, allow conditional use of some chemical measurements which may not always be available. For example, if the molecular weight of an unknown is known, classification can often be considerably more accurate than when it is not available. Similarly, if some part of the spectrum were omitted (e.g. low m/z values), it would be useful if some estimate of the probabilities were still possible.

THE DECISION-TREE APPROACH

Recent developments in pattern recognition research provide an approach which satisfies these demands — a decision-tree approach. A very simple

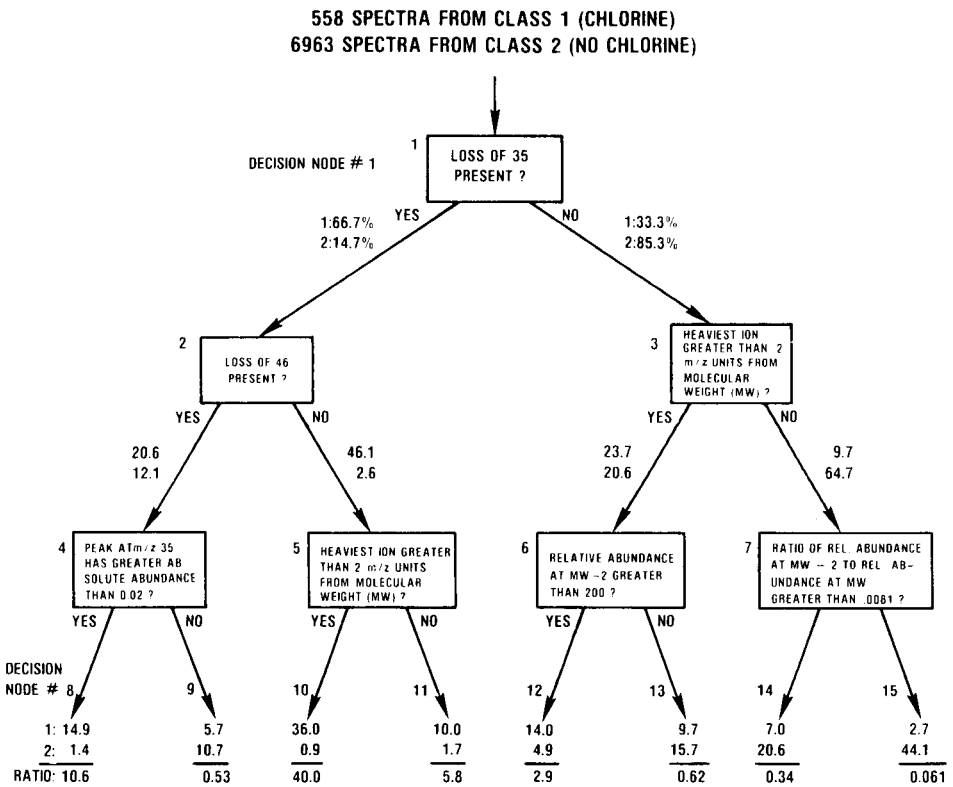


Fig. 1. Decision tree to indicate the likelihood of the feature "chlorine atoms present".

decision tree is illustrated in Fig. 1. In this tree, which will be fully discussed later in this paper, the molecular weight (MW) of the unknown is assumed to be available. The tree is designed to identify the likelihood of the feature "chlorine atom(s) present"; class 1 is "chlorine present" and class 2 is "no chlorine present". The numbers used in Fig. 1 are valid; they were obtained by applying the trees to spectra that were not used in its creation.

A decision tree consists of a series of decision nodes connected by branches. In order to classify a given spectrum, the decision tree is used in a sequential manner. In Fig. 1, a spectrum would first be examined to see if a loss of 35 a.m.u. was present. (Loss is defined as from the assumed molecular weight.) Depending on the answer, different aspects of the spectrum would be examined. If "yes", a loss of 46 is sought in the spectrum; if "no", it is necessary to determine whether the highest *m/z* value present in the spectrum is more than two *m/z* units from the MW. In either case, the tree is followed along branches determined by the spectrum until a terminal node is reached at the bottom of the tree. At the terminal node, the relative likelihood of the feature being present in the molecule (in this case chlorine atom(s)) is given.

The relative likelihood is expressed as follows. The first two numbers indicated at each terminal node in Fig. 1 are the percentages of each class which would reach that terminal node. For the terminal node 8, 14.9% of class 1 and 1.4% of class 2 have the features corresponding to that node (a loss of 35, a loss of 46, and a peak at 35 with greater absolute abundance than 0.02 relative to the base peak of the spectrum). Thus, if compounds were chosen from an environment where chlorine was as likely to be present as not, for the node the odds would be 10.6 to 1, i.e., 14.9/1.4, that chlorine was present. This ratio is listed at each terminal node in Fig. 1. Assumptions as to the relative abundance of chlorine in the environment, i.e., the a priori probability of chlorine being present in the compound (irrespective of the spectrum), can be incorporated into this measurement of likelihood if deemed appropriate. The use of a priori probabilities is discussed below. The immediate discussion of ratios assumes equal a priori probabilities of each class. When the ratio is less than 1, as in node 9, the percentage of class 2 compounds reaching the node is greater. Thus, for that node, the odds are 1.9 to 1 against the presence of chlorine (1/0.53 = 1.9).

In a classical pattern recognition approach, the terminal nodes at the bottom of the tree would simply be consigned to class 1 if the ratio exceeded 1, and to class 2 if not. From this classical viewpoint, nodes 5 and 7 achieve no purpose, as after the division indicated by those nodes, the classification (in the separation of classes sense) is the same irrespective of the outcome of that decision node. Thus, the present purpose of these nodes is to refine the estimate of confidence in the decision. It is particularly attractive to have very clear decisions such as in node 10, where the ratio is 40; here, one can be very confident that chlorine is present.

This particular small tree is intended to illustrate the decision-tree approach, not to reflect the full scope of the approach. A larger tree can use more characteristics of the spectrum and can be considerably more accurate. Further, the aspects of the spectrum considered could be considerably more complex than those illustrated. This tree is consistent with the objectives set earlier for a decision-aiding scheme based on pattern recognition. First, the approach yields an estimate of the relative probability with which the feature is present. Secondly, the decision tree is computationally efficient. The features are calculated only when required. In the tree illustrated, no more than three decision nodes are required to evaluate any given spectrum, although there are a great deal more decision nodes which are potentially usable. Thirdly, the algorithm has potential for very accurate classification, as the types of features actually used in the decision nodes can be specific to the chemical under consideration. More aspects of the spectrum can be considered in the decision tree without running into limitations as to the number of example spectra, because in classifying any given compound, only a few of these aspects of the spectra are considered in making the decision. Thus, this form of the decision rule tends to make maximal use of the data available.

Fourthly, the algorithm allows for conditional use of features of the chemical spectrum. If a node down the tree requires information that is not available, the procedure can terminate at that point and still indicate the likelihood of the feature being present. As an example, consider node 5. If the information required to make the decision at node 5 were for some reason not available, the decision could be made on the percentages prior to node 5. Since there are 46.1% of class 1 and 2.6% of class 2 prior to that node, the ratio would then be about 18 to 1 (i.e., 46.1/2.6).

This form of classification algorithm is easily interpreted, in contrast to many other approaches. This is a great advantage in generating confidence in workers who are unconvinced of the value of pattern recognition techniques. This aspect of the decision-tree approach also allows the results to be used as a teaching aid in the classification of mass spectra.

THE USE OF A PRIORI PROBABILITIES

The a priori probability of a chemical characteristic is the frequency with which that characteristic occurs in the universe of chemicals under test. For example, if chlorine is present in 25% of the chemicals tested, the a priori probability of chlorine present is 0.25.

In discussing the decision tree of Fig. 1, the ratios at terminal nodes were treated as if the two classes "chlorine present" and "chlorine absent" had equal a priori probabilities. Thus, a ratio of 10 at a terminal node implies odds of 10:1 that chlorine is present only if the a priori probabilities are equal. If, for example, the a priori probability of chlorine being present was only 0.1, i.e., 1:10, then the implication of a ratio of 10:1 at a terminal node is that there is a 50% likelihood of the sample containing chlorine.

Suppose that a characteristic (e.g., "chlorine present") has the a priori probability P of occurring (and thus, the probability $1-P$ of not occurring). Then let the a priori ratio $r(0)$ be defined as: $r(0) = P/(1-P)$. If P is 1:11, then $r(0) = 0.1$, and the odds are 1:10 that any given compound will contain chlorine, irrespective of any consideration of the mass spectrum. The ratio R' for a terminal node, taking into account the a priori probabilities, is the ratio R calculated by assuming equal a priori probabilities, multiplied by $r(0)$:

$$R' = R^* r(0) \tag{1}$$

Thus, in this example, $r(0) = 0.1$ and $R = 10$, yielding $R' = 1$ (i.e., equal odds that chlorine is present or absent, considering both the a priori probability and the mass spectrum).

This provides a means of including the a priori probability in determining the ratio which allows assessment of the likelihood of a given feature being present. This raises two questions, however: when is it appropriate to use the a priori probabilities and how are they to be determined?

The latter question will be discussed first. The decision as to the appropriate a priori probabilities depends on the universe of chemicals from which

the unknown is drawn. For example, if the compounds in the MSSS [6] are representative, the frequency with which the feature appears in that data base could be used. However, that universe of chemicals may not be representative of those from which the unknown was drawn in any particular application. It may, in fact, be quite difficult to decide what the appropriate a priori probability should be. To put this issue in perspective, the purpose of a given decision tree is to provide a comparative evaluation of the likelihood of one feature being present in contrast to other characteristics. Thus, what is sought is the a priori probability which reflects the relative likelihood of a given feature being present. If it can be assessed that, for example, nitrogen is twice as likely to be present as chlorine on an a priori basis, this can be reflected in the a priori probabilities.

The appropriate use of the a priori probabilities can now be discussed on the assumption that there is an acceptable assessment of these probabilities. Suppose that there is a feature that occurs relatively infrequently compared with other features (e.g., the compound is a polychlorinated biphenyl). If one simply looked at a list of adjusted ratios, i.e., those which incorporate the a priori probabilities based on eqn. (1), one would probably never consider that low-probability feature because the adjusted ratio, R', would be small, even when the unadjusted ratio, R, was relatively large. When R is large, it suggests that the unknown has a much larger chance than usual of having the feature in question. Similarly, for a common feature of a mass spectrum, a relatively high value of R' can result when R is less than 1, i.e., when the feature is less likely than usual to be present. Therefore it is suggested that a system for aiding in the classification of an unknown from its mass spectrum should indicate the unadjusted ratio, R, irrespective of the a priori probabilities.

If a priori probabilities are ignored entirely, then one may spend an inordinate amount of effort in considering relatively unlikely features simply because they are more likely than usual. Thus, it is also proper to consider R' in judging which features to explore. As a simple guideline, the features (chlorine present, aromatic ring present, etc.) can be ranked by the estimated R and R' separately. Then a particular feature will have a rank by each criterion; for example, "chlorine present" may be ranked the second most likely feature by R and the third most likely feature by R'. The features can then be examined in order of their average rank (e.g., "chlorine present" would have an average rank of 2.5).

This approach to the use of pattern recognition in classifying mass spectra depends on having an algorithm which yields a measure of the likelihood of a characteristic which is comparable to a similar measure of likelihood generated by a separate tree for a different characteristic. Furthermore a likelihood measure is required that can be modified by the use of a priori probability. The decision-tree approach and the ratio measure satisfy these requirements. A specific example of this approach is given below.

EXPERIMENTAL EXAMPLE

A pilot study was carried out on a file of 4816 mass spectra from an older version of the NIH/EPA/MSDC MSSS master data base of 39,509 spectra [6]. The purpose of the study was to classify the compounds into two categories: N, i.e. those containing nitrogen (actually 2162 spectra or 45% of the file), and NN, i.e. those not containing nitrogen (actually 2654 spectra or 55% of the file).

All the non-nitrogen compounds contained oxygen, and some of the nitrogen compounds contained oxygen. While this is not an absolutely unbiased data set, it is probably very close to one. The molecular weights of the compounds in the file ranged from 55 to 320. The latter weight was an arbitrary cut-off point and was used primarily to limit the size of the data base to under 5000 spectra in this illustrative pilot run. The *m/z* values were all integers between 1 and 320. Intensities of all *m/z* values were rounded off to the nearest integer (i.e. 0.4% = 0% and 15.7% = 16%) so that all intensities were between 0 and 100%.

First, the mass spectral data were separated, at random, into a learning set consisting of 75% of the spectra (1622 N, 1990 NN), and a test set consisting of the remaining 25% (540 N, 664 NN).

A tree-growing program (Classification and Regression Tree, CART, which is a software package owned by Technology Service Corporation), was applied to the learning set using three different sets of features. The first set (Case 1) was restricted to *m/z* values between 1 and 200, along with the corresponding intensities. The second set (Case 2) consisted of the *m/z* values from 1 to 200, and the losses (defined as the known molecular weight peak) for values from 0 (molecular ion) to 100 (molecular weight, 100), along with their corresponding intensities, were rounded off as stated above. Finally, the third set (Case 3) was formed by adding the molecular weight to the second set of features. In this last case, the molecular weight served only to tell whether the compound has an odd or even molecular weight. Each test set was run through each of the three trees constructed by CART. Table 1 shows the overall results when all spectra were classified. These results are not definitive, and should be interpreted only as a pilot test study. However, the high accuracy of classification achieved is very encouraging.

This pilot analysis provided some interesting results. It was unexpected that use of the *m/z* 1—100 values along with the 1—100 losses and the corresponding intensities instead of the 1—200 *m/z* values and their intensities gave no improvement in the overall classification rate.

However, when the use of the molecular weight was explored, it was found that including a single decision node, based on whether the molecular weight was even or odd, resulted in a dramatic increase in the classification rate, as in the number of decision nodes required. It should be emphasized that in the three cases tested the information content varied considerably. The first set required only *m/z* values and the second and third sets

TABLE 1

Percentage of correctly classified compounds (test set in parentheses)

Class	Features used		
	Case 1	Case 2	Case 3
Nitrogen	86.3 (79.5)	88.2 (81.9)	89.0 (85.0)
Non-nitrogen	86.3 (77.6)	82.9 (75.8)	95.0 (89.9)
Overall	86.3 (78.4)	85.3 (77.4)	92.3 (87.7)

required the molecular weight to calculate losses; the third set also used the molecular weight to differentiate between even and odd values.

APPROACHES TO TREE CONSTRUCTION

The methodology for deriving such trees is now discussed, guided by spectra of known classification. The basic objective of the decision-tree pattern recognition method is to classify the samples as accurately and efficiently as possible. Currently, there are no published approaches to finding the tree which truly minimizes the misclassification rate of known samples for a given size of tree. Payne and Meisel [7] have described a method for finding the smallest tree equivalent to a tree derived by any other method. Current work which extends this approach to direct minimization of the misclassification rate is nearly complete but computational limitations will restrict its applicability.

Current methodology uses a sub-optimal, one-node-at-a-time approach; thus, a tree is "grown" by adding branches. To begin, one must choose the first node. Two choices must be made: (1) the variable to be used in the node, and (2) the quantitative form of the decision made at that node. If there are N potential variables $x(1), x(2), ..., x(n)$, and at each node decision rules are restricted to the form $x(j) \geqslant T$, then j and T must be chosen at each node.

The most straightforward criterion for choosing j and T consists of minimizing the misclassification rate of known samples. Figure 2 shows how the node divides the samples of each class into those which satisfy the condition and those which do not. Assume, for clarity, that samples of Class 1 predominate in the left branch and that samples of Class 2 predominate in the right branch. Then the assignment of classification which minimizes the number of samples misclassified is to call samples in the left node Class 1 and to call those in the right node Class 2. This results in $N(\text{error}) = n(2, L) + n(1, R)$ samples being misclassified.

For every choice of j and T, a different value of $N(\text{error})$ can easily be computed. One can then choose the values of j and T which give the minimum value of $N(\text{error})$., i.e., the minimum number of misclassified samples.

Once the first node has been chosen, this approach can be repeated at any given node. For example, in Fig. 2, the same analysis can be applied to the

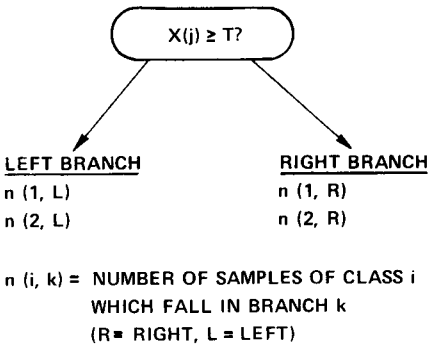


Fig. 2. Definition of notation for node-splitting.

$n(1, R)$ samples in Class 1 and the $n(2, R)$ of Class 2 in the right-hand branch, yielding (probably) different values of j and T and adding another branch to the tree.

A sequential procedure requires a stopping rule which indicates when the best values of j and T do not yield a statistically significant reduction in the misclassification rate. Most rules employed to date have been ad hoc, rather than based on theoretical significance computations. Typical approaches use a required minimum number of samples in each resulting branch and/or a maximum number of nodes in the tree.

The misclassification rate is the most easily understood criterion to use in choosing a decision rule at each node. Another possible criterion is an information measure; this was the measure used in deriving the nitrogen example above. Friedman [8] suggested the Kolmogorov—Smirnov distance. This distance was used in deriving the chlorine tree in Fig. 1. To date, no strong evidence of the superiority of any given measure has been found in this work.

Other methodology used in deriving trees has been outlined. Obviously, there are many variations on this theme. For example, the decision at each node need not be restricted to the use of only one variable; complex decision rules based on more than one variable can often be justified. When two variables are allowed, pairs of variables and the decision rule can often be chosen by looking at scatter plots as was done by Liles [9]; an interactive graphics package can be useful in this context.

Another extension of the methodology is to pick each node as optimal, looking ahead to the best rule at the resulting nodes. This two-steps-at-a-time approach involves much larger computation costs than the one-step-at-a-time approach.

One advantage of the iterative approach to tree construction described is that subjective considerations can be employed at each step. For example, at the top of the tree, one might limit consideration to variables which are seldom missing from a mass spectrum. Additionally, one might prefer a variable which came in a close second if it allowed a clearer physical interpretation than the variable which came in first.

REFERENCES

1 See, e.g., J. Zupan, Anal. Chim. Acta, 103 (1978) 273.
2 R. S. Heller, G. W. A. Milne, R. J. Feldmann and S. R. Heller, J. Chem. Inf. Comp. Sci., 16 (1976) 176.
3 S. Sasaki, CHEMICS-F in Information Chemistry, University of Tokyo, 1975, p. 227.
4 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
5 See, e.g., D. H. Smith (Ed.), Computer-Assisted Structure Elucidation, ACS Symp. Ser. 54, Washington, D.C., 1977.
6 Details of the latest MSSS data base available from: Dr. L. Gevantman, N.B.S., O.S.R.D., A323/221, Washington, D.C. 20234.
7 H. J. Payne and W. S. Meisel, IEEE Trans. Comput., C26 (1977) 905.
8 J. H. Friedman, IEEE Trans. Comput., C26 (1977) 404.
9 W. C. Liles, Interactive Decision Structuring System Manual, TSC publication, 1976.

Analytica Chimica Acta, 112 (1979) 417—421
Computer Techniques and Optimization

PARTIAL LEAST-SQUARES PATH MODELLING WITH LATENT VARIABLES

ROBERT W. GERLACH, BRUCE R. KOWALSKI* and HERMAN O. A. WOLD**

Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, Washington 98195 (U.S.A.)

(Received 20th June 1979)

SUMMARY

A partial least-squares treatment of multivariate data related through a complex model allows simultaneous evaluation of the interactions between large numbers of features. Results are given for a model in which water sources flow together; each source is represented by water quality data to allow the influence of the various sources to be evaluated with respect to their importance on the resulting flow downstream.

When the goal of a study is to understand the inter-relationship among several parts of a complex system, statistical procedures are often employed to analyse features from sets of samples collectively used to represent each part. All too often, the number of features and/or parts is larger than the number of samples, and many multivariate statistical procedures are not useful. A simple example is the case where one set of independent features has to be related to only one dependent feature by multiple regression analysis, represented as Model I in Fig. 1. The calculation can give a perfect but possibly meaningless fit if the number of features is greater than the number of samples. For the establishment of a predictive model, this problem is normally overcome by the use of stepwise regression analysis. However, in this analysis the regression coefficients are not informative with respect to understanding of the model, and the results do not provide information about the utility of the omitted features, which may be only a little less informative than those chosen to provide the best fit.

Consider the case where multiple blocks of data, each block consisting of several features obtained over several samples, are to be inter-related by a complex scheme or path model. When only one block of features is to be related to a second block of features (Model II, Fig. 1), a canonical correlation analysis [1] or target-transformation analysis [2] can be carried out. For more than two blocks of data, various multidimensional scaling techniques have been developed [3] which relate blocks of features along axes preserving the

**Present address: Department of Statistics, University of Uppsala, Uppsala, Sweden.

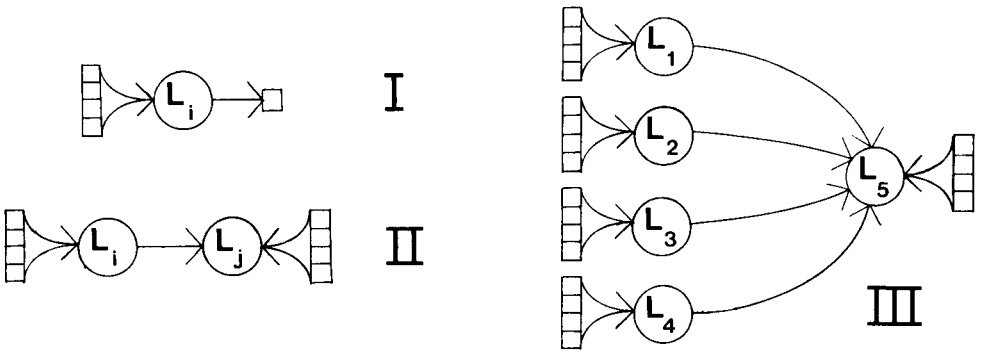


Fig. 1. Model I represents a multiple regression analysis of one matrix onto a single feature. Model II depicts two matrices of features related to one another. Model III shows the particular multi-matrix path model dealt with through a partial least-squares analysis. In Model III the four matrices on the left represent sources of flow in a watershed which combine to form the flow represented by the fifth matrix.

maximum amount of all interblock information at once. However, when not all interconnections between blocks are desired or relevant, more flexible methodology is required.

This methodology, herein called the PLS (Partial Least Squares) approach to path modelling by latent variables, has been developed by Wold [4–8]. This important tool allows blocks of features to be represented by unobservable or "latent" variables indirectly observed. The latent variables are then related to one another by a path predetermined by the user. The latent variables are found by an iterative procedure involving simple and multiple regression analysis so that they simultaneously and optimally (in the PLS sense) represent the measured features and provide the best fit to the path model. The method is so general that principal component analysis, multiple regression analysis, and canonical correlation analysis are included as special cases. The first application of this method to the physical sciences — an analysis of water chemistry measurements to assess the environmental impact of mine spoils drainage — is reported here.

In order to understand the impact of coal mining on local water quality, Skogerboe and co-workers [9] monitored several water quality parameters at numerous sites on Trout Greek in Colorado. Data taken at monthly intervals from October 1973 to July 1976 were provided by Skogerboe [10] for this study. Five sites best characterized the environmental impact and were selected for the present analysis. Site 1 was upstream from run-off influenced by spoils of the Midway Edna Coal Mine, which is adjacent to the stream. Sites 2–4 represented the run-off from strip mine spoils, i.e., mining activity during the periods 1930–40, 1940–50, and 1960 to the present, respectively. Run-off from these sites entered the stream in the order given above. Site 5 was downstream from the mine. Only 25 months of data were included in this study because occasionally several features at a site were not determined

in certain months. At each site, the data set was composed of eleven features: pH, Cl^-, SO_4^{2-}, Ca^{2+}, Fe^{2+}, K^+, Mg^{2+}, Mn^{2+}, Na^+, Zn^{2+}, and HCO_3^-, all but pH reported in mg l^{-1}. The final data set had ca. 8% of its values missing; these were filled in so as to minimize any deviation from the known data structure of a particular site [11].

The goal of the present work was to establish a path model based on all five sites. Each site, represented by a data matrix of 11 features sampled over 25 months, was used in the model as a separate entity. In the case considered here, the path model is clearly that shown as Model III in Fig. 1. The only relationship possible is that site 1, the upstream site, and sites 2, 3, and 4 mix to form site 5, the downstream site.

In order to consider the effect of all features at once, the method forms latent variables, $L_k = \Sigma_{i=1}^{N_k} a_{k,i} x_{k,i}$, at each site; here N_k is the number of features being considered at site k, $x_{k,i}$ is the value of feature i, and $a_{k,i}$ is a coefficient determined in the course of the analysis. The $a_{i,k}$ values for each of the upstream sites are estimated from a multiple regression of all the features at a particular site to the downstream latent variable, L_5, as indicated in Model III of Fig. 1. All coefficients $a_{k,i}$ are then scaled so that the latent variables L_k have unit variance. Next, L_5 is regressed on the upstream latent variables to estimate the $P_{k,5}$ values in the expression $L_5 = \Sigma_{k=1}^{4} P_{k,5} L_k$. When the $P_{k,5}$ and L_k values are used to estimate L_5, a multiple regression of the features of site 5 is done on it in order to estimate the $a_{5,i}$ values. From the newly found $a_{5,i}$ values, a new L_5 is formed which is scaled to unit variance, and the entire procedure is repeated until all $a_{k,i}$ and $P_{k,5}$ converge. All calculations were initiated with all $a_{k,i}$ and $P_{k,5}$ set to one. A similar series of path models can be developed to analyse any number of blocks of variables connected by any set of paths.

When all 11 features in each block were used, the calculation of Model III converged with an overall fit of 0.99. The square of the fit correlation coefficient, R^2, gives the relative amount of information at L_5 accounted for by the other four latent variables and is calculated from $R^2 = \Sigma_{k=1}^{4} P_{k,5} R_{k,5}$ where $R_{k,5}$ is the correlation between L_k and L_5. The site contributions to R^2 are given in Table 1. It can be seen that the good fit is primarily due to a strong relation between sites 4 and 5. The contributions of each individual feature to the fit were calculated and showed that the high correlation was due largely to a fit between HCO_3^- at site 4 and Ca^{2+} and Mg^{2+} at site 5. Although only a small amount of the total variance in all of the data is accounted for by this relationship, it is a rather striking one as HCO_3^- introduced by site 4 strongly buffers the Ca^{2+} and Mg^{2+} concentration.

A principal component analysis of the features at site 5 yielded two readily interpretable components. The first component represented the major salt load Ca^{2+}, Mg^{2+}, Na^+, K^+, SO_4^{2-}, and Cl^- on the creek and the second component represented primarily the trace metals zinc and manganese. Thus, a more directed analysis targeting on the principal components was suggested. Results of Model III calculations where L_5 is represented by an individual

TABLE 1

$(P_{k,s}) \times (R_{k,s})$ values for sites 1 through 4 and the corresponding R^2 for models where L_s is described in column 1; PCs are principal components

	Site				R^2
	1	2	3	4	
11 features	0.02	−0.04	0.06	0.93	0.97
PC 1	0.35	0.69	−0.16	0.03	0.91
PC 2	0.21	0.59	0.00	0.11	0.91
Cl^-	0.09	0.58	−0.08	0.29	0.88

principal component are also shown in Table 1. The first component is modeled by the upstream values of site 1 and the first source of mine drainage represented by site 2. These results indicate that site 2 has by far the most dramatic effect on water quality. Similar results were obtained for the second principal component with an additional smaller contribution from site 4.

Model III calculations were also done for the case when L_5 represents only one of the features from site 5, a non-iterative calculation. An example based on chloride is also shown in Table 1. Though the concentrations of chloride and the other major species at sites 2, 3, and 4 are comparable in magnitude [9], drainage from site 2 is obviously the dominant influence on the downstream chloride concentration. Drainage represented by site 4 also perturbs the downstream chloride concentration, most likely because it represents flow from the newest spoils, which have a greater concentration of the more soluble salts. The lack of influence from site 3 shows that drainage by this site is not different enough or large enough to alter the chloride composition set at site 2.

From the above it is clear that quantitative estimates of the effect of stream components contributing to the load at the downstream site can be made. In addition, detailed information can be obtained on each component. For example, for many species which have a high concentration at an upstream site but fail to be used in modelling the downstream site, it is believed that some form of buffering or precipitation action may be taking place. In these cases, the PLS analyses show where more extensive investigation should be directed if the stream chemistry is to be fully understood. Conclusions arrived at by using the PLS path modelling scheme are compatible with those obtained in this laboratory by using a battery of standard multivariate techniques on a more extensive data set of which the present data are a subset [12]

The above results show how PLS path modelling based on latent variables can provide insight into the inter-relationships between groups of features. It is especially important to note that the treatment of groups of features as a unit allows one to include many more features in the analysis than would normally be allowed by more conventional techniques when one is

confronted with limited quantities of data. In all the above calculations, 44 features were considered in sites 1 through 4 and consistently interpretable results were obtained with only 25 sets of data. This form of analysis can be a powerful aid to anyone confronted with blocks of features which are related to one another along a set of logical paths.

This work was supported in part by the Office of Naval Research. Many enlightening conversations with Maynarhs da Koven are gratefully acknowledged.

REFERENCES

1 R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York, 1977, p. 69.
2 P. H. Weiner, E. R. Malinowski and A. R. Levinstone, J. Phys. Chem., 74 (1970) 4537.
3 R. N. Shepard, A. K. Romney and S. B. Nerlove (Eds.), Multidimensional Scaling, Vol. 1, Seminar Press, New York, 1972.
4 H. Wold, in F. N. David (Ed.), Research Papers in Statistics, Festschrift for J. Neyman, Wiley, New York, 1966, p. 411.
5 H. Wold, in P. R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 1966, p. 391.
6 H. Wold, in H. M. Blalock (Ed.), Quantitative Sociology, Academic Press, New York, 1975, p. 307.
7 H. Wold, in J. Gani (Ed.), Perspectives in Probability and Statistics, Academic Press, New York, 1975, p. 117.
8 H. Wold, in R. Henn and O. Moeschlin (Eds.), Mathematical Economics and Game Theory; Essays in Honor of Oskar Morgenstern, Springer, Berlin, 1977, p. 536.
9 D. B. McWhorter, R. K. Skogerboe and G. V. Skogerboe, Environ. Prot. Technol. Ser., Publication 670, U.S.E.P.A., Washington, D. C., 1975.
10 R. K. Skogerboe, personal communication.
11 S. Wold, Pattern Recognition, 8 (1976) 127.
12 S. D. Brown, Ph.D. Thesis, University of Washington, Seattle, Washington, 1978.

Analytica Chimica Acta, 112 (1979) 423—442
Computer Techniques and Optimization

FACTOR ANALYSIS OF CHEMICAL MIXTURES

Non-negative Factor Solutions for Spectra of Cereal Amino Acids*

HARALD MARTENS
Norwegian Food Research Institute, P.O. Box 50, N-1432 Aas-NLH (Norway)
(Received 13th February 1979)

SUMMARY

From spectral data for a set of mixtures of unknown compounds, the spectra and the amounts of the pure components can be estimated without physical separation of the compounds. Spectra for the amino acid content of whole finger millet grain samples are used as the example. Different methods of factor analysis and weighting were compared. The number of relevant "pure components" (i.e. protein groups) was found to be 3 in finger millet grain grown under widely varying fertilizer conditions. Ranges of acceptable spectra of these"pure components" and ranges of their amounts were found by applying non-negativity criteria to the factor analysis solutions. The spectra were then estimated concisely by performing the factor analysis on the data scaled to different units in which all components except one remained constant and were excluded from the factor solution in turn. The amounts of the three "pure components" were estimated by multiple regression. Thus the rotationally ambiguous factor analysis solution was converted to a physically meaningful description of the unknown compounds in the mixtures.

Spectral data from mixtures readily lend themselves to factor analysis, whereby quantitative information about the underlying pure compounds may be obtained [1, 2], but the factor analysis solution contains a rotational ambiguity, which makes direct physical interpretation difficult. The present paper illustrates some methods of overcoming this problem.

The input data for a "mixtures model" may be of widely different nature as long as they behave additively, from digitized visible absorption spectra of coloured compounds to discrete monomer composition "spectra" of polymers. Data for the latter type, amino acids from proteins, are used as illustrations here. The mixtures which are suitable for this kind of analysis may be composed of a few (typically 2—4) pure unknown compounds varying more or less independently from sample to sample, or they may consist of a larger number of pure unknown compounds distributed in a few stoichiometrically well defined groups, each group appearing as a "pure component". Thus a "pure component", as used here, may imply a single compound or a well de-

*Presented at the International Conference on Computers and Optimization in Analytical Chemistry, Amsterdam, April 1978.

defined group of compounds. The amino acid "spectra" were obtained for a very complex system, barley grain. In cereal seeds, very many different types of protein are present, but their quantitative variations are such that they may be grouped into a few protein groups.

For many types of measurement, interaction effects between variables are negligible. Such measurements may be described by the following model of linearly additive mixtures:

$$x_{ik} = \sum_{j=1}^{n_c} a_{ij} \cdot p_{jk} + e_{ik} \tag{1}$$

where x_{ik} represent experimental data for variable i ($i = 1, 2, \ldots, n_a$); a_{ij} is the relative contribution of variable i in pure component j ($j = 1, 2, \ldots, n_c$); p_{jk} is the amount of pure component j in sample k ($k = 1, 2, \ldots, n_s$); and e_{ik} is the error in x_{ik}, plus possible interaction terms, non-linearity terms and other model errors.

In matrix form, a simpler equation is obtained:

$$X = AP + E \tag{2}$$

where ${}_{n_a}X_{n_s}$ consists of columns of spectra of n_s mixtures; ${}_{n_a}A_{n_c}$ consists of columns of spectra of n_c pure components; ${}_{n_c}P_{n_s}$ consists of columns of amounts of the n_c pure components in the n_s mixture samples; and ${}_{n_a}E_{n_s}$ represents the errors of X.

Each pure spectrum may for practical purposes be normalized so that it adds up to 100 ("unit I" data). If e_{ik} is assumed to have an expectancy of zero, the sum of the amounts of the pure components in each mixture k, should be related to the sum of its measured spectrum by

$$\sum_{j=1}^{n_c} p_{jk} = \sum_{i=1}^{n_a} x_{ik}/100 \quad (k = 1, 2, \ldots, n_s) \tag{3}$$

For many types of spectral measurements, the following non-negativity constraints apply:

$x_{ik} \geqslant 0$ } non-negative spectra (e.g. amino acid compositions) (4a)

$a_{ij} \geqslant 0$ } (4b)

$p_{jk} \geqslant 0$ non-negative amounts (e.g. of proteins) (4c)

Regressions from the mixtures model

If the n_a-dimensional spectra in A (eqn. 2) are known for the n_c compounds present in one or more mixtures, X, the amounts of the compounds, P, may be estimated, as long as $n_c < n_a$ [3—5]. In the regression method used here, each variable is first weighted by some estimate d_i, which is inversely proportional to its analytical error: $w_{ik} = x_{ik}\, d_i$. The mixture model (eqn. 2) becomes $W = GP + F$, where F is the weighted error matrix, and $g_{ij} = a_{ij}\, d_i$. P is estimated by, e.g.,

$$\hat{P} = (G^T G)^{-1} \cdot G^T W \tag{5}$$

(Alternative methods of weighting are available [6]. In order to ensure non-negative P (eqn. 4c), iterative fitting [1] may be used).

Equation (3) should be satisfied by $\hat{P}$. A large inequality in eqn. (3) for a sample is a practical indication that the model does not fit the data of that sample, because of errors either in the model, A, or in the sample, x_k. $\hat{P}$ represents the estimated "molar" amounts of the pure components. If weight ratios are required instead, the differences in "molar weight" between the components must be allowed for. In the present study the protein groups were assumed, for simplicity, to have identical molar weight per 100 amino acid units.

Equation (5) may be used for simplification of many analytical situations, e.g. in food chemistry. By submitting one or a few samples of mixtures to chemical fractionation, the spectra, A, of the pure components may be obtained. Their apparent amounts, P, may then be estimated in other mixtures, thereby reducing the need for quantitative chemical fractionation of the pure components in every mixture. Conversely, if the amount P of the n_c components is known in n_s mixtures X, the data for one or more spectrum variables A may be found by an analog method as long as $n_c < n_s$. However, if very little is known about the underlying components that constitute a set of mixture data, these simple regression techniques are not applicable. Instead the technique of factor analysis is an alternative.

Factor analysis model

The model of mixtures (eqn. 2) is in principle identical with the factor analysis model:

$$X = BS + E \tag{6}$$

where ${}_{n_a}X_{n_s}$ is the experimental data matrix defined in eqn. (2) now described by n_f factors; ${}_{n_a}B_{n_f}$ consists of n_f columns, factor loading vectors, analogous to abstract "spectra"; ${}_{n_f}S_{n_s}$ consists of n_f rows, factor score vectors, each giving the "amount" of the respective abstract "spectrum" in the n_s mixtures; and ${}_{n_a}E_{n_s}$ represents the residuals in X.

In factor analysis the data matrix X is represented by a simplified model, consisting of the product of two smaller matrices, BS. Several different procedures are available [7, 8], but they all have the primary aim of representing the experimental data by a minimum number of factors n_f, but yet with the maximum amount of information retained (rank reduction, which is the first step in factor analysis). In the case of factor analysis of spectra from chemical mixtures, the solution obtained should be rotated from the initial, somewhat arbitrary solution to factors which are physically interpretable in terms of loading "spectra" and score "amounts" vectors (axis rotation, the second step in factor analysis).

Geometrical representation

An experimentally obtained spectrum, measured as a column of n_a variables,

constitutes a single point in an n_a-dimensional space. Thus a series of n_s mixture spectra, X, may be represented by n_s points in this space. Geometrically, the first step in factor analysis may be visualized as finding the smallest n_f-dimensional subspace which gives an acceptable description of these points in terms of signal/noise separation.

The abstract factor analysis yields, for each factor dimension j, a loading vector b_j which describes a direction axis in the n_a-space, and a score vector $s_j{}^T$, which describes the positions of the n_s mixture points along this factor axis. The rather abstract first-step factor solution BS, representing directions B and positions S, may in itself be an aid for interpretation of a large table of spectra. A plot of the first two score vectors s_j usually gives a simplified visualization of the experimental data X.

The ambiguous factor analysis solution

In the second step of factor analysis, the abstract factor solution BS should ideally be converted to the model of mixtures of actual unknown compounds, AP. Indeed the abstract factor solution BS obtained can be transformed to an estimate of the "true" unknown model, AP. But BS can also be transformed to very many other related solutions which give equally good fits to the data X. In addition to the more trivial affine transformation phenomenon (e.g. 1000 g = 1 kg), the original orthogonal factor solution BS may be rotated at will, orthogonally or obliquely [8]; thus $BS = BH \cdot H^{-1}S$, where ${}_{n_f}H_{n_f}$ is any non-singular matrix, causing an orthogonal or oblique rotation of the loading axes. Such rotations may be visualized on a geographical map as, for example, rotating the usual north/south, east/west axis system to north-west/south-east, east/west (oblique solution) or north-west/south-east, south-west/north-east (orthogonal solution).

Factor rotation methods

If a rotated factor analysis solution is to represent a physically meaningful mixture model, both A and P must be non-negative (eqn. 4), and reasonably concisely determined. The present work was designed to propose methods for the rotation of an abstract factor solution BS into a meaningful mixture model AP, giving the spectrum of the unknown pure components A, and the amounts of each of these pure components P, in each mixture.

The practical example used here involves a complex system: amino acid data from proteins in finger millet cereal grains [9]. The methods have also been applied to other complex food science data, e.g. barley protein data [10] and fatty acid data from mixtures of lipids [11]. Factor analysis and related methods have been used in chemistry for the analysis of multivariate spectral data obeying eqn. (1) [12—26]. However, little has been published on the analysis of discrete "monomer" spectra from polymer mixtures, e.g. amino acid spectra of protein mixtures [5, 11, 27—30].

Several papers have been published on the rotation of an abstract principal component factor solution to obtain non-negative loadings (spectra, A) and

scores (amounts, P). Lawton et al. [13—15] estimated possible ranges of the absorption spectra of two coloured components from light spectra of mixtures of them. Ohta [31] similarly developed spectral ranges for three compounds from their mixture spectra. The width of the spectral ranges obtained and thus their usefulness depend on the degree to which the mixture data fill the n_f-subspace, as will be shown below. Simulations [32] on data from Lawton and Sylvestre [13] showed that the ranges became narrower and the solution thus more "concise" with increasing errors in the input data — an obvious paradox.

Leggett [19] used the existence of equilibrium equations in addition to the non-negativity criteria, to determine the exact absorption spectra of two pH indicators, thus giving an example of the use of additional information with the factor analysis to obtain physically meaningful factor solutions.

Several workers have converted an initial, abstract factor solution to a physically meaningful solution by target transformation [21]. This technique is probably the most direct way of obtaining physically meaningful factor solutions, but it requires good hypotheses for the vectors of either the spectrum (A) or the amounts (P) for some or all of the unknown compounds. If such hypotheses are unavailable and equilibrium relations between the pure components are unknown or non-existent, rotation of the factor analysis model to a concise, physically meaningful solution has hitherto been difficult. The "constant-factor method" and the "simplex intersect method" allow the problem to be overcome for certain types of spectral mixtures.

A protein is a genetically-fixed combined "spectrum" of about 20 different amino acids, most of which may be quantified by hydrolysis and subsequent column chromatography. Whole cereal grain contains a very large number of different proteins, each with a different spectrum. These proteins have widely different functions (enzymes, structure proteins, storage proteins, etc.) and have different nutritional value. From conventional solubility fractioning studies they are known to respond differently to growth conditions [33, 34]. However, such studies are very time-consuming and somewhat ambiguous, and there is a need for methods of obtaining a simple system of protein "fractions" which relate directly to the nutritional value of the seed, rather than to more or less irrelevant solubility or mobility characteristics.

THEORY

Transformation of amino acid spectra before factor analysis

Centering the variables. Factor solution of the original, non-negative data, X, requires one factor to represent the distance from origin to the mean point of the "cloud" of data points in the n_a-space, and n_c factors to span the space of the n_c independently varying protein groups, yielding a total number of factors $n_f = n_c + 1$. However, the origin may be moved to the mean point by means of the equation $y_{ik} = x_{ik} - \bar{x}_i$, where $\bar{x}_i = \Sigma_{k=1}^{n_s} x_{ik}/n_s$. Instead of eqn. (6) the factor analysis model then becomes:

$$Y = BS + E \tag{7}$$

Factor analysis of Y instead of X yields a first-step factor solution somewhat easier to interpret, because $n_f = n_c$ and a large, trivial variance is withdrawn with the mean vector $\bar{x}$. This vector may be added back into the factor solution in the subsequent rotational procedures (step two) in order to obtain a meaningful spectrum:

$$a_{im} = \bar{x}_i + \sum_{j=1}^{n_f} b_{ij} \cdot s_{jm} \tag{8}$$

where m is the spectrum number.

Weighting the variables. Principal component analysis was the primary rank reduction used because of its computational simplicity and easy geometrical interpretation. However, this factor analysis method is somewhat sensitive to noise in the data [7, 23], especially in variables with the largest variances.

Glutamic acid is usually present in cereal proteins at concentrations 20 times higher than those of some other amino acids, and with a correspondingly higher absolute analytical uncertainty. Therefore a weighting of the amino acid data was introduced so as to ensure that each amino acid variable pulled the factor loading axes in the direction of that particular amino acid in proportion to its relative information content.

Conventionally, standardized variables are used in principal component analysis, but this implies giving equal weight to variables with low and high signal/noise ratios. Therefore a different procedure was adopted for comparison. The amino acid data were assumed to contain approximately equal relative levels of analytical uncertainty for the different amino acid variables. Therefore the mean value for each amino acid was used as an estimate proportional to its analytical uncertainty $d_i = 1/\bar{x}_i$ $(i = 1, 2, \ldots n_a)$ where $\bar{x}_i = \Sigma_{k=1}^{n_s} x_{ik}/n_s$. The data submitted to factor analysis were thus weighted analogously to the equation $w_{ik} = x_{ik}\, d_i$:

$$z_{ik} = y_{ik} \cdot d_i = (x_{ik} - \bar{x}_i) \cdot d_i \tag{9}$$

or expressed in terms of the original data:

$$z_{ik} = x_{ik}/\bar{x}_i - 1 \tag{10}$$

The transformed data matrix, Z, is then factor-analyzed in the same way as Y in eqn. (7). The elements of the weighted loading matrix, L, are defined as $l_{ij} = b_{ij}\, d_i$, which in the present case becomes $l_{ij} = b_{ij}/\bar{x}_i$. Other estimates of the weights d_i than those expressed by $d_i = 1/\bar{x}_i$ may be used.

The complete factor analysis model is analogous to the equations $W = GP + F$ and $Y = BS + E$ mentioned above:

$$Z = LS + F \tag{11}$$

where L consists of the weighted factor loading vectors, and F is the weighted residual matrix. The elements in a spectrum vector a_{im} are calculated by eqn. (8) from the loading vectors (directions) in L and scores (positions) in S after the conversion of L to B by $b_{ij} = l_{ij}/d_i$.

Factor analysis: weighted principal component analysis

The transformed data matrix, Z, was submitted to singular value decomposition, whereby

$$Z = L \cdot \Lambda \cdot V + F \tag{12}$$

where Z and F are as defined in eqn. (11); ${}_{n_f}\Lambda_{n_f}$ is the diagonal matrix containing the square roots of the n_f first eigenvalues of ZZ^T; ${}_{n_a}L_{n_f}$ contains the n_f first eigenvectors of ZZ^T; and ${}_{n_f}V_{n_s}$ contains the n_f first eigenvectors of ZZ^T. For plotting purposes, the scores, V, are premultiplied by Λ, the relative size of the factors, to yield the score vectors S of eqn. (11):

$$S = \Lambda \cdot V \tag{13}$$

Principal component analysis (eqns. 12, 13) is sensitive to the weighting of the variables (eqn. 9). An alternative factor analysis method, maximum likelihood factor analysis [8] operates on the same basic factor model (eqn. 7), but this method is insensitive to variable weighting. Thus the computationally simple former method may be checked by the more complicated latter method; the factor solutions $Y = BS + E$ should be similar. Plots of the scores S along, for example, the two first factors will illustrate the relative importance of the two factors. The number of relevant factors, n_f, must be determined. This involves statistical problems which will not be discussed here [24, 35, 36]. Two pragmatic methods were used. First, the relative deviation, $CV_{i,\,n_f}$ for each amino acid after n_f factors was calculated from

$$CV_{i,\,n_f} = \left[\left(\sum_{k=1}^{n_s} f^2{}_{ik}\right) \Big/ (n_s - n_f - 1)\right]^{\frac{1}{2}} \quad (n_f = 1, 2, \ldots\ldots) \tag{14}$$

When these relative deviations approached the expected level of experimental errors for all the amino acids, the approximate number of factors n_f became available. Secondly, the exact number of factors n_f was chosen after a close inspection of two-dimensional loading and score plots. Factors with small eigenvalues and apparently irrelevant information content could thus be ignored.

After this first rank-reduction step had been done, the factor analysis solution (eqn. 11) was rotated into a physically meaningful and unambiguous mixture solution (eqn. 2). Different methods were then required.

Rotation methods: simplex theorem

Factor analysis of centralized data (eqn. 11) yields one factor for each independently varying protein group. If the mixture data, X, are given in a unit yielding a constant sum (e.g. gram of amino acid/100 gram of the n_a amino acids included, "unit I"), one of the amino acids is a linear function of the $n_a - 1$ other amino acids and one of the protein groups is likewise a linear function of the $n_c - 1$ other protein groups (eqn. 3). This loss of information is, however, offset by the applicability of the simplex theorem for the identification of the pure protein groups.

The simplex theorem states that for "unit I" data spectra of mixtures of n_c

pure components (i.e. protein groups) must fall inside an $(n_c - 1)$-dimensional polyhedron (simplex) in the n_a-dimensional variable space [37]. The n_c pure components constitute the n_c extreme points ("corners") of this simplex. In practice, this means that for amino acid data scaled to yield a total of 100% in each spectrum, all mixtures of two pure protein groups must lie on the straight line segment connecting the two pure protein groups in the n_a-dimensional amino acid space. Mixtures of three such groups must lie inside the triangle connecting the three groups, etc. Conversely, if factor analysis with just one factor yields a satisfactory description of the amino acid spectra of a set of grain samples, this implies that only two main independent protein groups are present. Generally, n_f factors imply $n_c = n_f + 1$ independent groups.

Non-negativity requirements. The concise positions of the simplex corners (i.e., the amino acid spectra of unknown major protein groups) are not found by the first rank-reduction step of factor analysis, owing to rotational ambiguity. Some information may be obtained from the non-negativity criteria.

(a) For the relative amounts of all the protein groups in all the samples to be non-negative (eqn. 4c), the n_f-dimensional unknown simplex must encompass every mixture point. Thus the "cloud" of mixture data points represents the "inner limits" of the scores of the $n_f + 1$ protein groups.

(b) For the amino acid spectrum of all the protein groups to be non-negative (eqn. 4b), all the corners of the simplex must lie inside the first quadrangle of the n_a-space. These "outer limits" of the scores can be easily estimated by a trial-and-error search for scores (eqn. 8) or by a more analytical approach. These inner and outer score limits correspond to an easily visualized version of the method of Lawton et al. [13] and Ohta [31] (see below). A more or less narrow range of scores is thus obtained for each protein group. The corresponding ranges of amino acid spectra, A, are calculated by eqn. (8), and the corresponding ranges of amounts P are then found by eqn. (5). However, the ranges for A and P thus obtained by non-negativity analysis may not be sufficiently narrow to yield satisfactory accuracy in the estimation of A and P. Two different approaches for a more concise estimate may therefore be needed.

The constant-factor method. Factor analysis on centered data (eqn. 7) yields one factor for each independently varying protein group in a set of mixture samples. If the input spectral data are given in a unit in which one of the protein groups remains in a constant amount in all the samples, the effect of this constant group will be withdrawn in the centering process instead of yielding a factor. If there are only two main protein groups present in a set of samples, and one of these groups remains constant, then factor analysis yields only one factor, representing "that which varies" in the samples. The loading vector of this factor will be non-negative, so that it can be deweighted and normalized to yield a sum of 100%, whereby the amino acid spectrum of that protein group is found directly. Such a single-factor solution may even be estimated by simple difference analysis instead of factor analysing; subtracting one mixture spectrum in the above-mentioned unit from another one,

and normalizing the difference vector to a sum of 100% [10] yields a crude estimate of the spectrum directly.

It has been shown chemically [33, 34] that non-storage proteins in cereal grain usually remain at constant levels on a dry matter basis when the total protein content is increased because of fertilizers, etc. Thus in grain samples containing only two protein groups (the storage and the non-storage group) the amino acid spectrum of the storage protein group can be estimated directly as follows. The "unit I" amino acid data (adding to 100 in each spectrum) are converted to a dry matter basis ("unit II data"):

$$x''_{ik} = x_{ik} \cdot h_k \tag{15}$$

where x''_{ik} are "unit II" data of amino acid i, sample k; x_{ik} are "unit I" data of amino acid i, sample k; and h_k is the percent total protein in dry matter (here, Kjeldahl-N × 6.25). These "unit II" data are submitted to factor analysis by eqns. (10)—(13). The non-storage protein group spectrum may be similarly estimated after conversion of the input data to a different basis where the storage protein group remains constant ("unit III"). In this study, the sum of water-soluble albumins, globulins and insoluble residual protein contents from the low-nitrogen fertilizer grain samples [39] was used as an estimate of the amount of non-storage proteins in the grain. The amount of storage protein was estimated by the difference between the amounts of total and of non-storage proteins. "Unit III" was thus defined:

$$x'''_{ik} = x''_{ik}/(h_k - c) = x_{ik} \cdot h_k/(h_k - c) \tag{16}$$

where x'''_{ik} are "unit III" data of amino acid i, sample k; and c is the amount of non-storage proteins in the dry matter (here constant at 3%).

If a set of mixtures contain three or more components, a sub-set where only two components are present must be found before the constant-factor method can be used. In the present case, the grain samples could be divided into a low-nitrogen fertilizer sub-set and a high-nitrogen fertilizer sub-set.

The simplex intersect method. If it is assumed that a set of protein mixtures can be explained by two factor axes, the set will contain three protein groups (1, 2 and 3). The three protein groups constitute the unknown corners of some triangle in the n_a-space. If some of the samples in this set contain only groups 1 and 3, these samples lie on the 1—3 side of the unknown simplex triangle. If some of the other samples contain only groups 2 and 3, they lie on the 2—3 side of the triangle. If the spectral data of these samples along the two sides are sufficiently different and sufficiently accurate, the scores for 3 are found at the intersection of these two sides of the triangle. In a two-dimensional plane the two lines, 1—3 and 2—3, will always intersect. In a higher-dimensional plot, they will probably not quite intersect, e.g. because of analytical errors in the spectra. Then an iterative least-squares algorithm can be used to find the positions of closest approach between the two lines, instead of the true intersect [38]. The spectrum A of protein group 3 may thus be found by eqn. (8). Similar reasoning can be developed for the other two

corners, and for mixtures of four or more protein groups. The intersect method requires certain types of mixture patterns in the data, and can only be used in systems with 3 or more pure components.

RESULTS AND DISCUSSION

Input data

An East African finger millet (*Eleusine Coracana v. Gaertner*) was grown under widely varying conditions of nitrogen (N) and sulphur (S) fertilizer in a total of 27 field and pot experiments. The seeds were analyzed [9] for total crude protein content (Kjeldahl-N × 6.25) and for 16 amino acids. The results from the field experiment and some of the pot experiments are given in Table 1; the total number of samples used in the calculations was $n_s = 27$. It was originally concluded [9] that the finger millet amino acid spectrum varied strongly with fertilizer condition. Increasing nitrogen fertilizer levels resulted in increased protein content of the dry matter of the seed, but the seed protein simultaneously lost some of its nutritional value, because of shifts in its amino acid spectrum (decreasing amounts of LYS, MET, CYS, etc. and increasing amounts of GLU). High N:S fertilizer ratios resulted in S-deficiency syndromes of the plants (yellow leaves) and significant levels of free amino acids, mainly ASP, in the seed.

The protein data in Table 1 were subjected to factor analysis with the aim of characterizing the major underlying protein groups in the seed.

Comparison of methods of rank reduction by factor analysis

In Table 2, columns 8 and 9 show the means and relative standard deviations for the 16 amino acids, to indicate their differences in level and variability. Each of the sample spectra was normalized to "unit I", centralized, weighted (eqn. 10), and submitted to principal component analysis (eqn. 12). The percentage variance from the two first factors is given in Table 2, column 1. This number of factors was chosen as the most illustrative for comparison of the factor analysis methods.

Significant amounts of free ASP, from the S-deficiency, e.g. in columns 4 and 5 (Table 1), create "noise" in the factor solution. This complication can be avoided either by eliminating ASP completely from the normalization, weighting and principal component analysis ($n_a = 15$, $n_s = 27$, column 2, Table 2) or by eliminating the four samples with the most drastic S-deficiency ($n_a = 16$, $n_s = 23$, column 3, Table 2). A similar increase in the percentage of explained variances (Table 2, bottom) is obtained in both cases, as expected.

The effect of weighting is evident from a comparison of column 3 (mean-weighted solution, $d_i = 1/x_i$), column 4 (the corresponding unweighted solution, $d_i = 1, 0$) and column 5 (the standardized solution, $d_i = 1/\text{s.d.}_i$). The unweighted solution gives a 100.0% fit for the amino acid with the largest values, GLU, and a poorer fit for small amino acids like HIS, CYS and MET.

TABLE 1

Input data. Crude protein contents and amino acid spectra of finger millet as influenced by nitrogen and sulphur fertilizers (Amino acid data are given as percent of recovered amino acids in crude protein (N × 6.25). Columns 2—12 show results from some of the 22 pot samples. Columns 13—17 show results for all field samples, representing increasing nitrogen fertilizer levels.)

Nitrogen (ppm)	10			250				1000								
Sulphur (ppm)	5	25	125	5	12.5	25	125	50	100	200	500					
S:N ratio	1:2	5:2	25:2	1:50	1:20	1:10	1:2	1:20	1:10	1:5	1:2					
Crude protein (%)[a]	5.16	5.16	4.80	7.32	7.08	7.22	7.04	14.81	15.31	14.81	14.50	6.0	7.5	9.0	11.6	13.5
Lysine	3.61	3.57	3.58	2.67	2.83	2.99	3.06	1.84	2.00	1.99	2.07	3.01	2.66	2.42	2.03	1.80
Histidine	2.39	2.34	2.30	2.48	2.40	2.41	2.34	2.30	2.34	2.49	2.39	2.25	2.32	2.36	2.27	2.22
Arginine	5.54	5.80	5.94	3.62	4.14	4.52	4.64	3.46	3.65	3.82	3.99	4.78	4.38	4.17	3.58	3.36
Aspartic acid	7.42	8.01	7.81	16.95	10.10	7.24	7.00	6.56	6.33	6.18	6.24	6.87	6.63	6.36	6.07	5.78
Threonine	4.41	4.63	4.41	3.23	3.77	4.16	4.38	3.86	3.86	3.83	3.99	4.46	4.52	4.65	4.34	3.97
Serine	5.37	5.37	5.26	4.74	5.19	5.21	5.13	5.85	5.58	5.65	5.75	5.42	5.88	5.91	5.79	5.54
Glutamic acid	16.95	16.75	16.55	21.50	21.90	21.75	21.20	25.90	27.45	26.50	25.75	19.10	20.90	23.40	25.50	25.80
Glycine	4.92	5.13	5.04	3.60	4.01	4.26	4.43	3.04	3.14	3.13	3.21	4.34	3.93	3.62	3.16	2.87
Alanine	5.91	6.14	6.03	5.09	5.64	5.91	6.31	5.87	5.82	5.74	5.99	6.08	6.11	6.28	6.12	5.78
Cystine	2.24	2.19	2.29	1.05	1.57	2.17	2.24	1.58	1.60	1.68	1.68	2.06	2.05	2.03	1.85	1.61
Valine	5.69	5.73	5.66	5.69	6.24	6.43	6.32	6.97	7.16	7.03	6.89	6.20	6.37	6.76	7.01	6.91
Methionine	3.54	3.51	3.43	1.41	2.07	3.58	3.56	2.19	2.36	2.47	2.57	3.54	3.52	3.38	2.79	2.38
Isoleucine	3.77	3.91	3.76	3.93	4.47	4.35	4.50	5.09	4.95	4.91	4.86	4.20	4.45	4.81	4.74	4.75
Leucine	8.31	8.34	8.15	8.70	9.43	10.00	9.62	11.40	11.20	11.15	10.90	9.17	9.95	10.95	11.05	11.05
Tyrosine	2.93	3.01	2.75	3.10	3.53	3.81	3.69	4.11	3.83	3.69	4.08	3.32	3.59	3.61	3.83	3.72
Phenylalanine	4.97	5.02	4.76	4.45	5.03	5.34	5.42	5.48	5.48	5.58	5.36	5.23	5.59	5.77	5.54	5.26
Sum	87.97	89.45	87.72	84.38	90.31	94.13	93.84	95.50	96.75	95.84	95.69	90.03	92.85	96.48	95.67	92.80

[a] %N × 6.25.

TABLE 2

Comparison of different factor analysis methods. Rank reduction of finger millet amino acid data, given as percent amino acid variances explained by two factors (Factor analysis algorithms: principal component analysis (p.c.a.) and maximum likelihood analysis (m.l.a.).)

Factor analysis	1	2	3	4	5	6	7	8	9
	P.c.a.	P.c.a.	P.c.a.	P.c.a.	P.c.a.	M.l.a.	P.c.a.	Mean	(
Sum[a]	100	100	100	100	100	100	Recovered N × 6.25		
Weight[b]	Mean[c]	Mean[c]	Mean[c]	1.0[d]	*S.d.*[e]	*S.d.*[e]	Mean[c]		
n_a[f]	16	15	16	16	16	15[i]	16	16	
n_s[g]	27	27	23	23	23	23	23	23	
LYS	98.5	99.3	99.2	98.3	99.2	99.3	98.9	2.89	
HIS	60.1	74.6	62.5	59.0	61.0	64.4	12.7	2.55	
ARG	93.5	96.2	98.1	98.3	95.5	96.3	97.1	4.67	
ASP	98.3	(72.3)[h]	97.1	97.8	95.2	96.5	92.2	7.22	
THR	90.9	91.9	86.3	90.5	94.4	85.8	76.4	4.58	
SER	49.7	6.8	3.0	14.2	56.5	3.1	58.5	6.01	
GLU	98.2	98.8	98.9	100.0	98.9	98.7	97.6	23.77	
GLY	99.0	99.2	99.4	99.0	99.1	99.7	99.1	4.25	
ALA	91.9	91.4	87.0	89.9	90.0	86.9	58.7	6.47	
CYS	97.8	98.3	98.2	95.2	92.4	98.0	97.6	2.13	
VAL	95.5	93.7	96.4	96.2	96.0	96.6	96.5	6.95	
MET	93.1	99.4	99.2	94.7	84.3	98.1	98.9	3.36	
ILE	88.6	83.5	90.1	88.8	90.1	89.9	91.3	4.82	
LEU	98.4	97.8	98.7	98.4	98.7	98.9	97.7	10.77	
TYR	81.4	76.4	82.4	81.6	81.5	79.2	88.0	3.87	
PHE	78.1	39.7	57.5	63.9	70.4	(58.2)[i]	83.2	5.73	
Total (weighted)[j]									
Factor 1	66.3	86.6	93.5	97.2	79.6	82.6	90.8		
Factors 1 + 2	95.9	97.0	97.2	98.8	87.7	86.1	96.0		

[a]Sum of the n_a amino acids used, (for "unit I" data sum = 100).
[b]Divisor of the amino acids, d_i^{-1} (eqn. 9).
[c]Mean of each amino acid over the n_s samples used: $\bar{x} = \Sigma_{k=1}^{n_s} x_{ik}/n_s$.
[d]1.0 = unweighted.
[e]Standard deviation from the mean.
[f]Number of amino acids used in normalization and factor analysis.
[g]Number of grain samples used in normalization and factor analysis.
[h]ASP excluded intentionally from the normalizations and factor analysis to avoid the free amino acid.
[i]PHE excluded by the computer program because of linear dependence on the other 15 amino acids.
[j]Measure of the overall fit of the one-factor and two-factor solutions to the data.

In contrast, the standardized principal component analysis appears to give too high a weighting to amino acids with small signal/noise ratios like SER and PHE, with a corresponding loss of fit for many other amino acids.

The weighting problem can apparently be eliminated by using maximum

likelihood analysis for rank reduction. This method is insensitive to variable weighting [7]. Column 6 gives the corresponding two-factor fit [40] of normalized, standardized data. As can be seen, this solution is very similar to the best results by principal component analysis (column 3) despite the lack of proper weighting. Finally, principal component analysis of un-normalized data, directly as given in Table 1, with sums between 84 and 96 (column 7), is compared to the corresponding analysis of "unit I" data (sum = 100, column 3). Only small differences are observed in the fit, except for the lowest signal/noise amino acids (HIS, TYR, SER, ALA and PHE). Thus the method employed in column 3 was used in subsequent work, since it is computationally simple, gives satisfactory weighting and satisfies the simplex normalizing criterion.

Table 3 gives a closer description of principal component analysis solution from column 3 of Table 2, with respect to the amino acid variables. Columns 1—4 describe the rank reduction by the three first factors in terms of the initial (100%) and explained percent sum-of-squares, and columns 5—8 describe this in terms of initial and residual coefficient of variation (eqn. 14). Columns 9—12 give the actual factor solution in terms of the amino acid mean vector $\bar{x}$ and the weighted direction vectors of these axes (columns of L, eqn. 11). Deweighted directions vectors (columns of B, eqn. 7) may be found by solving the equation $l_{ij} = b_{ij}\, d_i$ for b_{ij}.

One-factor solution

Non-negativity method. Of the total variance (sum-of-squares) for the results shown in Table 3, 93% was accounted for by the first axis (column 10) alone. Figure 1 shows the projection of this axis b_1 into the LYS—GLU plane, together with the individual measurements. This pair of amino acids represents the two dimensions which first give negative loading elements along the axis, since they have the highest positive and negative values in the first loading vector. With the weighting elements $d_i = 1/\bar{x}_i$, the calculation of these "outer limits" is very simple; the limits are defined by $s_{min} = -1/l_{min}$ and $s_{max} = -1/l_{max}$. For this first axis $s_{min} = -1/l_{GLU} = -1/-1.03 = 0.97$, and $s_{max} = -1/l_{LYS} = -1/2.13 = -0.47$. The amino acid compositions of these two "outer limits" are given in Table 4 (columns 1 and 6). Similarly, the scores of the two most extreme experimental samples ($s = 0.23$ and -0.16, respectively) define the "inner limits"; these spectra are also given in Table 4 (columns 3 and 4). Thus an amino acid spectrum range is defined for each of the two main protein groups.

Constant-factor method. The same experimental data were converted to dry matter basis ("unit II", eqn. 15) and submitted to principal component analysis (eqn. 12) whereby a physically meaningful "storage" protein group was estimated by the constant-factor method. In order to check the uncertainty of the results, the 23 samples were split at random into two independent sub-sets prior to the analyses. Point S in Fig. 1 and column 5 of Table 4 gives the means of the replicates with their standard deviation. As can be seen, the

TABLE 3

Determination of number of factors in finger millet amino acid data
(All 16 amino acids and 23 grain samples included; 4 S-deficient samples excluded. Amino acid data in percentage of 16 amino acids ("unit I"). Measures of goodness-of-fit for factors 1, 2 and 3 are: for columns 1—4, a "relative " measure, i.e. percent variance (sum-of-squares) explained by each factor; and for columns 5—8, an "absolute" measure, i.e. percent residual CV remaining after the factors (eqn. 14). Factor solutions with respect to amino acid means and loadings are described in columns 9—12.)

Amino acid	1	2	3	4	5	6	7	8	9	10	11	12
	Relative to initial variation				Relative to initial mean				Factor loading solution			
	Total SS	% of total SS explained by factors			Total *CV*	Residual *CV* (%) after factors			Mean	Loading vectors of factors		
		1	2	3		1	2	3		1	2	3
LYS	2.00	99.0	0.2	0.6	30.1	3.2	2.8	1.5	2.89	2.13	−0.52	−1.65
HIS	0.07	58.1	4.4	16.9	5.7	3.8	3.7	2.8	2.56	0.31	−0.43	1.64
ARG	1.26	94.7	3.5	1.2	23.9	5.6	3.4	1.9	4.67	1.65	−1.59	−1.85
ASP	0.44	92.9	4.1	0.0	14.1	3.9	2.6	2.6	7.22	0.97	−1.02	0.00
THR	0.15	79.4	6.9	7.6	8.2	3.8	3.2	2.2	4.58	0.52	0.76	−1.56
SER	0.02	1.4	1.7	6.9	3.3	3.3	3.4	3.2	6.01	−0.03	0.15	−0.60
GLU	0.47	98.2	0.6	0.1	14.6	2.0	1.6	1.5	23.77	−1.03	−0.42	0.39
GLY	1.23	99.4	0.0	0.2	23.7	1.8	1.9	1.5	4.25	1.68	−0.10	0.74
ALA	0.05	84.8	2.2	1.5	4.9	2.0	1.9	1.8	6.47	0.32	0.26	−0.43
CYS	0.48	91.4	6.8	0.0	14.8	4.4	2.0	2.1	2.13	1.01	1.38	0.19
VAL	0.06	96.4	0.0	0.3	5.3	1.0	1.1	1.0	6.95	−0.37	0.01	−0.20
MET	0.73	79.0	20.2	0.0	18.3	5.0	1.7	1.7	3.36	1.15	2.92	−0.20
ILE	0.11	89.1	1.0	0.0	6.9	2.3	2.3	2.3	4.82	−0.46	0.25	−0.03
LEU	0.18	98.2	0.5	0.0	9.1	2.1	1.1	1.1	10.77	−0.64	0.23	−0.14
TYR	0.19	80.1	2.3	7.4	9.4	4.3	4.1	3.2	3.84	−0.59	0.50	1.77
PHE	0.02	27.9	29.6	5.0	3.0	2.6	2.1	2.0	5.73	−0.11	0.58	−0.47
Mean	0.46	93.5[a]	3.7[a]	1.0[a]	12.2	3.1	2.4	2.0	6.25	1.0[b]	1.0[b]	1.0[b]

[a]Total (weighted); eigenvalues as percent of total sum of squares (7.46).
[b]Root mean squares.

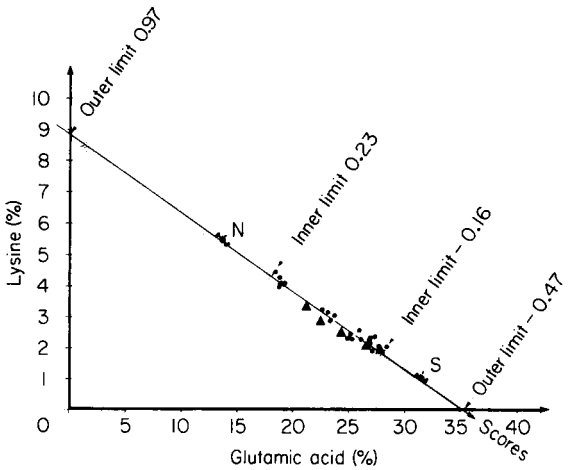


Fig. 1. Factor analysis of finger millet amino acid data (a) One-factor solution: amino acid data and first loading vector (Table 3, column 10) projected into the LYS—GLU plane; (•) grain grown in pot experiments; (▲) grain grown in field experiment. The non-negative ranges between the inner and outer score limits of the storage protein group (—0.16, —0.47) and the non-storage protein group (0.23, 0.97) are shaded. (b) Constant factor solutions: (S) Storage protein group found by factor analysis of data in "unit II"; (N) non-storage protein group found by factor analysis of data in "unit III".

numerically-obtained "storage" protein falls close to the expected line segment, and is highly reproducible.

The "Non-storage" protein group is discussed first in terms of non-negativity. It is known that its score lies along the first factor line, inside the other range, implying an amino acid spectrum range. Figure 2 shows the corresponding range of estimated amounts (eqn. 5), on a dry matter basis, as a function of total crude protein content of the grain, with S as the fixed storage protein spectrum. Figure 2 shows that the non-storage protein group may occur in constant amounts but at different levels, between 2 and 4%, depending on the position score inside the available line segment. By assuming 3% of non-storage proteins in the dry matter, the amino acid spectrum of this protein group can be estimated concisely (N, Fig. 1, Table 4, column 2) by the constant-factor method on "unit III" data (eqn. 16). Mean and standard deviation of N were estimated by analysis of two independent halves of the sample set. Like S, N is highly reproducible.

Two-factor solution

Non-negativity method. Table 3 shows that the first factor alone does not give a completely satisfactory description of the 23 samples. The second factor (column 11, Table 3) is necessary in order to account for the systematic variation in, e.g., ARG, ASP, CYS and MET. After these two factors, the residual coefficient of variation of all the amino acids is close to the expected

TABLE 4

Amino acid spectra obtained from the one-factor solution in "unit I"
(Inner and outer limits along the first factor axis are given for the non-storage (columns 1, 3) and the storage (columns 4, 6) protein group. The spectra obtained by the constant factor method for the non-storage (N, column 2) group and the storage (S, column 4) group are also given, as means of the spectra from two independent sub-sets of samples. Standard errors of the means are included.)

Protein group	Non-storage protein group, N			Storage protein group, S		
	1	2	3	4	5	6
	Outer limit	Obtained by constant-factor method	Inner limit	Inner limit	Obtained by constant-factor method	Outer limit
Score	0.97		0.23	−0.16		−0.47
LYS	8.88	5.52 ± 0.18	4.30	1.90	1.06 ± 0.04	0.00
HIS	3.33	3.01 ± 0.02	2.70	2.42	2.36 ± 0.04	2.18
ARG	12.18	7.76 ± 0.07	6.44	3.43	2.60 ± 0.02	1.04
ASP	19.99	10.19 ± 0.21	8.81	6.09	5.42 ± 0.12	3.94
THR	6.87	5.48 ± 0.09	5.12	4.19	3.66 ± 0.12	3.47
SER	5.84	5.96 ± 0.13	5.97	6.03	6.01 ± 0.06	6.08
GLU	0.00	13.91 ± 0.42	18.17	27.72	31.69 ± 0.02	35.26
GLY	11.19	7.19 ± 0.19	5.88	3.09	2.09 ± 0.05	0.90
ALA	8.48	7.40 ± 0.10	6.95	6.14	5.76 ± 0.10	5.50
CYS	4.21	2.92 ± 0.06	2.61	1.78	1.34 ± 0.03	1.13
VAL	4.45	5.85 ± 0.00	6.36	7.36	7.75 ± 0.04	8.15
MET	7.12	4.75 ± 0.01	4.24	2.73	1.78 ± 0.16	1.54
ILE	2.64	3.91 ± 0.02	4.31	5.17	5.48 ± 0.09	5.86
LEU	4.07	7.88 ± 0.03	9.20	11.88	12.77 ± 0.21	14.01
TYR	1.61	2.85 ± 0.00	3.32	4.21	4.44 ± 0.05	4.90
PHE	5.11	5.43 ± 0.25	5.59	5.33	5.77 ± 0.01	6.04

coefficient of variation in these particular data (2%). The third factor (column 12) gives a small further refinement for a few amino acids (e.g. LYS). This factor distinguished mainly between field and pot experiments. In the following paragraphs, it is ignored because it is small, and the first two factors are regarded as a satisfactory representation of all the information in the 23 samples, accounting for 97.2% of the variances. LYS and GLU were previously found to give the "outer limit" scores along the first factor axis. Likewise, ARG and MET gave outer limits along the second axis, since they have the most extreme loading elements in column 11 of Table 3.

Figure 3 shows the complete "outer limit" hull of the two-dimensional solution. As can be seen, the distance from the "cloud" of experimental data to this hull is rather large, and the information obtained by the non-negativity constraints is limited for these data. The 23 experimental samples are distributed along an open V-shaped track. This is better illustrated in Fig. 4, which represents the corresponding two-factor scores for only 15 amino acids; ASP was excluded entirely from the scaling, weighting and analysis. All 27 samples

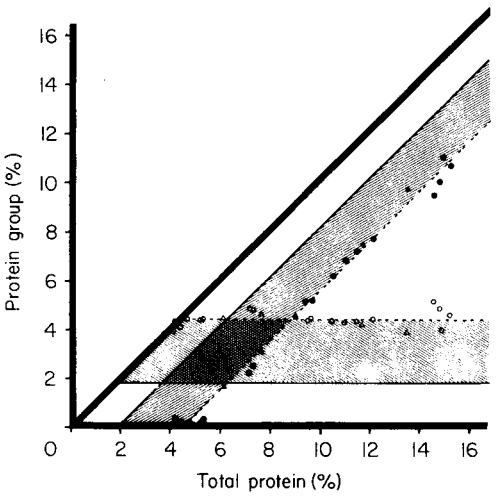


Fig. 2. Non-negativity ranges of protein amounts. Amounts of the storage protein group and the non-storage protein group in dry seed (estimated by weighted regression, see *Regressions from the Mixtures Model*) vs. total crude protein content in the dry seed (Kjeldahl-N × 6.25). Range of permissible amounts of the two protein groups from the one-factor model, limited by non-negativity restrictions (Fig. 1) on the non-storage group (○, △). Table 4, column 5 was used for the storage protein (●, ▲). (○, ●) Grain grown in pot experiments; (△, ▲) grain grown in field experiment. Thick diagonal corresponds to total amount of protein (N × 6.25). The thinner lines correspond to the amounts of the protein groups and were drawn by linear regression through the 23 sample points: (—) obtained with non-storage protein at outer limit (score: 0.97, 0% GLU in protein): (---) obtained with non-storage protein at inner limit (score: 0.23; 0% storage protein in the most extreme grain sample).

were used, including the sulphur-deficient samples, which contain free ASP. Its goodness-of-fit was shown in Table 2, column 2.

The first three loading vectors of this 15 amino acids/27 samples analysis were very similar to those shown in Table 3. A detailed study of the residuals, F (eqn. 12), showed that within the analytical error limits in the data, the three A samples with N:S fertilizer ratios of 20:1 (Table 1) belonged to the same two-factor plane (Fig. 4) as the 23 "normal" samples, although they lie off the V-shaped track which the "normal" grain follows with increased nitrogen fertilizer. A fourth, severely sulphur-deficient sample (N:S fertilizer ratio of 100:1) had atypical residuals which showed that it fell outside this two-factor plane; it was therefore ignored.

Constant-factor method. Two different storage protein groups were obtained by analyzing 11 low-N and 12 high-N fertilizer samples separately; each group was again split into two independent sub-sets. Their mean amino acid spectra were somewhat different [9]; the low-N fertilizer "storage" group (S1, Fig. 4) had more of the sulphur-containing amino acids (CYS and MET) and less of the nitrogen-rich acids (LYS, HIS and ARG) than the high N-fertilizer group

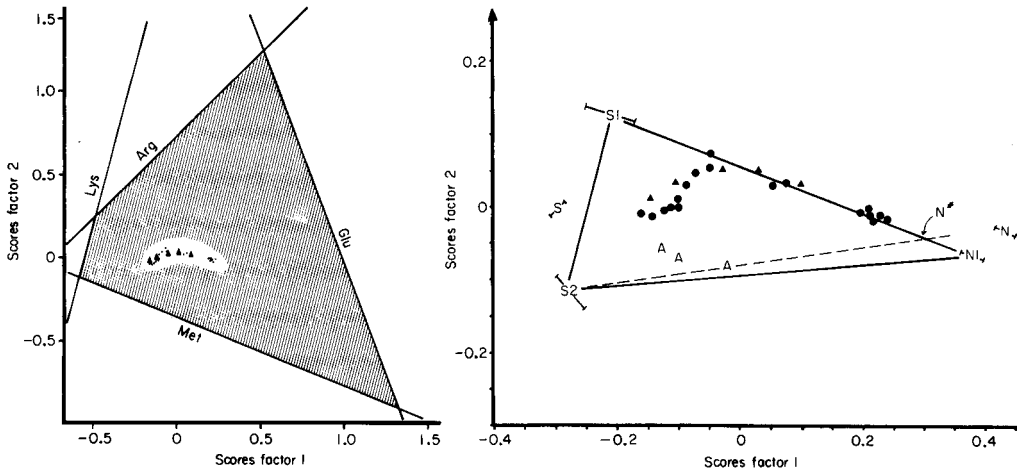


Fig. 3. Non-negativity range in two-factor plane. Outer score limits of the range of permissible amino acid compositions in the two-factor model (16 amino acids, 23 samples). The lines mark the limits outside which the respective amino acid becomes negative, which is impossible. Thus, the unknown protein groups must lie within the shaded area. Circles and triangles show the positions of the pot and field samples, which represent the corresponding inner limit range; choosing a protein group inside this range implies negative amounts of protein group, which is also physically impossible.

Fig. 4. Two-factor scores plane. Scores of the two factors found for 15 amino acids (27 samples. (•) Pot samples; (A) pot samples from S-deficient plants; (▲) field experiment samples. S1 and S2 are storage protein groups found by the constant factor method ("unit II data" on the 11 low-protein and 12 high-protein samples, respectively). N* is the non-storage protein group found by extrapolation of two sides of the simplex. N1 is the non-storage protein group found by the constant factor method ("unit III data") on the 11 low-protein samples.

(S2). The position scores of S1 and S2 in Fig. 4 were calculated by solving eqn. (7) for S in analogy to eqn. (5) for the calculation of P. An estimate of the non-storage protein group (N1) is found by factor analysis of "unit III" data for the same 11 samples used to find S1, assuming 3% of this group in all 11 samples.

Simplex intersect method. Since the three sulphur-deficient A samples lie in the same plane as the 23 "normal" samples (Fig. 4), they all contain the same three protein groups, according to the simplex theory. However, it is obvious that sulphur deficiency leads to a higher S2/S1 storage protein ratio at a given N-fertilizer level. If one of the A samples had contained only S2 and non-storage proteins, a line through S2 and this sample would have to pass through the non-storage group. This sample can therefore be estimated at the point where this line intersects the regression line from S1 through the low-nitrogen samples. N* represents the non-storage group thus obtained. It implies lower LYS and higher GLU contents than were found for N1, although the two estimates are fairly similar. The difference probably arises because the

A sample used contain some of storage group S1. The positions of the two groups N and S found previously are also given, in order to illustrate the effect of including the second factor in the factor analysis model.

Amounts of the protein groups. The amino acid spectra of S1, S2 and N1 were used to calculate their amounts in the dry matter of the different samples (eqn. 5). At low protein contents, S1 rose linearly from 1 to 6% while S2 remained absent, with increasing total crude protein content. With total crude protein contents above 10%, S1 remained more or less constant, while S2 rose linearly. In the sulphur-deficient samples, S1 was more or less replaced by the poor S2, as expected. The non-storage protein group, N1, remained virtually constant at 3% in all samples, including the sulphur-deficient ones.

These results agree well with what was found by chemical fractionation of the proteins in the same finger millet samples [39]. Corresponding factor analysis studies of six barley varieties [10] also gave similar results.

CONCLUSIONS

Physically meaningful "pure components" (protein groups) can be estimated in terms of number, n_c, spectra, A, and amounts, P, based on factor analysis of spectra of their mixtures (whole cereal grain). The problems of negative or ambiguous spectra and amounts were overcome by application of the methods of constant-factor factor analysis and simplex intersect.

These methods may be applicable to many different types of spectral data. For mixtures where such a numerical "fractionation" is possible, expensive and time-consuming chemical fractionation may be reduced. In systems of complex mixtures, new insight may also be obtained concerning the relationships between the many compounds present.

Dr. A. Stabursvik, Norwegian Agricultural College, is warmly thanked for experimental data of unusual precision and for stimulating discussions. Prof. E. Spjøtvoll, Norwegian Agricultural College, and Mr. R. Volden, Norwegian Computing Center, are thanked for their contributions to the statistical theory and computational practice.

REFERENCES

1 H. Martens, NINF Report 13, Norwegian Food Res. Inst., 1432 Aas-NLH, Norway (1978).
2 R. Volden and H. Martens, NINF Report 6, Norwegian Food Res. Inst., 1432 Aas-NLH, Norway (1978).
3 B. Sheldrick, Biochem. J., 123 (1971) 996.
4 B. Lindqvist, J. Østgren and I. Lindberg, Z. Lebensm. Unders.-Forsch., 159 (1975) 15.
5 R. J. M. Gold, H. S. Tenenhouse and L. S. Adler, Biochem. J., 159 (1976) 157.
6 S. R. Searle, Linear Models, Wiley, New York, 1974, p. 87.
7 J. B. Kruskal, in W. H. Kruskal and J. M. Tanúr (Eds.), International Encyclopedia of Statistics, Free Press, New York, 1978, p. 307.
8 E. Weber, Einführung in die Faktorenanalyse, Veb. G. Fisher, Jena, 1974.

9 A. Stabursvik and O. M. Heide, Plant Soil, 41 (1974) 549.
10 H. Martens and K. E. Bach Knudsen, 5th Int. Cong. Food Sci. and Technol., Kyoto, September 1978, Cereal Chem., in press.
11 H. Martens, NINF Report 26, Norwegian Food Res. Inst., 1432 Aas-NLH, Norway (1978).
12 See, e.g. P. S. Schoenfeld and J. R. DeVoe, Anal. Chem., 48 (1976) 403R.
13 W. H. Lawton and E. A. Sylvestre, Technometrics, 13 (1971) 617.
14 W. H. Lawton, E. A. Sylvestre and M. S. Maggio, Technometrics, 14 (1972) 513.
15 E. A. Sylvestre, W. H. Lawton and M. S. Maggio, Technometrics, 16 (1974) 353.
16 G. L. Ritter, S. R. Lowry, T. L. Isenhour and C. L. Wilkings, Anal. Chem., 48 (1976) 591.
17 T. Hirschfeld, Anal. Chem., 48 (1976) 721.
18 E. H. Lane, S. D. Christian and F. Garland, J. Phys. Chem., 80 (1976) 690.
19 D. J. Leggett, Anal. Chem., 49 (1977) 276.
20 M. Takatsuki and K. Yamaoka, J. Sci. Hiroshima Univ., Ser. A, 40 (1976) 387.
21 See, e.g., E. R. Malinowski and M. McCue, Anal. Chem., 49 (1977) 284.
22 I. M. Warner, G. D. Christian, E. R. Davidson and J. B. Callis, Anal. Chem., 49 (1977) 564.
23 R. N. Cochran and F. H. Horne, Anal. Chem., 49 (1977) 846.
24 E. R. Malinowski, Anal. Chem., 49 (1977) 606.
25 S. M. McCown, H. H. Land and C. M. Earnest, Anal. Chem., 50 (1978) 1362.
26 G. T. Rasmussen and T. L. Isenhour, Anal. Chim. Acta, 103 (1978) 213.
27 K. G. Bergstrand, FAO/IAEA Panel on New Approaches in Breeding for Plant Protein Improvement, Röstonga, Sweden, 1968, p. 117.
28 M. Denic, J. Dumanovic, K. G. Bergstrand and L. Ehrenberg, Genetika, 1 (1969) 25.
29 H. Martens, Y. Solberg, L. Roer and E. Vold, Potato Res., 18 (1975) 515.
30 H. M. Shapiro, Biochim. Biophys. Acta, 236 (1970) 725.
31 N. Ohta, Anal. Chem., 45 (1973) 553.
32 E. Spjotvoll, H. Martens and R. Volden (1979), submitted for publication.
33 G. Michael, B. Blume und M. Faust, Z. Pflanzenernär. Düng. Bodm., 92 (1961) 106.
34 G. Michael, Qual. Plant. Mater. Veg., 10 (1963) 248.
35 S. Wold, Pattern Recognition, Pergamon, Oxford, 1976, p. 127.
36 D. L. Duewer, B. R. Kowalski and J. L. Fasching, Anal. Chem., 48 (1976) 2002.
37 J. C. Gower, The Statistician, 17 (1966) 13.
38 H. Martens and K. E. Bach Knudsen (1979), submitted for publication.
39 E. Poulsson, Sci. Rep. Agric. Univ. Norway, 54 (1975) 5, 1.
40 BMDP4M Factor Analysis, Double Precision version, Health Sci. Comput. Facility, Univ. of Calif., L. A., USA, 1975.

Analytica Chimica Acta, 112 (1979) 443—448
Computer Techniques and Optimization

Short Communication

A COMPUTER PROGRAM FOR THE AUTOMATIC IDENTIFICATION OF EDIBLE OILS BY GAS—LIQUID CHROMATOGRAPHY

JURGEN L. KACPRZAK* and VICKI R. HIGGINS

Division of Analytical Laboratories, P.O. Box 162, Lidcombe, New South Wales, 2141 (Australia)

(Received 26th March 1979)

SUMMARY

A computer program is given for the identification of edible oils from their fatty acid ratios. The program was used with the Hewlett-Packard 3354 Laboratory Automation System in autocall mode to identify a number of vegetable oils from the fatty acid ratios determined by gas chromatography of the trans-esterified oil samples.

Fats and oils may be identified by the determination of fatty acid ratios obtained by trans-esterificaton of the oil and subsequent gas—liquid chromatography (g.l.c.) of the fatty acid methyl esters. These results are then compared with specified limits of composition for particular oils and fats. The program described is written in LAB BASIC II for use with the Hewlett-Packard 3354 Laboratory Automation System and identifies an oil by reading the processed data file generated at the end of an analytical run by a normalized system method (i.e., one which uses relative response factors for quantifying the fatty acid ratios). The method is calibrated with a synthetic mixture of fatty acid methyl esters of known concentration and is based on methyl oleate as the reference compound. The samples are identifed automatically as the method has been instructed to call the program at the end of the analytical run (autocall mode). The result of the identification appears as an added printout at the end of the system report.

Experimental

Equipment. A Hewlett-Packard Model 5730A gas chromatograph with Model 7670A auto-injector was used. The gas chromatograph had a hydrogen flame ionization detector and was fitted with a silanized glass column (3m long, 4mm i.d.) packed with 7% CS-10 (Alltech Associates, 202 Campus Drive, Arlington Heights, Ill.) on 100—120 mesh Chromosorb W-HP (Alltech Associates). Data processing was done by a Hewlett-Packard 3354 Laboratory Automation System.

Operating conditions were: column temperature, 190°C; injector and detector temperatures, 250°C; nitrogen flow rate, 20 ml min^{-1}.

Trans-esterification. The methyl esters were formed by reacting 10 drops (ca. 100 mg) of oil with 5 ml of 0.025 M sodium methoxide solution (prepared by dissolving sodium metal in absolute methanol) in a sealed glass ampoule at about 100°C for 2 h. The reaction was complete when no oil phase remained. An aliquot of the methanolic solution of the methyl esters was transferred to the automatic sampling bottles for g.l.c.

Computer program

The program compares 13 fatty acid ratios of a sample (obtained by analysis) with a fixed set of data representing specified ranges of the same fatty acids used to define or characterize a particular oil. Two items of data (a minimum and maximum) are required to give a range for each fatty acid ratio and therefore 26 data points are required to define each oil. The program currently contains data for 12 common edible oils. It could be extended to include more oils and/or fatty acid ratios to meet individual needs. The data used to define the various oils in the program (Table 1) are based on values given by Spencer et al. [1], except for the data for rapeseed and olive oils which are based on values from other sources [2, 3]. Data may be updated or varied according to currently accepted values or as required. Allowance is made for error in the g.l.c. analysis. A tolerance of ±2% has been arbitrarily selected for each fatty acid ratio in the sample. As not all 13 fatty acids will always be found in a sample, the program initializes fatty acid ratio values to zero and only replaces them when they are found in the processed data file.

The program prints the name of every oil whose data ranges conform to the fatty acid ratio values of the sample. The result is that more than one name can be assigned to a sample; an example is the typical system report given in Fig. 1. The flow diagram (Fig. 2) illustrates the working of this part of the program. The program also calculates the theoretical iodine value from the unsaturated fatty acid ratios (not illustrated). The program can easily be modified to permit fatty acid ratios to be entered manually by using a keyboard terminal (terminal mode) instead of the autocall mode.

Results and discussion

Thirty oil samples were analyzed by the method and 28 were correctly identified as labelled. The 30 samples comprised 8 sunflower oils, 8 safflower oils, 4 soya bean oils, 4 maize oils, 3 olive oils, 2 sesame seed oils and 1 cottonseed oil. Six of the sunflower oils satisfied the data for safflower oil as well as sunflower oil and so were reported as both; one sunflower oil was identified as soya bean oil only, and this sample gave a positive Richard's test [3]. The two sesame seed oils were correctly identified, but the results were also within the limits for sunflower oil; in addition, the results for one sample were also within the limits for maize oil. A qualitative test for sesame seed oil, the Pavolini test [3], was positive for these two samples. The sample of cottonseed oil was not identified. The myristic acid content was much

TABLE 1

Lower and upper limits of the fatty acid ratios used for the identification of oils

Fatty acid	Oil											
	Sunflower	Safflower	Maize	Arachis	Olive	Cottonseed	Rapeseed	Sesame seed	Mustard seed	Soya bean	Lard and pork fat	Premier Jus and tallow
Lauric (C-12)	0—0.1	0—0.1	0—0.1	0—0.1	0—0.1	0—0.1	0—0.5	0—0.1	0—0.5	0—0.1	0—0.5	0—0.1
Myristic (C-14)	0—0.5	0—1.0	0—1.0	0—0.1	0—2.4	0.5—2.0	0—1.0	0—0.5	0—1.0	0—0.5	0.5—2.5	1.4—6.3
Palmitic (C-16)	3.0—10	2.0—10	8.0—19	6.0—15.5	7.1—21.1	17—29	0.5—5.0	7.0—12	0.5—4.5	7.0—12	20—32	20—37
Palmitoleic (C-16:1)	0—1.0	0—0.5	0—0.5	0—1.0	0.2—5.5	0.5—1.5	0—1.0	0—0.5	0—0.5	0—0.5	1.7—5.0	0.7—8.8
Stearic (C-18)	1.0—10	1.0—10	0.5—4.0	1.3—6.5	0.3—3.8	1.0—4.0	0.5—3.0	3.5—6.0	0.5—2.0	2.0—5.5	5.0—24	6.0—40
Oleic (C-18:1)	14—65	7.0—42	19—50	36—72	54—93.5	13—44	9.0—40	35—50	8.0—23	19—30	35—62	26—50
Linoleic (C-18:2)	20—75	55—81	34—62	13—45	1.0—23.6	33—58	11—29	35—50	10—24	48—58	3.0—16	0.5—5.0
Linolenic (C-18:3)	0—0.7	0—1.0	0—2.0	0—1.0	0—1.8	0.1—2.1	5.0—12	0—1.0	6.0—18	4.0—10	0—1.5	0—2.5
Arachidic (C-20)	0—1.0	0—0.5	0—1.0	1.0—2.5	0—1.3	0—0.5	0—1.5	0—1.0	0—1.5	0—1.0	0—1.0	0—0.5
Eicosenoic (C-20:1)	0—0.5	0—0.5	0—0.5	0.5—2.1	0—4.9	0—0.5	5.0—15	0—0.5	5.0—13	0—1.0	0—1.0	0—0.5
Behenic (C-22)	0—1.0	0—0.5	0—0.5	1.5—4.8	0—0.8	0—0.5	0—1.5	0—0.5	0.2—2.5	0—0.5	0—0.1	0—0.1
Erucic (C-22:1)	0—0.5	0—0.1	0—0.1	0—0.1	0—0.1	0—0.5	30—60	0—0.1	22—50	0—0.1	0—0.1	0—0.1
Lignoceric (C-24)	0—0.5	0—0.1	0—0.5	1.0—2.5	0—0.1	0—0.5	0—2.0	0—0.1	0—0.5	0—0.1	0—0.1	0—0.1

```
REPORT:     8.30  CHANNEL: 11                    FATTY ACID RATIOS

SAMPLE: F737          INJECTED AT  10:59:13 ON FEB  2, 1979

NORM METHOD: FAMEN2   BTL:  2


ACTUAL RUN TIME:  35.950 MINUTES

RUN ABORTED
BL > 10 MV
ENDED NOT ON BL


   RT        AREA        WT %          NAME

  5.66     1051665 BV    7.124    C16 PALMITIC
  8.56      657304 VV    4.326    C18 STEARIC
 10.07     2526444 VV   17.458   *C18 OLEIC
 12.79     9996988 VV   70.763    C18 LINOLEIC
 15.56        6187 VV     .040    C20 EICOSENOIC
 16.34        5311 VV     .040    C18 LINOLENIC
 21.04       32790 VV     .202    C22 BEHENIC
 33.89        8709 VF     .048    C24 LIGNOCERIC

TOTAL AREA =      14285402        TOTAL WT %   =     100.000

PROCESSED DATA FILE: PR2          RAW DATA FILE: *RAW11


SUNFLOWER OIL
SAFFLOWER OIL
IODINE VALUE = 138.6 WIJS UNITS
```

Fig. 1. Typical system report.

higher than the normal myristic acid content of cottonseed oil. Experience shows that several conditions must be satisfied in order to identify an oil correctly. Identification depends on the range of fatty acid concentrations that has been chosen to define the oil; consequently, if a sample is only fractionally outside the chosen range (plus the allowed error of 2%) for a single fatty acid in the program, the sample will not be identified. The range of values used to define an oil requires some consideration; this particularly applies to the higher fatty acids which are usually absent or present in only trace quantities. A value of 0.1% was allowed as the maximum content for acids which are not usually present. It was found to be advantageous to keep the quantity of oil used for methylation within defined limits so that large changes in concentrations would not affect retention times. If only five or six types of oils are of interest, it may be advantageous to have two or three different sets of ranges for the same oil included in the program to allow various "degrees of identification", thus permitting an oil to be identified with a greater or lower level of certainty. The method of analysis was found to be reproducible and coefficients of variation for fatty acid ratios in concentrations as low as 0.5% were around 2%. For fatty

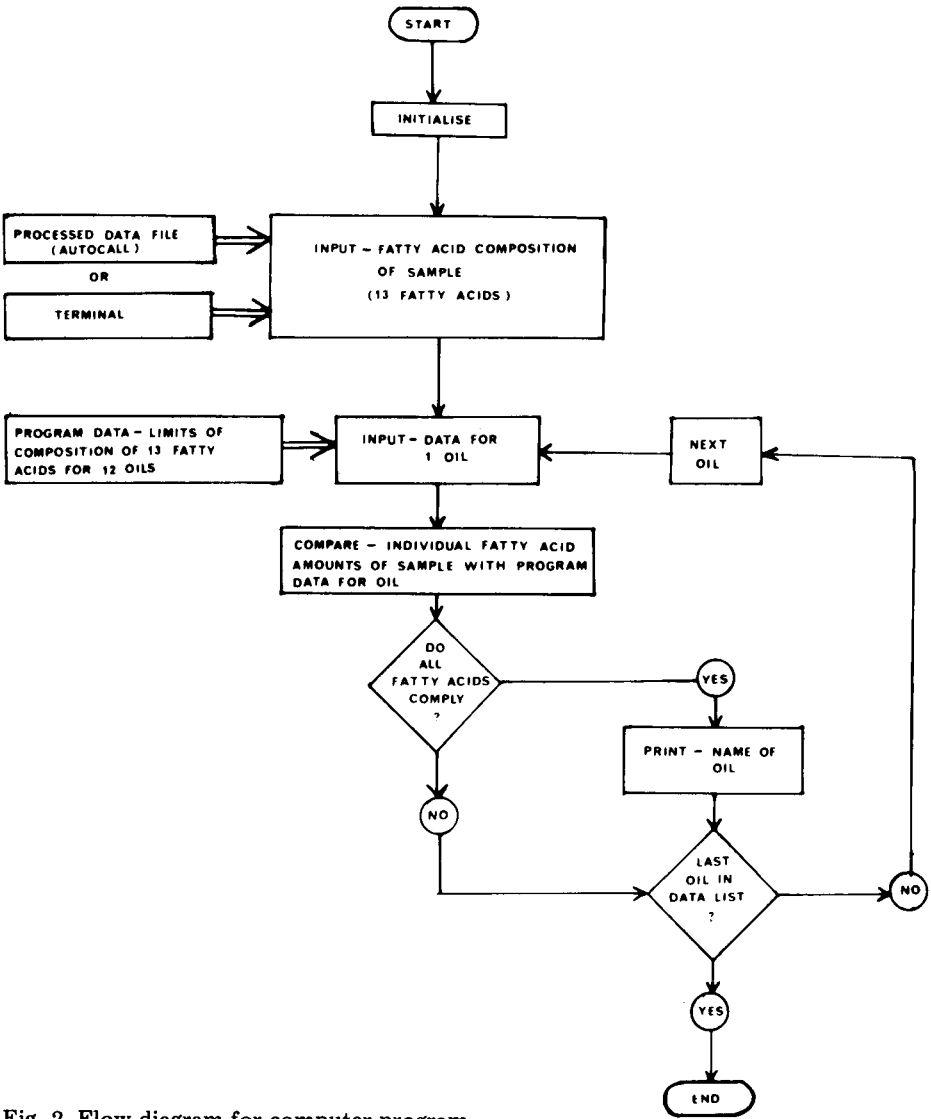


Fig. 2. Flow diagram for computer program.

acid ratios exceeding 6%, the coefficients of variation were less than 0.5%. The method is rapid and 36 samples can be completed in one day with overnight running of the chromatograph. A listing of the program in LAB BASIC II is available upon request from the authors.

Acknowledgement is made to the New South Wales Government Analyst and The Health Commission of New South Wales for permission to publish this paper.

REFERENCES

1 G. F. Spencer, S. F. Herb and P. J. Gormisky, J. Am. Oil Chem. Soc., 53 (1976) 94.
2 Codex Alimentarius Commision Codex Committee on Fats and Oils, Revised Identity Characteristics for Fats and Oils based on Gas Liquid Chromatography, CX/FO 70/1, 1970.
3 H. A. Boekenoogen, Analysis and Characterization of Oils, Fats and Fat Products, Interscience, New York, 1968.

Analytica Chimica Acta, 112 (1979) 449—453
Computer Techniques and Optimization

Short Communication

NON-LINEAR DETECTOR RESPONSE IN GAS AND LIQUID CHROMATOGRAPHY — QUANTIFICATION WITH A LABORATORY AUTOMATION SYSTEM

GUIDO JANSSENS

Instituut voor Hygiëne en Epidemiologie, Departement Farmacotoxicologie, J. Wytsmanstraat, 14, B-1050 Brussel (Belgium)

(Received 18th May 1979)

SUMMARY

A Laboratory Automation System (Hewlett-Packard 3354 B) minimizes systematic errors in quantification problems when detector responses are not linear. The calibration curves for the various products are stored and results are calculated by linear segmental interpolation. The complete operating cycle is controlled by the System.

Successful application of external standards and internal standards in quantitative gas and liquid chromatography implies a linear relation between the amount of sample injected and the corresponding peak area. This linear range is often limited: deviations may occur for both small and large quantities. Typical linear ranges for the most common detectors in gas chromatography have been reported by Driscoll [1]. In practice, however, the range may be smaller for some compounds than is often accepted in the literature. In order to minimize systematic errors resulting from non-linear detector responses, various resources are available.

With a non-linear detector, the response factor is a function of the amount of analyte in the calibration solution. Roughly equating the concentrations for the calibration and in the sample gets around this problem. However, the sample concentration is often unpredictable, or a series of samples may have strongly varying concentrations, thus this stratagem cannot be considered as a general solution.

If the non-linearity is due not to the detector itself, but to incomplete recovery of the analyte as a consequence of preliminary purifications and separations, good results may be obtained by adding an appropriate internal standard, with chemical properties analogous to the analyte, prior to any treatment. The presence of various components having different chemical structures therefore requires various internal standards, each analyte being determined by comparison with its chemical analogue. This method is most frequently applied in liquid chromatography for clinical chemistry [2]. Yet, the choice of internal standards is not always easy, and much time may be needed to elaborate an analytical method.

If a disk-based memory system is available, an obvious solution consists in storage of concentration—area pairs, and quantification by linear interpolation of the measured area between adjacent reference points. Hewlett-Packard provide notes on this procedure for application on its Data System 3352 C. Since the system is not equipped with an external memory, the capabilities and flexibility of the system are restricted. The unpopularity of this simple method is probably due to the fact that establishment of the calibration curves for all the analytes is time-consuming. A Laboratory Automation System (LAS) that not only treats detector signals, but also controls an automatic liquid sampler may be used to eliminate most human interventions. The application of a Hewlett-Packard 3354 B Laboratory Automation System controlling all calibration and calculation processes is reported here.

System description

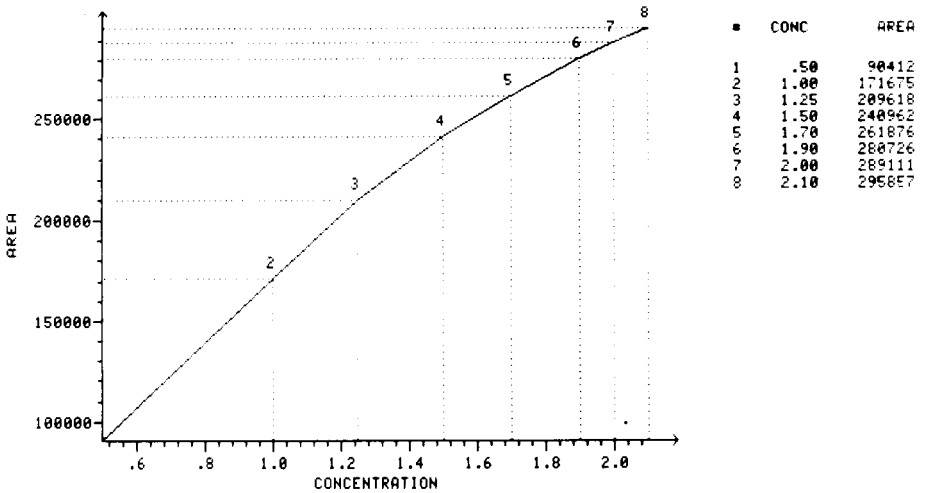
The main components of the 3354 B Laboratory Automation System are a computer (e.g. HP 2108 A) with 64 Kbytes memory and a disc with 15—20 Mbytes storage capacity. The g.c.-detector signals are integrated for time intervals determined by the user. The series of digitized values are stored in raw data files on the disc. The standard HP software provides for peak integration with sophisticated baseline corrections, peak identification with updating of the calibration table, quantification and report routing to various output devices. Complete automation is obtained by using an automatic liquid sampler (HP 7671 A) interfaced with a sampler control module (HP 18653 A).

Computer programs

On the basis of the system capabilities, five programs (SECAL:GJ, SECAL2:GJ, SEPOL:GJ, SEPOL2:GJ and CHROMA:GJ) were developed in the simple Lab Basic II language. The programs allow storage of calibration curves for 11 compounds, each with up to 20 reference points. Reference points are considered as being different if their concentration values do not correspond. (Copies of the programs are available from the author by forwarding a HP-9162-0061 tape cartridge or by requesting the listings).

Manipulations in the reference table. The content of the reference table can be read and if necessary modified by running program SECAL:GJ in terminal mode on a Hewlett-Packard terminal of the 2600-series. Characteristic features of this group of terminals such as "inverse video" and "soft keys" are applied in programs SECAL:GJ, SECAL2:GJ, and CHROMA:GJ, which are thus exclusively intended for applications on the HP display stations.

Figure 1 shows an example of a calibration curve generated by calling the appropriate subprogram via soft key f7. The program provides automatic scaling and labelling of both axes. Any errors occurring during the calibration process can immediately be seen from the image. Moreover, some idea of the linearity of the detector response to the analyte can be obtained almost immediately. If the curvature in a particular range appears to be too pronounced, extra reference points can be added (see below). If the supplied table name cannot be found, a table of the right size (11 recordings) is automatically created.



#	CONC	AREA
1	.50	90412
2	1.00	171675
3	1.25	209618
4	1.50	240962
5	1.70	261876
6	1.90	280726
7	2.00	289111
8	2.10	295857

Fig. 1. Computer-generated calibration curve showing non-linear response for caffeine in the range 0.5—2.1 mg ml^{-1}.

Quantification. Concentrations are calculated by the post-run program SEPOL:GJ. The designation of the reference table used for the calculations should be supplied as the method parameter file. Table 1 gives an example of a quantitative report. Concentrations of substances which are not matched in the table or have only one reference point were calculated by the standard HP software; these values are indicated by a minus sign. The other results were calculated by linear interpolation of the area between adjacent reference points, or by linear extrapolation of the closest reference points if the area is outside the calibration range.

Automatic storage of reference points. Besides the manual input of data into the calibration table by means of program SECAL:GJ, the automatic storage of reference points is a more flexible way of producing a response curve. If the program SEPOL:GJ observes that the sample name of the analytical run starts with character string CALIB, the storage program SEPOL2:GJ is brought into use. The number following CALIB indicates the amount injected. These values, together with the measured peak areas, are automatically added to the calibration table at the end of the run. When the same concentration is already present for a substance, old and new values are averaged. Table 2 shows a post-run report giving information about the addition of calibration data. The program checks on overflow of peak names and reference points. Peaks X, Y and F were not appended because the calibration table already contained the maximum number of 11 peaks; the reference point of peak D was also not included, as peak D contained the maximum number of 20 reference points. Calibration reports and quantitative reports are sent to the output devices specified in the method.

Chromatogram. At the end of the analysis, the chromatogram is automatically

TABLE 1

Example of quantitative report obtained for non-linear detector response by LAB BASIC II for raw data file RSAM5 and processed data file PSAM5
(Results calculated by the standard HP software are indicated by a minus sign)

REPORT# 124 CHAN# 2 ISTD METHOD: MTEST

SEQUENCE: STEST:GJ BOTTLE# 18

SAMPLE: IHE-213/VW INJECTED AT 13:58:44 ON FEB 8, 1979

RT (min)	AREA (μV s)	CONC (ng μl^{-1})	NAME
20.22	70118	19.63	PEAK A
25.85	68047	4.98	PEAK B
26.11	83145	60.28	PEAK C
42.27	56491	118.79	PEAK D
50.07	74186	−1.00	#PEAK Y
56.06	82153	99.13	PEAK G
56.86	54475	19.45	PEAK H
63.84	30501	−9.18	PEAK F

TABLE 2

Example of a calibration extension run
(Raw data file: RTEST; processed data file: PTEST)

REPORT# 112 CHAN# 2 ISTD METHOD: MTEST

SEQUENCE: STEST:GJ BOTTLE# 6

SAMPLE: CALIB 200 INJECTED AT 15:29:31 ON FEB 7, 1979

Reference points added to table REFTAB

NAME	RT	CONC.	AREA
PEAK A	20.22	200.0	70118
PEAK X NOT APPENDED: MAX # PEAKS IN TABLE			
PEAK C	26.11	200.0	83415
PEAK D NOT ADDED — MAX # REF PTS			
&PEAK Y NOT APPENDED : MAX # PEAKS IN TABLE			
PEAK G	56.06	200.0	82153
PEAK H	56.86	200.0	54475
PEAK F NOT APPENDED : MAX # PEAKS IN TABLE			

reconstructed on a HP-2648A graphics terminal according to parameters determined by the user [3]. Before the injection cycle is started, these parameters are stored in the sequence parameter file by the interactive program CHROMA:GJ. An automatic slice-averaging is included in the program, as well as the choice between a fixed scale or a normalization with the highest peak as reference, an amplitude correction factor in case of area slice-averaging, distribution of the chromatogram over various sheets, and automatic scaling and labelling of both axes.

Results and discussion

The calculation procedure is based on external standards, so that particular attention must be paid to the reproducibility of the injected sample size. When injections are done by the automatic liquid sampler, variations in sample size are insignificant. Five consecutive injections of 1.8 μg of caffeine (cf. Fig. 1) gave an average concentration of 1.81 μg with a standard deviation of 0.02 μg. The gain in accuracy is evident from comparison with values obtained by the conventional internal standard method, by which a result of 1.68 μg was obtained when the response factor was related to a concentration of 1.5 μg (reference point 4 in Fig. 1) and a result of 1.49 μg when the calibration was done at 0.5 μg (reference point 1).

In addition to the peak areas, the baseline-corrected peak heights are available in the new software release (revision 1905). These data are to be preferred if peak shoulders are not detected by the system and wrongly contribute to the integrated peak area. Program adjustment consists of modifying and adding only some statements related to the reading of the processed data file.

REFERENCES

1 J. N. Driscoll, Am. Lab., 9 (1976) 71.
2 See, e.g. (a) K. H. Dudley, and (b) K. H. Dudley, D. L. Bius, B. L. Kraus and L. W. Boyles, in C. E. Pippenger, J. K. Penry and H. Kutt (Eds.), Antiepileptic Drugs: Quantitative Analysis and Interpretation, Raven Press, New York, 1978, pp. 19, 35.
3 G. Janssens, J. High Resol. Gas Chrom. Chrom. Comm., 2 (1979) 84.

ANALYTICA CHIMICA ACTA, VOL. 112 (1979)
(Computer Techniques and Optimization, Vol. 3, No. 4)

AUTHOR INDEX

\nnouncing two new volumes in the series:

;tudies in Environmental Science

olume 4

'OTENTIAL INDUSTRIAL CARCINOGENS AND ,IUTAGENS

AWRENCE FISHBEIN, *National Center for Toxicological Research, Jefferson, AR, U.S.A.*

his work provides detailed information on reported industrial carcinogens and mutagens nd arranges them by structural categories in order to highlight their potential risks nd to help predict the hazards of new agents considered for introduction into the nvironment. It includes information on such topics as: the synthesis of these agents, ature of their trace impurities, environmental occurrence, chemical and biological :activity, TLV's and MAC's, test systems, combination effects in chemical carcinogenesis, pidemiology, and risk-assessment.

his volume will therefore be of great interest to scientists involved in toxicology, arcinogenesis and mutagenesis studies, genetics, and environmental health. In addition, will provide valuable assistance to officials working in public health and environmental rotection agencies.

eb. 1979 x + 534 pages US $66.75/Dfl. 150.00 ISBN 0-444-41777-X

olume 2

\IR POLLUTION REFERENCE MEASUREMENT ,IETHODS AND SYSTEMS

roceedings of the International Workshop, Bilthoven, December 12-16, 1977

rganized by The National Institute of Public Health, Bilthoven, The Netherlands **o-sponsored by The World Health Organization. T. SCHNEIDER,** *The National ıstitute of Public Health, The Netherlands,* **H. W. DE KONING,** *WHO, Geneva, witzerland, and* **L. J. BRASSER,** *TNO, The Netherlands* (Editors).

particularly valuable feature of this work is the presentation of recommendations nd follow-up projects, including international projects that will contain and apply ıe reference principles discussed during the workshop. The book will serve as an p-to-date review of the status of Air Pollution Reference Methods and Systems for :chnicians involved in air pollution and will also provide useful background information ɔr those involved in air pollution activities in general. It is hoped that this work ill stimulate greater international cooperation in the development of good reference /stems.

ec. 1978 vii + 168 pages US $35.50/Dfl. 80.00 ISBN 0-444-41764-8

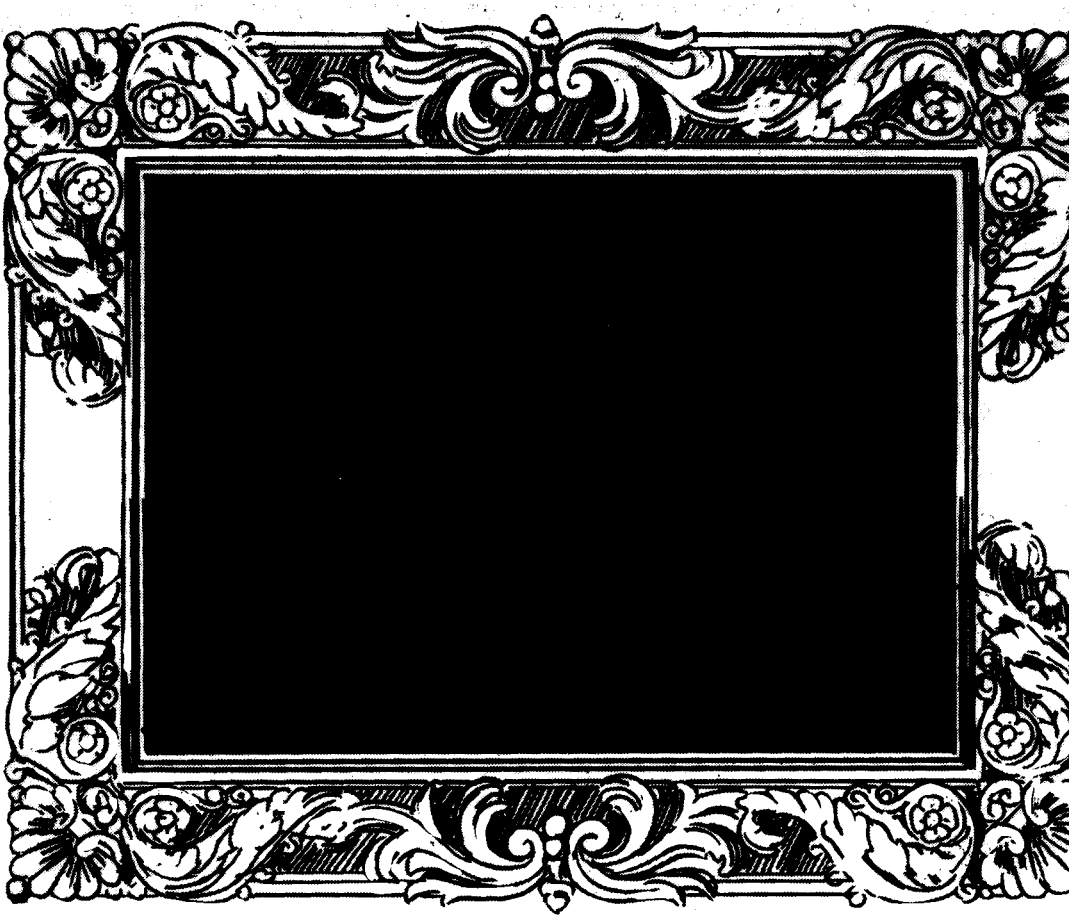
ELSEVIER

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

52 Vanderbilt Ave
New York, N.Y. 10017

7125

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



What Industry would be without computers.

Computers?
You as an engineer don't need them as toys. Playing is for kids and Einsteins. But you need them badly as tools.
You want to forget about them as soon as you can. But not sooner. You and your company thrive on them. Or worry about them.
Read Computers in Industry, from cover to cover, 4 times a year and subscribe to it. You need it more than you know. And so does your Industry.

COUPON FOR A FREE COPY

For a free copy of the first issue of **Computers in Industry,** please write or complete the coupon and return it directly to the publisher:

North-Holland Publishing Company
Attn: Mr. J. Dirkmaat
P.O. Box 211,
1000 AE Amsterdam, The Netherlands

Name: _____

_____

Address: _____

_____

_____

VOL. 112, No. 4 ANALYTICA CHIMICA ACTA December 15, 1979

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

CONTENTS

Printed in The Netherlands.

JOURNAL

INTERNATIONAL JOURNAL OF
MICROPROCESSING AND MICROPROGRAMMING

SCOPE

The main purpose of the Journal is to facilitate the international flow of information in the field of microprocessing and microprogramming.

Industrial Seminar papers from EUROMICRO's 1978 Symposium (Munich, October 17-19, 1978) have been published as a Special Issue of the Journal (Vol. 5, no. 1). A Special Issue to give a comprehensive *Status Report* on microprocessing and microprogramming in Europe has been scheduled for Vol. 5, no. 3. (This is meant as an update and extension of the previous Status Report which has appeared as Vol. 3, no. 4 of the Journal).

The third special issue of 1979 will be devoted to Personal Computing.

The Pan-European coverage of Euromicro's Board of Directors provides the basis for a special Reports Section intended to give an overview of national though important events (conferences, symposia, workshops, etc) and articles or special issues of national journals.

The Journal contains, in addition to reports on recent research and technological progress, comprehensive surveys, state-of-the-art reports, short communications and announcements pertinent to the micro field.

The scope of the Journal includes in particular:
- theory - design - research - languages - simulation
- emulation - teaching aids - evaluation/diagnostic methods,

relating to:
- microprogramming - microprocessing - microprocessor systems and networks - distributed computing - MSI/LSI components - computer structures - modular systems
- integrated hardware/software design - microarchitecture of computer systems.

EUROMICRO JOURNAL is the forum for EUROMICRO, The European Association for Microprocessing and Microprogramming. It commenced publication as "Euromicro Newsletter" in 1974 on a quarterly basis.
In 1978 it was renamed EUROMICRO JOURNAL. Since then, EUROMICRO JOURNAL appears as a bi-monthly.
The Journal is published and distributed by North-Holland both for the EUROMICRO membership and institutional subscribers.

EDITORIAL ORGANIZATION

North-Holland Publishing Company

P.O. Box 211, 1000 AE Amsterdam, The Netherlands *or:* 52 Vanderbilt Avenue, New York, N.Y. 10017

VOLUME 5, NUMBER 1

Industrial Seminar papers from Euromicro's 1978 Symposium

Guest Editor's Introduction, C. Aléonard, France
MICROPROCESSOR COMPONENT FAMILIES
The Total Software Upward/Downward Compatible R 65XX Microcomputer Family, G.G. Wiese, Fed. Rep. Germany
The 8X330 Floppy Disk Controller Chip, E.D. van Veldhuizen, The Netherlands
Operation of the Signetics 2652 Multi-Protocol Communication Controller, P. James, U.K.
DEBUGGING AND TESTING
The 8002 Microprocessing Development Aid System, W. Edel, Fed. Rep. Germany
Microprocessor Systems Testing - a Review and Future Prospects, C. Robach, G. Saucier and C. Aléonard, France
Efficient Manufacturing of Complex Digital Boards through the Use of a Low-Cost Universal Tester, J. Le Gars, France
MICROPROCESSOR APPLICATIONS
Speech recognition and Speech Synthesis, M. Wasmeier, Fed. Rep. Germany
A Reliable Low-Cost Graphic System, R.D. Klein, Fed. Rep. Germany
MULTIMICROPROCESSOR SYSTEMS
SMS 201 - A Powerful Parallel Processor with 128 Microcomputers, C. Kuznia, Fed. Rep. Germany
Solving Linear Equations with the SMS 201 Parallel Processor, K. Nagel, Fed. Rep. Germany

VOLUME 4, NUMBER 6

This issue contains a Special Section on Microprocessors in Biomedical Applications

The Use of Microprocessors in Biomedical Applications - an overview, F. Pinciroli, Italy
Microprocessors and Bioengineering, P. Morasso and V. Tagliasco, Italy
Microprocessors and Instrumentation for Intensive Care Units and Surgical Halls, M. Mirabel, France
Microprocessors and Instrumentation for Chemical Pathology, A. Musetti, Italy
Prospects of the Microprocessor in the Biomedical Instrumentation Field, P. Martinelli and A. Torsoli, Italy

VOLUME 4, NUMBER 1

Industrial Seminar papers from EUROMICRO's 1977 Symposium

Controlling the hidden costs in microcomputer system design, N. Jinadasa, Belgium
Realizing the economic promise of microprocessors, C.O. Simpson, Belgium
Development tools for bit-slice microprocessors, K. Schneider, Switzerland
The technology race in microprocessor application, P.G. Harrison, U.K.
Linking FORTRAN and assembly language programs, H. Tireford, Switzerland
The SEQUEL software language, C. Dye, U.K.
Low-speed data communications and your micro, M. Repko, The Netherlands
Microprocessor-controlled videogames, W.P. van Deursen, The Netherlands, and K. Li, U.S.A.
Analogue to microprocessor in easy stage, N.A. Justice, U.K.

VOLUME 3, NUMBER 4

This issue contains a collection of status reports on microprocessing and microprogramming in Europe and in Japan.

European Section

Microprogramming and Microprocessing in Austria, H. Schauer, G. Chroust, Austria
Activities in Microprocessors Field in LSI Manufacturers and Universities in France, F. Anceau, France
Status Report on Microprogramming and Microprocessing Activities in Germany, L. Richter, Germany
Status Report The Netherlands, J. Wilmink, The Netherlands
Status Report Portugal, A. Steiger Garcao, Portugal
Microprocessor Status Report - U.K., Thompson, U.K.

Japanese Section

Microcomputer Applications, R. Mori, I. Morishita, Y. Kita and Y. Okada, Japan
Microprocessors in Japan, R. Mori, H. Tajima, M. Tajima and Y. Okada, Japan

VOLUME 3, NUMBER 3

Industrial Seminar papers from EUROMICRO's 1976 Symposium

Evaluation and Debug of Microprocessor Systems Using a Reconfigurable Logic State analyzer, D.I. Kolody, U.S.A.
A Fast Microprocessor for Control Applications, Dr. H. Hoffman, Dr. J. Nemec, The Netherlands and U.S.A.
Microcomputer Control of a Batch Chemical Process Plant, J. Gallacher, U.K.
Memory Interface for the 2650 Microprocessor, H. Schutte, The Netherlands
Developments in Development Systems, H. Kornstein, U.K.

SUBSCRIPTION INFORMATION

The journal will be published in 1979 in one volume of six issues. The subscription price for libraries amounts to US $61.00/Dfl. 125.00. Library subscriptions are available from the Publisher or a local agent.
Membership of the Euromicro Association includes a Personal Subscription. Membership fee for 1979: FF 95,—. These membership subscriptions are for individual use only and should not be made available to libraries or circulated within institutions and industrial companies. Membership subscriptions can only be ordered through the Euromicro Association. (Please contact the Treasurer to this end.)

SUBSCRIPTION ORDER FORM

To your usual supplier or to:
North-Holland Publishing Company
P.O. Box 211, 1000 AE Amsterdam,
The Netherlands

EUROMICRO JOURNAL

☐ Please note my order for an **institutional** subscription to Vol. 5, 1979: US $61.00/Dfl. 125.00 (including postage and handling)
☐ Please send me a specimen copy first
☐ Payment enclosed:

Name of library/institution/industrial company:

_____

Please forward to the attention of:

Mr./Ms _____

Address _____

_____

City _____

State/Country _____

Date _____ Signature

EUROMICRO MEMBERSHIP APPLICATION

To: **Dr. P. le Beux**
Treasurer, EUROMICRO ASSOCIATION
14-18 Rue Planchat
75020 Paris
France

Please enter my membership to the EUROMICRO ASSOCIATION for 1979 at FF 95,— I understand that my membership subscription to the EUROMICRO JOURNAL is meant for **personal** use only

Name _____

Occupation _____

Personal address _____

_____

_____

Country _____

Telephone _____

Date _____

Signature

Payment enclosed ☐ Invoice ☐

EUROMICRO SYMPOSIUM DOCUMENTATION

Large Scale Integration

Technology, Applications and Impacts

Proceedings of the Fourth EUROMICRO Symposium on Microprocessing and Microprogramming, Munich, Germany, October 17-19, 1978.

edited by **HAROLD W. LAWSON, Jr.,** *Linköping University, Sweden,* **HELMUT BERNDT,** *Siemens AG, Munich, Germany, and* **GUNNAR HERMANSON,** *Saab - Scania AB, Linköping, Sweden.*

1979 xiii + 380 pages
Price: US$48.75/Dfl. 100.00
ISBN 0-444-85249-2

The rapid pace of technological breakthroughs and developments of Large Scale Integration circuits has been described as one of the most significant developments since the beginning of the industrial revolution.
The purpose of the scientific program at EUROMICRO 78 was to present the state of the art of Large Scale Integration technology, its application and impacts on society with particular emphasis upon implications for the labour force.

As the leading European forum concerned with microprocessor systems and microprogramming, the EUROMICRO symposium attracted contributions from nearly all European countries, the United States and the Far East. The program contained a variety of sessions relating to the implications of inexpensive computer systems, special and general purpose computer architectures, memory systems and their utilization, reliability and testing development methods, programming languages, input/output control, interfacing and various applications.
A more detailed examination was made of three areas considered especially important - the social implications of microprocessors, microprocessor software and personal computing.
The papers presented in this book blend a superb scientific program with the social implications of current computer system technology and will be of interest to computer scientists in both research and industry.

ABBREVIATED CONTENTS: Keynote Session: Invited papers: Large scale integration utilization (Present and Future) *(A. Prommer).* System development methodology. *(S.S. Husson).* New directions in computer systems architecture *(E.I. Organick).* Impact of computer utilization upon the world's labor force *(K. Nygaard).* SESSIONS: Applications for the handicapped. Central processing unit architecture. Social implications of microprocessors. Signal processing. Fault detection and testing. Synchronization in multiprocessor systems. Programming languages. Memory structures and technologies. Reliability. Input/output control. Microprocessor software. Interfacing. Microprogramming languages. Applications. Multiprocessing. Sequential memory applications. System development methods. Computer architecture. Personal computing.

Microcomputer Architectures

Proceedings of the Third EUROMICRO Symposium on Microprocessing and Microprogramming, October 3-6, 1977, Amsterdam.

edited by **JEAN-DANIEL NICOUD,** *Swiss Federal Institute of Technology, Lausanne, Switzerland,* **JAN WILMINK,** *Technical University, Twente, The Netherlands, and* **RODNAY ZAKS,** *Sybex, Paris.*

1978 xiv + 280 pages
Price: US$44.00/Dfl. 90.00
ISBN 0-444-85097-X

The proceedings of the EUROMICRO 77 Conference cover the broad field of microcomputer architectures, including microprocessors, microprogramming, multiprocessor systems and software support. Applications range from computers embedded in consumer devices, to sophisticated fault-tolerant systems.

These proceedings consist of both theoretical contributions and application papers, and with each group of two or three papers, the introduction given by the Session Chairman is included. They provide a wealth of information for those concerned with keeping up-to-date with the latest developments in the theory and practice of original architectures and design techniques for microcomputers.
As the successor of the proceedings volume of the Second EUROMICRO Symposium, held in Venice in October 1976, this book illustrates the subsequent trends, now more particularly towards Micro*processor* architectures. The more extensive use of one-chip microprocessors, both as input-output controllers, and as part of a processing unit including many processors, is clearly demonstrated.

Microprocessing and Microprogramming

Proceedings of the Second EUROMICRO Symposium, Venice, Italy, October 12-14, 1976

edited by **J. WILMINK, M.G. SAMI** *and* **R. ZAKS**

1978 xx + 336 pages
Price: US$36.50/Dfl. 75.00
ISBN 0-7204-0557-2

The present work provides a state-of-the-art report, covering the full range of microprocessing and microprogramming. Amidst the wealth of papers included, the recent dramatic advent of LSI and MSI techniques is particularly well reflected in the papers on special custom-tailored hardware and software, which shows that the hardware-software balance is exploited deeper than ever before. In many papers the synchronization of multiprocessor systems is a main topic. Software support for microprocessor systems, the application of microprocessors in the fields of data transmission and signal processing and the increasing importance of micorprogramming, due to the present availability of user microprogrammable computer systems and bit slice processors, are only a selection of the subjects treated. This report, the substantially enlarged version of the Proceedings of the 2nd Symposium on Microprocessing and Microprogramming organized by EUROMICRO in Venice, October 1976, is required reading for anybody involved in any way in the fields of microprocessing and microprogramming.

Microarchitecture of Computer Systems

Proceedings of the First EUROMICRO Workshop, Nice, 23-25 June, 1975

edited by **R. HARTENSTEIN** *and* **R. ZAKS**

1975 x + 294 pages 20 tables 219 illus. over 300 lit. refs
Price: US$29.75/Dfl. 60.00
ISBN 0-7204-2842-4

The papers presented at this meeting provide an overview of the important new developments made in computer architecture and microarchitecture.

Particular emphasis was placed on the need to bridge the communication gap between the hardware and software engineering sectors, and between other specialized disciplines such as operating systems theory, computer organization, logic design, and design and application of LSI microprocessors. A number of papers also report on recent developments in the organization of micro-programmable hardware. Of primary interest to programmers designers, researchers, teachers and students of both hardware and software, the book is also intended for practitioners who are looking for new developments and trends to improve their own methods of design.

". . . certainly this set of papers is worth reading by people who are interested in microprocessors and microprogramming as a part of computer science in its truest sense or who are interested in machines based on LSI technology."

The Computer Bulletin

Fundamentals of Structured Hardware Design

A Design Language Approach at Register Transfer Level

by **REINER W. HARTENSTEIN,** *University of Karlsruhe, Germany.*

1977 xvi + 324 pages
Price: US $41.50/Dfl. 85.00
ISBN 0-444-85007-4

This book introduces an easily readable register transfer level hardware design language, and explains its applications in the description of digital computer structures, subsystems, and integrated circuit components. In parallel, it also introduces a block diagram language which is directly mapped into the symbolic version of the language.

The second part of the book after discussing elementary use of the language, develops advanced methods of writing structured representations of complex data paths, and wiring patterns. It also examines modular design and the description of complex structures at different levels of abstraction. Finally, the book describes a very concise and highly user-oriented behavioral specification of digital integrated circuits and of LSI circuits.
This specific circuit replaces conventional conglomerates of more or less formal notations, used for IC circuits.

ORDER FORM

Send to your usual supplier or

In the U.S.A. and Canada:
ELSEVIER NORTH-HOLLAND, INC.
52 Vanderbilt Ave., New York,
N.Y. 10017

In Australia:
D.A. BOOK DEPOT PTY. LTD
11-13 Station Street, Mitcham,
Vic. 3132

In all other countries:
NORTH-HOLLAND PUBLISHING COMPANY
P.O. Box 211, 1000 AE Amsterdam
The Netherlands

Please send me the following book(s):

☐ Large Scale Integration (Lawson/Berndt/Hermanson)	US $48.75/Dfl.	100.00
☐ Microcomputer Architectures (Nicoud/Wilmink/Zaks)	US $44.00/Dfl.	90.00
☐ Microprocessing and Microprogramming (Wilmink/Sami/Zaks)	US $36.50/Dfl.	75.00
☐ Microarchitecture of Computer Systems (Hartenstein/Zaks)	US $29.25/Dfl.	60.00
☐ Fundamentals of Structured Hardware Design (Hartenstein)	US $41.50/Dfl.	85.00

Orders from individuals must be accompanied by a remittance following which books will be supplied postfree.

I enclose

my personal cheque	bank draft	UNESCO coupons

Name: _____

Address: _____

_____

_____

_____

Date: _____

Signature: